



# MACHINE LEARNING FOR PEPTIDE STRUCTURE, FUNCTION, AND DESIGN

EDITED BY: Ruiquan Ge, Chuan Dong, Juexin Wang and Yanjie Wei  
PUBLISHED IN: Frontiers in Genetics



# frontiers

## Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence.

The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714

ISBN 978-2-83250-395-9

DOI 10.3389/978-2-83250-395-9

## About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: [frontiersin.org/about/contact](https://frontiersin.org/about/contact)

# MACHINE LEARNING FOR PEPTIDE STRUCTURE, FUNCTION, AND DESIGN

Topic Editors:

**Ruiquan Ge**, Hangzhou Dianzi University, China

**Chuan Dong**, Wuhan University, China

**Juexin Wang**, Indiana University: Purdue University Indianapolis, United States

**Yanjie Wei**, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences (CAS), China

**Citation:** Ge, R., Dong, C., Wang, J., Wei, Y., eds. (2022). Machine Learning for Peptide Structure, Function, and Design. Lausanne: Frontiers Media SA.  
doi: 10.3389/978-2-83250-395-9

# Table of Contents

- 05 Editorial: Machine Learning for Peptide Structure, Function, and Design**  
Ruiquan Ge, Chuan Dong, Juexin Wang and Yanjie Wei
- 08 Network Analyses Based on Machine Learning Methods to Quantify Effects of Peptide–Protein Complexes as Drug Targets Using Cinnamon in Cardiovascular Diseases and Metabolic Syndrome as a Case Study**  
Yingying Wang, Lili Wang, Yinhe Liu, Keshen Li and Honglei Zhao
- 19 Systematic Modeling, Prediction, and Comparison of Domain–Peptide Affinities: Does it Work Effectively With the Peptide QSAR Methodology?**  
Qian Liu, Jing Lin, Li Wen, Shaozhou Wang, Peng Zhou, Li Mei and Shuyong Shang
- 29 SSH2.0: A Better Tool for Predicting the Hydrophobic Interaction Risk of Monoclonal Antibody**  
Yuwei Zhou, Shiyang Xie, Yue Yang, Lixu Jiang, Siqi Liu, Wei Li, Hamza Bukari Abagna, Lin Ning and Jian Huang
- 38 i2APP: A Two-Step Machine Learning Framework For Antiparasitic Peptides Identification**  
Minchao Jiang, Renfeng Zhang, Yixiao Xia, Gangyong Jia, Yuyu Yin, Pu Wang, Jian Wu and Ruiquan Ge
- 47 Refined Contact Map Prediction of Peptides Based on GCN and ResNet**  
Jiawei Gu, Tianhao Zhang, Chunguo Wu, Yanchun Liang and Xiaohu Shi
- 58 Ensemble-AHTPpred: A Robust Ensemble Machine Learning Model Integrated With a New Composite Feature for Identifying Antihypertensive Peptides**  
Supatcha Lertampaiporn, Apiradee Hongsthong, Warin Wattanapornprom and Chinae Thammarongtham
- 74 Inter-Residue Distance Prediction From Duet Deep Learning Models**  
Huiling Zhang, Ying Huang, Zhendong Bei, Zhen Ju, Jintao Meng, Min Hao, Jingjing Zhang, Haiping Zhang and Wenhui Xi
- 88 BBPpredict: A Web Service for Identifying Blood-Brain Barrier Penetrating Peptides**  
Xue Chen, Qian Yue Zhang, Bowen Li, Chunying Lu, Shanshan Yang, Jinjin Long, Bifang He, Heng Chen and Jian Huang
- 98 ProtTrans-Glutar: Incorporating Features From Pre-trained Transformer-Based Models for Predicting Glutarylation Sites**  
Fatma Indriani, Kunti Robiatul Mahmudah, Bedy Purnama and Kenji Satou
- 109 DTI-BERT: Identifying Drug-Target Interactions in Cellular Networking Based on BERT and Deep Learning Method**  
Jie Zheng, Xuan Xiao and Wang-Ren Qiu



**121 Profiling a Community-Specific Function Landscape for Bacterial Peptides Through Protein-Level Meta-Assembly and Machine Learning**

Mitra Vajjala, Brady Johnson, Lauren Kasperek, Michael Leuze and Qiuming Yao

**132 AttnTAP: A Dual-input Framework Incorporating the Attention Mechanism for Accurately Predicting TCR-peptide Binding**

Ying Xu, Xinyang Qian, Yao Tong, Fan Li, Ke Wang, Xuanping Zhang, Tao Liu and Jiayin Wang



## OPEN ACCESS

EDITED AND REVIEWED BY  
Richard D. Emes,  
University of Nottingham,  
United Kingdom

\*CORRESPONDENCE  
Ruiquan Ge,  
gespring@hdu.edu.cn

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

RECEIVED 30 July 2022  
ACCEPTED 17 August 2022  
PUBLISHED 20 September 2022

CITATION  
Ge R, Dong C, Wang J and Wei Y (2022),  
Editorial: Machine learning for peptide  
structure, function, and design.  
*Front. Genet.* 13:1007635.  
doi: 10.3389/fgene.2022.1007635

COPYRIGHT  
© 2022 Ge, Dong, Wang and Wei. This is  
an open-access article distributed  
under the terms of the [Creative  
Commons Attribution License \(CC BY\)](#).  
The use, distribution or reproduction in  
other forums is permitted, provided the  
original author(s) and the copyright  
owner(s) are credited and that the  
original publication in this journal is  
cited, in accordance with accepted  
academic practice. No use, distribution  
or reproduction is permitted which does  
not comply with these terms.

# Editorial: Machine learning for peptide structure, function, and design

Ruiquan Ge<sup>1,2\*</sup>, Chuan Dong<sup>3</sup>, Juexin Wang<sup>4</sup> and Yanjie Wei<sup>5</sup>

<sup>1</sup>School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China, <sup>2</sup>Hangzhou Institute of Advanced Technology, Hangzhou, China, <sup>3</sup>Key Laboratory of Combinatorial Biosynthesis and Drug Discovery, Ministry of Education, and School of Pharmaceutical Sciences, Wuhan University, Wuhan, China, <sup>4</sup>Department of BioHealth Informatics, Indiana University Purdue University Indianapolis, Indianapolis, IN, United States, <sup>5</sup>Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

## KEYWORDS

machine learning, functional peptides, deep learning, drug design, peptide therapeutics

## Editorial on the Research Topic

### Machine learning for peptide structure, function, and design

Peptides with a length from 2 to 50 amino acids play important roles in the biological process and functions. Because of their wonderful variety of biological properties, peptide-based therapy has been a potential treatment for many diseases for decades. Meanwhile, peptide sequence, structure, and function are closely related, especially the relationship between structure and function. However, obtaining the structure or function of the peptides with wet experiments is costly, laborious, and time-consuming. In recent years, because of the obvious advantages of traditional machine learning and deep learning technology, these methods have been widely used in various protein or peptide structure and function predictions such as many kinds of site prediction, various interactions prediction, drug-targets prediction, and so on.

This Research Topic explores the new technologies and applications of machine learning on peptide structure and function prediction. We are pleased to see that the authors of the 12 accepted papers introduce the research progress and application of the latest machine learning techniques in peptide-related problems. They are related to peptide treatment of disease, inter-residue distance or contact, binding site prediction, drug targets, community-specific function landscape for peptides, and related discussions.

Peptide-based therapy has become a new potential method of disease treatment in recent decades years. Compared with traditional disease treatments, such as radiation therapy and chemotherapy, therapeutic peptides could avoid the obvious side effects of traditional disease treatments to guarantee precise treatment. Furthermore, most therapeutic peptides have the characteristics of high specificity, low production cost, low toxicity, easy synthesis, modification, etc. In this topic, three tools are

proposed to predict therapeutic peptides, which are blood-brain barrier penetrating peptides (BBPpredict), antihypertensive peptides (Ensemble-AHTPpred), antiparasitic peptides (i2APP). Comparing with the experimental results of nine classifiers on the five-fold cross-validation and independent testing datasets, [Chen et al.](#) (BBPpredict) use a random forest method with optimal features selected by three feature scoring methods to predict the blood-brain barrier penetrating peptides (BBPs). In addition, they construct an online web service of BBPpredict to help researchers predict and find novel BBPs to accelerate the development of new drugs to treat central nervous system (CNS) diseases. Furthermore, [Lertampaiporn et al.](#) propose a robust ensemble machine learning model to identify antihypertensive peptides (Ensemble-AHTPpred). Ensemble-AHTPpred integrates various computed features and optimally weighted classifiers to improve the performance of the model. Moreover, i2APP proposed by [Jiang et al.](#) employs a two-step machine learning framework to identify antiparasitic peptides (APPs). It utilizes multi-feature extraction, feature selection with maximum information coefficient, and random down-sampling technology to improve the performance of models to identify APPs efficiently.

In addition, six papers pay attention to the inter-residue relationship, interaction, and binding sites. [Zhang et al.](#) propose DueDis to predict the inter-residue distance with duet deep learning models. DuetDis use the 1D and 2D complementary feature sets and high-quality multiple sequence alignment (MSA) to improve the prediction performance in the fused features. Peptide inter-residue contact maps determine its topological structure. [Gu et al.](#) utilize graph convolutional neural networks (GCN) and two different dimensional residual neural network architectures (1D ResNet and 2D ResNet) to capture global and local information, respectively. The compared experiments demonstrate its effectiveness on four different test datasets exceptionally on the long-range contact types. Furthermore, drug–target interactions (DTIs) are a hot topic in new drug discovery. [Zheng et al.](#) develop DTI-BERT to predict DTIs based on pre-trained Bidirectional Encoder Representations from Transformers (BERT) and deep learning methods. In the DTI-BERT model, sequence features are extracted by the pre-trained BERT for the proteins. And drug information is generated by Discrete Wavelet Transform (DWT) from drug molecular fingerprints. Then, a deep learning network is employed to judge the interaction using contrastive loss and cross-entropy loss in a few target families. In addition, [Zhou et al.](#) develop SSH2.0 to predict the hydrophobic interaction risk of monoclonal antibodies. SSH2.0 trains a new support vector machine-based ensemble model with the selected CKSAAGP features. Compared to the previous SSH, SSH2.0 performs

better and may be a good web tool for researchers. In addition, protein post-translational modifications (PTMs) play crucial roles in diverse biological processes, affecting the protein's function. Nowadays, various computational tools are developed to identify disease-associated PTM sites. In this issue, [Indriani et al.](#) propose ProtTrans-Glutar model to predict whether a protein sequence includes a glutarylation site. ProtTrans-Glutar extracts several kinds of feature sets such as the distribution feature, enhanced amino acid composition (EAAC), and ProtT5-XL-UniRef50, a pre-trained transformer-based model. Meanwhile, random under-sampling and XGBoost classifiers are used to train the model. Besides, [Xu et al.](#) propose AttnTAP to predict the binding of T cell receptor (TCR) and peptide with a dual-input deep learning framework to precisely predict the TCR-peptide binding. For AttnTAP, a bi-directional long short-term memory model (BiLSTM) model and attention mechanism with different weights for amino acids are employed to predict TCR-peptide binding effectively.

The remaining three articles analyze and discuss peptide-related problems from a relatively broad perspective. [Vajjala et al.](#) develop a metaBP toolkit to construct a community-specific function landscape for bacterial peptides from metagenomic samples. The toolkit metaBP and metaBP-ML can discover and annotate bacterial peptides from a natural microbial community. It may give us a new research perspective to better understand the characteristics of bacterial peptides. For another research work, [Liu et al.](#) reveal and verify that traditional peptide quantitative structure-activity relationship (pQSAR) strategies only model the genome-wide domain–peptide interaction (DPI) qualitatively or semi-quantitatively because of disordered peptide conformation and potential interactions between peptide residues. For the last work, [Wang et al.](#) design a three-step pipeline to discover drug targets using cinnamon in cardiovascular diseases and metabolic syndrome. Through pathway filter, combined network construction, and biomarker prediction and validation to quantitative analysis of the effects of peptide-protein complexes as drug targets, 17 peptide-protein complexes are identified as the cinnamon targets in 6 peptides and 4 proteins. The pipeline based on network analyses using machine learning may foster new drug discovery based on peptides.

In conclusion, this special issue involves several hot topics in solving peptide-related problems using currently popular machine learning techniques. These efforts will help accelerate the development of vaccines and new drugs. Additionally, we hope these works can attract more researchers to focus on the related fields. Moreover, we thank all the reviewers and authors for their efforts and contributions to this special issue.

## Author contributions

RG wrote the manuscript draft. CD, JW, and YW helped to review and edit the paper. All authors have approved the final version of the editorial.

## Funding

This work is supported by the Zhejiang Provincial Natural Science Foundation of China (No. LY21F020017), National Natural Science Foundation of China (No. 61702146).

## Conflict of interest

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

## Publisher's note

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



# Network Analyses Based on Machine Learning Methods to Quantify Effects of Peptide–Protein Complexes as Drug Targets Using Cinnamon in Cardiovascular Diseases and Metabolic Syndrome as a Case Study

Yingying Wang<sup>1,2</sup>, Lili Wang<sup>3</sup>, Yinhe Liu<sup>3</sup>, Keshen Li<sup>1,2\*</sup> and Honglei Zhao<sup>3\*</sup>

<sup>1</sup>Department of Neurology and Stroke Center, The First Affiliated Hospital of Jinan University, Guangzhou, China, <sup>2</sup>Clinical Neuroscience Institute, The First Affiliated Hospital of Jinan University, Guangzhou, China, <sup>3</sup>Fuwai Hospital Chinese Academy of Medical Sciences, Shenzhen, China

## OPEN ACCESS

### Edited by:

Ruiquan Ge,  
Hangzhou Dianzi University, China

### Reviewed by:

Shihua Zhang,  
Wuhan University of Science and  
Technology, China  
Lei Liu,  
Harbin Medical University, China

### \*Correspondence:

Keshen Li  
likeshen1971@126.com  
Honglei Zhao  
939924240@qq.com

### Specialty section:

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

**Received:** 16 November 2021

**Accepted:** 07 December 2021

**Published:** 24 December 2021

### Citation:

Wang Y, Wang L, Liu Y, Li K and  
Zhao H (2021) Network Analyses  
Based on Machine Learning Methods  
to Quantify Effects of Peptide–Protein  
Complexes as Drug Targets Using  
Cinnamon in Cardiovascular Diseases  
and Metabolic Syndrome as a  
Case Study.  
Front. Genet. 12:816131.  
doi: 10.3389/fgene.2021.816131

Peptide–protein complexes play important roles in multiple diseases such as cardiovascular diseases (CVDs) and metabolic syndrome (MetS). The peptides may be the key molecules in the designing of inhibitors or drug targets. Many Chinese traditional drugs are shown to play various roles in different diseases, and comprehensive analyses should be performed using networks which could offer more information than results generated from a single level. In this study, a network analysis pipeline was designed based on machine learning methods to quantify the effects of peptide–protein complexes as drug targets. Three steps, namely, pathway filter, combined network construction, and biomarker prediction and validation based on peptides, were performed using cinnamon (CA) in CVDs and MetS as a case. Results showed that 17 peptide–protein complexes including six peptides and four proteins were identified as CA targets. The expressions of AKT1, AKT2, and ENOS were tested using qRT-PCR in a mouse model that was constructed. AKT2 was shown to be a CA-indicating biomarker, while E2F1 and ENOS were CA treatment targets. AKT1 was considered a diabetic responsive biomarker because it was down-regulated in diabetic but not related to CA. Taken together, the pipeline could identify new drug targets based on biological function analyses. This may provide a deep understanding of the drugs' roles in different diseases which may foster the development of peptide–protein complex–based therapeutic approaches.

**Keywords:** peptide–protein complexes, network analyses, metabolic syndrome, cardiovascular, cinnamon

## INTRODUCTION

Peptide–protein complexes are the key components of protein–protein interaction (PPI) networks. Nearly 15–40% PPIs are mediated by these short linear peptides (Neduvu et al., 2005). The peptide–protein complexes are proven to play important roles predominantly in both signaling and regulatory pathways, implicating that the peptides are involved in many human diseases

(Pawson and Nash, 2003). As a result, the peptides are attracting more attention in drug research fields since they may be the key molecules in the designing of inhibitors or drug targets (Parthasarathi et al., 2008; London et al., 2010).

Due to the characters of peptide-protein interactions, it is reasonable to perform network analyses based on machine learning methods since the relationships between those peptides and proteins could be illustrated clearly in the form of graphs (Zhu et al., 2020; Ji et al., 2021; Yingying et al., 2021). Similarly, some complex diseases are found to be similar based on network analyses, indicating that more relationships between different diseases could be predicted using bioinformatics pipelines (Wang et al., 2019). Many Chinese traditional drugs are shown to play various roles in different diseases, and comprehensive analyses should be performed using networks which may offer more information than results generated from a single level. However, no pipeline aiming to predict the peptide-protein complex as drug targets in different diseases had been proposed. In this study, a network analysis pipeline was designed based on machine learning methods to quantify the effects of peptide-protein complexes as drug targets.

In this pipeline, diseases with at least 20 related genes and drugs with at least one related biological functional term could be used as analysis objects. Diseases that are similar to each other on at least one level (such as medical or biological level) are recommended. The candidate drugs do not need to be proven useful in the diseases analyzed since predicting new roles of the candidate drugs is also one application of the pipeline. Based on the abovementioned concerns, two types of diseases (cardiovascular diseases (CVDs) and metabolic syndrome (MetS)) and a Chinese traditional drug (cinnamon) as a case were chosen.

Cinnamon (*Cinnamomum zeylanicum* and Cinnamon cassia, CA) is one of the most important spices used daily (Hariri and Ghiasvand, 2016). Cinnamaldehyde is one of the main resinous ingredients found in CA, which is commonly used as a Chinese medicine for blood circulation disturbance and inflammation (Sheng et al., 2008; Cao et al., 2010; Yang et al., 2015). It was shown that cinnamaldehyde played important roles in both CVDs and MetS (patients suffering from type 2 diabetes (T2D), and glucose/insulin metabolism disturbance or insulin resistance, and was involved with at least two of the following four items: hypertension, dyslipidemia, obesity, and microalbuminuria defined by the WHO criteria) (Mollazadeh and Hosseinzadeh, 2016). CVDs and MetS are not independent since MetS is one of the most undeniable reasons of CVDs. Besides, there are multiple types of biomarkers identified as common features of CVDs and MetS, such as non-coding RNAs, proteins, and metabolites (Das et al., 2020).

It is of great importance to explore the mechanism of CA since this drug could participate in both of the disease types at the same time (Sheng et al., 2008; Yang et al., 2015). One possible reason may be its antidiabetic action by modulating the insulin and insulin-like growth factor (IGF1) signaling pathways (Schriner et al., 2014) since insulin resistance was

proven to play a fundamental key role for MetS complications (Khan et al., 1990). Besides, CA was shown to retard the progression of cardiac hypertrophy and fibrosis *via* blocking the ERK signaling pathway (Zhang et al., 2015; Xiao et al., 2017). However, functional analysis for CA in a system way, especially based on biological pathways, is still lacking.

As an integration of molecular interaction; genetic, cellular, and environmental information processing; and metabolism reactions, biological pathways are often used in systematic analyses of complex diseases such as CVDs, T2D, and cancers (Salt and Hardie, 2017; Kakiuchi-Kiyota et al., 2019; Kaku, 2019). Peptide-protein complexes were also proven to be the key components in pathways. It was postulated that there may be associations between the common pathways shared by CVD/MetS and CA which could be detected based on peptide-protein complex analyses. In this study, a new network analysis pipeline was proposed based on machine learning methods to identify common drug targets in different diseases.

## MATERIALS AND METHODS

The analyses were performed using the following three steps: (as shown in **Figure 1**).

Step1: Pathway filter. The similarity between any two selected diseases was calculated and used to filter the disease pairs. Enrichment analyses were performed for the related genes of the disease pairs. Meanwhile, the CA-related pathways were found through literature searching. Common pathways were then filtered and used as the inputs for step 2.

Step2: Combined network construction. All the common pathways were then converted into networks. The network structure similarities were calculated using two types of machine learning methods, and an integrate score was designed to measure the similarity between any two common pathways on the structural level in order to explore the potential correlations of these pathways. The pathways were then merged into a combined pathway network. Proteins in the pathways were merged into a combined protein network.

Step3: Biomarker prediction and validation based on peptides. The nodes in the combined protein network were first ranked according to the network topological characters. Then protein-peptide complexes containing these top proteins as receptors were selected, and the peptides were then clustered. The top genes with peptides clustered into the same clusters were selected as candidate biomarkers and validated using qRT-PCR in a mouse diabetic model that was constructed.

## Disease Similarity Calculation

Methods that can calculate the distances between any two diseases based on any biological or medical level could be used. In this case, a module-based method (Menche et al., 2015) was used. The similarity  $S_{ij}$  between two diseases  $i$  and  $j$  was calculated as follows:

$$S_{ij} \equiv \langle d_{ij} \rangle - \frac{\langle d_{ii} \rangle + \langle d_{jj} \rangle}{2}.$$

Of which,  $\langle d_{ii} \rangle$  and  $\langle d_{jj} \rangle$  represented the average shortest distances inside diseases  $i$  and  $j$ , respectively, while  $\langle d_{ij} \rangle$  represented the pairwise average shortest distance between disease  $i$  and  $j$ . The shortest distances were calculated for any protein pairs inside/between diseases using the relationships integrated from multiple molecular interaction levels including protein, regulatory, and metabolic pathways, and kinase substrate.

A z-score was calculated based on the random control networks by 1000 permutations of disease lists preserving randomization. A  $p$ -value for each  $S_{ij}$  score was calculated using the Mann-Whitney U test. Then FDR was used to obtain the  $q$ -values.

### Information Converting From Genes to Biological Pathways

The information conversions from genes to biological pathways were performed using the DAVID EASE score (Huang et al., 2009a; Huang et al., 2009b), which was a modified Fisher exact  $p$ -value. For any disease-related gene list  $l_i$  and biological pathway  $w_a$ , the EASE score was calculated as follows:

$$e(l_i, w_a) = 1 - \sum_{i=0}^{GH-1} \frac{\binom{OH}{i} \binom{OT-OH}{GT-i}}{\binom{OT}{GT-1}}.$$

Of which, the calculation methods of GH (gene hits), GT (gene total), OH (genome hits), and OT (genome total) are shown in the following 2\*2 table:

An e-value not above the threshold supported the alternative hypothesis that the probability of the first cell in the 2\*2 table was actually greater than that expected under the null hypothesis that the two variables were independent. The conclusion was that there was an association between the row and the column variables in the table, which meant the proportions of those genes falling into each category were different among groups.

### From Biological Pathways to Graphs

The information conversions from biological pathways to protein-protein networks were performed using the R package “graphite” (Sales et al., 2012). The algorithm in this package kept the information of protein complexes, gene families, and

removing chemical compounds from the final graphs, which was especially important in the peptide complex analyses of this study.

### Network Structure Similarity Calculation

The network structure similarity calculation algorithms could be divided into two types: alignment-free and alignment-based network comparison (Frigo et al., 2021). In this pipeline, it was recommended to use at least one alignment-free algorithm and one alignment-based algorithm to compare the different networks and combine the scores together.

### Alignment-Free Algorithm Based on Graphlet Degree Distribution Agreement

The alignment-free network comparison algorithms performed the network similarity analyses by quantifying the overall topological similarity between networks, irrespective of node mappings between the networks, and without any conserved edges or subgraph identification. In this pipeline, the algorithm named GDD agreement was chosen, which performed the structural similarity (SS) between networks based on the graphlet degree distribution as follows (Przulj, 2007):

The similarity between any two networks  $G$  and  $W$  was calculated as follows:

$$S_{GDD}(G, W) = \frac{1}{n} \sum_{j=0}^{n-1} S_{GDD}^j(G, W).$$

Of which,

$$S_{GDD}^j(G, W) = 1 - \left( \sum_{k=1}^{\infty} \left[ \frac{d_G^j(k)}{k} / \sum_{k=1}^{\infty} \frac{d_G^j(k)}{k} - \frac{d_W^j(k)}{k} / \sum_{k=1}^{\infty} \frac{d_W^j(k)}{k} \right]^2 \right)^{\frac{1}{2}}.$$

Of which,  $d_G^j(k)$  is the sample distribution of the number of nodes in network  $G$  touching the appropriate graphlet  $k$  times. The range of  $S_{GDD}$  is  $[0,1]$ ; a higher score meant the two networks compared were more similar to each other.

### Alignment-Based Algorithm Based on the Hungarian Method

The alignment-based network comparison methods referred to a series of algorithms aiming to find a mapping between the nodes of at least two networks that preserved edges and a large subgraph between the networks. In this pipeline, an alignment-based algorithm was chosen based on a Hungarian method as follows:



The network alignment scores, that is,  $S_{AE}(G, W)$  (between any two networks  $G$  and  $W$ ), were performed using the Hungarian method (Kuhn, 1955) on a square distance matrix  $C$  (if the sizes of the two networks were different, the larger number of nodes was used), which was calculated as follows:

$$C_{ab} = \sqrt{\sum_{t \in T} (M_{G_{a,t}} - M_{W_{b,t}})^2}.$$

Of which,  $M_{G_{a,t}} = \frac{-\sum_{j=1}^{N_G} A_{G_{a,j}}^{(t)} \ln(A_{G_{a,j}}^{(t)})}{H(I^{N_G})}$ , where  $A_{G_{a,j}}^{(t)}$  is a transition matrix of network  $G$ , which was constructed by converting the raw square transition matrix into Markov processes by normalizing each row sum to unity.  $A_G^{(t)}$  contained probabilities of edges transferring information from the  $i$ th to the  $j$ th member of the system in exactly  $t$  units of time. For  $t \in \{2^y\} \forall y \in \mathbb{N}$ ,  $t_{\max} \geq 2D$  and  $t_{\max-1} < 2D$ , where  $D$  is the max diameter of the two networks  $G$  and  $W$  being compared, and the R packages “igraph” and “netcom” were used to perform the calculation.

## Integrated Network Similarity Score

The integrated network similarity scores between the two networks  $G$  and  $W$  were calculated as follows:

$$S(G, W) = S_{GDD}(G, W) + S_{AE}(G, W).$$

A higher  $S$  score indicated that the two networks compared were more similar to each other on the structural level.

## Biomarker Prediction and Validation

The prediction and validation of the biomarkers were performed using the following steps:

- 1) The proteins in the combined protein network were ranked according to the network topological characters. For each node, degree and node betweenness were calculated. The edge betweenness was calculated for each edge using the R package “igraph.”
- 2) The protein–peptide complexes containing these top proteins as receptors were selected, and the peptides were then clustered. The high-resolution structures of protein–peptide complexes containing genes in the combined network as receptors were downloaded from the Protein Data Bank (PDB).
- 3) The top proteins with peptides clustered into the same clusters were selected as candidate biomarkers. The peptide sequences of these complexes were then classified using Hammock (1.2.0) (Krejci et al., 2016), which used hidden Markov model profiles for peptide sequence clustering. The consensus sequence for each cluster was generated using ClustalW (Thompson et al., 1994) and WebLogo (Crooks et al., 2004).
- 4) The candidate biomarkers were validated using qRT-PCR in a mouse diabetic model constructed as follows:

Fifty-nine male C57 mice (14–16g/28–35 days) were purchased from Guangdong Medical Experimental Animal

Center (Certificate No.: 44007200062167, License No.: scxk (Guangdong) 2018-0002, SPF clean grade).

The mice were divided into four groups as follows: 1) Group A (Control + vehicle): 5 mice were given solvent control (0.5% carboxymethyl cellulose solution (CMC)) by gavage; 2) Group B (Control + CA): 6 mice were given CA by gavage (the dose was 20 mg/kg/BW); 3) Group C (T2D + vehicle): 24 diabetic mice were given solvent by gavage; 4) Group D (T2D + CA): 24 diabetic mice were given CA by gavage.

Of which, the models of 48 diabetic mice were constructed using streptozotocin (STZ) using the following steps: 1) pretreatment: all the mice were made to starve 12 h before modeling; 2) model construction: STZ was intraperitoneally injected at a dose of 150 mg/kg/BW; 3) model test: the blood glucose value was measured continuously after 3 days of STZ injection. If the random blood glucose was  $>16.7$  mmol/L, the model was considered successful. Otherwise, another injection of STZ was administered until the random blood glucose was  $>16.7$  mmol/L.

Drug treatment (Groups C and D) was started 5 weeks after modeling. After 7 weeks of administration, all animals were killed, and the hearts of mice were treated with TRIzol and stored at  $-80^\circ\text{C}$ . Then qRT-PCR was performed for the candidate genes (the top proteins were mapped to their coding genes). The animal experiment was approved and recognized by the experimental Animal Ethics Committee of Shenzhen Sun Yat sen Cardiovascular Hospital (Approval No.: rye2019102806).

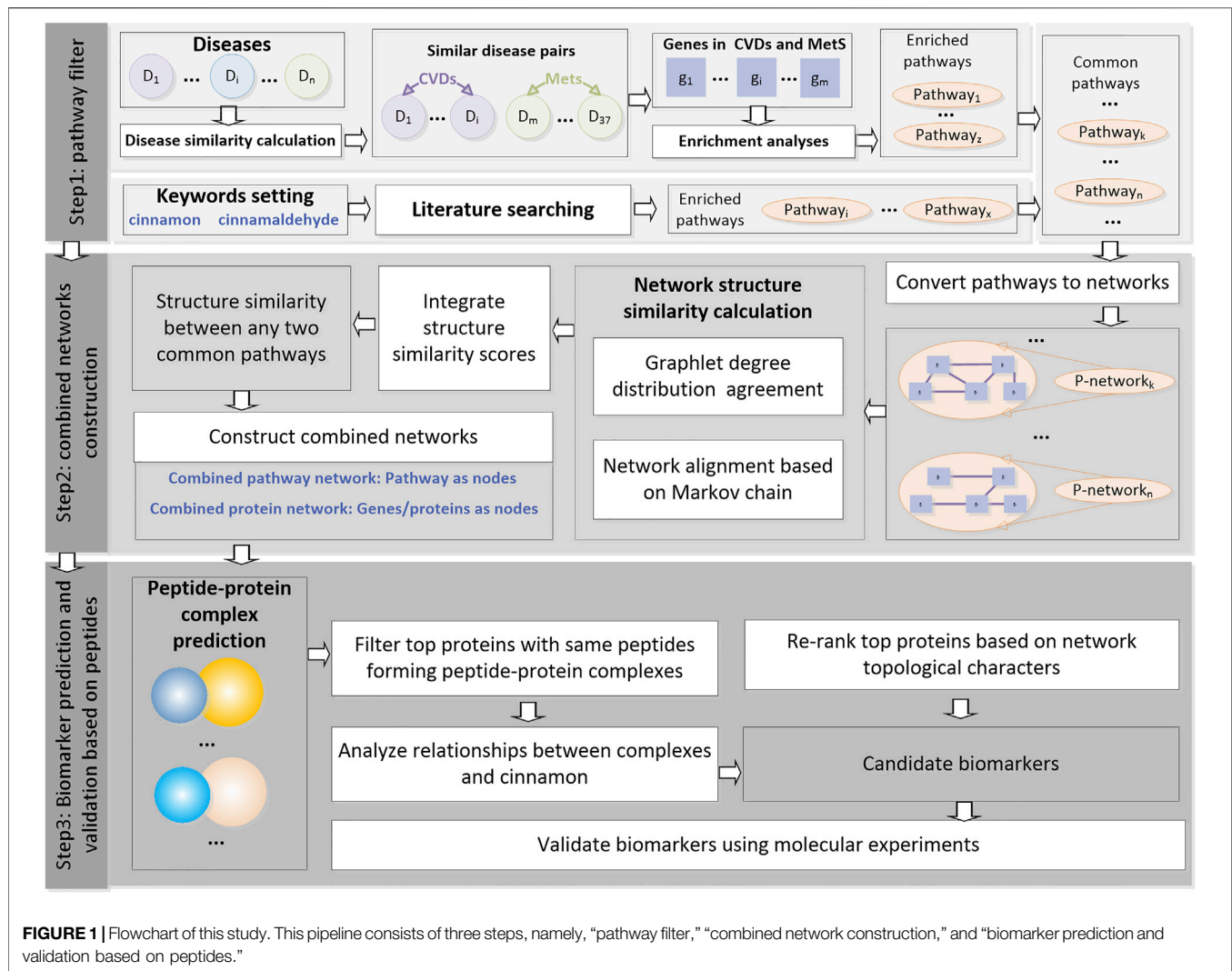
## RESULTS

### Pathway Filtrations

CVD (such as coronary disease) and MetS (such as diabetes mellitus) lists were extracted from Medical Subject Heading (MeSH) ontology with at least 20 disease-related genes from either OMIM or GWAS (listed in **Supplementary Table S1**). 553 disease pairs were shown to be similar with each other with a  $z$ -score  $\geq 1.6$  and  $q$ -value  $\leq 0.001$ . The 19 CVDs and 18 MetS comprising the 553 disease pairs were selected as HM (HeartMetS) datasets. As shown in **Figure 2A**, the average numbers of genes related to MetS (179.0556) were 2-fold of CVDs (86.42105). This indicated that MetS may be more complex than CVDs since these diseases involve the abnormality of multiple systems, such as endocrine, digestive, and immune systems.

The gene lists of each disease were then used as inputs of the information converting calculation. 179 pathways in KEGG (Kanehisa et al., 2017) and Biocarta with at least one  $e$ -value not above 0.05 were selected as HM-enriched pathways (see **Supplementary Table S2**). The common pathways in the two databases were named using KEGG ID. Otherwise, if there exists any difference between the two pathways, both of the pathways were kept.

CA was shown to play important roles through biological pathways in reducing metabolic syndrome complications and CVDs as reviewed in the former research (Yang et al., 2015;



Mollazadeh and Hosseinzadeh, 2016). The words “cinnamon” and “cinnamaldehyde” were used for literature searching through the NCBI PubMed to find the related pathways since 32 CA-related pathways were selected and binned into two groups according to their effects on diseases types: antidiabetic (including 28 pathways) and antihypertensive (including four pathways).

## Combined Pathway Network Analyses

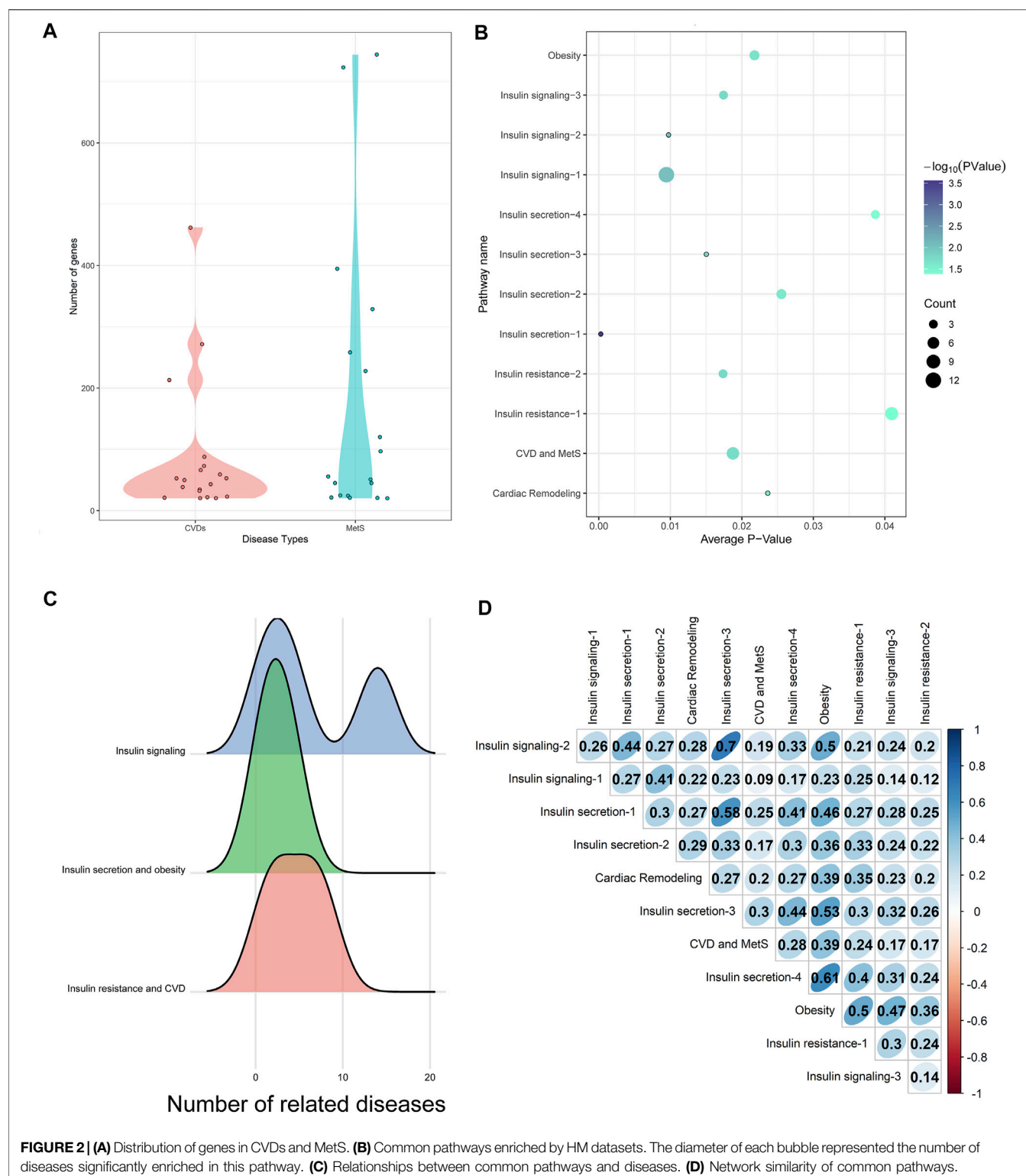
As shown in **Figure 2B**, there were 12 common pathways between the 179 HM-enriched pathways and 32 CA-related pathways. This indicated the dynamical roles CA played in different diseases or disease stages, including diabetes mellitus, obesity, MetS, and CVDs. The combined pathway network was built using the 12 common pathways as nodes. Sixty-seven connections were built if the two pathways shared at least one gene/protein.

The 12 common pathways were enriched by different numbers of diseases in the HM datasets. Of which, the insulin signaling pathway (hsa04910, marked as “Insulin

signaling-1”) was enriched by 12 diseases, while the following three pathways were only enriched by one disease: IL-2 receptor beta chain in T-cell activation (h\_il2rbPathway, marked as “Cardiac Remodeling”), the IGF-1 receptor and longevity (h\_longevityPathway, marked as “Insulin secretion-1”), multiple antiapoptotic pathways from IGF-1r signaling lead to bad phosphorylation (h\_igf1rPathway, marked as “Insulin secretion-3”), and sprouty regulation of tyrosine kinase signals (hsa04911, marked as “Insulin signaling-2”).

The 12 common pathways could be divided into three types according to their contributions to CVD and MetS (as shown in **Figure 2C**).

- 1) Insulin signaling: CA could enhance the insulin signaling pathway in the skeletal muscle by increasing the tyrosine phosphorylation level (Qin et al., 2003). Three pathways were involved in this stage, including the insulin signaling pathway (enriched by MetS and CVDs), tyrosine metabolism (enriched by MetS), and sprout regulation of



**FIGURE 2 | (A)** Distribution of genes in CVDs and MetS. **(B)** Common pathways enriched by HM datasets. The diameter of each bubble represented the number of diseases significantly enriched in this pathway. **(C)** Relationships between common pathways and diseases. **(D)** Network similarity of common pathways.

tyrosine kinase signals (enriched by CVD). It was interesting to see that the “insulin signaling pathway” was closely connected not only to MetS such as diabetes mellitus but also to CVDs, such as heart diseases. This may be explained by the fact that insulin signaling was an integral pathway

regulating the life span of laboratory organisms (Schriner et al., 2014).

- 2) Insulin secretion and obesity: Since impaired insulin secretion was one of the pathophysiological abnormalities in type 2 diabetes, IGF (insulin-like growth factors)-I, which was

**TABLE 1** | Pathway similarity results of different pathways.

Type of pathway sets	Number of pathways	$S_{GDD}$	$S_{AE}$
Common	12	0.257095099	0.047247541
CA-related	32	0.355600743	0.091677581
HM-related	179	0.324632623	0.061221968

**TABLE 2** | List of top 10 nodes and edges in the combined protein network.

Topological character	Protein symbols/protein-protein pairs
Degree	IRS1, MAOA, MAOB, AMPK1, AMPK2, PRKAB1, PRKAB2, PRKAG1, PRKAG2, and PRKAG3
Node betweenness	IRS1, OGT, AKT2, INS, AKT1, RAPGEF4, INSR, PDE3B, PTPA, and GNAS
Edge betweenness	PTPA-AKT2, IRS1-IGF1R, PPARGC1A-OGT, AKT1-E2F1, IGF1R-RAF1, AKT2-PDE3B, OGT-AKT1, E2F1-IL2RA, PRKCE-INSR, and NOS3-IRS1

shown to inhibit insulin secretion, would play a key role in the process (Leahy and Vandekerckhove, 1990; Pørksen et al., 1997). CA could increase the phosphorylation levels of the IGF-I receptor and its downstream signaling molecules (Takasao et al., 2012). It was interesting that binding IGF-I to its receptor could cause the activation of the tyrosine kinase, leading to autophosphorylation of the intrinsic tyrosines, which transduced the IGF-I signal to a complex network that was ultimately responsible for cell proliferation, modulation of tissue differentiation, and protection from apoptosis (Laviola et al., 2007).

- 3) Insulin resistance and CVD: The study showed that the insulin action on cAMP was severely impaired in insulin-resistant patients (Laviola et al., 2007). The cyclic-AMP signaling pathway was shown to be modulated by CA to exhibit antidiabetic action (Schriner et al., 2014). “Regulation of lipolysis in adipocytes” (marked as “Obesity”) was closely linked to MetS since the variations of insulin resistance severity may be related to the regulation of lipolysis in adipocytes (Guilherme et al., 2008). The “AMPK signaling pathway” was proven to be a master regulator of key molecular effectors involved in both metabolic processes and cardiovascular homeostasis by modulating the mTOR signaling and IGF-1 pathway (Salminen and Kaarniranta, 2012). The pathway “II-2 receptor beta chain in T-cell activation” was proven to significantly attenuate ventricular remodeling by reducing infarct size and improving left ventricular (LV) function (Zeng et al., 2016).

The  $S_{GDD}$  and  $S_{AE}$  were calculated for all the 32 CA-related pathways and the 179 HM-enriched pathways. Overall, the average intra-similarity (pathways of the same types including “insulin signaling,” “insulin secretion and obesity,” and “insulin resistance and CVD” as illustrated above) in either CA-related or HM-enriched pathways was similar: higher  $S_{GDD}$  score and lower  $S_{AE}$  scores (see **Table 1** for details). This indicated that these pathways may have small similar structures instead of the whole network. Each pathway may be an up or downstream event in a disease since the biological processes inducing diseases were complex. There may be local similar

structures between two pathways, especially the adjacent ones, that may help transform the information quickly.

The combined pathway network similarity scores between the 12 common pathways are shown in **Figure 2D**. Of which, the pathway “Regulation of lipolysis in adipocytes (hsa04923)” (marked as “Obesity” in **Figure 2D**) got the highest average combined network similarity score (0.438) in the 12 common pathways. As illustrated above, this pathway was involved in the “Insulin resistance and CVD” processes of CVD and MetS, which was the downstream event of CVD and MetS, indicating that more cross-talks may exist between this pathway and the upstream events through the similarity network structures. Compared with this, the pathway “Insulin signaling pathway (hsa04910/h\_insulinPathway)” (marked as “Insulin signaling-1” in **Figure 2D**) got the smallest average combined network similarity score (0.218). Interestingly, this pathway was the node with the highest degree 15 in the combined pathway network. Considering the biological character of this pathway, these indicated that this upstream event in MetS and CVD may play a triggering role regardless of structure similarities to other downstream pathways.

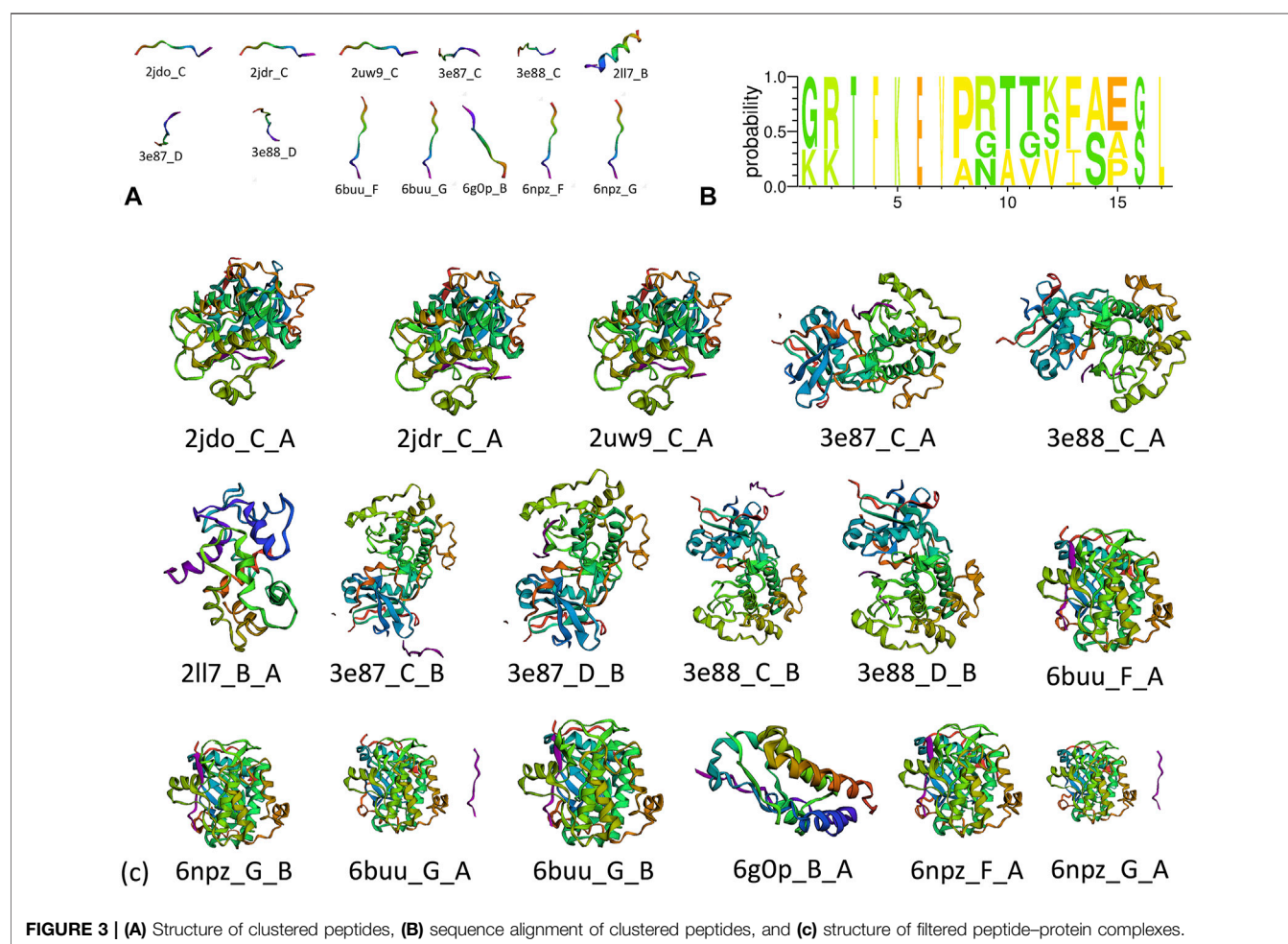
## Peptide-Protein-Based Drug Targets Selection

The combined protein network was built using all the proteins of the 12 common pathways. The network comprised 335 nodes and 1793 edges. The proteins with top 10 degree, node betweenness, and edge betweenness are listed in **Table 2** and selected as raw candidate biomarkers. The degree of a node indicated the importance of a node in the network. A higher degree meant more connections with other nodes; thus, the proteins with higher degree may be the key targets of CA. Five of the top 10 degree proteins had been proven to be regulated by CA, including IRS1, AMPK1, AMPK2, PRKAB1, and PRKAB2. The other five proteins could be divided into two groups: monoamine oxidase (MAOA and MAOB) and protein kinase AMP-activated non-catalytic subunit gamma (PRKAG1, PRKAG2, and PRKAG3) which may be the candidate targets of CA. Cinnamon extracts (CEs) were shown to increase insulin sensitivity by increasing the mRNA expression of INSR (insulin receptor) (Anderson et al.,



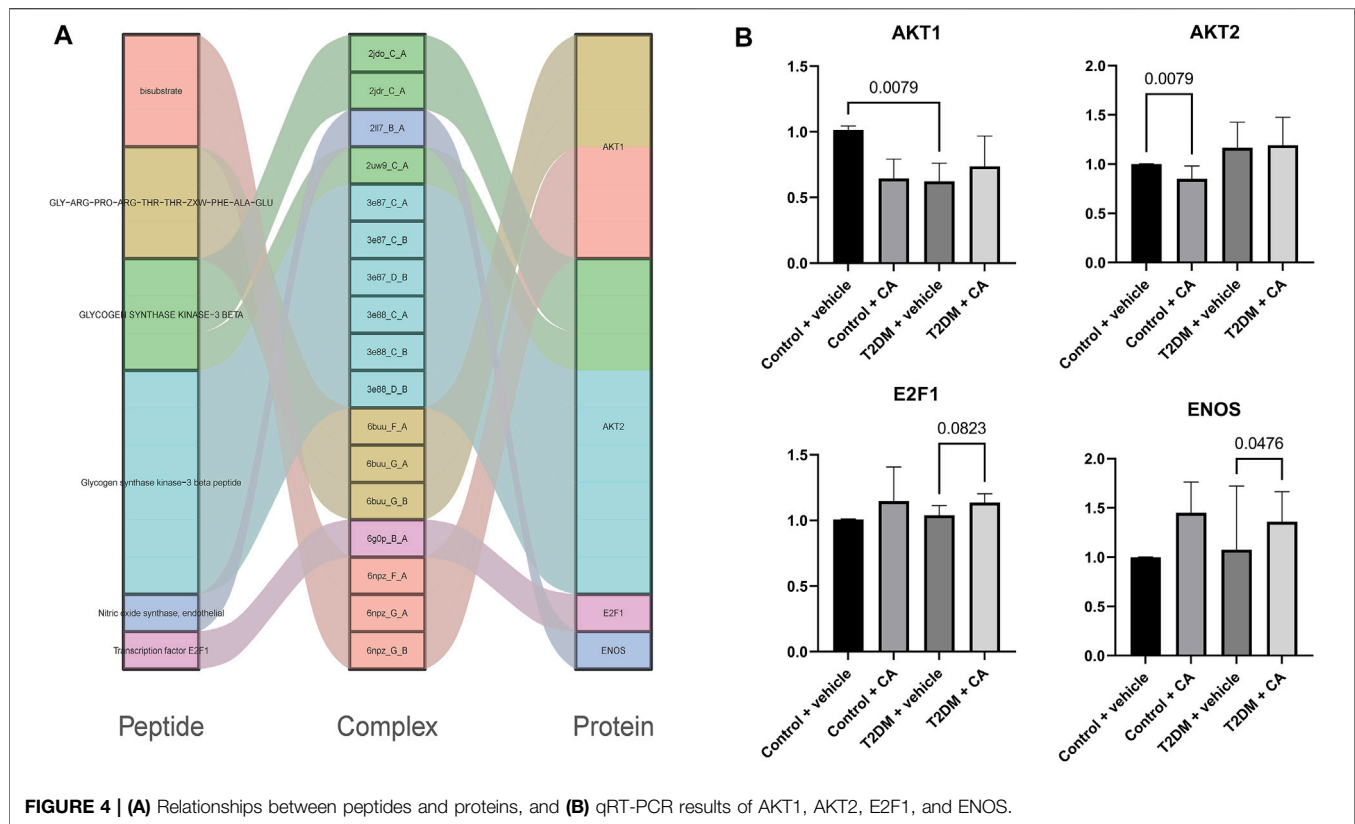
**TABLE 3** | Peptide in the CA-related cluster.

PDB	Peptide chain	Peptide size	Peptide sequence	Peptide description	Peptide molecular weight	Peptide aromaticity	Peptide instability	Peptide isoelectric point
6buu	F	11	GRPRTTXFAEX	GLY-ARG-PRO-ARG-THR-THR-ZXW-PHE-ALA-GLU	—	0.09	—	9.6
6buu	G	11	GRPRTTXFAEX	GLY-ARG-PRO-ARG-THR-THR-ZXW-PHE-ALA-GLU	—	0.09	—	9.6
6npz	F	11	GRPRTTXFAEX	Bisubstrate	—	0.09	—	9.6
6npz	G	11	GRPRTTXFAEX	Bisubstrate	—	0.09	—	9.6
2jdo	C	10	GRPRTTSFAE	Glycogen synthase kinase-3 beta	1121.2	0.1	20.72	9.6
2jdr	C	10	GRPRTTSFAE	Glycogen synthase kinase-3 beta	1121.2	0.1	20.72	9.6
2uw9	C	10	GRPRTTSFAE	Glycogen synthase kinase-3 beta	1121.2	0.1	20.72	9.6
3e87	C	10	GRPRTTSFAE	Glycogen synthase kinase-3 beta peptide	1121.2	0.1	20.72	9.6
3e87	D	10	GRPRTTSFAE	Glycogen synthase kinase-3 beta peptide	1121.2	0.1	20.72	9.6
3e88	C	10	GRPRTTSFAE	Glycogen synthase kinase-3 beta peptide	1121.2	0.1	20.72	9.6
3e88	D	10	GRPRTTSFAE	Glycogen synthase kinase-3 beta peptide	1121.2	0.1	20.72	9.6
6gOp	B	9	PGXGVXSPG	Transcription factor E2F1	—	0	—	5.96
2lI7	B	17	KKTFKEVANAVKISASL	Nitric oxide synthase, endothelial	1834.16	0.06	1.14	10

**FIGURE 3** | (A) Structure of clustered peptides, (B) sequence alignment of clustered peptides, and (C) structure of filtered peptide-protein complexes.

2013), promoting IRS1 (insulin receptor substrate 1) phosphorylation (Liu et al., 2016), and activating AMPK1/2 (protein kinase AMP-activated catalytic subunit alpha 1/2) (Hu

et al., 2013). On the contrary, CE was shown to decrease the expression of genes encoding insulin signaling pathway proteins, including IGF1R (Cao et al., 2010). INS-encoded insulin and trimer



**FIGURE 4 | (A)** Relationships between peptides and proteins, and **(B)** qRT-PCR results of AKT1, AKT2, E2F1, and ENOS.

procyanidins in CE were shown to contribute to the INS-1 pancreatic  $\beta$ -cell protection (Sun et al., 2016).

Compared with degree, the measure “betweenness” reflected the importance of proteins/protein–protein pairs in the interplays between different pathways/diseases. Three of the top 10 betweenness proteins, including IRS1, INS, and INSR, were validated to be regulated by CA. Six of the 10 betweenness edges contained at least one validated CA target. It was found that the two nodes forming the edge IRS1-IGF1R were CA targets; however, IRS1 was upregulated, while IGF1R was downregulated, indicating there may exist complex interactions between CA targets.

A total of 67 protein–peptide complexes containing these top proteins as receptors were selected, and the peptides in these were then aligned and clustered. In total, 13 peptides were grouped in the CA-related cluster, their characters are listed in **Table 3**, and the structures are shown in **Figure 3A,B**. In total, 17 peptide–protein complexes were then filtered (see **Figure 3C** for the complexes’ structures), see **Figure 4A** for the relationships between these peptides and proteins.

Four of the raw candidate biomarkers (AKT1, AKT2, E2F1, and ENOS) were receptors of the abovementioned 17 peptide–protein complexes. qRT-PCR was performed on the four genes (see **Supplementary Table S2** for details).

The candidate biomarkers were divided into three groups according to their expression changing pattern in the qRT-PCR results as follows: see **Figure 4B** for details 1) The genes differentially expressed between Group A (Control + vehicle) and Group B (Control + CA) were named as CA-indicating biomarkers since the two groups were under normal condition, while the only

difference between the two groups was the drug CA. 2) The genes differentially expressed between Group A (Control + vehicle) and Group C (T2D + vehicle) were named as T2D responsive biomarkers since these genes were significantly differentially expressed between T2DM and controls but were not related to the drug CA. 3) The genes differentially expressed between Group C (T2D + vehicle) and Group D (T2D + CA) were named as CA treatment targets since the samples of the two groups were all T2D, while the only difference between them was the treatment of CA. Of which, AKT2 was a CA-indicating biomarker and AKT1 was a T2D responsive biomarker, while E2F1 and ENOS were CA treatment targets. E2F1 and ENOS were shown to cooperate with each other in the treatment of hypertension (Li et al., 2019). Combined with results from this study, the two genes might also cooperate with each other in T2D and become the targets of CA. Besides, the two genes were found to be targeted by SARS-CoV-2-encoded miRNAs in recent research (Aydemir et al., 2021). As a result, CA may be a potential candidate drug to help reduce or prevent the complications since CVDs were one of the most common complications in COVID-19 patients.

## DISCUSSION

The analysis pipeline that was proposed in this study was based on the related genes of multiple diseases. In this study, these genes were collected from OMIM and GWAS results; however, the updates of the gene lists might only influence the results slightly since the analyses were performed on pathway levels. The information conversion from genes to pathways could capture most of the

functional characters of the disease, which may not be changed by adding or deleting a small number of genes. CA was shown to play roles in a wide disease spectrum, which was the character of many Chinese traditional medicines. Thus, the drug targets of these diseases may share some similar characters reflected by peptide clusters. The pipeline proposed in this study could be applied to other diseases and drugs. Pathways were commonly used in biological and medical analyses which could gain deep understanding of diseases. However, other biological terms that could be converted into networks could also be used in this pipeline.

The portability of the pipeline was shown in all the three steps. In step 1 (pathway filter), the similarity calculation methods between different disease pairs could be replaced by any suitable distance measures. The disease-related and drug-related pathways could be selected using any suitable scores or ways. Other functional resources and transcriptional information such as GO terms, transcriptional factors–targets, or miRNA targets could also be used. However, pathways were recommended as the primary choice because the biological pathways were widely used in biological and medical analyses since they could reflect the molecular connections in the form of graphs, which could be analyzed using multiple computational methods. Besides, the correlations between pathways and peptides were closer than those between other types of functional resources. In step 2 (combined network construction), the network structure similarities could be measured using one alignment-free and one alignment-based algorithms. In step 3 (biomarker prediction and validation based on peptides), the peptide clustering algorithms could be replaced by any other suitable alignment method.

## CONCLUSION

In this study, a new pipeline was proposed to discover drug targets based on peptides. The network analyses based on machine learning methods could quantify the effects of peptide–protein complexes with similar structures as drug targets in multiple diseases.

## REFERENCES

- Anderson, R. A., Qin, B., Canini, F., Poulet, L., and Roussel, A. M. (2013). Cinnamon Counteracts the Negative Effects of a High Fat/high Fructose Diet on Behavior, Brain Insulin Signaling and Alzheimer-Associated Changes. *PLoS One* 8 (12), e83243. doi:10.1371/journal.pone.0083243
- Aydemir, M. N., Aydemir, H. B., Korkmaz, E. M., Budak, M., Cekin, N., and Pinarbasi, E. (2021). Computationally Predicted SARS-COV-2 Encoded microRNAs Target NFKB, JAK/STAT and TGFB Signaling Pathways. *Gene Rep.* 22, 101012. doi:10.1016/j.genrep.2020.101012
- Cao, H., Graves, D. J., and Anderson, R. A. (2010). Cinnamon Extract Regulates Glucose Transporter and Insulin-Signaling Gene Expression in Mouse Adipocytes. *Phytomedicine* 17 (13), 1027–1032. doi:10.1016/j.phymed.2010.03.023
- Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). WebLogo: A Sequence Logo Generator. *Figure 1. Genome Res.* 14 (6), 1188–1190. doi:10.1101/gr.849004
- Das, S., Shah, R., Dimmeler, S., Freedman, J. E., Holley, C., Lee, J. M., et al. (2020). Noncoding RNAs in Cardiovascular Disease: Current

## DATA AVAILABILITY STATEMENT

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

## ETHICS STATEMENT

The animal study was reviewed and approved by the experimental Animal Ethics Committee of Shenzhen Sun Yat sen Cardiovascular Hospital.

## AUTHOR CONTRIBUTIONS

YW, KL, and HZ contributed to the conception and design of the study. YW performed the bioinformatics analysis. LW constructed the animal model. YL performed the illustration of results. All authors contributed to manuscript revision, and read and approved the submitted version.

## FUNDING

This study was supported by the China Postdoctoral Science Foundation (2020M683188), the Fund of “Sanming” Project of Medicine in Shenzhen (SZSM201911019), and the National Natural Science Foundation of China (81971079 and 61702496).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.816131/full#supplementary-material>

- Knowledge, Tools and Technologies for Investigation, and Future Directions: A Scientific Statement from the American Heart Association. *Circ. Genom. Precis. Med.* 13, e000062–372. doi:10.1161/HCG.0000000000000062
- Frigo, M., Cruciani, E., Coudert, D., Deriche, R., Natale, E., and Deslauriers-Gauthier, S. (2021). Network Alignment and Similarity Reveal Atlas-Based Topological Differences in Structural Connectomes. *Netw. Neurosci.* 5 (3), 711–733. doi:10.1162/netn\_a\_00199
- Guilherme, A., Virbasius, J. V., Puri, V., and Czech, M. P. (2008). Adipocyte Dysfunctions Linking Obesity to Insulin Resistance and Type 2 Diabetes. *Nat. Rev. Mol. Cell Biol.* 9 (5), 367–377. doi:10.1038/nrm2391
- Hariri, M., and Ghiasvand, R. (2016). Cinnamon and Chronic Diseases. *Adv. Exp. Med. Biol.* 929, 1–24. doi:10.1007/978-3-319-41342-6\_1
- Hu, N., Yuan, L., Li, H. J., Huang, C., Mao, Q. M., Zhang, Y. Y., et al. (2013). Anti-Diabetic Activities of Jiaotaiwan in Db/db Mice by Augmentation of AMPK Protein Activity and Upregulation of GLUT4 Expression. *Evid. Based Complement. Alternat. Med.* 2013, 180721. doi:10.1155/2013/180721
- Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics Enrichment Tools: Paths toward the Comprehensive Functional Analysis of Large Gene Lists. *Nucleic Acids Res.* 37 (1), 1–13. doi:10.1093/nar/gkn923



- Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat. Protoc.* 4 (1), 44–57. doi:10.1038/nprot.2008.211
- Ji, C., Chen, H., Wang, R., Cai, Y., and Wu, H. (2021). *Smoothness Sensor: Adaptive Smoothness-Transition Graph Convolutions for Attributed Graph Clustering*. *IEEE Trans Cybern.* doi:10.1109/TCYB.2021.3088880
- Kakiuchi-Kiyota, S., Schutten, M. M., Zhong, Y., Crawford, J. J., and Dey, A. (2019). Safety Considerations in the Development of Hippo Pathway Inhibitors in Cancers. *Front. Cell Dev. Biol.* 7, 156. doi:10.3389/fcell.2019.00156
- Kaku, K. (2019). A New Concept of GLP-1 Signaling Pathway on Pancreatic Insulin Secretion. *J. Diabetes Investig.* 11, 265–267. doi:10.1111/jdi.13136
- Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs. *Nucleic Acids Res.* 45 (D1), D353–D361. doi:10.1093/nar/gkw1092
- Khan, A., Bryden, N. A., Polansky, M. M., and Anderson, R. A. (1990). Insulin Potentiating Factor and Chromium Content of Selected Foods and Spices. *Biol. Trace Elem. Res.* 24 (3), 183–188. doi:10.1007/BF02917206
- Krejci, A., Hupp, T. R., Lexa, M., Vojtesek, B., and Muller, P. (2016). Hammock: a Hidden Markov Model-Based Peptide Clustering Algorithm to Identify Protein-Interaction Consensus Motifs in Large Datasets. *Bioinformatics* 32 (1), 9–16. doi:10.1093/bioinformatics/btv522
- Kuhn, H. W. (1955). The Hungarian Method for the Assignment Problem. *Naval Res. Logistics Q.* 2 (1–2), 83–97. doi:10.1002/nav.3800020109
- Laviola, L., Natalicchio, A., and Giorgino, F. (2007). The IGF-I Signaling Pathway. *Cpd* 13 (7), 663–669. doi:10.2174/138161207780249146
- Leahy, J. L., and Vandekerckhove, K. M. (1990). Insulin-Like Growth Factor-I at Physiological Concentrations Is a Potent Inhibitor of Insulin Secretion\*. *Endocrinology* 126 (3), 1593–1598. doi:10.1210/endo-126-3-1593
- Li, H., Li, Q., Zhang, Y., Liu, W., Gu, B., Narumi, T., et al. (2019). Novel Treatment of Hypertension by Specifically Targeting E2F for Restoration of Endothelial Dihydrofolate Reductase and eNOS Function under Oxidative Stress. *Hypertension* 73 (1), 179–189. doi:10.1161/hypertensionaha.118.11643
- Liu, Y., Li, X., Xie, C., Luo, X., Bao, Y., Wu, B., et al. (2016). Prevention Effects and Possible Molecular Mechanism of Mulberry Leaf Extract and its Formulation on Rats with Insulin-Insensitivity. *PLoS One* 11 (4), e0152728. doi:10.1371/journal.pone.0152728
- London, N., Raveh, B., Movshovitz-Attias, D., and Schueler-Furman, O. (2010). Can Self-Inhibitory Peptides Be Derived from the Interfaces of Globular Protein-Protein Interactions? *Proteins* 78 (15), 3140–3149. doi:10.1002/prot.22785
- Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., et al. (2015). Uncovering Disease-Disease Relationships through the Incomplete Interactome. *Science* 347 (6224), 1257601. doi:10.1126/science.1257601
- Mollazadeh, H., and Hosseinzadeh, H. (2016). Cinnamon Effects on Metabolic Syndrome: a Review Based on its Mechanisms. *Iran J. Basic Med. Sci.* 19 (12), 1258–1270. doi:10.22038/ijbms.2016.7906
- Neduvu, V., Linding, R., Su-Angrand, I., Stark, A., Masi, F. d., Gibson, T. J., et al. (2005). Systematic Discovery of New Recognition Peptides Mediating Protein Interaction Networks. *Plos Biol.* 3 (12), e405. doi:10.1371/journal.pbio.0030405
- Parthasarathi, L., Casey, F., Stein, A., Aloy, P., and Shields, D. C. (2008). Approved Drug Mimics of Short Peptide Ligands from Protein Interaction Motifs. *J. Chem. Inf. Model.* 48 (10), 1943–1948. doi:10.1021/ci800174c
- Pawson, T., and Nash, P. (2003). Assembly of Cell Regulatory Systems through Protein Interaction Domains. *Science* 300 (5618), 445–452. doi:10.1126/science.1083653
- Pørksen, N., Hussain, M. A., Bianda, T. L., Nyholm, B., Christiansen, J. S., Butler, P. C., et al. (1997). IGF-I Inhibits Burst Mass of Pulsatile Insulin Secretion at Supraphysiological and Low IGF-I Infusion Rates. *Am. J. Physiol.* 272 (3 Pt 1), E352–E358. doi:10.1152/ajpendo.1997.272.3.E352
- Przulj, N. (2007). Biological Network Comparison Using Graphlet Degree Distribution. *Bioinformatics* 23 (2), e177–e183. doi:10.1093/bioinformatics/btl301
- Qin, B., Nagasaki, M., Ren, M., Bajotto, G., Oshida, Y., and Sato, Y. (2003). Cinnamon Extract (Traditional Herb) Potentiates *In Vivo* Insulin-Regulated Glucose Utilization via Enhancing Insulin Signaling in Rats. *Diabetes Res. Clin. Pract.* 62 (3), 139–148. doi:10.1016/s0168-8227(03)00173-6
- Sales, G., Calura, E., Cavaliere, D., and Romualdi, C. (2012). Graphite - a Bioconductor Package to Convert Pathway Topology to Gene Network. *BMC Bioinformatics* 13, 20. doi:10.1186/1471-2105-13-20
- Salminen, A., and Kaarniranta, K. (2012). AMP-activated Protein Kinase (AMPK) Controls the Aging Process via an Integrated Signaling Network. *Ageing Res. Rev.* 11 (2), 230–241. doi:10.1016/j.arr.2011.12.005
- Salt, I. P., and Hardie, D. G. (2017). AMP-activated Protein Kinase. *Circ. Res.* 120 (11), 1825–1841. doi:10.1161/circresaha.117.309633
- Schriner, S. E., Kuramada, S., Lopez, T. E., Truong, S., Pham, A., and Jafari, M. (2014). Extension of *Drosophila* Lifespan by Cinnamon through a Sex-specific Dependence on the Insulin Receptor Substrate chico. *Exp. Gerontol.* 60, 220–230. doi:10.1016/j.exger.2014.09.019
- Sheng, X., Zhang, Y., Gong, Z., Huang, C., and Zang, Y. Q. (2008). Improved Insulin Resistance and Lipid Metabolism by Cinnamon Extract through Activation of Peroxisome Proliferator-Activated Receptors. *PPAR Res.* 2008, 581348. doi:10.1155/2008/581348
- Sun, P., Wang, T., Chen, L., Yu, B.-w., Jia, Q., Chen, K.-x., et al. (2016). Trimer Procyanidin Oligomers Contribute to the Protective Effects of Cinnamon Extracts on Pancreatic  $\beta$ -cells *In Vitro*. *Acta Pharmacol. Sin.* 37 (8), 1083–1090. doi:10.1038/aps.2016.29
- Takasao, N., Tsuji-Naito, K., Ishikura, S., Tamura, A., and Akagawa, M. (2012). Cinnamon Extract Promotes Type I Collagen Biosynthesis via Activation of IGF-I Signaling in Human Dermal Fibroblasts. *J. Agric. Food Chem.* 60 (5), 1193–1200. doi:10.1021/jf2043357
- Thompson, J. D., Higgins, D. G., Gibson, T. J., and Clustal, W. (1994). CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-specific gap Penalties and Weight Matrix Choice. *Nucl. Acids Res.* 22 (22), 4673–4680. doi:10.1093/nar/22.22.4673
- Wang, Y., Huang, X., Liu, J., Zhao, X., Yu, H., and Cai, Y. (2019). A Systems Analysis of the Relationships between Anemia and Ischemic Stroke Rehabilitation Based on RNA-Seq Data. *Front. Genet.* 10, 456. doi:10.3389/fgene.2019.00456
- Xiao, Y., Wu, Q. Q., Jiang, X. H., and Tang, Q. Z. (2017). Cinnamaldehyde Attenuates Pressure Overload-Induced Cardiac Fibrosis via Inhibition of Endothelial Mesenchymal Transition. *Zhonghua Yi Xue Za Zhi* 97 (11), 869–873. doi:10.3760/cma.j.issn.0376-2491.2017.11.015
- Yang, L., Wu, Q. Q., Liu, Y., Hu, Z. F., Bian, Z. Y., and Tang, Q. Z. (2015). Cinnamaldehyde Attenuates Pressure Overload-Induced Cardiac Hypertrophy. *Int. J. Clin. Exp. Pathol.* 8 (11), 14345–14354.
- Yingying, W., Yu, Y., Jianfeng, L., and Keshen, L. (2021). Construction of Anatomical Structure-specific Developmental Dynamic Networks for Human Brain on Multiple Omics Levels. *Curr. Bioinformatics* 16 (9), 1133–1142. doi:10.2174/1574893616666210331115659
- Zeng, Z., Yu, K., Chen, L., Li, W., Xiao, H., and Huang, Z. (2016). Interleukin-2/ Anti-Interleukin-2 Immune Complex Attenuates Cardiac Remodeling after Myocardial Infarction through Expansion of Regulatory T Cells. *J. Immunol. Res.* 2016, 8493767. doi:10.1155/2016/8493767
- Zhang, L. Q., Zhang, Z. G., Fu, Y., and Xu, Y. (2015). Research Progress of Trans-cinnamaldehyde Pharmacological Effects. *Zhongguo Zhong Yao Za Zhi* 40 (23), 4568–4572.
- Zhu, R., Ji, C., Wang, Y., Cai, Y., and Wu, H. (2020). Heterogeneous Graph Convolutional Networks and Matrix Completion for miRNA-Disease Association Prediction. *Front. Bioeng. Biotechnol.* 8, 901. doi:10.3389/fbioe.2020.00901

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

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

Copyright © 2021 Wang, Wang, Liu, Li and Zhao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Systematic Modeling, Prediction, and Comparison of Domain–Peptide Affinities: Does it Work Effectively With the Peptide QSAR Methodology?

Qian Liu<sup>1</sup>, Jing Lin<sup>1</sup>, Li Wen<sup>1</sup>, Shaozhou Wang<sup>1</sup>, Peng Zhou<sup>1\*</sup>, Li Mei<sup>2\*</sup> and Shuyong Shang<sup>3</sup>

<sup>1</sup>Center for Informational Biology, School of Life Science and Technology, University of Electronic Science and Technology of China (UESTC), Chengdu, China, <sup>2</sup>Institute of Culinary, Sichuan Tourism University, Chengdu, China, <sup>3</sup>Institute of Ecological Environment Protection, Chengdu Normal University, Chengdu, China

## OPEN ACCESS

### Edited by:

Juexin Wang,  
University of Missouri, United States

### Reviewed by:

Supratik Kar,  
Jackson State University,  
United States  
Fei He,  
Northeast Normal University, China

### \*Correspondence:

Peng Zhou  
p\_zhou@uestc.edu.cn  
Li Mei  
meili520777@126.com

### Specialty section:

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

**Received:** 24 October 2021

**Accepted:** 14 December 2021

**Published:** 14 January 2022

### Citation:

Liu Q, Lin J, Wen L, Wang S, Zhou P, Mei L and Shang S (2022) Systematic Modeling, Prediction, and Comparison of Domain–Peptide Affinities: Does it Work Effectively With the Peptide QSAR Methodology? *Front. Genet.* 12:800857. doi: 10.3389/fgene.2021.800857

The protein–protein association in cellular signaling networks (CSNs) often acts as weak, transient, and reversible domain–peptide interaction (DPI), in which a flexible peptide segment on the surface of one protein is recognized and bound by a rigid peptide-recognition domain from another. Reliable modeling and accurate prediction of DPI binding affinities would help to ascertain the diverse biological events involved in CSNs and benefit our understanding of various biological implications underlying DPIs. Traditionally, peptide quantitative structure–activity relationship (pQSAR) has been widely used to model and predict the biological activity of oligopeptides, which employs amino acid descriptors (AADs) to characterize peptide structures at sequence level and then statistically correlate the resulting descriptor vector with observed activity data *via* regression. However, the QSAR has not yet been widely applied to treat the direct binding behavior of large-scale peptide ligands to their protein receptors. In this work, we attempted to clarify whether the pQSAR methodology can work effectively for modeling and predicting DPI affinities in a high-throughput manner? Over twenty thousand short linear motif (SLiM)-containing peptide segments involved in SH3, PDZ and 14-3-3 domain-mediated CSNs were compiled to define a comprehensive sequence-based data set of DPI affinities, which were represented by the Boehringer light units (BLUs) derived from previous arbitrary light intensity assays following SPOT peptide synthesis. Four sophisticated MLMs (MLMs) were then utilized to perform pQSAR modeling on the set described with different AADs to systematically create a variety of linear and nonlinear predictors, and then verified by rigorous statistical test. It is revealed that the genome-wide DPI events can only be modeled qualitatively or semiquantitatively with traditional pQSAR strategy due to the intrinsic disorder of peptide conformation and the potential interplay between different peptide residues. In addition, the arbitrary BLUs used to characterize DPI affinity values were measured *via* an indirect approach, which may not very reliable and may involve strong noise, thus leading to a considerable bias in the modeling. The  $R_{pred}^2 = 0.7$  can be considered as the upper limit of external generalization ability of the pQSAR methodology working on large-scale DPI affinity data.

**Keywords:** computational peptidology, peptide quantitative structure–activity relationship, domain–peptide interaction, amino acid descriptor, statistical modeling, machine learning

## 1 INTRODUCTION

Protein–protein interactions play a key role in cell life. Through formation of the functionally complicated complexes between two or more interacting protein partners, they participate in a variety of signal cascades in cells, thereby regulating the life activities of cells and individuals (Slater et al., 2020). In cell signaling network, intrinsically disordered proteins (IDP) often interacts specifically with the peptide-recognition domain of target protein through a flexible peptide segment on its own surface (Dyson and Wright 2005). In this way, flexible peptides tend to spontaneously fold into regular secondary structures, and then the specific recognition and interaction between peptide-recognition domains (PRDs) and flexible peptides were created in a folding-on-binding or binding-on-folding manner (Dyson and Wright 2002). Different from the permanent and stable complexes that are commonly formed by binding with global rigid globulins, the domain–peptide complexes are generally transient and reversible due to the limited number of residues and small contact area involved in the complex interfaces. This feature makes domain–peptide interactions (DPIs) very suitable to serve as molecular switches in biological signaling pathways that require exquisitely dynamic regulation and are closely related to various cellular processes and major diseases.

Although the high-throughput synthesis techniques such as combinatorial library, phage display and peptide microarray have considerably promoted DPI discovery over the past decades (Engelmann et al., 2014; Gray and Brown 2014; Zambrano-Mila et al., 2020), it is still time-consuming and expensive to practice full systematic screening against all potential peptide segment candidates in the human genome. In addition, a variety of peptide-recognition domains existed in cells also largely intensify the challenge of systematic screening. To tackle this issue, we previously suggested the *computational peptidology* as a new and attractive area to rationally investigate and design bioactive peptides or peptidic agents with *in silico* assistance (Zhou et al., 2013), in which the peptide quantitative structure–activity relationship (pQSAR) is one of the most widely used strategies to model the statistical correlation between peptide structure and biological activity (or toxicity, efficacy and potency) at sequence level (Zhou et al., 2008a). Machine learning has been widely used to perform the pQSAR modeling, but most of previous studies were focused on specific domains and/or limited samples, and thus unable to systematically evaluate the feasibility and applicability of pQSAR methodology in predicting DPI affinities. For example, Hou et al. deployed a series of works to characterize the 3D-structurally physiochemical properties of peptide binding to SH3 domain by using dynamics simulation, molecular field analysis and interaction energy component decomposition, and then they employed support vector machine (SVM) to create the pQSAR relationship between the characterized property parameters and measured DPI affinities (Hou et al., 2006; Hou et al., 2008; Hou et al., 2009). Jin et al. used random forest (RF) to perform structure-based pQSAR study of DPI binding behavior by dissecting residue interaction profile at the complex interface of PDZ domain with its peptide ligands (Jin et al., 2013). We also proposed the Gaussian process (GP) as a promising machine learning approach to predict the binding affinities and biological

activities of diverse peptides against different proteins and domains (Zhou et al., 2008b; Zhou et al., 2010).

The key to the development of rapid pQSAR virtual screening technology for genome-wide DPIs is the characterization of interaction binding behavior and the construction of multivariate statistical model. The former parameterizes the sequence, structural, physicochemical and/or energetic properties of DPIs into a set of multidimensional numerical vectors that can be readily processed in computer, and the latter generates a regression relationship by statistically associating the vector set with corresponding DPI affinities with supervised machine learning approach. Recently, we have given a systematic review on the application of machine learning methods (MLMs) to quantitative DPI affinity prediction and its implications for therapeutic peptide design (Li et al., 2019), in which we pointed out that, although a number of pQSAR works have been reported to address the DPI affinity prediction problem, there was no comprehensive evaluation and systematic comparison of the pQSAR modeling performance between the different combinations of peptide-recognition domain types, MLMs and structural characterization strategies, thus lacking a general conclusion for the applicability of pQSAR methodology in DPI affinity modeling and prediction. In this study, we attempted to create, examine and compare a variety of pQSAR predictors built with PLS, SVM, RF and GP on >20,000 SLiM peptides involved in SH3, PDZ and 14-3-3 domain-mediated cell signaling networks. These peptide structures were characterized at traditional sequence level using classical amino acid descriptors (AADs) and their affinities were determined consistently by SPOT peptide syntheses and arbitrary light intensity assays. This work would shed light on the general purpose of pQSAR-based DPI affinity modeling and prediction.

## 2 MATERIALS AND METHODS

### 2.1 Four Machine Learning Methods That Have Ever Been Applied in Peptide Quantitative Structure–Activity Relationship

Four sophisticated MLMs that have ever been applied in the pQSAR study of DPIs and other protein–peptide binding phenomena were considered in this work, including one linear partial least squares (PLS) and three nonlinear support vector machine (SVM), random forest (RF) and Gaussian process (GP) (Geladi and Kowalski 1986; Cortes and Vapnik 1995; Breiman 2001; Obrezanova et al., 2007). The PLS is a widely used multivariate statistical technique in the QSAR community, which has been intrinsically integrated into the famous 3D-QSAR methods of comparative molecular field analysis (CoMFA) and comparative molecular similarity indices (CoMSIA) as standard modeling tool to perform pQSAR analysis of SH3–peptide interactions at molecular field level (Hou et al., 2006). The method provides a multi-dependent variable to multi-independent variable regression, which can better deal with the problems difficult to be solved by least square regression. The SVM has also been successfully employed to characterize the SH3- and PDZ-mediated DPIs involved in the human genome (Hou et al., 2008; Hou et al.,

**TABLE 1** | Four MLMs used in this study.

MLM	Type	Variable standardization	Model parameter	
			Parameter	Optimization
PLS	Linear	Autoscaling	NLV: number of latent variables	Increase of cumulative cross-validation $q^2$ is below 0.097
SVM	Nonlinear	[−1, +1] scaling	$\epsilon$ : $\epsilon$ -insensitive loss function C: penalty factor $\sigma^2$ : kernel radial	Systematic grid search for minimizing cross-validation RMSE <sub>cv</sub>
RF	Nonlinear	[−1, +1] scaling	ntree: number of trees mtry: size of descriptor subset	Systematic grid search for minimizing cross-validation RMSE <sub>cv</sub>
GP	Linear/nonlinear	Autoscaling	$\Theta$ : hyperparameter set	Automatic determination

2009; Li et al., 2011). The method converts quadratic convex programming problem into the corresponding duality problem for solving by Lagrange multiplier method, and constructs a series of kernel functions by using Mercer theorem to realize the high-dimensional inner product operation in the original space (Cortes and Vapnik 1995). In addition, the RF and GP were also introduced previously by our group to investigate DPIs (Zhou et al., 2008b) and other peptide-related issues such as enzyme-inhibitory activity (Zhou et al., 2010) and chromatographic retention behavior (Tian et al., 2009; Zhou et al., 2009). The former is an ensemble learning algorithm based on decision tree proposed by Breiman (2001), which also provides additional features such as variable importance and out-of-bag (OOB) validation that increase its utility for statistical modeling. The latter is based on the Bayesian non-parametric model that has a strict statistical learning theory basis and a strong generalization ability to adjust the model's flexibility and achieve a certain transparency through so-call “hyperparameters” rather than conventional parameters to avoid fixed basis function in the traditional sense (Obrezanova et al., 2007).

The details of these machine learning modeling processes can be found in our previous publications (Rasmussen and Williams 2006). Briefly, the input variables were standardized by autoscaling for PLS and RF or [−1, +1] scaling for SVM and GP. The model parameters such as the number of latent variables (NLV) for PLS, and the  $\epsilon$ -insensitive loss function, penalty factor (C) and kernel radial ( $\sigma^2$ ) for SVM, the number of trees (ntree) and the optimal size of the variable subset (mtry) for RF and the hyperparameter set ( $\Theta$ ) for GP need to be determined before modeling, and we employed consistent strategies as summarized in **Table 1** to optimize these parameters. Here, the PLS, SVM, RF and GP modeling and parameter optimization were carried out with in-house Matlab tool box ZP-explore (Zhou et al., 2009). In addition, the SVM regression was also carried out using the sophisticated LibSVM program (Chang and Lin 2011) for comparison purpose.

## 2.2 Curation of Comprehensive Sequence-Based Domain–Peptide Interaction Data Set With a Consistent Affinity Expression

A variety of peptide-recognition domains that can specifically recognize and interact with diverse short linear motifs (SLiMs) on

their partner protein surfaces have been discovered over the past decades (Kuriyan and Cowburn 1997), including but not limited to SH3, SH2, WW, PDZ, PTB, 14-3-3, EH, GYF, PH, EVH1, UEV, VHS, FHA, WD40 and so on. Here, we mainly selected three most common domain categories with considerably different SLiM properties but highly consistent affinity data for this study, namely, SH3, PDZ and 14-3-3; they can be further divided into different subtypes in terms of their parent proteins. The SH3 domain was first identified in the non-receptor tyrosine kinase *c*-Src and can specifically binds PxxP-containing polyproline-II (PPII) helix peptide segments (Li et al., 2005). The PDZ domain targets the C-terminal free peptide segments of substrate proteins with a plastic pattern (Ivarsson 2012). The 14-3-3 domain has been widely found in hundreds of signaling proteins to mediate protein–protein interactions by recognizing peptide segments of phosphoserine or phosphothreonine residues (Aitken et al., 1995).

Here, we curated totally 21,704 SLiM-containing peptides that separately target ten SH3 domains, seven PDZ domains and one 14-3-3 domain from previous reports (Boisguerin et al., 2004; Landgraf et al., 2004; Vouilleme et al., 2010; Panni et al., 2011) to define a comprehensive sequence-based DPI affinity data set consisting of 18 panels. These peptides were produced using SPOT peptide synthesis technology on cellulose membranes and then their binding affinities to different domains were consistently indicated by Boehringer light units (BLUs) derived from arbitrary light intensity assays (Volkmer et al., 2012). This protocol can fast yield various peptide candidates in a short time scale and test their domain binding in a high-throughput manner, and thus have been widely used to measure DPI affinities. By further excluding few invalid samples such as no binders or no affinity values, we consequently obtained 21,399 valid peptides; their information are summarized in **Table 2**, and their sequences and BLU values are tabulated in **Supplementary Tables S1–S3**.

## 2.3 Statistical Verification of Peptide Quantitative Structure–Activity Relationship Models With Internal and External Validations

The built pQSAR predictive models should pass rigorous statistical test before practical applications to examine their effectiveness, illness and generalization ability. Here, we used a



**TABLE 2 |** Summary of 21,704 SLiM-containing peptide samples binding to ten SH3, seven PDZ and one 14-3-3 domains.

Panel	Domain	Parent protein	Domain Number	Species	Peptide number
1	SH3	Amphiphysin	1/1	Human	884 Landgraf et al. (2004)
2		Amphiphysin	1/1	Yeast	2032 Landgraf et al. (2004)
3		Boi1	1/1	Yeast	1336 Landgraf et al. (2004)
4		Boi2	1/1	Yeast	1312 Landgraf et al. (2004)
5		Endophilin	1/1	Yeast	1998 Landgraf et al. (2004)
6		Myosin5	1/1	Yeast	1139 Landgraf et al. (2004)
7		Rvs167	1/1	Yeast	1369 Landgraf et al. (2004)
8		Sho1	1/1	Yeast	1015 Landgraf et al. (2004)
9		Yfr024	1/1	Yeast	1282 Landgraf et al. (2004)
10		Yhr016c	1/1	Yeast	1348 Landgraf et al. (2004)
11	PDZ	CALP	1/1	Human	80 Vouilleme et al. (2010)
12		NHERF1	1/2	Human	77 Vouilleme et al. (2010)
13		NHERF1	2/2	Human	80 Vouilleme et al. (2010)
14		NHERF2	1/2	Human	80 Vouilleme et al. (2010)
15		NHERF2	2/2	Human	80 Vouilleme et al. (2010)
16		SYNA1	1/1	Human	56 Vouilleme et al. (2010)
17		PSD95	1/1	Human	6068 Boisguerin et al. (2004)
18	14-3-3	14-3-3	1/1	Yeast	1163 Panni et al. (2011)

combination of internal and external validations to verify the statistical stability and predictive power of the models. Internal validation includes goodness-of-fit and 10-fold cross-validation on training set, while for the external validation we randomly divided each sample panel into  $\sim 2/3$  as a training set for building pQSAR model, and the remaining  $\sim 1/3$  as a test set for blind testing of the built model. In a highly cited paper, Golbraikh and Alexander (2002) pointed out that the internal validation is only a necessary but not sufficient condition to measure the reliability of a QSAR model, and the model predictability must be confirmed further through external validation.

## 2.4 Structural Characterization of Peptide Sequences Using Amino Acid Descriptors

Amino acid descriptors (AADs) are a classical approach to characterize peptide structure at sequence level, which utilize a  $n$ -dimensional vector to represent each of 20 amino acids and are commonly derived from a large number of original amino acid properties such as topological, physicochemical, 3D-structural and quantum-chemical, by using multivariate statistical techniques such as principal components analysis (PCA) and factor analysis (FA) (Zhou et al., 2008a). An  $n$ -mer peptide can be parameterized by in turn replacing its each amino acid residue to a corresponding  $m$ -dimensional AAD array, consequently resulting in  $n \times m$  descriptors for the peptide, which define the independent variable space  $X$  and can be further correlated statistically with independent variable  $y$  (affinity) using machine learning regression. Recently, we have systematically evaluated totally 80 AADs in pQSAR modeling and identified a number of AADs with good performance (Zhou et al., 2021), from which we herein selected four different types of AADs to characterize the 21,399 SLiM-containing peptides listed in Table 2, including MolSurf (quantum-chemical) (Norinder and Svensson 1998), ST\_scale (topological) (Yang et al., 2010), VHSE (physicochemical) (Mei et al., 2005) and

VSGETAWAY (3D-structural) (Tong and Zhang 2007). Their values are tabulated in **Supplementary Tables S4–S7**.

## 3 RESULTS AND DISCUSSION

There are several indicators that can be used to represent the binding affinity of DPIs, such as the  $K_d$  that can be determined by fluorescence polarization (FP) and surface plasmon resonance (SPR) to indicate the apparent dissociation constant for domain–peptide complex formation, and the  $\Delta G$  that can be measured using isothermal titration calorimetry (ITC) to denote free energy change upon the complex binding. However, neither  $K_d$  nor  $\Delta G$  can be obtained in a high-throughput manner, and thus they are not feasible for characterizing the large-scale DPI affinity data. In recent years, the SPOT peptide synthesis in conjunction with light intensity assays has been used to rapidly screen effective domain binders against massive peptide candidates, where peptides matching the defined patterns were synthesized at high density on cellulose membranes by SPOT synthesis technology and the membranes were probed with GST-fused domain protein, which were then revealed by an anti-GST antibody and by a secondary anti-IgG antibody coupled to horseradish peroxidase (POD) to derive the intensity of each SPOT quantitatively in Boehringer light unit (BLU) as an arbitrary light intensity unit (Landgraf et al., 2004). In this study, all DPI affinity data were expressed consistently as the BLU values collected from Refs (Boisguerin et al., 2004; Landgraf et al., 2004; Vouilleme et al., 2010; Panni et al., 2011).

### 3.1 Effect of Machine Learning Methods on Peptide Quantitative Structure-Activity Relationship Modeling

The PLSR, GP, RF, SVM and LibSVM regressions were employed to create three types of DPI affinity predictors for 18 DPI panels

**TABLE 3** | Comparison of different MLMs on different DPI samples<sup>a</sup>.

MLM	DPI <sup>b</sup>	Training set				Test set	
		$R_{fit}^{2c}$	$RMSE_{fit}^d$	$R_{cv}^{2c}$	$RMSE_{cv}^d$	$R_{prd}^{2c}$	$RMSE_{prd}^d$
PLS	SH3	0.8641	0.4765	0.8335	0.5275	0.3072	0.5851
	PDZ	0.9312	0.1062	0.1077	0.3823	0.2263	0.3276
	14-3-3	0.4344	0.7048	0.3341	0.7687	0.3625	0.7446
GP	SH3	0.8668	0.4719	0.8349	0.5252	0.3147	0.5808
	PDZ	0.6984	0.2223	0.1953	0.3631	0.3391	0.3028
	14-3-3	0.4334	0.7091	0.3548	0.7566	0.3669	0.7420
RF	SH3	0.9470	0.2975	0.2074	1.1509	0.4973	0.4987
	PDZ	0.8191	0.1722	0.4005	0.3134	0.3824	0.2928
	14-3-3	0.8116	0.4088	0.2562	0.8124	0.3456	0.7715
SVM	SH3	0.8772	0.4530	0.8352	0.5248	0.3091	0.5843
	PDZ	0.7242	0.2126	0.1880	0.3647	0.2689	0.3594
	14-3-3	0.5211	0.6519	0.3614	0.7527	0.3886	0.7279
LibSVM	SH3	0.7008	0.2971	0.6817	0.3144	0.4254	0.3693
	PDZ	0.8778	0.0813	0.1295	0.1528	0.2766	0.1189
	14-3-3	0.4003	0.6085	0.3097	0.6702	0.3025	0.6342

<sup>a</sup>VHSE, descriptor was used to characterize peptide sequences.

<sup>b</sup>Human amphiphsin SH3 (1/1), human SYNA1 PDZ (1/1) and yeast 14-3-3 (1/1) are selected as case analysis.

<sup>c</sup> $R_{fit}^{2c}$ ,  $R_{cv}^{2c}$  and  $R_{prd}^{2c}$  are the determination coefficients of internal fitting in training set, internal cross-validation on training set, and external blind prediction on test set, respectively.

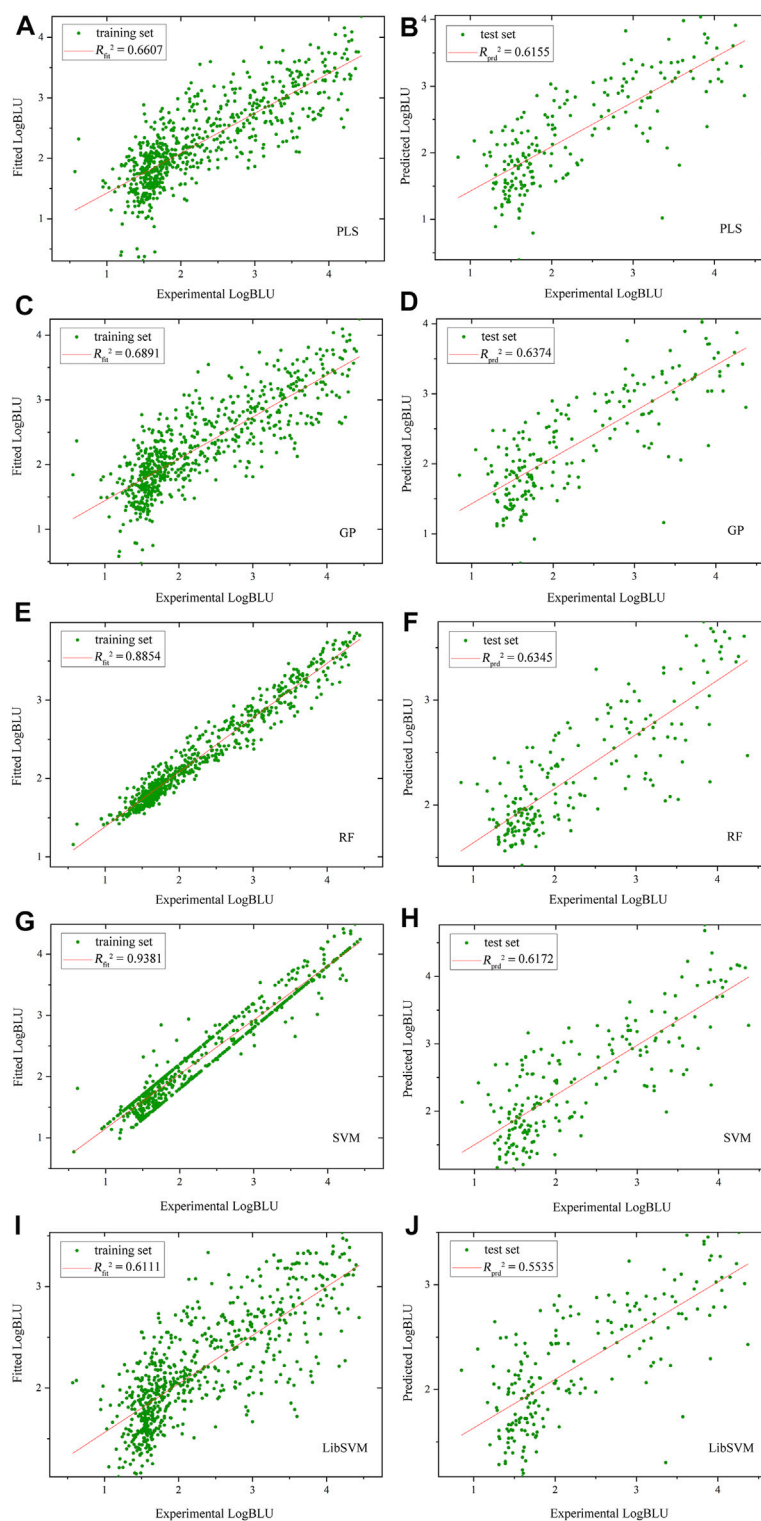
<sup>d</sup> $RMSE_{fit}$ ,  $RMSE_{cv}$  and  $RMSE_{prd}$  are the root-mean-square errors of internal fitting in training set, internal cross-validation on training set, and external blind prediction on test set, respectively.

based on training samples, which were then used to blindly predict test samples (all resulting statistics are tabulated in the **Supplementary Tables S8–S10**). In order to compare different MLMs in the modeling and prediction of DPI affinities, we selected three kinds of samples binding separately to human amphiphsin SH3 (1/1), human SYNA1 PDZ (1/1) and yeast 14-3-3 (1/1) domains, and compared their fitting determination coefficient  $R_{fit}^{2c}$  on the training set, cross-validation determination coefficient  $R_{cv}^{2c}$  on the training set and predictive determination coefficient  $R_{prd}^{2c}$  on test set. As can be in **Table 3**, the performance of obtained pQSAR models varies considerably over MLMs and domain types. These models have high internal fitting ability but generally exhibit moderate or modest internal stability and external predictability, with  $R_{fit}^{2c} > 0.6$  but  $R_{cv}^{2c} < 0.6$  and  $R_{prd}^{2c} < 0.5$ . Among the three types of DPI affinity predictors the predictive power  $R_{prd}^{2c}$  of nonlinear GP, RF, SVM and LibSVM is generally better than that of linear PLS, suggesting that the DPI events are complicated dynamic process that involve many nonlinear factors, which can be better handled by nonlinear than linear methods. Even so, the modeling performance of both the linear and nonlinear methods is generally moderately, indicated by the high internal fitting ability but relatively low internal stability and predictability, imparting an overfitting phenomenon may exist in these regression models.

The optimal models were built on human amphiphsin SH3 (1/1)-binding peptide panel with MolSurf characterization. Here, the scatter plots of fitted/predictive against experimental LogBLU values over 884 peptide samples using different MLMs are shown in **Figure 1**. As can be seen, the resulting external predictive  $R_{prd}^{2c}$  values are generally larger than 0.5, indicating a good generalization ability on this panel. In addition, the internal fitting  $R_{fit}^{2c}$  values of all these MLMs (except LibSVM) are

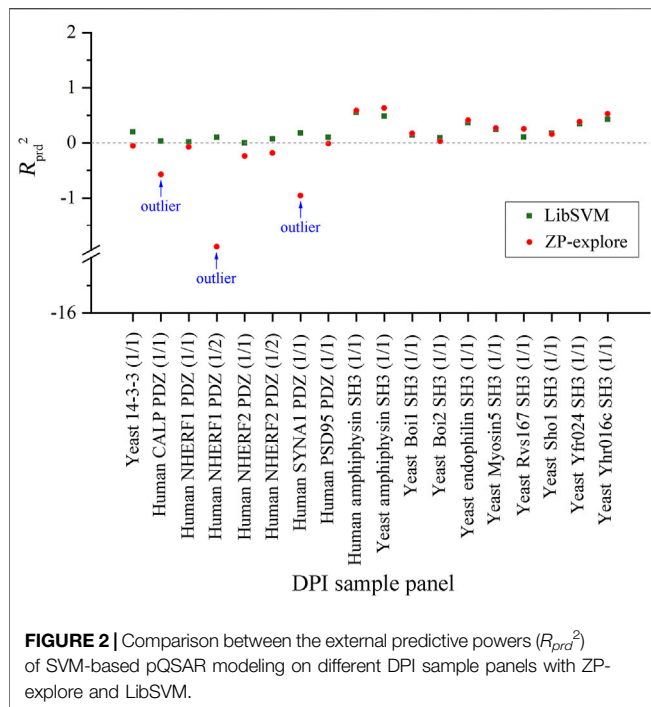
significantly higher than 0.65, in which the RF and SVM perform much better than others. However, there are no essential difference between the predictive powers of RF and SVM with PLS and GP ( $R_{prd}^{2c} > 0.6$ ), but are moderately better than LibSVM ( $R_{prd}^{2c} < 0.6$ ). The nonlinear GP, RF and SVM seem to have a good generalization ability relative linear PLS, albeit the difference is not very significant, suggesting that both the linear and nonlinear approaches exhibit similar predictability on test set, although the nonlinear methods can give stronger fitting on training set than linear one. This is also explain why the linear PLS has been successfully used in previous pQSAR modeling of DPI affinities, which can perform similarly but are easier to operate and more readily interpretable than those nonlinear modeling.

By comparing the SVM regressions modeled by in-house ZP-explore toolbox (Zhou et al., 2009) and sophisticated LibSVM program (Chang and Lin 2011), it is revealed that the former can perform considerably better than the latter, although both of them used the same machine learning method (SVM), worked on the same data panel (human amphiphsin SH3 (1/1)-binding peptides) and characterized the same AAD (MolSurf). This finding suggested that the pQSAR modeling of DPI affinities are sensitive to not only the data sets measured, but also the software used. This issue is usually neglected by the pQSAR community and previous works have no systematic examination of different tools/programs/software used in modeling. Therefore, we herein further compared the external predictive powers ( $R_{prd}^{2c}$ ) of SVM regressions modeled by ZP-explore and LibSVM on all the 18 DPI sample panels in **Figure 2**. It is revealed that the prediction can achieve a generally consistent power for some panels (e.g., human amphiphsin SH3 (1/1)- and human Biol1 SH3-binding peptides), but varies considerably for some others (e.g.,



**FIGURE 1 |** Scatter plots of fitted/predictive against experimental LogBLU values over 884 human amphiphysin SH3 (1/1)-binding peptides with MolSurf characterization and using different MLMs (A,B), PLSR, (C,D), GP; (E,F), RF, (G,H), SVM and (I,J), LibSVM.

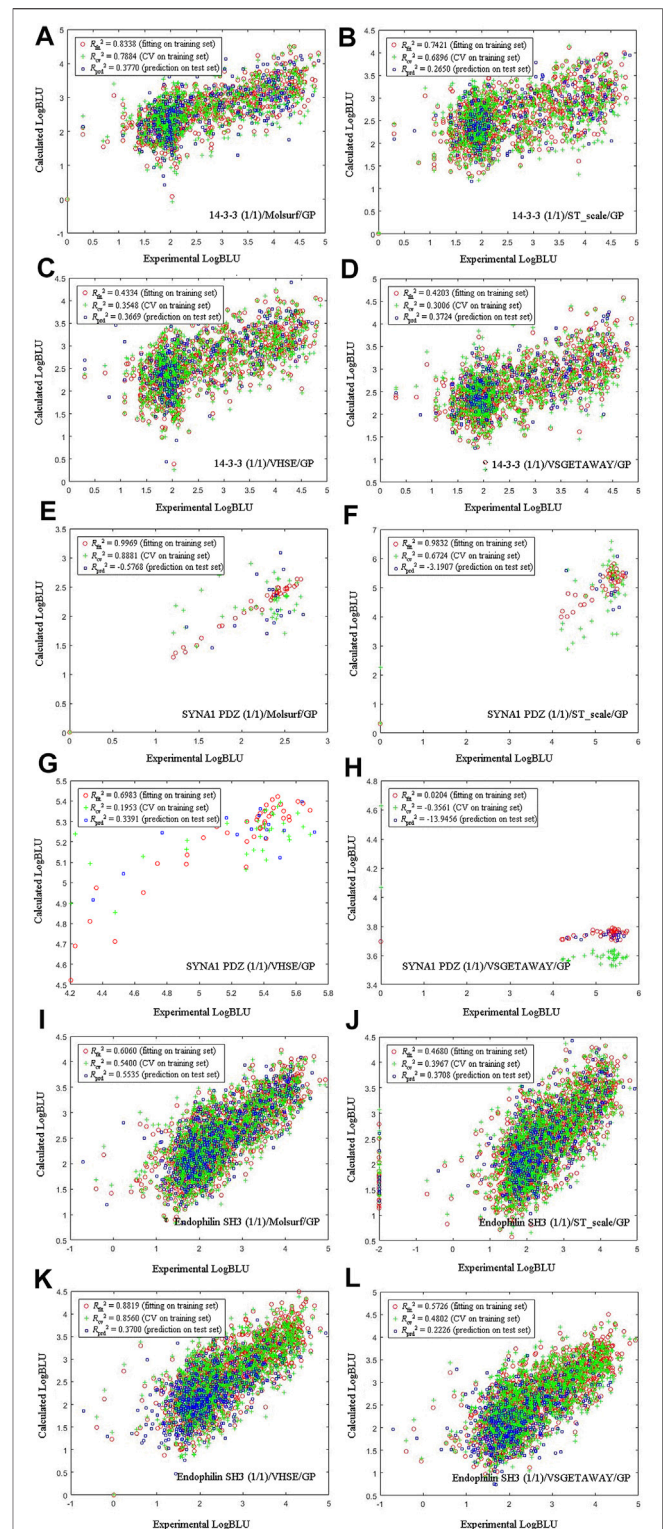


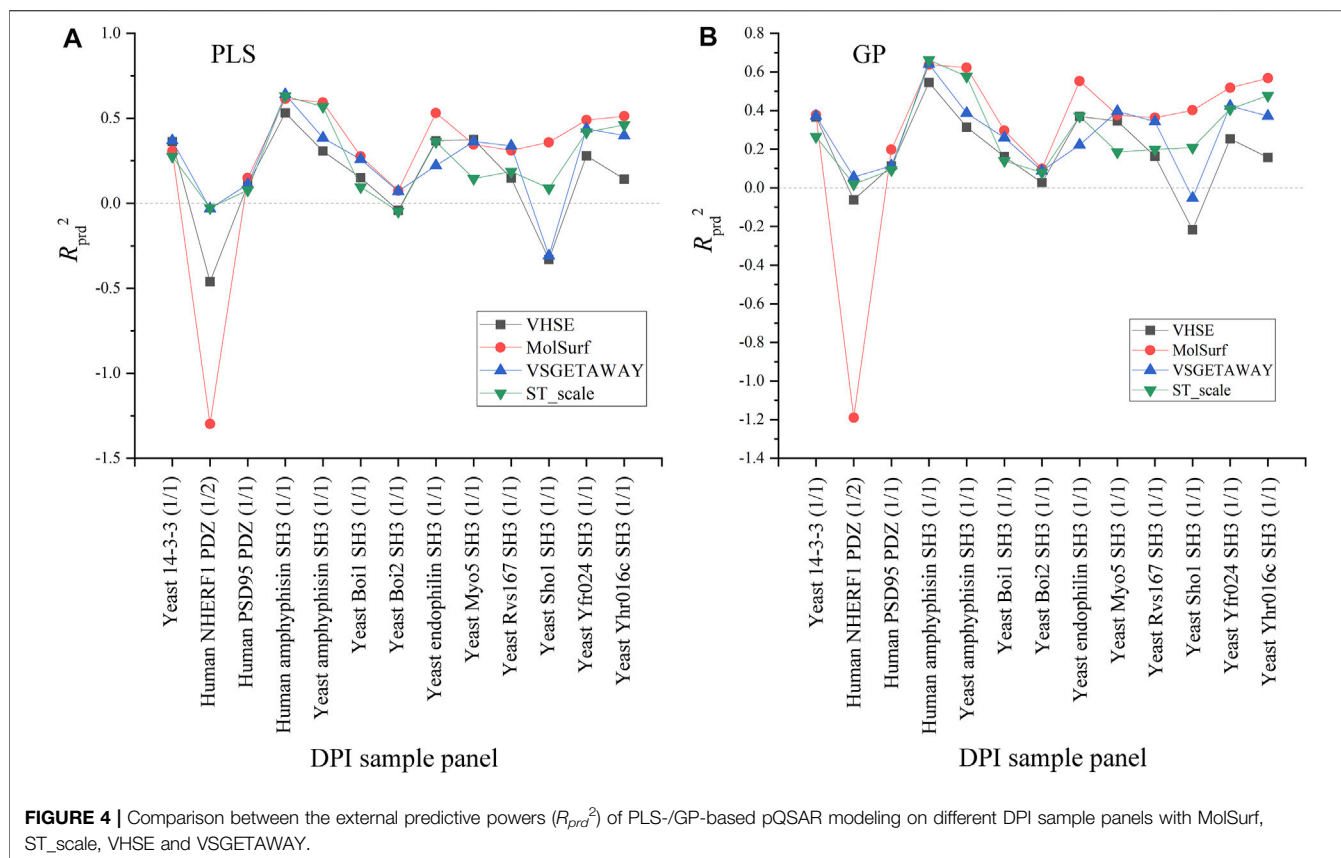


human NHERF1 PDZ (1/2)- and human SYNA1 PDZ (1/1)-binding peptides). It is worth noting that, although the ZP-explore can yield a better prediction on certain panels than LibSVM, the latter appears to be more stable than the former, as characterized by the ZP-explore predictive outliers for three PDZ panels in **Figure 2**, although for most panels the two tools can work similarly in their predictive behavior.

### 3.2 Effect of Amino Acid Descriptors on Peptide Quantitative Structure-Activity Relationship Modeling

Four amino acid descriptors characterizing different properties of amino acids, namely MolSurf (quantum-chemical), ST\_scales (topological), VHSE (physicochemical) and VSGETAWAY (3D-structural), were used to parameterize peptide sequences, which were then correlated with experimental LogBLU values with GP modeling on three selected DPI sample panels: human 14-3-3 (1/1), human SYNA1 PDZ (1/1), and yeast endophilin SH3 (1/1), and the resulting scatter plots of calculated against experimental LogBLU values over these panels are shown in **Figure 3**. It is evident that the calculated results, including internal fitting ability  $R_{fit}^2$  on training set, internal cross-validation stability  $R_{cv}^2$  on training set, and external predictability  $R_{prd}^2$  on test set, vary considerably over pQSAR models built with different AADs. For the 56 human SYNA1 PDZ (1/1)-binding peptides, the  $R_{fit}^2$ ,  $R_{cv}^2$  and  $R_{prd}^2$  all exhibit considerable illness, indicating that the pQSAR models cannot work effectively on this panel, no matter which AADs were used. In contrast, pQSAR modeling seems to have a moderate or good performance on the 1193 human 14-3-3 (1/1)- and 2025 yeast





**FIGURE 4 |** Comparison between the external predictive powers ( $R^2_{prd}$ ) of PLS-/GP-based pQSAR modeling on different DPI sample panels with MolSurf, ST\_scale, VHSE and VSGETAWAY.

endophilin SH3 (1/1)-binding peptides, with a satisfactory profile of internal fitting ability and cross-validation stability ( $R^2_{fit} > 0.6$  and  $R^2_{cv} > 0.5$ ), albeit many have only a moderate or modest external predictive power ( $R^2_{prd} < 0.4$ ). In addition, for the same sample panels characterized using different AADs, the pQSAR models generally exhibit a similar performance on both training and test sets, suggesting that the descriptor types would not have significant effect on modeling performance. However, the change in sample panels can lead to a considerable variation on the performance, suggesting that the AADs are not primarily responsible for pQSAR modeling; instead, the sample panels are.

Effects of four AADs on the external predictive powers ( $R^2_{prd}$ ) of PLS-/GP-based pQSAR models are compared in **Figure 4**. As can be seen, the linear PLS (A) and nonlinear GP (B) have a similar profile of  $R^2_{prd}$  values over these panels, in which the prediction on human NHERF1 PDZ (1/2) and Yeast Sho1 SH3 (1/1) vary significantly and moderately over the four AADs, respectively, while these descriptors exhibit a generally consistent performance for predicting other sample panels. For human NHERF1 PDZ (1/2) panel, the quantum-chemical MolSurf performs much worse, and secondly the physiochemical VHSE, whereas other two descriptors can work normally on this panel. For Yeast Sho1 SH3 (1/1) panel, only the quantum-chemical MolSurf has a particularly low performance as compared to other three descriptors. Besides, the four AADs seem to have a consistent performance on other panels. Even so, the pQSAR  $R^2_{prd}$  values obtained with different descriptors on these panels mainly range between 0 and 0.6, imparting that the models have only a moderate or modest predictive power on most sample panels,

and the  $R^2_{prd}$  variation is primarily influenced by sample panels but not descriptor types.

### 3.3 Effect of Sample Size on Peptide Quantitative Structure-Activity Relationship Modeling

By systematically examining the influence of MLMs and AADs on pQSAR modeling of different DPI sample panels, it is revealed that these models can perform fairly well on the human PSD95 PDZ (1/1) panel, which contains totally 6,068 peptide samples. Here, the MolSurf was employed to characterize the structure of these peptides at sequence level and then we carried out pQSAR modeling on all the 6,068 samples and two subsets with PLS, GP, RF, SVM and LibSVM regressions. The two subsets separately contain 1,000 and 3,000 sample data extracted randomly from the intact panel. The modeling resulted in 15 pQSAR models, which represent the systematic combination between five MLMs and three subsets with different sample sizes. The external predictive power ( $R^2_{prd}$ ) of these models on test set is listed in **Table 4**. It is seen that the models with fullset-6068 can generally obtain a consistent predictability for most MLMs as compared to other two subsets, except the RF modeling on the subset-3000, which yielded the highest  $R^2_{prd}$  than subset-1000 and fullset-6068. In contrast, the pQSAR modeling on subset-1000 can only obtain a marginal prediction. However, the  $R^2_{prd}$  difference is not very significant between different subsets for the same MLMs, but different MLMs can lead to a

**TABLE 4 |** Change in external predictive power ( $R_{prd}^2$ ) of PLS/GP/RF/SVM/ LibSVM-based pQSAR modeling on human PSD95 PDZ (1/1) panel characterized by MolSurf with different sample sizes of subset-1000, subset-3000 and fullset-6068.

Size	$R_{prd}^2$				
	PLS	GP	RF	SVM	LibSVM
Subset-1000	0.014	0.080	0.362	0.071	0.084
Subset-3000	0.087	0.088	0.485	0.119	0.117
Fullset-6068	0.149	0.198	0.421	0.204	0.147

considerable variation in the  $R_{prd}^2$  value. In addition, all the five MLMs can reach the highest fitting ability ( $R_{fit}^2$ ) with the fullset-6068 relative to subset-1000 and subset-3000. Therefore, it is revealed that the pQSAR performance is primarily determined by MLMs used and, secondarily, sample size. The larger the size is, the higher the performance is. Even so, the  $R_{prd}^2$  values of pQSAR modeling on the fullset-6068 are all not above the 0.5, indicating that the absolute predictive power of different MLMs is improved with sample size increase, but the increase is quite limited.

## 4 CONCLUSION

More than 20,000 SLiM-containing peptides as the binders of 3 peptide-recognition domains (PDZ, SH3 and 14-3-3) and 18 domain subtypes were comprehensively collected to perform an investigation of the applicability of pQSAR methodology in peptide affinity prediction. With a systematic combination of five widely used MLMs and four informatively diverse AADs to perform the pQSAR modeling on these peptide samples it is revealed that the domains and MLMs have significant effects on modeling performance, whereas the AADs and sample size can only influence the performance moderately and modestly. However, at most conditions the predictive power of pQSAR models is generally below 0.5 and only very few can be above 0.6, no matter what the combinations of domains, MLMs, AADs and sample size are adopted. This can be attributed to the fact that the high-throughput detection of arbitrary light intensity is a very indirect approach to characterize DPI affinity and the obtained BLU can only give a qualitative or semi-quantitative measure of

the affinity values, thus causing a considerable bias in the pQSAR modeling and prediction. Instead, although some other affinity indicators such as  $K_d$  and  $\Delta G$  are quantitative and more reliable, they cannot be tested in a high-throughput manner and thus are normally unavailable for large-scale DPI samples. Therefore, it is suggested that only focus on pQSAR modeling by optimizing AADs and MLMs is not an essential solution to improve the modeling performance of DPI affinity. Instead, the source of affinity data used to perform the modeling is the current bottleneck to restrict the feasibility and applicability of pQSAR methodology in DPI affinity prediction.

## DATA AVAILABILITY STATEMENT

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

## AUTHOR CONTRIBUTIONS

QL and PZ came up with the idea for this study. PZ and LM supervised this work throughout. QL, JL, and LW analyzed the data. QL and SW contributed to the tables, figures and software. QL and PZ wrote the manuscript. LM and SS revised the manuscript. All authors read and approved the final article.

## FUNDING

This work was supported by the National Natural Science Foundation of China (PZ, No. 31671361) and the Scientific Research Fund of Sichuan Provincial Education Department (SS, No. 17ZA0052).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.800857/full#supplementary-material>

## REFERENCES

- Aitken, A., Jones, D., Soneji, Y., and Howell, S. (1995). 14-3-3 Proteins: Biological Function and Domain Structure. *Biochem. Soc. Trans.* 23, 605–611. doi:10.1042/bst0230605
- Boisguerin, P., Leben, R., Ay, B., Radziwill, G., Moelling, K., Dong, L., et al. (2004). An Improved Method for the Synthesis of Cellulose Membrane-Bound Peptides with Free C Termini Is Useful for PDZ Domain Binding Studies. *Chem. Biol.* 11, 449–459. doi:10.1016/j.chembiol.2004.03.010
- Breiman, L. (2001). Random Forests. *Machine Learn.* 45, 5–32. doi:10.1023/a:1010933404324
- Chang, C. C., and Lin, C. J. (2011). Libsvm. *ACM Trans. Intell. Syst. Technol.* 2, 1–27. doi:10.1145/1961189.1961199
- Cortes, C., and Vapnik, V. (1995). Support-vector Networks. *Mach. Learn.* 20, 273–297. doi:10.1007/bf00994018
- Dyson, H. J., and Wright, P. E. (2002). Coupling of Folding and Binding for Unstructured Proteins. *Curr. Opin. Struct. Biol.* 12, 54–60. doi:10.1016/s0959-440x(02)00289-0
- Dyson, H. J., and Wright, P. E. (2005). Intrinsically Unstructured Proteins and Their Functions. *Nat. Rev. Mol. Cell Biol.* 6, 197–208. doi:10.1038/nrm1589
- Engelmann, B. W., Kim, Y., Wang, M., Peters, B., Rock, R. S., and Nash, P. D. (2014). The Development and Application of a Quantitative Peptide Microarray Based Approach to Protein Interaction Domain Specificity Space. *Mol. Cell Proteomics* 13, 3647–3662. doi:10.1074/mcp.o114.038695
- Geladi, P., and Kowalski, B. R. (1986). Partial Least-Squares Regression: a Tutorial. *Analytica Chim. Acta* 185, 1–17. doi:10.1016/0003-2670(86)80028-9
- Golbraikh, A., and Tropsha, A. (2002). Beware of Q2! *J. Mol. Graphics Model.* 20, 269–276. doi:10.1016/s1093-3263(01)00123-1
- Gray, B. P., and Brown, K. C. (2014). Combinatorial Peptide Libraries: Mining for Cell-Binding Peptides. *Chem. Rev.* 114, 1020–1081. doi:10.1021/cr400166n



- Hou, T., McLaughlin, W., Lu, B., Chen, K., and Wang, W. (2006). Prediction of Binding Affinities between the Human Amphiphysin-1 SH3 Domain and its Peptide Ligands Using Homology Modeling, Molecular Dynamics and Molecular Field Analysis. *J. Proteome Res.* 5, 32–43. doi:10.1021/pr0502267
- Hou, T., Xu, Z., Zhang, W., McLaughlin, W. A., Case, D. A., Xu, Y., et al. (2009). Characterization of Domain–Peptide Interaction Interface. *Mol. Cell Proteomics* 8, 639–649. doi:10.1074/mcp.m800450-mcp200
- Hou, T., Zhang, W., Case, D. A., and Wang, W. (2008). Characterization of Domain–Peptide Interaction Interface: A Case Study on the Amphiphysin-1 SH3 Domain. *J. Mol. Biol.* 376, 1201–1214. doi:10.1016/j.jmb.2007.12.054
- Ivarsson, Y. (2012). Plasticity of PDZ Domains in Ligand Recognition and Signaling. *FEBS Lett.* 586, 2638–2647. doi:10.1016/j.febslet.2012.04.015
- Jin, R., Ma, Y., Qin, L., and Ni, Z. (2013). Structure-based Prediction of Domain–Peptide Binding Affinity by Dissecting Residue Interaction Profile at Complex Interface: a Case Study on CAL PDZ Domain. *Ppl* 20, 1018–1028. doi:10.2174/0929866511320090008
- Kuriyan, J., and Cowburn, D. (1997). Modular Peptide Recognition Domains in Eukaryotic Signaling. *Annu. Rev. Biophys. Biomol. Struct.* 26, 259–288. doi:10.1146/annurev.biophys.26.1.259
- Landgraf, C., Panni, S., Montecchi-Palazzi, L., Castagnoli, L., Schneider-Mergener, J., Volkmer-Engert, R., et al. (2004). Protein Interaction Networks by Proteome Peptide Scanning. *Plos Biol.* 2, e14. doi:10.1371/journal.pbio.0020014
- Li, N., Hou, T., Ding, B., and Wang, W. (2011). Characterization of PDZ Domain–Peptide Interaction Interface Based on Energetic Patterns. *Proteins* 79, 3208–3220. doi:10.1002/prot.23157
- Li, S. S. C. (2005). Specificity and Versatility of SH3 and Other Proline-Recognition Domains: Structural Basis and Implications for Cellular Signal Transduction. *Biochem. J.* 390, 641–653. doi:10.1042/bj20050411
- Li, Z., Miao, Q., Yan, F., Meng, Y., and Zhou, P. (2019). Machine Learning in Quantitative Protein–Peptide Affinity Prediction: Implications for Therapeutic Peptide Design. *Curr. Drug Metab.* 20, 170–176. doi:10.2174/1389200219666181012151944
- Mei, H., Liao, Z. H., Zhou, Y., and Li, S. Z. (2005). A New Set of Amino Acid Descriptors and its Application in Peptide QSARs. *Biopolymers* 80, 775–786. doi:10.1002/bip.20296
- Norinder, U., and Svensson, P. (1998). Descriptors for Amino Acids Using MolSurf Parametrization. *J. Comput. Chem.* 19, 51–59. doi:10.1002/(sici)1096-987x(19980115)19:1<51:aid-jcc4>3.0.co;2-y
- Obrezanova, O., Csányi, G., Gola, J. M. R., and Segall, M. D. (2007). Gaussian Processes: a Method for Automatic QSAR Modeling of ADME Properties. *J. Chem. Inf. Model.* 47, 1847–1857. doi:10.1021/ci7000633
- Panni, S., Montecchi-Palazzi, L., Kiemer, L., Cabibbo, A., Paoluzi, S., Santonico, E., et al. (2011). Combining Peptide Recognition Specificity and Context Information for the Prediction of the 14-3-3-mediated Interactome in *S. cerevisiae* and *H. Sapiens*. *Proteomics* 11, 128–143. doi:10.1002/pmic.201000030
- Rasmussen, C. E., and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Massachusetts, USA: The MIT Press.
- Slater, O., Miller, B., and Kontoyianni, M. (2020). Decoding Protein–Protein Interactions: An Overview. *Ctmc* 20, 855–882. doi:10.2174/1568026620666200226105312
- Tian, F., Yang, L., Lv, F., and Zhou, P. (2009). Predicting Liquid Chromatographic Retention Times of Peptides from the *Drosophila melanogaster* Proteome by Machine Learning Approaches. *Analytica Chim. Acta* 644, 10–16. doi:10.1016/j.aca.2009.04.010
- Tong, J., and Zhang, S. (2007). A New 3D-Descriptor of Amino Acids and its Application in Quantitative Structure Activity Relationship of Peptide Drugs. *Acta Phys. Chim. Sin.* 23, 37–43.
- Volkmer, R., Tapia, V., and Landgraf, C. (2012). Synthetic Peptide Arrays for Investigating Protein Interaction Domains. *FEBS Lett.* 586, 2780–2786. doi:10.1016/j.febslet.2012.04.028
- Vouilleme, L., Cushing, P. R., Volkmer, R., Madden, D. R., and Boisguerin, P. (2010). Engineering Peptide Inhibitors to Overcome PDZ Binding Promiscuity. *Angew. Chem. Int. Edition* 49, 9912–9916. doi:10.1002/anie.201005575
- Yang, L., Shu, M., Ma, K., Mei, H., Jiang, Y., and Li, Z. (2010). ST-scale as a Novel Amino Acid Descriptor and its Application in QSAM of Peptides and Analogues. *Amino Acids* 38, 805–816. doi:10.1007/s00726-009-0287-y
- Zambrano-Mila, M. S., Blacio, K. E. S., and Vispo, N. S. (2020). Peptide Phase Display: Molecular Principles and Biomedical Applications. *Ther. Innov. Regul. Sci.* 54, 308–317. doi:10.1007/s43441-019-00059-5
- Zhou, P., Chen, X., Wu, Y., and Shang, Z. (2010). Gaussian Process: an Alternative Approach for QSAM Modeling of Peptides. *Amino Acids* 38, 199–212. doi:10.1007/s00726-008-0228-1
- Zhou, P., Liu, Q., Wu, T., Miao, Q., Shang, S., Wang, H., et al. (2021). Systematic Comparison and Comprehensive Evaluation of 80 Amino Acid Descriptors in Peptide QSAR Modeling. *J. Chem. Inf. Model.* 61, 1718–1731. doi:10.1021/acs.jcim.0c01370
- Zhou, P., Tian, F., Chen, X., and Shang, Z. (2008b). Modeling and Prediction of Binding Affinities between the Human Amphiphysin SH3 Domain and its Peptide Ligands Using Genetic Algorithm–Gaussian Processes. *Biopolymers* 90, 792–802. doi:10.1002/bip.21091
- Zhou, P., Tian, F., Lv, F., and Shang, Z. (2009). Comprehensive Comparison of Eight Statistical Modelling Methods Used in Quantitative Structure–Retention Relationship Studies for Liquid Chromatographic Retention Times of Peptides Generated by Protease Digestion of the *Escherichia coli* Proteome. *J. Chromatogr. A* 1216, 3107–3116. doi:10.1016/j.chroma.2009.01.086
- Zhou, P., Tian, F., Wu, Y., Li, Z., and Shang, Z. (2008a). Quantitative Sequence–Activity Model (QSAM): Applying QSAR Strategy to Model and Predict Bioactivity and Function of Peptides, Proteins and Nucleic Acids. *Curr. Comput. Aid. Drug Des.* 4, 311–321. doi:10.2174/157340908786785994
- Zhou, P., Wang, C., Ren, Y., Yang, C., and Tian, F. (2013). Computational Peptidology: a New and Promising Approach to Therapeutic Peptide Design. *Curr. Med. Chem.* 20, 1985–1996. doi:10.2174/0929867311320150005

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

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

Copyright © 2022 Liu, Lin, Wen, Wang, Zhou, Mei and Shang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# SSH2.0: A Better Tool for Predicting the Hydrophobic Interaction Risk of Monoclonal Antibody

Yuwei Zhou<sup>1,2</sup>, Shiyang Xie<sup>1,2</sup>, Yue Yang<sup>1,2</sup>, Lixu Jiang<sup>1,2</sup>, Siqi Liu<sup>1,2</sup>, Wei Li<sup>1,2</sup>, Hamza Bukari Abagna<sup>1,2</sup>, Lin Ning<sup>3\*</sup> and Jian Huang<sup>1,2\*</sup>

<sup>1</sup>Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China, <sup>2</sup>School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China, <sup>3</sup>School of Healthcare Technology, Chengdu Neusoft University, Chengdu, China

## OPEN ACCESS

### Edited by:

Chuan Dong,  
Wuhan University, China

### Reviewed by:

Jin-Xing Liu,  
Qufu Normal University, China  
Chengchi Fang,  
Institute of Hydrobiology (CAS), China

### \*Correspondence:

Lin Ning  
NingLin@nsu.edu.cn  
Jian Huang  
hj@uestc.edu.cn

### Specialty section:

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

**Received:** 23 December 2021

**Accepted:** 31 January 2022

**Published:** 15 March 2022

### Citation:

Zhou Y, Xie S, Yang Y, Jiang L, Liu S,  
Li W, Abagna HB, Ning L and Huang J  
(2022) SSH2.0: A Better Tool for  
Predicting the Hydrophobic Interaction  
Risk of Monoclonal Antibody.  
Front. Genet. 13:842127.  
doi: 10.3389/fgene.2022.842127

Therapeutic antibodies play a crucial role in the treatment of various diseases. However, the success rate of antibody drug development is low partially because of unfavourable biophysical properties of antibody drug candidates such as the high aggregation tendency, which is mainly driven by hydrophobic interactions of antibody molecules. Therefore, early screening of the risk of hydrophobic interaction of antibody drug candidates is crucial. Experimental screening is laborious, time-consuming, and costly, warranting the development of efficient and high-throughput computational tools for prediction of hydrophobic interactions of therapeutic antibodies. In the present study, 131 antibodies with hydrophobic interaction experiment data were used to train a new support vector machine-based ensemble model, termed SSH2.0, to predict the hydrophobic interactions of antibodies. Feature selection was performed against CKSAAGP by using the graph-based algorithm MRMD2.0. Based on the antibody sequence, SSH2.0 achieved the sensitivity and accuracy of 100.00 and 83.97%, respectively. This approach eliminates the need of three-dimensional structure of antibodies and enables rapid screening of therapeutic antibody candidates in the early developmental stage, thereby saving time and cost. In addition, a web server was constructed that is freely available at <http://i.uestc.edu.cn/SSH2/>.

**Keywords:** therapeutic antibody, developability, hydrophobic interactions, support vector machine, prediction model

## INTRODUCTION

Antibodies play an indispensable role in the vertebrate immune defence system (Kapingidza et al., 2020). They also serve as essential agents in biomedical research and clinical diagnostic assays such as enzyme-linked immunosorbent assay, immunohistochemical assay, and immunoprecipitation assay. Furthermore, antibodies have been extensively used in clinical treatment of many types of cancers, autoimmune diseases, and infectious diseases including the coronavirus disease 2019, which is caused by the severe acute respiratory syndrome coronavirus 2 (Ning et al., 2021). Rapid development of the monoclonal antibody (mAb) technology has revolutionised pharmaceutical science and industry. Many proteins that cannot interact with small chemical molecules or are undruggable due to self-tolerance are considered efficient targets for antibody drugs. More than 550 therapeutic mAbs have been tested in phase I/II clinical trials worldwide, of which 79 mAbs have entered the final stage of development (Kaplon et al., 2020). Antibody drugs account for a large

market share in the pharmaceutical industry. In 2018, the therapeutic antibodies had a global value of United States \$115.2 billion, which is expected to reach \$300 billion by the end of 2025 (Lu et al., 2020). Moreover, the large-scale application of antibody phage display, single B-cell antibody, and next-generation sequencing technologies has resulted in the development of tens of thousands of preclinical therapeutic antibody drug candidates. However, the probability of a human or humanised antibody drug candidate, which is under clinical trials, being approved is low (approximately 15%) (Carter and Lazar, 2018). Many mAbs fail due to unfavourable physicochemical properties such as high viscosity, increased aggregation tendency, and susceptibility to chemical degradation (Jain et al., 2017b).

Protein aggregation has been considered as one of the major challenges in biological drug development. It poses challenges during different developmental processes from fermentation and purification to storage (Obrezanova et al., 2015). It not only reduces the effectiveness of a drug but also induces adverse immune responses in patients (Martinez Morales et al., 2019). Thus, identifying therapeutic antibody candidates with high aggregation tendency at the early developmental stage is essential. The factors that affect protein aggregation are either intrinsic (e.g., interaction between hydrophobic patches, van der Waals forces and electrostatic interactions) or extrinsic (e.g., pH, salt concentration, buffer type, and storage conditions). Among these factors, the presence of hydrophobic moieties on the protein surface is the strongest determinant (Hebditch et al., 2019). A few tools to predict the hydrophobicity of proteins including mAbs have been reported (Lienqueo et al., 2006; Mahn et al., 2009; Hanke et al., 2016; Jain et al., 2017a). However, most of these tools rely on protein structures and do not provide free web services. In our previous study, we developed a tool called SSH, which can predict the hydrophobic interaction risk of mAbs solely by using the mAb sequences (Dzisoo et al., 2020). The SSH tool was trained with the tripeptide composition (TPC), and the prediction accuracy of 91.226% was achieved through the voting strategy. However, the number of features used to build the SSH model is extremely higher than the number of its samples, causing concerns with overfitting and weak generalisation.

In the present study, we combined the experimental assay data to construct a novel *in silico* tool called SSH2.0 for the prediction of hydrophobic interaction risk of mAbs. The tool developed in this study predicted hydrophobic interaction risk of mAbs by using only the amino acid sequence. Compared with the previous version, SSH2.0 was trained with new features that were optimised using a new feature selection method. Overall, SSH2.0 was superior to the previous version in terms of performance.

## DATASET AND METHOD

### Dataset

The antibody dataset used in a study by Jain et al. (2017b) was selected in the present study. We linked the variable region in the form of “heavy chain–light chain” as the antibody sequences. The dataset comprised 137 antibody sequences (48 from approved

antibodies and 89 from clinical II/III trials) and data of 12 biophysical and binding assays. Six antibody sequences with conflicting records were eliminated, resulting in inclusion of 131 antibody sequences. The assays, namely stand-up monolayer adsorption chromatography (SMAC), salt-gradient affinity-capture self-interaction nanoparticle spectroscopy (SGAC-SINS), and hydrophobic interaction chromatography (HIC), were used to determine the risk of hydrophobic interaction. A threshold of 10% was employed according to a study by Jain et al. (2017b) (Table 1). The antibody was labelled with a fault flag if one of the aforementioned three assay values exceeded the set threshold. We obtained 94 negative samples (0 flag) and 37 positive samples (25 with one flag, 8 with two flags, and four antibodies with exactly three flags). Figure 1 shows the detailed labelling of each antibody. To solve the problem of the dataset imbalance, 94 negative samples were randomly divided into three groups, with each group containing 31, 31, and 32 antibodies. Each sub-dataset (Group 1, Group 2, Group 3) was combined with positive samples to train three sub-models (SSH\_a, SSH\_b, SSH\_c). Then, the results of the three sub-models was integrated, and an ensemble predictor was constructed using a voting strategy.

### Feature Extraction and Selection

To construct an efficient prediction tool, appropriate feature extraction methods for transforming sequence data into numerical expressions (ideally, without distortion), in addition to a reliable benchmark data set, are crucial. Features based on sequence information such as the amino acid composition and pseudo amino acid components (He et al., 2019; Dzisoo et al., 2020; Wang et al., 2020), displayed good performance in protein and peptide classification (He et al., 2016; Li et al., 2017; Kang et al., 2019). Based on a large number of experimental results, the CKSAAGP (composition of k-spaced amino acid group pairs) (Chen et al., 2009; Chen et al., 2018) demonstrated the best performance in the present study. In the CKSAAGP encoding scheme, 20 amino acids were divided into the following five groups according to their physicochemical properties: g1: aliphatic group (GAVLMI); g2: aromatic group (FYW); g3: positive charge group (KRH); g4: negative charged group (DE); g5: uncharged group (STCPNQ) (Chen et al., 2018). Then, the frequency of amino acid group pairs separated by k residues was calculated (the default maximum value of k was set as 5). CKSAAGP can be defined as follows:

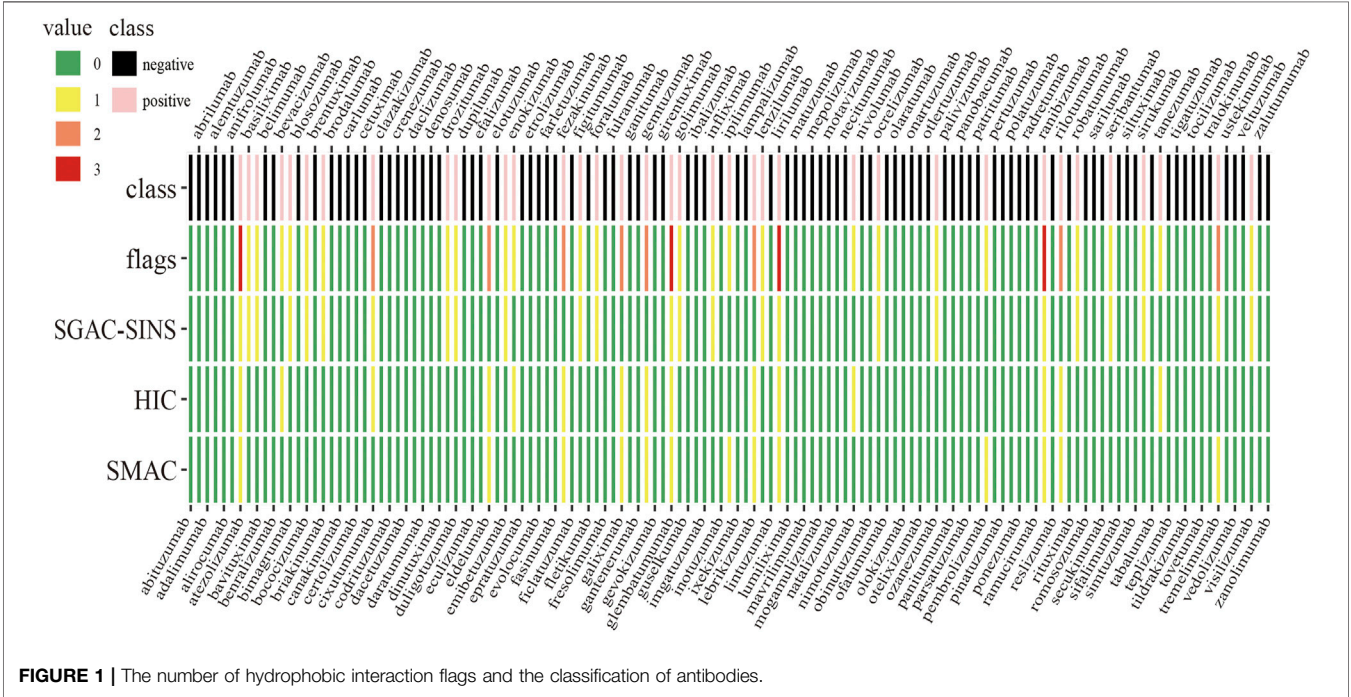
$$\left( \frac{N_{g1g1gap0}}{N_{total}}, \frac{N_{g1g2gap0}}{N_{total}}, \frac{N_{g1g3gap0}}{N_{total}}, \dots, \frac{N_{g5g4gap5}}{N_{total}}, \frac{N_{g5g5gap5}}{N_{total}} \right)$$

where  $N_{g1g1gap0}$  represents the number of times that the composition of the residue pair g1g1 is separated by 0 amino acids in the whole protein sequence;  $N_{total}$  represents the total number of k-spaced amino acid pairs. For a protein of length P, k = 0, 1, 2, 3, 4, and 5, and the values of  $N_{total}$  are P-1, P-2, P-3, P-4, P-5, and P-6, respectively. CKSAAGP can be used to encode unequal length sequences.

To compare the influence of different feature extraction algorithms, we used 19 feature extraction methods on the same

**TABLE 1 |** Three experimental thresholds for evaluating the hydrophobic interaction of antibodies (Jain et al., 2017b).

Assay	Worst 10% threshold	Units (flag)
Standup monolayer adsorption chromatography (SMAC)	12.8	Retention time (min) (>)
Salt-gradient affinity-capture self-interaction nanoparticle spectroscopy (SGAC-SINS)	370	Salt concentration (mM) (<)
Hydrophobic interaction chromatography (HIC)	11.7	Retention time (min) (>)



dataset and constructed 19 models. The feature extraction methods tested in this study are AAC, DPC, TPC, CKSAAP, DDE, GAAC, GDPC, GTPC, Moran, Geary, NMBroto, CTDC, CTDT, CTDD, CTriad, KSCTriad, SOCNumber, QSOrder, and PAAC. All feature extraction processes were performed using the iFeature (Chen et al., 2018) python package, which can be obtained from github (<https://github.com/Superzchen/iFeature/>).

High-dimensional small sample data usually cause the problem such as overfitting, longer training time and redundant features. In this study, an integrated method MRMD2.0 developed by He et al. (2020) was used for feature sorting and dimension reduction. MRMD2.0 represents different feature ranking with directed graph. Then the PageRank algorithm was used to obtain the new ranking. Finally, sequential forward selection (SFS) was used to select the optimal feature subset.

### Support Vector Machine Model Establishment

Owing to a high prediction accuracy and simple parameter optimisation, support vector machine (SVM) has been applied extensively in many fields such as protein–protein interactions (Romero-Molina et al., 2019), drug discovery (Patel et al., 2020),

and medical image processing (Yang et al., 2019). The basic idea of SVM is to determine the hyperplane with the largest interval in the space, which can divide positive and negative samples effectively and accurately. We employed LIBSVM (Chang and Lin., 2011) to construct the SVM sub-models. Among the given four kernel functions, we chose the radial basis function (RBF) kernel to obtain the optimal kernel parameter  $\gamma$  and penalty parameter  $C$ . Three sub-models were integrated through the voting strategy. The results of the three sub-models were integrated, and an antibody was predicted to have high risk of hydrophobic interaction if it was predicted as a positive sample by at least two models.

### Performance Evaluation

Leave-one-out cross-validation (LOOCV) was adopted to assess the performance of each sub-model. One sample in the sub-dataset was used as the test set, whereas the remaining samples constituted the training set. This process was repeated  $N$  times (where  $N$  is the number of samples). Eventually, the average prediction accuracy was considered as the final accuracy of the sub-model. The performance of the prediction models was evaluated using the common indicators, namely sensitivity ( $S_n$ ), specificity ( $S_p$ ), accuracy (ACC), and Matthews correlation coefficient (MCC). MCC is a relatively balanced



indicator for prediction that is mainly used to measure dichotomy. It comprehensively considers TP, TN, FP, and FN, which can avoid sample imbalance deviation. These indicators can be expressed as follows:

$$\begin{aligned} Sn &= \frac{TP}{TP + FN} \\ Sp &= \frac{TN}{TN + FP} \\ ACC &= \frac{TN + TP}{TP + FN + TN + FP} \\ MCC &= \frac{TN \times TP - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \end{aligned}$$

where TP and TN represent the number of positive data and negative data, respectively, that were predicted correctly, whereas FP and FN represent the number of positive data and negative data, respectively, that were erroneously predicted. In addition, AUC (area under the ROC curve) was used to illustrate the performance of the model. ROC curve is a TPR vs FPR plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. AUC value ranges from 0 to 1. A model whose prediction efficiency is 100% has an AUC value of 1.

## Developability Index (DI) Calculation

The developability index (DI) of each antibody in a study by Jain et al. (2017b) was computed using BIOVIA Discovery Studio 2019 (BIOINFORMATICS SOCIETY OF SICHUAN PROVINCE) with the default parameters pH = 6 and  $\beta$  = 0.05. The crystal structure of each antibody, if available, was downloaded from the PDB database. For the antibodies whose crystal structure was not available, we performed homology modelling to build their structure. Spearman rank correlation was used to explore the correlation between DI and 12 experiment assays (Jain et al., 2017b). Statistical analysis was performed with R4.1.0.

## Online Web Service

To facilitate the use of researchers, a user-friendly web server was developed. We used HTML, CSS, PHP, JavaScript to write the interface script for web service. The data processing process script was written using *Python*.

# RESULTS

## Feature Selection Based on CKSAAGP

From a total of 150 features, the optimal feature was selected using MRMD2.0. Finally, the three sub-datasets were respectively composed of 29, 31, and 35 features. **Figure 2** shows the variation of ACC with feature number during the sequential forward selection process. After feature selection, AUC was increased by at least 12% (Group 3) compared with the previous value. The prediction accuracy of the model increased with a decrease in the number of features. The small number of features also reduced the computational cost, model complexity, and the risk of overfitting. The feature dimensions of the sub-datasets were all

reduced by more than 70%, which demonstrated that the performance of MRMD2.0 was excellent.

## Model Evaluation

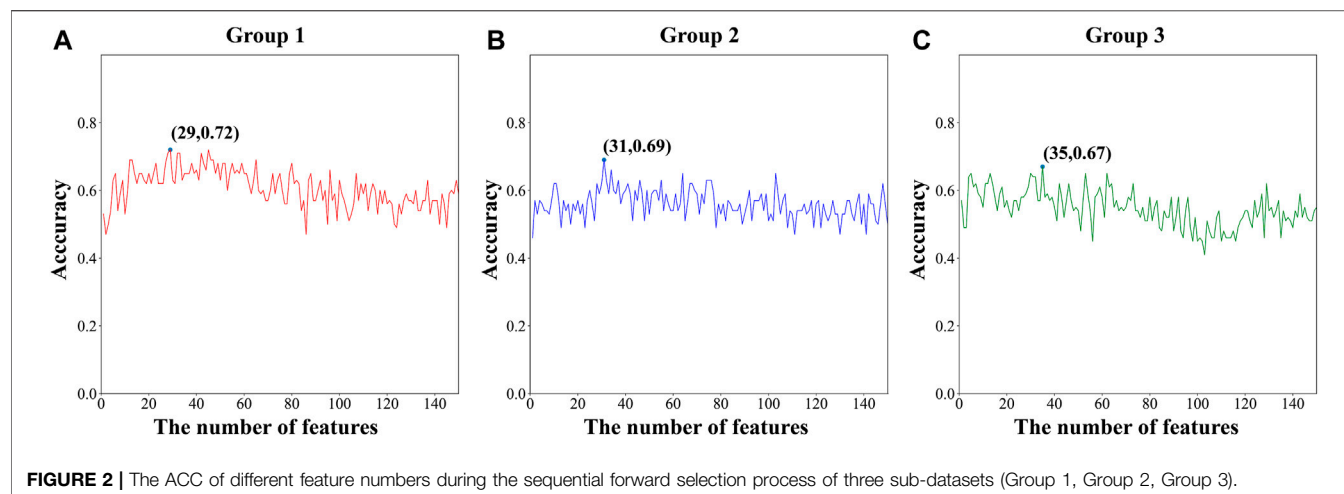
We trained three SVM sub-models based on LOOCV using the optimal features. As shown in **Table 2**, the accuracy rates of SSH\_a, SSH\_b and SSH\_c for the prediction of antibody hydrophobic interaction were 80.88, 77.94 and 75.36% respectively. By considering all samples as input of each sub-model, we obtained three prediction results. To visually demonstrate the ability of each sub-model to predict the hydrophobic interaction, a receiver operating characteristics (ROC) curve was drawn (**Figure 3**). The AUC value of SSH\_a, SSH\_b and SSH\_c reached 0.8583, 0.8956, and 0.8726, respectively. According to the aforementioned analysis, an ensemble model called SSH2.0 was constructed based on voting strategy. The sensitivity of the ensemble model was 100.00%, indicating that SSH2.0 can correctly identify all antibodies with a risk of hydrophobic interaction (**Table 2**).

## Comparison of Different Feature Extraction Methods

To comprehensively evaluate the effect of the CKSAAGP algorithm, we compared it with the other 19 feature extraction algorithms. **Figure 4** shows the feature dimension and dimension decline percentage obtained using all 20 algorithms after the reduction of MRMD2.0. The dimensions of multiple methods were reduced by more than 70%; however, the number of features varied among the three sub-datasets. For example, the number of TPC features decreased from 8,000 to 71 and 75 in Group 1 and Group 2, respectively, whereas that in Group 3 was 231. These results indicated that all feature extraction algorithms were affected by the samples, whereas CKSAAGP had smaller feature dimensions in all three sub-datasets with smaller variance, which was relatively robust. Furthermore, we assessed the ensemble model based on all 20 algorithms. As shown in **Table 3**, although the sensitivity of multiple features had reached 100%, CKSAAGP showed the highest specificity, accuracy, MCC and AUC of 77.66%, 83.97%, 0.7093, and 0.8883, respectively. Taken together, CKSAAGP was the most proper feature type for this problem, considering feature dimensions and the performance of sub-models and ensemble model.

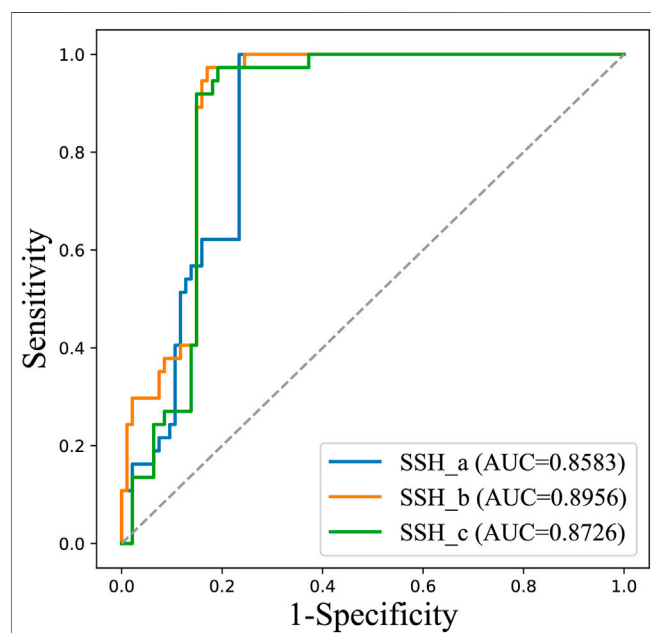
## CKSAAGP Features That Closely Related to the Hydrophobic Interaction

The properties of amino acid side chains are closely related to the structure and function of proteins. The nonpolar amino acids (aliphatic, and aromatic amino acids) are usually hydrophobic. Conversely, the polar amino acids (positively and negatively charged and uncharged amino acids) are hydrophilic. Among all the features in models, aliphatic, aliphatic.gap5, aromatic, aliphatic.gap3, negativecharger, aliphatic.gap1 were present in all sub-models, and only one of these features, namely aromatic, aliphatic.gap3, was in the top 10 features (**Table 4**). The binding of nonpolar amino acids with strong hydrophobicity increases the

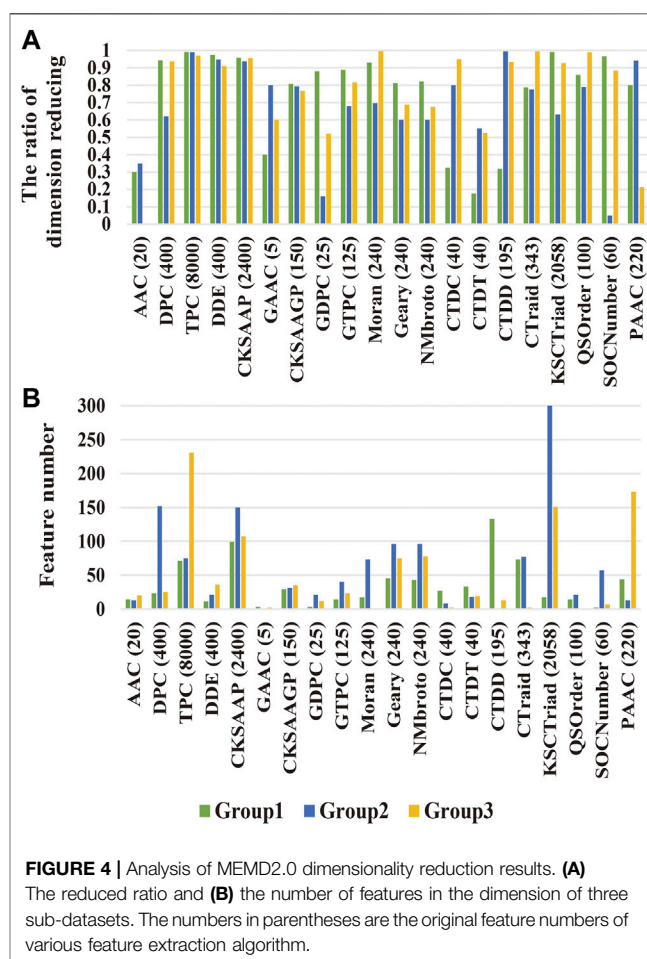


**TABLE 2 |** The prediction performance of three sub-models evaluated through leave-one-out cross-validation and that of the ensemble model evaluated through voting strategy.

Model	Sn(%)	Sp (%)	ACC(%)	MCC	AUC
SSH_a	81.08	80.64	80.88	0.6159	0.8086
SSH_b	81.08	74.19	77.94	0.5544	0.7763
SSH_c	78.37	71.87	75.36	0.5038	0.7513
SSH2.0	100.00	77.66	83.97	0.7039	0.8883



hydrophobicity of the protein. Interestingly, as shown in **Table 4**, the combination “polar + nonpolar” appeared frequently, which indicated that a polar amino acid and a nonpolar amino acid are



separated by several amino acids in space that probably enhances the hydrophobicity of the protein, although a single polar amino acid is hydrophilic. In summary, if the CKSAAGP features listed in **Table 4** appear frequently in an antibody sequence, the antibody should be excluded from early development.

**TABLE 3 |** The prediction performance of the ensemble model based on 20 feature extraction algorithms.

Feature	Sn (%)	Sp(%)	ACC(%)	MCC	AUC
CKSAAGP	100.00	77.66	83.97	0.7039	0.8883
CTriad	100.00	75.53	82.44	0.6825	0.8777
DPC	100.00	72.34	80.15	0.6518	0.8617
TPC	100.00	71.28	79.39	0.6419	0.8564
AAC	100.00	70.21	78.63	0.6322	0.8511
CKSAAP	100.00	69.15	77.86	0.6226	0.8457
NMBroto	97.30	69.15	77.10	0.5983	0.8322
DDE	100.00	65.96	75.57	0.5947	0.8298
GTPC	100.00	63.83	74.05	0.5767	0.8191
CTDC	97.30	65.96	74.81	0.5699	0.8163
CTDT	91.89	63.83	71.76	0.5021	0.7786
CTDD	97.30	56.38	67.94	0.4910	0.7684
Geary	100.00	53.19	66.41	0.4929	0.7660
SOCNumber	100.00	52.13	65.65	0.4850	0.7606
Moran	100.00	50.00	64.12	0.4693	0.7500
QSOrder	83.78	60.64	67.18	0.4003	0.7221
KSCTriad	100.00	40.43	57.25	0.4010	0.7021
GAAC	75.68	62.77	66.41	0.3464	0.6922
GDPC	100.00	30.85	50.38	0.3345	0.6543
PAAC	100.00	0.00	28.24	0.0000	0.5000

## Comparison Between the Previously Constructed SSH Model and DI Computational Tool

In our previous study, Dziso et al. (2020) provided a web-server named SSH based on TPC features to predict the hydrophobic interaction risk of mAbs. However, the number of features in SSH was far more than the number of samples, which indicated the probability of overfitting. In this study, we optimized the feature extraction algorithm and feature selection method to maintain the prediction accuracy with fewer features. We uniformly defined sensitivity as the ability to identify samples with hydrophobic interaction risk. As shown in **Table 5**, the number of each SSH sub-model features was more than 300, whereas the number of samples used for training was < 70. After using the CKSAAGP feature scheme and MRMD2.0 feature selection algorithm, the number of features in SSH2.0 reduced to one-tenth that of SSH. Although the ACC and AUC of the ensemble model decreased by 7.26% and 0.0737, respectively, we paid more attention to the performance to identify defective samples. The sensitivity of SSH2.0 reached 100.00%, which was 16.70% higher than that of SSH.

DI is another widely employed tool for assessing the aggregation propensity of proteins (Lauer et al., 2012). We performed the Spearman rank correlation test to explore the correlation between DI and 12 experimental assays. Surprisingly, the three most relevant assays were SMAC, SGAC-SINS and HIC (**Figure 5**), which we used to assess the hydrophobic interaction risk of mAbs in the current study. The result confirmed that protein aggregation is mainly driven by hydrophobic interactions (Hebditch et al., 2019). According to the methods based on the experimental data presented by Jain et al. (2017b), 37 antibodies were flagged with hydrophobic interaction warnings. We used this as the gold standard. Because high DI values correspond to low developability (Lauer et al., 2012), we sorted all the antibodies

**TABLE 4 |** The top 10 CKSAAGP features of three sub-models. The features marked in red indicate that they exist in at least two sub-models (neg: negative charged group; pos: positive charge group).

SSH_a	SSH_b	SSH_c
aromatic.uncharge.gap0	aromatic.aliphatic.gap1	aliphatic.pos.gap0
uncharge.uncharge.gap0	aliphatic.neg.gap3	uncharge.aliphatic.gap4
aromatic.aliphatic.gap3	pos.aliphatic.gap2	uncharge.aromatic.gap2
pos.neg.gap0	uncharge.uncharge.gap2	neg.aromatic.gap5
aliphatic.aromatic.gap5	aliphatic.pos.gap0	pos.uncharge.gap5
uncharge.uncharge.gap2	neg.uncharge.gap4	aliphatic.uncharge.gap5
pos.uncharge.gap0	aliphatic.aromatic.gap5	aromatic.aliphatic.gap3
pos.uncharge.gap4	neg.aliphatic.gap2	aliphatic.uncharge.gap1
neg.pos.gap2	aromatic.uncharge.gap2	aliphatic.aliphatic.gap2
aliphatic.uncharge.gap5	aromatic.pos.gap1	neg.neg.gap3

according to the descending order of their DI values. The top 37 antibodies with high DI values were predicted to have the hydrophobic interaction risk. However, the prediction performance of the DI method was inferior to that of SSH2.0. The accuracy rates of SSH2.0 and DI were 83.97 and 61.83%, respectively. The results suggest that owing to the low prediction accuracy, the application of DI to a screening platform would lead to many antibodies with a high aggregation risk being incorrectly selected.

## Web-Server Guidance

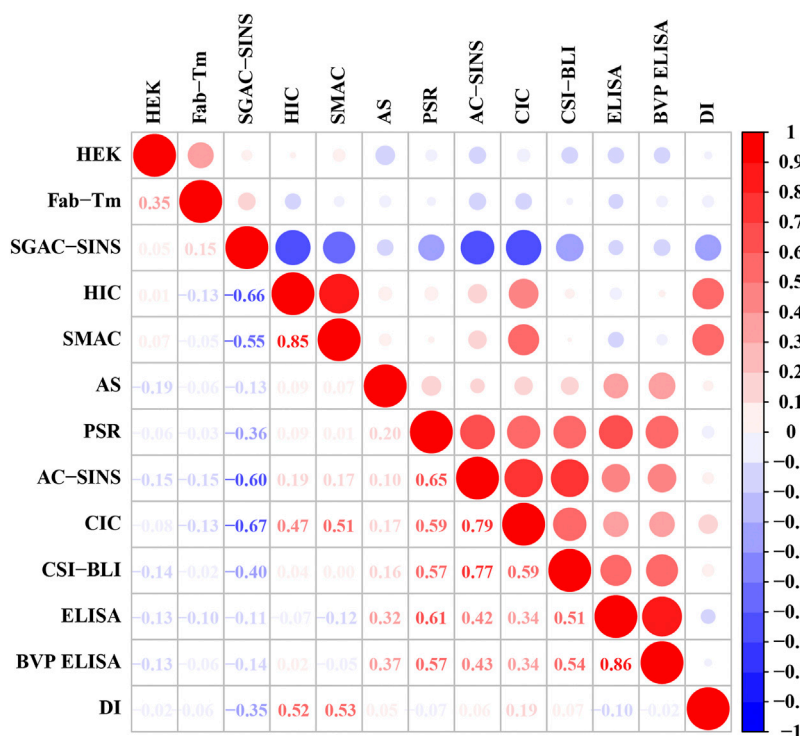
To serve the relevant researchers, we established a user-friendly web server for the prediction of hydrophobic interaction risk of mAbs. The server is freely accessible at <http://i.uestc.edu.cn/SSH2/>. The homepage of SSH2.0 is shown in **Figure 6A**. The variable region sequences of heavy chains and light chains were input separately. Because some antibodies only have one chain, the input consisting of single heavy or light chain were allowed. The submitted antibody sequences were in the FASTA format. The AbRSA tool can help in antibody numbering and CDR (complementarity-determining region) delimiting (Li et al., 2019). SSH2.0 allowed the detection of illegal characters, and only 20 common amino acids were found to be legal for sequence input. Illegal characters such as B, J, O, U, X, Z and the numbers 1–9 were forbidden (**Figure 6B**). **Figure 6C** shows the prediction results.

## DISCUSSION

The developability assessment is performed mainly to evaluate the biochemical and biophysical properties of mAbs and to select the lead antibody with ideal efficacy, safety, pharmacokinetic characteristics, and physicochemical characteristics to meet the technical requirements of the production and preparation processes (Xu et al., 2019). Various experimental strategies have been used to identify the unfavourable physicochemical properties of mAbs. However, experimental assays are time-consuming, expensive, and laborious. Computational methods can provide rapid and highly economic evaluation results and thus are expected to promote the development of antibodies (Krawczyk et al., 2017). DI is a well-known in silico tool for assessing the aggregation propensity of therapeutic antibodies

**TABLE 5** | Comparison of the feature and performance between SSH2.0 and SSH.

Model	Feature	Feature extraction method	Feature number of sub-models	Sn(%)	Sp(%)	ACC(%)	AUC
SSH	TPC	f -scores	313,315,315	84.30	96.39	91.23	0.9620
SSH2.0	CKSAAGP	MRMD2.0	29,31,35	100.00	77.66	83.97	0.8883

**FIGURE 5** | Correlation coefficient matrix of DI and 12 experimental assays. The lower triangle shows the spearman correlation coefficients, and the upper triangle represents the corresponding correlation values. The radius of the circles is proportional to the magnitude of the correlation coefficient. Red represents a positive correlation, and blue represents a negative correlation.

and it is based on the principles that protein aggregation is mainly driven by hydrophobic interactions. Regretfully, this tool relies on the antibody structure and runs slowly. Moreover, it is an expensive tool, which makes its application limited for high-throughput screening of mAbs at the early developmental stage.

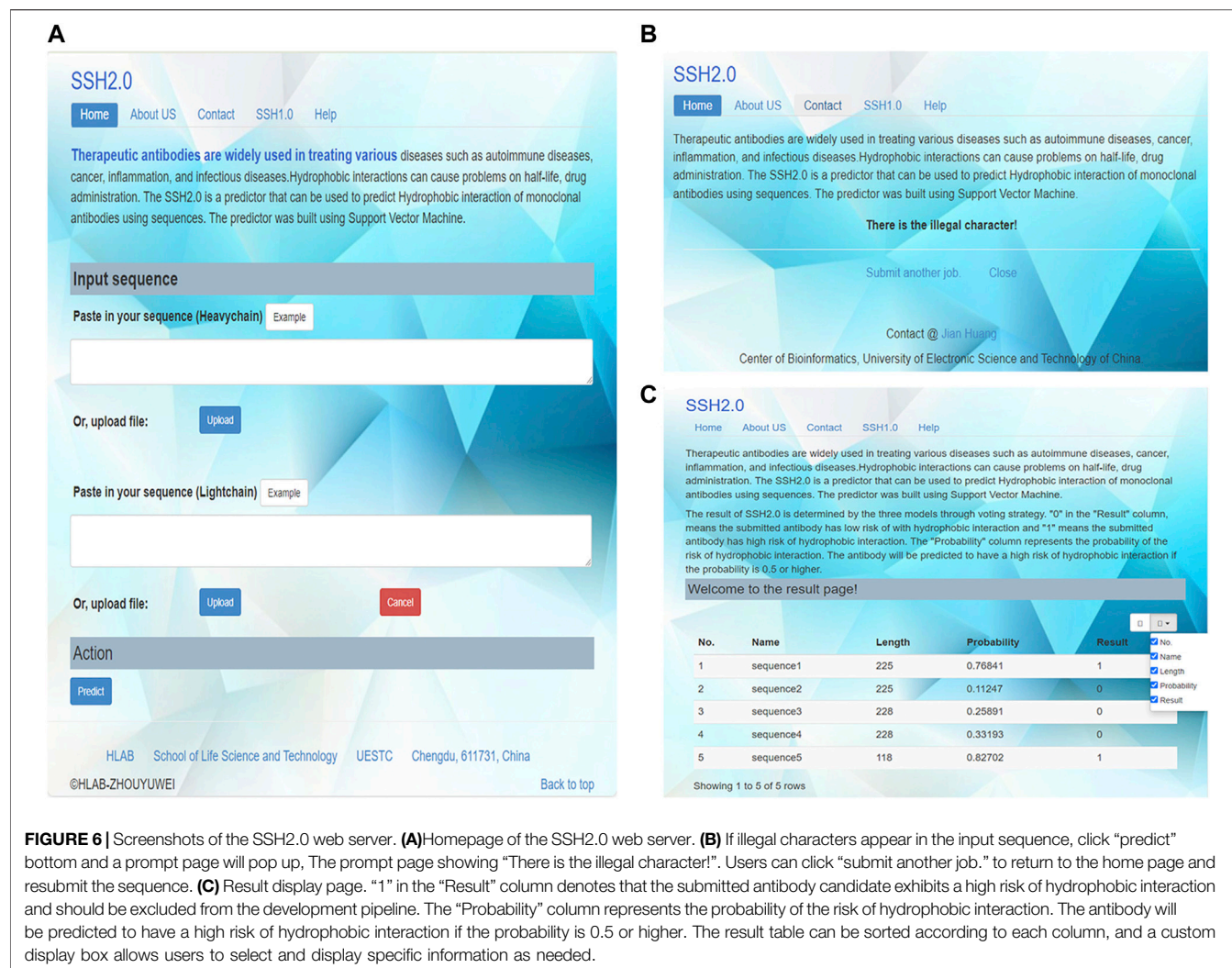
Currently, data mining and machine learning are widely applied in antibody development research (Dzisoo et al., 2021). Lecerf et al. (2019) confirmed that the sequence characteristics of the antibody variable region can determine the physicochemical properties of therapeutic antibodies. Obrezanova et al. (2015) constructed a model to predict the aggregation propensity based on the antibody sequence, and the AUC of the best AdaBoost model reached 0.76. Furthermore, Jain et al. (2017a) constructed a model to predict the solvent-accessible surface area of each amino acid residue in the variable region based on the amino acid sequence of the antibody and predicted the hydrophobic interaction of antibodies through simple logistic regression. However, aforementioned tools do not provide available model or sever.

The hydrophobic interaction prediction model constructed in the present study was trained on sequence only and eliminated

the requirement of 3D protein structure, thereby saving the computation resources. The high sensitivity usually corresponds to the low specificity. The sensitivity of SSH2.0 reached 100.00%, which indicated that the SSH2.0 prediction result may have more false positives. However, the high sensitivity of SSH2.0 is acceptable or even preferred because the main purpose of this tool is to exclude antibodies with a risk of unfavourable hydrophobic interactions. In addition, after the step of modern mAb discovery, usually tens of thousands of therapeutic antibody candidates remain to be evaluated, and the presence of even more false positives in SSH2.0 prediction results is affordable. In summary, we propose that SSH2.0 is an efficient model for predicting the hydrophobic interaction risk of mAbs.

The hydrophobic interaction risk predictor SSH2.0 constructed in this study for therapeutic mAb development is a powerful tool for selection of the antibody drug candidates with a high risk of hydrophobic interaction. This free tool based on the antibody sequence might be a better and faster alternative to the existing DI computational tool. We expect that the newer version





of this tool can be used to identify reasonable mutants with a decreased risk of hydrophobic interaction. Because the number of proven therapeutic antibodies is limited, and the experiment assays vary across batches, we also expect the tool can be assessed by an independent dataset in future.

## CONCLUSION

In this study, we developed SSH2.0, a SVM-based ensemble model trained with CKSAAGP features, for predicting the hydrophobic interaction risk of therapeutic mAbs. Compared with our previous model SSH and the widely used DI tool, SSH2.0 may be a better and robust predictor that achieved the maximum sensitivity of 100.00%, and ACC and AUC of 83.97 and 88.83%, respectively. We also developed a user-friendly web server, which is freely available at <http://i.uestc.edu.cn/SSH2/>. This tool offers a high-throughput and efficient assessment of the developability of antibodies from the perspective of hydrophobic interaction risk.

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: <https://www.pnas.org/content/114/5/944/tab-figures-data>.

## AUTHOR CONTRIBUTIONS

JH and LN conceived and designed this study. YZ and LJ wrote the manuscript. YZ, SL, and WL analyzed the data. YY wrote the interface script of web service. SX and HA drew the figures.

## FUNDING

This work was supported by grant from the National Natural Science Foundation of China (62071099).

## REFERENCES

- Carter, P. J., and Lazar, G. A. (2018). Next Generation Antibody Drugs: Pursuit of the 'high-Hanging Fruit'. *Nat. Rev. Drug Discov.* 17, 197–223. doi:10.1038/nrd.2017.227
- Chang, C.-C., and Lin, C.-J. (2011). Libsvm. *ACM Trans. Intell. Syst. Technol.* 2, 1–27. doi:10.1145/1961189.1961199
- Chen, K., Jiang, Y., Du, L., and Kurgan, L. (2009). Prediction of Integral Membrane Protein Type by Collocated Hydrophobic Amino Acid Pairs. *J. Comput. Chem.* 30, 163–172. doi:10.1002/jcc.21053
- Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., et al. (2018). iFeature: a Python Package and Web Server for Features Extraction and Selection from Protein and Peptide Sequences. *Bioinformatics* 34, 2499–2502. doi:10.1093/bioinformatics/bty140
- Dzisoo, A. M., Kang, J., Yao, P., Klugah-Brown, B., Mengesha, B. A., and Huang, J. (2020). SSH: A Tool for Predicting Hydrophobic Interaction of Monoclonal Antibodies Using Sequences. *Biomed. Res. Int.* 2020, 3508107. doi:10.1155/2020/3508107
- Dzisoo, A. M., Ren, L. P., Xie, S. Y., Zhou, Y. W., and Huang, J. (2021). Progress in Research on Evaluation of Developability of Therapeutic Antibody. *J. Univ. Electron. Sci. Techn. China* 50, 476–480.
- Hanke, A. T., Klijn, M. E., Verhaert, P. D. E. M., Van Der Wielen, L. A. M., Ottens, M., Eppink, M. H. M., et al. (2016). Prediction of Protein Retention Times in Hydrophobic Interaction Chromatography by Robust Statistical Characterization of Their Atomic-Level Surface Properties. *Biotechnol. Prog.* 32, 372–381. doi:10.1002/btpr.2219
- He, B., Kang, J., Ru, B., Ding, H., Zhou, P., and Huang, J. (2016). SABinder: A Web Service for Predicting Streptavidin-Binding Peptides. *Biomed. Res. Int.* 2016, 9175143. doi:10.1155/2016/9175143
- He, B., Chen, H., and Huang, J. (2019). PhD7Faster 2.0: Predicting Clones Propagating Faster from the Ph.D.-7 Phage Display Library by Coupling PseAAC and Tripeptide Composition. *PeerJ* 7, e7131. doi:10.7717/peerj.7131
- He, S., Guo, F., Zou, Q., and Ding, H. (2020). MRMD2.0: A Python Tool for Machine Learning with Feature Ranking and Reduction. *Curr. Bioinform.* 15, 1213–1221. doi:10.2174/1574893615999200503030350
- Hebdtich, M., Roche, A., Curtis, R. A., and Warwicker, J. (2019). Models for Antibody Behavior in Hydrophobic Interaction Chromatography and in Self-Association. *J. Pharm. Sci.* 108, 1434–1441. doi:10.1016/j.xphs.2018.11.035
- Jain, T., Boland, T., Lilov, A., Burnina, I., Brown, M., Xu, Y., et al. (2017a). Prediction of Delayed Retention of Antibodies in Hydrophobic Interaction Chromatography from Sequence Using Machine Learning. *Bioinformatics* 33, 3758–3766. doi:10.1093/bioinformatics/btx519
- Jain, T., Sun, T., Durand, S., Hall, A., Houston, N. R., Nett, J. H., et al. (2017b). Biophysical Properties of the Clinical-Stage Antibody Landscape. *Proc. Natl. Acad. Sci. USA* 114, 944–949. doi:10.1073/pnas.1616408114
- Kang, J., Fang, Y., Yao, P., Li, N., Tang, Q., and Huang, J. (2019). NeuroPP: A Tool for the Prediction of Neuropeptide Precursors Based on Optimal Sequence Composition. *Interdiscip. Sci.* 11, 108–114. doi:10.1007/s12539-018-0287-2
- Kapingidza, A. B., Kowal, K., and Chruszcz, M. (2020). Antigen-Antibody Complexes. *Subcell Biochem.* 94, 465–497. doi:10.1007/978-3-030-41769-7\_19
- Kaplon, H., Muralidharan, M., Schneider, Z., and Reichert, J. M. (2020). Antibodies to Watch in 2020. *MAbs* 12, 1703531. doi:10.1080/19420862.2019.1703531
- Krawczyk, K., Dunbar, J., and Deane, C. M. (2017). Computational Tools for Aiding Rational Antibody Design. *Methods Mol. Biol.* 1529, 399–416. doi:10.1007/978-1-4939-6637-0\_21
- Lauer, T. M., Agrawal, N. J., Chennamsetty, N., Egodage, K., Helk, B., and Trout, B. L. (2012). Developability index: a Rapid In Silico Tool for the Screening of Antibody Aggregation Propensity. *J. Pharm. Sci.* 101, 102–115. doi:10.1002/jps.22758
- Lecerf, M., Kanyavuz, A., Lacroix-Desmazes, S., and Dimitrov, J. D. (2019). Sequence Features of Variable Region Determining Physicochemical Properties and Polyreactivity of Therapeutic Antibodies. *Mol. Immunol.* 112, 338–346. doi:10.1016/j.molimm.2019.06.012
- Li, L., Chen, S., Miao, Z., Liu, Y., Liu, X., Xiao, Z. X., et al. (2019). AbRSA: A Robust Tool for Antibody Numbering. *Protein Sci.* 28, 1524–1531. doi:10.1002/pro.3633
- Li, N., Kang, J., Jiang, L., He, B., Lin, H., and Huang, J. (2017). PSBinder: A Web Service for Predicting Polystyrene Surface-Binding Peptides. *Biomed. Res. Int.* 2017, 5761517. doi:10.1155/2017/5761517
- Lienqueo, M. E., Mahn, A., Navarro, G., Salgado, J. C., Perez-Acle, T., Rapaport, I., et al. (2006). New Approaches for Predicting Protein Retention Time in Hydrophobic Interaction Chromatography. *J. Mol. Recognit.* 19, 260–269. doi:10.1002/jmr.776
- Lu, R.-M., Hwang, Y.-C., Liu, I.-J., Lee, C.-C., Tsai, H.-Z., Li, H.-J., et al. (2020). Development of Therapeutic Antibodies for the Treatment of Diseases. *J. Biomed. Sci.* 27, 1. doi:10.1186/s12929-019-0592-z
- Mahn, A., Lienqueo, M. E., and Salgado, J. C. (2009). Methods of Calculating Protein Hydrophobicity and Their Application in Developing Correlations to Predict Hydrophobic Interaction Chromatography Retention. *J. Chromatogr. A* 1216, 1838–1844. doi:10.1016/j.chroma.2008.11.089
- Martinez Morales, M., Zalar, M., Sonzini, S., Golovanov, A. P., Van Der Walle, C. F., and Derrick, J. P. (2019). Interaction of a Macrocyclic with an Aggregation-Prone Region of a Monoclonal Antibody. *Mol. Pharm.* 16, 3100–3108. doi:10.1021/acs.molpharmaceut.9b00338
- Ning, L., Abagna, H. B., Jiang, Q., Liu, S., and Huang, J. (2021). Development and Application of Therapeutic Antibodies against COVID-19. *Int. J. Biol. Sci.* 17, 1486–1496. doi:10.7150/ijbs.59149
- Obrezanova, O., Arnell, A., De La Cuesta, R. G., Berthelot, M. E., Gallagher, T. R., Zurdo, J., et al. (2015). Aggregation Risk Prediction for Antibodies and its Application to Biotherapeutic Development. *MAbs* 7, 352–363. doi:10.1080/19420862.2015.1007828
- Patel, L., Shukla, T., Huang, X., Ussery, D. W., and Wang, S. (2020). Machine Learning Methods in Drug Discovery. *Molecules* 25. doi:10.3390/molecules25225277
- Romero-Molina, S., Ruiz-Blanco, Y. B., Harms, M., Münch, J., and Sanchez-Garcia, E. (2019). PPI-detect: A Support Vector Machine Model for Sequence-Based Prediction of Protein-Protein Interactions. *J. Comput. Chem.* 40, 1233–1242. doi:10.1002/jcc.25780
- Wang, Y., Kang, J., Li, N., Zhou, Y., Tang, Z., He, B., et al. (2020). NeuroCS: A Tool to Predict Cleavage Sites of Neuropeptide Precursors. *Protein Pept. Lett.* 27, 337–345. doi:10.2174/092986652666619112150636
- Xu, Y., Wang, D., Mason, B., Rossomando, T., Li, N., Liu, D., et al. (2019). Structure, Heterogeneity and Developability Assessment of Therapeutic Antibodies. *MAbs* 11, 239–264. doi:10.1080/19420862.2018.1553476
- Yang, K., Zhou, B., Yi, F., Chen, Y., and Chen, Y. (2019). Colorectal Cancer Diagnostic Algorithm Based on Sub-patch Weight Color Histogram in Combination of Improved Least Squares Support Vector Machine for Pathological Image. *J. Med. Syst.* 43, 306. doi:10.1007/s10916-019-1429-8

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

The handling editor declared a past collaboration with one of the authors JH.

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

Copyright © 2022 Zhou, Xie, Yang, Jiang, Liu, Li, Abagna, Ning and Huang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# i2APP: A Two-Step Machine Learning Framework For Antiparasitic Peptides Identification

Minchao Jiang<sup>1†</sup>, Renfeng Zhang<sup>2†</sup>, Yixiao Xia<sup>1</sup>, Gangyong Jia<sup>1</sup>, Yuyu Yin<sup>1</sup>, Pu Wang<sup>3\*</sup>, Jian Wu<sup>4\*</sup> and Ruiquan Ge<sup>1\*</sup>

<sup>1</sup>School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China, <sup>2</sup>Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, China, <sup>3</sup>Computer School, Hubei University of Arts and Science, Xiangyang, China, <sup>4</sup>MyGenostics Inc., Beijing, China

## OPEN ACCESS

### Edited by:

Alfredo Pulvirenti,  
University of Catania, Italy

### Reviewed by:

Leyi Wei,  
Shandong University, China  
Piyush Agrawal,  
National Cancer Institute (NIH),  
United States

### \*Correspondence:

Pu Wang  
nywangpu@yeah.net  
Jian Wu  
jw2231@mygeno.cn  
Ruiquan Ge  
gespring@hdu.edu.cn

<sup>†</sup>These authors have Co-first authors

### Specialty section:

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

Received: 26 February 2022

Accepted: 11 April 2022

Published: 27 April 2022

### Citation:

Jiang M, Zhang R, Xia Y, Jia G, Yin Y,  
Wang P, Wu J and Ge R (2022) i2APP:  
A Two-Step Machine Learning  
Framework For Antiparasitic  
Peptides Identification.  
Front. Genet. 13:884589.  
doi: 10.3389/fgene.2022.884589

Parasites can cause enormous damage to their hosts. Studies have shown that antiparasitic peptides can inhibit the growth and development of parasites and even kill them. Because traditional biological methods to determine the activity of antiparasitic peptides are time-consuming and costly, a method for large-scale prediction of antiparasitic peptides is urgently needed. We propose a computational approach called i2APP that can efficiently identify APPs using a two-step machine learning (ML) framework. First, in order to solve the imbalance of positive and negative samples in the training set, a random under sampling method is used to generate a balanced training data set. Then, the physical and chemical features and terminus-based features are extracted, and the first classification is performed by Light Gradient Boosting Machine (LGBM) and Support Vector Machine (SVM) to obtain 264-dimensional higher level features. These features are selected by Maximal Information Coefficient (MIC) and the features with the big MIC values are retained. Finally, the SVM algorithm is used for the second classification in the optimized feature space. Thus the prediction model i2APP is fully constructed. On independent datasets, the accuracy and AUC of i2APP are 0.913 and 0.935, respectively, which are better than the state-of-arts methods. The key idea of the proposed method is that multi-level features are extracted from peptide sequences and the higher-level features can distinguish well the APPs and non-APPs.

**Keywords:** antiparasitic peptides, feature representation, maximum information coefficient, feature selection, T-distributed stochastic neighbor embedding

## INTRODUCTION

Parasites are a very common source of disease. Parasitic diseases can affect almost all living things, including plants and mammals. The effects of parasitic diseases can range from mild discomfort to death (Momčilović et al., 2019). It is estimated that one billion people worldwide are infected with ascariasis, although it is usually harmless. *Necator americanus* and *Ancylostoma duodenale* can cause hookworm infections in humans, resulting in anemia, malnutrition, shortness of breath and weakness. This infection affects about 740 million people in the developing countries, including children and adults (Diemert et al., 2018). Malaria is very harmful to humans. It causes 300 to 500 million illnesses and about 2 million deaths each year, with about half of those deaths occurring in



children under the age of 5 (Barber et al., 2017). The main method of treating parasitic diseases today is the use of antibiotics (Zahedifard and Rafati, 2018). However, frequent use of antibiotics can increase parasite resistance and even have some undetected side effects (Ertabaklar et al., 2020). Studies have found that anti-parasite peptide (APP) can effectively inhibit the growth of parasites and even kill them (Lacerda et al., 2016). Anti-parasite peptides are usually composed of 5–50 amino acids and are relatively short in length. They are usually changed by antimicrobial peptides (AMPs) (Mehta et al., 2014). APPs can kill parasites by destroying the cell membrane of the parasite or inhibiting the reductase in the parasite (Bell, 2011; Torrent et al., 2012). Therefore, it is very important to be able to identify APPs.

In the past few years, many methods for predicting functional peptides based on machine learning have been proposed, such as AAPred-CNN (Lin et al., 2022) for anti-angiogenic peptides, mAHTPred (Manavalan et al., 2019) for anti-hypertensive peptides, AVPIDen (Pang et al., 2021) for anti-viral peptides. PredictFP2 can predict fusion peptide domains in all retroviruses (Wu et al., 2019). AMPfun (Chung et al., 2020) and PredAPP (Zhang et al., 2021) are proposed for antiparasitic peptides identification. Based on random forests, the AMPfun tool can be used to identify anticancer peptides, APP, and antiviral peptides. AMPfun can be used to characterize and identify antimicrobial peptides with different functional activities, but the prediction results for APPs are not very good. In 2021, (Zhang et al., 2021) proposed PredAPP, a model for predicting antiparasitic peptides using an under sampling and ensemble approach. A variety of data under sampling methods are proposed for data balance. This model adopts an ensemble approach, combining 9 feature groups and 6 machine learning algorithms, and finally achieves good results, but there is still room for improvement.

In this work, we propose a new model named i2APP for identifying APPs, which uses a two-stage machine learning framework. In the first stage, we extract dozens of feature groups for each peptide sequence, and then build the first-layer classifiers with these feature groups. The outputs of the first-layer classifiers are used as the higher-level features. What's more, MIC (Kinney and Atwal, 2014; Ge et al., 2016) is used here to filter out the insignificant features. In the second stage, with the higher-level features, we build the second-layer classifier, whose outputs are the final results of identifying APPs. Through independent test, we will find that the proposed model is better than the state-of-the-arts methods in most metrics. The tool i2APP is available at <https://github.com/greyspring/i2APP>.

## MATERIALS AND METHODS

### Datasets

A benchmark dataset is the premise for an effective and reliable model. To train our model and compare it with others, the dataset studied by (Zhang et al., 2021) were used in this work, in which 301 APPs were used as positive samples and 1909 non-APPs were negative ones. For the positive samples, 301 APPs were taken out as positive training samples, and the remaining 46 APPs were used as positive testing

samples. 46 non-APPs were randomly selected from the negative samples as negative testing samples, and the remaining 1863 non-APPs were used as negative training samples. In this way, 255 APPs and 1863 non-APPs constituted the original training set, and 46 APPs and 46 non-APPs constituted the testing set. Since the samples in the training set are very unbalanced, we use random under sampling (Tahir et al., 2012; Stilianoudakis et al., 2021) on the training set and get 255 APPs and 255 non-APPs to constitute the final training set. For the sake of simplicity, the final training dataset is marked as T255p + 255n, and the testing dataset is marked as V46p + 46n.

We take out the 5 amino acids at the N-terminus and C-terminus of each peptide sequence to compare the differences between positive and negative samples by Two Sample Logo application (Schneider and Stephens, 1990; Crooks et al., 2004), which calculates and visualizes the differences between two sets of aligned samples of amino acids or nucleotides. At each position in the aligned groups of sequences, statistically significant amino acid symbols are plotted using the size of the symbol that is proportional to the difference between the two samples. It can be seen from the comparison in **Figure 1** that the amino acid composition at both ends of the APPs and non-APPs sequences have some differences, so it can be considered to extract features from both ends of peptide sequence to distinguish the two types of samples.

### Features Representation

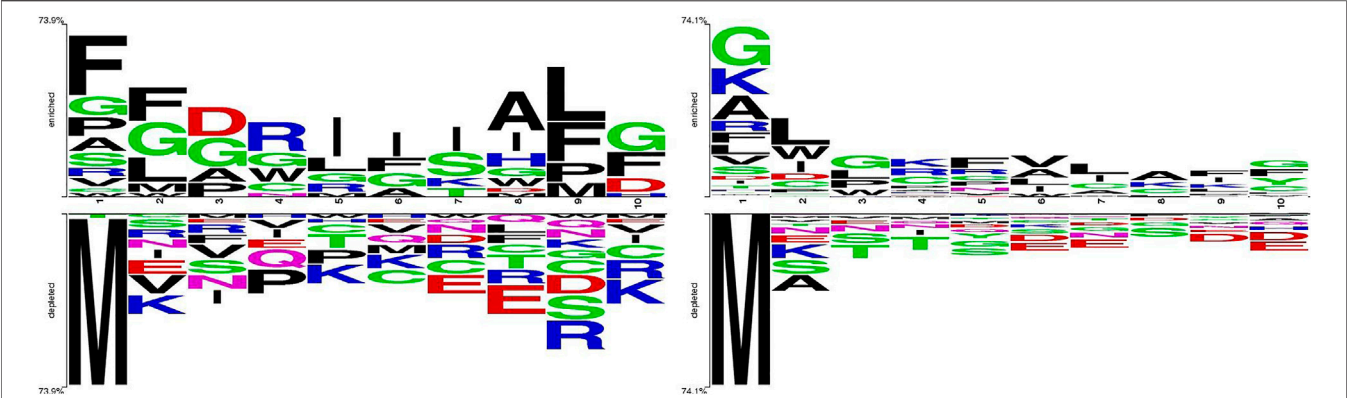
Good features are beneficial to the training of machine learning models and obtain good prediction performance. The classification of peptides mainly depends on the feature set constructed by the structural and functional properties. Extracting features from peptide sequences that effectively reflect their sequence pattern information is a challenging problem. In this study, we extract 18 kinds of physicochemical features from the peptide sequences, some of which contain very important information, such as functional domains, gene ontology and sequential evolution, etc (Liu et al., 2015; Liu et al., 2017). Thus 18 groups of sequence-based features will be obtained for each peptide sequence.

In addition, the N-terminus and C-terminus of a protein or peptide often have very important biological function, so we also extract features from the both ends of peptide sequence. In this study, we take out a fragment with three or five amino acids at the N-terminus or C-terminus of a peptide sequence, and use 12 types of feature extraction method for this fragment (Jing et al., 2019). In such a way, 48 groups of terminus-based features will be obtained for each peptide sequence.

All these feature extraction methods are listed in **Table 1**.

### Computational Models

As shown in **Figure 2**, the overall framework of i2APP includes four main steps. As a first step, the benchmark datasets are collected from various databases and literatures, and then divided into training dataset and testing dataset. To get a balanced training dataset, the random under sampling procedure is performed on the negative training samples. In the second step, we adopt 18 types of feature extraction methods on the whole peptide sequence to get 18 groups of



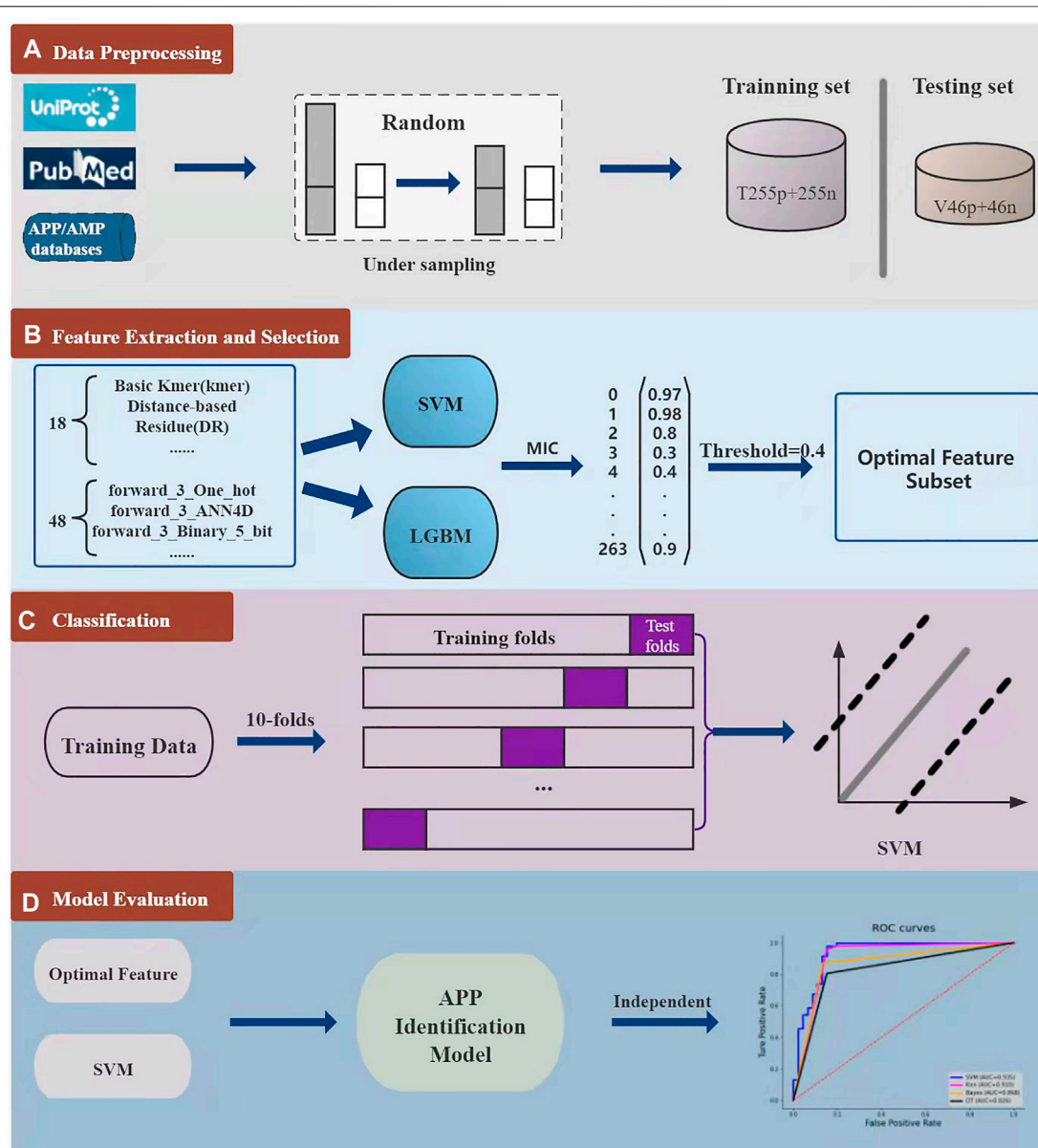
**FIGURE 1 |** Different distribution between APP and non-APP sequences. **(A)** V46p+46n **(B)** T255p+1863n.

**TABLE 1 |** Peptide sequence features.

	Features
Sequence-based	Basic Kmer (kmer) Distance-based Residue (DR) Distance Pair (DP) Auto covariance (feature-AC) Auto-cross covariance (ACC) Cross covariance (feature-CC) Physicochemical distance transformation (PDT) Parallel correlation pseudo amino acid composition (PC-PseAAC) Series correlation pseudo amino acid composition (SC-PseAAC) General parallel correlation pseudo amino acid composition (PC-PseAAC-General) General series correlation pseudo amino acid composition (SC-PseAAC-General) Select and combine the nmost frequenct aminoacids according to their frequencies (Top-n-gram) Profile-based Physicochemical distance transformation (PDT-Profile) Distance-based Top-n-gram (DT) Profile-based Auto covariance (AC-PSSM) Profile-based Cross covariance (CC-PSSM) Profile-based Distance-based Top-n-gram (PSSM-DT) Profile-based Auto-cross covariance (ACC-PSSM)
Terminus-based	One_hot One_hot_6_bit Binary_5_bit Hydrophobicity_matrix Meiler_parameters Acthely_factors PAM250 BLOSUM62 Miyazawa_energies Micheletti_potentials AESNN3 ANN4D

sequence-based features, and 12 types of feature extraction methods on the N-terminus and C-terminus of peptide sequence. Considering that all peptide sequences are at least 5 residues in length, we take 3 and 5 residues at both ends of the sequence. So, a total of 48 groups of terminal-based features are extracted. For each feature group, SVM and LGBM are trained respectively, and 132 probability outputs are got for each peptide sequence. These probabilities can be seemed as higher-level features for further classification. What’s more, the probability

greater than 0.5 is recorded as 1, and the probability less than 0.5 is recorded as 0. These binarized values help remove noise from the model. Stacking the probabilities and their binarized values, a total of 264 higher-level features are obtained. However, these higher-level features may have information redundancy, so a feature selection method is needed here to filter out the superfluous ones. In this study, the maximum information coefficient (MIC) is calculated for each feature, and the threshold is set to 0.4, that is, only the feature with the MIC



**FIGURE 2 |** The whole model consists of four parts. The first part is the collection, division and down sampling of the dataset. The second part is feature extraction and feature selection for each peptide sequence. The third part is to analyze the effect of different classifiers through 10-fold cross-validation. In the fourth part, the proposed model is evaluated through independent test.

value greater than 0.4 is retained. The third step is to use ten-fold cross-validation to select the best classifier based on the reduced higher-level feature set. The candidate include the popular classifiers, such as SVM, Bayes (Jahromi and Taheri, 2017), Decision Tree (DT) (Wang et al., 2019), K-Nearest Neighbor (KNN) (Wang et al., 2017), Random Forest (RF), Adaboost (Ada) and so on. In the fourth step, we test the effect of the proposed model on an independent test dataset, and compare its performance with other models. In this work, we used the scikit-learn package (Pedregosa et al., 2011) to implement all classifiers.

## Evaluation

In order to evaluate the results of the final classification and facilitate comparison with other models, we used five commonly used indicators in bioinformatics research (Luo et al., 2019; Yang et al., 2021), including specificity (SP), sensitivity (SN), F1 score (F1), Matthew correlation coefficient (MCC) and accuracy (ACC). The specific calculation formula of these measured values is as follows:

$$Sp = \frac{TN}{TN + FP}$$

**TABLE 2 |** The results of cross-validation on the training set with different classifiers.

	Model	ACC (%)	SN (%)	SP (%)	AUC	MCC	F1
Training Set	SVM	<b>90.0</b>	<b>93.2</b>	86.9	<b>0.952</b>	<b>0.803</b>	<b>0.900</b>
	Bayes	86.5	83.2	87.9	0.865	0.729	0.838
	Knn	86.3	93.0	80.5	0.893	0.736	0.867
	DT	82.7	82.0	84.5	0.833	0.660	0.824
	RF	87.5	91.9	83.7	0.951	0.753	0.877
	Ada	82.2	84.8	79.8	0.823	0.645	0.822

The bold values indicate the best performance.

**TABLE 3 |** The results of independent test on the testing set with different classifiers.

	Model	ACC (%)	SN (%)	SP (%)	AUC	MCC	F1
Testing Set	SVM	<b>91.3</b>	<b>97.8</b>	84.8	<b>0.935</b>	<b>0.833</b>	<b>0.918</b>
	Bayes	85.9	84.8	87.0	0.868	0.718	0.857
	Knn	89.1	97.8	80.4	0.910	0.800	0.900
	DT	82.6	80.4	84.8	0.826	0.653	0.822
	RF	88.0	93.5	82.6	0.931	0.765	0.887
	Ada	88.0	91.3	84.8	0.880	0.762	0.884

The bold values indicate the best performance.

$$Sn = \frac{TP}{TP + FN}$$

$$F1 = \frac{2TP}{2TP + FP + FN}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

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

Where TP means the number of APPs correctly predicted by the model; TN means the number of non-APPs that the model correctly predicts; FP means the number of non-APPs that the model mispredicts; FN means the number of APPs that the model mispredicts. In addition, we also use other metrics to evaluate the performance of i2APP, including receiver operating characteristic (ROC) curve (Fawcett, 2006), the area under the ROC curve (AUC) (Lobo et al., 2008), precision-recall (PR) curve (Davis and Goadrich, 2006), and the area under the PR curve (AUPR).

## RESULTS

### Effects of Different Classifiers

First, we fix the classifier of the second layer as SVM because it is very effective in small sample learning, and then compare the different classification models in the first layer. Through cross-validation experiments, it is found that the effects of SVM and LGBM are better, so we use these two classification models in the first layer. Now we can compare different classifiers in the second layer. As can be seen from **Table 2**, different classifiers are tested on the training dataset T255p + 255n through ten-fold cross-validation, and the final result is the average of ten evaluations. After parameter tuning, SVM is higher than other classifiers in most metrics, and reaches

90.0%, 0.952, 93.2%, 86.9%, 0.803, and 0.900% in ACC, AUC, SN, SP, MCC, and F1, respectively. Among all classifiers, ACC, AUC, SN, MCC, and F1 obtained by SVM achieved the first position. So we also focused on using SVM as a classifier for the independent test set.

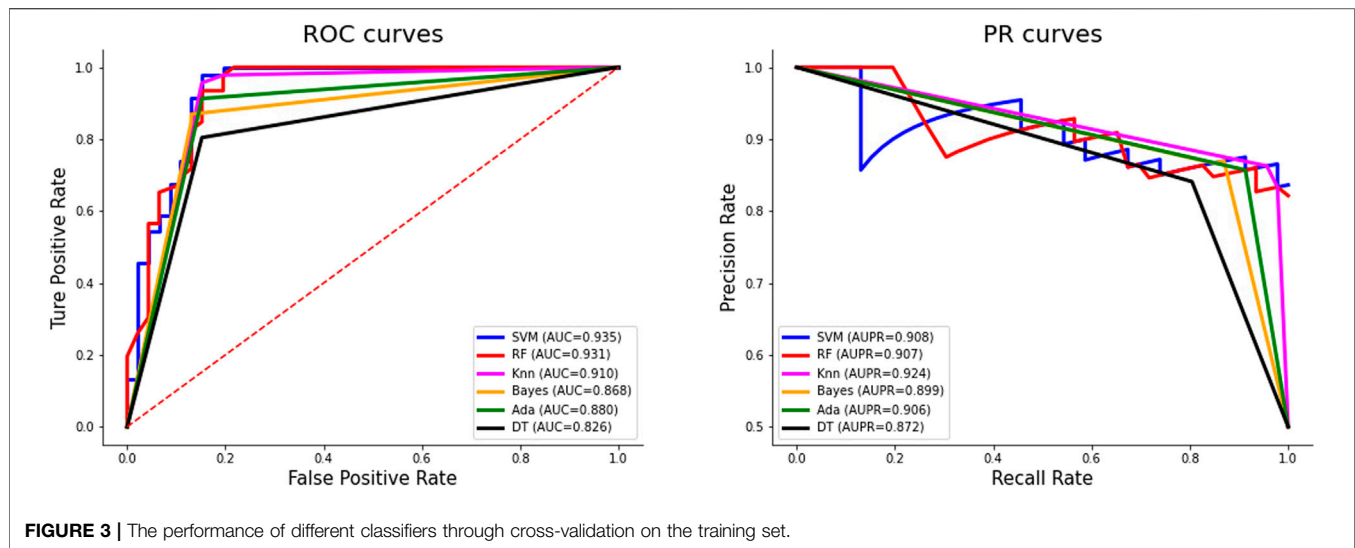
As can be seen from **Table 3**, SVM has a huge advantage over other classifiers on the independent test set V46p + 46n. The values of ACC, AUC, SN, SP, MCC, and F1 are 91.3%, 0.935, 97.8%, 84.8%, 0.833, and 0.918%, respectively. The values of ACC, AUC, SN, MCC, and F1 obtained by SVM all rank first among all classifiers. Especially MCC and AUC by SVM is 0.833 and 0.935 higher than the second-ranked classifier. The comparison of these results shows that SVM is the most suitable classifier in our work.

**Figure 3** shows the ROC curves and PR curves of different classifiers on the independent test set. The ROC curve of SVM is closest to the upper left corner, surpassing other classifiers. The AUC value of SVM is 0.935, which is the highest and 0.025 higher than the second-ranked classifier KNN. Although the AUPR value of SVM is not the largest, when the recall rate is 1, the precision rate of SVM reaches 0.836, which is the highest.

### Comparison With Other Methods

Our model is compared with others through ten-fold cross-validation on the training dataset, and the results are shown in **Table 4**. NM-BD and RUS-BD are both proposed in (Zhang et al., 2021), and the imbalanced training set was down sampled using NearMiss method (Mani and Zhang, 2003; Li et al., 2021) for the former, while the random under sampling method was used for the latter, which is also adopted in this study. Compared with RUS-BD, our model outperforms it on all metrics, with improvement of 1.8% on ACC, 0.7% on SN, 3% on SP, 1.8% on SP, 0.013 on F1, and 0.035 on MCC. When compared with NM-BD, our model is also the winner on nearly all metrics except SP. These results show that the performance of our model on the training set is better than the others on the whole.

To further verify the validity of the proposed model, we compare it with other models on an independent test dataset, and the results are shown in **Table 5**, from which we can see that the metrics of i2APP are nearly all better than that of other models. The values of ACC, SN, MCC and F1 are 17.4, 45.6, 0.302 and 0.251% higher than AMPfun, and the values of ACC, MCC, F1, and SP are 178 3.3, 0.107, 0.027, and 6.5% higher than PredAPP. All these results show that the proposed model has better generalization ability than the state-of-the-art models for APP prediction.



**TABLE 4 |** Comparison of our model with the existing methods through cross-validation on the training set.

Method	ACC (%)	SN (%)	SP (%)	MCC	F1
NM-BD	88.8	85.5	92.2	0.778	0.884
RUS-BD	88.2	92.5	83.9	0.768	0.887
<b>i2APP</b>	<b>90.0</b>	<b>93.2</b>	86.9	<b>0.803</b>	<b>0.900</b>

The bold values indicate the best performance.

**TABLE 5 |** Comparison of our model with the existing methods through independent test on the testing set.

Method	ACC (%)	SN (%)	SP (%)	MCC	F1
AMPfun	73.9	52.2	95.7	0.531	0.667
PredAPP	88.0	97.8	78.3	0.776	0.891
<b>i2APP</b>	<b>91.3</b>	<b>97.8</b>	84.8	<b>0.833</b>	<b>0.918</b>

The bold values indicate the best performance.

**TABLE 6 |** The results of ten-fold cross-validation on the balanced or unbalanced datasets.

Method	ACC (%)	SN (%)	SP (%)	MCC	F1
PredAPP (unbalanced)	91.9	52.5	97.3	0.574	0.609
i2APP (balanced)	90.0	93.2	86.9	0.803	0.900
i2APP (unbalanced)	96.5	76.7	99.3	0.826	0.839

## Impact of Dataset Balancing

We performed 10-fold cross-validation on the original dataset containing 255 APPs and 1863 non-APPs, and the results were listed in **Table 6**. It can be found that compared with the balanced dataset, the SP, MCC and ACC metrics have a greater improvement on the unbalanced dataset. However,

**TABLE 7 |** The results of independent test using the balanced or unbalanced datasets as the training set.

Method	ACC (%)	SN (%)	SP (%)	MCC	F1
i2APP (balanced)	91.3	97.8	84.8	0.833	0.918
i2APP (unbalanced)	93.5	100.0	87.0	0.877	0.939

because there are too few positive samples, the SE metric decreases a lot. In addition, our model achieves large improvements in various metrics compared to the model PredAPP (IMBD) (Zhang et al., 2021) using the same unbalanced dataset.

With the unbalanced dataset as the training set, we tested the proposed model on the independent test set including 46 APPs and 46 non-APPs and listed the results in **Table 7**, from which we can see that whether using balanced or unbalanced training sets, i2APP has good generalization ability.

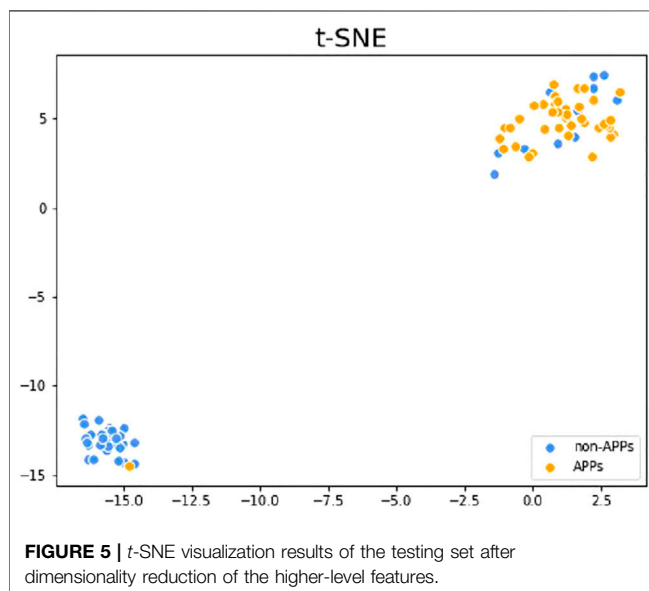
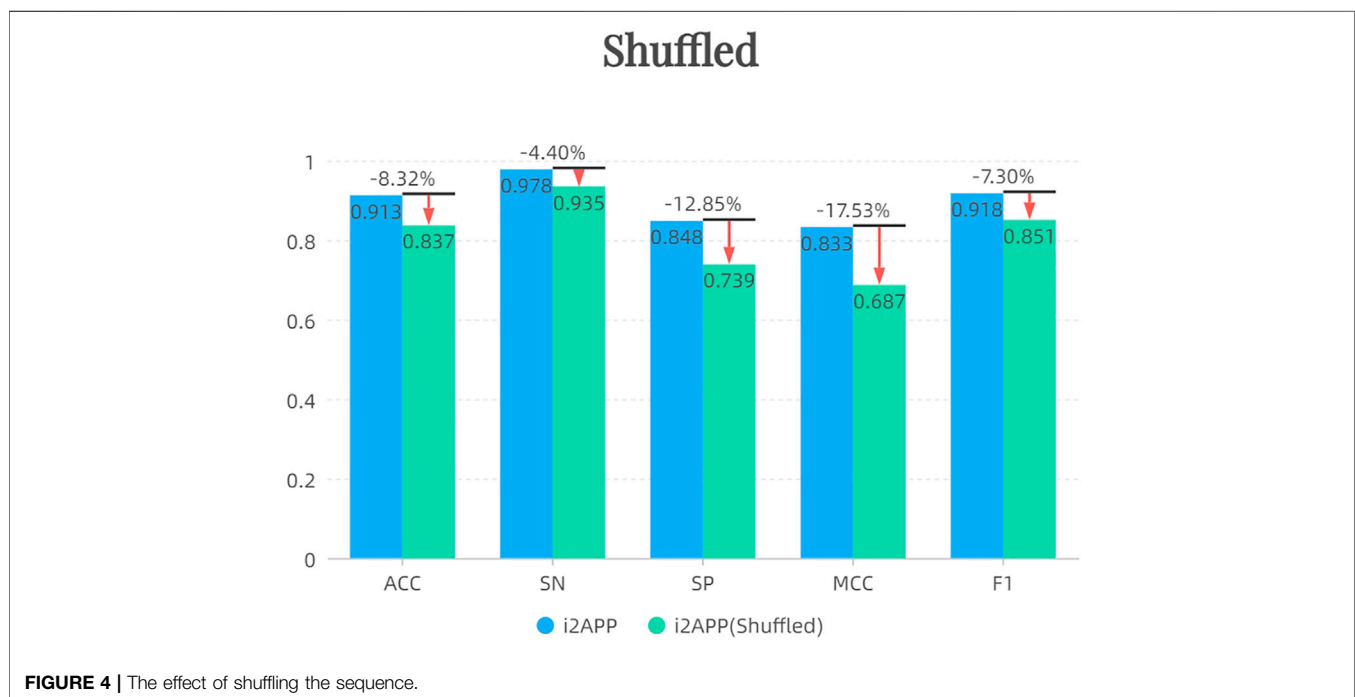
## Impact of Shuffled Sequence

After shuffling the sequence of negative samples in the training set, we randomly sampled 255 new negative samples to form the training set together with 255 positive samples. The results of independent test are shown in **Figure 4**. It can be seen that the performance of the model decreases after using the shuffled negative samples, probably because the effect of the terminus-based features is reduced after the sequence is shuffled.

## Interpretability Analysis

T-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten and Hinton, 2008) is a very popular data visualization tool that can reduce high-dimensional data to 2-3 dimensions, so as to draw samples on a plane or 3D space and observe the sample distribution. **Figure 5** shows the





visualization results of the test dataset V46p + 46n after dimensionality reduction on the higher-level features, which are the outputs of the first layer classification. The orange points in the figure are APPs, and the blue points are non-APPs. As can be seen from the figure, the two types of samples can be well distinguished with the higher-level features, so that our model can achieve better performance. What's more, it can be found that the aggregation degree of

APPs is higher than that of non-APPs, indicating that it is easier to identify APPs than non-APPs, so the metric SN in our model will be higher than SP.

## CONCLUSION

In this study, we propose a novel model named i2APP to identify APPs efficiently. The main structure of this work consists of four steps. Firstly, the random under sampling method is used to balance the training set. Secondly, a variety of sequence-based and terminus-based features are extracted from any peptide sequence, and then enter these raw features into the first layer classifiers, SVM and LGBM, to get the higher-level features. The maximum information coefficient (MIC) is calculated for each higher-level feature, and only the significant features are retained. Thirdly, based on the optimal feature subset, several popular classifiers are evaluated through cross-validation on the training dataset, and SVM is chosen as the second layer classifier. Finally, independent test is performed on the proposed model and the others, and we can see that i2APP has better generalization ability than the state-of-the-art models for APP prediction. The sequence features used in this paper are all extracted by hand, and some of them are quite complex. Although we simplify the model by two-step learning and feature selection, the overall model still looks complicated. In the future, as the amount of data increases, the RNN or Transformer model can be used for automatic feature learning, which may further improve the accuracy of APP recognition.

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: <https://github.com/greyspring/i2APP/tree/master/datasets>.

## AUTHOR CONTRIBUTIONS

RG and PW designed the method and Supervised the whole project. MJ and YX developed the prediction models. RZ, GJ, YY, and JW analysed the data and results. RZ and JW participated in the design, helped in writing the

manuscript. All authors have read and approved the revised manuscript.

## FUNDING

This work has been supported by the Zhejiang Provincial Natural Science Foundation of China (No. LY21F020017, 2022C03043), the National key research and development program of China (No. 2019YFC0118404, 2019YFC0118403), Joint Funds of the Zhejiang Provincial Natural Science Foundation of China (U20A20386), National Natural Science Foundation of China (No. 61702146).

## REFERENCES

- Barber, B. E., Rajahram, G. S., Grigg, M. J., William, T., and Anstey, N. M. (2017). World Malaria Report: Time to Acknowledge Plasmodium Knowlesi Malaria. *Malar. J.* 16 (1), 135. doi:10.1186/s12936-017-1787-y
- Bell, A. (2011). Antimalarial Peptides: the Long and the Short of it. *Cpd* 17 (25), 2719–2731. doi:10.2174/138161211797416057
- Chung, C.-R., Kuo, T.-R., Wu, L.-C., Lee, T.-Y., and Horng, J.-T. (2020). Characterization and Identification of Antimicrobial Peptides with Different Functional Activities. *Brief. Bioinformatics* 21 (3), 1098–1114. doi:10.1093/bib/bbz043
- Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). WebLogo: A Sequence Logo Generator: Figure 1. *Genome Res.* 14 (6), 1188–1190. doi:10.1101/gr.849004
- Davis, J., and Goadrich, M. (2006). “The Relationship between Precision-Recall and ROC Curves,” in Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, PA: Association for Computing Machinery, 233–240.
- Diemert, D., Campbell, D., Brelsford, J., Leasure, C., Li, G., Peng, J., et al. (2018). “Controlled Human Hookworm Infection: Accelerating Human Hookworm Vaccine Development,” in *Open Forum Infectious Diseases* 5 (5). doi:10.1093/ofid/ofy083
- Ertabaklar, H., Malatyali, E., Malatyali, E., and Ertug, S. (2020). Drug Resistance in Parasitic Diseases. *Eur. J. Ther.* 26, 1–5. doi:10.5152/eurjther.2019.18075
- Fawcett, T. (2006). An Introduction to ROC Analysis. *Pattern recognition Lett.* 27 (8), 861–874. doi:10.1016/j.patrec.2005.10.010
- Ge, R., Zhou, M., Luo, Y., Meng, Q., Mai, G., Ma, D., et al. (2016). McTwo: a Two-step Feature Selection Algorithm Based on Maximal Information Coefficient. *BMC bioinformatics* 17 (1), 142. doi:10.1186/s12859-016-0990-0
- Jahromi, A. H., and Taheri, M. (2017). “A Non-parametric Mixture of Gaussian Naive Bayes Classifiers Based on Local Independent Features,” in *Artificial Intelligence and Signal Processing Conference* (Shiraz, Iran: AISP IEEE), 209–212. doi:10.1109/aisp.2017.8324083
- Jing, X., Dong, Q., Hong, D., and Lu, R. (2019). Amino Acid Encoding Methods for Protein Sequences: a Comprehensive Review and Assessment. *Ieee/acm Trans. Comput. Biol. Bioinform* 17 (6), 1918–1931. doi:10.1109/TCBB.2019.2911677
- Kinney, J. B., and Atwal, G. S. (2014). Equitability, Mutual Information, and the Maximal Information Coefficient. *Proc. Natl. Acad. Sci. U.S.A.* 111 (9), 3354–3359. doi:10.1073/pnas.1309933111
- Lacerda, A. F., Pelegri, P. B., de Oliveira, D. M., Vasconcelos, É. A. R., and Grossi-de-Sá, M. F. (2016). Anti-parasitic Peptides from Arthropods and Their Application in Drug Therapy. *Front. Microbiol.* 7, 91. doi:10.3389/fmicb.2016.00091
- Li, M., Wu, Z., Wang, W., Lu, K., Zhang, J., Zhou, Y., et al. (2021). Protein-Protein Interaction Sites Prediction Based on an Under-Sampling Strategy and Random Forest Algorithm. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 1–1. doi:10.1109/tcbb.2021.3123269
- Lin, C., Wang, L., and Shi, L. (2022). AAPred-CNN: Accurate Predictor Based on Deep Convolution Neural Network for Identification of Anti-angiogenic Peptides. *Methods*. doi:10.1016/j.ymeth.2022.01.004
- Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K.-C. (2015). Pse-in-One: a Web Server for Generating Various Modes of Pseudo Components of DNA, RNA, and Protein Sequences. *Nucleic Acids Res.* 43 (W1), W65–W71. doi:10.1093/nar/gkv458
- Liu, B., Wu, H., and Chou, K.-C. (2017). Pse-in-One 2.0: an Improved Package of Web Servers for Generating Various Modes of Pseudo Components of DNA, RNA, and Protein Sequences. *Ns* 09 (04), 67–91. doi:10.4236/ns.2017.94007
- Lobo, J. M., Jiménez-Valverde, A., and Real, R. (2008). AUC: a Misleading Measure of the Performance of Predictive Distribution Models. *Glob. Ecol. Biogeogr.* 17 (2), 145–151. doi:10.1111/j.1466-8238.2007.00358.x
- Luo, F., Wang, M., Liu, Y., Zhao, X.-M., and Li, A. (2019). DeepPhos: Prediction of Protein Phosphorylation Sites with Deep Learning. *Bioinformatics* 35 (16), 2766–2773. doi:10.1093/bioinformatics/bty1051
- Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019). maHTPRED: a Sequence-Based Meta-Predictor for Improving the Prediction of Anti-hypertensive Peptides Using Effective Feature Representation. *Bioinformatics* 35 (16), 2757–2765. doi:10.1093/bioinformatics/bty1047
- Mani, I., and Zhang, I. (2003). “kNN Approach to Unbalanced Data Distributions: a Case Study Involving Information Extraction,” in Proceedings of Workshop on Learning from Imbalanced Datasets. Washington, DC: ICML, 1–7.
- Mehta, D., Anand, P., Kumar, V., Joshi, A., Mathur, D., Singh, S., et al. (2014). ParaPep: a Web Resource for Experimentally Validated Antiparasitic Peptide Sequences and Their Structures. *Database* 2014, bau051. doi:10.1093/database/bau051
- Momčilović, S., Cantacessi, C., Arsić-Arsenijević, V., Otranto, D., and Tasić-Otašević, S. (2019). Rapid Diagnosis of Parasitic Diseases: Current Scenario and Future Needs. *Clin. Microbiol. Infect.* 25 (3), 290–309. doi:10.1016/j.cmi.2018.04.028
- Pang, Y., Yao, L., Jhong, J. H., Wang, Z., and Lee, T. Y. (2021). AVPIDen: a New Scheme for Identification and Functional Prediction of Antiviral Peptides Based on Machine Learning Approaches. *Brief Bioinform* 22 (6), bbab263. doi:10.1093/bib/bbab263
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. *J. machine Learn. Res.* 12, 2825–2830. doi:10.48550/arXiv.1201.0490
- Schneider, T. D., and Stephens, R. M. (1990). Sequence Logos: a New Way to Display Consensus Sequences. *Nucl. Acids Res.* 18 (20), 6097–6100. doi:10.1093/nar/18.20.6097
- Stilianoudakis, S. C., Marshall, M. A., and Dozmorov, M. G. (2021). preciseTAD: a Transfer Learning Framework for 3D Domain Boundary Prediction at Base-Pair Resolution. *Bioinformatics* 38 (3), 621–630. doi:10.1093/bioinformatics/btab743
- Tahir, M. A., Kittler, J., and Yan, F. (2012). Inverse Random under Sampling for Class Imbalance Problem and its Application to Multi-Label Classification. *Pattern Recognition* 45 (10), 3738–3750. doi:10.1016/j.patcog.2012.03.014
- Torrent, M., Pulido, D., Rivas, L., and Andreu, D. (2012). Antimicrobial Peptide Action on Parasites. *Cdt* 13 (9), 1138–1147. doi:10.2174/138945012802002393
- Van der Maaten, L., and Hinton, G. (2008). Visualizing Data Using T-SNE. *J. machine Learn. Res.* 9 (86), 2579–2605.
- Wang, G., Li, X., and Wang, Z. (2016). APD3: the Antimicrobial Peptide Database as a Tool for Research and Education. *Nucleic Acids Res.* 44 (D1), D1087–D1093. doi:10.1093/nar/gkv1278
- Wang, J., Yang, B., An, Y., Marquez-Lago, T., Leier, A., Wilksch, J., et al. (2017). Systematic Analysis and Prediction of Type IV Secreted Effector Proteins by Machine Learning Approaches. *Brief. Bioinform.* 20 (3), 931–951. doi:10.1093/bib/bbx164

- Wang, P.-H., Tu, Y.-S., and Tseng, Y. J. (2019). PgpRules: a Decision Tree Based Prediction Server for P-Glycoprotein Substrates and Inhibitors. *Bioinformatics* 35 (20), 4193–4195. doi:10.1093/bioinformatics/btz213
- Wu, S., Wu, X., Tian, J., Zhou, X., and Huang, L. (2019). PredictFP2: a New Computational Model to Predict Fusion Peptide Domain in All Retroviruses. *Ieee/acm Trans. Comput. Biol. Bioinform* 17 (5), 1714–1720. doi:10.1109/TCBB.2019.2898943
- Yang, H., Wang, M., Liu, X., Zhao, X.-M., and Li, A. (2021). PhosIDN: an Integrated Deep Neural Network for Improving Protein Phosphorylation Site Prediction by Combining Sequence and Protein-Protein Interaction Information. *Bioinformatics* 37 (24), 4668–4676. doi:10.1093/bioinformatics/btab551
- Zahedifard, F., and Rafati, S. (2018). Prospects for Antimicrobial Peptide-Based Immunotherapy Approaches in Leishmania Control. *Expert Rev. anti-infective Ther.* 16 (6), 461–469. doi:10.1080/14787210.2018.1483720
- Zhang, W., Xia, E., Dai, R., Tang, W., Bin, Y., and Xia, J. (2021). PredAPP: Predicting Anti-parasitic Peptides with Undersampling and Ensemble Approaches. *Interdiscip. Sci. Comput. Life Sci.* 14 (1)–258268. doi:10.1007/s12539-021-00484-x

**Conflict of Interest:** Author JW is employed by MyGenostics Inc. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

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

Copyright © 2022 Jiang, Zhang, Xia, Jia, Yin, Wang, Wu and Ge. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Refined Contact Map Prediction of Peptides Based on GCN and ResNet

Jiawei Gu<sup>1</sup>, Tianhao Zhang<sup>1</sup>, Chunguo Wu<sup>1,2</sup>, Yanchun Liang<sup>1,2,3</sup> and Xiaohu Shi<sup>1,2,3\*</sup>

<sup>1</sup>College of Computer Science and Technology, University of Jilin, Changchun, China, <sup>2</sup>Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, Changchun, China, <sup>3</sup>School of Computer Science, Zhuhai College of Science and Technology, Zhuhai, China

Predicting peptide inter-residue contact maps plays an important role in computational biology, which determines the topology of the peptide structure. However, due to the limited number of known homologous structures, there is still much room for inter-residue contact map prediction. Current models are not sufficient for capturing the high accuracy relationship between the residues, especially for those with a long-range distance. In this article, we developed a novel deep neural network framework to refine the rough contact map produced by the existing methods. The rough contact map is used to construct the residue graph that is processed by the graph convolutional neural network (GCN). GCN can better capture the global information and is therefore used to grasp the long-range contact relationship. The residual convolutional neural network is also applied in the framework for learning local information. We conducted the experiments on four different test datasets, and the inter-residue long-range contact map prediction accuracy demonstrates the effectiveness of our proposed method.

**Keywords:** peptide inter-residue contact map prediction, deep learning, graph convolutional network, residual convolutional neural network, multiple sequence alignment

## OPEN ACCESS

### Edited by:

Juexin Wang,  
University of Missouri, United States

### Reviewed by:

Yi Xiong,  
Shanghai Jiao Tong University, China  
Yang Liu,  
Dana-Farber Cancer Institute,  
United States

### \*Correspondence:

Xiaohu Shi  
shixh@jlu.edu.cn

### Specialty section:

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

**Received:** 21 January 2022

**Accepted:** 23 March 2022

**Published:** 27 April 2022

### Citation:

Gu J, Zhang T, Wu C, Liang Y and Shi X  
(2022) Refined Contact Map Prediction  
of Peptides Based on GCN  
and ResNet.  
Front. Genet. 13:859626.  
doi: 10.3389/fgene.2022.859626

## 1 INTRODUCTION

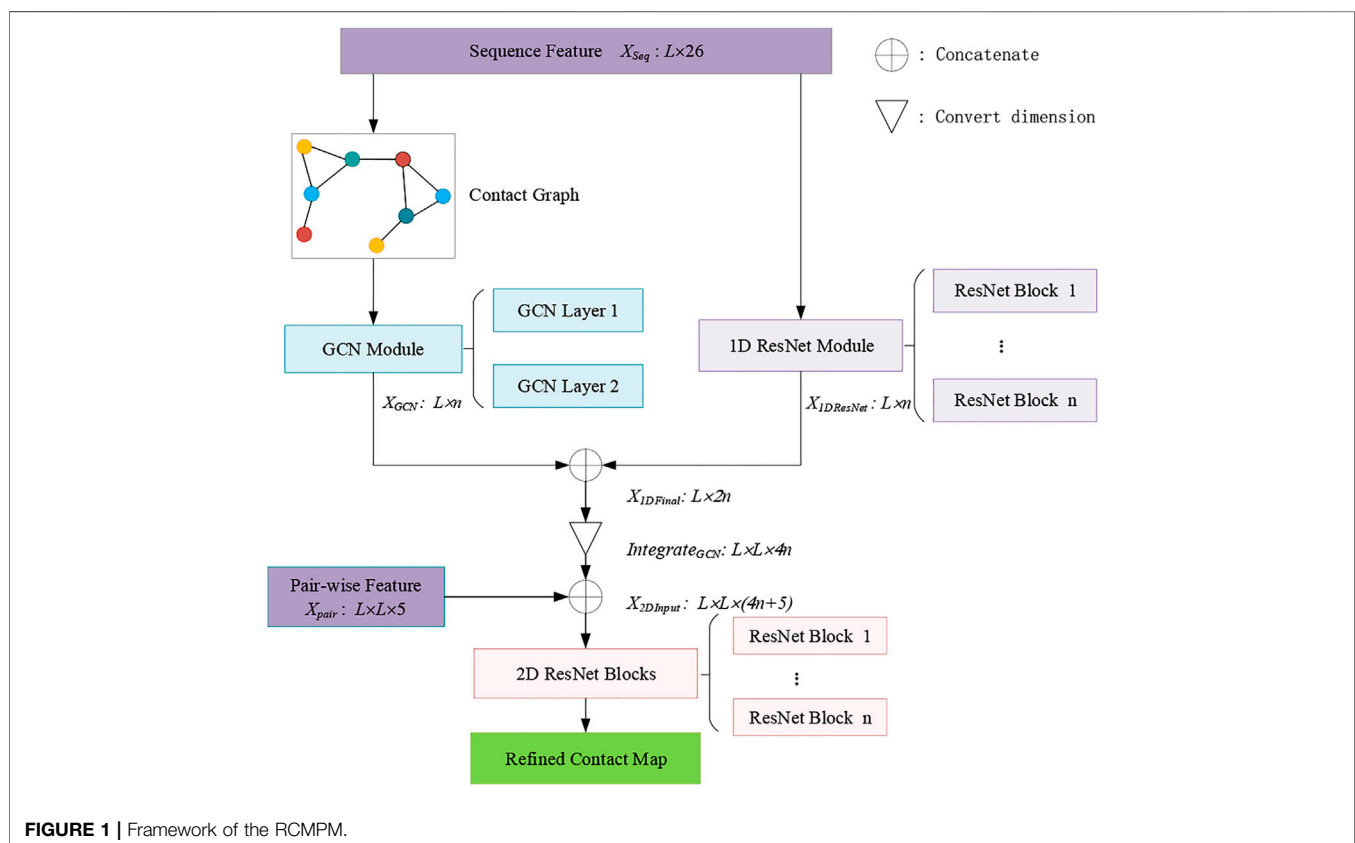
Peptides play an important role in computational and experimental biology (Torrissi et al., 2020), which motivates the development of accurate methods to predict their native conformations from the sequences. As a special kind of peptide, protein-related predictions from its amino acid sequence remain an open problem in the field of computational biology. Using biological experiments to determine the protein structure is very cumbersome and expensive. Therefore, it is very effective to use machine learning methods or deep learning methods to obtain a universal law from the amino acid sequence to the prediction of a protein's three-dimensional structure. The inter-residue contact map (Lena et al., 2012) is a two-dimensional representation of a protein's three-dimensional structure. The contact map constrains the conformation of protein structures; as a result, accurate prediction of the contact map can facilitate *ab initio* structure modeling, and the accuracy of the contact map affects the accuracy of the three-dimensional structure of the protein. Furthermore, contact maps have been widely used for model assessment and structure alignment.

The current contact map prediction methods are mainly based on direct coupling analysis (DCA) methods, machine learning methods, and deep learning methods. DCA-based methods mainly use multiple sequence alignment methods to determine the relationships between amino acid pairs. However, DCA-based methods assume that pairs of contacted residues are more likely to mutate simultaneously as the protein structure or function evolves and mainly use the multiple sequence

alignment (MSA) to determine the relationships between the amino acid pairs. Therefore, the accuracy of the DCA-based method depends on the number of homologous protein sequences in the protein sequence library. On the other hand, due to the existence of indirect evolutionary coupling information, the generated coupling information from the DCA might include “noise signal.” The common DCA-based methods include CCMpred (Seemayer et al., 2014), PSICOV (Jones et al., 2012), and GREMLIN (Kamisetty et al., 2013). CCMpred mainly uses Markov random field pseudo-likelihood maximization to learn the contacts between the protein inter-residues. When there are a large number of homologous proteins in the protein sequence, the accuracy of the contact prediction results is higher; however, when the sequences of the homologous protein are fewer, the accuracy is lower. On the other hand, machine learning-based and deep learning-based methods use a set of input features derived from multiple sequence alignments (MSAs) to predict the protein inter-residue contact map, including position-specific scoring matrices (PSSMs), secondary structure (SS) predictions, and solvent accessibility (SA) information. Machine learning-based methods are mainly based on support vector machines (SVMs) (Hearst et al., 1998) to learn the abovementioned features and common support vector machine (SVM) methods including SVMCon (Cheng and Baldi, 2007) and R2C (Yang et al., 2016). SVMCon used support vector machines (SVMs) and yields good

performance on medium- to long-range contact predictions. In recent years, deep learning methods have been mainly used to predict the contact map between the protein inter-residues and are mainly based on the structure of the convolutional neural network (CNN) and residual neural network (ResNet) (He et al., 2016). The ResNet structure further improves the CNN structure and solves the problem of reduced accuracy when there are too many convolutional layers through the skip connection mechanism. RaptorX-Contact (Wang et al., 2017) was the first model that used the ResNet structure for protein inter-residue contact map prediction tasks. Zhong Li et al. (Li et al., 2020) used ResNet and DenseNet (Huang et al., 2017) structures and a new protein sequence feature (PSFM) to improve the contact map prediction accuracy. DeepCov (Jones and Kandathil, 2018) applied the CNN to predict contact maps when limited evolutionary information is available, which has been trained on a very limited set of input features: pair frequencies and covariance. It is noticed that there are several similar studies predicting the distance matrix instead of the contact map, such as RaptorX structure prediction (Xu, 2018), PG-GNN (Xia and Ku, 2020), and AlphaFold (Senior et al., 2020).

However, there are two main difficulties in obtaining accurate contact predictions. First, many amino acid sequences lack a large number of homologous sequences, which limits the level of accuracy of predictions. On the contrary, the target sequences with many





homologous sequences might generate “noise signals” from the evolutionary coupling information. Second, most methods use convolutional neural network (CNN)-based models for inter-residue contact map prediction, leading to over-learning of the local information, but under-learning of the long-range information, which is reflected by a low long-range accuracy.

Therefore, eliminating “noise signals” is necessary to improve the residue contact prediction. Improving the inter-residue contact prediction has been of interest for many years due to its critical importance in structure bioinformatics, with either the sequence or structure template information. R2C (Yang et al., 2016) used SVM and PSICOV methods and used a dynamic fusion strategy to predict the contact map between amino acids and applied Gaussian noise filters for further denoising. Amelia Villegas-Morcillo et al. (Villegas-Morcillo et al., 2018) applied K-SVD (Aharon et al., 2006) and deep convolutional neural network (DCNN) methods specially designed for image denoising to solve the problem of Gaussian noise. DNCON2 (Adhikari et al., 2017) adopted the structure of the two-stage convolutional neural networks (CNNs) to improve the contact map prediction, which divides the prediction into two parts. The first part trains five CNNs to predict the contact map between the distances of 6, 7.5, 8, 8.5, and 10, respectively. The second part takes the input feature as the output of the first part and then utilizes a CNN structure for further prediction.

In the past few years, the graph neural network (Zhou et al., 2018) was raised to represent the protein structure in various deep learning-based methods and had succeeded in the computational biology area, such as protein interface prediction, protein solubility prediction, and protein function prediction. Fout et al., (2017) proposed a type of architecture for the task of predicting protein interfaces between the pairs of proteins using a graph representation of the underlying protein structure. GraphSol (Chen et al., 2020) was used to predict the protein residue solubility by combining the predicted contact maps, graph neural networks, and attention mechanisms. DeepFRI (Gligorijević et al., 2021) used an LSTM (Hochreiter and Schmidhuber, 1997) and a graph convolutional network to predict protein functions. PG-GNN (Xia and Ku, 2020) used a new convolution kernel to perform deep convolution to obtain the distance map, which was used to construct an inter-residue graph between the residues for obtaining the dihedral information between residues, and finally constructed a three-dimensional protein structure.

Here, to focus on getting more accurate contact maps, especially on the long-range level, we developed a novel refined contact map prediction model (RCMPM) to refine the rough contact map produced by the existing methods, which combines a graph convolution network (GCN) (Kipf and Welling, 2016) and residual convolution neural networks (ResNet) (He et al., 2016). The main contributions of the article are summarized as follows:

- The peptide contact map refinement task is modeled as a geometric 2D graph improvement, with nodes representing the amino acid residues and edges representing contacts

between the residues. The rough results of other models such as CCMpred and RaptorX-Contact are used to construct the inter-residue contact graph.

- Aiming at the challenges previously mentioned, a novel deep neural network framework is proposed for the inter-residue contact prediction by combining a graph convolution network (GCN) and residual convolution neural networks (1D ResNet and 2D ResNet), of which the GCN has a strong global information extraction ability, and hence can better capture the long-range contact relationships among the complex sequence inter-residues.
- The experiments are conducted on four different test datasets, and the inter-residue long-range contact map prediction accuracy demonstrates the effectiveness of our proposed method due to the new network architecture.

The rest of the article is organized as follows. **Section 2** details the materials and methods, including contact definition, graph construction, feature selection, and the proposed prediction model. **Section 3** reports the datasets used in our method, evaluation metrics, and experiments on four test datasets. **Section 4** concludes the article and discusses the directions for the future work.

## 2 MATERIALS AND METHODS

### 2.1 Contact Definition

In general, two residues are considered to be in contact if certain atoms are close enough to form a molecular interaction. In the Critical Assessment of protein Structure Prediction (CASP) experiment (Moult et al., 2014, 2016, 2018), the contact definition is based on the spatial distance of  $C_\beta$  atoms. For instance, assuming that  $v = \{v_1, v_2, \dots, v_i, \dots, v_j, \dots, v_L\}$  is the residue sequence, where  $L$  is the sequence length, and  $(x_{v_i}, y_{v_i}, z_{v_i})$  is the three-dimensional coordinates of amino acid residue  $v_i$ , then the equation for the distance between the residues  $v_i$  and  $v_j$  is

$$\begin{aligned} \text{Distance}(i, j) &= \text{Distance}(C_{\beta_i}, C_{\beta_j}) \\ &= \sqrt{(x_{v_i} - x_{v_j})^2 + (y_{v_i} - y_{v_j})^2 + (z_{v_i} - z_{v_j})^2}. \end{aligned} \quad (1)$$

If the Euclidean distance between the  $C_\beta$  atoms ( $C_\alpha$  for GLY) of two amino acids is less than a given threshold  $\gamma$ , then the two residues are said to be in contact.

### 2.2 Graph Construction

As mentioned in **Section 2.1**, we can use the other contact map prediction models, such as CCMpred (Seemayer et al., 2014) and RaptorX-Contact (Wang et al., 2017), to obtain a contact matrix  $CM$ . Assuming that the length of the peptide is  $L$ , then  $CM$  is an  $L \times L$  matrix, whose element  $CM_{ij}$  denotes whether the pair of residues  $i$  and  $j$  is contacted or not (1 or 0). Denote  $G = \{N, E\}$  is the contact graph of the peptide, where  $N$  is the node set including  $L$  amino acids, and  $E$  is the edge set. Then, the contact graph could be constructed as follows:

**Algorithm 1.** Graph construction.

---

```

 $E = \Phi$ 
for  $i = 1$  to  $L-1$  do
  for  $j = i + 1$  to  $L$  do
    if  $Distance_{ij} < \gamma$  then
       $E = E \cup l_{ij}$ 
    end if
  end for
end for

```

---

where  $l_{ij}$  is the edge between node  $i$  and node  $j$ , and the threshold  $\gamma$  is set as  $8\text{\AA}$  in this article.

**2.3 Feature Selection****2.3.1 Sequence Features**

We devised three groups of sequence features to train our model, namely, the position-specific scoring matrix (PSSM), secondary structure (SS), and solvent accessibility (SA). The PSSM is a widely used sequence feature, which is produced by executing PSI-BLAST (Altschul et al., 1997) on the UniRef90 database (Suzek et al., 2015) with 0.001 e-value after the three iterations, which is a 20-dimensional profile feature for each residue. The secondary structure and solvent accessibility describe the arrangement of the protein backbone, which are also very important for the contact prediction. The secondary structure and solvent accessibility are predicted by the RaptorX-Property (Wang et al., 2016) program (<http://raptorx.uchicago.edu/StructurePropertyPred/predict/>). The secondary structure is divided into three categories, namely, helix (H), strand (E), and coil (C), and the solvent accessibility is also classified into three types, namely, buried, medium, and exposed. The PSSM is represented as a two-dimensional matrix of  $L \times 20$ , while both the secondary structure and solvent accessibility are represented as a two-dimensional matrix of  $L \times 3$ ; therefore, the concatenation sequence embedding vector  $X_{seq}$  is obtained with the  $L \times 26$  dimension, where the order of the splicing input is [PSSM, secondary structure, and solvent accessibility].

**2.3.2 Pairwise Features**

Pairwise features are the information that characterizes the relationship between the pairs of residues, including the co-evolutionary information, statistical information, and so on. Four groups of pairwise features are used to train our model, namely, RaptorX-Contact prediction, CCMpred prediction, mutual information (Dunn et al., 2008), and contact potential (Betancourt and Thirumalai, 1999), which provide the co-evolutionary information for each pair of alignment columns. RaptorX-Contact and CCMpred prediction are mainly used as inter-residue scores. RaptorX-Contact prediction results can be obtained by model training, the source code of which can be downloaded from <https://github.com/j3xugit/RaptorX-Contact>. CCMpred prediction results can be obtained by the CCMpred program, which could be accessed at <https://github.com/soedinglab/CCMpred>. However, CCMpred requires the homologous sequence result of the multiple sequence alignment (MSA) as the input, which is produced by executing the HHblits program (Remmert et al., 2012) on the Uniclust30

database (Mirdita et al., 2017) with 0.001 e-value after three iterations. Both RaptorX-Contact and CCMpred output an inter-residue score for each residue pair. After the MSA profile is obtained, the mutual information could be defined by

$$MI_{ij} = \sum_{x,y \in R} p_{ij}(x,y) \ln \frac{p_{ij}(x,y)}{p_i(x)p_j(y)}, \quad (2)$$

where  $R$  is the set of amino acid types,  $x$  and  $y$  are the elements in column  $i$  and column  $j$ , respectively,  $p_i(x)$  and  $p_j(y)$  indicate the probabilities of residue  $x$  in column  $i$  and residue  $y$  in column  $j$ , and  $p_{ij}(x,y)$  is the probability that residue  $x$  is in column  $i$  and residue  $y$  is in column  $j$ , respectively. Normalized mutual information, namely, average product correction (APC) mutual information is also used in our method, which is defined by

$$MI_{ij}^{APC} = MI_{ij} - APC_{ij}, \quad (3)$$

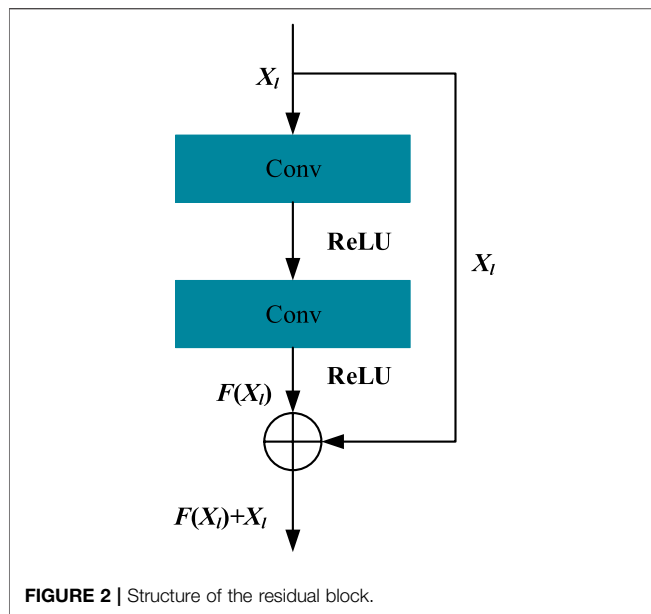
$$APC_{ij} = \frac{\sum_{j \neq i} MI_{ij} \sum_{i \neq j} MI_{ij}}{\sum_{i,j (i \neq j)} MI_{ij}}. \quad (4)$$

The contact potential is computed by averaging the contact potential terms across the two alignment columns. Mutual information and contact potential are generated by alnstats in the MetaPSICOV (Jones et al., 2015) program, which also requires the homologous sequence as the input. For RaptorX-Contact prediction, CCMpred prediction, mutual information, APC mutual information, and contact potential, all are represented as a three-dimensional matrix of  $L \times L \times 1$ ; therefore, the concatenation pair-wise embedding features  $X_{pair}$  are obtained with the  $L \times L \times 5$  dimension, where the order of the splicing input is [RaptorX-Contact prediction, CCMpred prediction, MI, APC MI, and contact potential].

**2.4 Prediction Model****2.4.1 The Framework of the RCMPM Model**

Residual networks (ResNets) are very helpful for accurate peptide contact map prediction, which has been demonstrated in the RaptorX-Contact model (Wang et al., 2017). Therefore, ResNet architecture is retained in our proposed refined contact map prediction model (RCMPM). On the other hand, the rough contact map obtained by the other methods could be well utilized by transferring it into an amino acid graph, and therefore, the graph convolution network (GCN) could handle the graph topology very well. Hence, the proposed RCMPM model includes a GCN module, a 1D ResNet module, and a 2D ResNet module, respectively.

**Figure 1** shows the framework of the RCMPM model, which has two types of features, namely, sequence features and pair-wise features. The GCN module is used to learn the global structural features of the inter-residue contact graphs, whose input is the node representation of the sequence features, and the output is a dense global structural embedding vector for each amino acid node. 1D ResNet module is used to handle the one-dimensional sequence feature and output a sequence embedding vector for each amino acid. 2D ResNet module integrates the above two modules' outputs and the pair-wise features as well and finally generates the refined contact map.



The following part of this section will describe these three modules in detail.

## 2.4.2 GCN Module

Given a sequence with  $L$  residues, the residue graph can be represented by a contact map, that is, the nodes of the graph are the residues of the peptide, and the features of the nodes are represented by the attributes of the residues. The edges of the contact graph indicate whether there are connections between the amino acid nodes, and the weight of the edge represents the probability of contact. We used the graph convolution network (GCN) to obtain the global structural features of the graph.

The graph convolutional layer in the prediction model uses the following equation:

$$H^{(l+1)} = \sigma(\tilde{A}H^{(l)}W^{(l)}), \quad (5)$$

where  $\tilde{A} = A + I_L$  is the variant of the adjacency matrix by adding the self-loop identity matrix  $I_L$  on the original adjacency matrix  $A$ , and  $H^{(l)}$  is the hidden matrix learned by the  $l$ th layer, initial of which is the hidden matrix  $H^{(0)} = X_{seq}$ .  $W^{(l)}$  is a weight matrix of the layer-specific trainable parameters and is used to map the iterations to a low-dimensional rich information space, and  $\sigma$  is a nonlinear activation function, which is taken as the ReLU function in our model. We also use normalization to map the input feature of each layer  $H^{(l)}$  to  $[0,1]$  to improve the data performance and reduce errors. Finally, we used a 2-layer graph convolutional network to learn the global structural features of the contact graph containing amino acid node features. Hence, the final output of the GCN module in the RCMPM model uses the following equation:

$$X_{GCN} = RELU(\tilde{A}ReLU(\tilde{A}X_{seq}W^{(0)})W^{(1)}). \quad (6)$$

## 2.4.3 1D ResNet Module

A 1D ResNet module is used to handle the one-dimensional sequence feature and outputs a sequence embedding vector for each residue, which is stitched together by the residual blocks. A residual block consists of two convolutional layers and two activation layers, which can be defined as follows:

$$X_{l+1} = F(X_l, W_l) + X_l, \quad (7)$$

where  $X_l$  and  $X_{l+1}$  are the input and output vectors of the residual block, respectively, and the initial hidden matrix  $X_0 = X_{seq}$ . Here,  $W_l$  is the weight matrix in convolutional layers of the  $l$ th block, and  $F(X_l, W_l)$  represents the result after the action of the convolutional layer and activation function layer. Here, the operation of the convolutional layer is implemented by the `conv1d` function of the tensorflow framework. Here, we used the `ReLU` function as the activation function of our method and also used normalization to map the data to  $[0,1]$  to improve data performance and reduce errors. We kept the dimension of  $X_{l+1}$  larger than  $X_l$  because the higher dimension can carry more information. For a residual block, the  $F(X_l, W_l)$  function can be expressed as shown in **Figure 2**.

Finally, the output of the 1D ResNet module in the RCMPM model could be described as follows:

$$X_{1DResNet} = \sum_{l=0}^n F(X_l, W_l) + X_l. \quad (8)$$

In our 1D ResNet module, the number of residual blocks is selected as 3.

## 2.4.4 2D ResNet Module

The 2D ResNet module is used to learn the final contact relationship for each residue pair by integrating the aforementioned two modules, namely, that it takes the input of the output feature  $X_{GCN}$  of the GCN module and the output feature  $X_{1DResNet}$  of the 1D ResNet module and the pairwise feature  $X_{pair}$  as well. Different with the 1D ResNet module, the 2D ResNet module is dealing with two-dimensional feature maps. The pairwise features  $X_{pair}$  is of  $L \times L \times 5$  dimension, as described in **section 2.3.2**, while the output features  $X_{GCN}$  and  $X_{1DResNet}$  are the one-dimensional feature map with the same dimension  $L \times n$ , which should be converted to a two-dimensional feature map. Similarly with the method used in (Wang et al., 2017),  $X_{GCN}$  and  $X_{1DResNet}$  are first concatenated on the second dimension, obtaining an  $L \times 2n$  feature map  $X_{1DFinal}$ :

$$X_{1DFinal} = X_{1DResNet} \oplus X_{GCN}. \quad (9)$$

Then, it is converted to a 2-dimensional feature map. Redefined  $X_{1DFinal}$  from  $L \times 2n$  to  $L \times 1 \times 2n$  dimension by adding a second-order dimension with 1, then duplicate  $X_{1DFinal}$   $L$  times to extend the second order from 1 to  $L$ , getting an  $L \times L \times 2n$  tensor  $TX_{GCN_{1D}}$ .  $TX'_{GCN}$  is denoted as the transpose of  $TX_{GCN_{1D}}$  on the first two orders;  $X_{GCN}$  and  $X_{1DResNet}$  are finally integrated as  $Integrate_{GCN}$ :

$$Integrate_{GCN} = TX_{GCN_{1D}} \oplus TX'_{GCN_{1D}}, \quad (10)$$

**TABLE 1** | Contact map results by four different methods on the PDB25 testing dataset.

Method	Long-range				Medium-range				Short-range			
	L/10	L/5	L/2	L	L/10	L/5	L/2	L	L/10	L/5	L/2	L
CCMpred	0.528	0.475	0.361	0.257	0.456	0.356	0.222	0.148	0.356	0.275	0.175	0.121
R2C	0.666	0.667	0.648	0.449	0.591	0.590	0.322	0.176	0.597	0.408	0.201	0.119
RaptorX-Contact	0.774	0.739	0.633	0.497	0.758	0.675	0.469	0.300	0.756	0.641	0.404	0.241
RCMPM (CCMpred)	0.718	0.685	0.582	0.446	0.707	0.622	0.421	0.262	0.685	0.576	0.355	0.208
RCMPM (RaptorX-Contact)	0.784	0.748	0.646	0.508	0.761	0.679	0.473	0.300	0.754	0.645	0.403	0.237

where  $\oplus$  represents concatenation on the third-order dimension; therefore  $Integrate_{GCN_{1D}}$  is of  $L \times L \times 4n$  dimension. Afterward, it should be combined with the pairwise features  $X_{pair}$  by

$$X_{2DInput} = Integrate_{GCN} \oplus X_{pair}, \quad (11)$$

where  $X_{2DInput}$  is of  $L \times L \times (4n + 5)$  dimension finally.

We also used the same residual network block structure with that of the 1D ResNet (Figure 2) module to stack the 2D ResNet module. The difference is that the 2D ResNet module is dealing with 2D feature maps and utilizing *conv2d* function of the tensorflow framework for the convolution operation. The final output  $X_{2DResNet}$  of the 2D ResNet module could be expressed by

$$X_{2DResNet} = \sum_{l=0}^n F(X_l, W_l) + X_l, \quad (12)$$

where  $X_l$  is the input feature of the  $l$ th residual block, being initialized by  $X_0 = X_{2DInput}$ .  $W_l$  is the weight matrix in the convolutional layers of the  $l$ th block,  $F()$  is the mapping function with the same meaning of that in the 1D ResNet block, and  $n$  is the block number, which is set as 30 in the 2D ResNet module. Hence, the output of the 2D ResNet  $X_{2DResNet}$  will go through the softmax layer and obtain the inter-residue contact label:

$$y = Softmax(X_{2DResNet}), \quad (13)$$

where  $y \in \{0, 1\}^{L \times L}$ , the element  $y_{ij}$  means whether the pair of residue  $i$  and residue  $j$  is contacted according to the model (1 for contacted and 0 for uncontacted).

To train the model, the cross-entropy function averaged over all the residue pairs is used as the loss function:

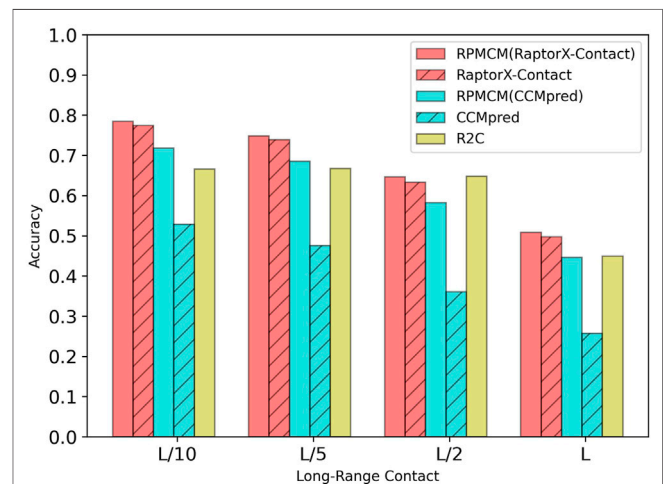
$$E(t, y) = -\frac{1}{L^2} \sum_i \sum_j t_{ij} \log y_{ij}, \quad (14)$$

where  $t_{ij}$  is the true contact label, and  $y_{ij}$  is the predicted contact label between residues  $i$  and  $j$ , and  $L$  is the length of the peptide. For the training process, stochastic gradient descent optimization is utilized, and the learning rate is set as 0.01.

## 3 RESULTS

### 3.1 Training and Test Datasets

In our experiment, we used one training dataset to train our proposed RCMPM model and four different testing datasets to test its performance.

**FIGURE 3** | Comparison of method accuracy for the long-range contact on the PDB25 testing dataset.

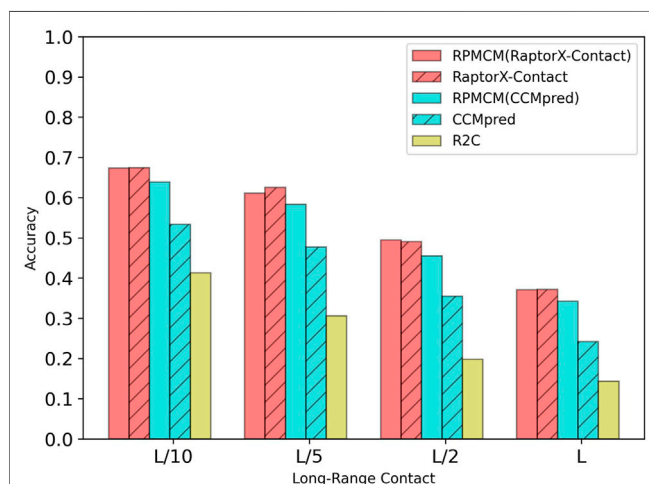
The training dataset is a subset of PDB25 extracted from the PDB database (<http://www.rcsb.org/pdb/home/home.do>) with homology reduction at 25% level of sequence identity, resulting in 6767 non-homologous protein sequences. The number of amino acids of each training protein ranges from 26 to 300. To avoid overfitting, 400 proteins are randomly chosen for validation and the remaining others for training.

To evaluate the performance of our model, it is applied to four testing datasets. The first testing dataset is the PDB25 dataset, which contains 500 nonhomologous protein sequences. The training set, validation dataset, and testing dataset of the abovementioned PDB25 dataset can be downloaded from <http://raptorx.uchicago.edu/ContactMap/>. The other three datasets were obtained from three CASP (Critical Assessment of Structure Prediction) competitions (CASP10 (Moult et al., 2014), CASP11 (Moult et al., 2016), and CASP12 (Moult et al., 2018)). For the three CASP datasets, we used the same screening method as that used in the R2C method (Yang et al., 2016). For the CASP10 dataset, the sequence data could be accessed on the website of [https://predictioncenter.org/download\\_area/CASP10/targets/](https://predictioncenter.org/download_area/CASP10/targets/). The total number of the sequences is 123. However, seven short sequences are removed (T0651-D3, T0675-D1, T0675-D2, T0677-D1, T0700-D1, T0709-D1, and T0711-D1), and the constructed CASP10 test dataset size is 116. CASP11 and CASP12 datasets are also publicly available on the websites of [https://predictioncenter.org/download\\_](https://predictioncenter.org/download_)



**TABLE 2 |** Contact map results by four different methods on the CASP10 testing dataset.

Method	Long-range				Medium-range				Short-range			
	L/10	L/5	L/2	L	L/10	L/5	L/2	L	L/10	L/5	L/2	L
CCMpred	0.533	0.477	0.355	0.242	0.512	0.417	0.272	0.185	0.418	0.313	0.197	0.137
R2C	0.413	0.306	0.198	0.143	0.540	0.425	0.278	0.191	0.571	0.511	0.373	0.264
RaptorX-Contact	0.674	0.625	0.490	0.372	0.699	0.629	0.458	0.318	0.638	0.540	0.368	0.233
RCMPM (CCMpred)	0.639	0.583	0.455	0.342	0.646	0.593	0.426	0.290	0.571	0.486	0.316	0.198
RCMPM (RaptorX-Contact)	0.673	0.611	0.495	0.371	0.681	0.612	0.452	0.312	0.630	0.530	0.360	0.225

**FIGURE 4 |** Comparison of method accuracy for the long-range contact on the CASP10 testing dataset.

area/CASP11/targets/ and [https://predictioncenter.org/download\\_area/CASP12/targets/](https://predictioncenter.org/download_area/CASP12/targets/), with 105 and 55 sizes, respectively. After removing the three short sequences from CASP11 (T0759-D1, T0820-D1, and T0820-D2), the final sizes of CASP11 and CASP12 datasets are 102 and 55, respectively.

### 3.2 Evaluation Metrics

By using the same evaluation criteria as the CASP competition, we evaluated the accuracies of the top  $L/k$  ( $k = 10, 5, 2, 1$ ) predicted contacts, where  $L$  is the protein sequence length. Accuracy is the proportion of true positive samples in the total number of predicted positive samples, which is defined by

$$Accuracy = \frac{TP}{TP + FP} \quad (15)$$

where  $TP$  is the number of predicted contacted pairs being actually contacted, and  $FP$  is the number of predicted contacted pairs not being actually contacted, respectively. Residue-residue contacts are categorized into three types according to the residue distances in sequence: short-range, medium-range, and long-range corresponding to the distances between 6 and 11, 12 and 23, and at least 24 residues, respectively. It should be noted that a long-range contact places strong constraints on the conformation of peptides and is particularly important for the peptide structure and function study, which is also the main focus of this article.

### 3.3 Performance on PDB25 Testing Datasets and CASP Testing Datasets

In our experiment, we used top  $L/k$  ( $k = 10, 5, 2, 1$ ) in the long-range contact to evaluate the prediction accuracy of contact maps. Here,  $L$  is the length of the sequence, and the prediction accuracy rates are given in three kinds of contact, namely, long-range, medium-range, and short-range.

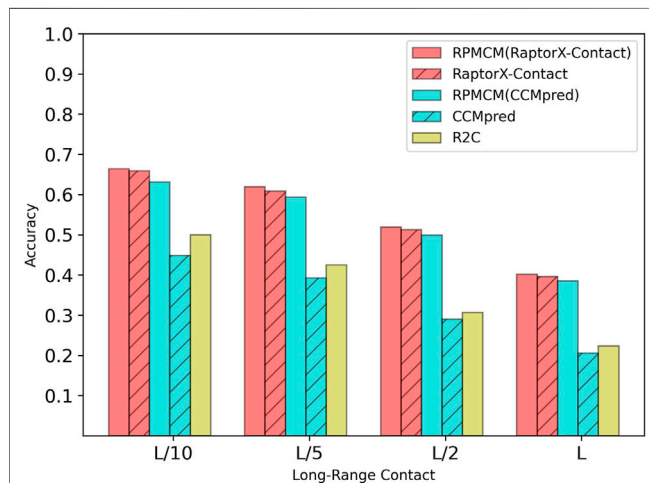
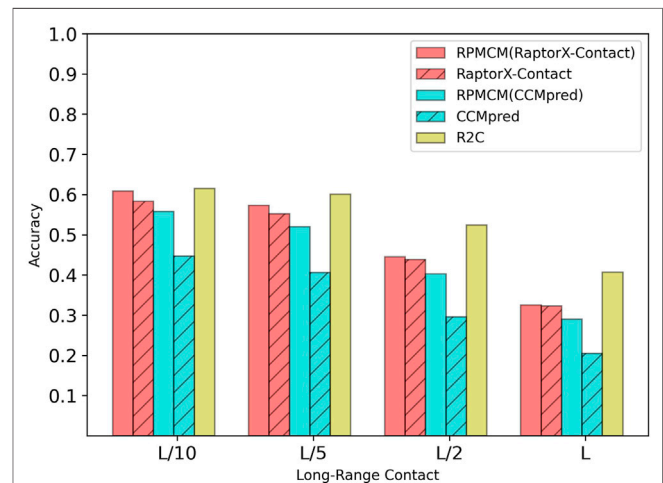
The datasets used in our experiment are PDB25, CASP10, CASP11, and CASP12 datasets. To examine the performance of our proposed RCMPM model, three state-of-the-art methods are used for comparison, namely, CCMpred (based on Markov random field pseudo-likelihood maximization, MSA), R2C (based on SVM), and RaptorX-Contact (based on ResNet), respectively. We have realized CCMpred and RaptorX-Contact models and trained them under the same environments with that of the RCMPM and hence obtained the experiment results of the two models by ourselves. The results of R2C on CASP10 and CASP11 datasets are cited from the reference (Villegas-Morcillo et al., 2018), while the results on the other two datasets are calculated through its webserver (<http://www.csbio.sjtu.edu.cn/bioinf/R2C/>). For comparison, two rough contact maps, produced by CCMpred and RaptorX-Contact models, are used to construct the amino acid graph in the proposed RCMPM, respectively.

**Table 1** shows the comparison results on the PDB25 dataset. For the long-range contact type prediction, the results of the RCMPM by using the CCMpred outputs as the rough contact map (RCMPM (CCMpred)) are significantly better than those of CCMpred, with 19.9%, 22.2%, 38.0%, and 73.5% improvements on top  $L/10$ ,  $L/5$ ,  $L/2$ , and  $L$  levels, respectively. Compared to R2C, it improves by 7.8% and 2.7% on the top  $L/10$  and  $L/5$  levels, respectively, and decreased by 10.2% and 0.7% on the top  $L/2$  and  $L$  levels, respectively. RaptorX-Contact is an excellent algorithm, the results of which are better than those of RCMPM (CCMpred). However, when the RCMPM model uses the output of RaptorX-Contact as the rough contact map (RCMPM (RaptorX-Contact)), it outperforms RaptorX-Contact on all the four top levels with 1.3%, 1.2%, 2.1%, and 2.2% improvements, respectively. The results of RCMPM (RaptorX-Contact) are also significantly better than CCMpred, R2C, and RCMPM (CCMpred), with the only exception being slightly below R2C at the top  $L/2$  level. **Figure 3** shows the comparison results of five methods on the long-range contact type prediction. For the medium-range contact type, both RCMPM (CCMpred) and RCMPM (RaptorX-Contact) are significantly superior to CCMpred and RaptorX-Contact, both outperforming their opponents at the four top levels. Among all the five comparison methods, RCMPM (RaptorX-Contact) performs



**TABLE 3** | Contact map results by four different methods on the CASP11 testing dataset.

Method	Long-range				Medium-range				Short-range			
	L/10	L/5	L/2	L	L/10	L/5	L/2	L	L/10	L/5	L/2	L
CCMpred	0.448	0.393	0.290	0.206	0.376	0.298	0.187	0.132	0.318	0.251	0.162	0.118
R2C	0.500	0.425	0.307	0.223	0.397	0.296	0.192	0.138	0.314	0.228	0.146	0.115
RaptorX	0.659	0.608	0.512	0.396	0.677	0.608	0.447	0.296	0.683	0.598	0.405	0.249
RCMPM (CCMpred)	0.631	0.593	0.499	0.385	0.644	0.593	0.431	0.277	0.646	0.577	0.380	0.224
RCMPM (RaptorX-Contact)	0.664	0.619	0.519	0.402	0.670	0.608	0.450	0.299	0.682	0.601	0.406	0.245

**FIGURE 5** | Comparison of method accuracy for the long-range contact on the CASP11 testing dataset.**FIGURE 6** | Comparison of method accuracy for the long-range contact on the CASP12 testing dataset.

best at all the four levels. For the short-range contact type, RCMPM (CCMpred) is greatly better than CCMpred, while RCMPM (RaptorX-Contact) performs similarly with RaptorX-Contact, both significantly outperforming the other three methods.

**Table 2** shows the comparison results on the CASP10 dataset. For the long-range contact type prediction, the results of RCMPM by using CCMpred outputs as the rough contact map (RCMPM (CCMpred)) are significantly better than those of CCMpred, with 19.8%, 22.2%, 28.2%, and 41.3% improvements on top L/10, L/5, L/2 and L levels, respectively. Compared to R2C, it improved by 54.7%, 90.5%, 129.8%, and 139.2% at the four top levels, respectively. When the RCMPM uses the RaptorX-Contact outputs as the rough contact map (RCMPM (RaptorX-Contact)), it performs similarly with the RaptorX-Contact, with -0.1%, 1.0%, -2.2% and -0.2% variations at the top levels, respectively. Both of them significantly outperform CCMpred and R2C, and a little better than RCMPM (CCMpred). RCMPM (RaptorX-Contact) increases by 26.3%, 28.1%, 39.4%, and 53.3% compared to CCMpred and increases by 63.0%, 99.7%, 150.0%, and 159.4% compared to R2C at the four top levels, while compared to RCMPM (CCMpred), the improvements are 5.3%, 4.8%, 8.8%, and 8.5%, respectively. **Figure 4** shows the comparison results of the five methods on the long-range contact type prediction. The results are similar for both the medium-range contact and short-range contact types, with RCMPM (CCMpred)

being significantly superior to CCMpred, while RCMPM (RaptorX-Contact), despite its lower performance than RaptorX-Contact, had a weak gap.

**Table 3** shows the comparison results on the CASP11 dataset. For the long-range contact type prediction, the results of the RCMPM by using the CCMpred outputs as the rough contact map (RCMPM (CCMpred)) are significantly better than those of CCMpred, with 40.8%, 50.9%, 72.1%, and 86.9% improvements at the four top levels. Compared to R2C, it outperforms at all the four top levels with 26.2%, 39.5%, 62.5%, and 72.6%. RaptorX-Contact is better than RCMPM (CCMpred). However, when the RCMPM uses the output of RaptorX-Contact as the rough contact map (RCMPM (RaptorX-Contact)), it improves by 0.8%, 1.8%, 1.4%, and 1.5% on the four top levels, respectively. The results of RCMPM (RaptorX-Contact) are also significantly better than CCMpred and R2C. The results are similar for both the medium-range contact and short-range contact types, with RCMPM (CCMpred) being significantly superior to CCMpred, while RCMPM (RaptorX-Contact), despite its lower performance than RaptorX-Contact, had a weak gap. **Figure 5** shows the comparison results of the five methods on the long-range contact type prediction. For both the medium-range contact and short-range contact types' results, we can draw the following conclusions: among the three existing state-of-the-art (SOTA) methods, RaptorX-Contact performs the best; RCMPM

**TABLE 4 |** Contact map results by four different methods on the CASP12 testing dataset.

Method	Long-range				Medium-range				Short-range			
	L/10	L/5	L/2	L	L/10	L/5	L/2	L	L/10	L/5	L/2	L
CCMp <sub>pred</sub>	0.447	0.406	0.296	0.205	0.421	0.339	0.205	0.136	0.355	0.256	0.165	0.119
R2C	0.615	0.601	0.524	0.407	0.622	0.545	0.399	0.259	0.584	0.502	0.323	0.205
RaptorX	0.583	0.552	0.438	0.323	0.616	0.545	0.371	0.247	0.581	0.488	0.331	0.222
RCMPM (CCMp <sub>pred</sub> )	0.558	0.520	0.403	0.290	0.586	0.492	0.329	0.213	0.525	0.438	0.278	0.177
RCMPM (RaptorX-Contact)	0.608	0.573	0.445	0.325	0.606	0.530	0.372	0.245	0.591	0.484	0.333	0.215

**TABLE 5 |** Contact map results by the comparison between our network structures.

Method	Long-range				Medium-range				Short-range			
	L/10	L/5	L/2	L	L/10	L/5	L/2	L	L/10	L/5	L/2	L
RCMPM (without GCN)	0.775	0.741	0.635	0.498	0.761	0.676	0.471	0.299	0.755	0.642	0.402	0.238
RCMPM	0.784	0.748	0.646	0.508	0.761	0.679	0.473	0.300	0.754	0.645	0.403	0.237

**TABLE 6 |** Comparison results for feature combinations by using the rough RaptorX-Contact contact map.

Method	Long-range				Medium-range				Short-range			
	L/10	L/5	L/2	L	L/10	L/5	L/2	L	L/10	L/5	L/2	L
RCMPM (PSSM)	0.772	0.741	0.639	0.502	0.760	0.674	0.467	0.297	0.761	0.642	0.403	0.238
RCMPM (PSSM+SS)	0.777	0.742	0.639	0.505	0.760	0.673	0.469	0.298	0.756	0.643	0.401	0.237
RCMPM (PSSM+SS+SA)	0.784	0.748	0.646	0.508	0.761	0.679	0.473	0.300	0.754	0.645	0.403	0.237

**TABLE 7 |** Comparison results for feature combinations by using the rough CCM<sub>pred</sub> contact map.

Method	Long-range				Medium-range				Short-range			
	L/10	L/5	L/2	L	L/10	L/5	L/2	L	L/10	L/5	L/2	L
RCMPM (PSSM)	0.657	0.604	0.461	0.308	0.614	0.499	0.299	0.180	0.581	0.438	0.238	0.138
RCMPM (PSSM+SS)	0.712	0.670	0.570	0.437	0.692	0.609	0.411	0.254	0.680	0.569	0.346	0.201
RCMPM (PSSM+SS+SA)	0.718	0.685	0.582	0.446	0.707	0.622	0.421	0.262	0.685	0.576	0.355	0.208

(CCMp<sub>pred</sub>) is significantly superior to CCM<sub>pred</sub>; RCMPM (RaptorX-Contact) obtains similar results with RaptorX-Contact, while CCM<sub>pred</sub> is much lower than RaptorX-Contact; and RCMPM (CCMp<sub>pred</sub>) has yielded results comparable to RCMPM (RaptorX-Contact).

**Table 4** shows the comparison results on the CASP12 dataset. For the long-range contact type prediction, the results of the RCMPM by RCMPM (CCMp<sub>pred</sub>) are significantly better than those of CCM<sub>pred</sub>, with 24.8%, 28.1%, 36.1%, and 41.5% improvements at the top L/10, L/5, L/2, and L levels, while RCMPM (RaptorX-Contact) outperforms RaptorX-Contact with 4.3%, 3.8%, 1.6%, and 0.6% improvements on the four top levels, respectively. The results of RCMPM (RaptorX-Contact) are also significantly better than those of CCM<sub>pred</sub> and RCMPM (CCMp<sub>pred</sub>). **Figure 6** shows the comparison results of the five methods on the long-range contact type prediction. For both the medium-range contact and short-range contact types'

results, we can draw the following conclusions: among the three existing SOTA methods, R2C performs the best; RCMPM (CCMp<sub>pred</sub>) is significantly superior to CCM<sub>pred</sub>; RCMPM (RaptorX-Contact) obtains similar results with RaptorX-Contact; CCM<sub>pred</sub> is much lower than RaptorX-Contact, but the gap between RCMPM (CCMp<sub>pred</sub>) and RCMPM (RaptorX-Contact) is greatly reduced.

To summarize the long-range results of the four datasets, it could be found that our proposed RCMPM method is significantly superior to the other methods on PDB25 and CASP11. For CASP10, RCMPM performs much better than CCM<sub>pred</sub> and R2C, and although it does not perform as well as RaptorX, the gap is very small. For CASP12, the accuracy of RCMPM is higher than that of CCM<sub>pred</sub> and RaptorX, and is slightly lower than that of R2C. Therefore, it can be concluded that our proposed method performs best overall on the four datasets and is the most stable one as well.

### 3.4 Ablation Study

#### 3.4.1 Evaluation of the GCN Module of the Model Structure

In order to examine the effectiveness of our proposed method, we used two network structures to construct different network structures, the original RCMPM and the RCMPM removal GCN module (RCMPM (without GCN)). **Table 5** shows the comparison results on the PDB25 dataset. Compared to the RCMPM (without GCN), the RCMPM improves by 1.2%, 0.9%, 1.7%, and 2% on the top  $L/10$ ,  $L/5$ ,  $L/2$ , and  $L$  levels, respectively, while the RCMPM performs much similar with the RCMPM (without GCN) on both the short-range and medium-range levels. This is because the graph neural network module can utilize the output of the existing methods, especially on the global information level, and therefore reflected by improvements on the long-range level contact prediction.

#### 3.4.2 Evaluation of Different Feature Combinations

In order to verify the effectiveness of the sequence features on the long-range contact map prediction, we used three different feature combinations as the input of the 1D ResNet module and GCN module, including PSSM, PSSM, and secondary structure (PSSM+SS), including the PSSM, secondary structure, and solvent accessibility (PSSM+SS+SA), a total of  $L \times 26$  dimensional features. **Table 6** shows the comparison results by using the RaptorX-Contact outputs as the rough contact map on the PDB25 dataset. From **Table 6**, it could be found that on the long-range contact type, RCMPM (PSSM+SS+SA) is improved by 0.9%, 0.8%, 1.7%, and 0.6% at the four top levels compared to RCMPM (PSSM+SS) and 1.6%, 0.9%, 1.1%, and 1.2% on the four top levels compared to RCMPM (PSSM). Meanwhile, on the medium-range contact type, although the trend is the same as the long-range type, the increase is very small. On the short-range contact type, the results of the three methods are even very close. **Table 7** shows the comparison results by using the CCMpred outputs as the rough contact map on the PDB25 dataset. From **Table 7**, it could be found that on the long-range contact type, RCMPM (PSSM+SS+SA) is improved by 0.8%, 2.2%, 2.1%, and 2.1% at the four top levels compared to RCMPM (PSSM+SS) and 9.3%, 13.4%, 26.2%, and 44.8% at the four top levels compared to RCMPM (PSSM). On the medium-range and short-range contact types, the results of the RCMPM (PSSM+SS+SA) are also better than the other two types' results. The results show that the PSSM is a very important feature for the contact prediction, and the secondary structure and solvent accessibility are also beneficial. When the initial contact map is used as RaptorX-Contact, the secondary structure and solvent accessibility have a limited effect on the medium- and short-range contact type predictions, in part because the RCMPM uses the output of RaptorX-Contact, which already contains the secondary structure and solvent accessibility information.

## 4 DISCUSSION

In this article, we formulated the peptide contact map refinement task as a geometric 2D graph improvement and proposed a novel

refined contact map prediction model (RCMPM) to refine the protein inter-residue contact map predictions using graph convolutional neural networks (GCNNs) and one-dimensional and two-dimensional residual neural network (1D ResNet and 2D ResNet) architectures. Our method combines the residual neural networks for learning the local information with the graph convolutional neural networks for learning the global information, which can better capture the long-range contact relationship between the complex sequence inter-residues. The experimental results show that our method can refine the contact map greatly for the long-range contact type, that is to say, by using CCMpred outputs as the rough contact map, the RCMPM is significantly better than CCMpred, and by using the RaptorX-Contact outputs as the rough contact map, the RCMPM is significantly better than RaptorX-Contact as well. For the medium-range contact prediction, the degree of improvement is significantly reduced, and for the short-range contact prediction, there is not even a significant improvement. The main reason is that the GCN module of the RCMPM can utilize the outputs of the existing methods, which are highly reflected on the global information level, and therefore, the RCMPM model makes improvements mainly on the long-range contact types. By using a larger protein database in HHblits or PSI-BLAST to calculate the homology features of protein sequences and combining more effective features as inputs, we can expect to further improve the precision.

## DATA AVAILABILITY STATEMENT

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

## AUTHOR CONTRIBUTIONS

XS, CW, and YL contributed to conception and design of the study. JG performed the statistical analysis. JG wrote the first draft of the manuscript. JG, TZ, and XS wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version. The handling editor (JW) declared a past co-authorship with the author (YL).

## FUNDING

This research was funded by the National Natural Science Foundation of China (61972174), the Key Research and Development Project of Jilin Provincial Science and Technology Department (20210201080GX), Jilin Province Development and Reform Commission (2021C044-1), Guangdong Science and Technology Planning (2020A0505100018), Guangdong universities' innovation team (2021KCXTD015), and Guangdong key discipline (2021ZDJS138) projects.

## REFERENCES

- Adhikari, B., Hou, J., and Cheng, J. (2017). Dncon2: Improved Protein Contact Prediction Using Two-Level Deep Convolutional Neural Networks. *bioRxiv* 2017, 222893. doi:10.1093/bioinformatics/btx781
- Aharon, M., Elad, M., and Bruckstein, A. (2006).  $\ell_1$  K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *IEEE Trans. Signal. Process.* 54, 4311–4322. doi:10.1109/tsp.2006.881199
- Altschul, S., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped Blast and Psi-Blast: a New Generation of Protein Database Search Programs. *Nucleic Acids Res.* 25, 3389–3402. doi:10.1093/nar/25.17.3389
- Betancourt, M. R., and Thirumalai, D. (1999). Pair Potentials for Protein Folding: Choice of Reference States and Sensitivity of Predicted Native States to Variations in the Interaction Schemes. *Protein Sci.* 8, 361–369. doi:10.1110/ps.8.2.361
- Chen, J., Zheng, S., Zhao, H., and Yang, Y. (2020). Structure-aware Protein Solubility Prediction from Sequence through Graph Convolutional Network and Predicted Contact Map. *bioRxiv*.
- Cheng, J., and Baldi, P. (2007). Improved Residue Contact Prediction Using Support Vector Machines and a Large Feature Set. *BMC Bioinformatics* 8, 113. doi:10.1186/1471-2105-8-113
- Di Lena, P., Nagata, K., and Baldi, P. (2012). Deep Architectures for Protein Contact Map Prediction. *Bioinformatics* 28, 2449–2457. doi:10.1093/bioinformatics/bts475
- Dunn, S. D., Wahl, L. M., and Gloor, G. B. (2008). Mutual Information without the Influence of Phylogeny or Entropy Dramatically Improves Residue Contact Prediction. *Bioinformatics* 24, 333–340. doi:10.1093/bioinformatics/btm604
- Fout, A., Byrd, J., Shariat, B., and Ben-Hur, A. (2017). “Protein Interface Prediction Using Graph Convolutional Networks,” in *Neural Information Processing Systems*.
- Glorigrijević, V., Renfrew, P. D., Kosciolk, T., Leman, J. K., Berenberg, D., Vatanen, T., et al. (2021). Structure-based Protein Function Prediction Using Graph Convolutional Networks. *Nat. Commun.* 12, 3168. doi:10.1038/s41467-021-23303-9
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). “Deep Residual Learning for Image Recognition,” in *Computer Vision and Pattern Recognition*. doi:10.1109/cvpr.2016.90
- Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support Vector Machines. *IEEE Intell. Syst. Their Appl.* 13, 18–28. doi:10.1109/5254.708428
- Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.* 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735
- Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2017). “Densely Connected Convolutional Networks,” in *Computer Vision and Pattern Recognition*. doi:10.1109/cvpr.2017.243
- Jones, D. T., Buchan, D. W. A., Cozzetto, D., and Pontil, M. (2012). Psicov: Precise Structural Contact Prediction Using Sparse Inverse Covariance Estimation on Large Multiple Sequence Alignments. *Bioinformatics* 28, 184–190. doi:10.1093/bioinformatics/btr638
- Jones, D. T., and Kandathil, S. M. (2018). High Precision in Protein Contact Prediction Using Fully Convolutional Neural Networks and Minimal Sequence Features. *Bioinformatics* 34, 3308–3315. doi:10.1093/bioinformatics/bty341
- Jones, D. T., Singh, T., Kosciolk, T., and Tetchner, S. (2015). Metapsicov: Combining Coevolution Methods for Accurate Prediction of Contacts and Long Range Hydrogen Bonding in Proteins. *Bioinformatics* 31, 999–1006. doi:10.1093/bioinformatics/btu791
- Kamisetty, H., Ovchinnikov, S., and Baker, D. (2013). Assessing the Utility of Coevolution-Based Residue-Residue Contact Predictions in a Sequence- and Structure-Rich Era. *Proc. Natl. Acad. Sci. U.S.A.* 110, 15674–15679. doi:10.1073/pnas.1314045110
- Kipf, T., and Welling, M. (2016). Semi-supervised Classification with Graph Convolutional Networks. *arXiv: Learn*.
- Li, Z., Lin, Y., Elofsson, A., and Yao, Y. (2020). Protein Contact Map Prediction Based on Resnet and Densenet. *Biomed. Res. Int.* 2020, 7584968. doi:10.1155/2020/7584968
- Mirdita, M., von den Driesch, L., Galiez, C., Martin, M. J., Söding, J., and Steinegger, M. (2017). Uniclust Databases of Clustered and Deeply Annotated Protein Sequences and Alignments. *Nucleic Acids Res.* 45, D170. doi:10.1093/nar/gkw1081
- Moult, J., Fidelis, K., Kryshchak, A., Schwede, T., and Tramontano, A. (2014). Critical Assessment of Methods of Protein Structure Prediction (CASP) - Round X. *Proteins* 82, 1–6. doi:10.1002/prot.24452
- Moult, J., Fidelis, K., Kryshchak, A., Schwede, T., and Tramontano, A. (2018). Critical Assessment of Methods of Protein Structure Prediction (CASP)-round Xii. *Proteins* 86, 7–15. doi:10.1002/prot.25415
- Moult, J., Fidelis, K., Kryshchak, A., Schwede, T., and Tramontano, A. (2016). Critical Assessment of Methods of Protein Structure Prediction: Progress and New Directions in Round Xi. *Proteins* 84, 4–14. doi:10.1002/prot.25064
- Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). Hhblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment. *Nat. Methods* 9, 173–175. doi:10.1038/nmeth.1818
- Seemayer, S., Gruber, M., and Söding, J. (2014). CCMpred-Fast and Precise Prediction of Protein Residue-Residue Contacts from Correlated Mutations. *Bioinformatics* 30, 3128–3130. doi:10.1093/bioinformatics/btu500
- Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2020). Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature* 577, 706–710. doi:10.1038/s41586-019-1923-7
- Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., and Wu, C. H. (2015). Uniref Clusters: a Comprehensive and Scalable Alternative for Improving Sequence Similarity Searches. *Bioinformatics* 31, 926–932. doi:10.1093/bioinformatics/btu739
- Torrini, M., Pollastri, G., and Le, Q. (2020). Deep Learning Methods in Protein Structure Prediction. *Comput. Struct. Biotechnol. J.* 18, 1301–1310. doi:10.1016/j.csbj.2019.12.011
- Villegas-Morcillo, A., Morales-Cordova, J. A., Gomez, A. M., and Sanchez, V. (2018). “Improved Protein Residue-Residue Contact Prediction Using Image Denoising Methods,” in 2018 26th European Signal Processing Conference (EUSIPCO) (Rome, Italy: IEEE), 1167–1171. doi:10.23919/eusipco.2018.8553519
- Wang, S., Li, W., Liu, S., and Xu, J. (2016). Raptorx-property: a Web Server for Protein Structure Property Prediction. *Nucleic Acids Res.* 44, W430–W435. doi:10.1093/nar/gkw306
- Wang, S., Sun, S., Li, Z., Zhang, R., and Xu, J. (2017). Accurate De Novo Prediction of Protein Contact Map by Ultra-deep Learning Model. *Plos Comput. Biol.* 13, e1005324. doi:10.1371/journal.pcbi.1005324
- Xia, T., and Ku, W.-S. (2020). Deep Multi-Attribute Graph Representation Learning on Protein Structures. *arXiv: Learn*.
- Xu, J. (2018). Distance-based Protein Folding Powered by Deep Learning. *bioRxiv* 2018, 465955.
- Yang, J., Jin, Q.-Y., Zhang, B., and Shen, H.-B. (2016). R2c: Improving Ab Initio Residue Contact Map Prediction Using Dynamic Fusion Strategy and Gaussian Noise Filter. *Bioinformatics* 32, 2435–2443. doi:10.1093/bioinformatics/btw181
- Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., et al. (2018). Graph Neural Networks: A Review of Methods and Applications. *arXiv: Learn*.

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

The handling editor JW declared a past co-authorship with the author YL.

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

Copyright © 2022 Gu, Zhang, Wu, Liang and Shi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Ensemble-AHTPpred: A Robust Ensemble Machine Learning Model Integrated With a New Composite Feature for Identifying Antihypertensive Peptides

Supatcha Lertampaiporn<sup>1</sup>, Apiradee Hongsthong<sup>1</sup>, Warin Wattanapornprom<sup>2</sup> and Chinae Thammarongtham<sup>1\*</sup>

<sup>1</sup>Biochemical Engineering and Systems Biology Research Group, National Center for Genetic Engineering and Biotechnology, National Science and Technology Development Agency at King Mongkut's University of Technology Thonburi, Bangkok, Thailand, <sup>2</sup>Applied Computer Science Program, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi, Bangkok, Thailand

## OPEN ACCESS

### Edited by:

Yanjie Wei,  
Shenzhen Institutes of Advanced  
Technology (CAS), China

### Reviewed by:

Deepika Mathur,  
Icahn School of Medicine at Mount  
Sinai, United States  
Piyush Agrawal,  
National Cancer Institute,  
United States

### \*Correspondence:

Chinae Thammarongtham  
chinae@biotec.or.th

### Specialty section:

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

**Received:** 25 February 2022

**Accepted:** 04 April 2022

**Published:** 28 April 2022

### Citation:

Lertampaiporn S, Hongsthong A,  
Wattanapornprom W and  
Thammarongtham C (2022)  
Ensemble-AHTPpred: A Robust  
Ensemble Machine Learning Model  
Integrated With a New Composite  
Feature for Identifying  
Antihypertensive Peptides.  
Front. Genet. 13:883766.  
doi: 10.3389/fgene.2022.883766

Hypertension or elevated blood pressure is a serious medical condition that significantly increases the risks of cardiovascular disease, heart disease, diabetes, stroke, kidney disease, and other health problems, that affect people worldwide. Thus, hypertension is one of the major global causes of premature death. Regarding the prevention and treatment of hypertension with no or few side effects, antihypertensive peptides (AHTPs) obtained from natural sources might be useful as nutraceuticals. Therefore, the search for alternative/novel AHTPs in food or natural sources has received much attention, as AHTPs may be functional agents for human health. AHTPs have been observed in diverse organisms, although many of them remain underinvestigated. The identification of peptides with antihypertensive activity in the laboratory is time- and resource-consuming. Alternatively, computational methods based on robust machine learning can identify or screen potential AHTP candidates prior to experimental verification. In this paper, we propose Ensemble-AHTPpred, an ensemble machine learning algorithm composed of a random forest (RF), a support vector machine (SVM), and extreme gradient boosting (XGB), with the aim of integrating diverse heterogeneous algorithms to enhance the robustness of the final predictive model. The selected feature set includes various computed features, such as various physicochemical properties, amino acid compositions (AACs), transitions, n-grams, and secondary structure-related information; these features are able to learn more information in terms of analyzing or explaining the characteristics of the predicted peptide. In addition, the tool is integrated with a newly proposed composite feature (generated based on a logistic regression function) that combines various feature aspects to enable improved AHTP characterization. Our tool, Ensemble-AHTPpred, achieved an overall accuracy above 90% on independent test data. Additionally, the approach was applied to novel experimentally validated AHTPs, obtained from recent studies, which did not overlap with the training and test datasets, and the tool could precisely predict these AHTPs.



**Keywords:** antihypertensive, prediction, classification, ACE inhibitor, ACE inhibitory peptide, ensemble machine learning

## INTRODUCTION

Hypertension is a global health issue due to its worldwide incidence and association with increased mortality and morbidity (Mills et al., 2020). Chronic hypertension is a substantial risk factor for heart diseases, stroke, cardiovascular diseases, congestive heart failure, glomerulonephritis, arteriosclerosis, and other diseases (Zhou et al., 2021).

The renin-angiotensin system (RAS) or the renin-angiotensin-aldosterone system (RAAS) is responsible for blood pressure regulation. The RAS regulates blood pressure and cardiac output by controlling the flow of blood through the heart (Wu et al., 2018).

One of the most important enzymes in the RAS system, angiotensin-converting enzyme (ACE), regulates blood pressure and fluid/salt homeostasis (He et al., 2014; Balgir and Sharma 2017). In the RAS, renin transforms angiotensinogen into angiotensin-I (ANG I), and subsequently, ACE transforms the inactive decapeptide angiotensin-I (ANG I) into the vasoconstrictor octapeptide angiotensin-II (ANG II). Excessive ACE activity results in the production of excessive amounts of angiotensin II and, as a result, an increase in blood pressure (i.e., it upregulates blood pressure) (Zhu et al., 2021).

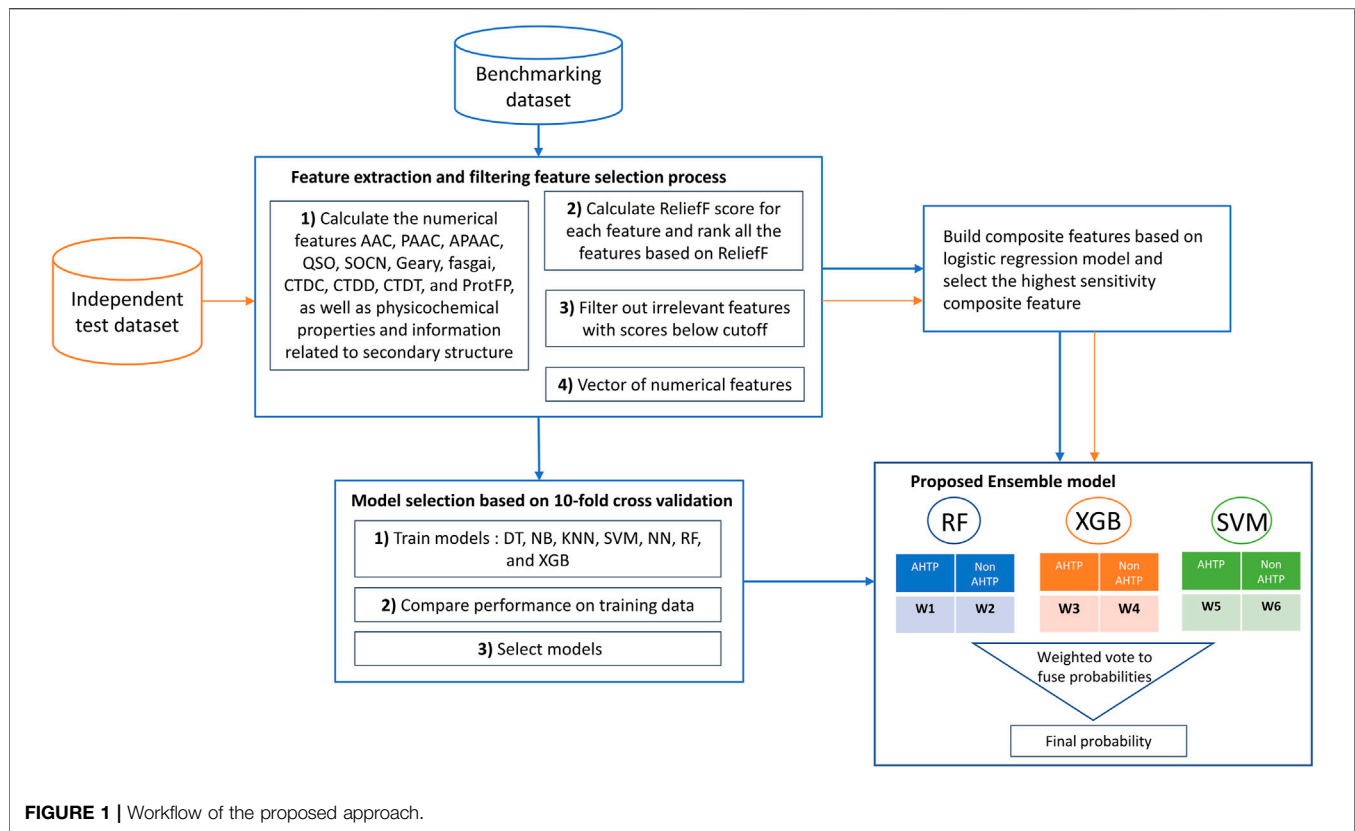
ACE inhibition is a well-established technique for developing pharmaceuticals for the treatment of hypertension. Synthetic ACE inhibitors such as captopril, enalapril, cilazapril, benazepril, and lisinopril are typically used in clinical hypertension treatments (Daskaya-Dikmen, et al., 2017). However, the long-term treatment of hypertension with these drugs is accompanied by severe or mild adverse effects, such as cough, headache, diarrhea, dizziness, fatigue, angioedema, hyperkalemia, hypotension, or, in rare cases, renal impairment (De Leo et al., 2009; Nguyen et al., 2010; Norris and FitzGerald, 2013; Daskaya-Dikmen et al., 2017; Abachi et al., 2019; Festa et al., 2020).

Antihypertensive peptides (AHTPs) are bioactive peptides obtained from natural foods that have the effects/activities of ACE inhibitors against hypertension and are considered safe for consumption, with fewer adverse side effects than synthetic ACE inhibitor drugs or even no side effects. These natural ACE inhibitory bioactive peptides are highly desired for the development of functional foods, nutraceuticals and pharmaceuticals for the prevention and treatment of hypertension (Norris and FitzGerald, 2013; de Castro and Sato, 2015; Kumar et al., 2015; Abachi et al., 2019; Pujiastuti et al., 2019; Jiang et al., 2021; Zaky et al., 2022). Peptides are often multifunctional and may exhibit several health-promoting bioactivities, such as antioxidative, antihypertensive, anti-inflammatory, cytoprotective, and antimicrobial effects (He et al., 2019; Jakubczyk et al., 2020). Emerging evidence indicates that AHTPs may mediate antihypertensive effects by interacting with RAS-related renin, AT-II receptors, arginine-nitric oxide pathway, endothelin system, or  $\text{Ca}^{2+}$

channels in addition to ACE inhibition (Udenigwe and Mohan, 2014; Aluko 2015). AHTPs have major potential as functional ingredients (dietary compounds) in a daily diet aimed at helping prevent and safely manage hypertension and enhancing human health (Norris and FitzGerald, 2013; Jakubczyk et al., 2020). Therefore, the identification of new, nontoxic bioactive peptides derived from food or natural sources has received significant attention. As a consequence, an increasing number of food-derived antihypertensive peptides have been studied and reported (Martínez-Maqueda et al., 2012; Kumar et al., 2014; Abachi et al., 2019; Lee and Hur 2019; Pujiastuti et al., 2019; Lu et al., 2021). Finding new AHTPs in various organisms is currently a significant research topic. However, large-scale identification through wet laboratory experiments is a costly, time consuming, and labor-intensive approach (Li-Chan 2015; Pujiastuti et al., 2019; Festa et al., 2020). The use of bioinformatics and *in silico* methods for the identification of potential candidate AHTPs for subsequent experimental assays is necessary to shorten the process. The development of efficient computational approaches will facilitate the processes of discovery and screening, allowing potential novel AHTP candidates to be identified in a cost-, resource- and time-effective manner.

A few existing machine learning-based computational approaches are available for predicting AHTPs. mAHTPred is a meta-predictor that employs a two-step feature selection methodology (Manavalan et al., 2019). PAAP is an RF classification model approach based on varied combinations of amino acids, dipeptides, and pseudo amino acid composition descriptors (Win et al., 2018). AHTpin was developed to screen, predict, and design AHTPs by using an SVM-based regression model for tiny peptides and SVM-based classification models for small, medium and large peptides (Kumar et al., 2015). Additionally, an SVM prediction tool was recently built by using convolutional neural network (CNN) deep learning-based encoding features derived from amino acid compositions (AACs) and dipeptide composition features (Rauf et al., 2021).

Although certain tools for AHTP prediction are available, the development of our ensemble method is different from that of the existing approaches in several ways. First, we developed a weighted voting method for integrating the strengths of three independent machine learning models, each of which has high levels of performance in different aspects. Second, a new composite feature called comF2 was developed based on a logistic regression statistical framework. In both the RF and extreme gradient boosting (XGB) feature importance plots, this feature was ranked as the most significant. In addition, a Shapley additive explanations (SHAP) analysis revealed consistent results, showing that comF2 was the top-ranked feature and was capable of explaining large samples in the model; therefore, it could capture characteristics for most of the AHTPs in the training data. Third, our ensemble method



outperformed previously developed methods in terms of robustness and accuracy when predicting independent testing datasets, with an enhanced accuracy of 90.4%. Last, the technique could also correctly classify many novel unseen, and experimental AHTPs collected from recent studies.

## MATERIALS AND METHODS

### Workflow

The workflow of Ensemble-AHTPpred is shown in **Figure 1**.

### Datasets

In this study, we employed two nonredundant datasets from mAHTPpred (Manavalan et al., 2019): a benchmarking dataset and an independent testing dataset. The balanced benchmarking dataset contained 913 unique AHTPs and 913 unique non-AHTPs. The 913 AHTPs were experimentally validated on the publicly available AHTPDB (Kumar et al., 2015) and BIOPEP (Minkiewicz et al., 2008; Iwaniak et al., 2016) databases. Note that experimentally validated non-AHTPs were not available as a public non-AHTP database. Therefore, the non-AHTPs were random peptides generated from Swiss-Prot proteins. Considering random sequences as a negative dataset is a routinely used standard procedure in many peptide-based prediction methods (Sharma et al., 2013; Kumar et al., 2015; Chen et al., 2016; Usmani et al., 2018; Manavalan et al., 2019) with the assumption that the probability of finding a random sequence

to be positive is very low. Positive and negative training datasets have similar length distributions. The AHTPs in the benchmarking dataset have a length between 5 and 81 amino acids, with an average length of 7.7 amino acids. The non-AHTPs in the benchmarking dataset have a length between 5 and 45, with an average length of eight amino acids.

Another dataset, an independent dataset, was composed of 386 nonredundant, experimentally validated AHTPs (Win et al., 2018; Yi et al., 2018) and 386 random peptides generated from Swiss-Prot as negative samples. The AHTPs in the independent testing dataset have a length between 5 and 24 amino acids, with an average length of 6.48 amino acids. The non-AHTPs in the independent testing dataset have a length between 5 and 29, with an average length of 15.42 amino acids.

### Features

The peptide properties that were relevant for predicting AHTPs were determined and encoded as a vector of 431 numerical features. The features can be grouped into seven main types as follows.

- 1) AAC descriptors: These descriptors were used as the fractions of each amino acid type within a protein sequence. The fractions of all 20 natural amino acids {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}, were calculated. (**AAC1-AAC20: 20 dimensions**).
- 2) Chou's pseudo amino acid composition (PseAAC) was generated in various modes: Chou's PseAAC (Chou, 2005)

has been widely used to convert complicated protein sequences with various lengths to fixed-length numerical feature vectors that incorporate sequence-order information. In comparison with an AAC, a PseAAC is more informative and capable of representing a protein sequence and incorporating information about its sequence order. Hence, it has been widely used for diverse amino acid sequence-based prediction problems (Chou, 2011). The PseAACs were calculated by using parameters of  $\lambda = 3$  and  $w = 0.05$  (**PAAC1-PAAC23: 23 dimensions**). PseAACs in parallel correlations (**Pse\_PC1-Pse\_PC22: 22 dimensions**), PseAACs in series correlations (**Pse\_SC1-Pse\_SC26: 26 dimensions**), and amphiphilic pseudo AACs with hydrophobicity correlation functions (**APAAC1\_1-APAAC1\_23: 23 dimensions**) and hydrophilicity correlation functions (**APAAC2\_1-APAAC2\_23: 23 dimensions**) were also calculated.

- 3) Composition/transition/distribution (C/T/D): The three descriptors based on the grouped AACs (Dubchak et al., 1995) [composition (**CTDC1-CTDC21: 21 dimensions**), transition (**CTDT1-CTDT21: 21 dimensions**) and distribution (**CTDD1-CTDD105: 105 dimensions**) descriptors] were calculated. C/T/D was calculated using the *protr* R package (Xiao et al., 2015). All amino acid residues were divided into three groups according to seven types of physicochemical properties, as defined in Dubchak et al. (1999). The seven physicochemical properties used for calculating these features were hydrophobicity, normalized van der Waals volume, polarity, polarizability, charge, secondary structures, and solvent accessibility.
- 4) Quasi-sequence-order descriptors: The quasi-sequence-order descriptors were derived from the distance matrix of the 20 amino acids (Chou 2000). Quasi-sequence-order descriptors (**QSO1-QSO46: 46 dimensions**) and sequence-order-coupling numbers (**SOCN1-SOCN6: 6 dimensions**) ( $\text{lag} = 3$ ,  $w = 0.1$ ) were calculated.
- 5) Various physicochemical and topological property-based features: The Crucian properties covariance index (**Crucian1-Crucian3: 3 dimensions**) (Cruciani et al., 2004), Z-scales based on physicochemical properties (**zscales1-zscales5: 5 dimensions**) (Sandberg et al., 1998), factor analysis scales of generalized amino acid information (**fsgai1-fsgai6: 6 dimensions**) (Liang and Li 2007), T-scales based on physicochemical properties (**tScales1-tScales5: 5 dimensions**) (Tian et al., 2007), VHSE-scales (principal component score vectors of hydrophobic, steric, and electronic properties) (**vhscscales1-vhscscales8: 8 dimensions**) (Mei et al., 2005), protFPs (**protFP1-protFP8: 8 dimensions**) (van Westen et al., 2013), ST-scales based on physicochemical properties (**stscscales1-stscscales8: 8 dimensions**) (Yang et al., 2010), MS-WHIM scores (**mswhimscore1-mswhimscore3: 3 dimensions**) (Zaliani and Gancia 1999), aliphatic indices of proteins (**aIndex: 1 dimension**) (Ikai, 1980), Geary autocorrelations (**geary1-geary12: 12 dimensions**), the autocovariance index (**autcov: 1 dimensions**) (Ikai, 1980), the potential protein interaction index (**Boman: 1 dimension**) (Boman, 2003), the net charge (**Charge: 1 dimension**), cross-covariance indices (**Crosscov1-Crosscov2: 2 dimensions**), instability indices (**Instaindex: 1 dimension**) (Guruprasad et al., 1990), the hmoment alpha helix (**Hmoment1: 1 dimensions**), the hmoment beta sheet (**Hmoment2: 1 dimensions**), BLOSUM matrix-derived descriptors (**Blosum1-8: 8 dimensions**), and the isoelectric point (**pI: 1 dimension**) were calculated by using the peptide R package (Osorio et al., 2015).
- 6) Occurrence of selected k-mer motifs: The YP, HLP, IYP, LHL, LPP, LRP, VPP, PEV, PFP, QTP, VLP, VYP, and YPF motifs (**13 dimensions**) were determined. First, we generated all 2-mers (400 dimensions) and all 3-mers (8000 dimensions). Then, we searched for the k-mer that was overrepresented in the positive and underrepresented in the negative datasets by calculating the log odds ratio score of the frequency of each k-mers in the positive versus negative datasets. Next, we ranked the discriminant k-mers based on the calculated log-odds score. Finally, we retained the top 2-mer and the top 12 3-mers as selected k-mer motif features that still need to be determined (the heatmap of log odds scores of 2-mers is shown in Figure 5).
- 7) Secondary structure conformation-related features: The aggregation, amyloid, turn, alpha-helix, helical aggregation, and beta-strand conformation secondary structure propensities were calculated using the Tango program (**tango1-tango6: 6 dimensions**) (Fernandez-Escamilla et al., 2004).

To further improve the prediction process with new informative features, we proposed a composite feature generation method *via* the fusion of the various selected features by using a logistic regression model. Various composite features based on various combinations of informative selected features were built by using logistic regression based on the benchmarking data and then compared through a 10-fold cross-validation process. The detailed process of building composite features is described in the hybrid feature section of ensemble-AMPPred (Lertampaiporn et al., 2021). A combination of features was used to fit a logistic regression model, which is represented by the following equation:

$$\text{Prob. } (Y = \text{AHTPs}|x) = \text{logistic}(x) \\ = \left( \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_n X_n}} \right)$$

Logit transformation (the logarithm of the odds ratio that Y is in the AHTP category) was applied to link a function with the logistic regression. The logit function is defined as

$$\text{Logit}(x) = \log \left( \frac{P(Y = \text{AHTP}|X = X)}{P(Y = \text{nonAHTP}|X = X)} \right) \\ = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_n X_n$$

Therefore, the composite feature was defined by the following equation:

$$\text{Composite feature} = \beta_0 + \beta_1 \text{feature}_1 + \beta_2 \text{feature}_2 \\ + \beta_3 \text{feature}_3 + \dots + \beta_n \text{feature}_n$$

where  $\beta_0$  is the intercept;  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$ , and  $\beta_n$  represent the regression coefficients for each selected feature in the equation; and  $\text{feature}_1$ ,  $\text{feature}_2$ , ..., and  $\text{feature}_n$  are the component features in the composite feature.

## Feature Selection

A feature selection procedure based on ReliefF (Kononenko, 1994) scores was used as a preprocessing step to filter irrelevant features with a cutoff score. The ReliefF score for a feature was calculated based on how well the feature could distinguish between instances that were near each other. The ReliefF evaluation criterion selected features that aided in the separation of the samples from different classes and gave higher weights to the features that discriminated the samples from the neighborhoods of different classes.

Recursive feature elimination (RFE) (Tolosi and Lengauer, 2011) is a wrapper-type feature selection algorithm. RFE starts with all features in the training dataset and then searches for a subset of features by removing features through recursive elimination to eliminate the least relevant features one by one and refitting the model. This process is repeated until the optimal number of features is reached, ensuring that the classifier can achieve high performance.

## Models

To select base classifiers for constructing an ensemble, seven machine learning algorithms were considered in our algorithm selection experiment—a naïve Bayes (NB) model, a neural network (NN), a support vector machine (SVM), k-nearest neighbors (kNN), a decision tree (DT), a random forest (RF) and an extreme gradient boost (XGB). Each algorithm has a different inductive bias and different learning hypotheses that can provide a potentially more independent and diverse set of predictions through the ensemble. For the hyperparameters, we used a grid search to find the optimal parameters.

The NB classifier is a simple probabilistic classifier based on Bayes' theorem and substantial independence assumptions between the features.

The NN was a multilayer perceptron (MLP). An MLP is a neural network with at least three layers: an input layer, a hidden layer, and an output layer (parameters: number of epochs: 500; learning rate: 0.3; and momentum for updating weights: 0.2).

The SVM model is a supervised learning model with associated learning algorithms for data classification and regression analysis. The SVM assigns training examples to coordinates in a high-dimensional space to widen the distance between the two classes and separates the two classes with a simple hyperplane (parameters:  $C = 36.0$ ; kernel = 'Radial Basis Function'; and  $\gamma = 0.119$ ).

The KNN method is a well-known nonparametric technique used in statistical pattern classification due to its simplicity, intuitiveness, and effectiveness. The essential principle is that an unclassified object is assigned to the class to which the majority of its  $k$  nearest neighbors belong (parameters:  $k = 7$  and distance = inverse weight).

The DT is another nonparametric supervised learning method used for classification and regression. It develops a model that accurately predicts the value of a target variable by inferring basic decision rules from data attributes. A tree can be thought of as an approximation to a piecewise constant (parameter: confidence factor = 0.25).

The RF algorithm is one of the most commonly used bagging ensemble algorithms because of its flexibility and ease of use. This algorithm can produce good results without hyperparameter tuning. The RF approach is an ensemble technique with the ability to achieve high accuracy and prevent overfitting by making use of voting with multiple decision trees (parameters: no. estimators = 350 and  $\text{max\_depth} = 12$ ).

The XGB algorithm is a gradient boosting ensemble algorithm. The boosting algorithm adjusts the model weights according to a differential loss function and then uses the adjusted weights in the next training iteration [parameters: no. estimators (nrounds) = 800;  $\text{max\_depth} = 10$ ;  $\text{eta} = 0.01$ ; and  $\text{subsample} = 0.8$ ].

The proposed method was implemented by using Perl, Python, and R scripts. The program was run on a Fedora Linux-based machine. All the data, the trained models and the standalone program are available to download at [http://ncrna-pred.com/Ensemble\\_AHTPpred.htm](http://ncrna-pred.com/Ensemble_AHTPpred.htm).

We adopted 10-fold cross-validation to investigate the classification performance of the various models on the benchmarking dataset. Based on the 10-fold cross validation results, model selection processes were performed. Then, the best-performing models were selected based on their diverse measurements and later used as the base classifiers of the ensemble model. Thereafter, the individual base classifiers were iteratively trained to find the optimal weight for each class of each classifier. The probability weight set ( $w_1, w_2, w_3, w_4, w_5, w_6$ ) was estimated by using the level of confidence in predicting each class (AHTP or non-AHTP), which fluctuated among the classes. The probabilities acquired from the base classifiers were aggregated through weighted voting to obtain the final prediction of the ensemble model.

Probability-weighted voting = ( $W_1 \cdot \text{Prob. (RF}_{\text{class=AHTP}})$ ) + ( $W_2 \cdot \text{Prob. (RF}_{\text{class=non-AHTP}})$ ) + ( $W_3 \cdot \text{Prob. (XGB}_{\text{class=AHTP}})$ ) + ( $W_4 \cdot \text{Prob. (XGB}_{\text{class=non-AHTP}})$ ) + ( $W_5 \cdot \text{Prob. (SVM}_{\text{class=AHTP}})$ ) + ( $W_6 \cdot \text{Prob. (SVM}_{\text{class=non-AHTP}})$ ).

To evaluate the classification performance of the model, the following metrics were used:

$$\text{ACC} = \frac{TP + TN}{(TP + TN + FP + FN)}$$

$$\text{Sn} = \frac{TP}{(TP + FN)}$$

$$\text{Sp} = \frac{TN}{(TN + FP)}$$

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

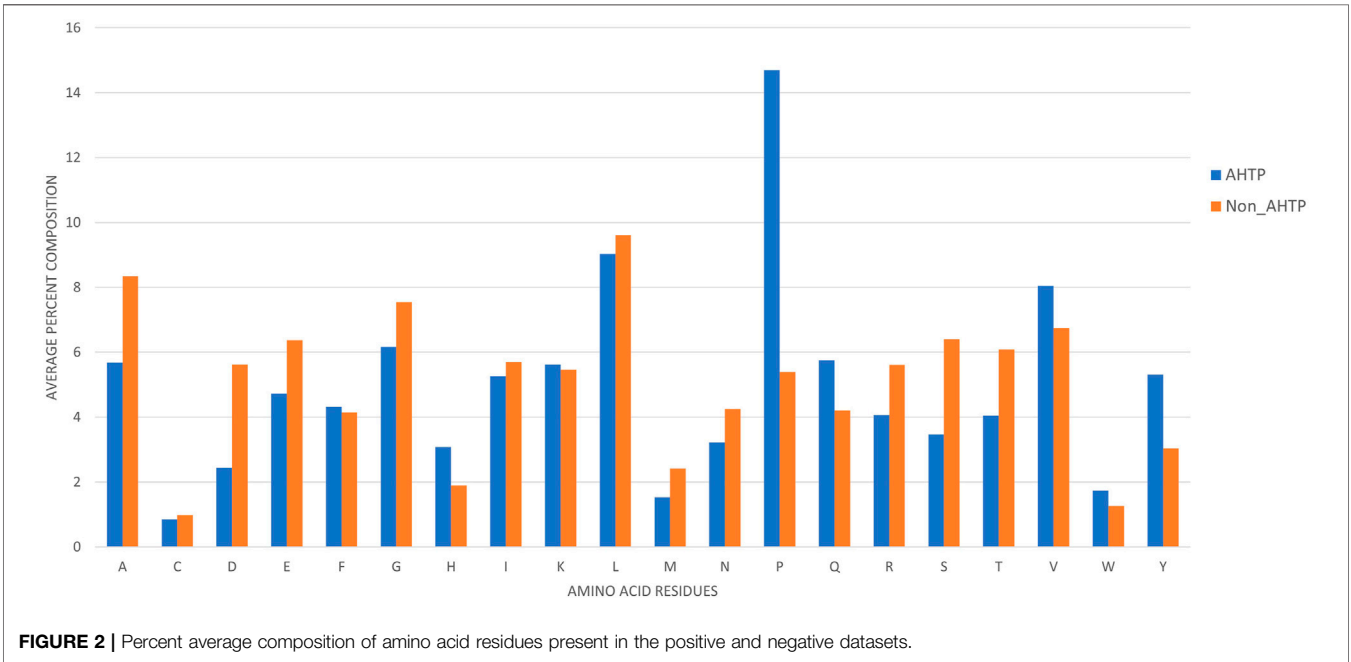
where ACC, Sn, Sp, and MCC are accuracy, sensitivity, specificity, and Matthew's coefficient correlation, respectively. These measurements were calculated based on the numbers of true



**TABLE 1 |** Physicochemical property-based composition of amino acids.

Physicochemical property-based composition of amino acids	Positive dataset (AHTPs)	Negative dataset (non-AHTPs)
Molecular weight of the peptide (Da)	888.2	<b>912.5</b>
Number of amino acids in the sequence	7.75	<b>8.05</b>
% Composition of charged residues (DEKHR)	19.91	<b>24.94</b>
% Composition of aliphatic residues (ILV)	<b>22.34</b>	22.04
% Composition of aromatic residues (FHWY)	<b>14.42</b>	10.32
% Composition of polar residues (DERKQN)	25.81	<b>31.49</b>
% Composition of neutral residues (AGHPSTY)	<b>43.44</b>	37.68
% Composition of hydrophobic residues (CVLIMFW)	30.75	<b>30.83</b>
% Composition of positively charged residues (HKR)	12.75	<b>12.96</b>
% Composition of negatively charged residues (DE)	7.16	<b>11.98</b>
% Composition of tiny residues (ACDGST)	22.65	<b>34.97</b>
% Composition of small residues (EHILKMNPQV)	<b>61.94</b>	51
% Composition of large residues (FRWY)	<b>15.41</b>	14.03

The higher values, between the two datasets, are shown in bold.



**FIGURE 2 |** Percent average composition of amino acid residues present in the positive and negative datasets.

positives (TPs), true negatives (TNs), false positives (FPs) and false negatives (FNs). The area under the receiver operating characteristic (ROC) curve (AUC) was calculated to assess the tradeoff between the sensitivity and specificity performance of the different methods. The ROC curve is a plot of the TP vs. FP rates at different thresholds. For a perfect predictor, the AUC is equal to 1.

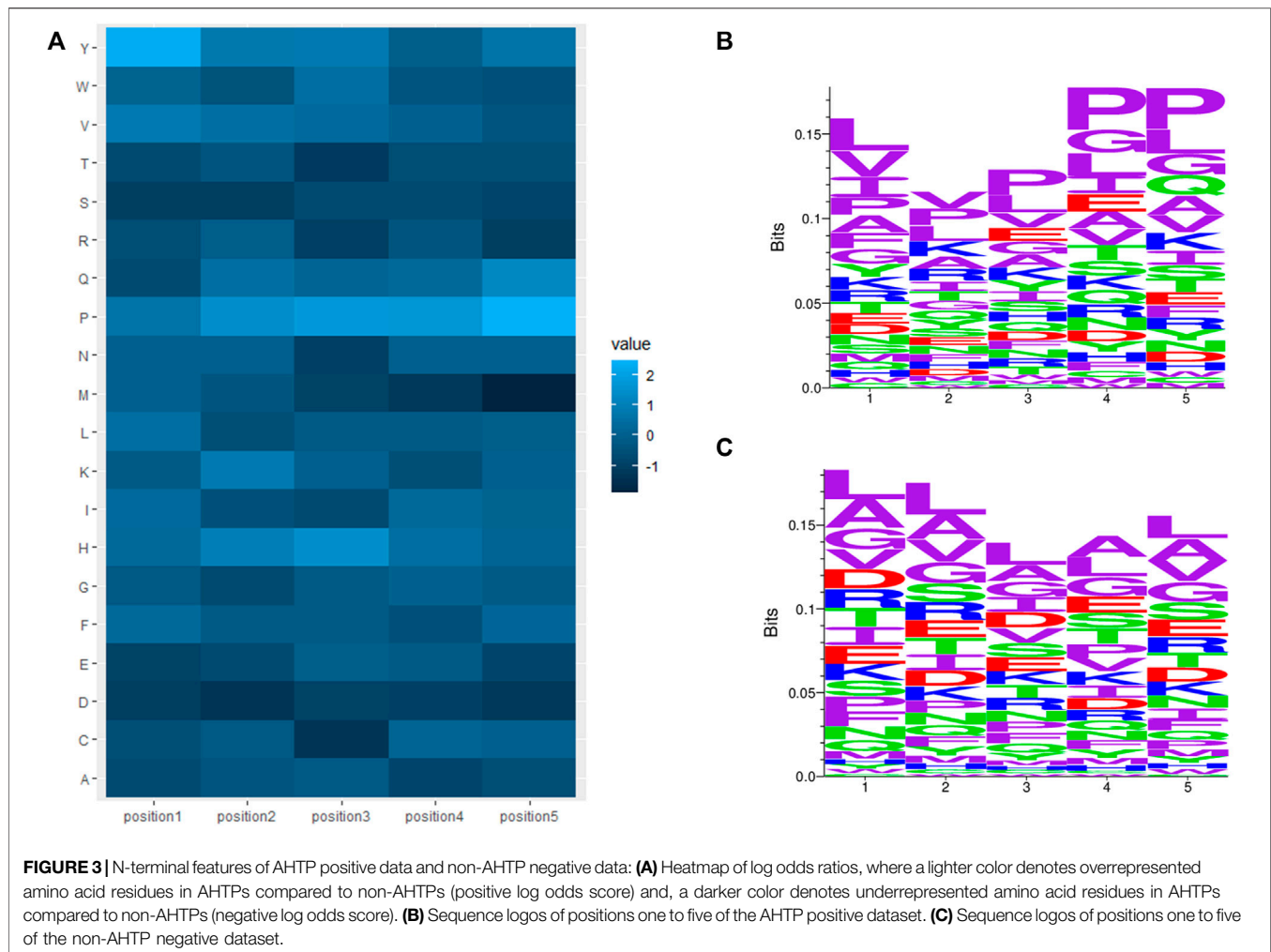
## RESULTS AND DISCUSSION

### Amino Acid Composition and Positional Residue Analysis

The activity of peptides depends on their structure and amino acid composition. To understand the relation between the

composition and antihypertensive function of a peptide, the composition of AHTPs and non-AHTPs were analyzed/ investigated. Generally, most antihypertensive peptides are relatively short peptide residues with lengths that vary from 2 amino acids to 20 amino acids. The amino acid composition is a quantitative measure of the fraction of each amino acid type within a protein. The percent amino acid composition based on the physicochemical properties of amino acids (whole peptides) was computed and calculated using COPid (Kumar et al., 2008) and includes the composition of charged (DEKHR), aliphatic (ILV), aromatic (FHWY), polar (DERKQN), neutral (AGHPSTY), hydrophobic (CVLIMFW), positively charged (HKR), negatively charged (DE), tiny (ACDGST), small (EHILKMNPQV) and large (FRWY) residues, as summarized in **Table 1** (a category with higher composition is shown in bold).



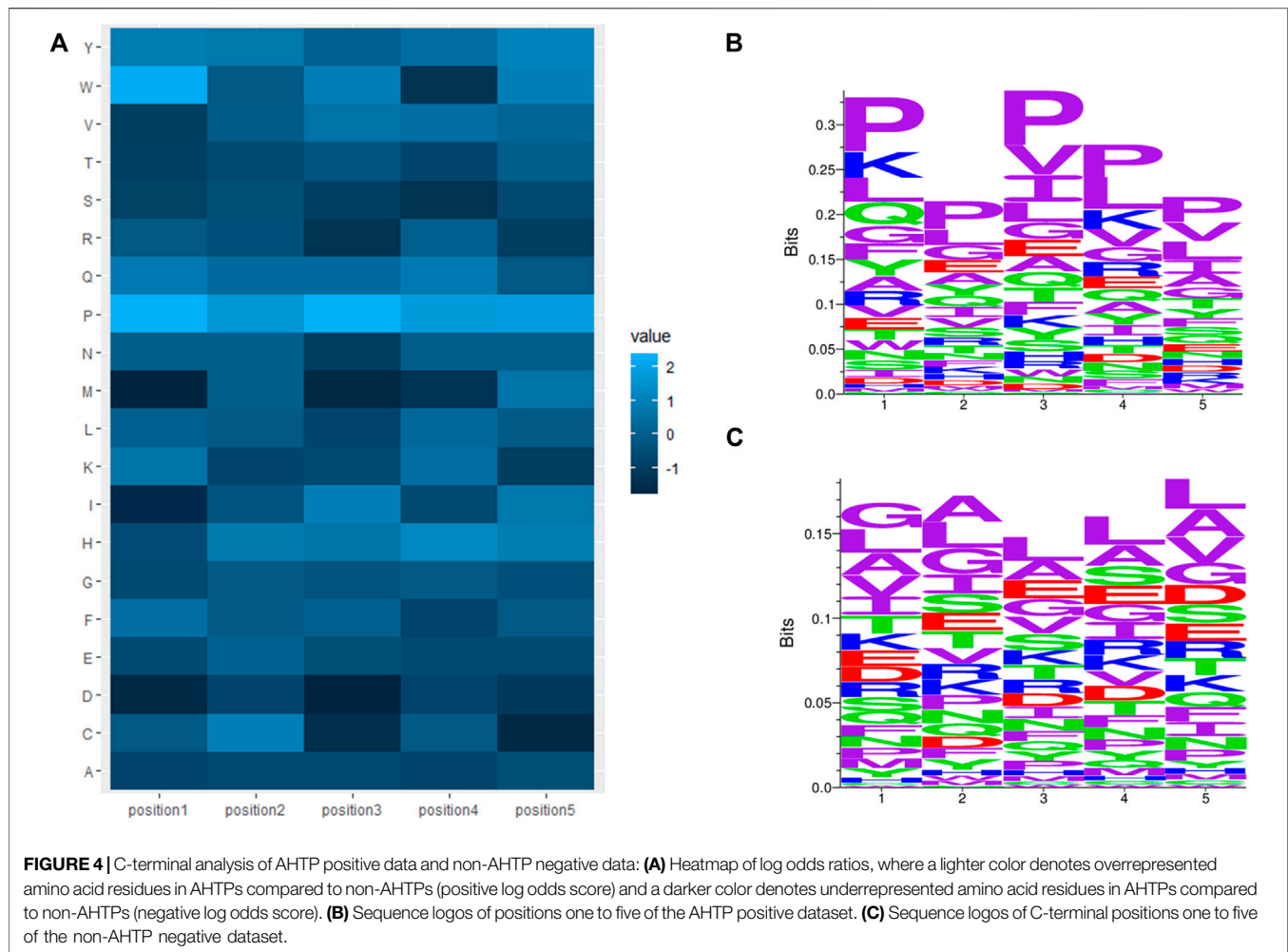


When comparing positive and negative of benchmarking datasets, we can see that AHTPs include more aliphatic (ILV), aromatic (FHWY), and neutral (AGHPSTY) amino acid residues than non-AHTP sequences.

Amino acid residues present in AHTPs and non-AHTPs were compared, as shown in **Figure 2**. Histidine (H), proline (P), glutamine (Q), valine (V), tryptophan (W) and tyrosine (Y) more frequently occurred in AHTPs than in non-AHTPs, especially proline (P), which is highly abundant in AHTPs. In contrast, certain residues such as cysteine (C), aspartic acid (D), methionine (M), and tryptophan (W) occurred rarely in AHTPs. Certain types of residues occurred frequently in both AHTPs and non-AHTPs, such as leucine (L) and valine (V). Amino acids such as alanine (A), aspartic Acid (D), and serine (S) were less frequent in AHTPs than in non-AHTPs.

C-terminal and N-terminal positional residue analysis was also performed by calculating the average amino acid composition of position one to position five of the N- and C-termini in AHTPs (positive) and non-AHTPs (negative). The log odds ratios between positive and negative N- and C-termini were calculated. The log-odds ratios of positive versus negative termini were calculated as  $[\log_2 (P_a/N_a)]$ ,

where  $P_a$  and  $N_a$  are the observed frequencies of amino acid  $a$  in the positive and negative training datasets, respectively. Heatmaps of log odds ratios were plotted for the N-terminal and C-terminal regions, as shown in **Figures 3A, 4A**. The sequence logos of positions one to five of the N- or C-terminus were generated by using Seq2Logo (Thomsen and Nielsen, 2012). **Figures 3B,C** display N-terminal positional sequence logos of AHTPs and non-AHTPs, respectively. (In sequence logos, specific colors were assigned to amino acids as follows, purple represents nonpolar sidechains (G A V L I M F W P), blue represents basic amino acid (K R H), Red represents acidic amino acid (D E), and green represents polar sidechains (S T C Y N Q); the height of the amino acids is proportional to their frequency at that position.) The most abundant amino acids in the N-terminus of AHTPs were Leu (9.069%), Pro (14.896%), Tyr (5.214%) and Val (8.697%). The most abundant amino acids in the C-terminus of AHTPs were Leu (9.003%), Pro (16.605%), and Val (7.338%). The most abundant amino acids in the N- and C-termini of non-AHTPs were Leu, Ala, Gly, and Val. The most abundant 2-mers in the N-terminus of AHTPs were YP, LP, PF, PP, and VP, while the most abundant 2-mers in the C-terminus of AHTPs were IP,



FP, PL, PP, PV, QP and VP. The most abundant 2-mers in the N- and C-termini of non-AHTPs were AA, LA, AL, LG, LE, and AR.

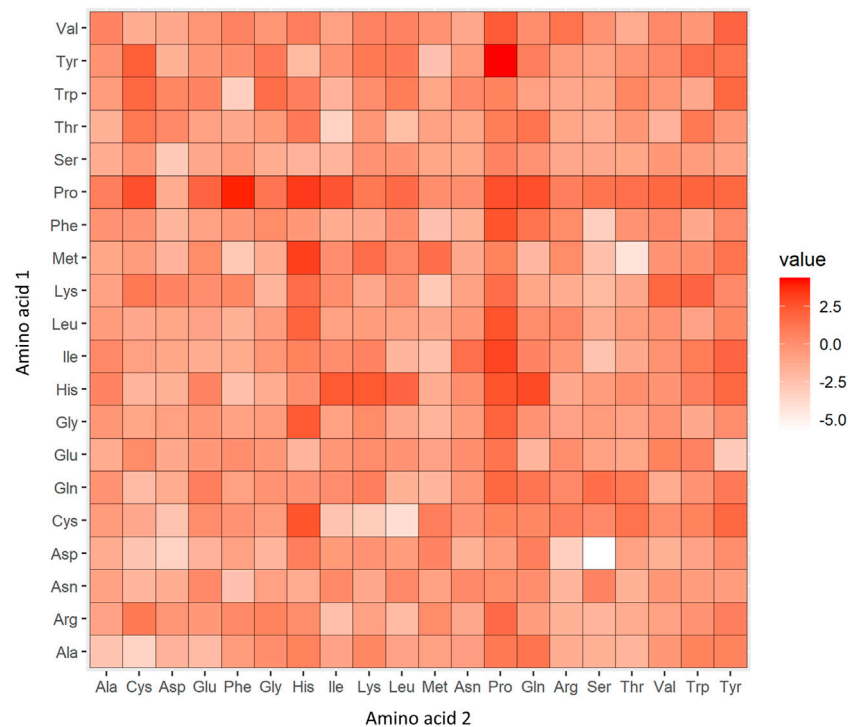
In addition, a heatmap of the log odds score of occurrences of the 2-mer motif in the whole sequence of AHTPs vs. in the whole sequence of non-AHTPs was also plotted, as shown in **Figure 5**. TyrPro (log odds = 4.393), ProPhe (log odds = 3.896) and ProHis (log odds = 3.340) were overrepresented in AHTP positive data compared to non-AHTP negative data. In contrast, AspSer (log odds = -5.708), MetThr (log odds = -4.292) and CysLeu (log odds = -4.070) were overrepresented 2-mers in non-AHTP negative data relative to AHTP positive data.

### Performance Evaluation Based on the Benchmarking Dataset to Select the Base Models for the Ensemble

Before training a prediction model, feature extraction and feature selection are two important steps for extracting various numerical features to represent biological sequences and then selecting relevant and discriminative features so that a

machine learning model can further analyze and detect the generalized pattern of the data of interest. In this work, we extracted a total of 431 numerical features to represent peptide sequences.

Since we collected as many features that could explain the peptides as possible, these 431 extracted features may have contained irrelevant and noninformative features with respect to explaining the AHTPs. Feature selection is required to eliminate irrelevant and redundant features that do not explain the target class. Furthermore, feature selection mitigates the curse of dimensionality (by reducing the number of dimensions) and prevents overfitting. Filter, wrapper, and embedding techniques are the three primary feature selection methods. Both the wrapper and embedding methods are tightly coupled with specific classification algorithms. The wrapper requires one predetermined classification algorithm and relies on its performance to evaluate and select the feature subset. This approach seeks the features that are best suited to the predetermined algorithm. As a result, these methods first necessitated determining the classification algorithm to be used. However, we intended to create an ensemble consisting of multiple



**FIGURE 5 |** Heatmap of the log odds scores of 2-mers abundant in the positive versus negative datasets. In the heatmap, a red color (high log odds score) denotes 2-mers overrepresented in AHTPs compared to non-AHTPs, and a white color (low log odds score) denotes 2-mers underrepresented in AHTPs compared to non-AHTPs.

**TABLE 2 |** Classification performance of different trained models.

	DT	NB	KNN	NN	SVM	XGB	RF
ACC (%)	73.494%	74.465%	74.918%	76.177%	80.504%	78.925%	<b>80.668%</b>
Sn	0.714	0.696	0.690	0.721	0.758	<b>0.789</b>	0.752
Sp	0.756	0.814	0.808	0.803	0.852	0.791	<b>0.861</b>
AUC	0.766	0.793	0.791	0.831	<b>0.878</b>	0.861	0.877

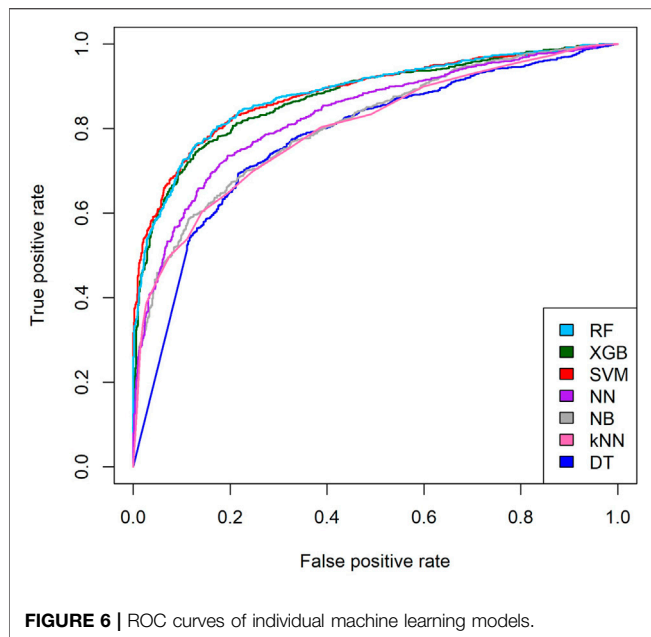
The highest values are in bold.

classification algorithms. Therefore, the filtering procedure was used initially to remove irrelevant features during this step. Note that the filtering method may not eliminate redundant features. We applied the filtering method based on ReliefF scores. After applying the filtering method, a total of 379 features had scores that were higher than the cutoff score. The vector containing these 379 numerical features was then used to train the 7 algorithms.

The training process was carried out *via* 10-fold cross-validation on a benchmarking dataset to investigate the classification performance of different trained models. **Table 2** shows the performance of the individual trained models. Different algorithms were able to take advantage of different characteristics and relationships contained in a given dataset. In this process, we

detected and combined the strengths of distinct algorithms to form a resilient and stable ensemble. The findings support the “no free lunch” theorem, which states that there is no single best algorithm that is superior in terms of every metric. The ROC curves of individual classification model performance are plotted in **Figure 6**.

Based on the performance obtained during the training process, **Table 2** shows that XGB had the highest sensitivity (0.789), followed by the SVM (0.758). The AUC provides a measure for evaluating which models are better on average by weighing the tradeoff between sensitivity and specificity. For the AUC metric, the SVM model achieved the highest score of 0.878, followed by the RF model (0.877), indicating that these two models achieved a good balance between positive and negative prediction. The RF model had the highest classification accuracy of 80.668% among the seven



trained models. Accordingly, based on the evaluation, we chose the SVM, the RF, and XGB as the ensemble members because of their superior performance in terms of different metrics.

Note that the input vectors for the SVM model were drawn from a separate collection of features. Because the RF and XGB have built-in feature selection, we used the complete 379-feature vector as the input feed. However, for the SVM-based model, we used RFE as an additional wrapper feature selection step to remove redundant features and reduce the computational time and memory. As a result, the feature subset used as the input vector for the SVM model was reduced from 379 to 256 attributes.

Each model was assigned a weight, which was proportional to the model classification accuracy across all classes. In addition, the capacities for classification and prediction on different classes may have been unequal. Therefore, the classifier with the highest prediction confidence was given greater weight for that class. Subsequently, the training process was conducted *via* 10-fold cross validation to find the optimal class weights for each classifier/predictor in the ensemble. Thereafter, the individual classifiers (SVM, RF, and XGB) were aggregated through weighted voting to obtain the final probability and prediction.

## The New Composite Feature is Significant for Improving the Sensitivity of the Method

We propose utilizing a logistic regression equation to create additional composite features, based on the fusion of two or more existing features. In contrast to sophisticated black-box classification models, regression is a powerful way to determine the unique relationships between a large number of features and a target class. In this work, we created a number of composite features and selected the two with the highest sensitivity, which

we refer to as comF and comF2. These features were merged into the feature vector as the input of the ensemble model.

The comF feature is defined as

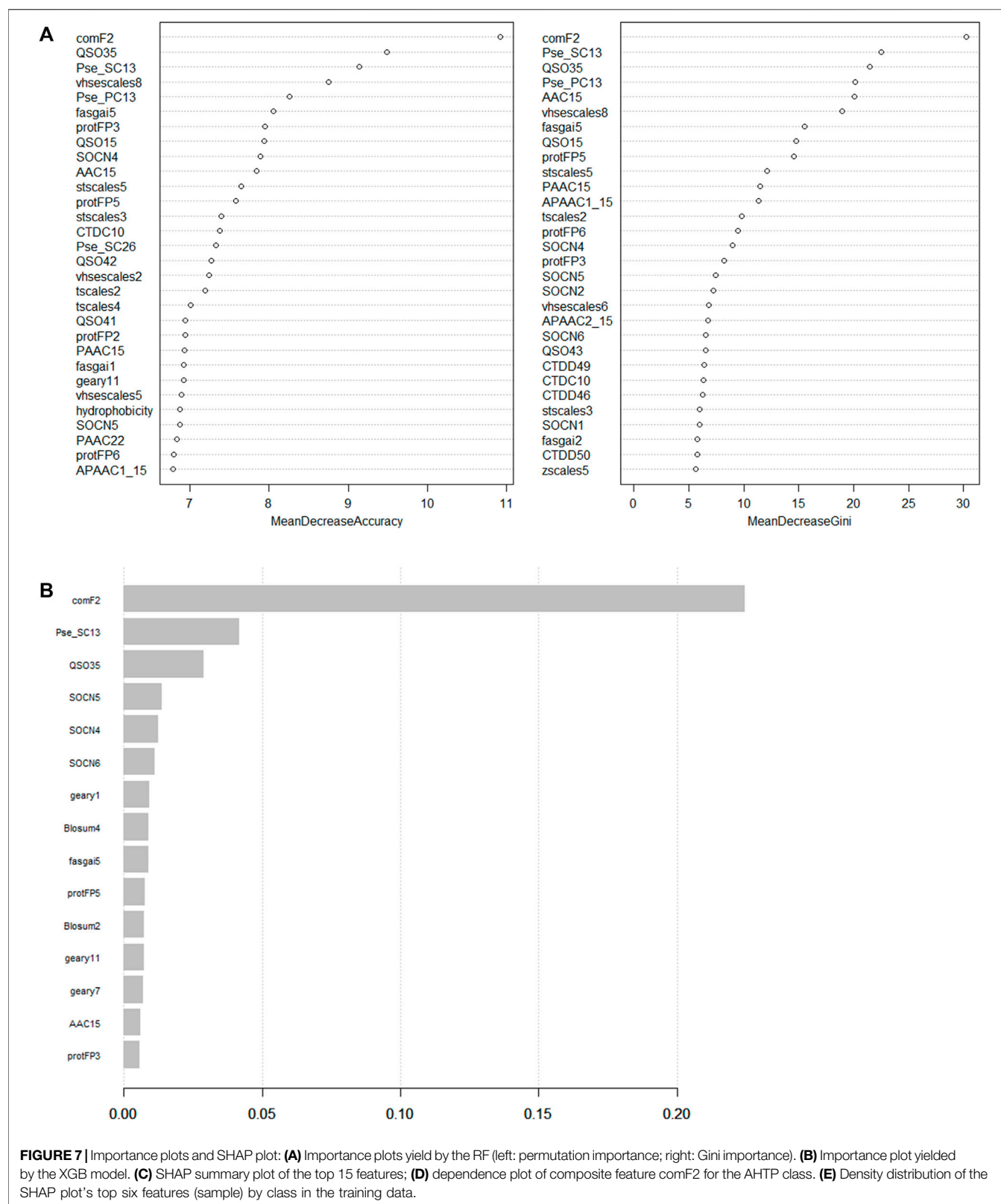
$$\begin{aligned} \text{comF} = & 0.8634 - 0.157\text{tscales4} - 0.154\text{CTDC19} - 0.135\text{protFP6} \\ & + 0.133\text{CTDC21} - 0.132\text{fasgai4} + 0.122\text{mswhimscore1} \\ & - 0.12\text{hydrophobicity} \end{aligned}$$

The comF2 feature is defined as

$$\begin{aligned} \text{comF2} = & 0.1786 + 0.1522\text{APAAC1\_15} - 2.2951\text{CTDC10} \\ & - 0.6069\text{CTDC19} - 0.0065\text{CTDD49} + 0.2176\text{QSO19} \\ & + 0.9747\text{fasgai4} + 0.3691\text{ProtFP3} \\ & + 2.0823\text{Pse\_PC13} \end{aligned}$$

where APAAC1\_15 denotes the amphiphilic PseAAC of amino acid R (the sequence-order coupling mode was used along a protein sequence *via* a hydrophobicity correlation function; the hydrophobic properties of amino acids were taken into account) and CTDC10 denotes the percentage of a particular amino acid in the polarizability group 1 (polarization between 0 and 1.08: amino acids G, A, S, D, and T) relative to protein length. CTDC19 is the percentage of a particular amino acid in solvent access group 1 (buried: amino acids A, L, F, C, G, I, V, and W) relative to the protein length. CTDD49 is the percentage of a particular amino acid in polarization group 1 (polarization between 0 and 1.08: amino acids G, A, S, D, and T) located in 75% of the residues of the protein chain. QSO19 is the quasi-sequence order of the normalized occurrence of amino acid Y, fasgai4 is a descriptor that reflects compositional characteristics, ProtFP3 is the scales-based descriptor derived from the amino acid properties of all AA indices (protein fingerprint 3), and Pse\_PC13 is the parallel correlation PseAAC of amino acid P.

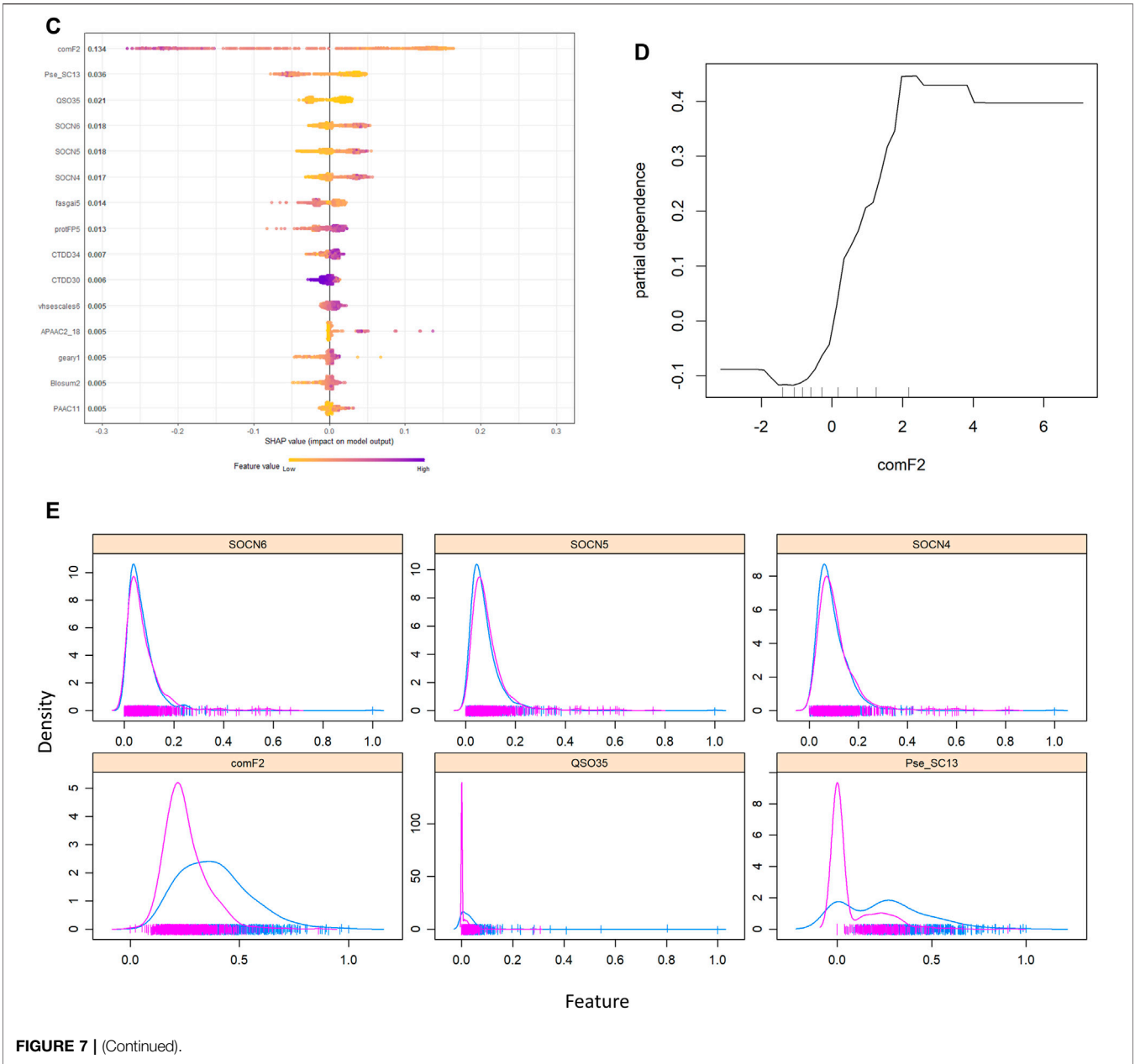
Interestingly, we discovered some intriguing aspects within the comF2 composite feature. Particular component properties of comF2, such as the distant locations of certain amino acids Y, R, and P, had beneficial impacts on the equation; this is consistent with the results of many research papers demonstrating that certain residues are dominant in the C-termini or N-termini of potent AHTPs. Hydrophobic residues with aliphatic side chains at the C-terminus promoted ACE inhibitory activity (Nimalaratne et al., 2015; Asoodeh et al., 2016; Jiang et al., 2021; Wang et al., 2021). Other studies have demonstrated that the positively charged lysine and arginine amino acids (K and R) contribute to the strong potency of ACE inhibitory peptides (Wei et al., 2019; Maky and Zendo, 2021). The richness of proline (P) and its number of occurrences in a sequence positively influenced the potency of ACE inhibition (Abachi et al., 2019; Festa et al., 2020; Pavlicevic et al., 2020). The presence of a polar amino acid at the C-terminus along with hydrophobic amino acids at the N-terminus may have contributed to the activity (Ryan et al., 2011; Udenigwe et al., 2012; de Castro and Sato, 2015). Moreover, the equation was adversely affected (according to the minus sign) by component



properties involving low-polarization amino acids (CTDC10 and CTDD49) and those with restricted solvent access (CTDC19; buried structure).

Because the RF and XGB have built-in feature importance analysis mechanisms, we discovered that the composite feature comF2 was the highest-ranking feature in both models based on





their importance plots (as shown in **Figures 7A,B**). It is well known that the value of a feature (as measured by information gain) varies depending on how frequently it is employed at the

leaf nodes. We also conducted SHAP (Shapley Additive exPlanations) analysis as a follow-up to our initial investigation. SHAP is a game-theoretic framework for

TABLE 3   Performance evaluation of the proposed method using benchmarking dataset.						
Method	ACC	Sn	Sp	MCC	AUC	
CNN + SVM (Rauf et al., 2021)	0.958	0.996	0.920	0.920	0.958	
mAHTPred (Manavalan et al., 2019)	0.848	0.821	0.874	0.697	0.903	
PAAP (Win et al., 2018)	0.791	0.865	0.780	0.585	NA	
AHTpin_AAC (Kumar et al., 2015)	0.785	0.777	0.793	0.567	NA	
AHTpin_ATC (Kumar et al., 2015)	0.785	0.783	0.787	0.573	NA	
Our ensemble	0.858	0.832	0.885	0.718	0.926	

TABLE 4   Performance evaluation of the proposed method using independence testing dataset.						
Method	ACC	Sn	Sp	MCC	AUC	
CNN + SVM (Rauf et al., 2021)	0.895	0.948	0.841	0.795	0.895	
mAHTPred (Manavalan et al., 2019)	0.883	0.894	0.873	0.767	0.951	
PAAP (Win et al., 2018)	NA	NA	NA	NA	NA	
AHTpin_AAC (Kumar et al., 2015)	0.800	0.821	0.780	0.601	0.852	
AHTpin_ATC (Kumar et al., 2015)	0.820	0.798	0.842	0.641	0.888	
Our ensemble	0.904	0.920	0.889	0.809	0.965	

**TABLE 5** | Performance evaluation of the proposed method using recently reported novel AHTPs.

Peptide sequence	IC <sub>50</sub>	Source	References	Correctly identify by our method (Yes/No)
YLYELR	9.37 $\mu$ M	Scorpion venom	Setayesh-Mehr et al. (2021)	Yes
AFPYYGHHLG	17.22 $\mu$ M	Scorpion venom	Setayesh-Mehr et al. (2021)	Yes
LVLPGE	13.5 $\mu$ M	Broccoli protein	Pei et al. (2021)	Yes
IPPAYTK	23.5 $\mu$ M	Broccoli protein	Dang et al. (2019)	Yes
LVLPGELAK	184 $\mu$ M	Broccoli protein	Dang et al. (2019)	Yes
TFQGPYPHGIQVER	3.4 $\mu$ M	Broccoli protein	Dang et al. (2019)	Yes
LIIPQH	120.1 $\mu$ M	Rice wine lees	He et al. (2021)	Yes
LIPPEH	60.49 $\mu$ M	Rice wine lees	He et al. (2021)	Yes
QTDEYGNPPR	210.03 $\mu$ M	Black tea	Lu et al. (2021)	Yes
AGFAGDDAPR	178.91 $\mu$ M	Black tea	Lu et al. (2021)	No
IDESLR	196.31 $\mu$ M	Black tea	Lu et al. (2021)	No
IQDKEGIPPDQQR	121.11 $\mu$ M	Black tea	Lu et al. (2021)	Yes
DAFGSFLYEYSE	-	Ricotta cheese	Pontonio et al. (2021)	No
RHPYFYAPELLYYANK	-	Ricotta cheese	Pontonio et al. (2021)	Yes
VERGRRITSV	6.82 $\mu$ M	Walnut Glutelin-1	Wang et al. (2021)	No
VIENPITPA	6.36 $\mu$ M	Walnut Glutelin-1	Wang et al. (2021)	Yes
LSGYGP	2.57 $\mu$ M	Tilapia	Chen et al. (2020)	Yes
LVPPIA	414.88 $\mu$ M	Radix Astragali	Wu et al. (2020)	Yes
SAGGYIW	0.002 $\mu$ M	Wheat gluten	Zhang et al. (2020)	Yes
APATPSFW	0.875 $\mu$ M	Wheat gluten	Zhang et al. (2020)	Yes
PPNNNPASPDFSSS	-	Soy protein	Daliri et al. (2019)	Yes
GPKALPII	-	Soy Protein	Daliri et al. (2019)	Yes
IIRCTGC	-	Soy protein	Daliri et al. (2019)	No
IGPGPFSSR	47.22 $\mu$ M	Mussel lamellidens	Ankhi et al. (2022)	Yes
FHAPWK	16.83 $\mu$ M	Cassia obtusifolia seeds	Shih et al. (2019)	Yes

explaining the output of any machine learning model. It correlates optimal credit allocation with local explanations by using classic Shapley values (Lundberg and Lee 2017). Since it averages the marginal contributions across all permutations, the performance of SHAP is notably more consistent than that of the information gain technique. The SHAP summary plot in **Figure 7C** is somewhat consistent with the information gain-based importance plot, which shows that comF2 was the most significant feature, followed by Pse\_SC13 and QSO35. According to the SHAP plot, the comF2 feature had an effect on the likelihoods for a larger model sample. Every dot in the SHAP plot represents a sample from the data. For each sample, the color of the corresponding dot refers to the value of the associated feature. The x-axis represents the feature's influence on the model's prediction. The high spread of comF2 indicates that it could capture and provide more useful information to the model to predict/identify the classes. Moreover, the partial dependence plot (PDP) of comF2 presents the impact of this feature on the predicted outcome, as shown in **Figure 7D**, allowing for a better understanding of the feature's interdependence with the target class (AHTP). According to the comF2 PDP illustrated in **Figure 7D**, the higher the value of the comF2 feature is, the higher the chance of the sample being classified into the AHTP class by the model (comF2 greater than two likelihoods of being in the AHTP class). Additionally, **Figure 7E** depicts the distribution of the top six features. A substantial distribution difference was observed between the AHTP and non-AHTP classes in the histogram of the comF2 feature. However, some overlap occurred between the two

classes' territories. The functionality of comF2 can be enhanced, resulting in an increase in prediction performance.

## Comparison With Existing Prediction Methods

To evaluate the performance of the proposed method, we used the benchmarking dataset and the independence testing dataset (as shown in **Tables 3, 4**, respectively), and then we compared and evaluated our ensemble method with the available prediction tools based on the results reported in (Manavalan et al., 2019; Rauf et al., 2021). As shown in **Table 3**, our technique achieved 85.8% accuracy on the benchmarking dataset or training dataset, outperforming most of the other methods. However, while the CNN + SVM technique surpassed our ensemble for the training dataset, our ensemble performed substantially better on the independent dataset.

When testing was performed on the independent data, accuracies of 90.4% were achieved, as shown in **Table 4**, and our method significantly outperformed the other methods.

## Performance Evaluation of Our Model With Novel Antihypertensive Peptides From Recent Studies

Novel AHTPs derived from food or natural sources are receiving significant attention. Therefore, an increasing number of food-derived or natural sources AHTPs have been researched and reported. To further assess the

generalization performance and robustness of the proposed method on new unseen data, we collected various experimental AHTPs from recent studies. These published AHTPs have been validated by *in vitro* or *in vivo* experimental assays. The results are summarized in **Table 5**. Note that these peptides did not overlap with our training data. Our ensemble model correctly classified these novel AHTPs from different sources with an accuracy of 80%.

## CONCLUSION

In this work, an ensemble model with a combination of XGB, RF, and SVM machine learning algorithms integrated by weighted voting was developed to achieve improved sensitivity and reduce the false positive rate in terms of predicting AHTPs. A new composite feature for AHTPs, comF2, was proposed and incorporated to improve the sensitivity of the developed method. The components of the comF2 feature were selected by a machine learning process based solely on a single training dataset (benchmarking dataset). However, we hypothesize that this new feature can be improved and adjusted to be more sensitive by combining novel knowledge or the information contained in the structure-function relationships (structure-activity relationships) of AHTPs reported in recent studies or by experts/biologists in the field. This knowledge can be expanded by incorporating more recent information or new significant features found in the future to further improve the proposed approach.

Currently, deep learning (DL) has become very prominence because of its ability to identify patterns in large volumes of raw data (scalability) and its ability in perform automatic feature extraction from raw data (feature encoding/learning). However, DL does not have an explicit feature engineering step because it has automated feature extraction. We are interested in feature engineering, extraction, and selection; therefore, we apply machine learning, including DL-related algorithms so called

neural nets. We exploited various features that are more explainable in terms of biological meaning, and we tried to capture an explainable relationship in the hybrid feature that may be an advantage in AHTP design in the future. We used the ensemble method, which is well-known to ensure generalization and to reduce the problem of overfitting of individual models. For precision of classification tools, both positive and negative dataset are important for model training. Availability of experimentally validated negative datasets, particularly sequences with similar amino acid compositions to those of AHTPs, will be beneficial for further improvement. Moreover, additional negative datasets containing other classes of peptides, for example, antioxidant, antimicrobial, and anticancer peptides and neuropeptides, which have been experimentally confirmed for their activities and do not show any antihypertensive activity will be more advantageous. To make this tool more useful, implementation as a webserver will be more accessible to bioactive peptide research communities.

## DATA AVAILABILITY STATEMENT

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

## AUTHOR CONTRIBUTIONS

Conceptualization: AH, CT, and SL. Formal analysis: SL. Methodology: SL. Writing—original draft: SL. Writing—review and editing: AH, CT, WW, and SL. Funding acquisition: WW.

## FUNDING

This research was funded by King Mongkut's University of Technology Thonburi, Thailand.

## REFERENCES

- Abachi, S., Bazinet, L., and Beaulieu, L. (2019). Antihypertensive and Angiotensin-I-Converting Enzyme (ACE)-Inhibitory Peptides from Fish as Potential Cardioprotective Compounds. *Mar. Drugs* 17 (11), 613. doi:10.3390/md17110613
- Aluko, R. E. (2015). Antihypertensive Peptides from Food Proteins. *Annu. Rev. Food Sci. Technol.* 6, 235–262. PMID: 25884281. doi:10.1146/annurev-food-022814-015520
- Ankhi, H., Madhushrita, D., K, D. T. D., Pubali, D., and Jana, C. (2022). Isolation of an Antihypertensive Bioactive Peptide from the Freshwater Mussel *Lamellidens Marginalis*. *Int. J. Food Nutr. Sci.* 11, 1–8. doi:10.54876/ijfans\_01-08
- Asodeh, A., Homayouni-Tabrizi, M., Shabestarian, H., Emtenani, S., and Emtenani, S. (2016). Biochemical Characterization of a Novel Antioxidant and Angiotensin I-Converting Enzyme Inhibitory Peptide from *Struthio camelus* Egg white Protein Hydrolysis. *J. Food Drug Anal.* 24 (2), 332–342. doi:10.1016/j.jfda.2015.11.010
- Balgir, P. P., and Sharma, M. (2017). Biopharmaceutical Potential of ACE-Inhibitory Peptides. *J. Proteomics Bioinform.* 10, 171–177. doi:10.4172/jpb.1000437
- Boman, H. G. (2003). Antibacterial Peptides: Basic Facts and Emerging Concepts. *J. Intern. Med.* 254 (3), 197–215. doi:10.1046/j.1365-2796.2003.01228.x
- Chen, J., Ryu, B., Zhang, Y., Liang, P., Li, C., Zhou, C., et al. (2020). Comparison of an angiotensin-I-converting Enzyme Inhibitory Peptide from tilapia (*Oreochromis niloticus*) with Captopril: Inhibition Kinetics, *In Vivo* Effect, Simulated Gastrointestinal Digestion and a Molecular Docking Study. *J. Sci. Food Agric.* 100 (1), 315–324. doi:10.1002/jsfa.10041
- Chen, W., Ding, H., Feng, P., Lin, H., and Chou, K.-C. (2016). iACP: a Sequence-Based Tool for Identifying Anticancer Peptides. *Oncotarget* 7 (13), 16895–16909. doi:10.18632/oncotarget.7815
- Chou, K.-C. (2000). Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Order Effect. *Biochem. Biophysical Res. Commun.* 278, 477–483. doi:10.1006/bbrc.2000.3815
- Chou, K.-C. (2011). Some Remarks on Protein Attribute Prediction and Pseudo Amino Acid Composition. *J. Theor. Biol.* 273, 236–247. doi:10.1016/j.jtbi.2010.12.024
- Chou, K.-C. (2005). Using Amphiphilic Pseudo Amino Acid Composition to Predict Enzyme Subfamily Classes. *Bioinformatics* 21, 10–19. doi:10.1093/bioinformatics/bth466
- Cruciani, G., Baroni, M., Carosati, E., Clementi, M., Valigi, R., and Clementi, S. (2004). Peptide Studies by Means of Principal Properties of Amino Acids

- Derived from MIF Descriptors. *J. Chemometrics* 18, 146–155. doi:10.1002/cem.856
- Daliri, E. B.-M., Ofosu, F. K., Chelliah, R., Kim, J.-H., Oh, D.-H., and Oh, D. H. (2019). Development of a Soy Protein Hydrolysate with an Antihypertensive Effect. *Ijms* 20 (6), 1496. doi:10.3390/ijms20061496
- Dang, Y., Zhou, T., Hao, L., Cao, J., Sun, Y., and Pan, D. (2019). *In Vitro* and *In Vivo* Studies on the Angiotensin-Converting Enzyme Inhibitory Activity Peptides Isolated from Broccoli Protein Hydrolysate. *J. Agric. Food Chem.* 67, 6757–6764. doi:10.1021/acs.jafc.9b01137
- Daskaya-Dikmen, C., Yucetepe, A., Karbancioglu-Guler, F., Daskaya, H., and Ozcelik, B. (2017). Angiotensin-I-Converting Enzyme (ACE)-Inhibitory Peptides from Plants. *Nutrients* 9 (4), 316. doi:10.3390/nu9040316
- de Castro, R. J. S., and Sato, H. H. (2015). Biologically Active Peptides: Processes for Their Generation, Purification and Identification and Applications as Natural Additives in the Food and Pharmaceutical Industries. *Food Res. Int.* 74, 185–198. doi:10.1016/j.foodres.2015.05.013
- De Leo, F., Panarese, S., Gallerani, R., and Ceci, L. (2009). Angiotensin Converting Enzyme (ACE) Inhibitory Peptides: Production and Implementation of Functional Food. *Cpd* 15 (31), 3622–3643. doi:10.2174/138161209789271834
- Dubchak, I., Muchnik, I., Holbrook, S. R., and Kim, S. H. (1995). Prediction of Protein Folding Class Using Global Description of Amino Acid Sequence. *Proc. Natl. Acad. Sci. U.S.A.* 92, 8700–8704. doi:10.1073/pnas.92.19.8700
- Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I., and Kim, S.-H. (1999). Recognition of a Protein Fold in the Context of the Scop Classification. *Proteins* 35, 401–407. doi:10.1002/(sici)1097-0134(19990601)35:4<401::aid-prot3>3.0.co;2-k
- Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J., and Serrano, L. (2004). Prediction of Sequence-dependent and Mutational Effects on the Aggregation of Peptides and Proteins. *Nat. Biotechnol.* 22, 1302–1306. doi:10.1038/nbt1012
- Festa, M., Sansone, C., Brunet, C., Crocetta, F., Di Paola, L., Lombardo, M., et al. (2020). Cardiovascular Active Peptides of Marine Origin with ACE Inhibitory Activities: Potential Role as Anti-hypertensive Drugs and in Prevention of SARS-CoV-2 Infection. *Ijms* 21 (21), 8364PMC7664667. doi:10.3390/ijms21218364.PMID:33171852
- Guruprasad, K., Reddy, B. V. B., and Pandit, M. W. (1990). Correlation between Stability of a Protein and its Dipeptide Composition: a Novel Approach for Predicting *In Vivo* Stability of a Protein from its Primary Sequence. *Protein Eng. Des. Sel* 4 (2), 155–161. doi:10.1093/protein/4.2.155
- He, R., Aluko, R. E., and Ju, X.-R. (2014). Evaluating Molecular Mechanism of Hypotensive Peptides Interactions with Renin and Angiotensin Converting Enzyme. *PLoS ONE* 9 (3), e91051. doi:10.1371/journal.pone.0091051
- He, R., Wang, Y., Yang, Y., Wang, Z., Ju, X., and Yuan, J. (2019). Rapeseed Protein-Derived ACE Inhibitory Peptides LY, RALP and GHS Show Antioxidant and Anti-inflammatory Effects on Spontaneously Hypertensive Rats. *J. Funct. Foods* 55, 211–219. doi:10.1016/j.jff.2019.02.031
- He, Z., Liu, G., Qiao, Z., Cao, Y., and Song, M. (2021). Novel Angiotensin-I Converting Enzyme Inhibitory Peptides Isolated from Rice Wine Lees: Purification, Characterization, and Structure-Activity Relationship. *Front. Nutr.* 8, 746113. doi:10.3389/fnut.2021.746113
- Ikai, A. (1980). Thermostability and Aliphatic index of Globular Proteins. *J. Biochem.* 88 (6), 1895–1898. doi:10.1093/oxfordjournals.jbchem.a133104
- Iwaniak, A., Minkiewicz, P., Darewicz, M., Sieniawski, K., and Starowicz, P. (2016). BIOPEP Database of Sensory Peptides and Amino Acids. *Food Res. Int.* 85, 155–161. doi:10.1016/j.foodres.2016.04.031
- Jakubczyk, A., Karaś, M., Rybczyńska-Tkaczyk, K., Zielińska, E., and Zieliński, D. (2020). Current Trends of Bioactive Peptides-New Sources and Therapeutic Effect. *Foods*, 9(7). PMID, 846PMC7404774. doi:10.3390/foods9070846. PMID:32610520
- Jiang, Q., Chen, Q., Zhang, T., Liu, M., Duan, S., and Sun, X. (2021). The Antihypertensive Effects and Potential Molecular Mechanism of Microalgal Angiotensin I-Converting Enzyme Inhibitor-like Peptides: A Mini Review. *Ijms* 22 (8), 4068. doi:10.3390/ijms22084068
- Kononenko, I. (1994). “Estimating Attributes: Analysis and Extensions of RELIEF,” in *Machine Learning: ECML-94. ECML 1994. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence)*. Editors F. Bergadano and L. De Raedt (Berlin, Heidelberg: Springer), 784. doi:10.1007/3-540-57868-4\_57
- Kumar, M., Thakur, V., and Raghava, G. P. (2008). COPid: Composition Based Protein Identification. *Silico Biol.* 8 (2), 121–128.
- Kumar, R., Chaudhary, K., Singh Chauhan, J., Nagpal, G., Kumar, R., Sharma, M., et al. (2015). An *In Silico* Platform for Predicting, Screening and Designing of Antihypertensive Peptides. *Sci. Rep.* 5, 12512. doi:10.1038/srep12512
- Kumar, R., Chaudhary, K., Sharma, M., Nagpal, G., Chauhan, J. S., Singh, S., et al. (2014). AHTPDB: A Comprehensive Platform for Analysis and Presentation of Antihypertensive Peptides. *Nucleic Acids Res.* 43, D956–D962. doi:10.1093/nar/gku1141
- Lee, S. Y., and Hur, S. J. (2019). Purification of Novel Angiotensin Converting Enzyme Inhibitory Peptides from Beef Myofibrillar Proteins and Analysis of Their Effect in Spontaneously Hypertensive Rat Model. *Biomed. Pharmacother.* 116, 109046. doi:10.1016/j.biopha.2019.109046
- Lertampaiporn, S., Vorapreeda, T., Hongsthong, A., and Thammarongtham, C. (2021). Ensemble-AMPPred: Robust AMP Prediction and Recognition Using the Ensemble Learning Method with a New Hybrid Feature for Differentiating AMPs. *Genes* 12 (2), 137. doi:10.3390/genes12020137
- Li-Chan, E. C. (2015). Bioactive Peptides and Protein Hydrolysates: Research Trends and Challenges for Application as Nutraceuticals and Functional Food Ingredients. *Curr. Opin. Food Sci.* 1, 28–37. doi:10.1016/j.cofs.2014.09.005
- Liang, G., and Li, Z. (2007). Factor Analysis Scale of Generalized Amino Acid Information as the Source of a New Set of Descriptors for Elucidating the Structure and Activity Relationships of Cationic Antimicrobial Peptides. *QSAR Comb. Sci.* 26 (6), 754–763. doi:10.1002/qsar.200630145
- Lu, Y., Wang, Y., Huang, D., Bian, Z., Lu, P., Fan, D., et al. (2021). Inhibitory Mechanism of Angiotensin-Converting Enzyme Inhibitory Peptides from Black tea. *J. Zhejiang Univ. Sci. B* 22 (7), 575PMC8284085–589. PMID, 10.1631/jzus.B2000520.PMID:34269010
- Lundberg, S. M., and Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* 30. arXiv:1705.07874 [cs.AI].
- Maky, M. A., and Zendo, T. (2021). Generation and Characterization of Novel Bioactive Peptides from Fish and Beef Hydrolysates. *Appl. Sci.* 11, 10452. doi:10.3390/app112110452
- Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019). mAHTPred: A Sequence-Based Meta-Predictor for Improving the Prediction of Anti-hypertensive Peptides Using Effective Feature Representation. *Bioinformatics* 35, 2757–2765. doi:10.1093/bioinformatics/bty1047
- Martínez-Maqueda, D., Miralles, B., Recio, I., and Hernández-Ledesma, B. (2012). Antihypertensive Peptides from Food Proteins: a Review. *Food Funct.* 3 (4), 350–361. doi:10.1039/c2fo10192k
- Mei, H., Liao, Z. H., Zhou, Y., and Li, S. Z. (2005). A New Set of Amino Acid Descriptors and its Application in Peptide QSARs. *Biopolymers* 80 (6), 775–786. doi:10.1002/bip.20296
- Mills, K. T., Stefanescu, A., and He, J. (2020). The Global Epidemiology of Hypertension. *Nat. Rev. Nephrol.* 16 (4), 223–237. doi:10.1038/s41581-019-0244-2
- Minkiewicz, P., Dziuba, J., Iwaniak, A., Dziuba, M., and Darewicz, M. (2008). BIOPEP Database and Other Programs for Processing Bioactive Peptide Sequences. *J. AOAC Int.* 91 (4), 965–980. doi:10.1093/jaoac/91.4.965
- Nguyen, Q., Dominguez, J., Nguyen, L., and Gullapalli, N. (2010). Hypertension Management: An Update. *Am. Health Drug Benefits.* 3, 47–56.
- Nimalaratne, C., Bandara, N., and Wu, J. (2015). Purification and Characterization of Antioxidant Peptides from Enzymatically Hydrolyzed Chicken Egg white. *Food Chem.* 188, 467–472. doi:10.1016/j.foodchem.2015.05.014
- Norris, R., and J., R. (2013). “Antihypertensive Peptides from Food Proteins,” in *Bioactive Food Peptides in Health and Disease*. Editors B. Hernandez-Ledesma and C. Hsieh (IntechOpen). doi:10.5772/51710
- Osorio, D., Rondón-Villarreal, P., and Torres, R. (2015). Peptides: A Package for Data Mining of Antimicrobial Peptides. *R. J.* 7 (1), 4–14. doi:10.32614/rj-2015-001
- Pavlicevic, M., Maestri, E., and Marmiroli, M. (2020). Marine Bioactive Peptides-An Overview of Generation, Structure and Application with a Focus on Food Sources. *Mar. Drugs* 18 (8), 424. doi:10.3390/md18080424
- Pei, J., Hua, Y., Zhou, T., Gao, X., Dang, Y., and Wang, Y. (2021). Transport, *In Vivo* Antihypertensive Effect, and Pharmacokinetics of an Angiotensin-Converting Enzyme (ACE) Inhibitory Peptide LVLPG. *J. Agric. Food Chem.* 69 (7), 2149–2156. doi:10.1021/acs.jafc.0c07048



- Pontonio, E., Montemurro, M., De Gennaro, G. V., Miceli, V., and Rizzello, C. G. (2021). Antihypertensive Peptides from Ultrafiltration and Fermentation of the Ricotta Cheese Exhausted Whey: Design and Characterization of a Functional Ricotta Cheese. *Foods* 10 (11), 2573. doi:10.3390/foods10112573
- Pujiastuti, D. Y., Ghoyatul Amin, M. N., Alamsjah, M. A., and Hsu, J.-L. (2019). Marine Organisms as Potential Sources of Bioactive Peptides that Inhibit the Activity of Angiotensin I-Converting Enzyme: A Review. *Molecules* 24 (14), 2541. doi:10.3390/molecules24142541
- Rauf, A., Kiran, A., Hassan, M. T., Mahmood, S., Mustafa, G., and Jeon, M. (2021). Boosted Prediction of Antihypertensive Peptides Using Deep Learning. *Appl. Sci.* 11 (5), 2316. doi:10.3390/app11052316
- Ryan, J. T., Ross, R. P., Bolton, D., Fitzgerald, G. F., and Stanton, C. (2011). Bioactive Peptides from Muscle Sources: Meat and Fish. *Nutrients* 3 (9), 765–791. doi:10.3390/nu3090765
- Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., and Wold, S. (1998). New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *J. Med. Chem.* 41, 2481–2491. doi:10.1021/jm9700575
- Setayesh-Mehr, Z., Ghasemi, L. V., and Asoodeh, A. (2021). Evaluation of the *In Vivo* Antihypertensive Effect and Antioxidant Activity of HL-7 and HL-10 Peptide in Mice. *Mol. Biol. Rep.* 48 (7), 5571–5578. doi:10.1007/s11033-021-06576-7
- Sharma, A., Kapoor, P., Gautam, A., Chaudhary, K., Kumar, R., Chauhan, J. S., et al. (2013). Computational Approach for Designing Tumor Homing Peptides. *Sci. Rep.* 3, 1607. doi:10.1038/srep01607
- Shih, Y.-H., Chen, F.-A., Wang, L.-F., and Hsu, J.-L. (2019). Discovery and Study of Novel Antihypertensive Peptides Derived from Cassia Obtusifolia Seeds. *J. Agric. Food Chem.* 67 (28), 7810–7820. doi:10.1021/acs.jafc.9b01922
- Thomsen, M. C. F., and Nielsen, M. (2012). Seq2Logo: a Method for Construction and Visualization of Amino Acid Binding Motifs and Sequence Profiles Including Sequence Weighting, Pseudo Counts and Two-Sided Representation of Amino Acid Enrichment and Depletion. *Nucleic Acids Res.* 40, W281–W287. doi:10.1093/nar/gks469
- Tian, F., Zhou, P., and Li, Z. (2007). T-scale as a Novel Vector of Topological Descriptors for Amino Acids and its Application in QSARs of Peptides. *J. Mol. Struct.* 830, 106–115. doi:10.1016/j.molstruc.2006.07.004
- Tološi, L., and Lengauer, T. (2011). Classification with Correlated Features: Unreliability of Feature Ranking and Solutions. *Bioinformatics* 27, 1986–1994. doi:10.1093/bioinformatics/btr300
- Udenigwe, C. C., Li, H., and Aluko, R. E. (2012). Quantitative Structure-Activity Relationship Modeling of Renin-Inhibiting Dipeptides. *Amino Acids* 42, 1379–1386. doi:10.1007/s00726-011-0833-2
- Udenigwe, C. C., and Mohan, A. (2014). Mechanisms of Food Protein-Derived Antihypertensive Peptides Other Than ACE Inhibition. *J. Funct. Foods* 8, 45–52. doi:10.1016/j.jff.2014.03.002
- Usmani, S. S., Bhalla, S., and Raghava, G. P. S. (2018). Prediction of Antitubercular Peptides from Sequence Information Using Ensemble Classifier and Hybrid Features. *Front. Pharmacol.* 9, 954. doi:10.3389/fphar.2018.00954
- van Westen, G. J., Swier, R. F., Wegner, J. K., IJzerman, A. P., van Vlijmen, H. W., and Bender, A. (2013). Benchmarking of Protein Descriptor Sets in Proteochemometric Modeling (Part 1): Comparative Study of 13 Amino Acid Descriptor Sets. *J. Cheminform* 5 (1), 41. doi:10.1186/1758-2946-5-41
- Wang, J., Wang, G., Zhang, Y., Zhang, R., and Zhang, Y. (2021). Novel Angiotensin-Converting Enzyme Inhibitory Peptides Identified from Walnut Glutelin-1 Hydrolysates: Molecular Interaction, Stability, and Antihypertensive Effects. *Nutrients* 14 (1), 151. doi:10.3390/nu14010151
- Wei, D., Fan, W., and Xu, Y. (2019). *In Vitro* Production and Identification of Angiotensin Converting Enzyme (ACE) Inhibitory Peptides Derived from Distilled Spent Grain Prolamin Isolate. *Foods* 8 (9), 390. doi:10.3390/foods8090390
- Win, T. S., Schaduagrat, N., Prachayasittikul, V., Nantasenamat, C., and Shoombutong, W. (2018). PAAP: A Web Server for Predicting Antihypertensive Activity of Peptides. *Future Med. Chem.* 10, 1749–1767. doi:10.4155/fmc-2017-0300
- Wu, C. H., Mohammadmoradi, S., Chen, J. Z., Sawada, H., Daugherty, A., and Lu, H. S. (2018). Renin-Angiotensin System and Cardiovascular Functions. *Arterioscler Thromb. Vasc. Biol.* 38 (7), e108–e116. doi:10.1161/ATVBAHA.118.311282
- Wu, J.-S., Li, J.-M., Lo, H.-Y., Hsiang, C.-Y., and Ho, T.-Y. (2020). Anti-hypertensive and Angiotensin-Converting Enzyme Inhibitory Effects of Radix Astragali and its Bioactive Peptide AM-1. *J. Ethnopharmacology* 254, 254112724. doi:10.1016/j.jep.2020.112724
- Xiao, N., Cao, D.-S., Zhu, M.-F., and Xu, Q.-S. (2015). Protr/ProtrWeb: R Package and Web Server for Generating Various Numerical Representation Schemes of Protein Sequences. *Bioinformatics* 31, 1857–1859. doi:10.1093/bioinformatics/btv042
- Yang, L., Shu, M., Ma, K., Mei, H., Jiang, Y., and Li, Z. (2010). ST-scale as a Novel Amino Acid Descriptor and its Application in QSAM of Peptides and Analogues. *Amino acids* 38 (3), 805–816. doi:10.1007/s00726-009-0287-y
- Yi, Y., Lv, Y., Zhang, L., Yang, J., and Shi, Q. (2018). High Throughput Identification of Antihypertensive Peptides from Fish Proteome Datasets. *Mar. Drugs* 16, 365. doi:10.3390/md16100365
- Zaky, A. A., Simal-Gandara, J., Eun, J.-B., Shim, J.-H., and Abd El-Aty, A. M. (2022). Bioactivities, Applications, Safety, and Health Benefits of Bioactive Peptides from Food and By-Products: A Review. *Front. Nutr.* 8, 815640. doi:10.3389/fnut.2021.815640
- Zaliani, A., and Gancia, E. (1999). MS-WHIM Scores for Amino Acids: a New 3D-Description for Peptide QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* 39 (3), 525–533. doi:10.1021/ci980211b
- Zhang, P., Chang, C., Liu, H., Li, B., Yan, Q., and Jiang, Z. (2020). Identification of Novel Angiotensin I-Converting Enzyme (ACE) Inhibitory Peptides from Wheat Gluten Hydrolysate by the Protease of *Pseudomonas aeruginosa*. *J. Funct. Foods* 65, 103751. doi:10.1016/j.jff.2019.103751
- Zhou, B., Perel, P., Mensah, G. A., and Ezzati, M. (2021). Global Epidemiology, Health burden and Effective Interventions for Elevated Blood Pressure and Hypertension. *Nat. Rev. Cardiol.* 18 (11), 785–802. doi:10.1038/s41569-021-00559-8
- Zhu, J., Li, J., Guo, Y., Quaisie, J., Hong, C., and Ma, H. (2021). Antihypertensive and Immunomodulatory Effects of Defatted Corn Germ Hydrolysates: An *In Vivo* Study. *Front. Nutr.* 8, 679583. doi:10.3389/fnut.2021.679583

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

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

Copyright © 2022 Lertampaiporn, Hongsthong, Wattanapornprom and Thammarongtham. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# Inter-Residue Distance Prediction From Duet Deep Learning Models

Huiling Zhang<sup>1,2</sup>, Ying Huang<sup>1,2</sup>, Zhendong Bei<sup>1,2</sup>, Zhen Ju<sup>1,2</sup>, Jintao Meng<sup>1,2</sup>, Min Hao<sup>3</sup>, Jingjing Zhang<sup>1,2</sup>, Haiping Zhang<sup>2</sup> and Wenhui Xi<sup>1,2\*</sup>

<sup>1</sup>Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, <sup>2</sup>University of Chinese Academy of Sciences, Beijing, China, <sup>3</sup>College of Electronic and Information Engineering, Southwest University, Chongqing, China

## OPEN ACCESS

### Edited by:

Ruiquan Ge,  
Hangzhou Dianzi University, China

### Reviewed by:

Leyi Wei,  
Shandong University, China  
Jun Wang,  
Nanjing University, China  
Duolin Wang,  
University of Missouri, United States

### \*Correspondence:

Wenhui Xi  
wh.xi@siat.ac.cn

### Specialty section:

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

**Received:** 01 March 2022

**Accepted:** 30 March 2022

**Published:** 16 May 2022

### Citation:

Zhang H, Huang Y, Bei Z, Ju Z,  
Meng J, Hao M, Zhang J, Zhang H and  
Xi W (2022) Inter-Residue Distance  
Prediction From Duet Deep  
Learning Models.  
Front. Genet. 13:887491.  
doi: 10.3389/fgene.2022.887491

Residue distance prediction from the sequence is critical for many biological applications such as protein structure reconstruction, protein-protein interaction prediction, and protein design. However, prediction of fine-grained distances between residues with long sequence separations still remains challenging. In this study, we propose DuetDis, a method based on duet feature sets and deep residual network with squeeze-and-excitation (SE), for protein inter-residue distance prediction. DuetDis embraces the ability to learn and fuse features directly or indirectly extracted from the whole-genome/metagenomic databases and, therefore, minimize the information loss through ensembling models trained on different feature sets. We evaluate DuetDis and 11 widely used peer methods on a large-scale test set (610 proteins chains). The experimental results suggest that 1) prediction results from different feature sets show obvious differences; 2) ensembling different feature sets can improve the prediction performance; 3) high-quality multiple sequence alignment (MSA) used for both training and testing can greatly improve the prediction performance; and 4) DuetDis is more accurate than peer methods for the overall prediction, more reliable in terms of model prediction score, and more robust against shallow multiple sequence alignment (MSA).

**Keywords:** residue distance prediction, protein structure reconstruction, deep learning, residual network, multiple sequence alignment

## INTRODUCTION

Knowing the structure of a protein helps to understand the role of the protein, reveals how the protein performs its biological function, and also, sets the foundation for the protein's interaction with other molecules. Therefore, the knowledge of a protein's structure is very important for biology as well as for medicine and pharmacy. Since Anfinsen suggested that the advanced spatial structure of a protein is determined by its amino acid sequence (Anfinsen, 1973), it has been a "holy grail" for the computational biology community to develop an algorithm that can accurately predict a protein's structure from its amino acid sequence. Sequence-based residue contact/distance prediction plays a crucial role in protein structure reconstruction.

Residue-residue contacts refer to the residue pairs that are close within a specific distance threshold in the three-dimensional protein structure. The contact map of a protein tells the constraints between residues in a binary form. Unlike the contact map, the distance map of a protein contains fine-grained information and, thus, provides more physical constraints of a protein structure. Protein contact/distance maps are 2D representations of the 3D protein structure and are being considered as one of the most important components in modern protein structure prediction packages. The application of predicted contacts/distances has been extended to intrinsic disorder

region recognition (Schlessinger et al., 2007; Shimomura et al., 2019), protein–protein interaction prediction (Vangone and Bonvin, 2015; Du et al., 2016; Cong et al., 2019), protein design (Anishchenko et al., 2021), etc.

Contact prediction methods in the early stage are mainly based on mutual information (MI) (Pollock and Taylor, 1997; Dunn et al., 2007; Lee and Kim, 2009), integer linear programming (ILP) techniques (McAllister and Floudas, 2008; Rajgaria et al., 2009; Rajgaria et al., 2010; Wei and Floudas, 2011), traditional machine learning (ML) algorithms (Cheng and Baldi, 2007; Wu and Zhang, 2008; Tegge et al., 2009), or techniques combining ILP with ML (Wang and Xu, 2013; Zhang et al., 2016). These methods are generally considered as local strategies since a residue pair is treated statistically independent of others (Zhang et al., 2020). Breakthroughs were achieved by capturing the correlated pattern of coevolved residues by global statistical inference methods such as direct coupling analysis (DCA) (Weigt et al., 2009) and sparse inverse covariance estimation (PSICOV) (Jones et al., 2012). Methods developed based on the ideas of DCA include EVfold (mfDCA) (Morcos et al., 2011), plmDCA (Ekeberg et al., 2013), GREMLIN (Kamisetty et al., 2013), CCMpred (Seemayer et al., 2014), gDCA (Baldassi et al., 2014), and Freecontact (Kaján et al., 2014). These methods emphasize the importance of distinguishing between directly and indirectly correlated residues. Consensus-predictors like PconsC (Skwark et al., 2013), MetaPSICOV (Jones et al., 2014), and NeBcon (He et al., 2017) combine the output of different DCA-based or ML-based contact predictors to create consensus predictions. In recent years, the introduction of deep learning (DL) techniques has made tremendous progress for residue contact prediction. The DL-based contact map prediction algorithms are mainly based on convolutional neural networks (CNN) (such as DeepCov (Jones and Kandathil, 2018), DeepContact (Liu et al., 2018), and DNCON2 (Adhikari et al., 2018)), Unet [such as PconsC4 (Michel et al., 2019)], residual networks (ResNet) [such as DeepConPred2 (Ding et al., 2018), ResPRE (Li et al., 2019), MapPred (Wu et al., 2020) and TripletRes (Li et al., 2021)], ResNet combined with long short-term memory (LSTM) [such as SPOT-Contact (Hanson et al., 2018)] and transformers [such as ESM (Malinin and Gales, 2021) and SPOT-Contact-LM (Singh et al., 2022)]. COMTOP (Reza et al., 2021) uses the mixed ILP technique to combine different contact predictors (including several DL predictors) to further improve the prediction performance.

Although the predicted contacts have been successfully applied to the protein structure prediction packages (Marks et al., 2012; Michel et al., 2014; Adhikari et al., 2015; Gao et al., 2019), contact maps are still insufficient for accurate structure prediction. The reason is twofold. Most contact prediction methods use a cutoff of 8 Å between C $\beta$ –C $\beta$  atoms to determine whether two residues are in contact or not, resulting a contact/non-contact ratio of less than 0.1 for globular proteins and a ratio of around 0.02 for alpha-helical transmembrane proteins (Zhang et al., 2016). The definition of contacts means that the native distance information is insufficiently being distinguished. Furthermore, contact-assisted conformation

sampling may be misguided by several wrongly predicted contacts and needs a long time to generate good conformations for large proteins (Xu, 2019). In this context, inter-residue distance maps are more informative than residue–residue contact maps since distances are fine-grained or real numbers, while contacts are binary values.

The methods for inter-residue distance prediction can be roughly categorized into two groups, those based on multiclass classification with discrete values and those based on regression with continuous values. Early distance maps are mainly predicted from homologous proteins (Aszódi and Taylor, 1996) or from traditional machine learning techniques (Walsh et al., 2009; Zhao and Xu, 2012; Kukic et al., 2014). The introduction of deep learning technology has injected new life into distance prediction. Wang et al. (2017) pioneered the study of introducing residual network to multiclass distance prediction. The success of this approach can be partially attributed to the ability of deep learning to simultaneously consider the global set of pair-wise interactions instead of considering only one interaction at a time, thereby leading to more accurate discrimination between direct and indirect contacts. TripletRes (Li et al., 2021), which uses a similar deep learning architecture but with a unique set of features that include multiple coevolutionary coupling matrices directly deduced from deep multiple sequence alignment (MSA) without post-processing. GANProDist (Ding and Gong, 2020) predicts real value distance as a regression problem by generative adversarial network. PDNET (Adhikari, 2020), DeepDist (Wu et al., 2021), SDP (Rahman et al., 2022), and Li et al. (2021) (Li and Xu, 2021) predict both real-valued and binned distances from residual networks. DL-based distance prediction has recently demonstrated unprecedented ability to assist protein structure reconstruction such as DMPFold (Greener et al., 2019), RaptorX (Xu, 2019), trRosetta (Yang et al., 2020), and AlphaFold (Senior et al., 2020). However, further progress needs more accurate inter-residue distance prediction since the quality of a predicted protein structure highly depends on the accuracy of the distance prediction.

Shimomura et al. (2019) introduced a technique for predicting structurally disordered regions in proteins through average distance maps (AMD) based on statistics of average distances between residues. AMD first divides the residue pairs into different ranges according to their sequence separations, and calculates the distances of residue pairs within each range. AMD contact density maps were plotted against distance thresholds in different ranges. AMD technology detects the boundaries of structurally compact regions and finally predicts structurally disordered regions by calculating differences in density maps. The accuracy of AMD technology is comparable to the leading methods in the CASP competition such as PrDOS, DISOPRED, and Biome. Protein domains are subunits that can fold and function independently. Therefore, correct domain boundary assignment is a critical step to achieve accurate protein structure and function analysis. Zheng et al. (2020) proposed FUPred to detect protein domains based on contact maps predicted by deep learning. The core idea of this method is to retrieve domain boundary locations by maximizing the number of intra-domain contacts while minimizing the number of inter-

domain contacts from the contact map. FUPred was tested on a large-scale dataset consisting of 2,549 proteins and achieved a Matthews correlation coefficient (MCC) of 0.799 for single domain and multi-domain classification, which is 19.1% higher than the best machine learning-based method. For proteins with discontinuous domains, FUPred domain boundary detection and normalized domain overlap scores were 0.788 and 0.521, which were 17.3% and 23.8% higher than the best peer method. The results demonstrate that residue contact prediction provides a new way to accurately detect domains, especially discontinuous multi-domains. Cong et al. (2019) first compared the contact prediction methods based on mutual information, evolutionary coupling analysis, and deep learning in the prediction of residue contacts between protein complex chains and found that although the deep learning methods are outstanding for monomer contact prediction, they fail to outperform methods based on mutual information and evolutionary coupling analysis in inter-chain contact prediction. By identifying coevolving residue pairs between protein chains based on mutual information and evolutionary coupling analysis methods, 1,618 protein interactions (682 of which were unexpected) in *Escherichia coli*, and 911 protein interactions in *M. tuberculosis* (most of which were not identified in previous studies) were detected. The expected false positive rate for this study is between 10% and 20%, and the predicted interactions and networks provide a good starting point for further research. Anishchenko et al. (2021) investigated whether the residue distance information captured by deep neural networks is rich enough to generate new folded proteins. The study generated random amino acid sequences that were completely unrelated to the sequences of the native proteins used in the trRosetta training model, and fed them into the trRosetta structure prediction network to predict the starting residue distance map. Monte Carlo sampling is then performed in the amino acid sequence space to optimize the contrast between the network-predicted distribution of inter-residue distances and the background distribution averaged across all proteins. Optimization from different random starting points yields novel proteins spanning a broad range of sequences and predicted structures. Synthetic genes encoding 129 of the ‘network-hallucinated’ sequences were obtained, and the proteins were expressed and purified in *E. coli*; 27 of the proteins yielded monodisperse species with circular dichroism spectra consistent with the hallucinated structures. Three of the three-dimensional structures of the hallucinated proteins were determined by experiments, and these closely matched the hallucinated models. We can see that residue distance-assisted protein structure prediction methods can be inverted to *de novo* protein design.

In this study, we develop a method based on deep residual convolutional neural network, named DuetDis, to predict the full-length multiclass distance map from a sequence. DuetDis uses a modified ResNet module to build the network, and adopts two sets of complementary feature sets to further improve the prediction accuracy. The results by DuetDis suggest that prediction results from different feature sets show obvious differences and ensembles of different feature sets can improve

the prediction performance. DuetDis is also evaluated together with 11 widely used contact/distance prediction methods, and the results show that DuetDis is more accurate for the overall prediction, more reliable in terms of model prediction score, and more robust against shallow MSA. DuetDis is available at <http://hpcc.siat.ac.cn/hlzhang/DuetDis/>.

## MATERIALS AND METHODS

### Datasets

The test set is obtained from our previous work, containing 610 highly non-redundant protein chains (Zhang et al., 2021). The training set is obtained through culling from the whole PDB with the following criteria: 1) with maximum sequence identity of 30% against each chain in the training set and test set; 3) with structure resolutions better than 2.5 Å; 4) released before 1 May 2018 (before the beginning of CASP13). Finally, we get a non-redundant training set with 13,069 protein chains.

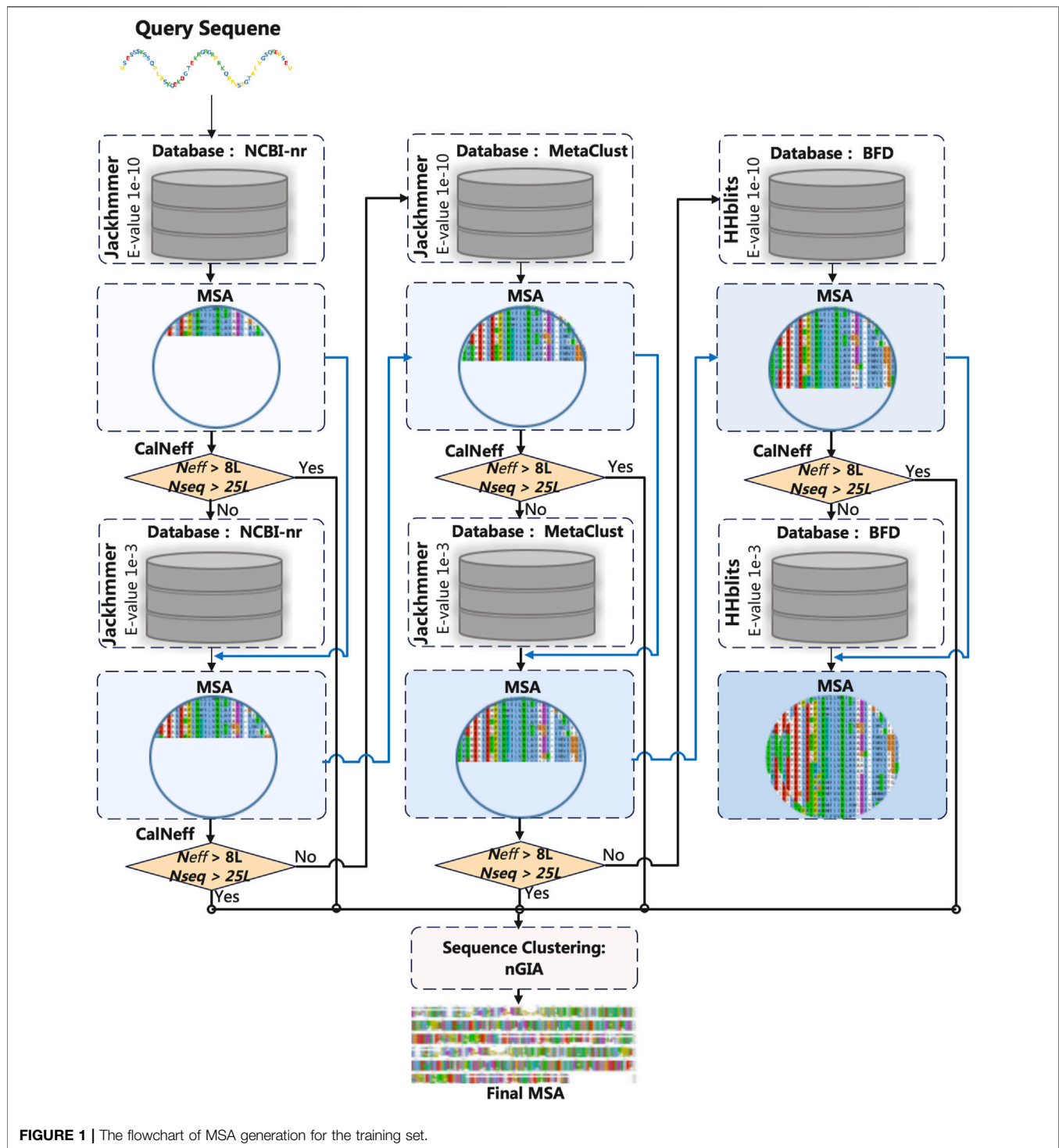
### Definition of Contact and Distance

In this study, the definition of contacts is directly taken from the CASP experiments. A pair of residues in the experimental structure is considered to be in contact if the distance between their C $\beta$  atoms (Ca for Gly) is less than or equal to 8 Å. For direct comparison, the multiclass distance definition is taken directly from trRosetta (Adhikari, 2020). The C $\beta$ –C $\beta$  distance of every pair of residues in a target protein is treated as a vector of probabilities. The distance range (2–20 Å) is binned into 36 equally spaced segments, 0.5 Å each, and one bin indicating that residues are not in contact, generating a distance vector of 37 bins for each residue pair.

Depending on the separation of two residues along the sequence (*seq\_sep*), the contacts are classified into four classes: all-range (*seq\_sep*  $\geq 6$ ), short-range ( $6 \leq \text{seq\_sep} < 12$ ), medium-range ( $12 \leq \text{seq\_sep} < 24$ ), and long-range (*seq\_sep*  $> 24$ ).

### Multiple Sequence Alignment Generation for Training and Test

Generating high-quality MSA is the first step for protein structure prediction based on the fact that interacting residue pairs are under evolutionary pressure to maintain the structure. The MSA used for model training is obtained as indicated in **Figure 1**. The target sequence in the training set is searched against NCBI-nr (Jackhmmer), MetaClust (Jackhmmer), and BFD (HHblits) respectively, with *E*-values of  $1e-10$  and  $1e-3$ . The search will stop if the target MSA has  $N_{seq} > 25 \times L$  (*L* is the sequence length) and  $N_{eff} > 8 \times L$ , where  $N_{seq}$  is the number of sequences (with sequence coverage  $> 50\%$ ) and  $N_{eff}$  [defined in (Zhang et al., 2021)] is the number of effective sequences in the MSA. After the search, the final MSA is obtained through sequence clustering (with sequence identity of 95%) using our in-house software nGIA (Ju et al., 2021).



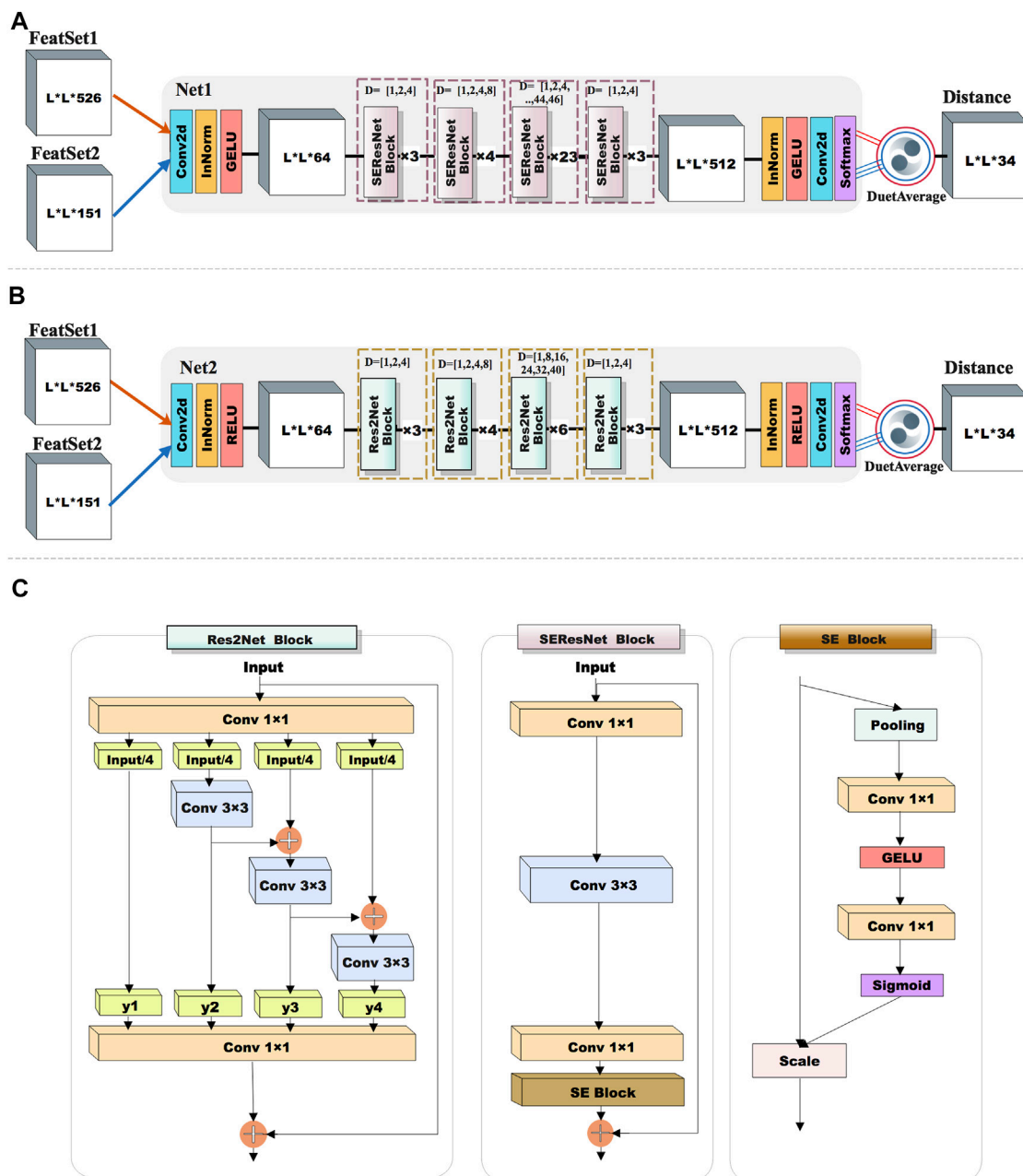
The MSA used for testing is obtained through searching JackHMMER (Johnson et al., 2010) against the NCBI-nr database with iteration = 3 and  $E$ -value = 0.0001.

## Input Features

We used two subsets of features as the inputs for the deep residual network of DuetDis. The first feature set contains 526 feature

channels: one-hot-encoder of the target sequence (1D features,  $20 \times 2$  channels); position-specific frequency matrix (1D features,  $21 \times 2$  channels, considering gap) and positional entropy (Yang et al., 2020) (1D features,  $1 \times 2$  channels); and coupling features (Yang et al., 2020) (2D features, 441 channels) derived from the inverse of the shrunk covariance matrix of MSA. The second feature set contains 151 feature channels: one-hot-encoder of the





**FIGURE 2 |** The network architecture used in this work. **(A)** The network used by DuetDis; **(B)** the reference network; **(C)** basic modules used in the networks; dilated convolution.

target sequence (1D features, 20\*2 channels), position-specific scoring matrix (Altschul et al., 1997) (1D features; 20\*2 channels; not considering gap), HMM profile (Remmert et al., 2012) (1D features, 30\*2 channels), secondary structure from SPOT-1D (Hanson et al., 2019) (1D features, 3\*2 channels), solvent accessible surface area from SPOT-1D (Hanson et al., 2019) (1D features, 1\*2 channels), CCMpred score (Seemayer et al., 2014) (2D features, 1 channel), mutual information (Zhang et al., 2022) (2D feature, 1 channel), and statistical pair-wise contact potential (Betancourt and Thirumalai, 1999) (2D feature, 1

channel). The first feature set, indicated as FeatSet1, is mainly composed of 2D direct coupling features (441 out of 526 total features) from the MSA, while the second feature set, indicated as FeatSet2, is mainly composed of 1D sequence-based features (148 out of 151 total features). Most of the features except the one-hot-encoder features in FeatSet1 and FeatSet2 are different, so the prediction results from the two feature sets can be complementary in a duet way (as indicated in the results).

Both FeatSet1 and FeatSet2 are widely used by previous works (Hanson et al., 2018; Yang et al., 2020; Jain et al., 2021; Su et al.,



**TABLE 1 |** The strategies used for the training of sub-models (N1\_M1/N1\_M2/N1\_M3/N1\_M4/N1\_M5 are used for DuetDis).

Sub-models	Network	Feature set	MSA	MSA shuffle
N1_M1	Net1	FeatSet1	MSA_All	Yes
N1_M2	Net1	FeatSet1	MSA_Top	No
N1_M3	Net1	FeatSet2	MSA_Top	No
N1_M4	Net1	FeatSet2	MSA_1	No
N1_M5	Net1	FeatSet2	MSA_2	No
N2_M1	Net2	FeatSet1	MSA_All	Yes
N2_M2	Net2	FeatSet1	MSA_Top	No
N2_M3	Net2	FeatSet2	MSA_Top	No
N2_M4	Net2	FeatSet2	MSA_1	No
N2_M5	Net2	FeatSet2	MSA_2	No

2021), showing their great efficacy in contact/distance prediction. The aim of DuetDis is not to design new feature types, but to evaluate the performance of previously widely used feature sets under the situation of unified input and identical network, as well to study how to complement the advantages of different types of features for better prediction performance.

## Deep Network Architectures and Model Training for Distance Prediction

The proposed method DuetDis implements residual neural networks (ResNet) (He et al., 2016) as the deep learning model. Compared to traditional convolutional networks, ResNet adds feedforward neural networks to an identity map of input, which helps enable the efficient training of extremely deep neural networks. ResNet has shown its power in successful residue contact/distance prediction (Xu, 2019; Li et al., 2021). The deep residual network of DuetDis is shown in **Figure 2A**. The basic module of DuetDis network is a combination of squeeze-and-excitation and ResNet (SEResNet). The DuetDis network is composed of 33 SEResNet modules. In order to observe the impact of different networks and features on the prediction performance, we also designed another reference network (**Figure 2B**), which has very different basic modules and backbones from **Figure 2A**. The reference network is composed of 16 Res2Net modules. In this work, both SEResNet and Res2Net use dilation convolutions, while SEResNet use gelu and Res2Net use relu as the activation functions. The networks in **Figures 2A and B** are indicated as Net1 and Net2, respectively. The final MSA obtained in **Figure 1** is indicated as MSA\_All, and a subset with top 10 L sequences (ranked with sequence identity against the target sequence) selected from MSA\_All is indicated as MSA\_Top, and two disjoint subsets with each containing 10 L sequences randomly selected from MSA\_All are indicated as MSA\_1 and MSA\_2, respectively. As described in **Table 1**, 10 sub-models are trained based on Net1 (the DuetDis network) and Net2 (the reference network) with different feature sets from different MSAs. “MSA Shuffle” in **Table 1** means that the MSA are constructed through randomly selecting 10 L sequences in MSA\_All. For each epoch, N1\_M1/N2\_M1 are trained through “MSA Shuffle” strategy, N1\_M2/N1\_M3/N2\_M2/N2\_M3 are trained with MSA\_Top, N1\_M4/N2\_M4 are trained with MSA\_1, and N1\_M5/N2\_M5

are trained with MSA\_2. The outputs of five sub-models are averaged to produce the final distance map, indicated as “DuetAverage” in **Figures 2A,B**.

The sub-models are generated by independent training branches. AdamW optimizer is performed with an initial learning rate of 0.0001 (multi-step decay is adopted as the learning rate decay strategy). Cross-entropy is used as the loss-function, and L2 regularization is used during the training process to correct overfitting. The training set is split into two parts: 600 protein chains are used as the validation set and the rest are used for training. The precision of top-L long-range contact predictions (multiclass distance map is converted to the binary contact map according to the definition in **Section 2.2**) on the validation dataset is calculated at each epoch, and the training process will stop when there is no update of the validation precision for 10 epochs. The training processes are implemented in Pytorch on TeslaV100 SMX2, and each independent training generally takes 5–10 days.

## Evaluation Metrics

- 1) The predicted distance map is a matrix of probability estimates. We analyze the performance of predictors on reduced lists of distances/contacts (sorted by the probability estimates) selected by either the probability threshold or the top-L/n (L is the sequence length, and  $n = 1, 2, 5$ ) criteria. The prediction performance is assessed using precision (accuracy in some references), coverage (recall in some references), and Matthew's Correlation Coefficient (MCC), defined as follows:

$$Precision = \frac{TP}{TP + FP}, \quad (1)$$

$$Coverage = \frac{TP}{TP + FN}, \quad (2)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \quad (3)$$

where  $TP$ ,  $FP$ ,  $TN$ , and  $FN$  are the number of true positive, false positive, true negative, and false negative contacts, respectively.

- 2) Standard deviation reflects the degree of dispersion among individuals within the group, which is defined as

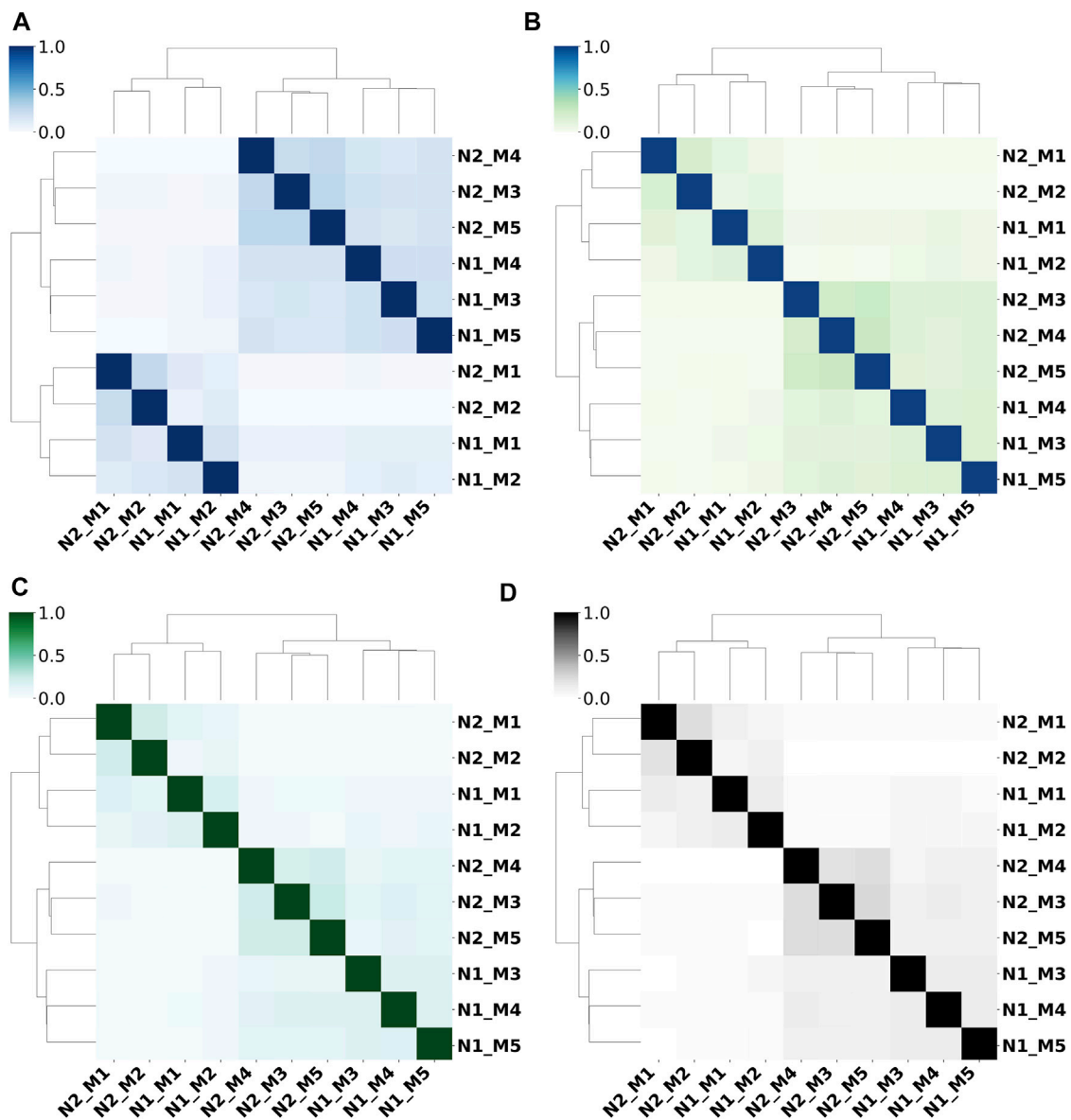
$$STD = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}, \quad (4)$$

where  $\bar{x}$  is the mean of the variable  $x$ . The standard deviation can be used to evaluate the dispersion of *Precision*, *Coverage*, and *MCC*.

- 3) Jaccard index (Jaccard similarity coefficient) measures the similarities between sets. It is defined as the size of the intersection divided by the size of the union of two sets.

$$J(X, Y) = |X \cap Y| / |X \cup Y|, \quad (5)$$

where  $X$  and  $Y$  are the set of predicted contacts from two different predictors,  $|X \cap Y|$  is the number of elements in the intersection of  $X$  and  $Y$  and the  $|X \cup Y|$  represents the number of elements in



**FIGURE 3 |** Prediction similarities between different sub-models for (A) all-range, (B) short-range, (C) mid-range, and (D) long-range contacts/distances.

the union of  $X$  and  $Y$ . The Jaccard index has values in the range of  $[0,1]$ , with the value of 0 for completely dissimilar ones and 1 for identical predictors.

## RESULTS

In this section, we assess the performance of DuetDis from different perspectives. **Section 3.1, 3.2** study the performance of sub-models, while **Section 3.3–3.5** focus on the comparison between DuetDis and peer methods. The peer methods used in this work are 4 DCA-based contact predictors (EVfold, FreeContact, gDCA, and CCMpred), 4 DL-based contact predictors (DeepCov, PconsC4, DNCON2, and SPOT-

Contact), and 3 DL-based distance predictors (TripletRes, trRosetta, and RaptorX). **Section 3.1–3.3** and **Section 3.5** use the results of top- $L/n$  ( $n = 1, 2, 5$ ) predictions, while **Section 3.4** considers the results given by specific probability/score threshold. All sub-models and peer-methods use the same MSA as input.

## Prediction Results From Different Feature Sets Show Obvious Differences

We use the Jaccard indices of prediction results from 10 sub-models (as described in **Table 1**) to study their prediction similarities. **Figure 3** shows the dendrogram heatmap of Jaccard indices using Ward's hierarchical clustering method on the independent test set. The Jaccard index between two methods

**TABLE 2 |** The prediction precisions of N1\_M1/N1\_M2/N1\_M3/N1\_M4/N1\_M5/ N1\_Ensemble for different sequence separations.

Range	Method	Top-L	Top-L/2	Top-L/5
All	N1_M1	0.7769	0.8717	0.9206
	N1_M2	0.7587	0.8475	0.8941
	N1_M3	0.7491	0.846	0.9027
	N1_M4	0.7256	0.8266	0.8888
	N1_M5	0.7319	0.8328	0.8942
	N1_Ensemble	0.7896	0.8786	0.9266
Short	N1_M1	0.2955	0.481	0.7389
	N1_M2	0.2928	0.4754	0.7287
	N1_M3	0.2948	0.4757	0.7374
	N1_M4	0.2824	0.4588	0.7109
	N1_M5	0.2947	0.473	0.7219
	N1_Ensemble	0.2988	0.4918	0.7633
Medium	N1_M1	0.3512	0.5477	0.7725
	N1_M2	0.3422	0.5336	0.7514
	N1_M3	0.342	0.5329	0.7533
	N1_M4	0.3306	0.5135	0.7275
	N1_M5	0.3371	0.5209	0.7352
	N1_Ensemble	0.3537	0.5592	0.7895
Long	N1_M1	0.6245	0.7696	0.865
	N1_M2	0.6062	0.7411	0.8273
	N1_M3	0.594	0.7308	0.8246
	N1_M4	0.5695	0.7091	0.8088
	N1_M5	0.5742	0.712	0.8121
	N1_Ensemble	0.6416	0.7797	0.8626

**TABLE 3 |** The prediction precisions of N2\_M1/N2\_M2/N2\_M3/N2\_M4/N2\_M5/ N2\_Ensemble for different sequence separations. -80

Range	Method	Top-L	Top-L/2	Top-L/5
All	N2_M1	0.7532	0.8562	0.9103
	N2_M2	0.7435	0.839	0.8938
	N2_M3	0.7148	0.8188	0.8828
	N2_M4	0.7091	0.8119	0.8768
	N2_M5	0.7071	0.8121	0.879
	N2_Ensemble	0.7590	0.8579	0.9153
Short	N2_M1	0.2864	0.4654	0.7172
	N2_M2	0.2901	0.4647	0.71
	N2_M3	0.2852	0.4583	0.7014
	N2_M4	0.2831	0.4547	0.6982
	N2_M5	0.2825	0.4548	0.7002
	N2_Ensemble	0.3449	0.5396	0.7367
Medium	N2_M1	0.3413	0.5325	0.755
	N2_M2	0.3428	0.5267	0.7395
	N2_M3	0.3298	0.5082	0.7206
	N2_M4	0.3281	0.5042	0.7152
	N2_M5	0.3283	0.5057	0.7159
	N2_Ensemble	0.3449	0.5396	0.7602
Long	N2_M1	0.6035	0.746	0.8473
	N2_M2	0.5997	0.7361	0.828
	N2_M3	0.5638	0.7022	0.8066
	N2_M4	0.5525	0.6877	0.7913
	N2_M5	0.5548	0.6917	0.7941
	N2_Ensemble	0.6136	0.7508	0.8473

is calculated by averaging the Jaccard index value of each protein on the whole test set. According to the clustering results, these 10 sub-models can be roughly divided into two categories, and each category contains two sub-categories. N1\_M1/ N1\_M2 and N2\_M1/ N2\_M2 trained by FeatSet1 are clustered into one category (Category\_1), while N1\_M3/ N1\_M4/ N1\_M5 and N2\_M3/ N2\_M4/ N2\_M5 trained by FeatSet2 form another category (Category\_2). N1\_M1/ N1\_M2 trained by Net1 and N2\_M1/ N2\_M1 trained by Net2 form two sub-categories in Category\_1, while N1\_M3/ N1\_M4/ N1\_M5 trained by Net1 and N2\_M3/ N2\_M4/ N2\_M5 trained by Net2 form two sub-categories in Category\_2. So, we can draw the conclusion that prediction results from different feature sets show obvious differences, and the conclusion is true for all-range, short-range, mid-range, and long-range contacts/distances. The feature set decides the similarity between models for typical architectures of networks.

## Ensembling Different Feature Sets Improves Prediction Performance

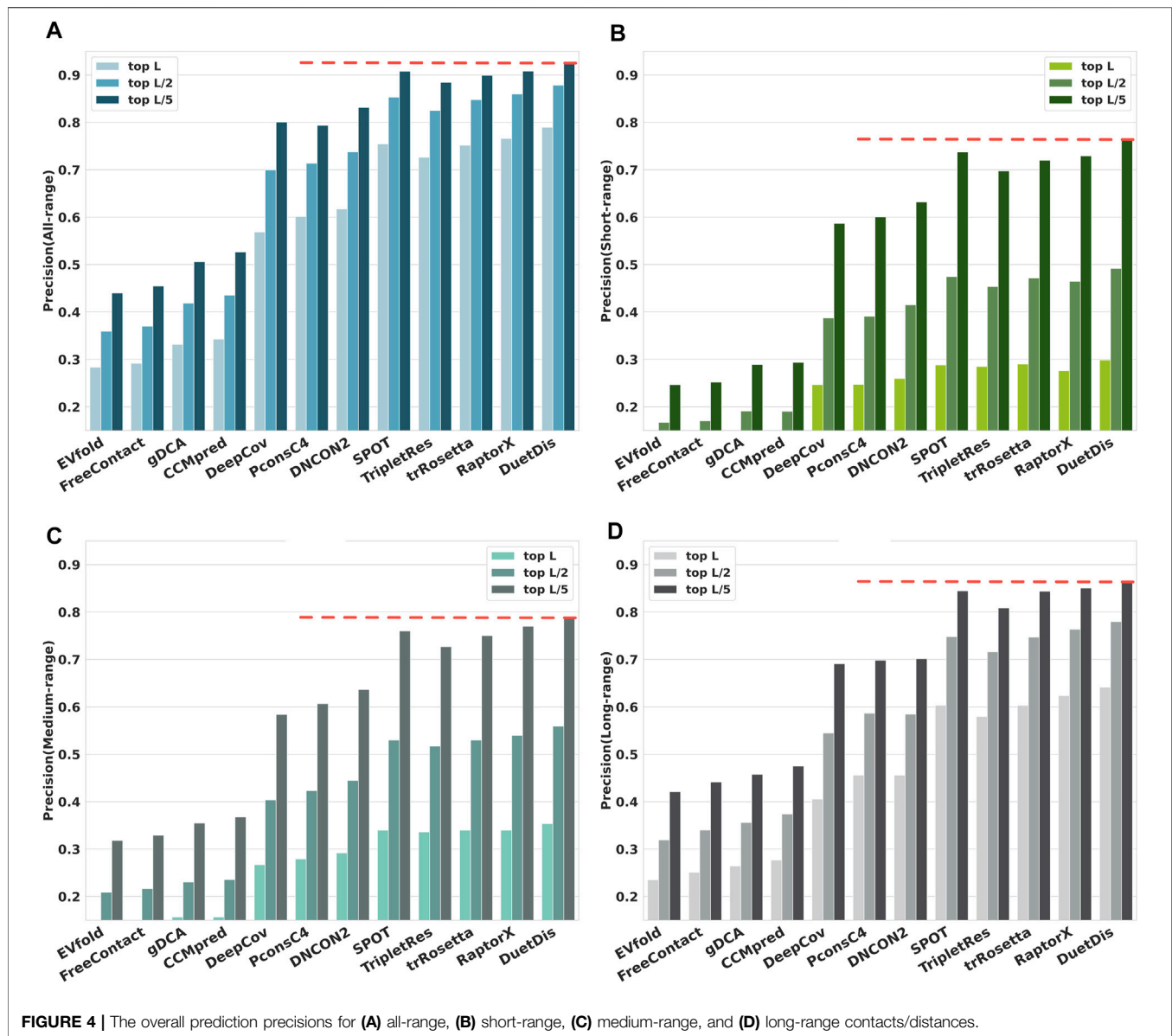
The prediction accuracies of N1\_M1/ N1\_M2/ N1\_M3/ N1\_M4/ N1\_M5/ N1\_Ensemble (obtained by averaging the five Net1 sub-models) and N2\_M1/ N2\_M2/ N2\_M3/ N2\_M4/ N2\_M5/ N2\_Ensemble (obtained by averaging the five Net2 sub-models) are listed in **Tables 2, 3**, respectively.

As we can see from **Table 2**, N1\_M1 trained through randomly shuffling MSA\_All can obtain the best performance, which is 1.8%/ 0.3%/ 0.9%/ 1.8%, 2.8%/ 0.1%/ 0.9%/ 3.0%, 5.1%/ 1%/ 2.1%/ 5.5%, and 4.5%/ 0.1%/ 2.1%/ 5% higher than N2\_M2/ N2\_M3/ N2\_M4/ N2\_M5 for top-L all-/ short-/ medium-/ long-range predictions.

Although using the same network and feature set, N1\_M1 shows superior prediction precisions than N1\_M2, implying that randomly shuffling MSA\_All in each epoch enables augmentation of the training set and thus, a better model can be obtained. N1\_M3 uses the same network and feature set as N1\_M4 and N1\_M5, but the prediction precisions of N1\_M3 are higher than N1\_M4 and N1\_M5, indicating that high-quality MSA used for training helps to boost the model performance. N1\_Ensemble outperforms the individual sub-models N1\_M1/ N1\_M2/ N1\_M3/ N1\_M4/ N1\_M5 by 1.3%/ 3.1%/ 4.0%/ 6.4%/ 5.8%, 0.3%/ 0.6%/ 0.4%/ 1.6%/ 0.4%, 0.3%/ 1.2%/ 1.2%/ 2.3%/ 1.7%, and 1.7%/ 3.5%/ 4.8%/ 7.2%/ 6.7% for top-L all-/ short-/ medium-/ long-range predictions, suggesting that ensembles of models trained on different feature sets can improve the overall prediction performance. Similar phenomenon can be observed and consistent conclusions can be drawn from the results in **Table 3**.

## The Overall Performance of DuetDis

The prediction precisions of all-/ short-/ medium-/ long-range contacts for DuetDis and other 11 peer methods on the independent test set are shown in **Figure 4**. In general, DL methods, which can capture the higher-order residue correlations and use nonlinear models with fewer parameters to be estimated from thousands of protein families (Rajgarja et al., 2010), significantly outperform DCA methods. Specifically, DuetDis shows the best overall performance. Compared with DeepCov/ PconsC4/ DNCON2/ SPOT/ TripletRes/ trRosetta/ RaptorX, DuetDis obtains 22.1%/ 18.8%/ 17.2%/ 3.5%/ 6.3%/ 3.8%/ 2.4%, 5.2%/ 5.2%/ 3.9%/ 1.0%/ 1.4%/ 0.8%/ 2.2%, 8.7%/ 7.5%/ 6.2%/ 1.4%/ 1.8%/ 1.4%/ 1.4%, and 2.4%/ 1.9%/ 3.8%/



6.2%/ 3.8%/ 1.8% higher precisions for all-range, short-range, medium-range, and long-range top-L predictions, as well as 12.5%/ 13.3%/ 9.5%/ 1.9%/ 4.2%/ 2.7%/ 1.9%, and 17.6%/ 16.3%/ 13.1%/ 2.6%/ 6.6%/ 4.3%/ 3.4% higher precisions for all-range, short-range, medium-range, and long-range top-L/5 predictions, respectively. The better performance of DuetDis is probably due to the high-quality MSAs used for training, the delicately designed deep residual network, and the effective integration of different features.

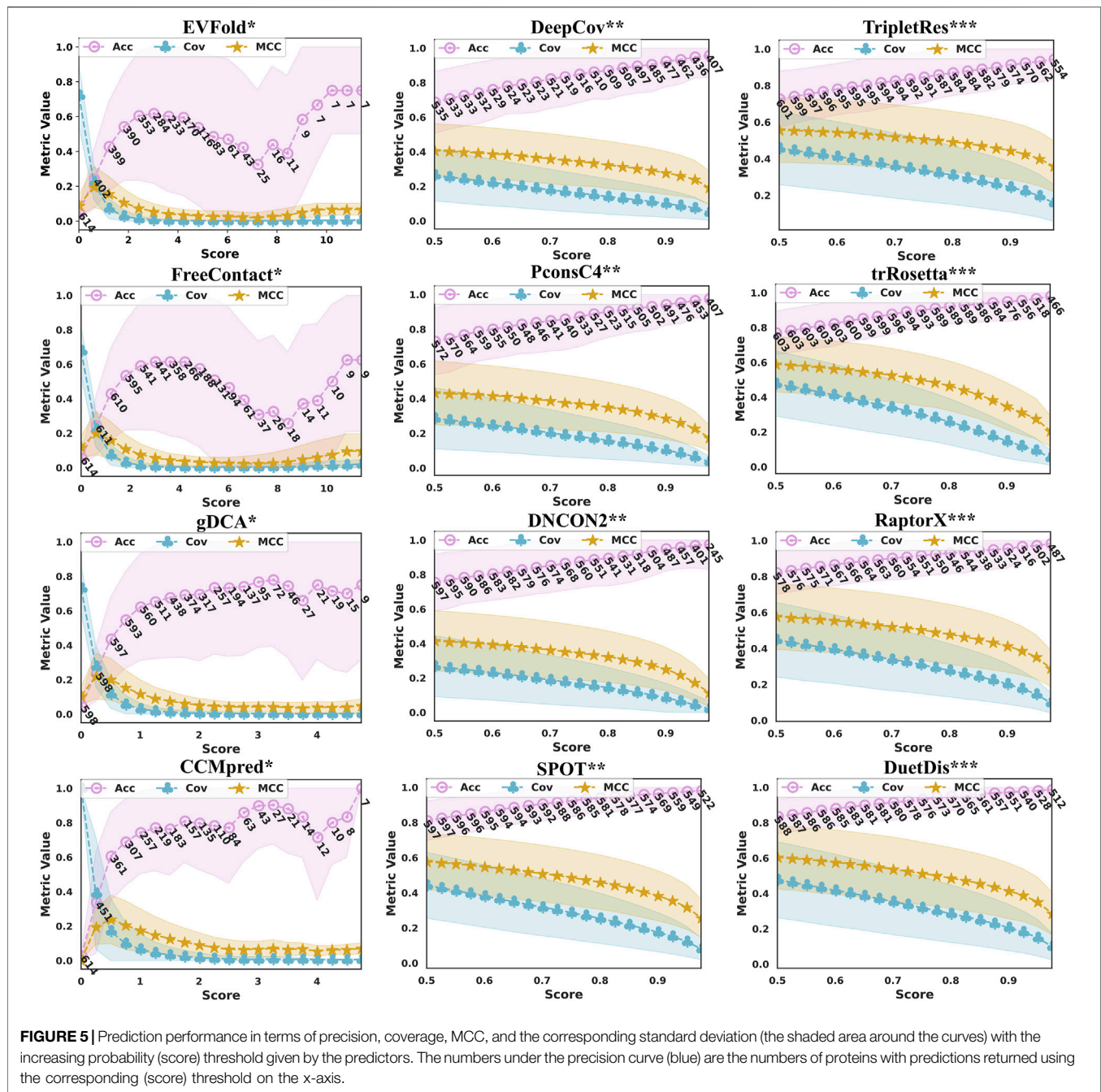
## DuetDis Embraces High Model Reliability in Terms of Prediction Score

The confidence of the probability (score) given by a DCA or DL model can greatly reflect the reliability of the corresponding model. The prediction probabilities (scores) given by EVfold, FreeContact,

gDCA, CCMpred, DeepCov, PconsC4, DNCON2, SPOT, TripletRes, trRosetta, RaptorX, and DuetDis are distributed at (0.000,1.309), (−2.537,17.931), (−1.243, 6.564), (0.000, 5.270), (0.0, 1.0), (0.0, 1.0), (0.0, 1.0), (0.0, 1.0), (0.0, 1.0), (0.0, 1.0), (0.0, 1.0), and (0.0, 1.0), respectively. For machine learning (both traditional and deep learning) applications, people usually use 0.5 as a threshold for classification. However, the threshold may be inaccurate for a complex problem like contact/distance prediction. Therefore, studying the scoring trend and the reliability of the model is of great benefit to understand the model performance.

Figure 5 illustrates the prediction performance in terms of precision/ coverage/ MCC with the increase in probability (score) threshold given by DuetDis and the peer methods. With the increase of the probability (score) threshold, the prediction coverages decrease monotonically for all methods. As the threshold increases, their precision curves go down at some

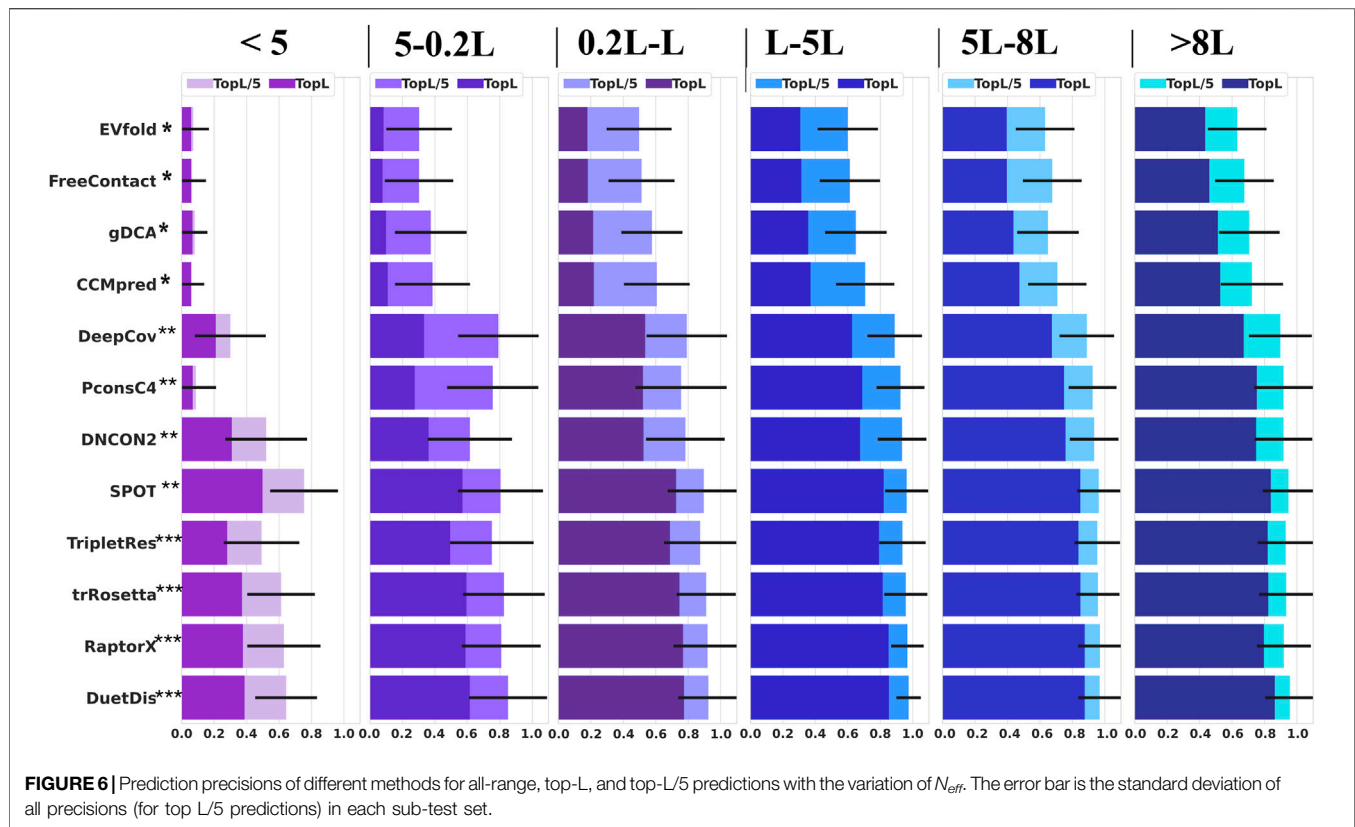




probability (score) value. The prediction precisions of all DL methods (DeepCov/ PconsC4/ DNCON2/ SPOT/ TripletRes/ trRosetta/ RaptorX) increase monotonically with the probability (score) threshold. However, the precision curves of DCA methods (EVfold/ FreeContact/ gDCA/ CCMpred) show turning points at some probability (score) values. Meanwhile, DCA methods also show much larger STDs on precisions and relatively lower coverages/MCCs compared with DL methods. The numbers under the precision curve in **Figure 4** are the numbers of proteins with predictions returned using the corresponding probability (score) threshold on the x-axis. It is

obvious that, as the probability (score) threshold increases, there are more proteins being predicted by DL methods than by DCA methods. Specifically, DuetDis achieves prediction precisions/coverages/ MCCs of 98.1%/ 15.0%/ 0.352 (calculated on the 523 proteins with prediction scores higher than 0.95) at the (score) threshold of 0.95, which are higher than that by DeepCov (94.7%/ 7.4%/ 0.240: 431 proteins), PconsC4 (96.3%/ 6.5%/ 0.228: 448 proteins), DNCON2 (96.8%/ 4.4%/ 0.173: 396 proteins), SPOT (97.5%/ 12.3%/ 0.318: 544 proteins), TripletRes (93.0%/ 19.6%/ 0.399: 557 proteins), trRosetta (96.5%/ 9.4%/ 0.276: 513 proteins), and RaptorX (97.2%/ 14.5%/ 0.352: 497 proteins). In summary,





DuetDis shows higher reliability in model probability (score) compared with peer methods.

## DuetDis Is Robust Against Shallow Multiple Sequence Alignment

Coevolutionary coupling signals extracted from MSA play central role in most modern contact/distance prediction methods. In this study, the independent test set is divided into six groups according to  $N_{eff}$  (<5, 5–0.2 L, 0.2 L–L, L–5 L, 5–8 L, and >8 L). The performance of different methods on these sub-groups of the test set is shown in **Figure 6**. DuetDis achieves prediction precisions of 64.4% for  $N_{eff}$  <5, 85.1% for  $N_{eff}$  = 5–0.2 L (2.5% higher than the second), 92.5% for  $N_{eff}$  = 0.2 L–L (0.5% higher than the second), 97.5% for  $N_{eff}$  = L–5 L (0.8% higher than the second), 96.9% for  $N_{eff}$  = 5–8 L (0.2% higher than the second), and 95.6% for  $N_{eff}$  = >8 L (0.9% higher than the second). For  $N_{eff}$  <5 L, DuetDis ranks the second in prediction precision; while for  $N_{eff}$  = 5–0.2 L, 0.2 L–L, L–5 L, 5–8 L and >8 L, DuetDis is in the leading position of prediction precision. For  $N_{eff}$  <5 L, PconsC4 shows a STD of 0.125 which is smaller than DuetDis, however, the smaller STD is because of lower overall precision by PconsC4 (the average prediction precisions are 8.7% for PconsC4 and 64.4% for DuetDis). Hence, DuetDis obtains the least STD among all DL methods for all sub-groups of the test set. In general, DuetDis shows leading precisions and the smallest STD for most ranges of  $N_{eff}$ , especially highlights its robustness in shallow MSA-based distance prediction.

## CONCLUSION

Proteins are considered as the molecular machines and perform many important functions of life (Zhang et al., 2017). Knowing the structure of a protein helps to understand the role of the protein, how the protein performs its biological function, and the interaction between the protein and the protein (or other molecules), which is very important for biology as well as for medicine and pharmacy. Residue distance prediction from the sequence is critical for many biological applications such as protein structure reconstruction. However, prediction of large distances and distances between residues with long sequence separation length still remains challenging.

In this paper, we propose DuetDis, which uses duet deep learning models for distance prediction. DuetDis adopts two complementary feature sets, one set is mainly composed of 2D coevolutionary couplings, and another set contains mainly 1D sequence-based features. We trained 10 sub-models using two different networks (Net1 and Net2), two different sets of features (FeatSet1 and FeatSet2), and four different MSAs (MSA\_All, MSA\_Top, MSA\_1, MSA\_2). By evaluating 10 sub-models based on the large-scale test set, we found that: 1) prediction results from different feature sets show obvious differences; 2) ensembling different feature sets can improve the prediction performance; and 3) high-quality MSA used for both training and testing can greatly improve the prediction performance. DuetDis is also compared with 11 widely used contact/distance predictors. The experimental results show that DuetDis outperforms the peer methods in terms of overall prediction precisions, model reliability, and robustness against shallow MSA.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

HZ, YH, and ZB conducted the experiments; all authors analyzed the data; HZ and WX wrote the manuscript.

## REFERENCES

- Adhikari, B. (2020). A Fully Open-Source Framework for Deep Learning Protein Real-Valued Distances. *Sci. Rep.* 10 (1), 13374. doi:10.1038/s41598-020-70181-0
- Adhikari, B., Bhattacharya, D., Cao, R., and Cheng, J. (2015). CONFOLD: Residue-Residue Contact-Guided Ab Initio Protein Folding. *Proteins* 83 (8), 1436–1449. doi:10.1002/prot.24829
- Adhikari, B., Hou, J., and Cheng, J. (2018). DNCON2: Improved Protein Contact Prediction Using Two-Level Deep Convolutional Neural Networks. *Bioinformatics* 34 (9), 1466–1472. doi:10.1093/bioinformatics/btx781
- Altschul, S., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs. *Nucleic Acids Res.* 25 (17), 3389–3402. doi:10.1093/nar/25.17.3389
- Anfinsen, C. B. (1973). Principles that Govern the Folding of Protein Chains. *Science* 181 (4096), 223–230. doi:10.1126/science.181.4096.223
- Anishchenko, I., Pellock, S. J., Chidyausiku, T. M., Ramelot, T. A., Ovchinnikov, S., Hao, J., et al. (2021). De Novo protein Design by Deep Network Hallucination. *Nature* 600 (7889), 547–552. doi:10.1038/s41586-021-04184-w
- Aszódi, A., and Taylor, W. R. (1996). Homology Modelling by Distance Geometry. *Folding Des.* 1 (5), 325–334.
- Baldassi, C., Zamparo, M., Feinauer, C., Procaccini, A., Zecchina, R., Weigt, M., et al. (2014). Fast and Accurate Multivariate Gaussian Modeling of Protein Families: Predicting Residue Contacts and Protein-Interaction Partners. *PLoS one* 9 (3), e92721. doi:10.1371/journal.pone.0092721
- Betancourt, M. R., and Thirumalai, D. (1999). Pair Potentials for Protein Folding: Choice of Reference States and Sensitivity of Predicted Native States to Variations in the Interaction Schemes. *Protein Sci.* 8 (2), 361–369. doi:10.1110/ps.8.2.361
- Cheng, J., and Baldi, P. (2007). Improved Residue Contact Prediction Using Support Vector Machines and a Large Feature Set. *BMC Bioinformatics* 8 (1), 113. doi:10.1186/1471-2105-8-113
- Cong, Q., Anishchenko, I., Ovchinnikov, S., and Baker, D. (2019). Protein Interaction Networks Revealed by Proteome Coevolution. *Science* 365 (6449), 185–189. doi:10.1126/science.aaw6718
- Ding, W., and Gong, H. (2020). Predicting the Real-Valued Inter-Residue Distances for Proteins. *Adv. Sci.* 7 (19), 2001314. doi:10.1002/advsc.202001314
- Ding, W., Mao, W., Shao, D., Zhang, W., and Gong, H. (2018). DeepConPred2: An Improved Method for the Prediction of Protein Residue Contacts. *Comput. Struct. Biotechnol. J.* 16, 503–510. doi:10.1016/j.csbj.2018.10.009
- Du, T., Liao, L., Wu, C. H., and Sun, B. (2016). Prediction of Residue-Residue Contact Matrix for Protein-Protein Interaction with Fisher Score Features and Deep Learning. *Methods* 110, 97–105. doi:10.1016/j.ymeth.2016.06.001
- Dunn, S. D., Wahl, L. M., and Gloor, G. B. (2007). Mutual Information without the Influence of Phylogeny or Entropy Dramatically Improves Residue Contact Prediction. *Bioinformatics* 24 (3), 333–340. doi:10.1093/bioinformatics/btm604
- Ekeberg, M., Lökvist, C., Lan, Y., Weigt, M., and Aurell, E. (2013). Improved Contact Prediction in Proteins: Using Pseudolikelihoods to Infer Potts Models. *Phys. Rev. E Stat. Nonlin Soft Matter Phys.* 87 (1), 012707. doi:10.1103/PhysRevE.87.012707
- Gao, M., Zhou, H., and Skolnick, J. (2019). DESTINI: A Deep-Learning Approach to Contact-Driven Protein Structure Prediction. *Sci. Rep.* 9 (1), 3514. doi:10.1038/s41598-019-40314-1
- Greener, J. G., Kandathil, S. M., and Jones, D. T. (2019). Deep Learning Extends De Novo Protein Modelling Coverage of Genomes Using Iteratively Predicted Structural Constraints. *Nat. Commun.* 10 (1), 3977. doi:10.1038/s41467-019-11994-0
- Hanson, J., Paliwal, K., Litfin, T., Yang, Y., and Zhou, Y. (2018). Accurate Prediction of Protein Contact Maps by Coupling Residual Two-Dimensional Bidirectional Long Short-Term Memory with Convolutional Neural Networks. *Bioinformatics* 34 (23), 4039–4045. doi:10.1093/bioinformatics/bty481
- Hanson, J., Paliwal, K., Litfin, T., Yang, Y., and Zhou, Y. (2019). Improving Prediction of Protein Secondary Structure, Backbone Angles, Solvent Accessibility and Contact Numbers by Using Predicted Contact Maps and an Ensemble of Recurrent and Residual Convolutional Neural Networks. *Bioinformatics* 35 (14), 2403–2410. doi:10.1093/bioinformatics/bty1006
- He, B., Mortuza, S. M., Wang, Y., Shen, H.-B., and Zhang, Y. (2017). NeBcon: Protein Contact Map Prediction Using Neural Network Training Coupled with Naïve Bayes Classifiers. *Bioinformatics* 33 (15), 2296–2306. doi:10.1093/bioinformatics/btx164
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). “Deep Residual Learning for Image Recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition. 17–19 June 1997. Juan, PR, USA. (IEEE). doi:10.1109/cvpr.2016.90
- Jain, A., Terashi, G., Kagaya, Y., Venkata Subramaniya, S. R. M., Christoffer, C., and Kihara, D. (2021). Analyzing Effect of Quadruple Multiple Sequence Alignments on Deep Learning Based Protein Inter-residue Distance Prediction. *Scientific Rep.* 11 (1), 1–13. doi:10.1038/s41598-021-87204-z
- Johnson, L. S., Eddy, S. R., and Portugaly, E. (2010). Hidden Markov Model Speed Heuristic and Iterative HMM Search Procedure. *BMC bioinformatics* 11 (1), 431. doi:10.1186/1471-2105-11-431
- Jones, D. T., Buchan, D. W. A., Cozzetto, D., and Pontil, M. (2012). PSICOV: Precise Structural Contact Prediction Using Sparse Inverse Covariance Estimation on Large Multiple Sequence Alignments. *Bioinformatics* 28 (2), 184–190. doi:10.1093/bioinformatics/btr638
- Jones, D. T., and Kandathil, S. M. (2018). High Precision in Protein Contact Prediction Using Fully Convolutional Neural Networks and Minimal Sequence Features. *Bioinformatics* 34 (19), 3308–3315. doi:10.1093/bioinformatics/bty341
- Jones, D. T., Singh, T., Kosciolk, T., and Tetchner, S. (2014). MetaPSICOV: Combining Coevolution Methods for Accurate Prediction of Contacts and Long Range Hydrogen Bonding in Proteins. *Bioinformatics* 31 (7), 999–1006. doi:10.1093/bioinformatics/btu791
- Ju, Z., Zhang, H., Meng, J., Zhang, J., Li, X., Fan, J., et al. (2021). “An Efficient Greedy Incremental Sequence Clustering Algorithm,” in *International Symposium on Bioinformatics Research and Applications* (Springer, Cham). doi:10.1007/978-3-030-91415-8\_50
- Kaján, L., Hopf, T. A., Kalaš, M., Marks, D. S., and Rost, B. (2014). FreeContact: Fast and Free Software for Protein Contact Prediction from Residue Coevolution. *BMC bioinformatics* 15 (1), 85. doi:10.1186/1471-2105-15-85
- Kamisetty, H., Ovchinnikov, S., and Baker, D. (2013). Assessing the Utility of Coevolution-Based Residue-Residue Contact Predictions in a Sequence- and

## FUNDING

This work was partly supported by the National Key Research and Development Program of China under Grant No. 2018YFB0204403, Strategic Priority CAS Project XDB38050100, the Key Research and Development Project of Guangdong Province under Grant No. 2021B0101310002, National Science Foundation of China under Grant No. U1813203, the Shenzhen Basic Research Fund under Grant Nos. RCYX2020071411473419, JCYJ20200109114818703, and JSGG20201102163800001, and CAS Key Lab under Grant No. 2011DPI73015.

- Structure-Rich Era. *Proc. Natl. Acad. Sci. U.S.A.* 110 (39), 15674–15679. doi:10.1073/pnas.1314045110
- Kucic, P., Mirabello, C., Tradigo, G., Walsh, I., Veltri, P., and Pollastri, G. (2014). Toward an Accurate Prediction of Inter-residue Distances in Proteins Using 2D Recursive Neural Networks. *BMC bioinformatics* 15 (1), 6–15. doi:10.1186/1471-2105-15-6
- Lee, B.-C., and Kim, D. (2009). A New Method for Revealing Correlated Mutations under the Structural and Functional Constraints in Proteins. *Bioinformatics* 25 (19), 2506–2513. doi:10.1093/bioinformatics/btp455
- Li, J., and Xu, J. (2021). Study of Real-Valued Distance Prediction for Protein Structure Prediction with Deep Learning. *Bioinformatics* 37 (19), 3197–3203. doi:10.1093/bioinformatics/btab333
- Li, Y., Hu, J., Zhang, C., Yu, D.-J., and Zhang, Y. (2019). ResPRE: High-Accuracy Protein Contact Prediction by Coupling Precision Matrix with Deep Residual Neural Networks. *Bioinformatics* 35 (22), 4647–4655. doi:10.1093/bioinformatics/btz291
- Li, Y., Zhang, C., Bell, E. W., Zheng, W., Zhou, X., Yu, D.-J., et al. (2021). Deducing High-Accuracy Protein Contact-Maps from a Triplet of Coevolutionary Matrices through Deep Residual Convolutional Networks. *Plos Comput. Biol.* 17 (3), e1008865. doi:10.1371/journal.pcbi.1008865
- Liu, Y., Palmedo, P., Ye, Q., Berger, B., and Peng, J. (2018). Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks. *Cell Syst.* 6 (1), 65–74. e3. doi:10.1016/j.cels.2017.11.014
- Malinin, A., and Gales, M. J. F. (2021). Uncertainty Estimation in Autoregressive Structured Prediction. 9th International Conference on Learning Representations, {ICLR} 2021, Virtual Event, Austria, May 3–7, 2021. Available at: <https://openreview.net/forum?id=jN5y-zb5Q7m>.
- Marks, D. S., Hopf, T. A., and Sander, C. (2012). Protein Structure Prediction from Sequence Variation. *Nat. Biotechnol.* 30 (11), 1072–1080. doi:10.1038/nbt.2419
- McAllister, S. R., and Floudas, C. A. (2008).  $\alpha$ -Helical Topology Prediction and Generation of Distance Restraints in Membrane Proteins. *Biophysical J.* 95 (11), 5281–5295. doi:10.1529/biophysj.108.132241
- Michel, M., Hayat, S., Skwark, M. J., Sander, C., Marks, D. S., and Elofsson, A. (2014). PconsFold: Improved Contact Predictions Improve Protein Models. *Bioinformatics* 30 (17), i482–i488. doi:10.1093/bioinformatics/btu458
- Michel, M., Menéndez Hurtado, D., and Elofsson, A. (2019). PconsC4: Fast, Accurate and Hassle-free Contact Predictions. *Bioinformatics* 35 (15), 2677–2679. doi:10.1093/bioinformatics/bty1036
- Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., et al. (2011). Direct-coupling Analysis of Residue Coevolution Captures Native Contacts across many Protein Families. *Proc. Natl. Acad. Sci. U S A.* 108 (49), E1293–E1301. doi:10.1073/pnas.1111471108
- Pollock, D. D., and Taylor, W. R. (1997). Effectiveness of Correlation Analysis in Identifying Protein Residues Undergoing Correlated Evolution. *Protein Eng. Des. Selection* 10 (6), 647–657. doi:10.1093/protein/10.6.647
- Rahman, J., Newton, M. A. H., Islam, M. K. B., and Sattar, A. (2022). Enhancing Protein Inter-residue Real Distance Prediction by Scrutinising Deep Learning Models. *Sci. Rep.* 12 (1), 787. doi:10.1038/s41598-021-04441-y
- Rajgaria, R., McAllister, S. R., and Floudas, C. A. (2009). Towards Accurate Residue-Residue Hydrophobic Contact Prediction for  $\alpha$  Helical Proteins via Integer Linear Optimization. *Proteins* 74 (4), 929–947. doi:10.1002/prot.22202
- Rajgaria, R., Wei, Y., and Floudas, C. A. (2010). Contact Prediction for Beta and Alpha-Beta Proteins Using Integer Linear Optimization and its Impact on the First Principles 3D Structure Prediction Method ASTRO-FOLD. *Proteins* 78 (8), 1825–1846. doi:10.1002/prot.22696
- Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment. *Nat. Methods* 9 (2), 173–175. doi:10.1038/nmeth.1818
- Reza, M. S., Zhang, H., Hossain, M. T., Jin, L., Feng, S., and Wei, Y. (2021). COMTOP: Protein Residue-Residue Contact Prediction through Mixed Integer Linear Optimization. *Membranes* 11 (7), 503. doi:10.3390/membranes11070503
- Schlessinger, A., Punta, M., and Rost, B. (2007). Natively Unstructured Regions in Proteins Identified from Contact Predictions. *Bioinformatics* 23 (18), 2376–2384. doi:10.1093/bioinformatics/btm349
- Seemayer, S., Gruber, M., and Söding, J. (2014). CCMpred-fast and Precise Prediction of Protein Residue-Residue Contacts from Correlated Mutations. *Bioinformatics* 30 (21), 3128–3130. doi:10.1093/bioinformatics/btu500
- Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2020). Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature* 577, 706–710. doi:10.1038/s41586-019-1923-7
- Shimomura, T., Nishijima, K., and Kikuchi, T. (2019). A New Technique for Predicting Intrinsically Disordered Regions Based on Average Distance Map Constructed with Inter-residue Average Distance Statistics. *BMC Struct. Biol.* 19 (1), 3–12. doi:10.1186/s12900-019-0101-3
- Singh, J., Litfin, T., Singh, J., Paliwal, K., and Zhou, Y. (2022). SPOT-Contact-LM: Improving Single-Sequence-Based Prediction of Protein Contact Map Using a Transformer Language Model. *Bioinformatics*. doi:10.1093/bioinformatics/btac053
- Skwark, M. J., Abdel-Rehim, A., and Elofsson, A. (2013). PconsC: Combination of Direct Information Methods and Alignments Improves Contact Prediction. *Bioinformatics* 29 (14), 1815–1816. doi:10.1093/bioinformatics/btt259
- Su, H., Wang, W., Du, Z., Peng, Z., Gao, S. H., Cheng, M. M., et al. (2021). Improved Protein Structure Prediction Using a New Multi-Scale Network and Homologous Templates. *Adv. Sci.* 8, 2102592. doi:10.1002/adv.202102592
- Tegge, A. N., Wang, Z., Eickholt, J., and Cheng, J. (2009). NNcon: Improved Protein Contact Map Prediction Using 2D-Recursive Neural Networks. *Nucleic Acids Res.* 37, W515–W518. doi:10.1093/nar/gkp305
- Vangone, A., and Bonvin, A. M. (2015). Contacts-based Prediction of Binding Affinity in Protein-Protein Complexes. *elife* 4, e07454. doi:10.7554/eLife.07454
- Walsh, I., Baù, D., Martin, A. J., Mooney, C., Vullo, A., and Pollastri, G. (2009). Ab Initio and Template-Based Prediction of Multi-Class Distance Maps by Two-Dimensional Recursive Neural Networks. *BMC Struct. Biol.* 9 (1), 5–20. doi:10.1186/1472-6807-9-5
- Wang, S., Sun, S., Li, Z., Zhang, R., and Xu, J. (2017). Accurate De Novo Prediction of Protein Contact Map by Ultra-deep Learning Model. *Plos Comput. Biol.* 13 (1), e1005324. doi:10.1371/journal.pcbi.1005324
- Wang, Z., and Xu, J. (2013). Predicting Protein Contact Map Using Evolutionary and Physical Constraints by Integer Programming. *Bioinformatics* 29 (13), i266–i273. doi:10.1093/bioinformatics/btt211
- Wei, Y., and Floudas, C. A. (2011). Enhanced Inter-helical Residue Contact Prediction in Transmembrane Proteins. *Chem. Eng. Sci.* 66 (19), 4356–4369. doi:10.1016/j.ces.2011.04.033
- Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., and Hwa, T. (2009). Identification of Direct Residue Contacts in Protein-Protein Interaction by Message Passing. *Proc. Natl. Acad. Sci. U.S.A.* 106 (1), 67–72. doi:10.1073/pnas.0805923106
- Wu, Q., Peng, Z., Anishchenko, I., Cong, Q., Baker, D., and Yang, J. (2020). Protein Contact Prediction Using Metagenome Sequence Data and Residual Neural Networks. *Bioinformatics* 36 (1), 41–48. doi:10.1093/bioinformatics/btz477
- Wu, S., and Zhang, Y. (2008). A Comprehensive Assessment of Sequence-Based and Template-Based Methods for Protein Contact Prediction. *Bioinformatics* 24 (7), 924–931. doi:10.1093/bioinformatics/btn069
- Wu, T., Guo, Z., Hou, J., and Cheng, J. (2021). DeepDist: Real-Value Inter-Residue Distance Prediction with Deep Residual Convolutional Network. *BMC Bioinform.* 22, 30. doi:10.1186/s12859-021-04269-3
- Xu, J. (2019). Distance-based Protein Folding Powered by Deep Learning. *Proc. Natl. Acad. Sci. U.S.A.* 116 (34), 16856–16865. doi:10.1073/pnas.1821309116
- Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. (2020). Improved Protein Structure Prediction Using Predicted Interresidue Orientations. *Proc Natl Acad Sci U S A.* 117(3), 1496–1503. doi:10.1073/pnas.1914677117
- Zhang, H., Bei, Z., Xi, W., Hao, M., Ju, Z., Saravanan, K. M., et al. (2021). Evaluation of Residue-Residue Contact Prediction Methods: From Retrospective to Prospective. *Plos Comput. Biol.* 17 (5), e1009027. doi:10.1371/journal.pcbi.1009027
- Zhang, H., Wu, H., Ting, H. F., and Wei, Y. (2020). “Protein Interresidue Contact Prediction Based on Deep Learning and Massive Features from Multi-Sequence Alignment,” in International Conference on Parallel and Distributed Computing: Applications and Technologies, Shenzhen, China, December 28–30 (Shenzhen: Springer).

- Zhang, H., Hao, M., Wu, H., Ting, H.-F., Tang, Y., Xi, W., et al. (2022). Protein Residue Contact Prediction Based on Deep Learning and Massive Statistical Features from Multi-Sequence Alignment. *Tsinghua Sci. Technol.* 27 (5), 843–854. doi:10.26599/tst.2021.9010064
- Zhang, H., Huang, Q., Bei, Z., Wei, Y., and Floudas, C. A. (2016). COMSAT: Residue Contact Prediction of Transmembrane Proteins Based on Support Vector Machines and Mixed Integer Linear Programming. *Proteins* 84 (3), 332–348. doi:10.1002/prot.24979
- Zhang, H., Xi, W., Hansmann, U. H. E., and Wei, Y. (2017). Fibril-Barrel Transitions in Cylindrin Amyloids. *J. Chem. Theor. Comput.* 13 (8), 3936–3944. doi:10.1021/acs.jctc.7b00383
- Zhao, F., and Xu, J. (2012). A Position-specific Distance-dependent Statistical Potential for Protein Structure and Functional Study. *Structure* 20 (6), 1118–1126. doi:10.1016/j.str.2012.04.003
- Zheng, W., Zhou, X., Wuyun, Q., Pearce, R., Li, Y., and Zhang, Y. (2020). FUPred: Detecting Protein Domains through Deep-Learning-Based Contact Map Prediction. *Bioinformatics* 36 (12), 3749–3757. doi:10.1093/bioinformatics/btaa217

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

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

Copyright © 2022 Zhang, Huang, Bei, Ju, Meng, Hao, Zhang, Zhang and Xi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



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

Xue Chen<sup>1</sup>, Qianye Zhang<sup>1</sup>, Bowen Li<sup>1</sup>, Chunying Lu<sup>1</sup>, Shanshan Yang<sup>1</sup>, Jinjin Long<sup>1</sup>, Bifang He<sup>1\*</sup>, Heng Chen<sup>1\*</sup> and Jian Huang<sup>2\*</sup>

<sup>1</sup>Medical College, Guizhou University, Guiyang, China, <sup>2</sup>School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China

## OPEN ACCESS

### Edited by:

Chuan Dong,  
Wuhan University, China

### Reviewed by:

Leyi Wei,  
Shandong University, China  
Zunnan Huang,  
Guangdong Medical University, China

### \*Correspondence:

Bifang He  
bthe@gzu.edu.cn  
Heng Chen  
hchen13@gzu.edu.cn  
Jian Huang  
hj@uestc.edu.cn

### Specialty section:

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

**Received:** 30 December 2021

**Accepted:** 30 March 2022

**Published:** 17 May 2022

### Citation:

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

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

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

## 1 INTRODUCTION

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

Blood-brain barrier penetrating peptides (BBPs) can cross the BBB through various mechanisms without destroying the integrity of BBB (Van Dorpe et al., 2012; Oller-Salvia et al., 2016). It has been reported that partial BBPs can transfer drugs into the brain, which provides a new avenue for the development of drugs for CNS diseases (Zhou et al., 2021). Furthermore, because of their



characteristics of easy synthesis, satisfactory effect, low toxicity and wide selectivity (Muttenthaler et al., 2021), BBPs show broad application prospects as carriers or therapeutic agents for CSN diseases treatment (Zhou et al., 2021). Nonaka et al. reported that IF7, an annexin A1-binding peptide, could overcome BBB and deliver chemotherapeutics to target brain tumors (Nonaka et al., 2020). Xie and coworkers demonstrated that d-peptide ligand of angiopep-2 modified nanoprobe could cross BBB and locate glioma sites (Xie et al., 2021). Lim and collaborators found that dNP2 peptide could penetrate BBB and deliver ctCTLA-4 protein to ameliorate autoimmune encephalomyelitis in mouse models (Lim et al., 2015). Kurzrock and Drappatz et al. showed that ANG1005 or GRN1005, a conjugate of angiopep-2 and paclitaxel, has reached clinical study for the treatment of glioma (Kurzrock et al., 2012; Drappatz et al., 2013).

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

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

## 2 MATERIALS AND METHODS

### 2.1 Datasets

In this work, we selected experimentally validated BBPs as candidate positive samples that were collected from Brainpeps (Van Dorpe et al., 2012), B3Pdb (Kumar et al., 2021b), public datasets of BBPpred (Dai et al., 2021) and B3Pred (Kumar et al.,

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

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

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

### 2.2 Training and Independent Testing Datasets

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

### 2.3 Feature Extraction

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

(i.e., “ACDEFGHIKLMNPQRSTVWY”). Feature extraction was implemented by an in-house script.

### 2.3.1 Amino Acid Composition

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

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

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

### 2.3.2 Dipeptide Composition

DPC gives 400 descriptors (i.e. “AA, AC, AD, … YY”) (Saravanan and Gautham, 2015). It is defined as:

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

where  $N_{rs}$  is the number of the dipeptide consisting of amino acids  $r$  and  $s$  in the peptide sequence.

### 2.3.3 Grouped Amino Acid Composition

For the GAAC encoding, 20 natural amino acids are firstly divided into five categories according to their physicochemical properties: amino acid groups  $g_1$  (GAVLMI),  $g_2$  (FYW),  $g_3$  (KRH),  $g_4$  (DE) and  $g_5$  (STCPNQ). Group  $g_1$  belongs to the aliphatic group,  $g_2$  aromatic group,  $g_3$  positive charge group,  $g_4$  negative charged group and  $g_5$  uncharged group, respectively. GAAC represents the frequency of each amino acid group (Lee et al., 2011) and can be described as:

$$f(g) = \frac{N(g_i)}{N}, i \in \{g_1, g_2, g_3, g_4, g_5\} \quad (3)$$

$$N(g_i) = \sum N(i), i \in \{g_1, g_2, g_3, g_4, g_5\}$$

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

### 2.3.4 Composition of $K$ -Spaced Amino Acid Group Pairs

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

$$\left( \frac{N_{g_1g_1}}{N_{total}}, \frac{N_{g_1g_2}}{N_{total}}, \frac{N_{g_1g_3}}{N_{total}}, \dots, \frac{N_{g_1g_5}}{N_{total}} \right)_{25} \quad (4)$$

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

### 2.3.5 Pseudo-Amino Acid Composition

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

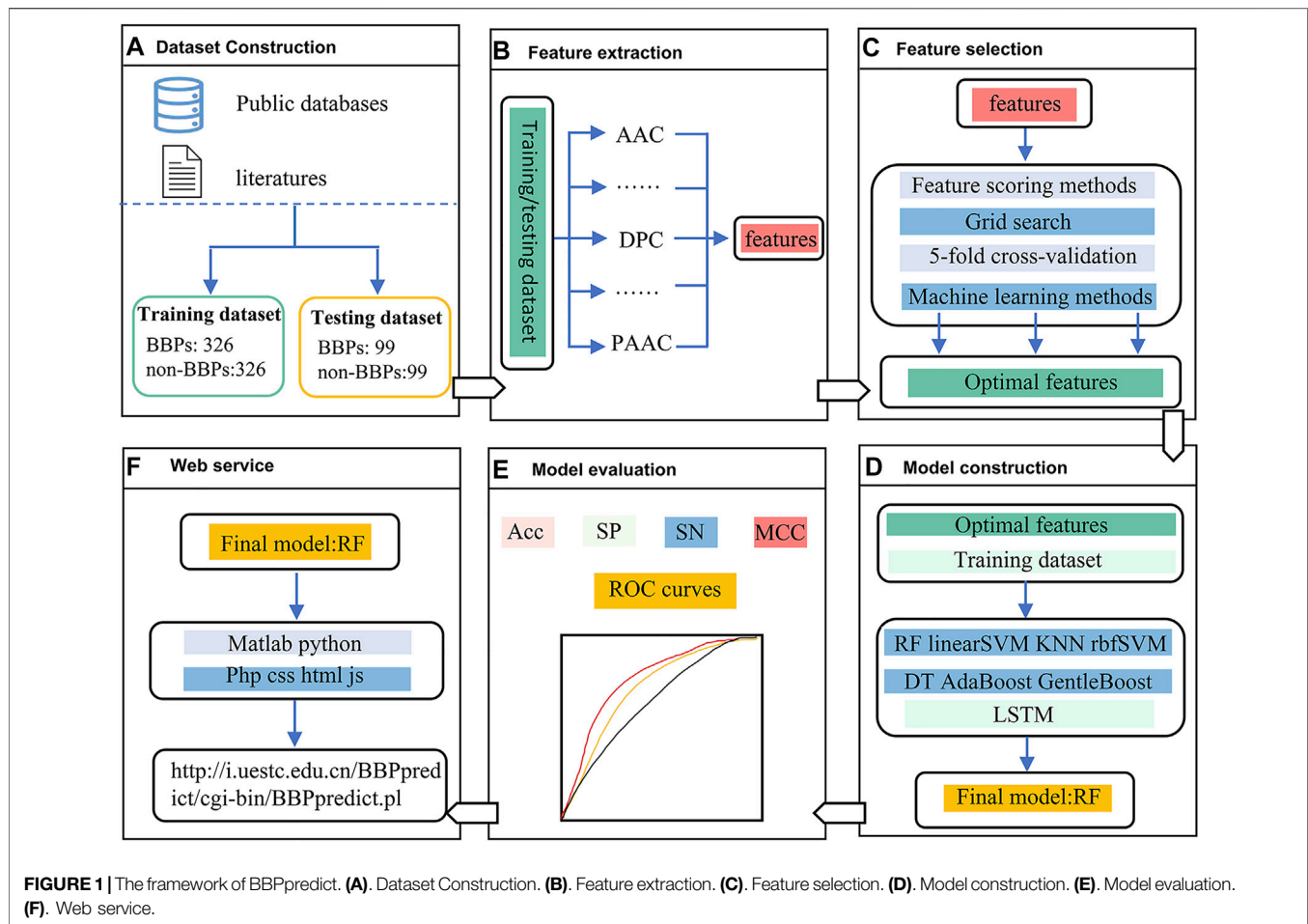
After feature extraction, each peptide was encoded by a 550-dimensional feature vector, which was generated by concatenating five types of feature vector.

## 2.4 Feature Scoring and Selection

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

## 2.5 Classification Model Construction

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



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

## 2.6 Prediction Assessment

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

$$SN = \frac{TP}{TP + FN} \quad (5)$$

$$SP = \frac{TN}{TN + FP} \quad (6)$$

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} \quad (7)$$

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

where  $TP$  describes the number of genuine BBPs which are predicted as BBPs.  $FN$  represents the number of genuine BBPs that are identified as non-BBPs. Denote  $TN$  as the number of true non-BBPs classified as non-BBPs and  $FP$  the number of true non-BBPs identified as BBPs.  $SN$  and  $SP$  primarily assess the ability of

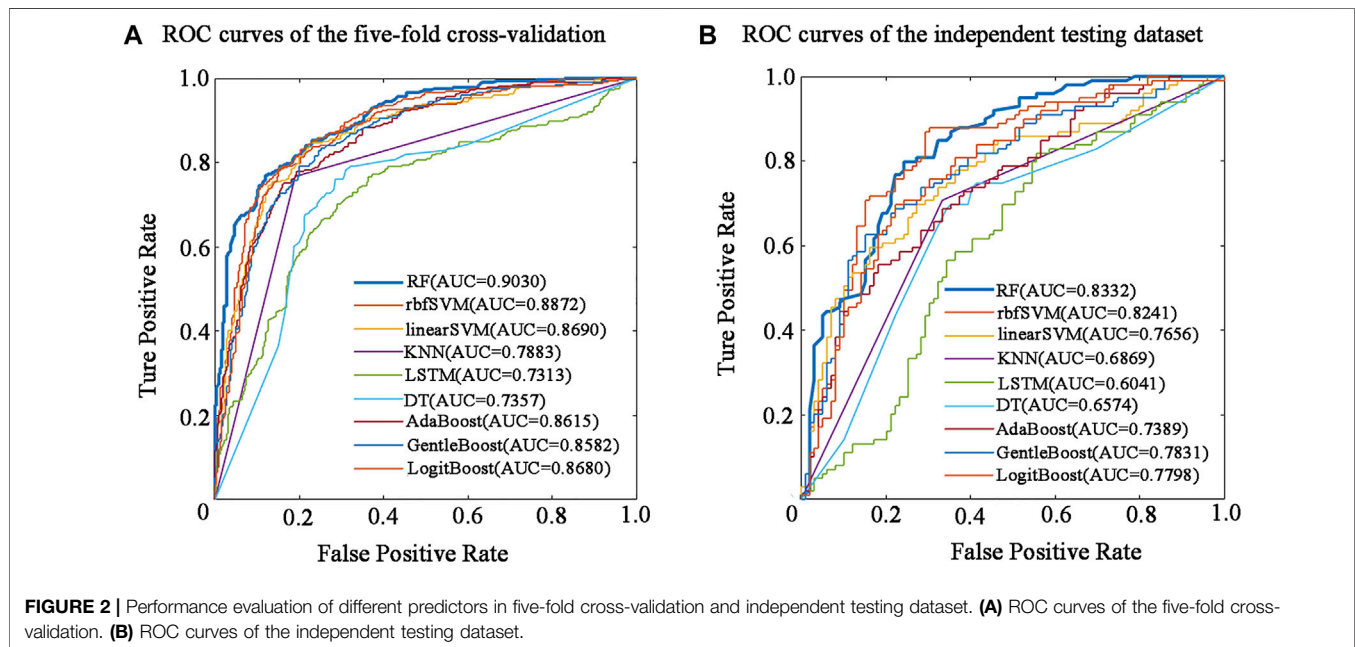
a predictive model to identify positive and negative samples respectively, while  $ACC$  and  $MCC$  investigate the comprehensive capacity of a prediction model to classify both positive and negative samples (Wang et al., 2019). The AUC score is often utilized to judge the merits and demerits of classifiers. In this study, we selected the optimal predictive model according to the AUC value. The model construction and evaluation were performed at a computational server (Sugon I840-G20, Dawning Information Industry Co., LTD., Beijing, China).

## 2.7 Reproducible Analysis

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

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

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



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

### 3 RESULT

#### 3.1 Overall Workflow

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

models was selected as the final model, which was implemented into a web server.

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

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



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

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

**TABLE 4 |** Comparison of datasets for three predictors.

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

### 3.3 Performance of Nine Classifiers on the Independent Testing Dataset

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

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

We also used the RF algorithm with optimal features selected by Pearson and Lasso feature scoring methods to construct prediction model. As shown in **Supplementary Tables S4,5**, the model under the combination of RF and F-score achieved the second highest AUC value in the nested five-fold cross-

**TABLE 5 |** The prediction performances of different predictors.

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


validation and the highest AUC value in the independent testing dataset. Therefore, we finally chose the combination of RF and F-score to build the final model based on 184 features and tree depth of 63.

### 3.5 Prediction Performance of Existing Predictors

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



A



[BBPpredict](#)
[Download](#)
[Citation](#)
[Help](#)

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

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


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

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

Or upload a sequence file:  No file chosen

Set the *tp*: 0.5

B



[BBPpredict](#)
[Download](#)
[Citation](#)
[Help](#)

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

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

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

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

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

### 3.6 Web Server Implementation

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

BBPpredict allows users to adjust the threshold of the probability value (*tp*) to distinguish between predicted positives and negatives, which can range from 0 to 1. As shown in **Table 6**,

with the increase of *tp*, the value of *SN* decreases, and the *SP* increases. When *tp* is 0.5, *ACC* achieves the highest score of 77.27%, *MCC* reaches the highest value of 0.5455.

## 4 DISCUSSION

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

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

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

We also reconstructed the BBPs/non-BBPs classification models with different machine learning methods using the new feature vectors that were generated from 16 feature extraction methods, including AAC, DPC, CKSAAGP, PAAC, GAAC, Grouped Di-Peptide Composition (GDPC) (Chen et al., 2018; Chen et al., 2020), Dipeptide Deviation from Expected Mean (DDE) (Chen et al., 2020), Composition (CTDC) (Dubchak et al., 1995; Dubchak et al., 1999; Chen et al., 2020), Transition (CTDT) (Dubchak et al., 1995; Dubchak et al., 1999; Chen et al., 2020), Distribution (CTDD) (Chen et al., 2020), Amphiphilic Pseudo-Amino Acid Composition (APAAC) (Chou, 2005; Jiao and Du, 2016), Quasi-sequence-order (QSOrder) (Chen et al., 2020), Normalized Moreau-Broto Autocorrelation (NMBroto) (Chen et al., 2018), Geary

correlation (Geary) (Chen et al., 2020), Moran correlation (Moran) (Feng and Zhang, 2000; Chen et al., 2020) and Sequence-Order-Coupling Number (SOCNumber) (Lim et al., 2015). The detailed description of the last 11 feature encoding approaches can be found in the **Supplementary Materials**. *F*-score was used for feature sorting, grid search with five-fold cross-validation was utilized to select the best feature parameters and the best classifier parameters for different classifiers. **Supplementary Tables S6,7** illustrated the detailed results of five-fold cross-validation and independent testing dataset of reconstructed classification models, respectively. However, the addition of feature encoding methods did not improve the classification performance of the model. We speculate that it is caused by the high correlation between the extracted features based on different feature extracting methods, which might induce highly correlated features in the final feature subset. As the feature number is limited, the highly correlated features might reduce useful information for model construction. Another possible reason might be the limited sample size, which might cause high false positive rate during the process of feature selection. The increase of feature size would lead to the increase of false positive features, which would affect the robustness of the predictive model.

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

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

## 5 CONCLUSION

In this study, we proposed an RF-based predictor for identifying BBPs, called BBPpredict, which is available for free at <http://i.uestc.edu.cn/BBPpredict/cgi-bin/BBPpredict.pl>. To find the optimal classifier, eight traditional machine learning algorithms and one deep learning algorithm were used for developing models. The RF algorithm was selected to construct BBPpredict after comparing the results of nine classifiers in the five-fold cross-validation and

independent test. The RF-based model reached an AUC of 0.9030 with an accuracy of 81.90% and an AUC of 0.8332 with an accuracy of 77.27% in the nested five-fold cross-validation and independent testing dataset, respectively. We also compared BBPpredict with two existing BBPs predictors, BBPpred and B3Pred. The results showed that BBPpredict was remarkably higher in accuracy, MCC, sensitivity and specificity than these two predictors. BBPpredict is a promising classification model, and we expect it to play a positive role in the discovery of BBPs to facilitate the development of drugs for CNS diseases.

## DATA AVAILABILITY STATEMENT

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

## AUTHOR CONTRIBUTIONS

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

conceived and designed the experiments, authored or reviewed drafts of the paper. All authors approved the final draft.

## FUNDING

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

## ACKNOWLEDGMENTS

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

## SUPPLEMENTARY MATERIAL

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

## REFERENCES

- Banks, W. A. (2016). From Blood-Brain Barrier to Blood-Brain Interface: New Opportunities for CNS Drug Delivery. *Nat. Rev. Drug Discov.* 15 (4), 275–292. doi:10.1038/nrd.2015.21
- Banks, W. A., and Kastin, A. J. (1996). Passage of Peptides across the Blood-Brain Barrier: Pathophysiological Perspectives. *Life Sci.* 59 (23), 1923–1943. doi:10.1016/s0024-3205(96)00380-3
- Bhadra, P., Yan, J., Li, J., Fong, S., and Siu, S. W. I. (2018). AmPEP: Sequence-Based Prediction of Antimicrobial Peptides Using Distribution Patterns of Amino Acid Properties and Random forest. *Sci. Rep.* 8 (1), 1697. doi:10.1038/s41598-018-19752-w
- Bhasin, M., and Raghava, G. P. S. (2004). Classification of Nuclear Receptors Based on Amino Acid Composition and Dipeptide Composition. *J. Biol. Chem.* 279 (22), 23262–23266. doi:10.1074/jbc.M401932200
- Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* 2 (3), 1–27. doi:10.1145/1961189.1961199
- Chen, K., Jiang, Y., Du, L., and Kurgan, L. (2009). Prediction of Integral Membrane Protein Type by Collocated Hydrophobic Amino Acid Pairs. *J. Comput. Chem.* 30 (1), 163–172. doi:10.1002/jcc.21053
- Chen, K., Kurgan, L. A., and Ruan, J. (2007b). Prediction of Flexible/rigid Regions from Protein Sequences Using K-Spaced Amino Acid Pairs. *BMC Struct. Biol.* 7, 25. doi:10.1186/1472-6807-7-25
- Chen, K., Kurgan, L. A., and Ruan, J. (2008). Prediction of Protein Structural Class Using Novel Evolutionary Collocation-Based Sequence Representation. *J. Comput. Chem.* 29 (10), 1596–1604. doi:10.1002/jcc.20918
- Chen, K., Kurgan, L., and Rahbari, M. (2007a). Prediction of Protein Crystallization Using Collocation of Amino Acid Pairs. *Biochem. Biophysical Res. Commun.* 355 (3), 764–769. doi:10.1016/j.bbrc.2007.02.040
- Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., et al. (2018). iFeature: a Python Package and Web Server for Features Extraction and Selection from Protein and Peptide Sequences. *Bioinformatics* 34 (14), 2499–2502. doi:10.1093/bioinformatics/bty140
- Chen, Z., Zhao, P., Li, F., Marquez-Lago, T. T., Leier, A., Revote, J., et al. (2020). iLearn: an Integrated Platform and Meta-Learner for Feature Engineering, Machine-Learning Analysis and Modeling of DNA, RNA and Protein Sequence Data. *Brief Bioinform.* 21 (3), 1047–1057. doi:10.1093/bib/bbz041
- Chou, K.-C. (2001). Prediction of Protein Cellular Attributes Using Pseudo-amino Acid Composition. *Proteins* 43 (3), 246–255. doi:10.1002/prot.1035
- Chou, K.-C. (2005). Using Amphiphilic Pseudo Amino Acid Composition to Predict Enzyme Subfamily Classes. *Bioinformatics* 21 (1), 10–19. doi:10.1093/bioinformatics/bth466
- Dai, R., Zhang, W., Tang, W., Wynendaele, E., Zhu, Q., Bin, Y., et al. (2021). BBPpred: Sequence-Based Prediction of Blood-Brain Barrier Peptides with Feature Representation Learning and Logistic Regression. *J. Chem. Inf. Model.* 61 (1), 525–534. doi:10.1021/acs.jcim.0c01115
- Drappatz, J., Brenner, A., Wong, E. T., Eichler, A., Schiff, D., Groves, M. D., et al. (2013). Phase I Study of GRN1005 in Recurrent Malignant Glioma. *Clin. Cancer Res.* 19 (6), 1567–1576. doi:10.1158/1078-0432.Ccr-12-2481
- Dubchak, I., Muchnik, I., Holbrook, S. R., and Kim, S. H. (1995). Prediction of Protein Folding Class Using Global Description of Amino Acid Sequence. *Proc. Natl. Acad. Sci. U.S.A.* 92 (19), 8700–8704. doi:10.1073/pnas.92.19.8700
- Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I., and Kim, S.-H. (1999). Recognition of a Protein Fold in the Context of the SCOP Classification. *Proteins* 35 (4), 401–407. doi:10.1002/(sici)1097-0134(19990601)35:4<401::aid-prot3>3.0.co;2-k
- Feng, Z.-P., and Zhang, C.-T. (2000). Prediction of Membrane Protein Types Based on the Hydrophobic index of Amino Acids. *J. Protein Chem.* 19 (4), 269–275. doi:10.1023/a:1007091128394
- He, B., Chen, H., and Huang, J. (2019). PhD7Faster 2.0: Predicting Clones Propagating Faster from the Ph.D.-7 Phage Display Library by Coupling PseAAC and Tripeptide Composition. *PeerJ* 7, e7131. doi:10.7717/peerj.7131
- He, B., Kang, J., Ru, B., Ding, H., Zhou, P., and Huang, J. (2016). SABinder: A Web Service for Predicting Streptavidin-Binding Peptides. *Biomed. Res. Int.* 2016, 1–8. doi:10.1155/2016/9175143
- Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.* 9 (8), 1735–1780. doi:10.1162/neco.1997.9.8.1735

- Jiao, Y.-S., and Du, P.-F. (2016). Predicting Golgi-Resident Protein Types Using Pseudo Amino Acid Compositions: Approaches with Positional Specific Physicochemical Properties. *J. Theor. Biol.* 391, 35–42. doi:10.1016/j.jtbi.2015.11.009
- Kumar, V., Agrawal, P., Kumar, R., Bhalla, S., Usmani, S. S., Varshney, G. C., et al. (2018). Prediction of Cell-Penetrating Potential of Modified Peptides Containing Natural and Chemically Modified Residues. *Front. Microbiol.* 9, 725. doi:10.3389/fmicb.2018.00725
- Kumar, V., Patiyl, S., Dhall, A., Sharma, N., and Raghava, G. P. S. (2021a). B3Pred: A Random-Forest-Based Method for Predicting and Designing Blood-Brain Barrier Penetrating Peptides. *Pharmaceutics* 13 (8), 1237. doi:10.3390/pharmaceutics13081237
- Kumar, V., Patiyl, S., Kumar, R., Sahai, S., Kaur, D., Lathwal, A., et al. (2021b). B3Pdb: an Archive of Blood-Brain Barrier-Penetrating Peptides. *Brain Struct. Funct.* 226 (8), 2489–2495. doi:10.1007/s00429-021-02341-5
- Kurzrock, R., Gabrail, N., Chandhasin, C., Moulder, S., Smith, C., Brenner, A., et al. (2012). Safety, Pharmacokinetics, and Activity of GRN1005, a Novel Conjugate of Angiopep-2, a Peptide Facilitating Brain Penetration, and Paclitaxel, in Patients with Advanced Solid Tumors. *Mol. Cancer Ther.* 11 (2), 308–316. doi:10.1158/1535-7163.Mct-11-0566
- Lee, T.-Y., Lin, Z.-Q., Hsieh, S.-J., Bretaña, N. A., and Lu, C.-T. (2011). Exploiting Maximal Dependence Decomposition to Identify Conserved Motifs from a Group of Aligned Signal Sequences. *Bioinformatics* 27 (13), 1780–1787. doi:10.1093/bioinformatics/btr291
- Li, F.-M., and Wang, X.-Q. (2016). Identifying Anticancer Peptides by Using Improved Hybrid Compositions. *Sci. Rep.* 6, 33910. doi:10.1038/srep33910
- Lim, S., Kim, W.-J., Kim, Y.-H., Lee, S., Koo, J.-H., Lee, J.-A., et al. (2015). dNP2 Is a Blood-Brain Barrier-Permeable Peptide Enabling ctCTLA-4 Protein Delivery to Ameliorate Experimental Autoimmune Encephalomyelitis. *Nat. Commun.* 6, 8244. doi:10.1038/ncomms9244
- Muttenthaler, M., King, G. F., Adams, D. J., and Alewood, P. F. (2021). Trends in Peptide Drug Discovery. *Nat. Rev. Drug Discov.* 20 (4), 309–325. doi:10.1038/s41573-020-00135-8
- Nance, E., Pun, S. H., Saigal, R., and Sellers, D. L. (2022). Drug Delivery to the central Nervous System. *Nat. Rev. Mater* 7 (4), 314–331. doi:10.1038/s41578-021-00394-w
- Nonaka, M., Suzuki-Anekoji, M., Nakayama, J., Mabashi-Asazuma, H., Jarvis, D. L., Yeh, J.-C., et al. (2020). Overcoming the Blood-Brain Barrier by Annexin A1-Binding Peptide to Target Brain Tumours. *Br. J. Cancer* 123 (11), 1633–1643. doi:10.1038/s41416-020-01066-2
- Oller-Salvia, B., Sánchez-Navarro, M., Giralte, E., and Teixidó, M. (2016). Blood-brain Barrier Shuttle Peptides: an Emerging Paradigm for Brain Delivery. *Chem. Soc. Rev.* 45 (17), 4690–4707. doi:10.1039/c6cs00076b
- Sánchez-Navarro, M., Giralte, E., and Teixidó, M. (2017). Blood-brain Barrier Peptide Shuttles. *Curr. Opin. Chem. Biol.* 38, 134–140. doi:10.1016/j.cbpa.2017.04.019
- Saravanan, V., and Gautham, N. (2015). Harnessing Computational Biology for Exact Linear B-Cell Epitope Prediction: A Novel Amino Acid Composition-Based Feature Descriptor. *OMICS: A J. Integr. Biol.* 19 (10), 648–658. doi:10.1089/omi.2015.0095
- Terstappen, G. C., Meyer, A. H., Bell, R. D., and Zhang, W. (2021). Strategies for Delivering Therapeutics across the Blood-Brain Barrier. *Nat. Rev. Drug Discov.* 20 (5), 362–383. doi:10.1038/s41573-021-00139-y
- Van Dorpe, S., Bronselaer, A., Nielandt, J., Stalmans, S., Wynendaele, E., Audenaert, K., et al. (2012). Brainpeps: the Blood-Brain Barrier Peptide Database. *Brain Struct. Funct.* 217 (3), 687–718. doi:10.1007/s00429-011-0375-0
- Walzer, M., and Vizcaíno, J. A. (2020). Review of Issues and Solutions to Data Analysis Reproducibility and Data Quality in Clinical Proteomics. *Methods Mol. Biol.* 2051, 345–371. doi:10.1007/978-1-4939-9744-2\_15
- Wang, J., Li, J., Yang, B., Xie, R., Marquez-Lago, T. T., Leier, A., et al. (2019). Bastion3: a Two-Layer Ensemble Predictor of Type III Secreted Effectors. *Bioinformatics* 35 (12), 2017–2028. doi:10.1093/bioinformatics/bty914
- Wei, L., Tang, J., and Zou, Q. (2017a). SkipCPP-Pred: an Improved and Promising Sequence-Based Predictor for Predicting Cell-Penetrating Peptides. *BMC Genomics* 18 (Suppl. 7), 742. doi:10.1186/s12864-017-4128-1
- Wei, L., Xing, P., Su, R., Shi, G., Ma, Z. S., and Zou, Q. (2017b). CPPred-RF: A Sequence-Based Predictor for Identifying Cell-Penetrating Peptides and Their Uptake Efficiency. *J. Proteome Res.* 16 (5), 2044–2053. doi:10.1021/acs.jproteome.7b00019
- Xie, R., Wu, Z., Zeng, F., Cai, H., Wang, D., Gu, L., et al. (2021). Retro-enantio Isomer of Angiopep-2 Assists Nanoprobes across the Blood-Brain Barrier for Targeted Magnetic Resonance/fluorescence Imaging of Glioblastoma. *Sig Transduct Target. Ther.* 6 (1), 309. doi:10.1038/s41392-021-00724-y
- Zhao, Y.-W., Lai, H.-Y., Tang, H., Chen, W., and Lin, H. (2016). Prediction of Phosphothreonine Sites in Human Proteins by Fusing Different Features. *Sci. Rep.* 6, 34817. doi:10.1038/srep34817
- Zhou, X., Smith, Q. R., and Liu, X. (2021). Brain Penetrating Peptides and Peptide-Drug Conjugates to Overcome the Blood-Brain Barrier and Target CNS Diseases. *WIREs Nanomed Nanobiotechnol* 13 (4), e1695. doi:10.1002/wnan.1695

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

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

Copyright © 2022 Chen, Zhang, Li, Lu, Yang, Long, He, Chen and Huang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# ProtTrans-Glutar: Incorporating Features From Pre-trained Transformer-Based Models for Predicting Glutarylation Sites

Fatma Indriani<sup>1,2\*</sup>, Kunti Robiatul Mahmudah<sup>3</sup>, Bedy Purnama<sup>4</sup> and Kenji Satou<sup>5</sup>

<sup>1</sup>Graduate School of Natural Science and Technology, Kanazawa University, Kanazawa, Japan, <sup>2</sup>Department of Computer Science, Lambung Mangkurat University, Banjarmasin, Indonesia, <sup>3</sup>Department of Postgraduate of Mathematics Education, Universitas Ahmad Dahlan, Yogyakarta, Indonesia, <sup>4</sup>School of Computing, Telkom University, Bandung, Indonesia, <sup>5</sup>Institute of Science and Engineering, Kanazawa University, Kanazawa, Japan

## OPEN ACCESS

### Edited by:

Ruiquan Ge,  
Hangzhou Dianzi University, China

### Reviewed by:

Hao Lin,  
University of Electronic Science and  
Technology of China, China  
Trinh Trung Duong Nguyen,  
University of Copenhagen, Denmark

### \*Correspondence:

Fatma Indriani  
f.indriani@gmail.com

### Specialty section:

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

Received: 28 February 2022

Accepted: 26 April 2022

Published: 31 May 2022

### Citation:

Indriani F, Mahmudah KR, Purnama B  
and Satou K (2022) ProtTrans-Glutar:  
Incorporating Features From Pre-  
trained Transformer-Based Models for  
Predicting Glutarylation Sites.  
Front. Genet. 13:885929.  
doi: 10.3389/fgene.2022.885929

Lysine glutarylation is a post-translational modification (PTM) that plays a regulatory role in various physiological and biological processes. Identifying glutarylated peptides using proteomic techniques is expensive and time-consuming. Therefore, developing computational models and predictors can prove useful for rapid identification of glutarylation. In this study, we propose a model called ProtTrans-Glutar to classify a protein sequence into positive or negative glutarylation site by combining traditional sequence-based features with features derived from a pre-trained transformer-based protein model. The features of the model were constructed by combining several feature sets, namely the distribution feature (from composition/transition/distribution encoding), enhanced amino acid composition (EAAC), and features derived from the ProtT5-XL-UniRef50 model. Combined with random under-sampling and XGBoost classification method, our model obtained recall, specificity, and AUC scores of 0.7864, 0.6286, and 0.7075 respectively on an independent test set. The recall and AUC scores were notably higher than those of the previous glutarylation prediction models using the same dataset. This high recall score suggests that our method has the potential to identify new glutarylation sites and facilitate further research on the glutarylation process.

**Keywords:** lysine glutarylation, protein sequence, transformer-based models, protein embedding, machine learning, binary classification, imbalanced data classification, post-translation modification

## 1 INTRODUCTION

Similar to the epigenetic modification of histones and nucleic acids, the post-translational modification (PTM) of amino acids dynamically changes the function of proteins and is actively studied in the field of molecular biology. Among various kinds of PTMs, lysine glutarylation is defined as an attachment of a glutaryl group to a lysine residue of a protein (Lee et al., 2014). This modification was first detected *via* immunoblotting and mass spectrometry analysis and later validated using chemical and biochemical methods. It is suggested that this PTM may be a biomarker of aging and cellular stress (Harmel and Fiedler, 2018). Dysregulation of glutarylation is related to some metabolic diseases, including type 1 glutaric aciduria, diabetes, cancer, and neurodegenerative diseases (Tan et al., 2014; Osborne et al., 2016; Carrico et al., 2018). Since the identification of



glutarylated peptides using proteomics techniques is expensive and time-consuming, it is important to investigate computational models and predictors to rapidly identify glutarylation.

Based on a survey of previous research, various prediction models have been proposed to distinguish glutarylation sites. The earliest one, GlutPred (Ju and He, 2018), constructs features from amino acid factors (AAF), binary encoding (BE), and the composition of k-spaced amino acid pairs (CKSAAP). The authors selected 300 features using the mRMR method. To overcome the problem of imbalance in this dataset, a biased version of support vector machine (SVM) was employed to build the prediction model. Another predictor, iGlu-Lys (Xu et al., 2018), investigated four different feature sets, physicochemical properties (AAIndex), K-Space, Position-Special Amino Acid Propensity (PSAAP), and Position-Specific Propensity Matrix (PSPM), in conjunction with SVM classifier. The feature set PSPM performed best in the 10-fold cross-validation and was therefore applied to the model. iGlu-Lys performed better than GlutPred in terms of accuracy and specificity scores. However, their sensitivity scores were lower. The next model proposed, MDDGlutar (Huang et al., 2019), divided the training set into six subsets using maximal dependence decomposition (MDD). Three feature sets were evaluated separately using SVM: amino acid composition (AAC), amino acid pair composition (AAPC), and CKSAAP. The best cross-validation score was the AAC feature set. The results of independent testing yielded a balanced score of 65.2% sensitivity and 79.3% specificity, but it had lower specificity and accuracy than those of the GlutPred model.

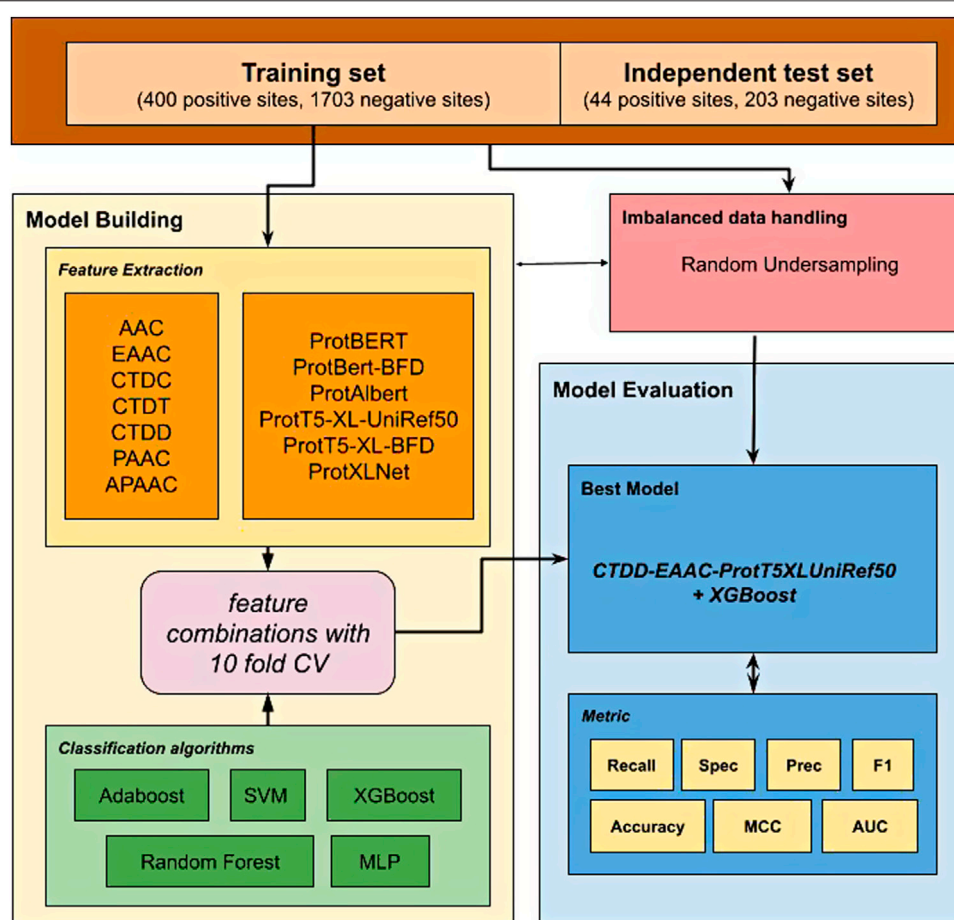
The next two predictors included the addition of new glutarylated proteins from *Escherichia coli* and HeLa cells for their training and test sets. RF-GlutarySite (Al-barakati et al., 2019) utilizes features constructed from 14 feature sets, reduced with XGBoost. The model's reported performance for independent testing was balanced, with 71.3% accuracy, 74.1% sensitivity, and 68.5% specificity. However, it is interesting to note that the test data was balanced by under-sampling, which did not represent a real-world scenario. iGlu\_Adaboost (Dou et al., 2021) sought to fill this gap by using test data with no resampling. This model utilizes features from 188D, enhanced amino acid composition (EAAC), and CKSAAP. With the help of Chi2 feature selection, 37 features were selected to build the model using SMOTE-Tomek re-sampling and the Adaboost classifier. The test result had good performance for recall, specificity, and accuracy metrics, but a lower Area Under the Curve (AUC) score than that of previous models.

Although many models have been built to distinguish between positive and negative glutarylation sites, the performance of these methods remains limited. One challenge to this problem is finding a set of features to represent the protein subsequence, which enables a correct classification of glutarylation site. BERT models (Devlin et al., 2019), and other transformer-based language models from natural language processing (NLP) research, show excellent performance for NLP tasks. These language models, having been adapted to biological sequences by treating them as sentences and then trained using large-scale

protein corpora (Elnaggar et al., 2021), also show promise for various machine learning tasks in the bioinformatics domain.

Previous studies have investigated the use of pre-trained language models from BERT and BERT-like models to show its effectiveness as protein sequence representation for protein classification. For example, Ho et al. (2021) proposed a new approach to predict flavin adenine dinucleotide (FAD) binding sites from transport proteins based on pre-training BERT, position-specific scoring matrix profiles (PSSM), and an amino acid index database (AAIndex). Their approach showed an accuracy score of 85.14%, which is an improvement over the scores of the previous methods. Another study (Shah et al., 2021) extracted features using pre-trained BERT models to discriminate between three families of glucose transporters. This method, compared to two well-known feature extraction methods, AAC and DPC, showed an improved performance of more than 4% in average sensitivity and Matthews correlation coefficient (MCC). In another study, Liu built a predictor for protein lysine glycation sites using features extracted from pre-trained BERT models, which showed improved performance in terms of accuracy and AUC score compared to previous methods (Liu et al., 2022). These studies demonstrate the suitability of utilizing BERT models to improve various protein classification tasks. Therefore, using embeddings from pre-trained BERT and BERT-like models has the potential to build an improved glutarylation prediction model.

In this study, we proposed a new prediction model to predict glutarylation sites (Figure 1) by incorporating features extracted from pre-trained protein models combined with features from handcrafted sequence-based features. A public dataset provided from Al-barakati et al. (2019) was used in this study. It was an imbalanced dataset with 444 positive sites and 1906 negative sites, and already separated into two sets for use in model building and independent testing. First, various feature sets were extracted from the dataset, consisting of two types of features. The first type consists of seven classic sequence-based features, and the second type consists of six embeddings from pre-trained protein language models. We evaluated the classifiers using a 10-fold cross-validation for the individual feature set. The next step was to combine two or more feature sets to evaluate further models, such as AAC-EAAC, AAC-CTDC, and AAC-ProtBert. For this, we limited the embedding features to a maximum of one in the combination. Five classification algorithms were included in the experiments: Adaboost, XGBoost, SVM (with RBF kernel), random forest (RF), and multilayer perceptron (MLP). Our best model combines the features of CTDD, AAC, and ProtT5-XL-UniRef50 with the XGBoost classification algorithm. This model, with the model of the best feature set from sequence-based feature groups and the model of the best feature set from the protein embedding feature group, was then evaluated with an independent dataset. For independent testing, the entire training set was used to develop a model. In both model building and independent testing, a random under-sampling method was used to balance the training dataset, while the testing dataset was not resampled to reflect performance in the real-world unbalanced scenario.



**FIGURE 1 |** Workflow strategy for the development of ProTrans-Glutar model.

**TABLE 1 |** Number of positive and negative sites in training and test set.

	Training set	Test set	
Positive sites	400	44	444
Negative sites	1703	203	1906
	2103	247	

## 2 MATERIALS AND METHODS

### 2.1 Dataset

This study utilized unbalanced benchmark datasets compiled by Al-barakati et al. (2019) to build their predictor, RF-GlutarySite. This dataset collected positive glutarylation sites from various sources, including PLMD (Xu et al., 2017) and (Tan et al., 2014) and consisted of four different species (*Mus musculus*, *Mycobacterium tuberculosis*, *E. coli*, and HeLa cells), for a total of 749 sites from 234 proteins. Homologous sequences that showed  $\geq 40\%$  sequence identity were removed using the CD-HIT tool. The remaining proteins were converted into peptides with a fixed length of 23, with glutarylated lysine as the central residue, and 11 residues each upstream and downstream.

Negative sites were generated in the same way, but the central lysine residue was not glutarylated. After removing homologous sequences, the final dataset consisted of 453 positive and 2043 negative sites. The distributions of the training and testing datasets are listed in **Table 1**. This dataset was also used by Dou et al. (2021) to build the proposed predictor model iGlu\_Adaboost (Dou et al., 2021).

### 2.2 Feature Extraction

The extraction of numerical features from protein sequences or peptides is an important step before they can be utilized by machine learning algorithms. In this study, we investigated two types of features: classic sequence-based features and features derived from pre-trained transformer-based protein embeddings. Classic sequence-based features were extracted using the *iFeature* Python package (Chen et al., 2018). After preliminary experiments, seven feature groups were chosen for further investigation: AAC, EAAC, Composition/Transition/Distribution (CTD), pseudo-amino acid composition (PAAC), and amphiphilic pseudo-amino acid composition (APAAC). The second type of feature, embeddings from pre-trained transformer-based models, was extracted using models trained

**TABLE 2 |** Physicochemical attributes and its division of the amino acids.

Attribute	Division		
Hydrophobicity_PRAM900101	Polar: RKEDQN	Neutral: GASTPHY	Hydrophobicity: CLVIMFW
Hydrophobicity_ARGP820101	Polar: QSTNGDE	Neutral: RAHCKMV	Hydrophobicity: LYPFIW
Hydrophobicity_ZIMJ680101	Polar: QNGSWTDERA	Neutral: HMCKV	Hydrophobicity: LPFYI
Hydrophobicity_PONP930101	Polar: KPDESNQT	Neutral: GRHA	Hydrophobicity: YMFVLCVI
Hydrophobicity_CASG920101	Polar: KDEQPSRNTG	Neutral: AHYMLV	Hydrophobicity: FIWC
Hydrophobicity_ENGD860101	Polar: RDKENQHYP	Neutral: SGTAW	Hydrophobicity: CVLIMF
Hydrophobicity_FASG890101	Polar: KERSQD	Neutral: NTPG	Hydrophobicity: AYHWMFLIC
Normalized van der Waals volume	Volume range: 0–2.78	Volume range: 2.95–94.0	Volume range: 4.03–8.08
	GASTPD	NVEQIL	MHKFRYW
Polarity	Polarity value: 4.9–6.2	Polarity value: 8.0–9.2	Polarity value: 10.4–13.0
	LIFWCMVY	PATGS	HQRKNED
Polarizability	Polarizability value: 0–1.08	Polarizability value: 0.128–120.186	Polarizability value: 0.219–0.409
	GASDT	GPNVEQIL	KMHFRYW
Charge	Positive: KR	Neutral: ANCQGHILMFSTWYV	Negative: DE
Secondary structure	Helix: EALMQKRH	Strand: VIYCWFT	Coil: GNPSD
Solvent accessibility	Buried: ALFCGIWW	Exposed: PKQEND	Intermediate: MPSTHY

and provided by Elnaggar et al. (2021). It consists of six feature sets from six protein models: ProtBERT, ProtBert-BFD, ProtAlberty, ProtT5-XL-UniRef50, ProtT5-XL-BFD, and ProtXLNet. The data for all extracted features are provided in the **Supplementary Material**.

### 2.2.1 Amino Acid Composition and Enhanced Amino Acid Composition

The AAC method encodes a protein sequence-based on the frequency of each amino acid (Bhasin and Raghava, 2004). For this type of feature, we used two variants.

The first variant is the basic AAC, in which the protein sequence is converted into a vector of length 20, representing the frequency of the 20 amino acids (“ACDEFGHIKLMNPQRSTVWY”). Each element is calculated according to Eq. 1, as follows:

$$f(t) = \frac{N(t)}{N} \quad (1)$$

where  $t$  is the amino acid type,  $N(t)$  is the total number of amino acids  $t$  appearing in the sequence, and  $N$  is the length of the sequence.

The second variant is EAAC, introduced by Chen et al. (2018). In this encoding, the EAAC was calculated using sliding windows, that is, from a fixed window size, moving from left to right. To calculate the frequency of each amino acid in each window, see Eq. 2:

$$f(t, win) = \frac{N(t, win)}{N(win)} \quad (2)$$

where  $N(t, win)$  represents the number of amino acids  $t$  that appear in the window  $win$  and  $N(win)$  represents the length of the window. To develop our model, a default window size of five was used. How these methods are applied to a protein sequence are provided in **Supplementary File S1**.

### 2.2.2 Composition/Transition/Distribution

The CTD method encodes a protein sequence-based on various structural and physicochemical properties (Dubchak et al.,

1995; Cai, 2003). Thirteen properties were used to build the features. Each property was divided into three groups (see **Table 2**). For example, the attribute “Hydrophobicity\_PRAM900101” divides the amino acids into polar, neutral, and hydrophobic groups.

The CTD feature comprises three parts: composition (CTDC), transition (CTDT), and distribution (CTDD). For composition, an attribute contributes to three values, representing the global distribution (frequency) of the amino acids in each of the three groups of attributes. The composition is computed as follows:

$$C(r) = \frac{N(r)}{N} \quad (3)$$

where  $N(r)$  is the number of occurrences of type  $r$  amino acids in the sequence and  $N$  is the length of the sequence.

For transition, an attribute also contributes to three values, each representing the number of transitions between any pair of groups. The transition is calculated as follows:

$$T(r, s) = \frac{N(r, s) + N(s, r)}{N - 1} \quad (4)$$

where  $N(r, s)$  represents the number of occurrences amino acid type  $r$  transit to type  $s$  (i.e., it appeared as “rs” in the sequence), and  $N$  is the length of the sequence. Similarly,  $N(s, r)$  is the reverse, that is, the number of “sr” occurrences in the sequence.

The distribution feature consists of five values per attribute group, each of which corresponds to the fraction of the sequence length at five different positions in the group: first occurrence, 25%, 50%, 75%, and 100%.

### 2.2.3 Pseudo Amino Acid Composition

Pseudo amino acid composition feature was proposed by Chou (2001). For protein sequence  $P$  with  $L$  amino acid residues  $P = (R_1R_2R_3 \dots R_L)$ , the PAAC features can be formulated as

$$P = [P_1, P_2, \dots, P_{20}, P_{20+1}, \dots, P_{20+\lambda}]^T, (\lambda < L) \quad (5)$$

where

$$p_u = \begin{cases} \frac{f_u}{\sum_{i=1}^{20} f_i + w \sum_{k=1}^{\lambda} \tau_k}, & (1 \leq u \leq 20) \\ \frac{w \tau_{u-20}}{\sum_{i=1}^{20} f_i + w \sum_{k=1}^{\lambda} \tau_k}, & (20 + 1 \leq u \leq 20 + \lambda) \end{cases} \quad (6)$$

$w$  is the weight factor and  $\tau_k$  is the  $k$ -th tier correlation factor, defined as

$$\tau_k = \frac{1}{L-k} \sum_{i=1}^{L-K} J_{i,i+k}, \quad (k < L) \quad (7)$$

and

$$J_{i,i+k} = \frac{1}{\Gamma} \sum_{q=1}^{\Gamma} [\Phi_q R_{i+k} - \Phi_q R_i]^2 \quad (8)$$

where  $\Phi_q(R_i)$  is the  $q$ -th function of the amino acid  $R_i$ , and  $\Gamma$  the total number of functions. In here  $\Gamma = 3$  and the functions used are hydrophobicity value, hydrophilicity value, and side chain mass of amino acid  $R_i$ .

A variant of PAAC called amphiphilic pseudo amino acid composition (APAAC) proposed in Chou (2005). A protein sample  $P$  with  $L$  amino acid residues  $P = (R_1 R_2 R_3 \dots R_L)$ , is formulated as

$$P = [p_1, p_2, \dots, p_{20}, p_{20+1}, \dots, p_{20+\lambda}, p_{20+\lambda}, \dots, p_{2\lambda}]^T, \quad (\lambda < L) \quad (9)$$

where

$$p_u = \begin{cases} \frac{f_u}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{2\lambda} \tau_j}, & (1 \leq u \leq 20) \\ \frac{w \tau_{u-20}}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{2\lambda} \tau_j}, & (20 + 1 \leq u \leq 20 + 2\lambda) \end{cases} \quad (10)$$

$\tau_j$  is the  $j$ -tier sequence-correlation factor calculated using the equations:

$$\left\{ \begin{array}{l} \tau_1 = \frac{1}{L-1} \sum_{i=1}^{L-1} H_{i,i+1}^1 \\ \tau_2 = \frac{1}{L-1} \sum_{i=1}^{L-1} H_{i,i+1}^2 \\ \tau_3 = \frac{1}{L-2} \sum_{i=1}^{L-2} H_{i,i+2}^1 \\ \tau_4 = \frac{1}{L-2} \sum_{i=1}^{L-2} H_{i,i+2}^2, \lambda < L \\ \dots \\ \tau_{2\lambda-1} = \frac{1}{L-1} \sum_{i=1}^{L-\lambda} H_{i,i+\lambda}^1 \\ \tau_{2\lambda} = \frac{1}{L-1} \sum_{i=1}^{L-\lambda} H_{i,i+\lambda}^2 \end{array} \right. \quad (11)$$

where  $H_{i,j}^1$  and  $H_{i,j}^2$  are hydrophobicity and hydrophilicity values of the  $i$ -th amino acid, described by the following equation:

$$\begin{aligned} H_{i,j}^1 &= h^1(R_i) \cdot h^1(R_j) \\ H_{i,j}^2 &= h^2(R_i) \cdot h^2(R_j) \end{aligned} \quad (12)$$

## 2.2.4 Pre-Trained Transformer Protein Embeddings

Protein language models has been trained from large protein corpora, using the state-of-the-art transformer models from the latest NLP research (Elnaggar et al., 2021). Six of the models were applied to extract features for our task of predicting glutarylation sites.

- ProtBERT and ProtBert-BFD are derived from the BERT model (Devlin et al., 2019), trained on UniRef100 and BFD corpora, respectively.
- ProtT5-XL-UniRef50 and ProtT5-XL-BFD are derived from the T5 model (Raffel et al., 2020), trained on UniRef50 and BFD corpora, respectively.
- ProtAlbert is derived from the Albert model (Lan et al., 2020) trained on UniRef100 corpora.
- ProtXLNet is derived from the XLNet model (Yang et al., 2020), trained on UniRef100 corpora.

Protein embeddings (features) were extracted from the last layer of this protein language model to be used for subsequent supervised training. This layer is a 2-dimensional array with a size of  $1024 \times \text{length of sequence}$ , except for the ProtAlbert model with an array size of  $4096 \times \text{length of sequence}$ . For the glutarylation prediction problem, this feature is simplified by summing the vectors along the length of the sequence; hence, each feature group is now one-dimensional, with a length of 4,096 for ProtAlbert and 1,024 for the rest.

## 2.2.5 The Feature Space

The features collected were of different lengths, as summarized in Table 3. These feature groups are evaluated either individually or using various combinations of two or more feature groups. As an example, for the combined feature group AAC-EAAC, a training sample will have  $20 + 380 = 400$ -dimensional features.

## 2.3 Imbalanced Data Handling

A class imbalance occurs when the number of samples is unevenly distributed. The class with a higher number of samples is called the majority class or the negative class, whereas the class with a smaller number is called the minority class. In the glutarylation dataset, the number of negative samples was nearly four times that of positive samples. This imbalance may affect the performance of classifiers because they are more likely to predict a positive sample as a negative sample (He and Garcia, 2009). A common strategy to solve this problem is by data re-sampling, either adding minority samples (over-sampling) or reducing majority samples (under-sampling). In this study, we implemented a random under-sampling strategy (He and Ma, 2013) after preliminary experiments with various re-sampling methods.

## 2.4 Machine Learning Methods

In this study, we used the XGBoost classifier (Chen and Guestrin, 2016) from the XGBoost package on the Python language platform (<https://xgboost.ai>). This is an implementation of a gradient-boosted tree classifier (Friedman, 2001). Gradient-

**TABLE 3 |** Features investigated for method development.

Group	Feature set	Length of features
Amino acid composition	AAC	20
	EAAC	380
C/T/D	CTDC	39
	CTDT	39
	CTDD	195
	PAAC	35
Pseudo amino acid composition	APAAC	50
Embeddings from pretrained transformer-based model	ProtBERT	1,024
	ProtBert-BFD	1,024
	ProtAlbert	4,096
	ProtT5-XL-UniRef50	1,024
	ProtT5-XL-BFD	1,024
	ProtXLNet	1,024

boosted trees are an ensemble classifier built from multiple decision trees, constructed one by one. XGBoost has been successfully used in various classification tasks, including bioinformatics (Mahmud et al., 2019; Chien et al., 2020; Zhang et al., 2020). In our experiments, several other popular classifiers are also compared and evaluated, including SVM, RF, MLP, and Adaboost, provided by the scikit-learn package (<https://scikit-learn.org>).

## 2.5 Model Evaluation

To achieve the model with the best prediction performance, the model was evaluated using 10-fold cross-validation and an independent test. For cross-validation, the training dataset was randomly split into 10 folds of nearly equal size. Nine folds were combined and then randomly under-sampled for training, and the 10th fold was used for evaluation. This process was performed with the other combination of folds (nine for training and one for testing). To remove sampling bias, the cross-validation process was repeated three times, and the mean performance was reported as the CV result. For independent testing, the entire training data were randomly under-sampled, then used to build the model, and later evaluated using the independent test set. Since the randomness in the under-sampling may affect to the performance result, this testing was repeated five times, and the mean performance was reported as an independent test result.

The performance of the cross-validation and independent test results was evaluated using seven performance metrics: recall (Rec), specificity (Spe), precision (Pre), accuracy (Acc), MCC, F1-score (F1), and area under the ROC curve (AUC). These metrics were calculated as follows:

$$\begin{aligned}\text{Rec} &= \frac{TP}{TP + FN} \\ \text{Spe} &= \frac{TN}{TN + FP} \\ \text{Pre} &= \frac{TP}{TP + FP} \\ \text{Acc} &= \frac{TP + TN}{TP + TN + FP + FN}\end{aligned}$$

$$\begin{aligned}\text{MCC} &= \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \\ \text{F1} &= 2 \times \frac{\text{Rec} \cdot \text{Pre}}{\text{Rec} + \text{Pre}}\end{aligned}\quad (13)$$

where *TP* is True Positive, *TN* is True Negative, *FP* is False Positive, and *FN* is False Negative.

The AUC metric is obtained by plotting recall against (1—specificity) for every threshold and then calculating the area under the curve.

## 3 RESULTS

### 3.1 Models Based on Sequence-Based Feature Set

We calculated the cross-validation performance for each sequence-based feature set using five supervised classifiers: AdaBoost, MLP, RF, SVM, and XGBoost. The performances of these classifiers are shown in **Table 4**. It can be observed that no classifier is the best for all feature groups. For example, using AAC features, MLP performs the best based on the AUC score. However, using EAAC features, the RF model has the best performance, whereas MLP has the poorest. Among the six different feature sets, the best model achieved was using EAAC features combined with RF, with an AUC score of 0.6999. This model also had the best specificity, precision, and accuracy compared to the other models.

### 3.2 Models Based on Embeddings From Pre-trained Transformer Models

Based on the embeddings extracted from the pre-trained transformer models, we evaluated the same five supervised classifiers. The performance results of the models are presented in **Table 5**. The combination of the ProtBERT model and SVM can match the recall score with the classic sequence-based feature result. However, all other metrics were lower. In this experiment, the best model with respect to the AUC score was a combination of features from the ProtAlbert model and SVM classifier (AUC = 0.6744). This model also had the



**TABLE 4** | Cross validation result of models from sequence-based features.

Feature groups	Classifier	Rec	Spe	Pre	Acc	MCC	F1	AUC
AAC	Adaboost	0.6120	0.6013	0.2654	0.6033	0.1690	0.3700	0.6433
	MLP	0.6520	0.6192	0.2864	0.6255	0.2150	0.3977	0.6864
	Random Forest	0.6190	0.5809	0.2575	0.5881	0.1576	0.3635	0.6378
	SVM	0.6395	0.5969	0.2714	0.6050	0.1868	0.3808	0.6651
	XGBoost	0.5917	0.5482	0.2353	0.5565	0.1102	0.3362	0.6101
EAAC	Adaboost	0.5983	0.6015	0.2608	0.6009	0.1584	0.3629	0.6384
	MLP	0.5850	0.5946	0.2530	0.5928	0.1422	0.3529	0.6323
	Random Forest	0.6450	0.6598	0.3089	0.6570	0.2450	0.4171	0.6999
	SVM	0.5967	0.6434	0.2821	0.6345	0.1923	0.3827	0.6571
	XGBoost	0.6408	0.6385	0.2945	0.6389	0.2230	0.4030	0.6834
CTDC	Adaboost	0.7050	0.5518	0.2699	0.5809	0.2019	0.3901	0.6641
	MLP	0.6867	0.6034	0.2905	0.6193	0.2300	0.4073	0.6912
	Random Forest	0.6408	0.5676	0.2579	0.5815	0.1639	0.3676	0.6556
	SVM	0.6842	0.5657	0.2705	0.5882	0.1966	0.3874	0.6765
	XGBoost	0.6367	0.5754	0.2605	0.5871	0.1672	0.3693	0.6450
CTDT	Adaboost	0.6208	0.5762	0.2566	0.5847	0.1556	0.3627	0.6261
	MLP	0.6408	0.5756	0.2622	0.5880	0.1708	0.3717	0.6439
	Random Forest	0.6025	0.5982	0.2603	0.5990	0.1588	0.3633	0.6241
	SVM	0.6425	0.5841	0.2661	0.5952	0.1787	0.3760	0.6493
	XGBoost	0.5783	0.5668	0.2390	0.5690	0.1147	0.3378	0.6015
CTDD	Adaboost	0.6358	0.6046	0.2744	0.6106	0.1904	0.3831	0.6531
	MLP	0.5942	0.5365	0.2434	0.5475	0.1120	0.3297	0.6065
	Random Forest	0.6967	0.6164	0.2994	0.6316	0.2476	0.4185	0.6987
	SVM	0.6675	0.6111	0.2877	0.6218	0.2206	0.4017	0.6794
	XGBoost	0.6675	0.6201	0.2927	0.6291	0.2282	0.4064	0.6847
PAAC	Adaboost	0.5942	0.6052	0.2611	0.6031	0.1581	0.3626	0.6253
	MLP	0.5958	0.5717	0.2462	0.5763	0.1321	0.3482	0.6261
	Random Forest	0.6375	0.5809	0.2633	0.5917	0.1723	0.3723	0.6413
	SVM	0.6617	0.5905	0.2752	0.6041	0.1990	0.3885	0.6745
	XGBoost	0.6217	0.5731	0.2554	0.5823	0.1537	0.3615	0.6375
APAAC	Adaboost	0.6125	0.5976	0.2634	0.6004	0.1662	0.3682	0.6367
	MLP	0.5658	0.5904	0.2450	0.5857	0.1237	0.3416	0.6162
	Random Forest	0.6458	0.5831	0.2671	0.5950	0.1805	0.3776	0.6464
	SVM	0.6650	0.5970	0.2794	0.6099	0.2069	0.3932	0.6777
	XGBoost	0.6425	0.5694	0.2596	0.5833	0.1668	0.3695	0.6375

highest cross-validation scores for precision, MCC, and F1-score. It can also be noted that out of the six models, SVM performed best on four of them compared to the other machine learning algorithms.

### 3.3 Models Based on Combination of Sequence-Based Feature and Pre-trained Transformer Models Feature Set

To obtain the best model, we tested various combinations of two or more feature sets to evaluate further models, such as AAC-EAAC, AAC-CTDC, and AAC-ProtBert. For this, we limited the embedding features to a maximum of one set in the combination. Similar to previous experiments, five classification algorithms were used: AdaBoost, XGBoost, SVM (RBF kernel), RF, and MLP.

Our best model, ProtTrans-Glutar, uses a combination of the features CTDD, EAAC, and ProtT5-XL-UniRef50 with the XGBoost classification algorithm. The performance of this model is shown in **Table 6**, with comparison to the best model from sequence-based features (EAAC with RF classifier) and the best model from embeddings of the protein model (ProtAlburt with SVM classifier). According to the cross-validation performance on training data, this model has the best AUC and recall compared with models with features from only one group. These three models were then evaluated using an independent dataset (**Figure 2**). This test result shows that ProtTrans-Glutar outperformed the other two models in terms of AUC, recall, precision, MCC, and F1-score. However, it is severely worse in terms of specificity and slightly worse in terms of accuracy compared to the EAAC + RF model.

**TABLE 5** | Cross validation result of models from pre-trained transformer models.

Feature groups	Classifier	Rec	Spe	Pre	Acc	MCC	F1	AUC
ProtBERT	Adaboost	0.5767	0.5680	0.2389	0.5697	0.1142	0.3374	0.5996
	MLP	0.5892	0.5608	0.2395	0.5662	0.1187	0.3396	0.6128
	Random Forest	0.5567	0.6426	0.2681	0.6262	0.1602	0.3616	0.6415
	SVM	0.7042	0.4775	0.2420	0.5207	0.1475	0.3578	0.6275
	XGBoost	0.6033	0.6007	0.2619	0.6012	0.1616	0.3649	0.6398
ProtBert-BFD	Adaboost	0.5433	0.5547	0.2231	0.5525	0.0773	0.3162	0.5776
	MLP	0.5900	0.5645	0.2420	0.5694	0.1218	0.3430	0.6076
	Random Forest	0.5383	0.6230	0.2510	0.6069	0.1289	0.3421	0.6122
	SVM	0.6242	0.5819	0.2595	0.5899	0.1626	0.3662	0.6420
	XGBoost	0.5908	0.5733	0.2453	0.5766	0.1295	0.3464	0.6142
ProtAlbert	Adaboost	0.5875	0.5753	0.2450	0.5776	0.1284	0.3456	0.6193
	MLP	0.5858	0.6189	0.2657	0.6126	0.1646	0.3615	0.6407
	Random Forest	0.5808	0.6316	0.2703	0.6220	0.1697	0.3687	0.6535
	SVM	0.6283	0.6136	0.2767	0.6164	0.1919	0.3840	0.6744
	XGBoost	0.6092	0.5927	0.2604	0.5958	0.1597	0.3646	0.6477
ProtT5-XL-UniRef50	Adaboost	0.5533	0.5655	0.2306	0.5632	0.0938	0.3254	0.5897
	MLP	0.6192	0.5633	0.2501	0.5739	0.1439	0.3558	0.6296
	Random Forest	0.5608	0.6171	0.2562	0.6064	0.1419	0.3515	0.6237
	SVM	0.6583	0.5710	0.2653	0.5876	0.1807	0.3777	0.6600
	XGBoost	0.5933	0.5807	0.2497	0.5831	0.1377	0.3509	0.6183
ProtT5-XL-BFD	Adaboost	0.5892	0.5600	0.2395	0.5656	0.1175	0.3405	0.5959
	MLP	0.6000	0.5768	0.2502	0.5812	0.1396	0.3529	0.6188
	Random Forest	0.5392	0.6163	0.2485	0.6017	0.1242	0.3399	0.6145
	SVM	0.6550	0.5625	0.2604	0.5801	0.1711	0.3724	0.6548
	XGBoost	0.5858	0.5862	0.2490	0.5862	0.1361	0.3489	0.6224
ProtXLNet	Adaboost	0.5125	0.5343	0.2057	0.5302	0.0369	0.2934	0.5421
	MLP	0.5325	0.5248	0.2081	0.5262	0.0450	0.2991	0.5463
	Random Forest	0.5050	0.5668	0.2152	0.5551	0.0568	0.3015	0.5511
	SVM	0.4742	0.5770	0.2103	0.5575	0.0408	0.2900	0.5460
	XGBoost	0.5642	0.5504	0.2274	0.5530	0.0902	0.3238	0.5652

**TABLE 6** | Performance comparison of the best models in each group.

Evaluation	Models	Length	Rec	Spe	Pre	Acc	MCC	F1	AUC
10-fold CV on Training Data	ProtTrans-Glutar <sup>a</sup>	1,599	0.6783	0.6277	0.3004	0.6374	0.2433	0.4158	0.7093
	ProtAlbert + SVM	4,096	0.6283	0.6136	0.2767	0.6164	0.1919	0.3840	0.6744
	EAAC + RF	380	0.6450	0.6598	0.3089	0.6570	0.2450	0.4171	0.6999
Independent Test Set	ProtTrans-Glutar <sup>a</sup>	1,599	0.7864	0.6286	0.3147	0.6567	0.3196	0.4494	0.7075
	ProtAlbert + SVM	4,096	0.6500	0.6286	0.2753	0.6324	0.2161	0.3866	0.6393
	EAAC + RF	380	0.6409	0.6739	0.2989	0.6680	0.2479	0.4076	0.6574

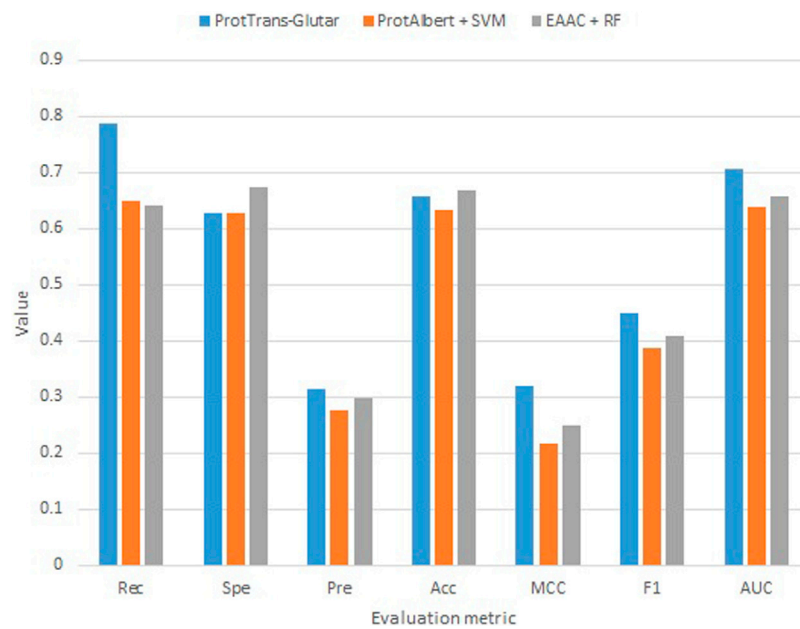
<sup>a</sup>Model uses combined features CTDD-EAAC-ProtT5XLUniRef50 with XGBoost classifier.

As shown in the ROC curves of the three models (**Figure 3**), EAAC + RF performed better for low values of FPR, but for larger values, ProtTrans-Glutar performed better. It is also noted that ProtAlbert + SVM performed worse for most values of FPR. Overall, ProtTrans-Glutar was the best model with an AUC of 0.7075.

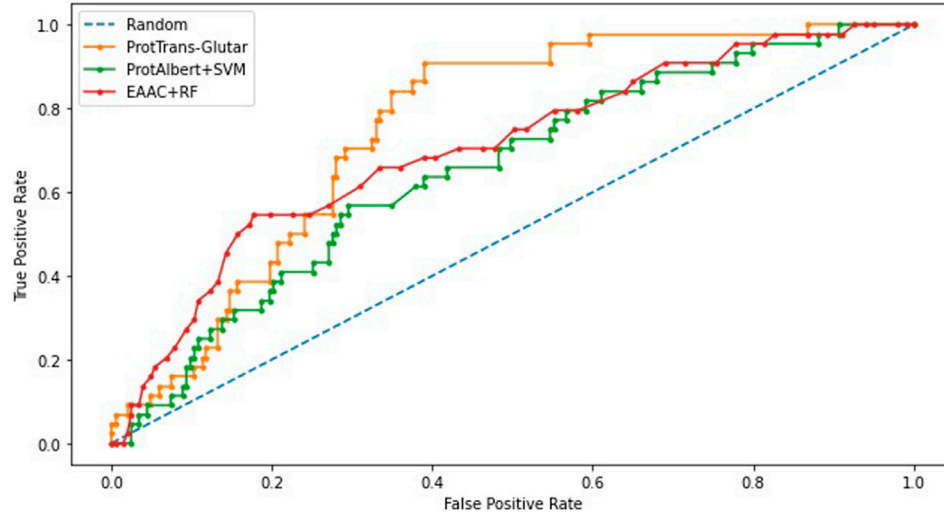
## 4 DISCUSSION

From our study, it was shown that building prediction models from traditional sequence-based features only provided limited performance (**Table 4**). It was also shown that using only embeddings from pre-trained protein models gave slightly worse results, except that the recall performance was almost

the same (**Table 5**). When we combined the features from these two groups, we found that the best performance was achieved by the combination of the features CTDD, EAAC, and ProtT5-XL-UniRef50 with the XGBoost classifier (independent test AUC = 0.7075). This indicated that ProtT5-XL-UniRef50 features on their own are not the best embedding model during the individual feature evaluation (see **Table 5**), but combined with CTDD and EAAC, it outperformed the other models. It is worth mentioning that Elnaggar et al. (2021), who developed and trained protein models, revealed that ProtT5 models outperformed state-of-the-art models in protein classification tasks, namely in prediction of localization (10-class classification) and prediction of membrane/other (binary classification), compared to other embedding models.



**FIGURE 2 |** Independent test evaluation of the best models from each group.



**FIGURE 3 |** ROC-Curve plot of best models in each group.

**TABLE 7 |** Performance comparison of existing models.

Models	Resources	Rec	Spe	Pre	Acc	MCC	F1	AUC
GlutPred	PLMD	0.5179	0.7850	0.2397	0.7541	0.2238	n/a	0.7663
iGlu-Lys	PLMD	0.5143	0.9531	n/a	0.8853	0.52	n/a	0.8842
MDDGlutar	PLMD	0.652	0.739	n/a	0.71	0.38	n/a	n/a
iGlu_AdaBoost	PLMD, NCBI, Swiss-Prot	0.7273	0.7192	0.3596	0.7207	0.36	0.48	0.6300
ProtTrans-Glutar	PLMD, NCBI, Swiss-Prot	0.7822	0.6286	0.3147	0.6567	0.3196	0.4494	0.7075

**TABLE 8** | Performance comparison with RF-GlutarySite using balanced train and test data.

Models	Resources	Rec	Spe	Pre	Acc	MCC	F1	AUC
RF-GlutarySite <sup>a</sup>	PLMD, NCBI, Swiss-Prot	0.741	0.685	0.72	0.713	0.43	0.72	0.72
ProtTrans-Glutar (balanced)	PLMD, NCBI, Swiss-Prot	0.7864	0.6455	0.6955	0.7159	0.4388	0.7358	0.7159

<sup>a</sup>RF-GlutarySite model balanced the training and testing dataset using undersampling.

For further evaluation, we compared our model with previous glutarylation site prediction models (Table 7). The first three models, GlutPred, iGlu-Lys, and MDDGlutar, used datasets that were different from our model and are shown for reference. The other model, iGlu\_Adaboost, utilized the same public dataset as for our model and contained glutarylation sites from the same four species. ProtTrans-Glutar outperformed the other models in terms of the recall performance (Rec = 0.7864 for unbalanced data). This high recall suggests that this model can be useful for uncovering new and potential glutarylation sites.

Furthermore, we also evaluated our model by using a balanced training and testing dataset using random under-sampling for comparison with the RF-GlutarySite model (Table 8), which uses the same dataset but is balanced before evaluating performance. Because the authors of RF-GlutarySite did not provide their data after the resampling process, we performed the experiments 10 times to handle variance from the under-sampling. The ProtTrans-Glutar model showed a higher recall score of 0.7864 compared to RF-GlutarySite (0.7410), in addition to a slightly higher accuracy, MCC, and F1-score. However, the specificity and precision scores were lower.

In summary, the model improved the recall score compared to the existing models but did not improve other metrics. However, we would like to point out that GlutPred, iGlu-Lys, and MDDGlutar based their glutarylation datasets on less diverse sources (two species only), whereas ProtTrans-Glutar with RF-GlutarySite and iGlu\_Adaboost utilized newer datasets (four species). The more diverse source of glutarylation sites in the data may present more difficulty in improving performance, especially in terms of specificity and accuracy. Compared with iGlu\_Adaboost, which used the same dataset, our model improved their recall and AUC scores. Despite this, the specificity is worse and will be a challenge for future research.

## 5 SUMMARY

In this study, we presented a new glutarylation site predictor by incorporating embeddings from pretrained protein models as features. This method, which is termed ProtTrans-Glutar, combines three feature sets: EAAC, CTDD, and ProtT5-XL-UniRef50. Random under-sampling was used in conjunction with the XGBoost classifier to train the model. The performance evaluations obtained from this model for recall, specificity, and AUC are 0.7864, 0.6286, and 0.7075, respectively.

Compared to other models using the same dataset of more diverse sources of glutarylation sites, this model outperformed the existing model in terms of recall and AUC score and could potentially be used to complement previous models to reveal new glutarylated sites. In the future, refinements can be expected through further experiments, such as applying other feature selection methods, feature processing, and investigating deep learning models.

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: <https://github.com/indriani/ProtTrans-Glutar/tree/main/dataset>.

## AUTHOR CONTRIBUTIONS

FI and KS conceived the study; FI and KM designed the experiments; FI, KM, and BP performed the experiments; KS supervised the study; FI wrote the draft article; FI, KM, and KS reviewed and revised the article. All authors have read and agreed to the published version of the manuscript.

## ACKNOWLEDGMENTS

FI would like to gratefully acknowledge the Directorate General of Higher Education, Research, and Technology; Ministry of Education, Culture, Research, and Technology of The Republic of Indonesia for providing the BPP-LN scholarship. In this research, the super-computing resource was provided by Human Genome Center, the Institute of Medical Science, the University of Tokyo. Additional computation time was provided by the super computer system in Research Organization of Information and Systems (ROIS), National Institute of Genetics (NIG).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2022.885929/full#supplementary-material>

## REFERENCES

- Al-barakati, H. J., Saigo, H., Newman, R. H., and Kc, D. B. (2019). RF-GlutarySite: A Random Forest Based Predictor for Glutarylation Sites. *Mol. Omics* 15, 189–204. doi:10.1039/C9MO00028C
- Bhasin, M., and Raghava, G. P. S. (2004). Classification of Nuclear Receptors Based on Amino Acid Composition and Dipeptide Composition. *J. Biol. Chem.* 279, 23262–23266. doi:10.1074/jbc.M401932200
- Cai, C. Z. (2003). SVM-prot: Web-Based Support Vector Machine Software for Functional Classification of a Protein from its Primary Sequence. *Nucleic Acids Res.* 31, 3692–3697. doi:10.1093/nar/gkg600
- Carrico, C., Meyer, J. G., He, W., Gibson, B. W., and Verdin, E. (2018). The Mitochondrial Acylome Emerges: Proteomics, Regulation by Sirtuins, and Metabolic and Disease Implications. *Cell Metab.* 27, 497–512. doi:10.1016/j.cmet.2018.01.016
- Chen, T., and Guestrin, C. (2016). “XGBoost: A Scalable Tree Boosting System,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA, August 2016 (ACM), 785–794. doi:10.1145/2939672.2939785
- Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., et al. (2018). iFeature: A Python Package and Web Server for Features Extraction and Selection from Protein and Peptide Sequences. *Bioinformatics* 34, 2499–2502. doi:10.1093/bioinformatics/bty140
- Chien, C.-H., Chang, C.-C., Lin, S.-H., Chen, C.-W., Chang, Z.-H., and Chu, Y.-W. (2020). N-GlycoGo: Predicting Protein N-Glycosylation Sites on Imbalanced Data Sets by Using Heterogeneous and Comprehensive Strategy. *IEEE Access* 8, 165944–165950. doi:10.1109/ACCESS.2020.3022629
- Chou, K.-C. (2001). Prediction of Protein Cellular Attributes Using Pseudo-amino Acid Composition. *Proteins* 43, 246–255. doi:10.1002/prot.1035
- Chou, K.-C. (2005). Using Amphiphilic Pseudo Amino Acid Composition to Predict Enzyme Subfamily Classes. *Bioinformatics* 21, 10–19. doi:10.1093/bioinformatics/bth466
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv1810.04805 Cs. doi:10.18653/v1/N19-1423
- Dou, L., Li, X., Zhang, L., Xiang, H., and Xu, L. (2021). iGlu\_AdaBoost: Identification of Lysine Glutarylation Using the AdaBoost Classifier. *J. Proteome Res.* 20, 191–201. doi:10.1021/acs.jproteome.0c00314
- Dubchak, I., Muchnik, I., Holbrook, S. R., and Kim, S. H. (1995). Prediction of Protein Folding Class Using Global Description of Amino Acid Sequence. *Proc. Natl. Acad. Sci. U.S.A.* 92, 8700–8704. doi:10.1073/pnas.92.19.8700
- Elnaggar, A., Heininger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., et al. (2021). ProtTrans: Towards Cracking the Language of Life's Code through Self-Supervised Deep Learning and High Performance Computing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1. doi:10.1109/TPAMI.2021.3095381
- Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* 29, 1189–1232. doi:10.1214/aos/1013203451
- Harmel, R., and Fiedler, D. (2018). Features and Regulation of Non-enzymatic Post-translational Modifications. *Nat. Chem. Biol.* 14, 244–252. doi:10.1038/nchembio.2575
- He, H., and Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* 21, 1263–1284. doi:10.1109/TKDE.2008.239
- H. He and Y. Ma (Editors) (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications* (Hoboken, New Jersey: John Wiley & Sons).
- Ho, Q.-T., Nguyen, T.-T. -D., Le, N. Q. K., and Ou, Y.-Y. (2021). FAD-BERT: Improved Prediction of FAD Binding Sites Using Pre-training of Deep Bidirectional Transformers. *Comput. Biol. Med.* 131, 104258. doi:10.1016/j.combiomed.2021.104258
- Huang, K.-Y., Kao, H.-J., Hsu, J. B.-K., Weng, S.-L., and Lee, T.-Y. (2019). Characterization and Identification of Lysine Glutarylation Based on Intrinsic Interdependence between Positions in the Substrate Sites. *BMC Bioinforma.* 19, 384. doi:10.1186/s12859-018-2394-9
- Ju, Z., and He, J.-J. (2018). Prediction of Lysine Glutarylation Sites by Maximum Relevance Minimum Redundancy Feature Selection. *Anal. Biochem.* 550, 1–7. doi:10.1016/j.ab.2018.04.005
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. ArXiv1909.11942 Cs. doi:10.48550/arXiv.1909.11942
- Lee, J. V., Carrer, A., Shah, S., Snyder, N. W., Wei, S., Venneti, S., et al. (2014). Akt-Dependent Metabolic Reprogramming Regulates Tumor Cell Histone Acetylation. *Cell Metab.* 20, 306–319. doi:10.1016/j.cmet.2014.06.004
- Liu, Y., Liu, Y., Wang, G.-A., Cheng, Y., Bi, S., and Zhu, X. (2022). BERT-kgly: A Bidirectional Encoder Representations from Transformers (BERT)-Based Model for Predicting Lysine Glycation Site for *Homo sapiens*. *Front. Bioinform.* 2, 834153. doi:10.3389/fbinf.2022.834153
- Mahmud, S. M. H., Chen, W., Jahan, H., Liu, Y., Sujan, N. I., and Ahmed, S. (2019). iDTi-CSsmoteB: Identification of Drug-Target Interaction Based on Drug Chemical Structure and Protein Sequence Using XGBoost with Over-sampling Technique SMOTE. *IEEE Access* 7, 48699–48714. doi:10.1109/ACCESS.2019.2910277
- Osborne, B., Bentley, N. L., Montgomery, M. K., and Turner, N. (2016). The Role of Mitochondrial Sirtuins in Health and Disease. *Free Radic. Biol. Med.* 100, 164–174. doi:10.1016/j.freeradbiomed.2016.04.197
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-To-Text Transformer. ArXiv1910.10683 Cs. doi:10.48550/arXiv.1910.10683
- Shah, S. M. A., Taj, S. W., Ho, Q.-T., Nguyen, T.-T. -D., and Ou, Y.-Y. (2021). GT-finder: Classify the Family of Glucose Transporters with Pre-trained BERT Language Models. *Comput. Biol. Med.* 131, 104259. doi:10.1016/j.combiomed.2021.104259
- Tan, M., Peng, C., Anderson, K. A., Chhoy, P., Xie, Z., Dai, L., et al. (2014). Lysine Glutarylation Is a Protein Posttranslational Modification Regulated by SIRT5. *Cell Metab.* 19, 605–617. doi:10.1016/j.cmet.2014.03.014
- Xu, H., Zhou, J., Lin, S., Deng, W., Zhang, Y., and Xue, Y. (2017). PLMD: An Updated Data Resource of Protein Lysine Modifications. *J. Genet. Genomics* 44, 243–250. doi:10.1016/j.jgg.2017.03.007
- Xu, Y., Yang, Y., Ding, J., and Li, C. (2018). iGlu-Lys: A Predictor for Lysine Glutarylation through Amino Acid Pair Order Features. *IEEE Trans. on Nanobioscience* 17, 394–401. doi:10.1109/TNB.2018.2848673
- Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2020). XLNet: Generalized Autoregressive Pretraining for Language Understanding. ArXiv1906.08237 Cs. doi:10.48550/arXiv.1906.08237
- Zhang, G., Liu, Z., Dai, J., Yu, Z., Liu, S., and Zhang, W. (2020). ItLnc-BXE: A Bagging-XGBoost-Ensemble Method with Comprehensive Sequence Features for Identification of Plant lncRNAs. *IEEE Access* 8, 68811–68819. doi:10.1109/ACCESS.2020.2985114

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

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

Copyright © 2022 Indriani, Mahmudah, Purnama and Satou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# DTI-BERT: Identifying Drug-Target Interactions in Cellular Networking Based on BERT and Deep Learning Method

Jie Zheng, Xuan Xiao\* and Wang-Ren Qiu\*

Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen, China

## OPEN ACCESS

### Edited by:

Juexin Wang,  
University of Missouri, United States

### Reviewed by:

Yang Liu,  
Dana-Farber Cancer Institute,  
United States  
Xing Chen,  
China University of Mining and  
Technology, China

### \*Correspondence:

Xuan Xiao  
jdzxiao@163.com  
Wang-Ren Qiu  
qiuone@163.com

### Specialty section:

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

Received: 21 January 2022

Accepted: 25 April 2022

Published: 08 June 2022

### Citation:

Zheng J, Xiao X and Qiu W-R (2022)  
DTI-BERT: Identifying Drug-Target  
Interactions in Cellular Networking  
Based on BERT and Deep  
Learning Method.  
Front. Genet. 13:859188.  
doi: 10.3389/fgene.2022.859188

Drug-target interactions (DTIs) are regarded as an essential part of genomic drug discovery, and computational prediction of DTIs can accelerate to find the lead drug for the target, which can make up for the lack of time-consuming and expensive wet-lab techniques. Currently, many computational methods predict DTIs based on sequential composition or physicochemical properties of drug and target, but further efforts are needed to improve them. In this article, we proposed a new sequence-based method for accurately identifying DTIs. For target protein, we explore using pre-trained Bidirectional Encoder Representations from Transformers (BERT) to extract sequence features, which can provide unique and valuable pattern information. For drug molecules, Discrete Wavelet Transform (DWT) is employed to generate information from drug molecular fingerprints. Then we concatenate the feature vectors of the DTIs, and input them into a feature extraction module consisting of a batch-norm layer, rectified linear activation layer and linear layer, called BRL block and a Convolutional Neural Networks module to extract DTIs features further. Subsequently, a BRL block is used as the prediction engine. After optimizing the model based on contrastive loss and cross-entropy loss, it gave prediction accuracies of the target families of G Protein-coupled receptors, ion channels, enzymes, and nuclear receptors up to 90.1, 94.7, 94.9, and 89%, which indicated that the proposed method can outperform the existing predictors. To make it as convenient as possible for researchers, the web server for the new predictor is freely accessible at: <https://bioinfo.jcu.edu.cn/dtibert> or <http://121.36.221.79/dtibert/>. The proposed method may also be a potential option for other DTIs.

**Keywords:** drug-target interactions, bidirectional encoder representations from transformers, BRL block, convolutional neural network, computational methods

## 1 INTRODUCTION

In the process of drug development, there are many important drug-related interaction directions, including drug-protein, drug-miRNA, drug-disease, drug-drug, etc. Small molecule therapeutic drugs typically exert their effects through binding to one or a few protein targets (Dubach et al., 2014; Lim et al., 2021), therefore identifying drug-protein interaction is an important part of genomic drug discovery (Yamanishi et al., 2014). Besides, several studies have indicated that although ncRNAs lack the potential to encode proteins, they play important roles in cellular functions, and their

deregulation heavily contributes to various pathological conditions. Among them, miRNAs are promising therapeutic targets for complex diseases (Wang and Chen, 2019; Yin et al., 2019; Zhou et al., 2020), it thus becomes important to understand the relationship between ncRNAs and drug targets, what's more, several databases and studies are actively promoting development (Chen et al., 2017). Drug-disease and drug-drug interaction play a crucial role in drug relocation, often serving as important information other than drug-target protein pairing and mainly based on a processing framework called a heterogeneous network. Qu et al. developed a novel computational model of HeteSim-based inference for SM-miRNA Association prediction by implementing a path-based measurement method of HeteSim on a heterogeneous network combined with known miRNA-SM associations, integrated miRNA similarity, and integrated SM similarity (Qu et al., 2019). Jin et al. combine drug features from multiple drug-related networks, and disease features from biomedical corpora with the known drug-disease association's network to predict the correlation scores between drug and disease (Qu et al., 2019). Drug-protein interactions play a key role in the field of biochemistry due to their scientific significance in drug discovery. This paper focuses on the identification of drug-protein interactions.

Drugs modulate the biological functions of proteins by interacting with target proteins, such as ion channels, nuclear receptors, enzymes, and G Protein-coupled receptors (GPCRs). For an in-depth understanding of the functions of drugs, the knowledge of their target protein is indispensable. Despite the substantial effort, only a few DTIs have been identified so far, since the experimental determination of drug-target interactions remains some defects, such as expensive, time-consuming, low accuracy, and so on (Haggarty et al., 2003). It is highly demanded to develop powerful computational tools, which are capable of detecting potential DTIs. Computational prediction of DTIs has emerged for 20 years as a research hotspot, which is not only for better understanding of the molecular mechanism of drug side effects but also for inventing new genomic drugs and identifying new targets for existing drugs (Wang et al., 2010; Kotlyar et al., 2012).

Knowledge of genomic space and chemical space is indispensable for identifying DTIs. With the coming of the post-genome era and the emergence of molecular medicine, transcriptome, and chemical compound, the rapidly increasing knowledge in the field of genomic space and chemical space enables researchers to study drug-target interaction problems (Dobson, 2004) on the basis of high-throughput experimental projects. Several different professional databases have been established, such as Drug Bank, which is consist of two parts information involving drug data and drug target information (Wishart et al., 2018); Therapeutic Target Database (TTD) provides comprehensive information about the drug resistance mutations, gene expressions, and target combinations data (Qin et al., 2014); BindingDB a public database of protein-ligand binding affinities (Liu et al., 2007); Kyoto Encyclopedia of Genes and Genomes (KEGG) including experimental knowledge on protein and their drug target, etc. These resources provide important materials for researchers to

predict drug-target interactions based on computational methods, it is time to develop more integrative approaches capable of taking genomic space, chemical space, and the available known drug-target network information into account simultaneously for the issue.

The development of identifying DTIs followed four main directions for research. Firstly, the most direct method is to use the docking simulation (Pujadas et al., 2008; Morris et al., 2009), which is a process of scoring favorable intermolecular interactions, the three-dimensional (3D) structures of proteins and chemical compounds are indispensable. With the development of techniques (e.g., X-ray crystallography, nuclear magnetic resonance), the rate of 3D protein structure determination is increasing every year, however, it is still not able to keep up with the exponential growth of sequence discovery, such as the PDB database only covers a small fraction of the ion channels and GPCRs, both are considered as the most pharmaceutically useful drug targets. Some programs and web servers provide the prediction of the protein structure, in practice, structure prediction is still relatively immature, and interaction prediction may be affected by the inaccurate structure. Secondly, based on the fact that similar molecules usually bind to similar proteins, it is most straightforward to apply the ligand-based approach (Keiser et al., 2007), for example, conducting Quantitative Structure-Activity Relationship (QSAR) studies that a new ligand can be categorized and compared to known proteins ligands. However, ligand-based approaches often present unreliable results due to available binding ligands of targets' insufficient number, and difficult to scientifically set thresholds to divide positive and negative samples (Butina et al., 2002). Thirdly, literature text mining could be used to extract DTIs from the related articles (Zhu et al., 2005), but this approach could not be used for new drugs and proteins. Fourthly, to overcome the drawbacks of the above-mentioned traditional approaches, chemogenomic approaches are universally studied directions. Chemogenomic approaches integrate information of chemical space, genomic space, and known drug-target interactions, which provide an architecture for deep learning approaches.

Chemogenomic approaches can be classified into three categories: graph-based approaches (Chen et al., 2012), network-based approaches (Alaimo et al., 2013), and learning-based approaches (Mousavian and Masoudi-Nejad, 2014). In the graph-based approach, drugs and targets are represented with graphs, in which nodes for chemical elements or amino acids and adjacency matrices for edges between nodes, adjacency matrices including atom/bond or residue/bond information (Lim et al., 2021). Drug and target graphs can be fed into Graph Neural Network (GNN); after a set of training iterations, information learned by Graph Convolutional Network (GCN) can be converted into vectors for DTIs prediction. Torng and Altman proposed a graph-convolutional framework to determine the interaction patterns (Torng and Altman, 2019). Karlov et al. used the message passing neural network to overcome the limitation of graph convolutional network by considering both nodes and edges (Karlov et al., 2020). Furthermore, the self-attention mechanism in Neural Networks is often coupled with

Graph convolutional network to predict DTIs better. But some research showed that there are difficulties in predicting the local non-covalent interactions between drugs and proteins (Li et al., 2020). Network-based approaches utilized the DTI network of identified edges between drugs and targets to identify new DTIs. Indeed, by constructing a heterogeneous network that includes information on drugs, proteins, diseases, and side-effects, the DTINet method can improve the accuracy of DTIs prediction (Luo et al., 2017), but the learning model only takes relatively simple log-bilinear functions, obtaining features may not be the inherent representations of drugs or targets for the final DTI prediction task (Wan et al., 2019). Supervised learning-based approaches are classified into similarity-based approaches and feature-based approaches (Chen et al., 2018). Similarity-based approaches generate the similarity matrixes for drugs and targets respectively, via various similarity measurement strategies such as chemical-based similarity (Haggarty et al., 2003), pharmacological-based similarity (Kim et al., 2013), therapeutic-based similarity, and drug-drug interaction similarity for drugs, and sequence-based similarity (Yamanishi et al., 2008), functional-bases similarity, protein-protein interaction similarity for targets. These similarity matrices have been used in bipartite local models (Mei et al., 2013), matrix factorization models (Ezzat et al., 2016), and the nearest neighbor methods (Zhang et al., 2016) to predict DTIs. The feature-based approaches extract more useful information from protein sequences and drug chemical structure, via the adequate support offered by the rapid development of algorithms.

Predicting DTIs with machine learning algorithms has recently become the focus of research. There are 1-D, 2-D, and 3-D representations of drugs (Rognan, 2007). Simplified Molecular Input Line Entry System (SMILES) string is a typical 1-D representation of the drug (Öztürk et al., 2016) that are commonly used descriptors (Kombo et al., 2013; Sawada et al., 2014). For targets, the sequences of protein are encoded by the physicochemical properties of amino acids, sequential evolution information formulation and general form of pseudo amino acid composition (Li et al., 2020). Lastly, machine learning algorithms are applied for decision-making. Recently, Wang et al. used a novel bag-of-words model and discrete Fourier transform to extract target sequence feature and molecular fingerprint pattern information, respectively, and then use a distance-weighted K-nearest-neighbor algorithm as a predictor (Wang et al., 2020). This paper motivates our work, that instead of using amino acid physicochemical properties to encode words and perform clustering, we can vectorization drugs and protein by using advanced methods such as word2vec and ProtBert (Elnaggar et al., 2021), which could map every word (amino acids are regarded as words) into the latent vector space where the geometric relationship can be used to characterize the semantic relationship between the words. And based on the present situation of identifying DTIs by the way of investigating a series of recently published articles (Keiser et al., 2007; Ezzat et al., 2016; Zhang et al., 2016) as well as some review papers (Rognan, 2007; Kombo et al., 2013; Öztürk et al., 2016), we have proposed a novel feature-based computational model for predicting drug-target interactions to enhance prediction

performance. The novelty of this proposed work 1) Compared with the end-to-end predictor, we treat DTIs task more flexibly. The protein sequences are regarded as natural language and vectorized by the state-of-art ProtBert model, and drug molecular is transformed by DWT, which is commonly used in signal processing. 2) Calculating the hybrid loss function (contrastive loss and cross-entropy loss), which can make the samples of the same interaction label closer, and the distance between different labels as far as possible and help the predictor achieve higher accuracy.

## 2 MATERIALS AND METHODS

### 2.1 Benchmark Dataset

Identifying DTIs can be regarded as a supervised prediction task to predict whether a pair of counterparts interact with each other or not in the drug-target networks. In this study, the benchmark dataset was taken from (He et al., 2010). There are mainly two reasons, 1) The information about the DTIs was collected from the DrugBanks, BRENDA, SuperTarget, and KEGG BRITE databases, which included four main drug target proteins of G Protein-coupled receptors (GPCR), enzymes (Ezy), ion channels (Chl), and nuclear receptors (NR). 2) In recent years, many researchers have been proposed to predict DTIs, which are based on this benchmark dataset, and hence will facilitate the comparison under the same condition. It can be summarized as follows:

$$\begin{cases} S = S_{GPCR-Drug} + S_{Chl-Drug} + S_{Ezy-Drug} + S_{NR-Drug} \\ S_{GPCR-Drug} = S_{GPCR-Drug}^+ (630) + S_{GPCR-Drug}^- (1240) \\ S_{Chl-Drug} = S_{Chl-Drug}^+ (1372) + S_{Chl-Drug}^- (2744) \\ S_{Ezy-Drug} = S_{Ezy-Drug}^+ (2719) + S_{Ezy-Drug}^- (5438) \\ S_{NR-Drug} = S_{NR-Drug}^+ (82) + S_{NR-Drug}^- (164) \end{cases} \quad (1)$$

There are 4,803 drug-target pairs in positive subsets, 2,719 for enzymes, 1,372 for ion channels, 630 for GPCRs, and 82 for nuclear receptors. Negative samples are randomly synthesized by separating each target and drug in  $S^+$ , and none of them appear in the corresponding positive dataset. The proportion of positive samples and negative samples was set as 1:2. For comparison with previously published papers, both our positive and negative samples are consistent with He et al. (He et al., 2010)

Check390 is a dataset constructed by Hu et al. It contains 130 pairs of positive samples from the KEGG database, and 260 negative samples generated using the above method (Hu et al., 2016). Each pair in Check390 cannot be found in  $S$ .

### 2.2 Framework of the Constructed Model

In this article, we construct a novel model for DTIs based on large-scale pre-trained Bidirectional Encoder Representations from Transformers (BERT) and the fully connected neural network-based module called the BRL block. **Figure 1** shows an overview of the DTIs model. The model has four modules: feature engineering, feature extraction, optimization, and decision-making. Firstly, in the feature engineering module, we use the auto-encoder ProtBert model, which is pre-trained on

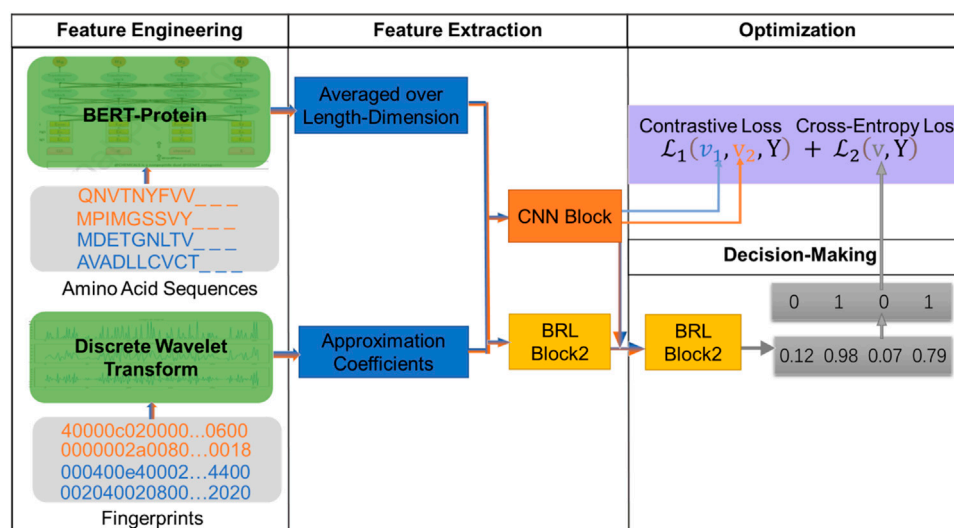


FIGURE 1 | Flowchart of the DTI-BERT model.

data from UniRef100 containing 216M protein sequences, to generate embedding vectors for protein sequences. As a result, the proteins can be represented via 1024-D vectors (dimensionality of the features extracted by the ProtBert model). Drug molecular fingerprints are represented by 128-D vectors through semi decomposition process discrete wavelet transform (DWT). Secondly, the 1152-D vectors (a concatenation of protein sequence feature and drug feature) are fed into the feature extraction model to generate interaction information through the first BRL block and CNN. Afterward, in the decision-making module, the second BRL block is used to map interaction features into a unified vector space. The optimization module contains a contrastive loss and a cross-entropy loss. The contrastive loss is used to calculate the interaction information (generated by CNN block), which can reduce the distance between samples with the same label, and increase the distance between samples with different labels, while the cross-entropy loss is computed as the loss of second BRL block, both are used to adapt weights in the module during the learning process by minimizing the total loss. At the end of model, we can obtain the interaction score (generated by a softmax layer after second BRL block, and range from 0-1), the pair is interaction if the prediction score is  $> 0.5$ .

### 2.2.1 Feature Extraction From Protein

Recently, many word-embedding methods have been used for protein feature extraction, for example, Zheng et al. identified the ion channel-drug interaction using both word2vec and node2vec as molecular representation learning methods (Zheng et al., 2021). However, there are still imperfect, like in these word-embedding methods may map every word with their unique vector, therefore this representation is context-independent. With the exponential growth of textual data, major progress has been made in the pre-training language representations (Peng et al., 2019; Bianchi et al., 2021). Bidirectional Encoder

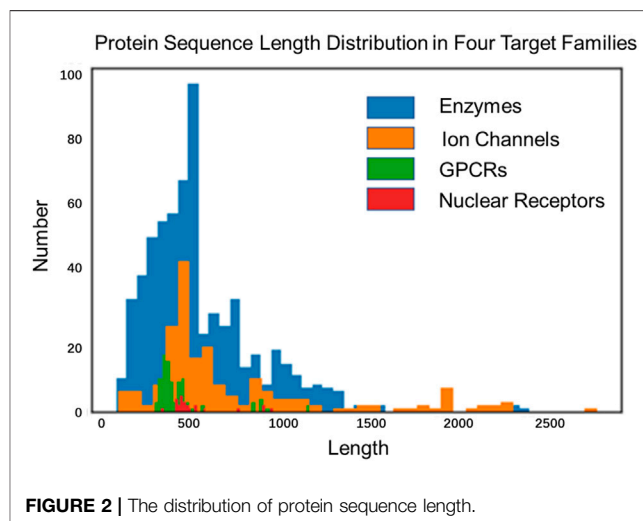


FIGURE 2 | The distribution of protein sequence length.

Representations from Transformers (BERT) was the first fine-tuning-based representation model (Devlin et al., 2018), which can generate different representations for the same word based on context (Devlin et al., 2018; Nozza et al., 2020).

Almost all sequence-based language models (e.g., context ELMo (Ilić et al., 2018), BERT (Devlin et al., 2018), Xlnet (Yang et al., 2019)) have been promoted the development of processing natural languages successfully, but model architectures and pre-training tasks may not be suitable for representing proteins. The primary reason is that proteins are more variable than sentences in length, and show many interactions in distant positions (due to their 3D structure). The length of English sentences is multiple, usually around 15-30 words (Brandes et al., 2022). Although the length limit of a sentence is not an issue in sentence-level NLP tasks (Dai et al., 2019; Brandes et al., 2022), however, many proteins are more than



20-times longer than nature sentences, reaching an average length of up to 600 residues in drug–the target benchmark dataset and over 20% of the sequences are longer than 1,000. The average length of GPCR, ion channel, enzyme and nuclear receptor are 470, 760, 570 and 540, the distribution of protein sequence length is shown in **Figure 2**.

For protein sequence representation, Elnaggar et al. released a model called ProtBert, which was trained on UniRef100 datasets (contained 216M protein sequences) (Elnaggar et al., 2021). In the ProtBert model, amino acids are set as single words and protein sequences as sentences. The model can deal with protein sequences up to 40k in length, and can download from: <https://github.com/agemagician/ProtTrans> (Elnaggar et al., 2021). In the current study, the protein sequence feature can be extracted by ProtBert based on transfer learning (Lee et al., 2019; Noorbakhsh et al., 2020).

The sequence expressed as an amino acid residue may be formulated in the following format:

$$G = R_1 R_2 R_3 \dots R_L \quad (2)$$

where  $R_1$  is the first residue in the protein sequence,  $R_2$  is the second residue,  $\dots$ ,  $R_L$  is the  $L$ -th residue.

The framework of ProtBert is similar to the original Bert publication, some special encoding symbols like [CLS] and [SEP] remain in the BERT model. [CLS] means classification, is added as the first token in the Bert sequence information. When designing the model, [CLS] token was considered as the representation of subsequent text classification. [SEP] means a separator, for example, the task was sentence-pair regression, the input for BERT consists of the two sentences, that would be separated by a special [SEP] token.

We add a [CLS] token at the beginning of the protein sequence marked as  $R_0$ , which acts as an aggregate sequence representation and is usually used for sequence classification tasks in the BERT model, and the [SEP] token at the end of the sequence, marked as  $R_{L+1}$ .

We get protein features from the last layer of ProtBert, and every amino acid can be converted to a 1024-dimensional vector  $B_{R_j}$ , and the protein can be represented as a feature matrix  $P_{BERT}$ :

$$B_{R_j} = [B_{R_j}^1 B_{R_j}^2 \dots B_{R_j}^i \dots B_{R_j}^{1024}] \quad (3)$$

$$P_{BERT} = \begin{bmatrix} B_{R_0}^1 & \dots & B_{R_0}^{1024} \\ \vdots & \ddots & \vdots \\ B_{R_{L+1}}^1 & \dots & B_{R_{L+1}}^{1024} \end{bmatrix} \quad (4)$$

It can be seen from **Eqs. 3, 4** that different protein has different size of  $P_{BERT}$ . To formulate the protein sequences with the same size mathematics formulation, the matrix was averaged (mean-pooled) over the vertical axis and a 1024-dimensional vector was obtained to be used as a representation of protein named BERT\_Mean:

$$b_n = \frac{\sum_{j=0}^{j=L+1} B_{R_j}^n}{L+2} \quad (1 \leq n \leq 1024) \quad (5)$$

$$P_{PROT} = [b_0 b_2 \dots b_n \dots b_{1024}] \quad (6)$$

## 2.2.2 Feature Extraction From Drug Molecule

A drug is saved as an MOL file (a file format that represents a compound in the form of a graph connection table) or SMILES in the database, both formats containing information about the molecule structure, and can be retrieved from the KEGG database (<http://www.kegg.jp/kegg/>) or ChEMBL (<https://www.ebi.ac.uk/chembl/>) according to drug IDs. We can also use the MOL file or SMILES as the input of the OpenBabel tool (<http://openbabel.org/>) to generate the molecular fingerprint file, including FP2, FP3, FP4, and MACSS. FP2 is an enumeration of linear fragments or ring substructures of one to seven connected atoms in a molecule, then maps them to a 256-bit hexadecimal string through a hash function. FP3, FP4, and MACSS use predefined structures to generate fingerprints. FP2 retains more sequence information, we use FP2 as molecular input.

The FP2 molecular fingerprint is represented by a 256-bit hexadecimal string, the hexadecimal char “0~F” can be converted to the number 0–15, drug molecule is represented as  $S_{FP2}$  in the following formulation:

$$S_{FP2} = [f_1 f_2 \dots f_{256}] \quad (7)$$

In previous studies, the FP2 can be further processed using some transposition functions, and Hu et al. (Hu et al., 2016) and Wang et al. (Wang et al., 2020) have confirmed the effectiveness of applying Discrete Fourier Transform (DFT). DFT can convert molecular fingerprints into frequency-domain values, reflecting the specific characteristics of drug molecules. DFT can freely choose frequency domain or time domain according to the needs of practical applications, however, it cannot obtain information in both cases simultaneously, and we cannot know the time when a signal occurs (in our study, it means sequence position information). To solve the local non-stationary components contained in the FP2, DWT was chosen to extract drug features. Daubechies family is the wavelet basis function in DWT, which can support discrete transformation and have good orthogonality and symmetry compared to other wavelet bases. In this paper, the specified wavelet basis function is used to decompose the fingerprint vector, and the approximation coefficients are used as the wavelet coefficients of the fingerprint vector.

After the transformation of DWT with the Daubechies family, 128 approximation coefficients can be obtained to form a vector:

$$S_A = [a_1 a_2 \dots a_{128}] \quad (8)$$

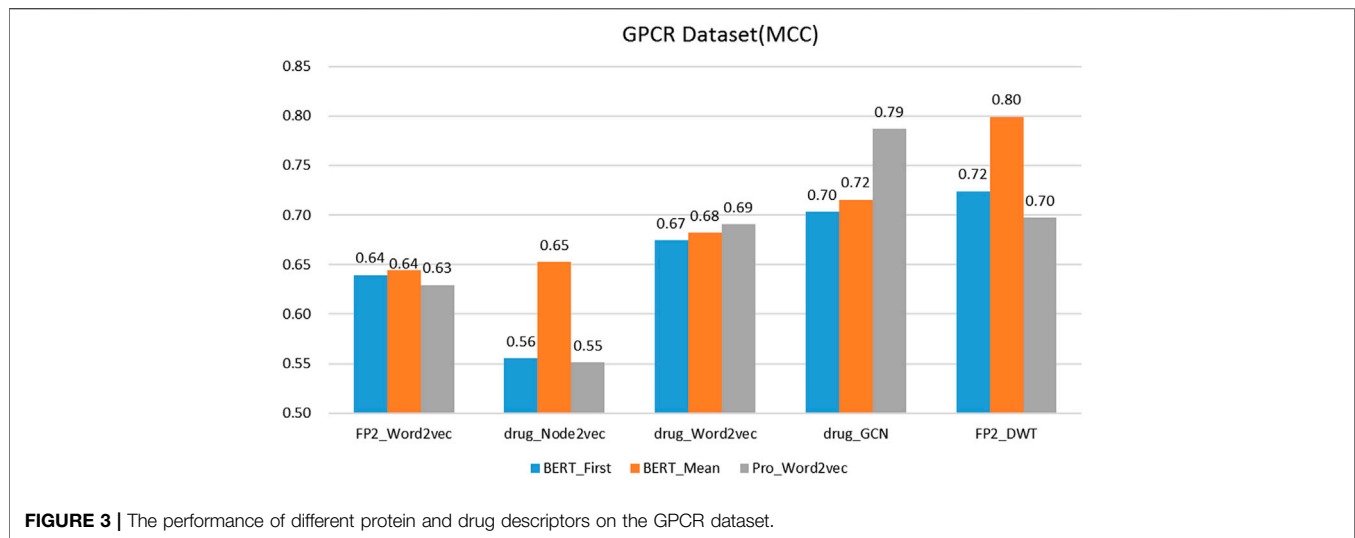
To better characterize the drug,  $S_A$  was subjected to a standard conversion as described by the following equation:

$$d_i = \frac{a_i}{\sum_{j=1}^{128} a_j} \quad (9)$$

$$D_{DWT} = [d_1 d_2 \dots d_{128}] \quad (10)$$

And  $D_{DWT}$  a 128-dimensional vector is obtained to be used as representation of drug. Finally, through the above several steps, a drug-protein pair can be represented with an 1152-D vector given by:





$$\Phi = [\Phi_1 \quad \Phi_2 \quad \cdots \quad \Phi_i \quad \cdots \quad \Phi_{1152}] \quad (11)$$

## 2.3 CNN Block

The CNN block includes a convolution layer, a rectified linear unit activation (ReLU), and a max-pooling layer. Instead of using multi-channels, we applied one channel only (Peng et al., 2018). In the convolution layer, apply a convolution kernel with a window size of  $h \times k$  to extract the DTIs features, then use the rectified linear unit activation function and performed max-pooling to get the most useful interaction feature from the feature matrix subsequently. Through this block, an output of input  $x$  is formulated as:

$$v = \max_{\text{pool}}(f(w \cdot x + b)) \quad (12)$$

where  $w \in R^{h \times k}$ , which is applied to a window of  $h = 18, k = 64$  to produce a new feature;  $b \in R$  is a bias term and  $f$  is a non-linear function.

## 2.4 BRL Block

The BRL is built as a special block in the neural network, where data is normalized and then mapped into a specific vector space. This block consists of three layers: a batch-norm layer (BN), a leaky rectified linear activation layer (Leaky ReLU), and a linear layer (Pedregosa et al., 2011).

The input data  $x$  is first Batch-normalized, which serves to increase the learning rates further, remove the dropout layer, and apply other modifications afforded by the batch normalization (Ioffe and Szegedy, 2015); then input to the Leaky ReLU activation layer, and finally linearly mapped. BRL block can mathematically be represented as:

$$\begin{aligned} X &= \text{Linear}(\text{LeakyReLU}(\text{BN}(x))) \\ &= W \times (\text{LeakyReLU}(\text{BN}(x))) + B \end{aligned} \quad (13)$$

where  $x$  is the input data, the BN transform is applied independently to each dimension of  $x$ ,  $W$  is the weight of the

linear layer, and  $B$  is the bias of the linear layer. The first BRL block and CNN block are used for capturing both global and local information to represent the drug-protein pair; the second BRL block is used for predicting DTIs.

The BRL block was implemented with PyTorch (version 1.6.0), and a fully connected layer was used for the linear mapping. The parameters of the first BRL block were set as: the number of input neurons and the batch normalized dimensions dimension were both 1,152, and the number of output neurons was set to 128. The parameters of the second BRL block were set as 192 (128-D from the first BRL block and 64-D from the CNN block), and two respectively. A softmax layer is applied after the second BRL block, which is used to generate the prediction score. Other hyperparameters used default values in Pytorch. The source code for the related methods is available on a GitHub repository at: <https://github.com/Jane4747/DTI-BERT>.

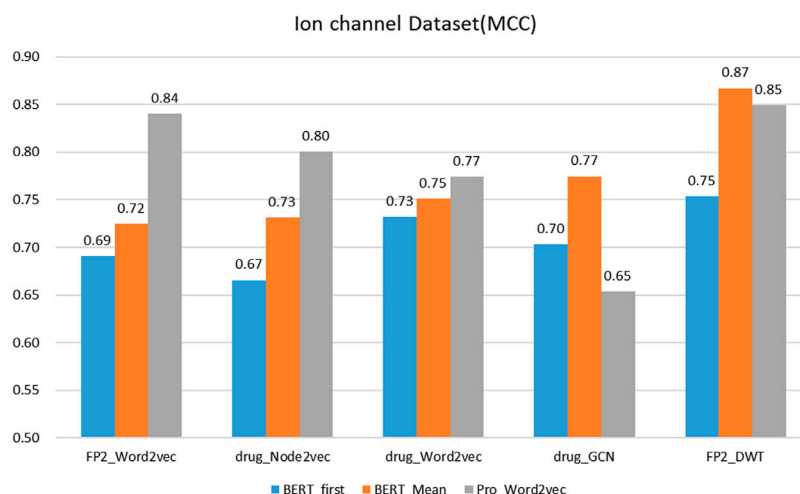
## 2.5 Optimization Module

In this frame, given two vectors  $v_1$  and  $v_2$ , input them into the same network in turn, the network will map the inputs to the new vector space where the similarity between two inputs can be evaluated by the distance measure function. Here, Euclidean distance was served as the distance measure, denoted as  $D(v_1, v_2)$ :

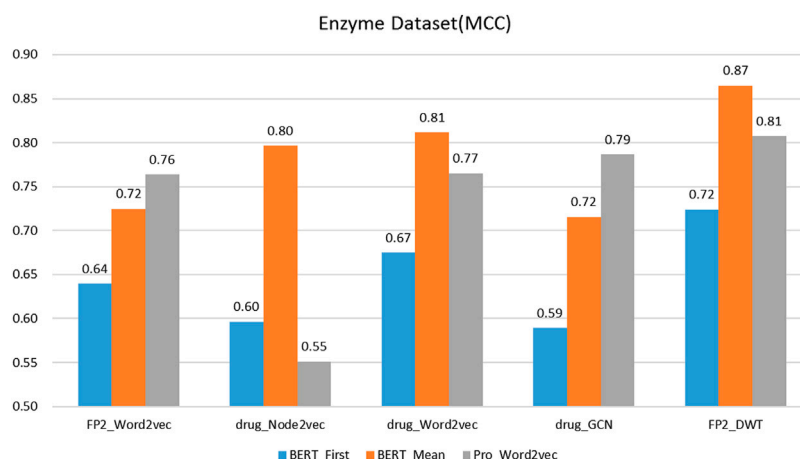
$$\mathcal{D}(v_1, v_2) = \|v_1 - v_2\|_2 \quad (14)$$

To make the samples of the same interaction label closer, and the distance between different labels as far as possible, the contrastive loss was applied as the loss function of the CNN network:

$$\begin{aligned} \mathcal{L}_1(v_1, v_2, Y) &= \frac{1}{2} (1 - Y) \mathcal{D}(v_1, v_2)^2 \\ &+ \frac{1}{2} Y \{ \max(0, m - \mathcal{D}(v_1, v_2))^2 \} \end{aligned} \quad (15)$$



**FIGURE 4 |** The performance of different protein and drug descriptors on the ion channel dataset.



**FIGURE 5 |** The performance of different protein and drug descriptors on the enzyme dataset.

where  $Y = 0$  if sequences  $v_1$  and  $v_2$  have the same label and  $Y = 1$  if they are different,  $m > 0$  is a margin. In other words, the margin defines a radius, and dissimilar pairs contribute to the loss function only if their distance is within the radius.

In this study, the second BRL block was used to convert the representation vector  $v$  to binary category outputs, the backpropagation algorithm was used to update network parameters, and the cross-entropy loss function was selected as the loss function of the second BRL block:

$$\mathcal{L}_2(v, Y) = -Y \log(\mathcal{D}(v)) - (1 - Y) \log(1 - \mathcal{D}(v)) \quad (16)$$

Therefore, the loss function of the DTI-BERT model is:

$$\mathcal{L}(v_1, v_2, Y, Y_1, Y_2) = \mathcal{L}_1(v_1, v_2, Y) + \mathcal{L}_2(v_1, Y_1) + \mathcal{L}_2(v_2, Y_2) \quad (17)$$

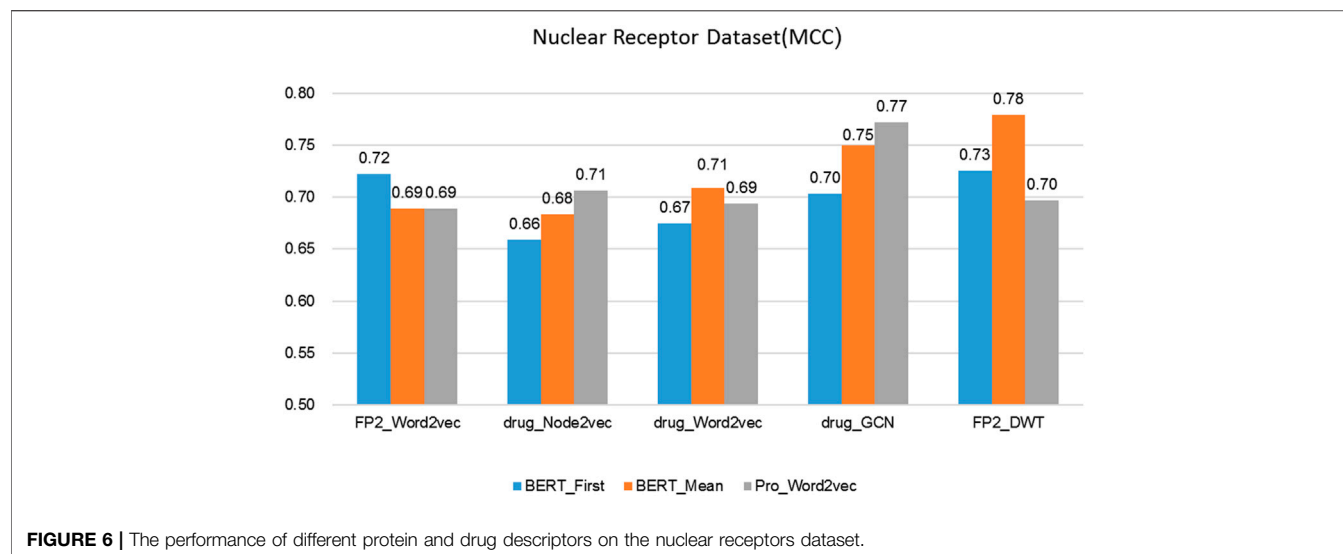
where  $Y_1$  and  $Y_2$  are the labels of  $v_1$  and  $v_2$ .

We implemented our model using Python three and Pytorch (version 1.6.0). Optimizer, training epochs and batch size are set with “Adam”, 70 and 64, respectively. In our work, the optimizing function, “Adam”, use its default parameters value. All codes and trained models can be found via <https://github.com/Jane4747/DTI-BERT>.

## 3 RESULTS AND DISCUSSION

### 3.1 Performance Metrics

The determination of a pair belongs to an interactive drug-target pair or non-interactive drug-target pair, is in the case of single-label classification. The metrics such as accuracy (ACC), sensitivity (Sn), Specificity (Sp), strength (str, the average of Sn and Sp) and Matthew’s correlation coefficient (MCC) are frequently used. The specific formulas are as follows:



**TABLE 1 |** Results of comparison with several traditional machine learning methods on four datasets.

Dataset	Method	Sn(%)	Sp(%)	ACC(%)	Str (%)	MCC
GPCR	MLP	86.8	75.3	82.8	81.5	0.61
GPCR	LightGBM	87.5	80.5	86.3	84.0	0.67
GPCR	BRL + CNN	<b>89.3</b>	<b>91.0</b>	<b>90.1</b>	<b>90.2</b>	<b>0.80</b>
Ion channel	MLP	93.3	83.1	89.6	88.2	0.77
Ion channel	LightGBM	92.7	89.3	91.7	91.0	0.81
Ion channel	BRL + CNN	<b>95.9</b>	<b>91.4</b>	<b>94.7</b>	<b>93.7</b>	<b>0.87</b>
Enzyme	MLP	92.2	86.0	90.1	89.1	0.79
Enzyme	LightGBM	92.8	90.5	92.4	91.7	0.83
Enzyme	BRL + CNN	<b>95.9</b>	<b>92.0</b>	<b>94.9</b>	<b>94.0</b>	<b>0.88</b>
NR	MLP	84.2	76.9	79.9	80.6	0.60
NR	LightGBM	84.4	83.1	82.7	83.8	0.65
NR	BRL + CNN	<b>92.5</b>	<b>85.2</b>	<b>89.0</b>	<b>88.9</b>	<b>0.78</b>

The best results for each metric are in bold.

**TABLE 2 |** Performance comparison on four datasets inaccuracy rate.

Method	GPCRs	Ion-Channels	Enzymes	NR
He et al. (2010)	78.5	80.8	85.5	88.4
DrugRPE Zhang et al. (2017)	85.2	89.0	90.0	<b>91.1</b>
Hu et al. (2019)	88.4	91.9	94.3	85.7
Our method	<b>90.1</b>	<b>94.7</b>	<b>94.9</b>	89.0

The best results for each metric are in bold.

**TABLE 3 |** Performance comparison on GPCR dataset over leave-one-out cross-validation.

Method	Sn(%)	Sp(%)	ACC(%)	Str (%)	MCC
IGPCR-Drug Xiao et al. (2013)	78.3	91.4	86.9	84.9	0.71
OET-KNN Hu et al. (2016)	77.8	88.7	85.0	83.3	0.67
QuickRBF Hu et al. (2016)	74.8	92.4	86.4	83.6	0.69
SVM Hu et al. (2016)	74.2	92.7	86.4	83.6	0.69
RF Hu et al. (2016)	76.5	92.9	87.3	84.7	0.71
RF + PP Hu et al. (2016)	79.7	92.8	88.3	86.3	0.73
DWKNN(Ensemble) Wang et al. (2020)	81.1	87.1	85.1	84.1	0.67
BOW-GBDT Qiu et al. (2021)	79.8	<b>93.1</b>	88.5	86.3	0.74
Our method	<b>92.2</b>	92.0	<b>91.9</b>	<b>90.1</b>	<b>0.84</b>

The best results for each metric are in bold.

$$\begin{cases}
 Acc = \frac{TP + TN}{TP + TN + FP + FN} \\
 Sn = \frac{TP}{TP + FN} \\
 Sp = \frac{TP}{TP + FP} \\
 Str = \frac{Sp + Sn}{2} \\
 MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN) + (TN + FP)(TN + FN)}}
 \end{cases} \quad (18)$$

where TP represents the true positive, FN the false negative, TN the true negative, FP the false positive.

### 3.2 Comparison of Several Classic Protein and Drug Feature Extraction Methods

On the protein representation task, auto-encoder models (word2vec and BERT) with different model parameters scales were tested. For the drug representation task, a variety of algorithms in various fields, including natural language

**TABLE 4 |** Performance comparison on Check390.

Method	Sn(%)	Sp(%)	ACC(%)	Str (%)	MCC
IGPCR-Drug Xiao et al. (2013)	80.8	66.9	71.6	73.9	0.45
OET-KNN Hu et al. (2016)	67.7	84.2	78.7	76.9	0.52
QuickRBF Hu et al. (2016)	76.2	77.7	77.2	77.6	0.52
SVM Hu et al. (2016)	76.2	78.9	78.0	77.6	0.53
RF Hu et al. (2016)	78.5	78.1	78.2	78.3	0.54
RF + PPP Hu et al. (2016)	83.1	79.6	80.8	81.3	0.60
DWKNN Wang et al. (2020)	83.9	80.0	81.3	81.9	0.61
DWKNN(Ensemble) Wang et al. (2020)	83.1	82.7	82.8	82.9	0.63
BOW-GBDT Qiu et al. (2021)	80.0	<b>90.0</b>	86.7	85.0	0.70
Our method	<b>87.1</b>	89.4	<b>88.4</b>	<b>88.3</b>	<b>0.76</b>

The best results for each metric are in bold.

processing (word2vec), graph (node2vec and GCN), and signal processing (DWT) were tested.

We evaluated the BERT\_Mean + DWT feature extraction method and compared it with several other classic protein and drug feature extraction methods, such as Pr ord2vec (a 64-D vector is obtained to represent the protein, it was extracted by an un-supervised word2vec model and implicated important biophysical and biochemical information (Yang et al., 2018; Zhang et al., 2020), BERT\_First (the first row of  $P_{BERT}$  is obtained to represented protein, it is a 1024-D vector) (Nambiar et al., 2020), FP2\_Word2vec (Jaeger et al., 2018), drug\_Node2vec (Grover and Leskovec, 2016; Tetko et al., 2020), drug\_Word2vec (Zhang et al., 2020; Zheng et al., 2021), drug\_GCN (Chen et al., 2020). **Figures 3–6** show the Matthews correlation coefficient (MCC) for the datasets  $S_{GPCR-Drug}$ ,  $S_{Chl-Drug}$ ,  $S_{Ezy-Drug}$ , and  $S_{NR-Drug}$  obtained for each approach in CNN + BRL classifier via 10-fold cross validation.

It was found that BERT\_Mean for the proteins and DWT for drugs can improve the performance of the classifier greatly in four datasets. The BERT\_Mean + DWT increased capacity for identifying DTIs compared to the using BERT\_First, PRO\_Word2vec, drug\_Node2vec, drug\_Word2vec, and drug\_GCN, and BERT\_Mean can find the most compact and informative features subsets which are deeply hidden in protein sequences. It is showed that word2vec for protein sequences and GCN for drugs in DTIs tasks, could also obtain good prediction results on three datasets ( $S_{GPCR-Drug}$ ,  $S_{Ezy-Drug}$ , and  $S_{NR-Drug}$ ), which inspires us that different protein representation methods need to consider different drug molecule representation methods, which need to be determined experimentally.

### 3.3 Comparison With Some Machine Learning Methods

In order to test the performance of the BRL + CNN and compare it with the existing machine learning methods, we use the same benchmark dataset (listed in **Eq. 1**) and the same BERT\_Mean + DWT feature as the input of the prediction model. The proposed BRL + CNN predictor and other commonly used classifiers provided by the Scikit-learn library, like Multi-Layer Perceptron (MLP) with two hidden layers (Pedregosa et al., 2011) and gradient boosting tree-based ensemble method called LightGBM (LGB) (Ke et al., 2017), were tested via 10-

fold cross-validation, the results are listed in **Table 1**. It was found that the proposed BRL + CNN predictor in this article has better performance than other classifiers in all metrics.

### 3.4 Comparison With Existing Predictor

To further demonstrate the power of the DTI-BERT predictor, we compared it with some existing methods. There are some new models for identifying DTIs trained with the datasets established by He et al. (He et al., 2010). For example, Hu et al. proposed a deep learning-based method to predict DTIs by using the information of drug structures and proteins sequences (Hu et al., 2019), this CnnDIT predictor has better prediction performance in predicting DTIs, and it has its own web server. Zhang et al. proposed a random projection ensemble approach DrugRPE to predict DTIs (Zhang et al., 2017), and several random projections build an ensemble REPTress system. In general, the method of fusing multiple predictors outperforms a single predictor. To facilitate comparison, the scores of accuracies (defined in **Eq. (18)**) obtained by these three predictors (He et al., 2010; Hu et al., 2016; Zhang et al., 2017) based on the benchmark datasets used in He et al. (He et al., 2010) via the 10-fold cross-validation test were listed in **Table 2**. Comprehensively, the comparative results showed that our model is more accurate than other existing methods.

GPCRs have proved to be one of the most important target families of modern drugs. Identifying the GPR-drug interaction is an important issue in bioinformatics, and a number of researchers have proposed effective predicted methods to identify GPCR-drug predictions. Our method was also compared with the performance of different methods which predicting GPCR-drug interaction on the training dataset  $S_{GPCR-Drug}$  over leave-one-out cross-validation, and validated in independent test dataset check390 (Xiao et al., 2013; Hu et al., 2016; Wang et al., 2020; Qiu et al., 2021). The results of the different methods tested on  $S_{GPCR-Drug}$  over leave-one-out cross-validation were shown in **Table 3**. The results of the other eight methods were reported in (Qiu et al., 2021). From **Table 3**, we can find that the MCC values of our method were 10% higher than others.

The generalization ability of machine learning models is usually evaluated through an independent test. The D92M is the GPCR-drug interaction dataset in (Wang et al., 2020), which is applied as a training dataset, and check390 as a validation

dataset. The results of the validation test on check390 were listed in **Table 4**, which demonstrated that our method almost outperform the others across the five metrics, except for BOW-GBDT achieves the highest value of Sp (93.1%). Compared with other state-of-the-art methods, the ACC value of our method is 3.4% higher, the MCC value is 6% higher than the second one. All these results demonstrate the effectiveness of the proposed methods.

## 4 CONCLUSION

In this work, we developed a powerful predictor based on the sequences of proteins and FP2 of drugs. We attempted to use pre-trained BERT to present proteins in DTIs and choose a useful representation for drugs via extensive experiments, including several state-of-art drug descriptions like drug\_Word2vec, drug\_Node2vec, drug\_GCN, FP2\_Word2vec, FP2\_DWT. The presenting results showed that FP2\_DWT is more efficient to present drug molecules than other descriptions. Furthermore, we used the deep learning method to generate interaction information and optimized the predicting network based on contrastive loss and cross-entropy loss, which performed much better than other common machine learning models. Moreover, compared with other existing predictors, DTI-BERT has better prediction performance in different target families of GPCRs, ion channels, enzymes and nuclear receptors, without any help of prior knowledge and handcrafted feature engineering. Overall, DTI-BERT can predict drug-target interactions that achieved high accuracy and we established a prediction web-server for the convenience of the most experienced scientists.

## REFERENCES

- Alaimo, S., Pulvirenti, A., Giugno, R., and Ferro, A. (2013). Drug-target Interaction Prediction through Domain-Tuned Network-Based Inference. *Bioinformatics* 29 (16), 2004–2008. doi:10.1093/bioinformatics/btt307
- Bianchi, F., Terragni, S., Hovy, D., and Assoc Computat, L. (2021). Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. in Joint Conference of 59th Annual Meeting of the Association-for-Computational-Linguistics (ACL)/11th International Joint Conference on Natural Language Processing (IJCNLP)/6th Workshop on Representation Learning for NLP (ReL4NLP), Aug 01–06 2021. (Electr Network), 759–766.
- Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. (2022). ProteinBERT: A Universal Deep-Learning Model of Protein Sequence and Function. *Bioinformatic*. 38 (8), 2102–2110. doi:10.1101/2021.05.24.445464
- Butina, D., Segall, M. D., and Frankcombe, K. (2002). Predicting ADME Properties In Silico: Methods and Models. *Drug Discov. today* 7 (11), S83–S88. doi:10.1016/s1359-6446(02)02288-2
- Chen, L., Tan, X., Wang, D., Zhong, F., Liu, X., Yang, T., et al. (2020). TransformerCPI: Improving Compound-Protein Interaction Prediction by Sequence-Based Deep Learning with Self-Attention Mechanism and Label Reversal Experiments. *Bioinformatics* 36 (16), 4406–4414. doi:10.1093/bioinformatics/btaa524
- Chen, R., Liu, X., Jin, S., Lin, J., and Liu, J. (2018). Machine Learning for Drug-Target Interaction Prediction. *Molecules* 23 (9), 2208. doi:10.3390/molecules23092208
- Chen, X., Sun, Y. Z., Zhang, D. H., Li, J. Q., Yan, G. Y., An, J. Y., et al. (2017). NRDTD: a Database for Clinically or Experimentally Supported Non-coding RNAs and Drug Targets Associations. *Database (Oxford)* 2017, bax057. doi:10.1093/database/bax057

The BERT model has very excellent general capabilities and has very outstanding feature extraction capabilities for DNA sequences (Le et al., 2021) and RNA sequences (Zhang et al., 2021). The DTIs prediction framework proposed in this paper has very good potential for predicting other drug targets as well.

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: <http://121.36.221.79/dtibert/download>.

## AUTHOR CONTRIBUTIONS

XX conceived and designed the experiments, JZ performed the extraction of features, model construction, model training, and evaluation. JZ drafted the manuscript, XX and W-RQ supervised this project and revised the manuscript. All authors read and approved the final manuscript.

## FUNDING

This work was supported by the grants from the National Natural Science Foundation of China (Nos 31860312, 62162032, and 62062043), Natural Science Foundation of Jiangxi Province, China (NO. 20202BAB202007), the International Cooperation Project of the Ministry of Science and Technology, China (NO. 2018-3-3).

- Chen, X., Liu, M.-X., and Yan, G.-Y. (2012). Drug-target Interaction Prediction by Random Walk on the Heterogeneous Network. *Mol. Biosyst.* 8 (7), 1970–1978. doi:10.1039/c2mb00002d
- Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., and Salakhutdinov, R. (2019). Acl: Transformer-xl: Attentive Language Models beyond a Fixed-Length Context. in 57th Annual Meeting of the Association-for-Computational-Linguistics (ACL): 2019, Florence, ITALY, Jul 28–Aug 02 2019, 2978–2988.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint. *arXiv:1810.04805* 2018.
- Dobson, C. M. (2004). Chemical Space and Biology. *Nature* 432 (7019), 824–828. doi:10.1038/nature03192
- Dubach, J. M., Vinegoni, C., Mazitschek, R., Fumene Feruglio, P., Cameron, L. A., and Weissleder, R. (2014). In Vivo imaging of Specific Drug-Target Binding at Subcellular Resolution. *Nat. Commun.* 5 (1), 3946–3949. doi:10.1038/ncomms4946
- Elnaggar, A., Heinzinger, M., Dallago, C., Rihawi, G., Wang, Y., Jones, L., et al. (2021). ProtTrans: Towards Cracking the Language of Life's Code through Self-Supervised Deep Learning and High Performance Computing. in IEEE Transactions on Pattern Analysis and Machine Intelligence 2021.
- Ezzat, A., Zhao, P., Wu, M., Li, X. L., and Kwok, C. K. (2016). Drug-target Interaction Prediction with Graph Regularized Matrix Factorization. *IEEE/ACM Trans. Comput. Biol. Bioinform* 14 (3), 646–656. doi:10.1109/TCBB.2016.2530062
- Grover, A., and Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks. in 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD): 2016, San Francisco, CA, Aug 13–17 2016, 855–864. doi:10.1145/2939672.2939754
- Haggarty, S. J., Koeller, K. M., Wong, J. C., Butcher, R. A., and Schreiber, S. L. (2003). Multidimensional Chemical Genetic Analysis of Diversity-Oriented



- Synthesis-Derived Deacetylase Inhibitors Using Cell-Based Assays. *Chem. Biol.* 10 (5), 383–396. doi:10.1016/s1074-5521(03)00095-4
- He, Z., Zhang, J., Shi, X.-H., Hu, L.-L., Kong, X., Cai, Y.-D., et al. (2010). Predicting Drug-Target Interaction Networks Based on Functional Groups and Biological Features. *PLoS one* 5 (3), e9603. doi:10.1371/journal.pone.0009603
- Hu, J., Li, Y., Yang, J.-Y., Shen, H.-B., and Yu, D.-J. (2016). GPCR-drug Interactions Prediction Using Random Forest with Drug-Association-Matrix-Based Post-processing Procedure. *Comput. Biol. Chem.* 60, 59–71. doi:10.1016/j.compbiolchem.2015.11.007
- Hu, S., Zhang, C., Chen, P., Gu, P., Zhang, J., and Wang, B. (2019). Predicting Drug-Target Interactions from Drug Structure and Protein Sequence Using Novel Convolutional Neural Networks. *BMC Bioinforma.* 20 (25), 689. doi:10.1186/s12859-019-3263-x
- Ilić, S., Marrese-Taylor, E., Balazs, J. A., and Matsuo, Y. (2018). Deep Contextualized Word Representations for Detecting Sarcasm and Irony. in *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2–7.
- Ioffe, S., and Szegedy, C. (2015). “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.” in *32nd International Conference on Machine Learning*: 2015, Lille, France, Jul 07–09 2015, 448–456.
- Jaeger, S., Fulle, S., and Turk, S. (2018). Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.* 58 (1), 27–35. doi:10.1021/acs.jcim.7b00616
- Karlov, D. S., Sosnin, S., Fedorov, M. V., and Popov, P. (2020). graphDelta: MPNN Scoring Function for the Affinity Prediction of Protein-Ligand Complexes. *ACS omega* 5 (10), 5150–5159. doi:10.1021/acsomega.9b04162
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). Lightgbm: A Highly Efficient Gradient Boosting Decision Tree. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. (Long Beach, CA, USA: Curran Associates Inc.), 3149–3157.
- Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J., and Shoichet, B. K. (2007). Relating Protein Pharmacology by Ligand Chemistry. *Nat. Biotechnol.* 25 (2), 197–206. doi:10.1038/nbt1284
- Kim, S., Jin, D., and Lee, H. (2013). Predicting Drug-Target Interactions Using Drug-Drug Interactions. *PLoS one* 8 (11), e80129. doi:10.1371/journal.pone.0080129
- Kombo, D. C., Tallapragada, K., Jain, R., Chewing, J., Mazurov, A. A., Speake, J. D., et al. (2013). 3D Molecular Descriptors Important for Clinical Success. *J. Chem. Inf. Model.* 53 (2), 327–342. doi:10.1021/ci300445e
- Kotlyar, M., Fortney, K., and Jurisica, I. (2012). Network-based Characterization of Drug-Regulated Genes, Drug Targets, and Toxicity. *Methods* 57 (4), 499–507. doi:10.1016/j.ymeth.2012.06.003
- Le, N. Q. K., Ho, Q. T., Nguyen, T. T., and Ou, Y. Y. (2021). A Transformer Architecture Based on BERT and 2D Convolutional Neural Network to Identify DNA Enhancers from Sequence Information. *Brief. Bioinform* 22 (5), bbab005. doi:10.1093/bib/bbab005
- Lee, C., Cho, K., and Kang, W. (2019). Mixout: Effective Regularization to Finetune Large-Scale Pretrained Language Models. in *International Conference on Learning Representations (ICLR)*: 2020. (International Conference on Learning Representations).
- Li, S., Wan, F., Shu, H., Jiang, T., Zhao, D., and Zeng, J. (2020). MONN: a Multi-Objective Neural Network for Predicting Compound-Protein Interactions and Affinities. *Cell Syst.* 10 (4), 308–322. doi:10.1016/j.cels.2020.03.002
- Lim, S., Lu, Y., Cho, C. Y., Sung, I., Kim, J., Kim, Y., et al. (2021). A Review on Compound-Protein Interaction Prediction Methods: Data, Format, Representation and Model. *Comput. Struct. Biotechnol. J.* 19, 1541–1556. doi:10.1016/j.csbj.2021.03.004
- Liu, T., Lin, Y., Wen, X., Jorissen, R. N., and Gilson, M. K. (2007). BindingDB: a Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Res.* 35 (Suppl. 1), D198–D201. doi:10.1093/nar/gkl999
- Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., et al. (2017). A Network Integration Approach for Drug-Target Interaction Prediction and Computational Drug Repositioning from Heterogeneous Information. *Nat. Commun.* 8 (1), 573. doi:10.1038/s41467-017-00680-8
- Mei, J.-P., Kwok, C.-K., Yang, P., Li, X.-L., and Zheng, J. (2013). Drug-target Interaction Prediction by Learning from Local Information and Neighbors. *Bioinformatics* 29 (2), 238–245. doi:10.1093/bioinformatics/bts670
- Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., et al. (2009). AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* 30 (16), 2785–2791. doi:10.1002/jcc.21256
- Mousavian, Z., and Masoudi-Nejad, A. (2014). Drug-target Interaction Prediction via Chemogenomic Space: Learning-Based Methods. *Expert Opin. drug metabolism Toxicol.* 10 (9), 1273–1287. doi:10.1517/17425255.2014.950222
- Nambiar, A., Heflin, M., Liu, S., Maslov, S., Hopkins, M., and Ritz, A. (2020). “Transforming the Language of Life: Transformer Neural Networks for Protein Prediction Tasks,” in *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 1–8.
- Noorbakhsh, J., Farahmand, S., Foroughi Pour, A., Namburi, S., Caruana, D., Rimm, D., et al. (2020). Deep Learning-Based Cross-Classifications Reveal Conserved Spatial Behaviors within Tumor Histological Images. *Nat. Commun.* 11 (1), 6367. doi:10.1038/s41467-020-20030-5
- Nozza, D., Bianchi, F., and Hovy, D. (2020). *What the [mask]? Making Sense of Language-specific BERT Models*. arXiv preprint. arXiv:200302912 2020.
- Öztürk, H., Ozkirimli, E., and Özgür, A. (2016). A Comparative Study of SMILES-Based Compound Similarity Functions for Drug-Target Interaction Prediction. *BMC Bioinforma.* 17 (1), 1–11. doi:10.1186/s12859-016-0977-x
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.
- Peng, Y., Rios, A., Kavuluru, R., and Lu, Z. (2018). Extracting Chemical-Protein Relations with Ensembles of SVM and Deep Learning Models. *Database: J. Biol. Databases curation* 2018, bay073. doi:10.1093/database/bay073
- Peng, Y., Yan, S., and Lu, Z. (2019). *Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets*. arXiv preprint. arXiv:190605474 2019.
- Pujadas, G., Vaque, M., Ardevol, A., Blade, C., Salvado, M., Blay, M., et al. (2008). Protein-ligand Docking: A Review of Recent Advances and Future Perspectives. *Cpa* 4 (1), 1–19. doi:10.2174/157341208783497597
- Qin, C., Zhang, C., Zhu, F., Xu, F., Chen, S. Y., Zhang, P., et al. (2014). Therapeutic Target Database Update 2014: a Resource for Targeted Therapeutics. *Nucl. Acids Res.* 42 (D1), D1118–D1123. doi:10.1093/nar/gkt1129
- Qiu, W., Lv, Z., Hong, Y., Jia, J., and Xiao, X. (2021). A GBDT Classifier Combining with Artificial Neural Network for Identifying GPCR-Drug Interaction Based on Wordbook Learning from Sequences. *Front. Cell Dev. Biol.* 8, 1789. doi:10.3389/fcell.2020.623858
- Qu, J., Chen, X., Sun, Y.-Z., Zhao, Y., Cai, S.-B., Ming, Z., et al. (2019). In Silico Prediction of Small Molecule-miRNA Associations Based on the HeteSim Algorithm. *Mol. Ther. - Nucleic Acids* 14, 274–286. doi:10.1016/j.omtn.2018.12.002
- Rognan, D. (2007). Chemogenomic Approaches to Rational Drug Design. *Br. J. Pharmacol.* 152 (1), 38–52. doi:10.1038/sj.bjp.0707307
- Sawada, R., Kotera, M., and Yamanishi, Y. (2014). Benchmarking a Wide Range of Chemical Descriptors for Drug-Target Interaction Prediction Using a Chemogenomic Approach. *Mol. Inf.* 33 (11–12), 719–731. doi:10.1002/minf.201400066
- Tetko, I. V., Karpov, P., Van Deursen, R., and Godin, G. (2020). State-of-the-art Augmented NLP Transformer Models for Direct and Single-step Retrosynthesis. *Nat. Commun.* 11 (1), 5575. doi:10.1038/s41467-020-19266-y
- Tornø, W., and Altman, R. B. (2019). Graph Convolutional Neural Networks for Predicting Drug-Target Interactions. *J. Chem. Inf. Model.* 59 (10), 4131–4149. doi:10.1021/acs.jcim.9b00628
- Wan, F., Hong, L., Xiao, A., Jiang, T., and Zeng, J. (2019). NeoDTI: Neural Integration of Neighbor Information from a Heterogeneous Network for Discovering New Drug-Target Interactions. *Bioinformatics* 35 (1), 104–111. doi:10.1093/bioinformatics/bty543
- Wang, C.-C., and Chen, X. (2019). A Unified Framework for the Prediction of Small Molecule-MicroRNA Association Based on Cross-Layer Dependency Inference on Multilayered Networks. *J. Chem. Inf. Model.* 59 (12), 5281–5293. doi:10.1021/acs.jcim.9b00667

- Wang, P., Huang, X., Qiu, W., and Xiao, X. (2020). Identifying GPCR-Drug Interaction Based on Wordbook Learning from Sequences. *BMC Bioinforma.* 21 (1), 150. doi:10.1186/s12859-020-3488-8
- Wang, Y.-C., Yang, Z.-X., Wang, Y., and Deng, N.-Y. (2010). Computationally Probing Drug-Protein Interactions via Support Vector Machine. *Lddd* 7 (5), 370–378. doi:10.2174/157018010791163433
- Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a Major Update to the DrugBank Database for 2018. *Nucleic acids Res.* 46 (D1), D1074–D1082. doi:10.1093/nar/gkx1037
- Xiao, X., Min, J.-L., Wang, P., and Chou, K.-C. (2013). iGPCR-Drug: A Web Server for Predicting Interaction between GPCRs and Drugs in Cellular Networking. *PLoS one* 8 (8), e72234. doi:10.1371/journal.pone.0072234
- Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., and Kanehisa, M. (2008). Prediction of Drug-Target Interaction Networks from the Integration of Chemical and Genomic Spaces. *Bioinformatics* 24 (13), i232–i240. doi:10.1093/bioinformatics/btn162
- Yamanishi, Y., Kotera, M., Moriya, Y., Sawada, R., Kanehisa, M., and Goto, S. (2014). DINIES: Drug-Target Interaction Network Inference Engine Based on Supervised Analysis. *Nucleic acids Res.* 42 (W1), W39–W45. doi:10.1093/nar/gku337
- Yang, K. K., Wu, Z., Bedbrook, C. N., and Arnold, F. H. (2018). Learned Protein Embeddings for Machine Learning. *Bioinformatics* 34 (15), 2642–2648. doi:10.1093/bioinformatics/bty178
- Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). “Xlnet: Generalized Autoregressive Pretraining for Language Understanding” in *Advances in Neural Information Processing Systems*. Editor H. Wallach, H. Larochelle, A. Beygelzimer, F. d. {e}-Buc, E. Fox, and R. Garnett (Curran Associates, Inc.) 32.
- Yin, J., Chen, X., Wang, C.-C., Zhao, Y., and Sun, Y.-Z. (2019). Prediction of Small Molecule-MicroRNA Associations by Sparse Learning and Heterogeneous Graph Inference. *Mol. Pharm.* 16 (7), 3157–3166. doi:10.1021/acs.molpharmaceut.9b00384
- Zhang, J., Zhu, M., Chen, P., and Wang, B. (2017). DrugRPE: Random Projection Ensemble Approach to Drug-Target Interaction Prediction. *Neurocomputing* 228, 256–262. doi:10.1016/j.neucom.2016.10.039
- Zhang, L., Qin, X., Liu, M., Liu, G., and Ren, Y. (2021). BERT-m7G: A Transformer Architecture Based on BERT and Stacking Ensemble to Identify RNA N7-Methylguanosine Sites from Sequence Information. *Comput. Math. Methods Med.* 2021, 7764764. doi:10.1155/2021/7764764
- Zhang, W., Zou, H., Luo, L., Liu, Q., Wu, W., and Xiao, W. (2016). Predicting Potential Side Effects of Drugs by Recommender Methods and Ensemble Learning. *Neurocomputing* 173, 979–987. doi:10.1016/j.neucom.2015.08.054
- Zhang, Y.-F., Wang, X., Kaushik, A. C., Chu, Y., Shan, X., Zhao, M.-Z., et al. (2020). SPVec: a Word2vec-Inspired Feature Representation Method for Drug-Target Interaction Prediction. *Front. Chem.* 7, 895. doi:10.3389/fchem.2019.00895
- Zheng, J., Xiao, X., and Qiu, W. R. (2021). iCD1-W2vCom: Identifying the Ion Channel-Drug Interaction in Cellular Networking Based on Word2vec and Node2vec. *Front. Genet.* 12, 738274. doi:10.3389/fgene.2021.738274
- Zhou, X., Dai, E., Song, Q., Ma, X., Meng, Q., Jiang, Y., et al. (2020). In Silico drug Repositioning Based on Drug-miRNA Associations. *Briefings Bioinforma.* 21 (2), 498–510. doi:10.1093/bib/bbz012
- Zhu, S., Okuno, Y., Tsujimoto, G., and Mamitsuka, H. (2005). A Probabilistic Model for Mining Implicit ‘chemical Compound-Gene’ Relations from Literature. *Bioinformatics* 21 (Suppl. 1\_2), ii245. doi:10.1093/bioinformatics/bti1141

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

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

Copyright © 2022 Zheng, Xiao and Qiu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Profiling a Community-Specific Function Landscape for Bacterial Peptides Through Protein-Level Meta-Assembly and Machine Learning

Mitra Vajjala<sup>1†</sup>, Brady Johnson<sup>1†</sup>, Lauren Kasperek<sup>1</sup>, Michael Leuze<sup>2</sup> and Qiuming Yao<sup>1\*</sup>

<sup>1</sup>School of Computing, University of Nebraska-Lincoln, Lincoln, NE, United States, <sup>2</sup>Nashville Biosciences, Nashville, TN, United States

## OPEN ACCESS

### Edited by:

Ruiquan Ge,  
Hangzhou Dianzi University, China

### Reviewed by:

Xuefeng Cui,  
Shandong University, China  
Rodrigo Bentes Kato,  
Federal University of Minas Gerais,  
Brazil

### \*Correspondence:

Qiuming Yao  
qyao3@unl.edu

<sup>†</sup>These authors have contributed  
equally to this work

### Specialty section:

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

**Received:** 03 May 2022

**Accepted:** 17 June 2022

**Published:** 22 July 2022

### Citation:

Vajjala M, Johnson B, Kasperek L,  
Leuze M and Yao Q (2022) Profiling a  
Community-Specific Function  
Landscape for Bacterial Peptides  
Through Protein-Level Meta-Assembly  
and Machine Learning.  
Front. Genet. 13:935351.  
doi: 10.3389/fgene.2022.935351

Small proteins, encoded by small open reading frames, are only beginning to emerge with the current advancement of omics technology and bioinformatics. There is increasing evidence that small proteins play roles in diverse critical biological functions, such as adjusting cellular metabolism, regulating other protein activities, controlling cell cycles, and affecting disease physiology. In prokaryotes such as bacteria, the small proteins are largely unexplored for their sequence space and functional groups. For most bacterial species from a natural community, the sample cannot be easily isolated or cultured, and the bacterial peptides must be better characterized in a metagenomic manner. The bacterial peptides identified from metagenomic samples can not only enrich the pool of small proteins but can also reveal the community-specific microbe ecology information from a small protein perspective. In this study, metaBP (Bacterial Peptides for metagenomic sample) has been developed as a comprehensive toolkit to explore the small protein universe from metagenomic samples. It takes raw sequencing reads as input, performs protein-level meta-assembly, and computes bacterial peptide homolog groups with sample-specific mutations. The metaBP also integrates general protein annotation tools as well as our small protein-specific machine learning module metaBP-ML to construct a full landscape for bacterial peptides. The metaBP-ML shows advantages for discovering functions of bacterial peptides in a microbial community and increases the yields of annotations by up to five folds. The metaBP toolkit demonstrates its novelty in adopting the protein-level assembly to discover small proteins, integrating protein-clustering tool in a new and flexible environment of RBiotoools, and presenting the first-time small protein landscape by metaBP-ML. Taken together, metaBP (and metaBP-ML) can profile functional bacterial peptides from metagenomic samples with potential diverse mutations, in order to depict a unique landscape of small proteins from a microbial community.

**Keywords:** bacterial peptide, machine learning, metagenomics, protein annotation, protein clustering

# 1 INTRODUCTION

Small proteins or peptides, translated from short open reading frames, largely exist in biological systems in both eukaryotes (Chen et al., 2020) and prokaryotes (Hemm et al., 2020; Orr et al., 2021). Historically, these small proteins were ignored or identified as non-coding elements (Storz et al., 2014) and were considered as “dark matter” due to the lack of genomic annotation (Garai and Blanc-Potard, 2020). Bacteria-derived small proteins can play diverse roles in microbial functions and host-microbe interactions, such as innate immunity (Huan et al., 2020), cell division, signal transduction, transporter regulation, enzymatic activity, and protein folding (Storz et al., 2014). Some of the bacterial peptides have the potential of being novel therapeutic candidates (Duval and Cossart, 2017).

Bacterial peptides are much harder to decompose and they annotate in a natural community. While detecting and testing a small gene can be difficult in a single organism, microbiome at community level brings additional challenges in the data complexity and sparsity for small protein detection, classification, and function annotation. Metagenomics from short gun sequencing provides information from the community-specific population to gene functions, but there haven't been many previous efforts specifically focusing on the role of bacterial peptides from a natural community. The lack of detection power and poor analytical resolution indicate the limitation from both the computation and experiment. First, the peptides detection from mass spectrometry needs abundant input materials and suffers from large search spaces in an unbiased and untargeted scenario. It usually requires a confident database from reference genomes or from metagenomes. Poor annotation of small genes in reference genomes is also an obstacle of the direct detection from mass spectrometry. Even by combining multiple types of omics data, the false positives can still be high in small bacterial peptides detection (Miravet-Verde et al., 2019). Second, the protein calling tools for metagenomics may require high quality of the assembly results. Especially some of them are optimized for long contigs and scaffolds (Hyatt et al., 2012). Recently, a large-scale study for bacterial peptides from metagenomic samples reported more than 4,000 novel small-protein families were found from human microbiome and less than 5% of the proteins could be mapped to known domains. However, they still used contigs as input data from the nucleotide-level metagenomic assembly, which can lose a large amount of original sequencing data due to the sample complexity and sparsity. Third, for homologous searching and function annotation (Cantalapiedra et al., 2021), there is not a specific tool designed for exploring and mapping to the space of small bacterial peptides.

In order to address the limitations from the nucleotide-level metagenomic assembly and the current shortages of small protein annotation from microbe communities, metaBP (Bacterial Peptides for metagenomic sample) has been developed as a comprehensive and user-friendly toolkit to explore the small protein universe in a more thorough and detailed way. The metaBP applies protein-level assembly from the metagenomic sequencing data to maximize the protein recovery and search from the open reading frames (Steinegger et al., 2019). The metaBP identifies confident small protein sequences and mutations in diverse homologous clusters

using the most current protein sequence clustering technique (Steinegger and Söding, 2018). The metaBP also contains a machine learning part, metaBP-ML, to address the sequence-based annotation integrating a natural language-based protein embedding model (Rives et al., 2021) with a million-sized database. Diverse small protein sequences and functions are demonstrated in various sets of samples, which cover mice, human, and environmental microbiome communities. The metaBP provides the capability to explore the small protein landscape both at the microbial community scale and at the base pair resolution.

# 2 MATERIALS AND METHODS

## 2.1 Toolkit Implementation Overview

The metaBP is an integrated and automated toolkit for identifying and annotating small proteins from the metagenomic sequencing data. MetaBP's implementation consists of three major modules (metaBP, metaBP-ML, and RBiotoools), and five main procedures (**Figure 1**): protein meta-assembly, protein clustering, mutation calling, protein embedding, and protein annotation. The first three procedures to identify small proteins along with mutations are from our major module, i.e., metaBP; the last two procedures to do protein embedding and annotation are integrated in our machine learning-based module metaBP-ML. The entire toolkit is implemented by both Python and R, and the machine learning module requires pyTorch. The most convenient way to install and use the metaBP, metaBP-ML and RBiotoools is to configure their individual Conda environment, which are described in our GitHub repository (see the data availability for our GitHub link).

## 2.2 Input and Output

The input data for metaBP is the raw sequencing reads (paired-end short gun sequencing) in a FASTQ format. The output data consists of mainly three parts of the information, protein clusters with mutations, small protein annotations, and a protein copy number table from annotations, which will be demonstrated in this study. For the purpose of this study, only the small protein analysis is mentioned and emphasized. In fact, the metaBP toolkit can also identify the entire proteome wide space of open reading frames, other than just small proteins.

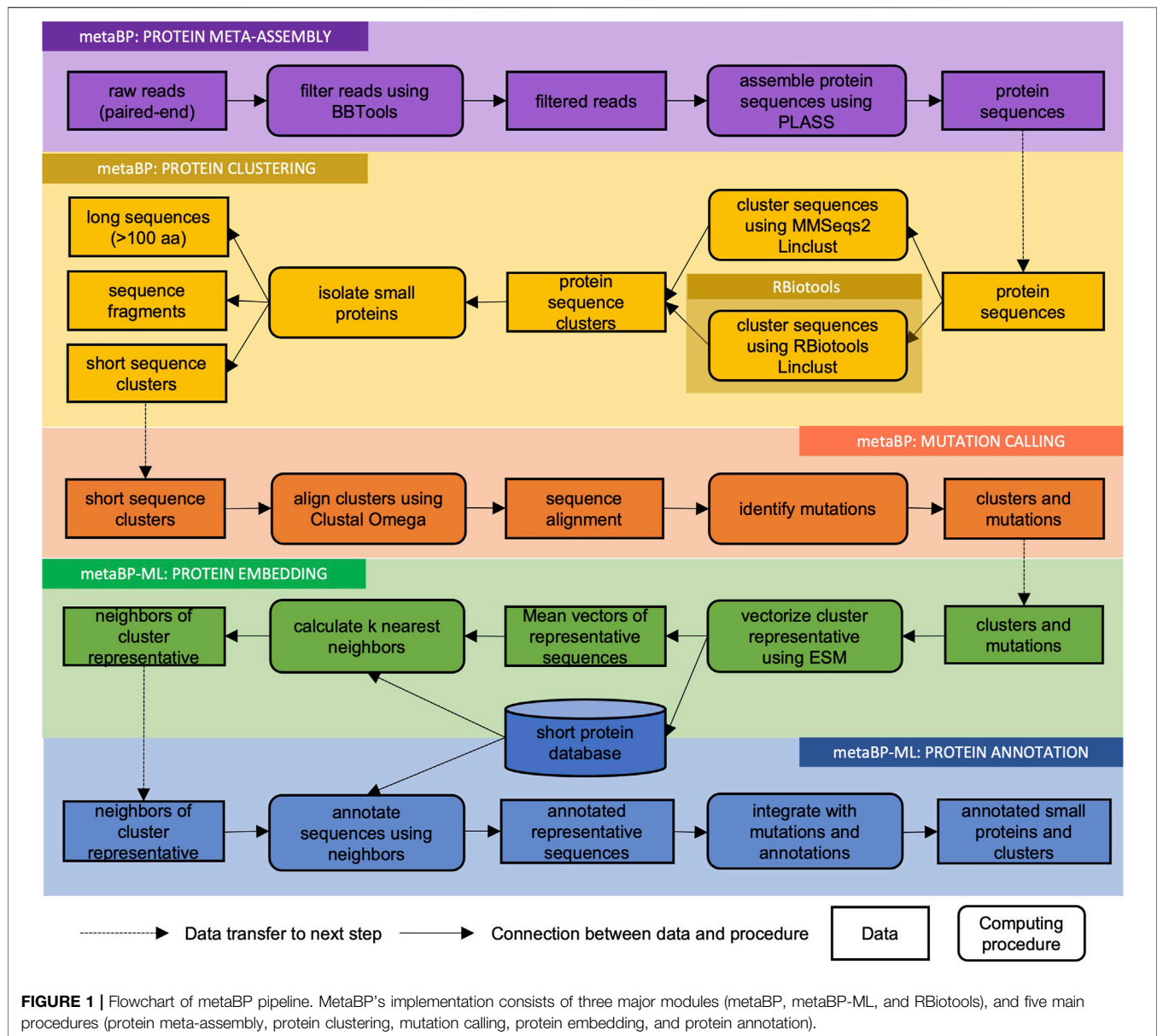
The raw FASTQ files used in this study are downloaded from NCBI SRA by the sra-toolkit. The sample IDs and general sequencing information are summarized in the supplementary table (**Supplementary Table S1**), with different read lengths and data volumes. This indicates our metaBP can be generalized to all types of metagenomics from natural environments.

## 2.3 Small Protein Identification and Clustering by MetaBP

### 2.3.1 Protein-Level Meta-Assembly

Raw sequencing reads in FASTQ files are pre-processed by BBTools (Bushnell et al., 2017). The pre-processing includes quality checking, read length trimming, and adaptor removal. The cleaned reads are used in the protein level assembly by PLASS





(Steinegger et al., 2019), which is reported to increase the protein yields by many-folds compared to the nucleotide-level metagenomic assembly. When an example data set from PLASS GitHub (see the data availability) is used to do protein level assembly, 99% identity in the sequence only yields 780 proteins, while 90% and 80% identity yield 1,217 and 1,267 proteins, respectively (**Supplementary Table S2**). Specifically, 80% of the sequence identity triples the number of non-single clusters. In order to maximize the initial protein throughput and capture the diversity inside average protein clusters, 80% identity in the sequence is recommended using in the metaBP toolkit. This setting can be changed by user's specific needs in terms of the protein recovery volume.

### 2.3.2 Protein Clustering and Mutation Calling

The assembled protein sequences are used in the clustering process. Linclust is one of the most recent protein clustering techniques that can approach both the good accuracy and linear time complexity (Steinegger and Söding, 2018). The metaBP has two ways to call Linclust procedure: one is from MMSeqs2 command line, and the other is from our independent implementation in RBiotoools. These two different ways to call Linclust provide different flexibility to the user side. The R version of Linclust inside the RBiotoools does require additional installation of the R environment, but it is more flexible for user to develop new applications and change parameters from the source code.



The protein clustering by Linclust has two purposes: one is to remove the redundancy from proteins and protein fragments, and the other is to group protein families by homology for mutations. The protein sequences generated from the PLASS example dataset are duplicated to test the effects of truncated sequences. Each protein is truncated up to 50% of the total length from either beginning or end of the sequence, mixed with their intact versions, and then they are clustered by Linclust at different settings. When the default parameter for Linclust is applied, the sequences truncated to 90% of the original length can still be clustered with its full-length version, but sequences truncated to 80% of the length cannot be clustered well. When using the customized setting in Linclust and setting the coverage rate to 50%, most of the truncated protein can still be clustered with the original full-length protein (**Supplementary Table S3**). For small protein clustering, the default parameters in Linclust are recommended to use the metaBP in order to make the small protein families more specific and sensitive. The user can always change the parameters to accommodate various protein lengths in a cluster. After the protein clusters are generated, the protein sequences are aligned by Clustal Omega (Sievers and Higgins, 2018), and the positions with conservative amino acids or the positions with potential mutations can be observed and reported. By randomly mutating the protein sequences, it is confirmed that Linclust can capture up to 5% of the sequence mutation in the same cluster (**Supplementary Table S4**). This implies that the final small-protein clusters obtained from the samples can represent a protein family with diverse sequences of at most five amino acid mutations.

The strategy to isolate small proteins from metagenomics data is as follows. First, sequences with longer than 100 amino acids are separated. Second, short sequences clustered with long sequences are removed so that the protein fragments can be minimized in the final output. Third, in this study, only protein clusters with four or more protein members are considered as confident protein families. This means that the same small protein should occur at least four times in a single sample. In addition, only protein clusters with a large size can display a meaningful sequence diversity. On average, after these criteria are applied to the datasets, less than 5% from the metagenomics data are small bacteria peptides, which is consistent with the study from MAGs (metagenome-assembled genomes) or contigs (Sberro et al., 2019).

## 2.4 Machine Learning-Based Annotation by metaBP-ML

### 2.4.1 Database Construction

The database for small protein sequences (not more than 100 amino acids) is constructed from the sequence files in the FASTA format downloaded from the Uniprot (Swiss-Prot and TrEMBL, November, 2021) (Bateman et al., 2021). In total, 16,565,616 sequences are downloaded for bacteria, 785,496 for archaea, 1,201,161 for virus, and 596,067 for metagenomics. Among these short sequences, 8,486,746 have species or function annotations. The rest of the 10,661,593 proteins without any annotation (“uncharacterized” or “unannotated”) is removed first.

Among annotated small proteins, Linclust is used to remove 80% of redundant sequences by clustering, and 3,682,960 proteins are eventually survived to form our final small protein sequence database.

As a transformer-based machine learning model inspired from natural language processing, ESM (Rives et al., 2021) is used to convert the sequences in the database to numerical vectors. In order to process 3 million of small proteins in the database, parallel computing with multiple threads is used to speed up the procedure. The resulted vectors for each small protein are 1,280 numerical values in length, and the principal components are computed in order to visualize the entire small protein database or landscape in a two-dimensional space. To our knowledge, before our study, this small protein landscape hasn't ever shown nor used in the small protein annotation.

### 2.4.2 Protein Embedding and Annotation

After the database is constructed with vectorized small protein sequences, small proteins from metagenomic samples must be processed in the same ESM model (Rives et al., 2021). Each of the small protein cluster is vectorized by its representative sequence and then it can be embedded to the entire small protein universe spanned by the database. For downstream protein annotation, user can select one of the two ways in metaBP-ML. The first one is to use an HMM based tool, i.e., eggNOG (Cantalapiedra et al., 2021), which is for general protein annotations as well as for small proteins. The second method is to search for  $k$  nearest neighbors (KNN) from our constructed database for each cluster representative. Since this requires calculating all pairs of vector distances, it can be time consuming for a larger  $k$ . From our simple test, using a mice gut microbial sample (Morissette et al., 2020), the newly recovered protein annotations drops to less than 10% when pursuing 10 neighbors (**Supplementary Table S5**). In metaBP-ML, top ten nearest neighbors are recommended for small protein annotations.

The final protein annotation strategy based on the ten nearest neighbors is heuristic. First, rule of thumb is used if there is a most frequent annotation in the neighborhood. Second, if there is no difference between annotation frequencies, the top annotation is always picked. Third, if useful annotation cannot be extracted from the top ten neighbors, the protein will be left as unannotated.

In this study, the enzyme commission (EC) number and the taxonomy information will be provided in the small protein annotation. For simplicity of this research, protein copy numbers are used to quantify the abundance of every annotation so that different samples are compared. The protein copy numbers are added together from different clusters with the same annotation. The protein copy numbers can be normalized by the total number of small protein copies in the data set. The normalized protein copy number [or counts, denoted as  $c(.)$ ] for a certain annotation  $A$  is calculated with the following formula, where  $s(C)$  is the size of cluster  $C$ . Analogous to transcriptome quantification, the normalized value can be multiplied by  $10^6$  to represent the copy numbers per million proteins.

**TABLE 1** | Data samples and statistics in metaBP analyses.

Sample	Biosample	Reads structure	# of reads (m)	# of assembled total proteins	# of small protein clusters with 4 or more members	Time for protein assembly (HH:MM:SS)	Time for metaBP-ML annotation (DD-HH:MM:SS)
1.1	Mice gut	2 × 100 bp	32.7	1847781	16475	00:56:55	01:06:28:23
1.2	Mice gut	2 × 100 bp	29.8	1969398	12725	00:59:14	23:04:32
1.3	Mice gut	2 × 100 bp	25.9	1714448	12916	00:48:12	01:00:06:49
1.4	Mice gut	2 × 100 bp	32.0	1705080	11870	00:56:49	22:07:28
1.5	Mice gut	2 × 100 bp	30.1	1662298	15140	00:49:58	16:48:22
1.6	Mice gut	2 × 100 bp	32.3	1999701	14555	01:03:55	01:15:22:07
1.7	Mice gut	2 × 100 bp	28.3	1525231	12758	00:49:56	01:00:56:54
1.8	Mice gut	2 × 100 bp	37.4	2483994	18411	01:14:25	12:58:13
1.9	Mice gut	2 × 100 bp	33.0	1948886	14380	01:00:10	01:02:03:12
1.10	Mice gut	2 × 100 bp	33.4	2808475	18745	01:23:08	01:08:10:31
1.11	Mice gut	2 × 100 bp	33.7	2167085	16087	01:02:30	01:07:40:41
1.12	Mice gut	2 × 100 bp	35.6	2376492	15779	01:12:33	01:04:21:47
1.13	Mice gut	2 × 100 bp	37.0	2974313	17935	01:39:31	01:07:51:20
1.14	Mice gut	2 × 100 bp	44.5	3626380	19818	01:53:51	12:02:22
1.15	Mice gut	2 × 100 bp	32.5	1965315	16659	01:02:31	01:05:44:15
1.16	Mice gut	2 × 100 bp	31.1	2322776	12481	01:12:24	22:28:17
2	Human gut	2 × 151 bp	37.5	29194727	171238	01:26:04.66	20:15:30:00
3	Human skin	2 × 150 bp	5.1	679115	2874	00:30:24.91	04:54:01
4	Meadowsoil	2 × 200 bp	24.7	14947269	23640	14:41:28.71	01:14:07:31
5	Marine	2 × 150 bp	13.1	4595883	23410	02:12:55.85	01:14:22:30
6	Human saliva	2 × 126 bp	26.8	133743	583	00:36:04.86	01:06:08

$$c(A) = \sum_{C \in A} s(C) \times 10^6 / \sum_C s(C).$$

## 3 RESULTS

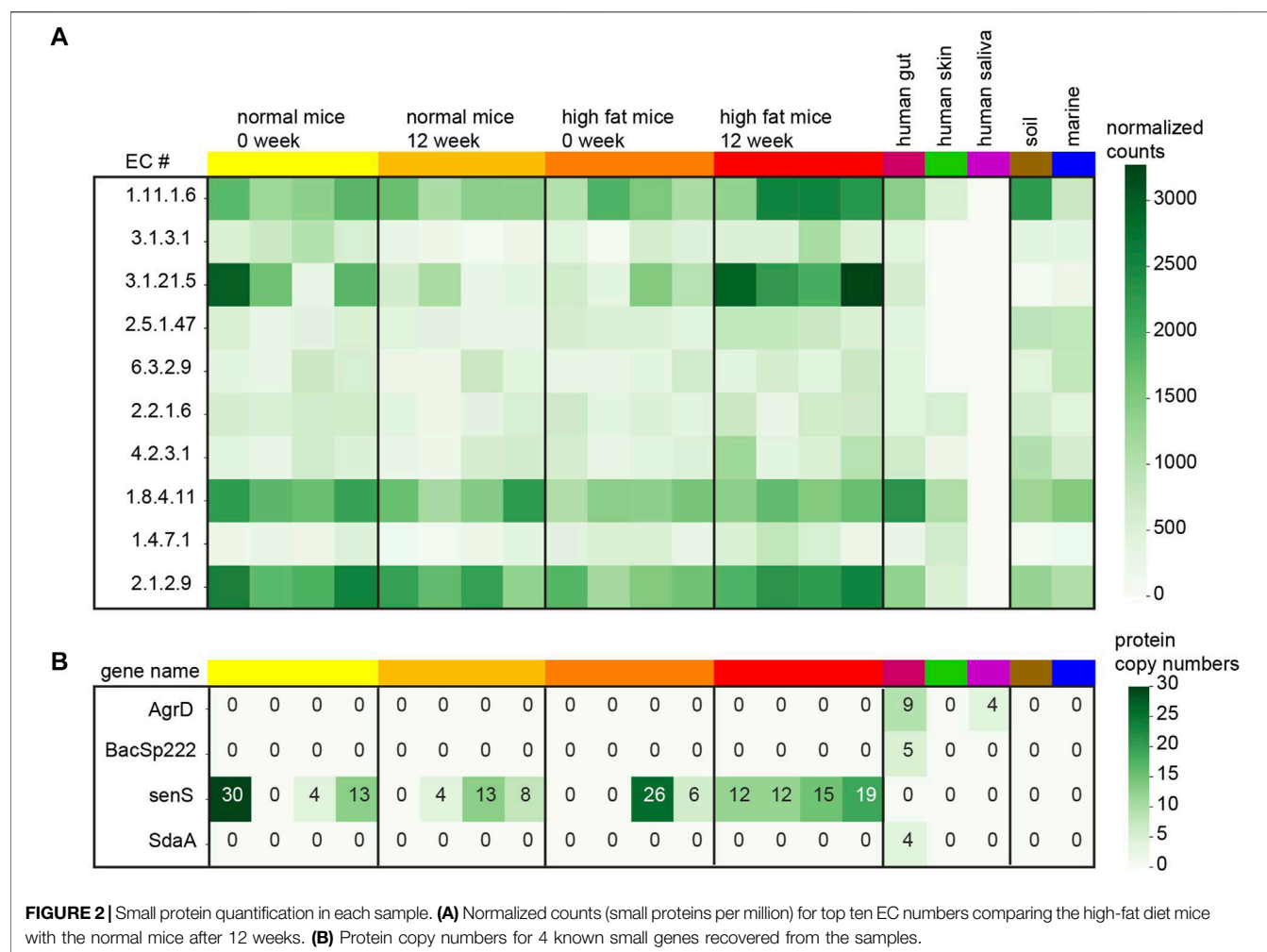
### 3.1 Small Protein Identification by metaBP in a Wide Range of Samples

The metaBP is applied on various metagenomic data sets, including sixteen mice gut samples (Morissette et al., 2020), one human gut sample (Lee et al., 2017), one human skin sample, one saliva sample, and environmental microbe samples in soil and marine (see **Supplementary Table S1** for NCBI Bioproject IDs). The data size varies from 5 million to 45 million of sequencing reads (**Table 1**; **Supplementary Table S1**). The sequencing read length varies between 100 and 200 base pairs from each end. On average, about one third of the resulted sequences are short sequences from the protein assembly results. However, only less than 5% of the sequences are in a cluster with at least four sequences, which is consistent with previously reported percentage of small open reading frames in metagenomic samples. Thus, clusters with at least four sequences are treated as reliable small protein families in the sample. Overall, about 5,000–8,000 clusters with their small representative proteins are generated within every million of the total assembled proteins. These clusters and representative sequences are sent to metaBP-ML (and/or eggNOG) for annotation, so that the taxonomy and enzyme commission (EC) information can be obtained and quantified for each sample.

The analysis from mice samples shows interesting enzyme activities. In order to compare EC numbers across samples, only those ECs existing in all the samples are used in this

analysis. First, the normalized counts of every EC number from mice samples are tested by ANOVA and the top ten important EC functions enriched in the high-fat diet of 12-week-old mice are presented in the heatmap (**Figure 2A**). The complete EC quantification table is available in the **Supplementary Table S6**. These top ten ECs corresponding to the high-fat diet mice show potential enzyme activities from small protein families. For example, the proteins marked with EC 1.11.1.6 belong to the catalase which is important for radical degradation. Catalases and antioxidant enzymes are known to increase in order to benefit the mice with a high-fat diet (Liang et al., 2015; Piao et al., 2017). It is necessary to mention that after the Benjamini–Hochberg *p*-value correction, none of the EC numbers are significant in the high-fat diet mice anymore. So, the EC numbers displayed in the heatmap are simply ranked by its original *p*-value (less than 5%). It is noticeable that the quantification pattern of these EC numbers from the human gut sample is more like the mice gut samples compared with the other samples. Human saliva samples do not have good yield of small proteins compared to the other samples.

In this study, 29 of the known short proteins derived from bacteria are searched from the metaBP output and only four of those are discovered in our samples (**Figure 2B**). The Uniprot IDs of these small genes are listed in the **Supplementary Table S7**. The most abundant small genes, *senS* are discovered in 12 of the 16 mice gut samples, but not in the human gut. The other three genes, *AgrD*, *BacSp222*, and *SdaA* are only recovered from the human gut sample. Indeed these 29 small genes are all from human associated microbes (Sberro et al., 2019) so that they may not be easily observed in the soil and marine samples. While metaBP-ML has discovered four of these 29 genes in our samples,



the annotation from eggNOG does not show any of these 29 genes.

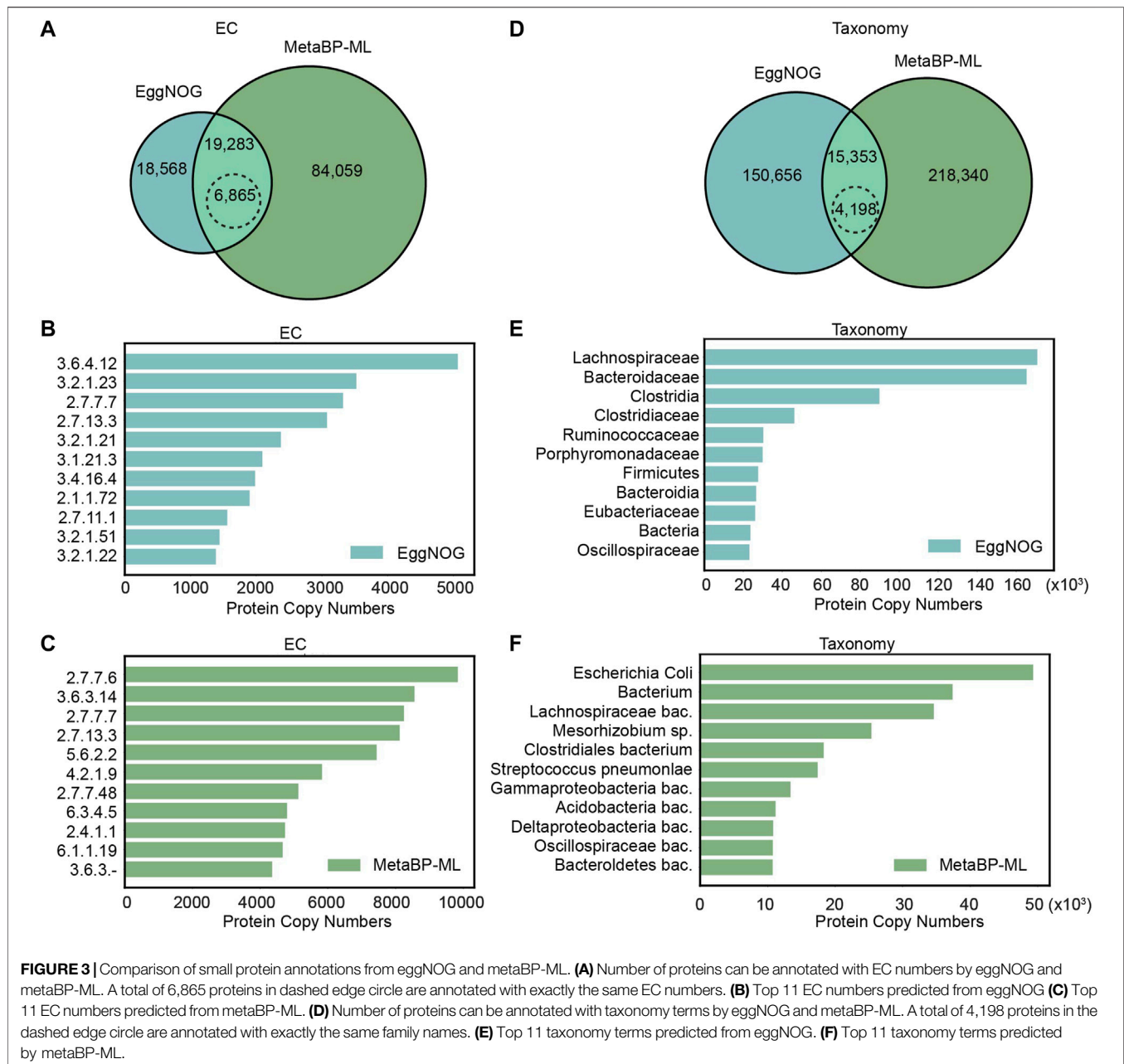
### 3.2 Small Protein Annotation by eggNOG and metaBP-ML

Besides the search for the known 29 small genes, sixteen mice gut samples are used to systematically compare the annotation outcomes from eggNOG and metaBP-ML for small protein families. As known that not many small proteins have clear enzyme activities, EC number annotation overall has lower yields compared with the taxonomy (organism group) annotation, no matter by eggNOG or metaBP-ML.

For the EC number annotation, metaBP-ML can annotate almost five times more proteins than eggNOG (**Figure 3A**). Both methods can annotate the same set of 19,283 proteins, but 6,865 proteins have the consensus EC annotation. Among the top 11, the most abundant EC numbers in eggNOG and metaBP-ML (**Figures 3B,C**), EC2.7.7.7 (DNA-directed DNA polymerase) and EC2.7.13.3

(histidine kinase), occur in both methods. However, it is hard to confirm if the small proteins can have these enzyme activities or not, since the functions are assigned only by the similarity computation.

For taxonomy annotation, metaBP-ML can annotate almost twice of the proteins than eggNOG (**Figure 3D**). In order to compare the predicted taxonomy labels directly, taxonomy IDs from both the methods are normalized to family IDs. This means among the same set of 15,353 proteins that gain the taxonomy annotation from both the methods, only 4,198 proteins have exactly the same family name from both the methods. The consensus rate is between 1/3 to 1/4 between two approaches. Top 11 abundant taxa from eggNOG are family names, order names or phylum names (extracted from the narrowest annotation from eggNOG results), while top 11 taxa from metaBP-ML, which can be as detailed as species level annotation (**Figures 3E,F**). From the top taxa lists obtained in both the methods, Lachnospiraceae, Oscillospiraceae, and Clostridia are the consensus. Overall, our metaBP-ML can provide more annotations with more



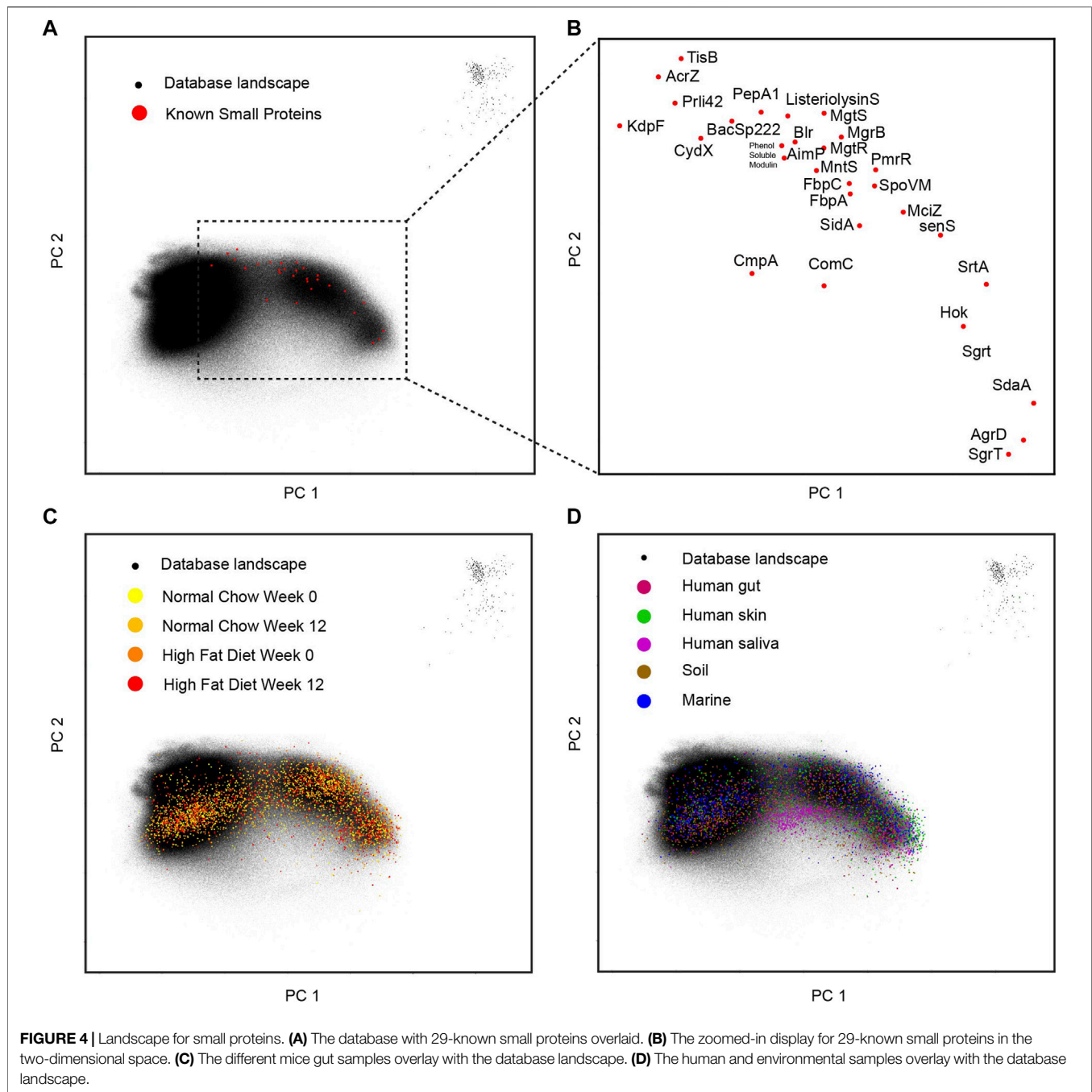
details mainly because of a very specific small protein database constructed.

### 3.3 Small Protein Landscape by metaBP-ML

As mentioned above, the entire small protein database composed of 3 million of short sequences are transformed into a 1,280-dimension vector space. In order to visualize the landscape within two dimensions, principal components analysis is performed, and the first two principal dimensions are shown in a dot plot (**Figure 4A**). The collected 29-known small genes are overlaid on this landscape and their relative locations and gene names are in a zoomed-in plot (**Figure 4B**). Surprisingly, within the first two principal components, the small protein landscape clearly shows

three clusters: left, right, and some outliers on the top right corner. It is hard to tell if this pattern of distribution reflects the true biology or some artifacts in the data collection, which requires future investigation. The known 29 small genes are mainly located on the right side of the landscape. When the mice samples are overlaid to this landscape (**Figure 4C**), there is no observable sample effects. When more samples are overlaid onto this landscape (**Figure 4D**), we can observe that the soil sample and skin sample are more on the right side while the human saliva sample is more located under the conjunction of the two parts. This entire landscape built from small proteins makes it possible to visualize the sample specific patterns from a natural microbial community.





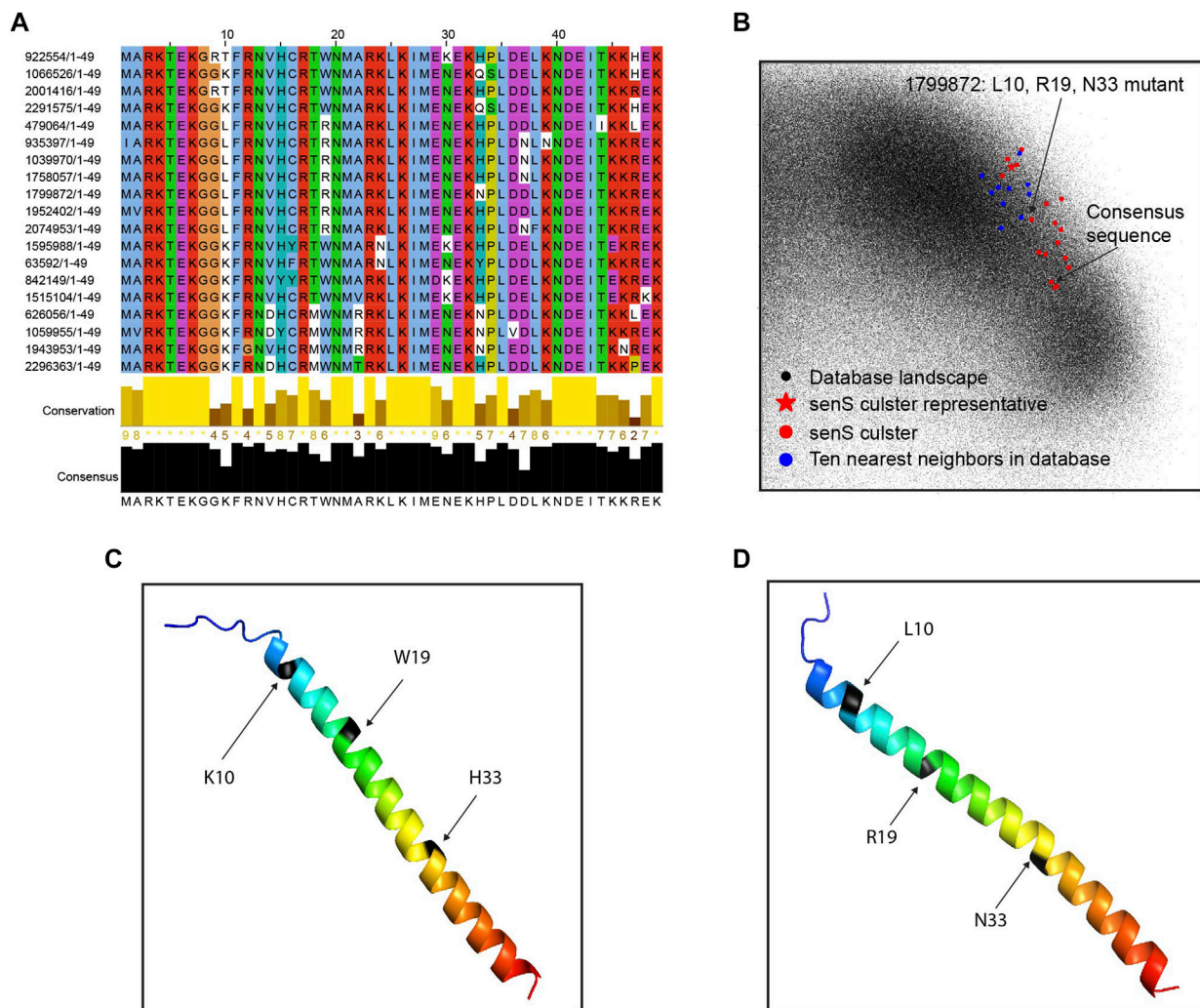
### 3.4 Sequence Diversity in Small Protein Clusters

To explore several interesting clusters identified in the mice gut samples, we pull out the protein cluster sequences from metaBP results and conduct further analyses. The clusters shown in this section are from the 12-week-old mice with a high-fat diet. One of the known small genes, *senS*, is widely discovered in the mice gut samples, and its sequence diversity is shown after the sequence alignment (**Figure 5A**). The *senS* protein sequences, including the consensus sequence and one of the mutants, are overlaid with all

small proteins (**Figure 5B**). This cluster is located on the right side of the landscape (**Supplementary Figure S1**). By using AlphaFold2 (Jumper et al., 2021), the consensus sequence of *senS* is predicted as an alpha helix structure (**Figure 5C**). Having three amino acids mutations, the structure for the mutant protein still shows a clear helix, but with a slightly bending effect (**Figure 5D**).

Another interesting cluster is from catalase EC1.11.1.6. The alignment of the sequences shows very few possible mutations are detected in the high-fat diet mice (**Supplementary Figure S2**). The structures predicted by alphaFold2, as well as the display of





**FIGURE 5 |** Sequence diversity of *senS* gene. **(A)** Sequence alignment and conservation of the *senS* proteins. **(B)** The *senS* cluster and ten neighbors overlay onto the database landscape. **(C)** The predicted structure for the consensus sequence of *senS*. **(D)** The predicted structure for a mutant of the consensus.

the protein landscape, show that the two amino acids substitution with longer side chains (R vs. G, N vs. H) help to make the loop region a little bit more structured, but not too much overall change. The structures show an alpha helix and beta sheet motif for this protein cluster.

## 4 DISCUSSION

The metaBP adopts protein level assembly by PLASS, and therefore it is not constraint by the requirement of long contigs or high-quality MAGs from the nucleotide level assembly. As we know, low-abundant rare species may overall constitute a large amount of the sequencing reads in the complex metagenomic samples but may not yield long contigs. When the sequencing depth is low, more than half of the data could be wasted as unassembled sequencing reads. But for small proteins this fragmented sequencing data should already provide sufficient

information for both the sequence and function. The metaBP together with metaBP-ML provide users with a complete toolkit to explore small proteins in natural metagenomic samples. For potential extension, the metaBP-ML does allow users to build their application specific models for protein annotation. In addition to metaBP-ML, we still provide eggNOG in the package to annotate proteins alternatively. In terms of the running time, eggNOG is more efficient with their pre-built reference database. The metaBP-ML is relatively taking more time when annotating proteins through vectorization and nearest neighbors. But due to our constructed small protein database, metaBP-ML can be very specific to identify and annotate small proteins. With the integration of both the tools, the metaBP can be used in various kinds of metagenomic data and annotate arbitrary protein classes.

However, there are still concerns and limitations from the current version of metaBP. Clusters with singletons at this moment are not used for the downstream analysis in the current metaBP. We assume

that only re-occurred sequences within the same cluster can indicate the reliability of small proteins and their mutations. Generally, high quality metagenomic data should be sufficient in the sequence depth. However, in many unexpected cases, metagenomic dataset can be sparse, and the clusters with lower number of protein members can also be informative for small proteins. Computationally, there has not been a perfect strategy to balance the false positives and false negatives without knowing the ground truth in the real data sets. But with the metaBP, we can at least provide a short list for the experimental detection through mass spectrometry and biochemical analysis.

The metaBP quantify the annotated features using the normalized protein copy numbers. Due to the protein level assembly, the protein copy numbers are the most straightforward quantification obtained from the data set. Although metaBP can recover more annotations than eggNOG, the quantification may not be sufficient to statistically recover significant features when comparing the samples. One future direction is to improve the resolution of the quantification using the original sequencing reads. The metaBP also displays the protein diversity by homologous protein clustering, but the current metaBP cannot quantify the confidence level of each amino acid mutation. So, the current metaBP is only for the discovery of the potential sequence diversity in a protein family, not for the strict quantification of mutation occurrence.

## 5 CONCLUSION

This study proposes a new and comprehensive toolkit, metaBP (and metaBP-ML), to discover and annotate the community specific bacterial (microbe derived) peptides from the metagenomic samples. It is built upon a new idea of direct protein level assembly and one of the current protein clustering tools, as well as machine learning based approaches. The exploration of the small protein landscape and the analyses of peptides annotation demonstrate the efficacy of this work and the value of machine learning.

## DATA AVAILABILITY STATEMENT

Publicly available metagenomic datasets were analyzed in this study. These data can be downloaded from NCBI SRA repository by the information provided in **Supplementary Table S1**. The small example data set for testing parameters are from PLASS

## REFERENCES

- Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., et al. (2021). UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* 49, D480. doi:10.1093/nar/gkaa1100
- Bushnell, B., Rood, J., and Singer, E. (2017). BBMerge - Accurate Paired Shotgun Read Merging via Overlap. *PLoS One* 12, e0185056. doi:10.1371/journal.pone.0185056
- Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-Mapper V2: Functional Annotation, Orthology
- GitHub (https://github.com/soedinglab/plass/tree/master/examples). The pipeline and tools are available through github for metaBP (https://github.com/yao-laboratory/metaBP), metaBP-ML (https://github.com/yao-laboratory/metaBP-ML) together with an integrated version of RBiTools. In metaBP-ML, ESM and its model are used. The source codes and models can be found from ESM GitHub: https://github.com/facebookresearch/esm. The pre-trained model for general purpose “esm1b\_t33\_650M\_UR50S” is used for this proposed embedding work.
- Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* 38, 5825–5829. doi:10.1093/molbev/msab293
- Chen, J., Brunner, A. D., Cogan, J. Z., Nuñez, J. K., Fields, A. P., Adamson, B., et al. (2020). Pervasive Functional Translation of Noncanonical Human Open Reading Frames. *Science* 367, 1140–1146. doi:10.1126/science.aay0262
- Duval, M., and Cossart, P. (2017). Small Bacterial and Phagic Proteins: An Updated View on a Rapidly Moving Field. *Curr. Opin. Microbiol.* 39, 81–88. doi:10.1016/j.mib.2017.09.010
- Garai, P., and Blanc-Potard, A. (2020). Uncovering Small Membrane Proteins in Pathogenic Bacteria: Regulatory Functions and Therapeutic Potential. *Mol. Microbiol.* 114, 710–720. doi:10.1111/mmi.14564

## AUTHOR CONTRIBUTIONS

The idea and framework of this toolkit were conceived and designed by QY. The implementation of the metaBP and the coordination of this project were carried out by MV. The machine learning module, metaBP-ML, was conducted by BJ. The RBiTools was developed independently by ML and integrated to metaBP by MV. The testing of the toolkit and the data generation of all samples were carried out by MV, BJ, and LK. The data analysis and the figures were provided by QY, MV, and BJ. The manuscript was initially drafted by QY and revised by the co-authors.

## FUNDING

Financial support was provided by the National Institutes of Health (NIH), grant no. P20GM104320. This research was also supported by UCARE program in University of Nebraska-Lincoln.

## ACKNOWLEDGMENTS

The authors acknowledge the Holland Computing Center (HCC) in the University of Nebraska-Lincoln providing computational resources and support.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2022.935351/full#supplementary-material>

- Hemm, M. R., Weaver, J., and Storz, G. (2020). *Escherichia coli* Small Proteome. *EcoSal Plus* 9. doi:10.1128/ecosalplus.esp-0031-2019
- Huan, Y., Kong, Q., Mou, H., and Yi, H. (2020). Antimicrobial Peptides: Classification, Design, Application and Research Progress in Multiple Fields. *Front. Microbiol.* 11, 582779. doi:10.3389/fmicb.2020.582779
- Hyatt, D., Locascio, P. F., Hauser, L. J., and Uberbacher, E. C. (2012). Gene and Translation Initiation Site Prediction in Metagenomic Sequences. *Bioinformatics* 28, 2223–2230. doi:10.1093/bioinformatics/bts429
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* 596, 583–589. doi:10.1038/s41586-021-03819-2
- Lee, S. T. M., Kahn, S. A., Delmont, T. O., Shaiber, A., Esen, Ö. C., Hubert, N. A., et al. (2017). Tracking Microbial Colonization in Fecal Microbiota Transplantation Experiments via Genome-Resolved Metagenomics. *Microbiome* 5, 50. doi:10.1186/S40168-017-0270-X
- Liang, L., Shou, X.-L., Zhao, H.-K., Ren, G.-q., Wang, J.-B., Wang, X.-H., et al. (2015). Antioxidant Catalase Rescues against High Fat Diet-Induced Cardiac Dysfunction via an IKK $\beta$ -AMPK-dependent Regulation of Autophagy. *Biochim. Biophys. Acta Mol. Basis Dis.* 1852, 343–352. doi:10.1016/j.bbadis.2014.06.027
- Miravet-Verde, S., Ferrar, T., Espadas-García, G., Mazzolini, R., Gharrab, A., Sabido, E., et al. (2019). Unraveling the Hidden Universe of Small Proteins in Bacterial Genomes. *Mol. Syst. Biol.* 15, e8290. doi:10.15252/msb.20188290
- Morissette, A., Kropp, C., Songpadith, J.-P., Junges Moreira, R., Costa, J., Mariné-Casadó, R., et al. (2020). Blueberry Proanthocyanidins and Anthocyanins Improve Metabolic Health through a Gut Microbiota-dependent Mechanism in Diet-Induced Obese Mice. *Am. J. Physiol. Endocrinol. Metabolism* 318, E965–E980. doi:10.1152/AJPENDO.00560.2019
- Orr, M. W., Mao, Y., Storz, G., and Qian, S.-B. (2021). Alternative ORFs and Small ORFs: Shedding Light on the Dark Proteome. *Nucleic Acids Res.* 48, 1029–1042. doi:10.1093/NAR/GKZ734
- Piao, L., Choi, J., Kwon, G., and Ha, H. (2017). Endogenous Catalase Delays High-Fat Diet-Induced Liver Injury in Mice. *Korean J. Physiol. Pharmacol.* 21, 317. doi:10.4196/kjpp.2017.21.3.317
- Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., et al. (2021). Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *Proc. Natl. Acad. Sci. U.S.A.* 118, e2016239118. doi:10.1073/pnas.2016239118
- Sberro, H., Fremin, B. J., Zlitni, S., Edfors, F., Greenfield, N., Snyder, M. P., et al. (2019). Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small, Novel Genes. *Cell* 178, 1245–1259. doi:10.1016/j.cell.2019.07.016
- Sievers, F., and Higgins, D. G. (2018). Clustal Omega for Making Accurate Alignments of Many Protein Sequences. *Protein Sci.* 27, 135–145. doi:10.1002/pro.3290
- Steinegger, M., and Söding, J. (2018). Clustering Huge Protein Sequence Sets in Linear Time. *Nat. Commun.* 9, 2542. doi:10.1038/s41467-018-04964-5
- Steinegger, M., Mirdita, M., and Söding, J. (2019). Protein-level Assembly Increases Protein Sequence Recovery from Metagenomic Samples Manyfold. *Nat. Methods* 16, 603–606. doi:10.1038/s41592-019-0437-4
- Storz, G., Wolf, Y. I., and Ramamurthi, K. S. (2014). Small Proteins Can No Longer Be Ignored. *Annu. Rev. Biochem.* 83, 753–777. doi:10.1146/annurev-biochem-070611-102400

**Conflict of Interest:** ML is affiliated with Nashville Biosciences. His development of the RBiotools package was done independently of any funding from Nashville Biosciences or from any other commercial funding source.

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

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

Copyright © 2022 Vajjala, Johnson, Kasperek, Leuze and Yao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



## OPEN ACCESS

EDITED BY  
Ruquan Ge,  
Hangzhou Dianzi University, China

REVIEWED BY  
Michael Birnbaum,  
Massachusetts Institute of Technology,  
United States  
Yushan Qiu,  
Shenzhen University, China

\*CORRESPONDENCE  
Jiayin Wang,  
wangjiayin@mail.xjtu.edu.cn

This study was submitted to  
Computational Genomics,  
a section of the journal Frontiers in  
Genetics.

RECEIVED 12 May 2022  
ACCEPTED 28 June 2022  
PUBLISHED 22 August 2022

CITATION  
Xu Y, Qian X, Tong Y, Li F, Wang K,  
Zhang X, Liu T and Wang J (2022),  
AttnTAP: A Dual-input Framework  
Incorporating the Attention Mechanism  
for Accurately Predicting TCR-  
peptide Binding.  
*Front. Genet.* 13:942491.  
doi: 10.3389/fgene.2022.942491

COPYRIGHT  
© 2022 Xu, Qian, Tong, Li, Wang, Zhang,  
Liu and Wang. This is an open-access  
article distributed under the terms of the  
[Creative Commons Attribution License](#)  
(CC BY). The use, distribution or  
reproduction in other forums is  
permitted, provided the original  
author(s) and the copyright owner(s) are  
credited and that the original  
publication in this journal is cited, in  
accordance with accepted academic  
practice. No use, distribution or  
reproduction is permitted which does  
not comply with these terms.

# AttnTAP: A Dual-input Framework Incorporating the Attention Mechanism for Accurately Predicting TCR-peptide Binding

Ying Xu<sup>1</sup>, Xinyang Qian<sup>1</sup>, Yao Tong<sup>1</sup>, Fan Li<sup>1</sup>, Ke Wang<sup>1,2</sup>,  
Xuanping Zhang<sup>1</sup>, Tao Liu<sup>1,2</sup> and Jiayin Wang<sup>1\*</sup>

<sup>1</sup>Department of Computer Science and Technology, School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China, <sup>2</sup>Geneplus Beijing Institute, Beijing, China

T-cell receptors (TCRs) are formed by random recombination of genomic precursor elements, some of which mediate the recognition of cancer-associated antigens. Due to the complicated process of T-cell immune response and limited biological empirical evidence, the practical strategy for identifying TCRs and their recognized peptides is the computational prediction from population and/or individual TCR repertoires. In recent years, several machine/deep learning-based approaches have been proposed for TCR-peptide binding prediction. However, the predictive performances of these methods can be further improved by overcoming several significant flaws in neural network design. The interrelationship between amino acids in TCRs is critical for TCR antigen recognition, which was not properly considered by the existing methods. They also did not pay more attention to the amino acids that play a significant role in antigen-binding specificity. Moreover, complex networks tended to increase the risk of overfitting and computational costs. In this study, we developed a dual-input deep learning framework, named AttnTAP, to improve the TCR-peptide binding prediction. It used the bi-directional long short-term memory model for robust feature extraction of TCR sequences, which considered the interrelationships between amino acids and their precursors and postcursors. We also introduced the attention mechanism to give amino acids different weights and pay more attention to the contributing ones. In addition, we used the multilayer perceptron model instead of complex networks to extract peptide features to reduce overfitting and computational costs. AttnTAP achieved high areas under the curves (AUCs) in TCR-peptide binding prediction on both balanced and unbalanced datasets (higher than 0.838 on McPAS-TCR and 0.908 on VDJDdb). Furthermore, it had the highest average AUCs in TPP-I and TPP-II tasks compared with the other five popular models (TPP-I: 0.84 on McPAS-TCR and 0.894 on VDJDdb; TPP-II: 0.837 on McPAS-TCR and 0.893 on VDJDdb). In conclusion, AttnTAP is a reasonable and practical framework for predicting TCR-peptide binding, which can accelerate identifying neoantigens and activated T cells for immunotherapy to meet urgent clinical needs.



## KEYWORDS

T-cell receptor, TCR-peptide binding prediction, deep learning framework, BiLSTM model, attention mechanism

## 1 Introduction

T-cell receptor (TCR) hypervariable regions are formed by complex recombination of genomic precursor elements that mediate recognition of antigens presented by peptide-major histocompatibility complex (pMHC) molecules (La Gruta et al., 2018; Joglekar and Li, 2021). Complementary determining region 3 (CDR3) is the key structural feature located within the TCR variable regions, and specific CDR3-pMHC complexes enable T cells to recognize and eliminate evolving pathogens or malignant cells (La Gruta et al., 2018; Joglekar and Li, 2021). Thus, the CDR3 region, derived from quasi-random mutations of V(D)J recombination, is considered to have a primary function in recognizing the endogenous and exogenous antigens in the immune-dominant T-cell process and resulting “TCR repertoire” in an individual, which defines a unique footprint of cellular immune protection (Chiffelle et al., 2020).

The high-throughput immune repertoire sequencing (IR-seq) can capture millions of sequencing reads derived from the hypervariable regions and produce detailed T-cell repertoires for individual or population analysis, such as epitope prediction (Warren et al., 2011; Woodsworth et al., 2013; Glanville et al., 2017). However, identifying epitopes from TCR repertoires by biomechanical experiments is a time-consuming and labor-intensive task. An epitope that is expanded in multiple T-cell clones is more likely to be exposed to the pMHC complex and can generally serve as a surface biomarker for immunotherapy or vaccine targets. Fortunately, the availability of immune-related TCR/BCR sequence databases, such as IEDB (Mahajan et al., 2018), VDJdb (Bagaev et al., 2020), and/or McPAS-TCR (Tickotsky et al., 2017), will serve as motivation to accelerate the development of well-integrated epitope prediction pipelines. As a result, it will be an ideal method that predicts an epitope from billions of TCR sequences and validates it with a biological experiment, greatly reducing time and cost consumption.

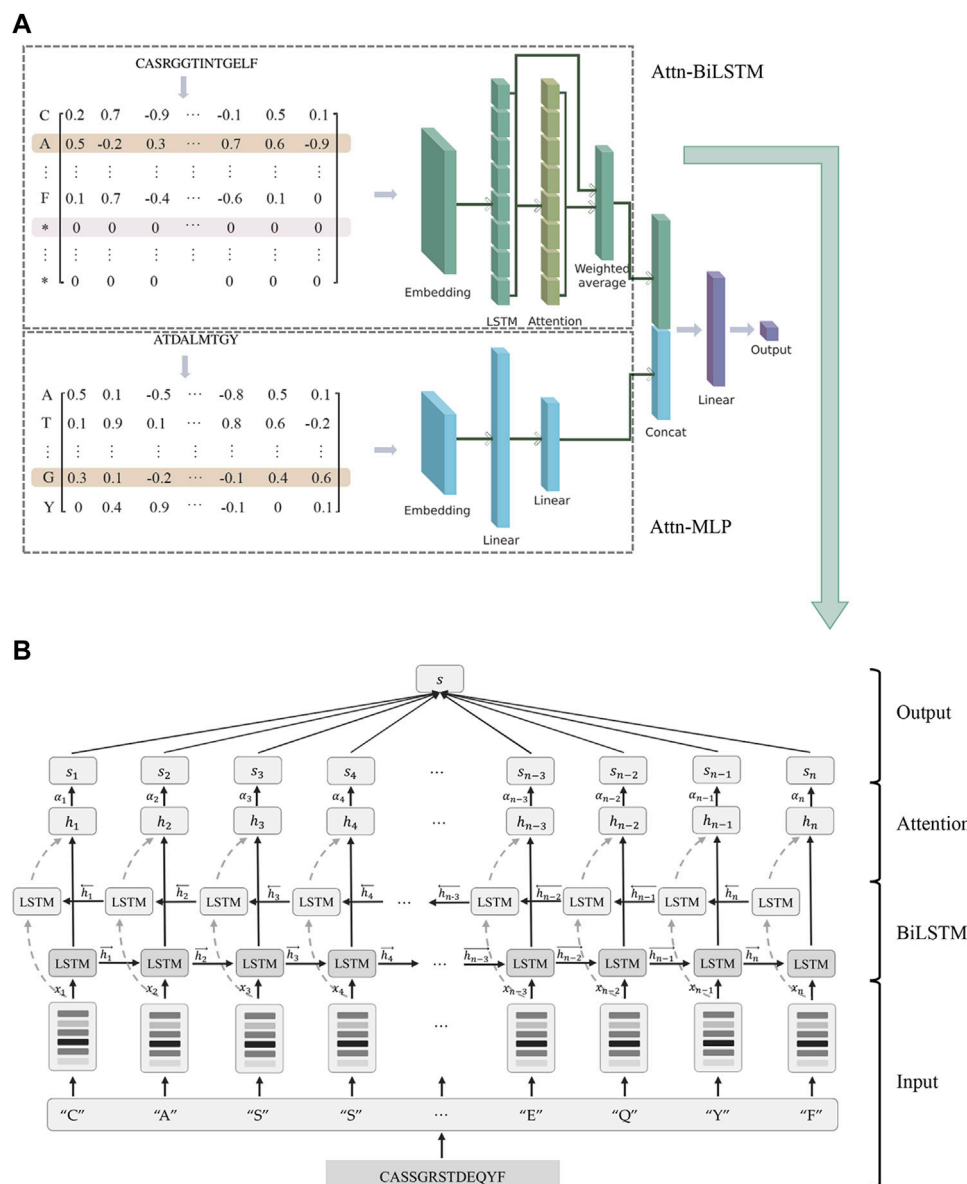
It is critical to introduce an appropriate prediction model to predict an epitope, as extracting fitness features from a highly variable and shortened amino acid chain is difficult (Bolotin et al., 2012). The length and positional characteristics of the subsequences are unknown, and the amino acids in the subsequences contribute to varying degrees. Unfortunately, the aforementioned public databases have an imbalanced epitope distribution (a high number of unseen epitopes) as well as a lack of high-quality labeled seen-epitope data (Moris et al., 2021). Deep machine learning (DL) models have significantly accelerated the epitope prediction task by automatically learning engineering features based on domain knowledge and

extracting unknown and implicit features from unprecedented amounts of TCR repertoire data using unprecedented scale models (LeCun et al., 2015; Zemouri et al., 2019; Tran et al., 2022).

Several cutting-edge TCR-peptide binding prediction approaches based on DL frameworks have been proposed in the last 2 years, and they were applicable to both seen and unseen-TCR epitopes. DLpTCR used a multi-model ensemble strategy comprised of three base classifiers in predicting the likelihood of interaction between TCR  $\alpha\beta$  chains and peptides (Xu et al., 2021). NetTCR-2.0 provided a 1-dimensional (1D) convolution neural network (CNN) architecture combining max-pooling for dealing with sequence length variations (Montemurro et al., 2021). The input TCR  $\alpha\beta$  chains and peptide sequences were encoded by the BLOSUM50 (Henikoff and Henikoff, 1992) matrix before being fed into a dense layer for prediction. ImRex used a four-layer convolution and two-layer max-pooling CNN architecture to predict the combined representation of CDR3 and peptide sequences, by extracting their physicochemical properties as features (Moris et al., 2021). ERGO employed a new multilayer perceptron (MLP) model to predict the likelihood of TCR-peptide binding. During the study, they provided two different encoding methods, a long short-term memory (LSTM) network, and an auto-encoder network to generate the corresponding models (ERGO-LSTM & ERGO-AE) (Springer et al., 2020).

The CNN architecture is widely used to extract the features of TCRs and make TCR-peptide prediction, such as DLpTCR, ImRex, NetTCR-2.0 and DeepLION (Xu et al., 2022), due to its superior capacity for image feature learning. However, the lack of CNN memory capability during the model process will reduce the feature extraction performance on short sequence data, especially TCRs. Due to the spatial folding of TCRs, amino acids in sequences may be related not only to their adjacent amino acids, but also to some more distant ones. When extracting sequence features, CNN only considered interrelationships between adjacent amino acids and ignored those between non-adjacent amino acids, which also play a significant role in TCR antigen-binding specificity. The LSTM architecture, used by the ERGO model, had memory capability and would reduce the information loss of non-adjacent amino acids. However, the ERGO model only used the last node output to represent the entire sequence, ignoring the contribution of previous node outputs to the final prediction. Furthermore, the existed start-of-art models could not pay more attention to the amino acids in sequences that contributed significantly to TCR antigen recognition. The complex framework would result in overfitting on TCR-peptide binding tasks, especially under unbalanced datasets with small labeled sample sizes. As a





**FIGURE 1**

AttnTAP improved the prediction accuracy of TCR-peptide binding. **(A)** AttnTAP was a dual-input deep learning framework, which included the feature extractors for TCR and peptide sequences, Attn-BiLSTM and Attn-MLP. The corresponding feature vectors extracted by the two models were then concatenated for predicting the likelihood of TCR and peptide binding using the multilayer perceptron network. **(B)** The feature extractor for TCR sequences, Attn-BiLSTM, was divided into four parts: the input layer, bi-directional long short term memory (BiLSTM) layer, attention layer, and output layer. Sequences were preprocessed and encoded into embeddings in the input layer. The embeddings (BiLSTM) layer, then fed into the BiLSTM and attention layers, respectively. The BiLSTM layer extracted the sequences' feature vectors, while the attention layer computed the weights of each position in the sequences. Finally, the output layer outputted the weighted feature vectors.

result, there were still some unresolved issues with existed models and their predictive performances can be further improved by overcoming several significant flaws in neural network designs.

Motivated by these, we proposed AttnTAP, a dual-input deep learning network that included the Attn-BiLSTM and Attn-MLP models, to improve the prediction of TCR-peptide binding (Figure 1). The bi-directional LSTM (BiLSTM) model with an

attention mechanism was used to extract the features of TCR sequences, as described in Section 2.2. The BiLSTM model considered the interrelationships between amino acids and their adjacent or non-adjacent precursors and postcursors. Moreover, due to the attention mechanism, all node outputs were used to represent the entire sequence after weighted calculation, with a focus on the key amino acids. Given that

TABLE 1 The datasets used for approach evaluation.

	Peptide type	TCR-peptide pair number	Positive sample size	Negative sample size
McPAS-TCR	25	9,597	9,597	9,597–143,955
VDJdb	56	38,134	38,134	38,134–572,010

very few known peptides in the public databases compared to the TCR sequences, a simple network, MLP, was used to extract peptide features to reduce the complexity of the network structure. A dual-input framework of CDR3 sequences and peptides was used to combine embedding matrices, and then the two output feature vectors were concatenated by the MLP network to predict the likelihood of a TCR recognizing a peptide. Finally, we evaluated the performance of AttnTAP and other start-of-art TCR-peptide binding prediction models, in terms of the prediction accuracy, computational cost, and space complexity.

## 2 Materials and methods

AttnTAP was a dual-input deep learning framework developed for predicting the TCR-peptide binding (Figure 1A). TCR CDR3 $\beta$  sequences, as one of the inputs, were extracted features using the BiLSTM model with an attention mechanism, named Attn-BiLSTM. The peptide sequences were extracted features using the MLP model, named Attn-MLP. Then, the corresponding features from Attn-BiLSTM and Attn-MLP models were concatenated to form a final feature that was used to predict the likelihood of TCR-peptide binding using the MLP network.

### 2.1 Data processing

The public TCR-peptide datasets used in this study were downloaded from the VDJdb (<https://vdjdb.cdr3.net/>) (Bagaev et al., 2020), IEDB (<http://www.iedb.org/>) (Mahajan et al., 2018), and McPAS-TCR (<http://friedmanlab.weizmann.ac.il/McPAS-TCR/>) (Tickotsky et al., 2017), respectively. The three datasets were used to train the word vectors for AttnTAP, and the VDJdb and McPAS-TCR datasets were used to evaluate the performance of binding prediction approaches. In all of the three datasets, the standard screening sequences are as follows: 1) We removed the duplicated sequences, too short (<6bp) or too long (>30bp) CDR3 $\beta$  sequences, incomplete sequences, and tag-less sequences; 2) The peptide sequences corresponding to less than 50 TCR sequences were also removed; 3) We retained only the correct sequences of the human TCR $\beta$  CDR3 and peptide sequences. As result, we obtained amounts of 181,436 CDR3 $\beta$  sequences from the three public datasets

(“CA ... F” sequences) to train the word vectors for AttnTAP (dataset one in this study). The length of CDR3 $\beta$  sequences ranges from 6 to 27 amino acids, with the majority containing 11–18 amino acids (Supplementary Figure S1).

Furthermore, after the screening process, we obtained 9,597 TCR-peptide pairs with 25 different peptide sequences from the McPAS-TCR database and 38,134 TCR-peptide pairs with 56 different peptide sequences from the VDJdb database as positive samples (Table 1, dataset two in this study). We analyzed these peptides in the datasets and their species, TCR counts, and abundances are shown in Supplementary Table S1. Negative samples were generated by randomly replacing the corresponding peptide in positive samples with other peptides (Springer et al., 2020). The procedure for generating negative samples is shown in Supplementary Algorithm S1. The ratio of negative samples to positive samples used in this study ranged from 1:1 to 15:1.

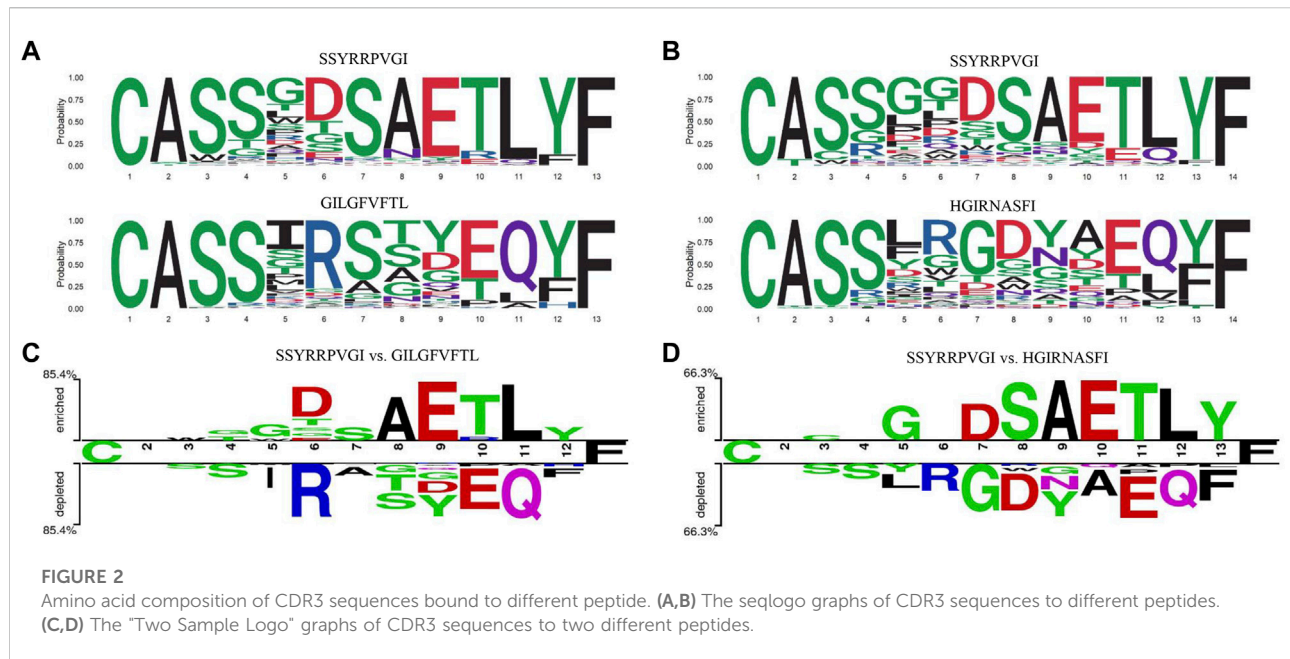
### 2.2 Attn-BiLSTM model

Attn-BiLSTM model was divided into four parts including the input layer, BiLSTM layer, attention layer, and output layer (Figure 1B). In the input layer, amino acid sequences were preprocessed and encoded into embeddings. Then, the embeddings were fed into both the BiLSTM and the attention layers. The feature vectors of sequences were extracted in the BiLSTM layer, while the weights of each position in the sequences were computed in the attention layer. Finally, the weighted feature vectors were output in the output layer.

#### 2.2.1 Input layer

According to the previous studies (Montemurro et al., 2021) and length-frequency statistics (Supplementary Figure S1), the maximum input length of CDR3 was 18 amino acids and the redundant part would be truncated to a longer sequence. For the shorter sequences, we completed them with a placeholder “X” to the maximum length.

Random initialization vectors and pre-training word vectors were available for Attn-BiLSTM to encode sequences. We used the character granularity vectors and word granularity vectors as pre-training word vectors, respectively. Each amino acid was viewed as a basic character, resulting in a total of 20 characters. Moreover, three consecutive amino acid residues in a sequence were considered as one word in word granularity vectors, also named triplet word vectors (Asgari and Mofrad, 2015). We used Word2vec (Mikolov et al., 2013) to train these word vectors.



### 2.2.2 BiLSTM layer

The LSTM model specializes in sequential data, reduces information loss and long-term dependency problems in the recurrent neural network, and performs well in TCR-peptide binding prediction (Springer et al., 2020). Compared to the LSTM, BiLSTM allows for more comprehensive and robust feature extraction because it takes into account both precursor and successor positions (Zhou et al., 2016). As a result, the BiLSTM model was used to extract the features of CDR3 sequences in this experiment. The encoded vector in the  $i$ th position  $x_i$  was fed into the forward LSTM (from left to right) and backward LSTM (from right to left) network, and the feature vectors  $\vec{h}_i$  and  $\overleftarrow{h}_i$  were output, respectively.

### 2.2.3 Attention mechanism

As an example, we plotted the seqlogo graphs of CDR3 sequences corresponding to the peptide sequences (Figures 2A,B) (Wagih, 2017), which indicated that the CDR3 sequences corresponding to different peptide sequences had similar patterns in upstream and downstream targets, but extremely distinct in the middle region. The difference between CDR3 sequences, corresponding to two different peptide sequences at various positions using "Two Sample Logo" (Figures 2C,D) (Schneider and Stephens, 2002; Crooks et al., 2004), also indicated that the amino acid composition of CDR3 sequences binding to different peptide sequences varies widely.

As shown in the aforementioned example, due to the significant differences in amino acid composition in the middle region of the CDR3 sequence, the attention mechanism could be used to focus on the amino acids that contributed to the antigen-binding specificity

and improve the feature extraction (Vaswani et al., 2017; Bahdanau et al., 2014). The weight of the feature vector in the  $i$ th position was calculated as

$$u_i = \text{Tanh}(W_A h_i + b_A), \quad (1)$$

$$a_i = \frac{e^{u_i^T u}}{\sum_t e^{u_t^T u}}, \quad (2)$$

where  $W_A$  and  $b_A$  were, respectively, the weight matrix and bias,  $\text{Tanh}(x)$  was the activation function, and  $a_i$  was the regularization of  $u_i$  using the Softmax function.

### 2.3 Attn-MLP model

Attn-MLP for peptide sequences consisted of the input layer and MLP layer. The input layer was the same as that in Attn-BiLSTM, and we set the maximum length of peptide sequences to nine in our study. We used a two-layer MLP model, a simple neural network model used in the majority of TCR-peptide binding prediction approaches (Springer et al., 2020; Montemurro et al., 2021; Moris et al., 2021; Xu et al., 2021), to extract the features of peptides. The operation process in each layer of the MLP model was given by

$$x' = \text{ReLU}(W_M \cdot x + b_M), \quad (3)$$

where  $W_M$  and  $b_M$  were, respectively, the weight matrix and bias, and  $\text{ReLU}(x)$  was the activation function to avoid gradient explosion or disappearance. To avoid overfitting, we used dropout (Srivastava et al., 2014) with a rate of 0.1.

TABLE 2 The selected representative TCR-peptide binding prediction approaches.

	Predictable TCR chain(s)	Model complexity	Input length constraint	Proposed date	Availability
ERGO-LSTM	TCR $\beta$	Medium	None	August 2020	<a href="https://github.com/louzounlab/ERGO/">https://github.com/louzounlab/ERGO/</a>
ERGO-AE	TCR $\beta$	Low	None	August 2020	<a href="https://github.com/louzounlab/ERGO/">https://github.com/louzounlab/ERGO/</a>
ImRex	TCR $\beta$	High	TCR: 10–20 & Epitope: 8–11	December 2020	<a href="https://github.com/pmoris/ImRex/">https://github.com/pmoris/ImRex/</a>
DLpTCR	TCR $\alpha$ & $\beta$	High	None	July 2021	<a href="https://github.com/jiangBiolab/DLpTCR/">https://github.com/jiangBiolab/DLpTCR/</a>
NetTCR-2.0	TCR $\alpha$ & $\beta$	Low	TCR: 8–18 & Epitope: 9	September 2021	<a href="https://github.com/mnielLab/NetTCR-2.0/">https://github.com/mnielLab/NetTCR-2.0/</a>

## 2.4 Multilayer perceptron network

The feature vectors of TCR and peptide sequences were concatenated into a final feature vector, which was used as the input of the latter MLP network for classification. The operation process of the MLP network was similar to Eq. 3, and the final prediction output was shown as

$$\tilde{Y} = P(Y = 1 | \{TCR_i, Peptide_j\}) = \text{ReLU}(W'_M \cdot x' + b'_M), \quad (4)$$

where  $\tilde{Y}$  denoted the probability that the  $i$ th TCR sequence binds to the  $j$ th peptide sequence. When  $\tilde{Y} > 0.5$ , we considered the TCR recognized the peptide and vice versa. The dropout with a rate of 0.1 was used to avoid overfitting. AttnTAP was end-to-end trainable, and the loss function was the log-likelihood function defined as

$$\mathcal{L} = -[\tilde{Y} \ln \tilde{Y} + (1 - \tilde{Y}) \ln (1 - \tilde{Y})]. \quad (5)$$

## 2.5 Performance evaluation approaches

We selected several state-of-the-art TCR-peptide combination prediction methods proposed in the last 2 years, which employed deep learning frameworks, to compare their performance with AttnTAP. As a result, ERGO (Springer et al., 2020), ImRex (Moris et al., 2021), DLpTCR (Xu et al., 2021), and NetTCR-2.0 (Montemurro et al., 2021) were selected for the comparison experiments (Table 2).

### 2.5.1 Two prediction tasks used for approach validation

Two different tasks, TCR-Peptide Pairing I (TPP-I) and TCR-Peptide Pairing II (TPP-II) as described in the previous study (Springer et al., 2020), were selected to estimate the performance of the binding prediction. In the TPP-I task, all of the TCRs and peptides both belong to the training and test sets, and TCR-peptide pairs were divided into disjoint training and test sets (dataset 2). We performed five-fold cross-validation (CV) for the TPP-I task. First, we sampled the

original dataset randomly and generated a new dataset (~10,000 TCR-peptide pairs). Then, the generated dataset was randomly divided into five equal parts, four of which were used as the training set and the rest as the test set. Three-quarters of the training data were used to train the model five times independently, and the rest were used as the validation data to select the final model.

The TPP-II was similar to TPP-I, except the TCRs contained in the pairs belonging to the training set could not belong to the test set. Considering that it was difficult to divide the dataset into five equal parts as required, we conducted independent replicate experiments 30 times to perform an unbiased estimation. The generated dataset was divided into a fixed ratio, the same as the five-fold CV in TPP-I, with a 4:1 ratio of training data to test data.

### 2.5.2 Metrics used for performance evaluation

In this study, we used the accuracy (ACC), recall (REC), precision (PRE), F1 score (F1), and area under the receiver operating characteristic curve (AUC), as the criteria for the performance evaluation of these six approaches. There were six values in these equations, including true (T), false (F), true positive (TP), true negative (TN), false positive (FP), and false-negative (FN), were used. The formulas were presented as follows:

$$ACC = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + FN + TN + FP}, \quad (6)$$

$$REC = \frac{TP}{P} = \frac{TP}{TP + FN}, \quad (7)$$

$$PRE = \frac{TP}{TP + FP}, \text{ and} \quad (8)$$

$$F1 = 2 \times \frac{PRE \times REC}{PRE + REC}. \quad (9)$$

Computational costs are always used in computer science to evaluate an algorithm. In this study, we considered the time complexity and the space complexity, which could be represented by the average running time and the required memory occupancy of the several algorithms in each model as previously described (Zhao et al., 2020).

**TABLE 3** The performance of AttnTAP with different encoding methods.

	McPAS-TCR		VDJdb	
	ACC <sup>a</sup>	AUC	ACC	AUC
Random initialization	<b>0.788</b>	<b>0.878</b>	0.843	0.910
Amino acid word vector	0.784	0.871	<b>0.847</b>	<b>0.911</b>
Triplet word vector	0.616	0.678	0.827	0.878

<sup>a</sup>Abbreviations: ACC: accuracy; AUC: area under the receiver operating characteristic curve.

**TABLE 4** The performance of AttnTAP under varied TCR feature extraction models.

		ACC <sup>a</sup>	REC	PRE	F1	AUC
McPAS-TCR	I <sup>b</sup>	0.736	0.803	0.708	0.752	0.827
	II	0.762	0.803	0.743	0.772	0.854
	III	0.766	0.807	0.747	0.775	0.857
	IV	0.774	0.755	0.755	0.758	0.861
	V	<b>0.781</b>	<b>0.818</b>	<b>0.762</b>	<b>0.789</b>	<b>0.869</b>
VDJdb	I	0.840	0.820	0.855	0.837	0.906
	II	0.839	0.799	0.868	0.832	0.901
	III	0.842	0.806	0.869	0.836	0.904
	IV	0.844	0.820	0.861	0.840	0.908
	V	<b>0.847</b>	<b>0.829</b>	<b>0.870</b>	<b>0.844</b>	<b>0.914</b>

<sup>a</sup>Abbreviations: ACC: accuracy; REC: recall; PRE: precision; F1: F1 score; AUC: area under the receiver operating characteristic curve.

<sup>b</sup>Model numbers: I: the multilayer perceptron model; II: the two-layer long short term memory (LSTM) model; III: the one-layer bi-directional LSTM model; IV: the two-layer LSTM model with attention mechanism; and V: Attn-BiLSTM model.

## 3 Results

### 3.1 AttnTAP model performance

#### 3.1.1 AttnTAP performance on different encoding methods

Three pre-training word vectors, random initialization vectors, amino acid word vectors, and triplet word vectors, were tested in the Attn-BiLSTM and Attn-MLP model, to validate their effectiveness on AttnTAP classification (Table 3). The ACC and AUC were used to evaluate the performance of the three different encoding methods on the balanced McPAS-TCR and VDJdb datasets. The random initialization vectors and amino acid word vectors showed better performance on two datasets, while the triplet word vector had the worst performance. The prediction accuracies of random initialization vectors, whose computational cost was much less, were similar to those of amino acid word vectors. Thus, the random initialization

vectors were used for sequence encoding to improve the prediction accuracy of AttnTAP.

#### 3.1.2 AttnTAP performance on five different TCR feature extraction models

To assess the ability of the feature extraction method at predicting accuracy, we tested the five different TCR extraction methods based on the balanced McPAS-TCR and VDJdb datasets. The five different TCR feature extraction methods were (I) the MLP model with the most suitable parameters by grid search algorithm; (II) the two-layer LSTM model used in ERGO; (III) the BiLSTM model with the same parameters as model II; (IV) the model II with an attention mechanism; and (V) Attn-BiLSTM, the model III with an attention mechanism. We summarized their performances under the AttnTAP framework with the TPP-I task. The five-fold CV results on McPAS-TCR and VDJdb datasets are shown in Table 4.

The results revealed that the BiLSTM model (model III) performed better than the MLP (model I) and LSTM (model II) on the McPAS-TCR dataset, and their three models had similar performance on the VDJdb dataset. The BiLSTM outperformed other feature extraction models without attention mechanism because it considered both precursor and successor amino acids, which extracted information on the interrelationships between amino acids in a more rational way. The models with attention mechanism, especially Attn-BiLSTM (model V), outperformed the other models without attention mechanism in terms of their ACC, REC, PRE, recall, F1 score, and AUC, which indicated that attention algorithms could focus on the key amino acids when processing large amounts of CDR3 information and improve the feature extraction. In AttnTAP, the BiLSTM layer and subsequent attention layer formed the main part of the CDR3 feature extraction model. The attention mechanism assigned various weights to the amino acid features output by the BiLSTM layer, correctly modeling the interrelationships between amino acids and paying more attention to the amino acids that contributed to the antigen-binding specificity (Supplementary Figure S2). As a result, Attn-BiLSTM achieved the highest, and balanced REC (mean 0.818 and 0.829 on McPAS-TCR and VDJdb, respectively) and PRE (mean 0.762 and 0.870 on McPAS-TCR and VDJdb, respectively) on two datasets. Furthermore, the AUC value of Attn-BiLSTM had reached as high as 0.869 and 0.914 on McPAS-TCR and VDJdb. To some extent, the BiLSTM model based on the attention mechanism could improve the performance of TCR-peptide prediction accuracy.

#### 3.1.3 AttnTAP performance on the unbalanced dataset

A real TCR repertoire usually contains more negative samples than positive samples. To validate the performance of the AttnTAP model on an unbalanced dataset and make it suitable for practice, we attempted to generate 14 unbalanced datasets (the ratio of negative to positive samples ranged from 2 to 15) using Supplementary



TABLE 5 The AUC of AttnTAP on unbalanced datasets.

Ratio	McPAS-TCR	VDJdb	Ratio	McPAS-TCR	VDJdb
1:1 <sup>a</sup>	0.838 <sup>b</sup>	0.908	1:9	0.865	0.914
1:2	0.853	0.910	1:10	0.870	0.911
1:3	0.854	0.912	1:11	0.872	0.912
1:4	0.863	0.913	1:12	0.873	0.912
1:5	0.862	0.911	1:13	0.872	0.913
1:6	0.871	0.909	1:14	0.870	0.912
1:7	0.867	0.909	1:15	0.868	0.913
1:8	0.870	0.912	-	-	-

<sup>a</sup>It denotes the ratio of positive samples to negative samples in the dataset.  
<sup>b</sup>We used the metric, area under the receiver operating characteristic curve, to evaluate the performance of the model.

TABLE 6 The performance evaluation of TPP-I task.

		ACC <sup>a,b</sup>	REC	PRE	F1	AUC
McPAS-TCR	ERGO-LSTM	0.748 ± 0.004	0.747 ± 0.013	0.748 ± 0.007	0.747 ± 0.006	0.831 ± 0.005
	ERGO-AE	0.734 ± 0.004	0.696 ± 0.020	0.754 ± 0.009	0.722 ± 0.008	0.808 ± 0.004
	ImRex	0.631 ± 0.003	0.625 ± 0.005	0.648 ± 0.005	0.636 ± 0.004	0.694 ± 0.003
	DLpTCR	0.502 ± 0.003	0.500 ± 0.004	<b>0.861 ± 0.003</b>	0.633 ± 0.003	0.529 ± 0.004
	NetTCR-2.0	0.728 ± 0.004	0.734 ± 0.010	0.715 ± 0.018	0.722 ± 0.006	0.799 ± 0.004
	AttnTAP	<b>0.758 ± 0.003</b>	<b>0.769 ± 0.013</b>	0.752 ± 0.007	<b>0.760 ± 0.005</b>	<b>0.840 ± 0.003</b>
VDJdb	ERGO-LSTM	0.834 ± 0.003	0.790 ± 0.004	0.864 ± 0.004	0.825 ± 0.003	0.889 ± 0.003
	ERGO-AE	0.837 ± 0.003	0.798 ± 0.006	0.864 ± 0.006	0.829 ± 0.004	0.891 ± 0.003
	ImRex	0.561 ± 0.004	0.556 ± 0.005	0.571 ± 0.006	0.564 ± 0.005	0.598 ± 0.004
	DLpTCR	0.482 ± 0.005	0.487 ± 0.004	0.861 ± 0.004	0.622 ± 0.004	0.503 ± 0.005
	NetTCR-2.0	0.832 ± 0.003	<b>0.851 ± 0.008</b>	0.802 ± 0.007	0.826 ± 0.003	0.890 ± 0.002
	AttnTAP	<b>0.839 ± 0.003</b>	0.801 ± 0.006	<b>0.865 ± 0.004</b>	<b>0.831 ± 0.003</b>	<b>0.894 ± 0.002</b>

<sup>a</sup>The results show 95% confidence intervals for all the validations (totally 30 validations for each cross-validation).  
<sup>b</sup>Abbreviations: ACC: accuracy; REC: recall; PRE: precision; F1: F1 score; AUC: area under the receiver operating characteristic curve.

Algorithm S1 in this section. The five-fold CV was used to evaluate the performance of AttnTAP on different unbalanced data (Table 5 and Supplementary Figure S3).

The average AUC on the McPAS-TCR dataset had been rising from 0.838 to 0.873 during the increased number of negative samples, while the average AUC on the VDJdb dataset had reached 0.9 across all the unbalanced data. The AUC performance results indicated that AttnTAP could consistently perform well on unbalanced datasets with an increased number of negative samples.

### 3.2 Performance evaluation of comparative approaches

#### 3.2.1 Performance evaluation of the TPP-I task

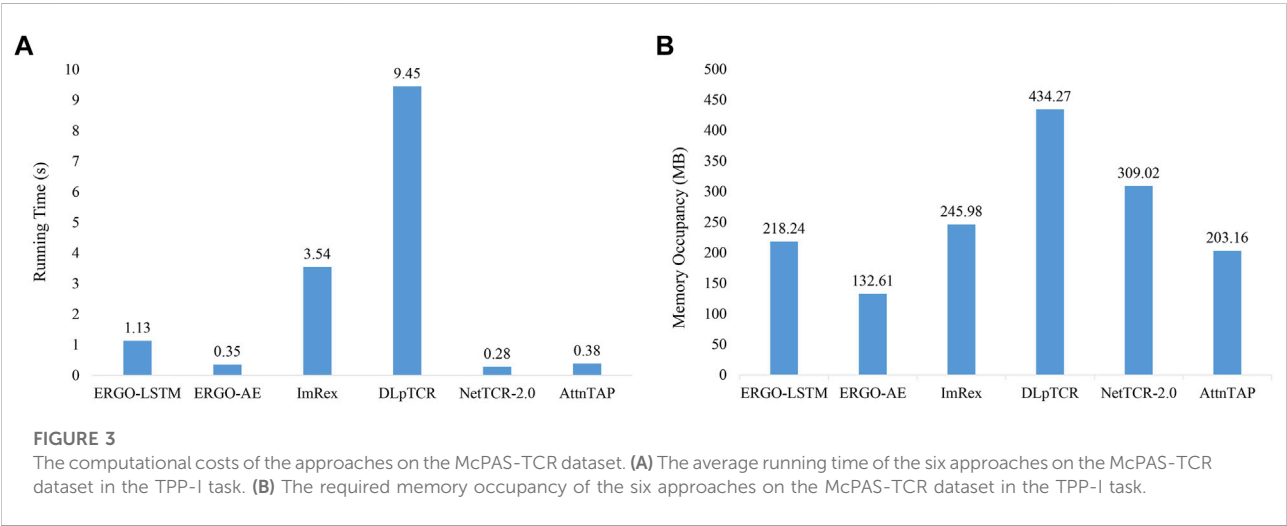
According to the requirements of the six deep neural networks (Table 2), we selected only the CDR3 β chains (without the α chains)

and discarded the extra amino acids of the sequences longer than the maximum length input. We performed five-fold CVs six times to reduce the unbiased evaluation. We trained the pre-training models of ERGO-LSTM, ERGO-AE, NetTCR-2.0, and AttnTAP, while the pre-training models of ImRex and DLpTCR were downloaded directly (<https://github.com/pmoris/ImRex/>; <https://github.com/jiangBiolab/DLpTCR/>) as previously described (Montemurro et al., 2021; Xu et al., 2021). We calculated the scores of five measurements for the different TCR-peptide binding prediction approaches across the two basic datasets. The ACC, REC, PRE, F1, and AUC values, with 95% confidence intervals, for a total of 30 validations experiments, were statistically analyzed (Table 6 and Supplementary Table S2). Briefly, among six prediction approaches, AttnTAP had the highest mean AUC values on both two datasets (the mean values were 0.84 on McPAS-TCR and 0.894 on VDJdb), and the AUC values ranged from 0.824 to 0.860 on McPAS-TCR and ranged from 0.882 to 0.905 on VDJdb (Supplementary Table S2). Moreover, AttnTAP outperformed all

TABLE 7 The performance evaluation of TPP-II task.

		ACC <sup>a,b</sup>	REC	PRE	F1	AUC
McPAS-TCR	ERGO-LSTM	0.735 ± 0.005	0.761 ± 0.016	0.724 ± 0.009	0.741 ± 0.006	0.818 ± 0.004
	ERGO-AE	0.731 ± 0.005	0.672 ± 0.022	0.764 ± 0.012	0.712 ± 0.009	0.800 ± 0.005
	ImRex	0.627 ± 0.004	0.621 ± 0.006	0.644 ± 0.007	0.632 ± 0.005	0.690 ± 0.005
	DLpTCR	0.501 ± 0.003	0.499 ± 0.003	<b>0.859 ± 0.004</b>	0.631 ± 0.003	0.524 ± 0.004
	NetTCR-2.0	0.731 ± 0.004	0.746 ± 0.008	0.699 ± 0.018	0.720 ± 0.008	0.804 ± 0.004
	AttnTAP	<b>0.755 ± 0.005</b>	<b>0.778 ± 0.011</b>	0.743 ± 0.006	<b>0.760 ± 0.006</b>	<b>0.837 ± 0.004</b>
VDJdb	ERGO-LSTM	0.832 ± 0.003	0.794 ± 0.007	0.860 ± 0.005	0.825 ± 0.004	0.891 ± 0.003
	ERGO-AE	0.836 ± 0.003	0.800 ± 0.009	0.864 ± 0.005	0.830 ± 0.004	0.888 ± 0.004
	ImRex	0.561 ± 0.005	0.560 ± 0.006	0.575 ± 0.006	0.568 ± 0.006	0.597 ± 0.006
	DLpTCR	0.488 ± 0.004	0.494 ± 0.004	0.862 ± 0.004	0.628 ± 0.004	0.510 ± 0.004
	NetTCR-2.0	0.832 ± 0.003	<b>0.860 ± 0.007</b>	0.794 ± 0.009	0.825 ± 0.004	0.891 ± 0.003
	AttnTAP	<b>0.838 ± 0.003</b>	0.794 ± 0.006	<b>0.872 ± 0.004</b>	<b>0.831 ± 0.004</b>	<b>0.893 ± 0.003</b>

<sup>a</sup>The results show 95% confidence intervals for totally 30 independent experiments.  
<sup>b</sup>Abbreviations: ACC: accuracy; REC: recall; PRE: precision; F1: F1 score; AUC: area under the receiver operating characteristic curve.



other methods overall with respect to the other four metrics, where, in particular, the REC and PRE of its prediction results on the datasets were balanced, indicating its good robustness and stability. Therefore, the AttnTAP was an optimal framework for predicting a TCR-peptide binding.

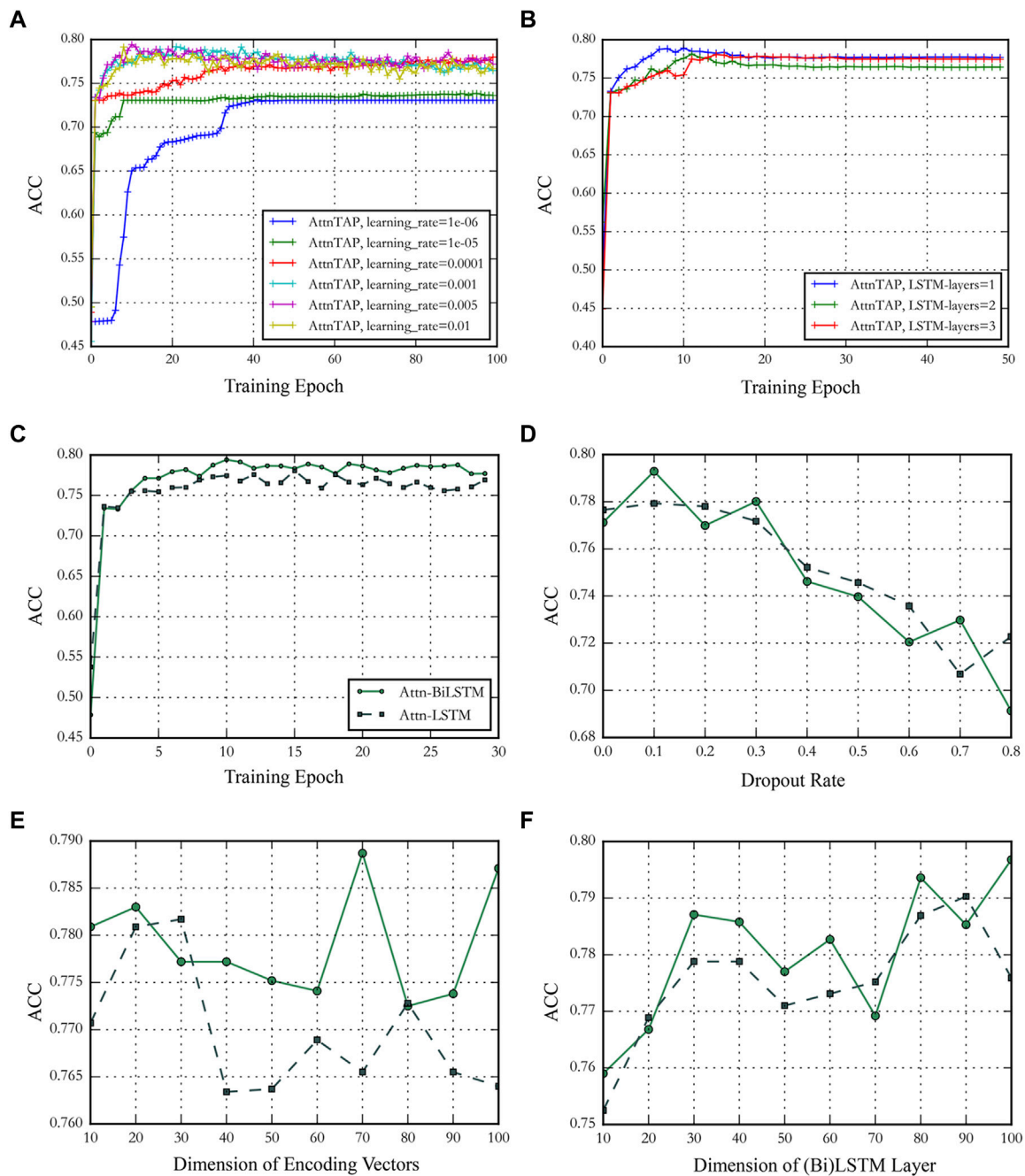
3.2.2 Performance evaluation of the TPP

To further validate the generalization performance of these methods, we evaluated them in the TPP-II task and conducted independent replicate experiments 30 times. Similar to the TPP-I task, the AttnTAP model achieved the highest AUC values (the mean values were 0.837 on McPAS-TCR and 0.893 on VDJdb) (Table 7), and the AUC values

ranged from 0.810 to 0.864 on McPAS-TCR and ranged from 0.873 to 0.908 on VDJdb in the TPP-II task (Supplementary Table S3). Moreover, it had better overall performance than other methods in terms of the other four metrics, with a balanced REC and PRE. As a result, compared with the existing methods, AttnTAP had better generalization and could perform better on new data.

3.2.3 Computational costs of approaches

In this study, the average running time was recorded 30 times independent experiments (Figure 3A and Supplementary Table S2). Figure 3A demonstrates that NetTCR-2.0, ERGO-AE, and AttnTAP had similar running times, which was much less than the other three

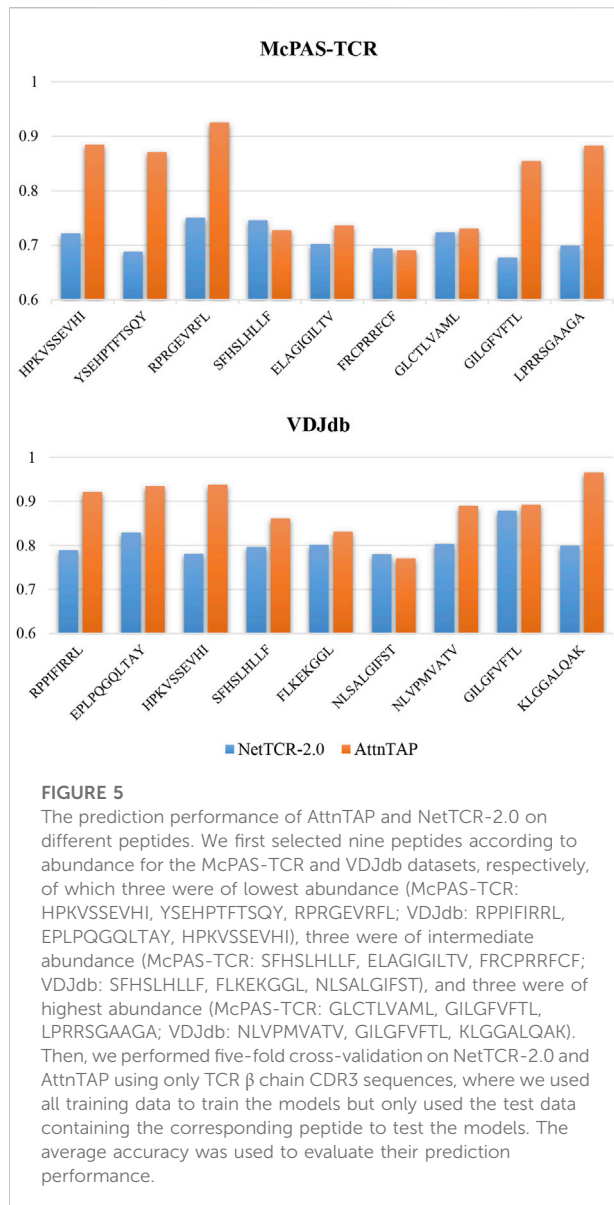


**FIGURE 4**

The performance of AttnTAP with different hyperparameters. (A,B) Panels showed the performance of AttnTAP with different learning rates and bi-directional long short-term (BiLSTM) layer numbers. (C–F) Panels depicted the performance of AttnTAP using LSTM/BiLSTM with different training epochs, dropout rates, dimensions of encoding vectors, and LSTM/BiLSTM layers, respectively.

approaches, while DLpTCR achieved the longest running time, which indicated that DLpTCR had a higher complexity of model configuration. The required memory occupancy of all the six approaches on the McPAS-TCR datasets was also recorded and averaged for comparison (Figure 3B and Supplementary Table S2).

The running ERGO-AE with the minimal space and followed by AttnTAP, whereas the DLpTCR had the largest space occupancy for its complex framework. Thus, AttnTAP improved the accuracy of TCR-peptide binding prediction while being quite efficient in terms of computational time and memory usage.



## 4 Discussion

The prediction of TCRs binding to the peptide is urgent in a clinical, but still extremely challenging, with highly cross-reactive TCRs and peptides, unseen peptides lack biological verification, and limited available training samples (Rudolph et al., 2006; Szeto et al., 2020; Moris et al., 2021). The breakthrough of deep convolutional neural networks in predicting TCR-peptide binding accuracy, accelerating well-integrated human immune repertoire, and potentially interacting peptides prediction pipelines. However, a few remaining issues led us to design this experiment. In this study, we designed the attention mechanism under the Attn-BiLSTM framework, considering the various contributions of

amino acids in CDR3 sequences. Then, a dual input of CDR3 sequences and peptides was needed to improve the prediction accuracy, instead of separate embedding steps ignoring the two protein molecular interactors. The experimental results also showed the AttnTAP achieved a good performance in TCR-peptide binding prediction.

Due to the high dimensionality, non-homogeneous, and sparsity of TCR repertoire data, we proposed a novel and unified architecture, which combined a bi-directional LSTM (BiLSTM), an attention mechanism, and a convolutional layer. The BiLSTM extracted TCR features by considering both the preceding and succeeding amino acid representations of a single CDR3 chain (Zhou et al., 2016). Moreover, an attention mechanism was employed to give a different focus to the information outputted from the hidden layers of BiLSTM. In Supplementary Figure S2, the weight of amino acids in a CDR3 chain varies greatly at different positions, with the color changed from light to dark. It is a biological truism that high weights (dark) tend to appear in the middle region of a CDR3 chain (Robins et al., 2009), and the weighting pattern displayed by AttnTAP on most CDR3 sequences was consistent with this truism. However, some sequences had special weighting patterns, showing strong weighting at the beginning or ending amino acids (N- or C- terminus of the CDR loop). We analyzed the attention weight condition of 1957 test samples from the VDJdb dataset in one five-fold CV test. We found that AttnTAP exhibited strong weighting for their beginning part only on 59 CDR3 sequences, which represented only 0.03 of all the samples, and these sequences corresponded to 31 different peptides. Furthermore, some CDR3 sequences showed strong weighting at the terminal amino acids (C- terminus) of the shorter sequences as well as the placeholders. Given that the attention mechanism may assign higher weights to the boundary part, where the anterior and posterior position features differ, AttnTAP focused on the terminal amino acid “F” and the placeholders, taking into account the sequence length feature. In addition, we also speculated that some CDR3 sequences had unexpected patterns due to the strong V or J region preferences or the dataset biases. Although most CDR3 sequences have a similar beginning or ending (e.g., beginning with “C” and ending with “F”), these similar beginnings and endings may still form specific combinations with highly variable amino acids in the middle of the sequences, which allows the sequences to possess antigen-binding specificity.

As is well-known, an adjustable hyperparameter, including the learning rate, the number of BiLSTM layers, the training epoch, and the dropout rate, could balance the latent channel capacity and improve the prediction accuracy (Graves et al., 2013; Zhou et al., 2016). We conducted a series of experiments on the McPAS-TCR dataset to validate the effect of different hyperparameters on model prediction

performances and determine the value of the hyperparameters based on the results. We used the metric ACC to evaluate the model prediction accuracy in the experiments (Figure 4). Four hyperparameters, including training epoch, dropout rate, dimension of encoding vectors, and the dimension of LSTM/BiLSTM layers, were used to compare the performance of Attn-LSTM and Attn-BiLSTM (Figures 4C–F). BiLSTM was an ideal model under the different hyperparameters conditions. Thus, in this study, we set the training epoch, the dropout rate, the dimensions of amino acid encoding vectors, and the BiLSTM layer to 10, 0.1, 70, and 80 for AttnTAP, respectively, according to the results. The ACC had deteriorated significantly when the learning rate was below 0.0001, thus we set the threshold to 0.001 for compatibility with the application in the various dataset (Figure 4A). There was no significant improvement in model performance as the number of BiLSTM layers increased, we used one-layer BiLSTM to reduce model complexity (Figure 4B).

ImRex and DLpTCR had lower prediction accuracies than the other four approaches under TPP-I and TPP-II tasks, maybe due to the overfitting caused by their complex model structures. We reduced the complexity of AttnTAP by using one-layer BiLSTM instead of multi-layer BiLSTM to extract TCR sequences features and the MLP model instead of the LSTM model to extract peptide features to avoid the overfitting. The results of AttnTAP in TPP-II were similar to those in TPP-I, which indicated that AttnTAP had a robust and good generalization in predicting an unseen TCR sequence binding to a peptide. Thus, the AttnTAP presented here could serve as an unseen TCR-peptide prediction method, for accelerating identifying neoantigens and activated T cells for immunotherapy clinically.

In addition to the performances of AttnTAP on the entire McPAS-TCR and VDJdb datasets, we also evaluated its performances on different peptides, especially the peptides with low abundance, in the TPP-I task. The abundance of peptides in the McPAS-TCR dataset ranged from 0.005 to 0.219, and from 0.001 to 0.356 in the VDJdb dataset (Supplementary Table S1). We selected nine peptides according to their abundances (high-, medium- and low-abundance accounted for one-third) for the McPAS-TCR and VDJdb datasets, respectively (Supplementary Table S1). Considering that NetTCR-2.0 is the latest method for TCR-peptide binding prediction and has high prediction accuracies with low computational cost, we selected it as the baseline model. We performed a five-fold CV on NetTCR-2.0 and AttnTAP using only TCR  $\beta$  chain CDR3 sequences and compared their performance by average ACC. In detail, we used all training data to train the models, while only used the test data containing the corresponding peptide to test the models (Figure 5 and Supplementary Table S4). On the McPAS-TCR dataset, the average ACCs of AttnTAP and NetTCR-2.0 were 0.894 and 0.720 for the lowest abundance peptides, 0.718 and 0.714 for the

intermediate abundance peptides, and 0.823 and 0.700 for the highest abundance peptides. Moreover, on the VDJdb dataset, their average ACCs were 0.932 and 0.800 for the lowest abundance peptides, 0.821 and 0.793 for the intermediate abundance peptides, and 0.916 and 0.828 for the highest abundance peptides, respectively. The results indicated that AttnTAP had higher ACCs than NetTCR-2.0 on most of the peptides and had similar performances to the latter on the other peptides (e.g., SFHSLHLLF and FRCPRRFCF in the McPAS-TCR dataset and NLSALGIFST in the VDJdb dataset). In our opinion, the AttnTAP framework had a good performance on TCR-peptide binding prediction, especially the low-abundance peptides, due to its BiLSTM model with attention mechanism in extracting CDR3 features, which validated that AttnTAP has good stability and robustness.

In conclusion, we successfully trained a dual-input model to predict the interactions between seen and unseen TCRs and peptides. Due to the limited training samples and known peptides we had available, we tried to reduce the complexity of the model to avoid overfitting on the premise of prediction accuracy. In the future, we will consider more information on TCR sequences, such as the CDR1 and CDR2, or TCR $\alpha$  chain when data become available, to train a good performance and more generalization prediction model to be suitable for multi-types data, meeting the urgent clinical needs.

## Data availability statement

AttnTAP is available on GitHub, at <https://github.com/Bioinformatics7181/AttnTAP/>, for academic use only. The publicly available data for this study can be found in VDJdb (<https://vdjdb.cdr3.net/>), IEDB (<http://www.iedb.org/>), and McPAS-TCR (<http://friedmanlab.weizmann.ac.il/McPAS-TCR/>). The pre-training models of ImRex and DLpTCR can be found on Github (<https://github.com/pmoris/ImRex/>; <https://github.com/jiangBiolab/DLpTCR/>); further inquiries can be directed to the corresponding author.

## Author contributions

JW, YX, and YT conceived and designed the experiments; XQ and FL performed the experiments; YX and XQ analyzed the data; YX contributed materials; JW, YX, and XQ wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

## Funding

This research was funded by the Natural Science Basic Research Program of Shaanxi, grant number 2020JC-01.



## Conflict of interest

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

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this study can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2022.942491/full#supplementary-material>

## References

- Asgari, E., and Mofrad, M. (2015). Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS One* 10 (11), e0141287. doi:10.1371/journal.pone.0141287
- Bagaev, D. V., Vroomans, R. M. A., Samir, J., Stervbo, U., Rius, C., Dolton, G., et al. (2020). VDJdb in 2019: Database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Res.* 48 (D1), D1057–D1062. doi:10.1093/nar/gkz874
- Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv*. doi:10.48550/arXiv.1409.0473
- Bolotin, D. A., Mamedov, I. Z., Britanova, O. V., Zvyagin, I. V., Shagin, D., Ustyugova, S. V., et al. (2012). Next generation sequencing for TCR repertoire profiling: Platform-specific features and correction algorithms. *Eur. J. Immunol.* 42 (11), 3073–3083. doi:10.1002/eji.201242517
- Chiffelle, J., Genolet, R., Perez, M. A., Coukos, G., Zoete, V., and Harari, A. (2020). T-cell repertoire analysis and metrics of diversity and clonality. *Curr. Opin. Biotechnol.* 65, 284–295. doi:10.1016/j.copbio.2020.07.010
- Crooks, G., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004). Weblogo: A sequence logo generator. *Genome Res.* 14 (6), 1188–1190. doi:10.1101/gr.849004
- Glanville, J., Huang, H., Nau, A., Hatton, O., Wagar, L. E., Rubelt, F., et al. (2017). Identifying specificity groups in the T cell receptor repertoire. *Nature* 547 (7661), 94–98. doi:10.1038/nature22976
- Graves, A., Mohamed, A. R., and Hinton, G. (2013). "Speech recognition with deep recurrent neural networks," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013 (IEEE), 6645–6649. doi:10.48550/arXiv.1303.5778
- Henikoff, S., and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.* 89, 10915–10919. doi:10.1073/pnas.89.22.10915
- Joglekar, A. V., and Li, G. (2021). T cell antigen discovery. *Nat. Methods* 18 (8), 873–880. doi:10.1038/s41592-020-0867-z
- La Gruta, N. L., Gras, S., Daley, S. R., Thomas, P. G., and Rossjohn, J. (2018). Understanding the drivers of MHC restriction of T cell receptors. *Nat. Rev. Immunol.* 18 (7), 467–478. doi:10.1038/s41577-018-0007-5
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521 (7553), 436–444. doi:10.1038/nature14539
- Mahajan, S., Vita, R., Shackelford, D., Lane, J., Schulten, V., Zarebski, L., et al. (2018). Epitope specific antibodies and T cell receptors in the immune epitope database. *Front. Immunol.* 9, 2688. doi:10.3389/fimmu.2018.02688
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*. doi:10.48550/arXiv.1301.3781
- Montemurro, A., Schuster, V., Povlsen, H. R., Bentzen, A. K., Jurtz, V., Chronister, W. D., et al. (2021). NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCR $\alpha$  and  $\beta$  sequence data. *Commun. Biol.* 4 (1), 1060. doi:10.1038/s42003-021-02610-3
- Moris, P., De Pauw, J., Postovskaya, A., Gielis, S., De Neuter, N., Bittremieux, W., et al. (2021). Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification. *Brief. Bioinform.* 22 (4), bbab318. doi:10.1093/bib/bbaa318
- Robins, H. S., Campregher, P. V., Srivastava, S. K., Wachter, A., Turtle, C. J., Kahsai, O., et al. (2009). Comprehensive assessment of T-cell receptor beta-chain diversity in alpha $\beta$  T cells. *Blood* 114 (19), 4099–4107. doi:10.1182/blood-2009-04-217604
- Rudolph, M. G., Stanfield, R. L., and Wilson, I. A. (2006). How TCRs bind MHCs, peptides, and coreceptors. *Annu. Rev. Immunol.* 24, 419–466. doi:10.1146/annurev.immunol.23.021704.115658
- Schneider, T., and Stephens, R. (2002). Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.* 18 (20), 6097–6100. doi:10.1093/nar/18.20.6097
- Springer, I., Besser, H., Tickotsky-Moskovitz, N., Dvorkin, S., and Louzoun, Y. (2020). Prediction of specific TCR-peptide binding from large dictionaries of TCR-peptide pairs. *Front. Immunol.* 11, 1803. doi:10.3389/fimmu.2020.01803
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15 (1), 1929–1958.
- Szeto, C., Lobos, C. A., Nguyen, A. T., and Gras, S. (2020). TCR recognition of peptide-MHC-I: Rule makers and breakers. *Int. J. Mol. Sci.* 22 (1), 68. doi:10.3390/ijms22010068
- Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E., and Friedman, N. (2017). McPAS-TCR: A manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics* 33 (18), 2924–2929. doi:10.1093/bioinformatics/btx286
- Tran, N. H., Xu, J., and Li, M. (2022). A tale of solving two computational challenges in protein science: Neoantigen prediction and protein structure prediction. *Brief. Bioinform.* 23 (1), bbab493. doi:10.1093/bib/bbab493
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst.* 3030 (NIPS), 1–15. doi:10.48550/arXiv.1706.03762
- Wagih, O. (2017). Ggseqlogo: A versatile R package for drawing sequence logos. *Bioinformatics* 33 (22), 3645–3647. doi:10.1093/bioinformatics/btx469
- Warren, R. L., Freeman, J. D., Zeng, T., Choe, G., Munro, S., Moore, R., et al. (2011). Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res.* 21 (5), 790–797. doi:10.1101/gr.115428.110
- Woodsworth, D. J., Castellarin, M., and Holt, R. A. (2013). Sequence analysis of T-cell repertoires in health and disease. *Genome Med.* 5 (10), 98. doi:10.1186/gm502
- Xu, Z., Luo, M., Lin, W., Xue, G., Wang, P., Jin, X., et al. (2021). DLpTCR: An ensemble deep learning framework for predicting immunogenic peptide recognized by T cell receptor. *Brief. Bioinform.* 22, 1–13. doi:10.1093/bib/bbab335
- Xu, Y., Qian, X., Zhang, X., Lai, X., Liu, Y., Wang, J., et al. (2022). DeepLION: Deep Multi-Instance Learning Improves the Prediction of Cancer-Associated T Cell Receptors for Accurate Cancer Detection. *Front. Genet.* 13:860510. doi:10.3389/fgene.2022.860510
- Zemouri, R., Zerhouni, N., and Racocanu, D. (2019). Deep learning in the biomedical applications: Recent and future status. *Appl. Sci. (Basel)*. 9 (8), 1526. doi:10.3390/app9081526
- Zhao, L., Liu, H., Yuan, X., Gao, K., and Duan, J. (2020). Comparative study of whole exome sequencing-based copy number variation detection tools. *BMC Bioinform.* 21 (1), 97. doi:10.1186/s12859-020-3421-1
- Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., et al. (2016). Attention-based bidirectional long short-term memory networks for relation classification. *Proc. 54th Annu. Meet. Assoc. Comput. Linguistics* 2, 207–212. doi:10.18653/v1/P16-2034

# Advantages of publishing in Frontiers



## OPEN ACCESS

Articles are free to read  
for greatest visibility  
and readership



## FAST PUBLICATION

Around 90 days  
from submission  
to decision



## HIGH QUALITY PEER-REVIEW

Rigorous, collaborative,  
and constructive  
peer-review



## TRANSPARENT PEER-REVIEW

Editors and reviewers  
acknowledged by name  
on published articles

## Frontiers

Avenue du Tribunal-Fédéral 34  
1005 Lausanne | Switzerland

Visit us: [www.frontiersin.org](http://www.frontiersin.org)

Contact us: [frontiersin.org/about/contact](http://frontiersin.org/about/contact)



## REPRODUCIBILITY OF RESEARCH

Support open data  
and methods to enhance  
research reproducibility



## DIGITAL PUBLISHING

Articles designed  
for optimal readership  
across devices



## FOLLOW US

@frontiersin



## IMPACT METRICS

Advanced article metrics  
track visibility across  
digital media



## EXTENSIVE PROMOTION

Marketing  
and promotion  
of impactful research



## LOOP RESEARCH NETWORK

Our network  
increases your  
article's readership