Machine learning empowered discovery of bioactive peptides from food proteins and beyond

by

Zhenjiao Du

B.E., Wuhan Polytechnic University, 2018
M.E., China Agricultural University, 2020

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Grain Science and Industry
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2024

Abstract

The increasing global demand for sustainable and natural health-promoting food ingredients has led to a growing interest in bioactive peptides derived from food proteins. These peptides exhibit various beneficial biological effects, such as antioxidant and antihypertensive activities. Traditional methods for producing, characterizing, and identifying bioactive peptides are often labor-intensive, time-consuming, and resource-intensive. The goal of this study was to employ advanced computational techniques for prediction model development to accelerate bioactive peptide discovery and extend applications into broader biological domains.

Specific objectives were to: 1) develop traditional chemometric methods for antioxidant peptide discovery: investigate predictive quantitative structure-activity relationship (QSAR) model development for antioxidant dipeptides and tripeptides, experimentally validate model performance, design an *in silico* hydrolysis simulation tool, and develop a workflow for applying these predictive models in peptide discovery from food proteins; 2) pre-train protein language models (pLM) for bioactive peptide discovery: build a universal architecture for peptide prediction model development and evaluate it across different peptide datasets; develop state-of-the-art (SOTA) models for allergenic proteins and peptides and investigate the effects of pLM sizes on prediction model performance; build a SOTA prediction model for high-activity angiotensin-I converting enzyme (ACE) inhibitory peptides and explore the superiority of pLMs over traditional peptide representation methods and assess the compatibility of different machine learning classifiers with pLMs; 3) fuse knowledge between language models for performance enhancement: develop multimodal prediction models for enzyme-substrate pair prediction using protein and chemistry language models to enhance performance through knowledge fusion; 4) build accessibility and user-friendliness of prediction models: deploy the developed prediction

models to user-friendly web servers to make them and their codes easily accessible for scientific discovery.

Two machine learning based QSAR models were developed with SOTA performance for predicting the antioxidant activity of tripeptides and dipeptides, respectively. Notably, the tripeptide model achieved an R^2 of 0.847 on the test dataset. The predicted antioxidant activities of peptides were confirmed through experimental validation. Subsequent feature analysis revealed that C-terminal residues in tripeptides and N-terminal residues in dipeptides contributed more significantly to antioxidant activity. Finally, a hydrolysis simulation tool (R-PeptideCutter) was developed to connect food proteins with predictive models for antioxidant peptide discovery from sorghum proteins, leading to the identification of a high-antioxidant-activity dipeptide, YR.

Utilizing pLMs, our universal architecture UniDL4BioPep outperformed existing SOTA models in 15 out of 20 peptide datasets, with accuracy improvements ranging from 0.7% to 7%, and demonstrated great potential for other peptide prediction tasks with custom datasets. To enhance usability, a well-annotated template was designed for non-programming users, and an advanced version was introduced for extremely imbalanced datasets. The pLM-based models exhibited SOTA performance in predicting high activity ACE inhibitory peptides with an accuracy of 88.3% and allergenic proteins/peptides with an accuracy of 95.1%, offering solutions to health concerns like hypertension and allergies. The pLMs demonstrated superiority over traditional peptide descriptors in peptide representation and showed strong compatibility with support vector machine, multilayer perceptron, and logistic regression in the ACE inhibitory peptide dataset. A positive correlation between pLM sizes and prediction model performance was observed in the allergenic protein and peptide dataset. All developed prediction models were deployed on user-friendly web servers for scientific exploration.

Beyond the application of pLMs in peptide discovery, a multimodal model, FusionESP, was presented for enzyme-substrate interaction prediction. By employing a novel projection head based on a contrastive learning strategy for knowledge fusion, FusionESP successfully achieved SOTA performance with an accuracy of 94.7% by integrating protein and chemistry language models, while requiring fewer computational resources and data points.

Overall, this work demonstrates the transformative potential of integrating advanced computational techniques—such as machine learning and deep learning—into bioactive peptide research and broader biological domains. The developed models and tools can significantly accelerate bioactive peptide discovery and provide valuable resources for scientific exploration. By making these tools accessible and user-friendly, this work contributes to sustainable agriculture, improved health outcomes, and fosters innovation across food science, biochemistry, and computational biology.

Machine learning empowered discovery of bioactive peptides from food proteins and beyond

by

Zhenjiao Du

B.E., Wuhan Polytechnic University, 2018
M.E., China Agricultural University, 2020

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Grain Science and Industry
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2024

Approved by:

Major Professor
Dr. Yonghui Li

Copyright

Abstract

The increasing global demand for sustainable and natural health-promoting food ingredients has led to a growing interest in bioactive peptides derived from food proteins. These peptides exhibit various beneficial biological effects, such as antioxidant and antihypertensive activities. Traditional methods for producing, characterizing, and identifying bioactive peptides are often labor-intensive, time-consuming, and resource-intensive. The goal of this study was to employ advanced computational techniques for prediction model development to accelerate bioactive peptide discovery and extend applications into broader biological domains.

Specific objectives were to: 1) develop traditional chemometric methods for antioxidant peptide discovery: investigate predictive quantitative structure-activity relationship (QSAR) model development for antioxidant dipeptides and tripeptides, experimentally validate model performance, design an *in silico* hydrolysis simulation tool, and develop a workflow for applying these predictive models in peptide discovery from food proteins; 2) pre-train protein language models (pLM) for bioactive peptide discovery: build a universal architecture for peptide prediction model development and evaluate it across different peptide datasets; develop state-of-the-art (SOTA) models for allergenic proteins and peptides and investigate the effects of pLM sizes on prediction model performance; build a SOTA prediction model for high-activity angiotensin-I converting enzyme (ACE) inhibitory peptides and explore the superiority of pLMs over traditional peptide representation methods and assess the compatibility of different machine learning classifiers with pLMs; 3) fuse knowledge between language models for performance enhancement: develop multimodal prediction models for enzyme-substrate pair prediction using protein and chemistry language models to enhance performance through knowledge fusion; 4) build accessibility and user-friendliness of prediction models: deploy the developed prediction

models to user-friendly web servers to make them and their codes easily accessible for scientific discovery.

Two machine learning based QSAR models were developed with SOTA performance for predicting the antioxidant activity of tripeptides and dipeptides, respectively. Notably, the tripeptide model achieved an R^2 of 0.847 on the test dataset. The predicted antioxidant activities of peptides were confirmed through experimental validation. Subsequent feature analysis revealed that C-terminal residues in tripeptides and N-terminal residues in dipeptides contributed more significantly to antioxidant activity. Finally, a hydrolysis simulation tool (R-PeptideCutter) was developed to connect food proteins with predictive models for antioxidant peptide discovery from sorghum proteins, leading to the identification of a high-antioxidant-activity dipeptide, YR.

Utilizing pLMs, our universal architecture UniDL4BioPep outperformed existing SOTA models in 15 out of 20 peptide datasets, with accuracy improvements ranging from 0.7% to 7%, and demonstrated great potential for other peptide prediction tasks with custom datasets. To enhance usability, a well-annotated template was designed for non-programming users, and an advanced version was introduced for extremely imbalanced datasets. The pLM-based models exhibited SOTA performance in predicting high activity ACE inhibitory peptides with an accuracy of 88.3% and allergenic proteins/peptides with an accuracy of 95.1%, offering solutions to health concerns like hypertension and allergies. The pLMs demonstrated superiority over traditional peptide descriptors in peptide representation and showed strong compatibility with support vector machine, multilayer perceptron, and logistic regression in the ACE inhibitory peptide dataset. A positive correlation between pLM sizes and prediction model performance was observed in the allergenic protein and peptide dataset. All developed prediction models were deployed on user-friendly web servers for scientific exploration.

Beyond the application of pLMs in peptide discovery, a multimodal model, FusionESP, was presented for enzyme-substrate interaction prediction. By employing a novel projection head based on a contrastive learning strategy for knowledge fusion, FusionESP successfully achieved SOTA performance with an accuracy of 94.7% by integrating protein and chemistry language models, while requiring fewer computational resources and data points.

Overall, this work demonstrates the transformative potential of integrating advanced computational techniques—such as machine learning and deep learning—into bioactive peptide research and broader biological domains. The developed models and tools can significantly accelerate bioactive peptide discovery and provide valuable resources for scientific exploration. By making these tools accessible and user-friendly, this work contributes to sustainable agriculture, improved health outcomes, and fosters innovation across food science, biochemistry, and computational biology.

Table of Contents

List of Figures

List of Tables

Acknowledgements

Time goes by extremely fast. I left this section for last while writing my dissertation, though it comes first in order. This journey has been long, filled with curiosity, challenges, support, and love. At times, it felt like everything was moving so quickly that I was not fully prepared to turn the page to this new chapter of my life. But as my manager at Colgate once told me, “You can never be fully prepared for anything. When you reach a point, just go ahead and move forward.” Just like the final scene in the drama *The Long Season* says, “Keep going, don’t look back.”

I would like to express my gratitude to the people and organizations who provided invaluable assistance and support in completing this study:

First and foremost, my deepest thanks go to Dr. Yonghui Li, my major professor, for offering me the opportunity to pursue a Ph.D. and for his unwavering support throughout my research. My research direction is not a mainstream one, and without his encouragement, guidance, patience, and understanding, I would not have been able to explore this interdisciplinary field. His support gave me the chance to pursue the research I truly want to dedicate my life to. As an international student, his help also extended to my daily life and career development, for which I am deeply grateful.

I would also like to thank Dr. Doina Caragea, Dr. Donghai Wang, and Dr. Scott Bean for serving on my supervisory committee. Their constructive feedback, guidance, and support have been instrumental in shaping my Ph.D. journey and career development.

Special thanks to Dr. Hongwei Shen for offering me the summer internship opportunity at Colgate-Palmolive's Global Technology Centers. This experience gave me a valuable glimpse into research roles in industry and introduced me to many inspiring people in that field.

I am grateful to Dr. William Hsu for introducing me to the concept of protein language models.

I would also like to thank Dr. Jeffrey Comer for helping me understand molecular docking, molecular dynamics simulations, and the usage of high-performance computing clusters.

A sincere thank you to my collaborators: Dr. Yanting Shen, Dr. Wenfei Tian, Dr. Xiaolong Guo, Dr. Yixiang Xu, Dr. Arslan Munir, Dr. Changqi Liu, Dr. Shan Hong, Dr. Jikai Zhao, Xingjian Ding, Hyukjin Kwon, and Weimin Fu.

To all my friends, Cereal Chemistry Lab, Grain science faculty, staff, and graduate students, thank you for your help, support, and friendship throughout this journey.

Lastly, I want to thank my beloved family for their endless understanding, patience, and love. A special thanks to my girlfriend, Yuandi Zhang, who crossed mountains and rivers to be by my side. Your support means the world to me.

Preface

The pursuit of sustainable agriculture and value addition to agricultural products and byproducts are fundamental objectives in the field of food science. Bioactive peptides derived from food proteins have emerged as significant contributors to this goal because of their positively biological effects, such as antioxidant activity, antihypertensive activity, etc. However, traditional approaches in bioactive peptide production, characterization, and identification are often laborious, time-consuming, and resource intensive. The integration of bioinformatics and advanced computational techniques into food science presents a transformative approach to overcome these challenges inherent in wet experiments.

This dissertation explores the application of cutting-edge computational methods to accelerate the scientific discovery from bioactive peptide prediction to protein-compound interaction prediction. By leveraging bioinformatic tools, quantitative structure-activity relationship (QSAR) models, machine learning, and deep learning algorithms, this work aims to enhance the efficiency of identifying peptide properties and design novel model architecture for scientific discovery in biological domain.

Chapter 1 lays the groundwork by providing a comprehensive overview of bioinformatics-driven studies on food-derived bioactive peptides. It highlights the roles of databases, proteolysis simulation, machine learning powered-quantitative structure-activity relationship (QSAR) models, molecular docking, and molecular dynamics simulations in advancing FBP research. The chapter underscores how these tools enable food scientists to utilize existing knowledge about identified peptides for novelty checking, bioactivity evaluation, and rational design of protein hydrolysates. This foundation emphasizes the critical role of bioinformatics in transforming food science research and applications.

Building on this foundation, in **Chapter 2** the focus shifts to antioxidant peptides, which are crucial due to their ability to combat oxidative stress linked to various chronic diseases. The chapter addresses the limitations of traditional wet-chemistry methods in peptide screening by introducing QSAR analysis as an efficient and cost-effective alternative. By comparing various feature selection and machine learning methods and employing amino acid indices from AAindex for peptide representation, the study identifies key features that influence antioxidant activity in tripeptides and successfully predict a few antioxidant peptides in experimental validation, highlighting the predictive power and practical applicability of the developed model. The development of high-performance predictive models not only streamlines the screening process but also enhances our understanding of the structural determinants of antioxidant efficacy.

Chapter 3 extends the application of QSAR analysis to dipeptides derived from sorghum kafirin, a type of low value protein. Utilizing a newly developed QSAR model and an improved hydrolysis simulation tool, R-PeptideCutter, the study accelerates the identification of high-activity dipeptides. The predictive models are also validated through synthesis and experimental testing of selected dipeptides, demonstrating how *in silico* approaches can significantly reduce the time and resources required for peptide discovery. This has direct implications for the utilization of agricultural byproducts, promoting sustainability and innovation in food ingredient development.

The advancement of deep learning models in bioactive peptide prediction is presented in **Chapter 4** with the introduction of UniDL4BioPep. This model architecture leverages pre-trained a protein language model (pLM) for peptide representation, facilitating high-performance classification model development for various peptide properties. The model architecture

outperforms existing state-of-the-art (SOTA) models across multiple bioactivity datasets, underscoring the potential of deep learning and pretrained pLMs based transfer learning. A user-friendly webserver, integrating 20 prediction models, was built for the usage of the prediction models to researchers from the wet lab end. UniDL4BioPep represents a crucial tool for researchers aiming to expedite the discovery of novel bioactive peptides.

Considering the need for high-activity peptides and the relatively limited availability of data points in food science, **Chapter 5** uses angiotensin-I converting enzyme (ACE) inhibitory peptides as an exemplar, given the relatively small dataset available and half maximal inhibitory concentration (IC50) of each peptide. As an extension of the previous chapter, we further explore the superiority of pLMs over traditional peptide representation methods, as well as the compatibility of different machine learning classifiers with pLMs. A SOTA model was ultimately obtained for predicting high-activity ACE inhibitory peptides, offering solutions to a prevalent health concern in hypertension.

Addressing the global health concern of allergies, **Chapter 6** presents pLM4Alg, a robust model for predicting allergenic proteins and peptides. Utilizing pre-trained pLMs with convolutional neural networks, the model achieves exceptional predictive performance after systematically evaluating the effects of pLM size on model performance. Notably, pLM4Alg can handle sequences with missing residues and non-standard amino acids, enhancing its utility in real-world applications. This advancement holds promise for improving allergen identification and risk assessment, ultimately contributing to public health safety.

Beyond bioactive peptide discovery and pLM-based transfer learning, **Chapter 7** introduces FusionESP, a multimodal architecture that fuses protein and chemistry language models using a contrastive learning strategy. Aimed at predicting enzyme-substrate pairs,

FusionESP demonstrates SOTA performance while requiring fewer computational resources and less training data. The promising performance highlights the effectiveness of the designed projection head based on contrastive learning for knowledge fusion from protein and chemistry language models, exhibiting great potential in protein-compound interaction prediction and broader biological domains.

Drawing the research together, **Chapter 8** summarizes the key findings and discusses future perspectives. It highlights how the integration of machine learning techniques has significantly enhanced the efficiency of bioactive peptide identification and extended into knowledge fusion for enzyme-substrate interaction prediction under multimodal conditions. Looking forward, Chapter 8 discusses future perspectives in the field, recommending areas for further research such as enhancing explainability and performance with graph neural networks, further exploration of contrastive learning strategies, development of foundation models for peptide discovery, establishment of high-quality databases for one-stop data retrieval, utilizing unlabeled data, and development of multi-label models. It also suggests potential avenues related to machine learning in biochemistry, such as enzyme annotation and regression model development with transformer models.

Collectively, this dissertation demonstrates the transformative potential of integrating bioinformatics, machine learning, and deep learning into food science research. By adopting these advanced computational methods, we can significantly accelerate scientific discovery. As a researcher working in such an interdisciplinary field, it is my hope that the methodologies and insights presented in this work will inspire further collaboration and innovation among food science, computer science, chemical engineering, and related disciplines—ultimately

contributing to sustainable agriculture, improved health outcomes, and the advancement of the field.

Chapter 1 - Bioinformatics approaches to discovering food-derived bioactive peptides: Reviews and perspectives*

Abstract

Food-derived bioactive peptides (FBPs) are gaining interest due to their great potential in agricultural byproduct valorization and high-activity peptide screening. The introduction of bioinformatics into FBP studies further enhances the prospects of this field. This review provides a comprehensive overview and critical insight into the latest advances in bioinformatics-driven FBPs studies. The roles of databases, proteolysis simulation, bioactivity potency evaluation, quantitative structure-activity relationships (QSAR) models, molecular docking, molecular dynamics simulation, and free energy calculation in FBP studies are covered. Furthermore, critical issues related to QSAR model development, molecular docking, and integrated bioinformatics strategies are highlighted. By leveraging these bioinformatics approaches, researchers can fully utilize existing knowledge about identified peptides for checking novelty, evaluating bioactivity potency as well as rational protein hydrolysate design. QSAR models and molecular docking enable efficient screening of thousands of peptide candidates and generate new insights into bioactivity mechanisms. Directions for future research and challenges in current studies are also discussed. The employment of bioinformatics will significantly accelerate the process from the identification of high-potential FBPs to product development, assist in wet chemistry experiment design for targeted protein hydrolysates preparation, and ultimately enhance the long-term development of nutraceutical, pharmaceutical, and cosmeceutical industries.

Keywords:

Food proteins, nutraceutical, database, QSAR, molecular docking, virtual screening

* Du, Z., Comer, J., & Li, Y. (2023). Bioinformatics approaches to discovering food-derived bioactive peptides: Reviews and perspectives. TrAC Trends in Analytical Chemistry, 162, 117051. https://doi.org/10.1016/j.trac.2023.117051

1.1 Introduction

Biologically active peptides (or bioactive peptides) are protein fragments that contain 2–20 amino acid residues joined by peptide bonds and exhibit positive biological effects. Food proteins are a sustainable source for bioactive peptide preparation and are gaining interests from academic researchers and industries (Tu, Cheng, et al., 2018; Udenigwe, 2014). The number of studies on food-derived bioactive peptides (FBPs) in 2022 are expected to quintuple the number ten years ago (Figure 1). The growing interest in FBPs is driven by the increasing demand for natural nutraceuticals, perceptions of the safety of synthetic products, sustainability, and, most importantly, the diverse bioactivities exhibited by FBPs and their potential to relieve the health burdens (FitzGerald et al., 2020; Iwaniak et al., 2019). Common bioactivities of FBPs include antioxidant, antihypertension, anti-diabetes, and anti-inflammation activities (Figure 1). Most FBPs exhibit their bioactivities at the protein level by inhibiting enzymes (e.g., angiotensin-converting enzyme inhibition for antihypertension) or through protein-ligand interactions (e.g., the Keap1–Nrf2 interaction for cytoprotective response regulation), although the bioactivities of a few FBPs might not involve proteins (e.g., capturing free radicals) (Du, Wang, et al., 2022; Y. Feng et al., 2021; L. Li et al., 2017; Nardo et al., 2020; Vukic et al., 2017; F. Wang & Zhou, 2020).

FBPs need to be liberated from protein precursors (animal, plant, edible insect protein) to exert their functions (Figure 1). Enzymatic hydrolysis is the most common method to generate FBPs. This approach requires only mildly controlled production conditions, results in good cleavage specificity and efficiency, and is free of organic solvents and toxic chemicals. Through wet chemistry experiments and in vitro and in vivo characterization, thousands of FBPs have

been identified and characterized in the last few decades, and some of them (e.g., Val-Pro-Pro) have been commercialized (FitzGerald et al., 2020; Gu et al., 2011). However, these conventional approaches suffer from low efficiency and high cost, and they also rely heavily on advanced instruments and trained personnel. Moreover, most mechanisms of protein–peptide interactions can only be elucidated experimentally by X-ray crystallography or nuclear magnetic resonance spectroscopy (NMR), both of which are hindered by sample preparation, and it is impossible to employ such protocols to exhaust all the interaction mechanisms between FBPs and protein receptors (Callaway, 2020).

To circumvent these limitations and accelerate FBP screening and mechanism explanation, some researchers turned to bioinformatics approaches (Agyei et al., 2018; Bo et al., 2021; Iwaniak et al., 2019; L. Li et al., 2017; Tao et al., 2020; Tu, Cheng, et al., 2018; F. Wang & Zhou, 2020). As an in silico method, bioinformatics can fully exploit known FBPs, protein sequences, cleavage sites of enzymes or chemicals, and computational chemistry knowledge to eliminate guesswork and guide experiment design (e.g., enzyme selection or purification condition setting) (Agyei et al., 2018; L. Li et al., 2017; Majumder & Wu, 2010; Y. Xiong et al., 2022). In addition, bioinformatics can be used to screen FBP bioactivities; explore interaction mechanisms; evaluate bioavailability, allergenicity, ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties, and taste-evoking properties (Aguilar-Toalá et al., 2022; Arámburo-Gálvez et al., 2022; Dai et al., 2014; X. Li et al., 2022; Udenigwe, 2014; Vidal-Limon et al., 2022; Y. Xiong et al., 2022; W. Zhao et al., 2023). The field of bioinformatics has made considerable progress, including newly developed algorithms and models, web servers, software, and docking strategies, yet its potential in FBPs discovery has not yet been fully explored.

To our knowledge, there is currently no comprehensive or in-depth review for readers who have expertise in conventional FBPs studies but lack knowledge in bioinformatics to accelerate their studies or who would like to explore more possibilities with advanced bioinformatics tools. To address this gap, this review presents the whole spectrum of bioinformatics-aided procedures from FBP preparation to bioactivity, bioavailability, taste, allergenicity, and ADMET evaluation. Specifically, the application of quantitative structure-activity relationships (QSAR), molecular docking, and molecular dynamics simulation in FBP virtual screening is highlighted, and commonly used and advanced databases, web servers, and software are summarized. Furthermore, suggestions for future research are given to accelerate FBPs studies from bench to market, including the needs in database construction, structure construction of unavailable targeted proteins, trends in QSAR model development and molecular docking & molecular dynamics simulation, utilization of hardware for computation-intensive tasks, ensemble strategies for accuracy improvement, dataset clean & augmentation, and multifunctional peptide screening. In short, this review gathers the latest advances in bioinformatics and FBPs studies to advance the long-term, synergistic developments of both fields and guide the sustainable future production of value-added products from agricultural products.

1.2 Application of databases and proteolytic simulation in FBPs screening and potency evaluation

In the past decades, millions of protein sequences and hundreds of FBPs were identified and characterized by in vitro and in vivo studies, most being composed of the 20 proteinogenic amino acids. These findings contributed to the creation of bioactive peptide databases for efficient retrieval of information. As of now, dozens of bioactive peptide databases have been

built, and each documents one or more bioactivities (e.g., BIOPEP) (Table 1). Users can use these databases to retrieve information (sources of origin, reference articles, bioactivities, etc.) about the peptide of interest using a one-letter sequence or three-letter sequence from the reported peptides (Iwaniak et al., 2020; Minkiewicz et al., 2019). In addition, these databases classify peptides into different categories (e.g., based on their source of origin, bioactivity, etc.), which simplifies the data mining procedure for other bioinformatics studies (e.g., QSAR analysis). It should be noted that no existing bioactive peptide database can be expected to contain all the latest data, so manual double-checking is always recommended (Deng et al., 2019; Uno et al., 2020; Y.-T. Wang et al., 2020). In addition, government-led databases for proteins (i.e., Uniprot (The Universal Protein Resource), NCBI (The National Center for Biotechnology Information), and RCSB PDB (Research Collaboratory for Structural Bioinformatics Protein Data Bank) provide one-stop portals for protein sequence retrieval with all available sequence information (Table 1). In silico proteolysis successfully connects the protein sequence databases and bioactive peptide databases to advance bioactive peptide discovery and bioactivity potency evaluation (Figure 2).

1.2.1 Database-driven approaches

The pure database-driven approach has three steps: 1. parent protein retrieval; 2. in silico proteolysis; 3. identification of bioactive peptides and additional evaluation (Figure 2). This technical route was adopted in most studies and was combined with QSAR or molecular docking methods for unknown FBPs (Amigo et al., 2020; Arámburo-Gálvez et al., 2022; Gomez et al., 2019, 2019; Iwaniak et al., 2020; Ji et al., 2019, 2020; Kartal et al., 2020; Panjaitan et al., 2018; Pooja et al., 2017; Tu, Liu, et al., 2018; Yu et al., 2018). As shown in Table 2, most database-driven FBPs studies are related to the angiotensin-I-converting enzyme (ACE) inhibitory activity

and dipeptidyl peptidase (DPP) IV inhibitory activity since these bioactivities are highly correlated to the most common and urgent chronic diseases (hypertension and diabetes). In addition, database-driven FBP studies tend to limit their data source to BIOPEP database where the number of peptides with ACE and DPP-IV inhibitory activity ranks first and fourth, respectively (Minkiewicz et al., 2019). This limitation in data can be improved by searching in other peptide databases (Table 1). This strategy was adopted in the study of Martini et al., where BIOPEP and MBPDB were combined to search the identified peptides from LC-MS for further selection (Martini et al., 2021).

A few studies adopted a partial database-driven approach which employed only peptide bioactivity databases in their FBP studies (Table 2) (Bechaux et al., 2020; Devita et al., 2021; Martini et al., 2022; Sayd et al., 2018). In the studies of Sayd et al. and Devita et al., liquid chromatography-mass spectroscopy (LC-MS) was used to identify the FBPs released from the protein precursor, and the BIOPEP database was employed to search the LC-MS identified FBPs (Sayd et al., 2018). Proteolytic simulation was used to guide enzyme selection or enzyme combination for protein substrates in order to maximize the value of desired functional food hydrolysis (Bechaux et al., 2020; Y. Fu et al., 2016; Luo et al., 2022). There are disadvantages to these studies. Usually, web servers can work only on one type of bioactivity for one protein subject to cleavage by one enzyme at a time, which makes it difficult to conduct large-scale analyses (Bechaux et al., 2020). Employing some automation of data mining tools (e.g., Selenium WebDriver) could significantly reduce repeated work. Besides, the order of enzyme addition for hydrolysis is not taken into consideration by most of the available proteolysis simulation tools (e.g., PeptideCutter, BIOPEP-UWM, etc.). Our lab recently developed and published an improved version (R-PeptideCutter) overcoming this defect. In addition, this

improved tool allows researchers to run large-scale simulation and generate results in an easily readable format with Python scripting (Du & Li, 2022a).

Additionally, QSAR web servers are widely employed in database-driven approaches to supplement the evaluation of bioactivity potential, physicochemical properties, allergenicity, toxicity, and ADMET. (Arámburo-Gálvez et al., 2022; Baskaran et al., 2021; Iram et al., 2022; Ji et al., 2020; Yu et al., 2018). However, most web servers might not update their built-in models with the latest reported data.

Overall, it is highly recommended that researchers employ two or more peptide databases and prediction tools built and developed by different research groups using different strategies to validate in silico results and provide more robust predictions. The most advanced and user-friendly web servers that can be used in database-driven studies are summarized in Table 3.

1.2.2 Bioactivity potency evaluation

Bioactivity potency evaluation of parent proteins is a common step after database searching (Table 2). A significant number of FBP studies aim to valorize byproducts from the food industry, and the most practical FBP production strategy using byproducts is to produce FBPs as a hydrolysis mixture instead of pure FBPs, because the latter incurs high purification cost (Du & Li, 2022b; Ji et al., 2019; Udenigwe, 2014). Some indicators—e.g., A (the occurrence frequency of peptides in a protein), AE (the occurrence frequency of released peptides in a protein under a specific enzyme), DHt (theoretical degree of hydrolysis)—were proposed to evaluate the potential bioactivity of protein hydrolysis (Table 2). The distinction between indicators A and AE as well as B (the potential bioactivity of a protein) and BE (the potential bioactivity of released peptides in a protein under a specific enzyme) should also be heeded. For example, A is based on the occurrence frequency of the FBPs in the protein

sequence regardless of the availability of cleavage needs from available enzymes. For some potential FBP sequences, there might be no enzymes capable of releasing them in practice (Agirbasli & Cavas, 2017; Coscueta et al., 2022; Iram et al., 2022; K. Lin et al., 2018; Panjaitan et al., 2018; Pooja et al., 2017). Modified versions of indicators (e.g., AE and BE) were proposed for such situations (Arámburo-Gálvez et al., 2022; Ji et al., 2020; Kartal et al., 2020). BE is valuable because it can be directly used for decision-making on the feasibility of specific protein substrates for bioactive hydrolysate production under optimal enzymes, and it takes into consideration cost, efficiency, and the expected positive effects on health. An interesting study of large-scale protein bioactivity potency evaluation based on BE was conducted for prediction of 40 pigeon proteins hydrolyzed by pepsin, papain, and thermolysin, and the theoretically generated peptides were predicted by modeling eight parameters using the random forest method; the eight parameters are frequencies of the six amino acid residues (A, P, V, G, L, F), hydrophobicity values, and AE (Boachie et al., 2019). Using only these eight parameters might have simplified the model development procedures, but also undermined the prediction power of the model (Boachie et al., 2019).

With reasonable indicators and proteolytic simulation, the last prerequisite for comprehensive bioactivity potency evaluation is an all-inclusive database that can be used to search for the bioactivity of the theoretical peptide. However, purifying or synthesizing all the theoretically possible peptides (i.e., 400 dipeptides, 8,000 tripeptides, and 160,000 tetrapeptides, etc.) is not possible in reality, let alone the evaluation of multiple possible distinct bioactivities. Therefore, there will always be FBPs with unknown bioactivity, which will undermine potency evaluation accuracy (Bo et al., 2021; Boachie et al., 2019; Zhu et al., 2022). The alternative

approach is to combine reported data and predictive data from QSAR models (see Section 3) for potency evaluation. Such attempts are expected to be used in in silico studies in the future.

1.2.3 Challenges in proteolysis simulation

In silico prediction was mainly employed as a qualitative tool for hypothesis proposal and experimental validation, without quantitative validation of predictions for peptides released by degradation of proteins (Amigo et al., 2020; Bechaux et al., 2020; Coscueta et al., 2022; Y. Fu et al., 2016; Gomez et al., 2019; Martini et al., 2021, 2022; Panjaitan et al., 2018; Pearman et al., 2020; Yu et al., 2018). With the help of QSAR models, database-driven approaches have the potential to become a quantitative tool for predicting the bioactivity of peptides released from proteins; however, there has not yet been a systematic quantitative validation of this approach. A big challenge is that the gap between peptides theoretically predicted using proteolysis simulation and those identified by wet chemistry needs to be bridged in order to advance the quantitative evaluation of protein bioactivities.

Some experiments have shown discordance between wet chemistry identification and in silico proteolysis (Chatterjee et al., 2015; Majumder & Wu, 2010; Nongonierma et al., 2018). A study comparing in silico and in vitro analysis conducted by Chatterjee et al. showed deviations not only between in vitro identified peptides and in silico predicted peptides but also between two different in silico strategies (Chatterjee et al., 2015). However, the details of this study limit generalization. The experimental sample used for the in vitro analysis, whey protein concentrate, was a mixture of various proteins, while the in silico proteolytic simulation considered only the two proteins α-lactalbumin and β-lactoglobulin, accounting for some of the disagreement (Chatterjee et al., 2015). Considering that purified forms of these two proteins are commercially available, further experiments could more directly compare the in silico and in vitro methods,

allowing for a better understanding of the origin of the disagreement and checking the limitations of current proteolytic simulation methods for predicting experimental results.

There are two major challenges for predicting the peptides released from a given protein source. First, protein naturally occurring in foods or agricultural by-products is invariably of a complex composition. This challenge could be addressed by obtaining better quantitative data on the protein composition of the protein source. Quantitative characterization of protein composition could be obtained from size exclusion chromatography (SEC) or sodium dodecyl-sulfate polyacrylamide gel electrophoresis (SDS-PAGE) (Shen et al., 2022; Shen & Li, 2021). In proteolytic simulation, this data could be used to assign a weight to each of the predominant protein sequences present in a source material.

The second challenge is from limitations of the in vitro experiments. Other components in the protein substrates (e.g., lipids and carbohydrates), environmental conditions, enzyme quality, and each protein's secondary, tertiary and quaternary structures contributed to inconsistency in different in vitro experiments (Y. Fu et al., 2016; Nongonierma & FitzGerald, 2018). Some attempts have been made to optimize in vitro experiments such as using microwave pre-treatment of protein substrates, heating, high hydrostatic pressure, pulsed electric fields, and ultrasound to unfold protein structure and improve protease accessibility during hydrolysis, and response surface methodology to select the hydrolysis conditions (Du & Li, 2022b; Nongonierma & FitzGerald, 2018). After hydrolysis, the smaller molecular weight fractions generally exhibited higher bioactivity, and the physicochemical properties of these peptides were the main factor for fractionation methods design. Some physicochemical property prediction web servers have been developed (Table 1), which can be used to guide extraction and determination protocols in wet chemistry experiments.

Such efforts have contributed to narrowing the gaps between in silico and in vitro experiments. Future studies are expected to integrate this progress, leading to more robust and reliable proteolytic simulation for complex protein sources. On the other hand, it should be noticed that bioinformatics software such as PEAKS X or Mascot, which is used for mass spectrometry (MS) data deconvolution and interpretation, is limited by predefined settings (e.g., parent protein databases and enzyme specificity) and also has difficulty in identifying small peptides, which are the majority of the peptides in the databases (Bechaux et al., 2020; Garzón et al., 2022; Nongonierma et al., 2018; Pearman et al., 2020; M. Zhang et al., 2022).

1.3 QSAR approaches and applications in FBP virtual screening

QSAR is a computational modeling method to interlink the structural characteristics and biological activity of a substance. It is the most popular bioinformatics approach in FBP virtual screening. QSAR models can be categorized as classification models or regression models based on their modeling methods (Table 3 & 4). A regression model can predict specific bioactivity values (e.g., IC50), while a classification model can only predict relative bioactivity among a group of samples (e.g., bioactivity potential) or binary classification (e.g., toxicity) (Agyei et al., 2018; Baskaran et al., 2021; F. Wang & Zhou, 2020; Zhou et al., 2021). The prerequisites for QSAR model development is a set of bioactive peptides with known bioactivity, so QSAR approaches are also classified as a ligand-based virtual screening method (Meng et al., 2011). There are three basic steps in QSAR modeling: collection of peptide bioactivity data; peptide representation by different descriptors; and model development (Figure 2).

1.3.1 Datasets in QSAR-driven virtual screening

As a knowledge-based method, QSAR modeling relies highly on datasets (Zhou et al., 2021). Therefore, dataset collection is a common issue that hinders QSAR model performance,

especially for researchers without a biochemistry background (Dai et al., 2014; Mahmoodi-Reihani et al., 2020; Zhou et al., 2021). At this time, there is no well-recognized and curated dataset for QSAR model development (Table 1 summarizes the databases used for FBP data retrieval). Some researchers directly retrieve the datasets used in their previous papers without any updates with the latest FBPs (Dai et al., 2014; Mahmoodi-Reihani et al., 2020; Zhou et al., 2021). For example, in the study of Zhou et al., the peptides in the 8 datasets were published in 1986–2008 (Zhou et al., 2021). Manually collecting FBP data from literature is laborious but effective for dataset enlargement and model performance improvement (Deng et al., 2019; Du, Wang, et al., 2022). Some efforts have been made to develop FBP databases. BIOPEP has an option for authors to upload their latest bioactivity data with published references; it then manually checks the data (Minkiewicz et al., 2019). However, there is still a long way to go.

In addition, the size of current FBP datasets is quite small. All the datasets used in the studies in Table 4 had fewer than 250 entries, although the dataset size slowly increased (N. Chen et al., 2018; Deng et al., 2019; Du, Wang, et al., 2022; Uno et al., 2020). In order to gain more peptides for QSAR model development, some researchers have synthesized many pure peptides using chemical approaches, which enlarged the FBP databases and contributed to QSAR model performance improvement (N. Chen et al., 2018; Deng et al., 2019; Uno et al., 2020; Zheng et al., 2016).

1.3.2 Peptide representation in QSAR-driven virtual screening

Peptide representation is an essential step in QSAR model development. Basically, there is a need to numerically represent peptide characteristics by different molecular descriptors. Peptide representation can be classified into different types based on the type of descriptors used (e.g., chemical descriptors), descriptor properties (e.g., 1D-, 2D-, and 3D-QSAR), or structure

separation (i.e., local descriptors and global descriptors) (Figure 3). Below, we further discuss peptide representation, considering categorization by structure separation because it can easily differentiate among FBP studies (Table 4), as well as the latest development in natural language processing (NLP)-based peptide representation (Bo et al., 2021; Patel et al., 2014).

1.3.2.1 Local descriptor-based peptide representation

In peptide representation, local descriptors are also known as amino acid descriptors (AADs) where peptide residues are separately characterized by their own AADs and assembled depending on the peptide sequence (N. Chen et al., 2018). For example, 5-z scale is a kind of AAD with 5 parameters for amino acids, and then a peptide with n residues will be represented as a vector of 5n components where the first 5 elements correspond to the 5 parameters of the first amino acid residue in the 5-z scale (Figure 2). This approach is the most popular peptide representation approach among FBP studies (N. Chen et al., 2018; Deng et al., 2017, 2019; Y. Fu et al., 2016; Nongonierma & FitzGerald, 2016; Qian et al., 2017; Xu & Chung, 2019). AADs are derived from the basic properties of amino acids, including physicochemical properties, topological properties, 3D structural information and others, using feature extraction methods such as principal component analysis (PCA) (Zhou et al., 2021). To date, there are up to 80 proposed AADs, and most of them were extracted by PCA for different properties or a mixture of different properties (Zhou et al., 2021). For instance, the 5-z scale is a physicochemical descriptor extracted from 26 original physicochemical properties, and the factor analysis scale of generalized amino acid information (FASGAI) is a mixture descriptor extracted from 335 properties including hydrophobicity, alpha and turn propensities, bulk properties, compositional characteristics, local flexibility, and electronic properties (Y. Fu et al., 2016; Qian et al., 2017; Zheng et al., 2016). Combining more AADs can provide more information and better represent

peptides. In light of this, the bootstrapping soft shrinkage (BOSS) method was employed to integrate 16 AADs for antioxidant, ACE inhibitory, and bitter peptide representation, and the performance of all QSAR models with integrated descriptors achieved better performance (Deng et al., 2017, 2019; Xu & Chung, 2019).

Meanwhile, the availability of the original properties for AAD extraction has been gradually improved with the development of amino acid property databases (e.g., AAIndex). Therefore, some researchers turned to extracting AADs from the original dataset (N. Chen et al., 2018; Du, Wang, et al., 2022; Mahmoodi-Reihani et al., 2020). For example, our lab conducted a comprehensive study (Du, Wang, et al., 2022) where 566 amino acid properties including different physicochemical and biological properties were screened by 6 feature selection methods and 14 machine learning methods for regression model development. The best tripeptide antioxidant activity regression model was obtained by a combination of new AADs extracted from original properties of amino acids by random forest regression and an eXtreme gradient boosting (XGBoost) regression model. This model performed better than the antioxidant activity prediction model that relies only on physiochemical properties and linear regression methods (N. Chen et al., 2018; Du, Wang, et al., 2022). The 566 amino acid properties and the 16 popular AADs compared in the study of Du, Wang et al., and Deng et al., are included in Supplementary 1 from the previous study (Du, Wang, et al., 2022).

The AAD-based QSAR approach does have an obvious disadvantage. Since descriptors are assigned to each residue, peptides with different lengths will have different feature dimensions after characterization, which results in dataset prerequisites (i.e., peptides with the same length). This limits the utilization of the latest data reported for peptides with different lengths and hinders building a universal QSAR model for a specific bioactivity (Y. Fu et al.,

2016; Guan & Liu, 2019; Xu & Chung, 2019; Zheng et al., 2016). Some researchers proposed modified AAD-based QSAR approaches that only consider residues close to the terminus because some studies showed that these residues had greater influence on bioactivity, and thus peptides with different lengths can be unified to the same feature dimension (Nongonierma & FitzGerald, 2016; Zhu et al., 2022). Peptides with different lengths could then be combined as a larger dataset for model development, although some peptides would suffer from property information loss. The other disadvantage is that the characterization does not consider the peptide as a whole and so does not consider the synergetic effects between different residues.

The information loss mentioned above in AAD-based QSAR modeling was shown in the study of Zhou et al, where support vector machine regression (SVMR) and partial least squares regression (PLSR) were used to build QSAR models based on 80 AADs from 8 different bioactivity datasets (Zhou et al., 2021). Specifically, two random descriptors generated by a standard normal distribution and a uniform distribution were also used in model development and did not result in a significantly worse performance (Zhou et al., 2021), which meant the prediction power was mainly derived from the modeling method. Even though the authors stated that their AAD-based modeling had almost reached theoretical limits, it should be noted that the datasets used in the study were very small and out-of-date, and accordingly, the study might not reflect the power of AADs very well.

There is an alternative way to overcome the peptide length limitation inspired by one-hot encoding and such attempts have been used to build a cutting-edge bioactivity prediction web server, AnOxPePred (Table 3). Since most reported antioxidant peptides in available antioxidant peptide dataset are composed of the 20 proteinogenic amino acids, a vector with 20 elements (nineteen elements are 0 and one element is 1) was used to represent any amino acids by the

position of 1 in the vector. A 20×30 matrix was created for peptide representation (ranging from 2 to 30 residues). A dipeptide was represented by a 20×30 matrix, but only the first and the second row were used. With a large dataset (1404 peptides) and the boost of a convolutional neural network (CNN) for classification model development, this peptide representation approach achieved great performance (accuracy around 80%) (Olsen et al., 2020). It should be noted that this approach is only feasible when a large dataset is available, and it is not suitable for regression model development.

1.3.2.2 Global descriptor-based peptide representation

Global descriptors characterize peptides as a whole, which can overcome the peptide length limitation mentioned above and also take the whole peptide structure into consideration for representative vector generation (Bo et al., 2021). Theoretically, global descriptors have more advantages than local descriptors.

Molecular descriptors for general chemicals might also be a great alternative for AADs. Descriptor calculation software for chemicals has been well developed (e.g., PaDEL, Dragon, MOE, etc.) (X. Li et al., 2022; Y.-T. Wang et al., 2020; T. Zhang et al., 2015; X. Zhao et al., 2011). In the study of Li et al., the QSAR model achieved amazing accuracy for IC50 (the half maximal inhibitory concentration) of DPP IV inhibitory activity among 39 cysteine-containing dipeptides by using a combination of PaDEL for descriptor calculation, a genetic algorithm for feature selection, and multiple linear regression for model development (X. Li et al., 2022). A similar attempt was also seen in the study of Wang et al., where 728 peptides with different lengths were collected (Y.-T. Wang et al., 2020).

Among FBP studies, three-dimensional (3D) QSAR is the most popular approach with global descriptor-based characterization because it reasonably represents 3D characteristics (Bo

et al., 2021; Qian et al., 2017; X. Zhao et al., 2011). In FBP studies, 3D QSAR mostly refers to Comparative Molecular Moment Analysis (CoMFA) and Comparative Molecular Similarity Indices Analysis (CoMSIA), which have some additional steps compared to AAD-based QSAR model development (Figure 2). There are also descriptors for 3D QSAR without the need for molecular superposition (e.g., CoMFA), but here we specifically focus on CoMFA and CoMSIA analysis. The first step of these analyses is to reconstruct the 3D structure of the peptides and then conduct energy minimization by calculation under force fields. Some software and web servers for structure construction are provided in Table 1. Energy minimization is essential because it makes structure conformation much more reasonable and representative and is generally achieved by search strategies: the steepest gradient descent method (fast) and conjugate gradient method (slow) (Vukic et al., 2017; F. Wang & Zhou, 2020; S. Wu et al., 2014). The next step, structural alignment, is the most critical step in developing a reliable 3D-QSAR model (F. Wang & Zhou, 2020). There are, generally speaking, three different alignment strategies: template ligand-based alignment, docking-based (receptor-based) alignment, and scaffold-based alignment (Table 4). The template ligand-based alignment aligns all the peptides in the dataset to the most potent peptides. The docking-based alignment retrieves peptide conformation from molecular docking results and then aligns peptides to the most potent peptides with the consideration of orientation, while scaffold-based alignment does not consider the orientation (Pissurlenkar et al., 2007; F. Wang & Zhou, 2020; X. Zhao et al., 2011). The last step is to calculate the molecular interaction field based on the 3D structure. Based on results from FBP studies (Table 4), it is difficult to judge which method is better: CoMFA analysis or CoMSIA analysis. Some researchers believe CoMSIA was often more robust because it contains

additional information from hydrogen bonding groups and hydrophobic regions (Bo et al., 2021; X. Zhao et al., 2011).

CoMFA and CoMSIA analyses have all the advantages of global descriptors over local descriptors (Bo et al., 2021; Vukic et al., 2017). Furthermore, 3D-QSAR can be easily combined with molecular docking to explain the peptide ligand-receptor interaction mechanism, because the descriptors (e.g., steric, electrostatic, hydrophobic, hydrogen bond donor, and hydrogen bond acceptor fields in CoMSIA) are the interaction forces in molecular docking and can be visualized as a contour map for the contribution distribution of different fields (F. Wang & Zhou, 2020). However, these approaches also face some challenges. First, peptide molecules are highly flexible, and energy minimization usually only samples a single energy minimum near the initial structure. If there are multiple biologically relevant structures, some may be missed. In addition, compared to AADs, global descriptors cannot reflect how an amino acid at a certain site affects bioactivity, so they cannot directly aid rational design at the level of amino acids (Bo et al., 2021; Du, Wang, et al., 2022; X. Zhao et al., 2011).

1.3.2.3 Natural language processing (NLP)-based peptide representation

Natural language processing (NLP) has been applied for encoding peptides by building pre-trained models for feature extraction from peptide sequences. One attempt (BERT4Bitter in Table 1 & 3) was conducted in 2021 to update the state-of-the-art classification model for bitter peptide prediction, in which three NLP-based peptide encoding methods were employed for peptide representation (Charoenkwan et al., 2021). A cutting-edge modeling architecture (bidirectional encoder representations from transformers (BERT)) is emerging recently for protein sequence encoding, which can transform the words in a sentence into a numerical vector with the consideration of the sentence context and has exhibited great performance in various

NLP tasks. Correspondingly, each amino acid residue in a protein sequence (sentence) can be considered as a word in the "sentence" (Brandes et al., 2022; Devlin et al., 2019). Such models are based on the comprehension of the available protein sequences through building self-supervised learning models. This approach relies on more than 200 million protein sequences and theoretically has a better capacity to internalize the information encrypted in protein sequences (Brandes et al., 2022; Elnaggar et al., 2022). Several BERT-based protein language models (pLMs) (e.g., ESM- 2, ProteinBERT, ProtTrans) have been released in the past two years and achieved great performance in downstream classification tasks, such as subcellular location, structure prediction, function prediction, etc. (Brandes et al., 2022; Du et al., 2023; Elnaggar et al., 2022; Lin et al., 2023). Very recently, our lab has developed a universal model architecture (UniDL4BioPep) by employing the latest protein language model ESM-2 with a CNN model. It was tested on 20 bioactivity dataset prediction tasks and achieved better performance than the respective state-of-the-art models in 15 out of 20 datasets (Du et al., 2023). Meanwhile, the model development for the bioactive peptide datasets based UniDL4BioPep does not require feature selection or hyperparameter tuning. This successful attempt supports the feasibility and superiority of pLM-based peptide representation. The great potential of pLMs in peptide representation for model performance improvement is expected to gain more attention in the future, though the model method lacks explainability.

Besides, recent progress in drug discovery might also help advance global descriptor-based characterization. Chemicals are often encoded by the simplified molecular input line entry system (SMILES), which is a string of letters, numbers, and symbols. These strings can in turn be subjected to NLP, which can categorize strings and build models to screen other chemicals (Irwin et al., 2022). The SMILES format provides sufficient information to construct the

molecules in atomic detail, and, therefore, it can be thought to represent the global structural information of the chemical. The latest transformer, Chemformer released in 2022 was pretrained through 100 million SMILES strings and can be used to encode the SMILES string of the peptide sequences. Unlike typical drug-like molecules, which are often smaller and contain mostly rigid groups, peptides with more than a few amino acids are relatively flexible and can have many accessible conformations. Thus, its practical application in peptides may be limited (Irwin et al., 2022). More information may need to be encoded in descriptors or strings to describe peptide conformational ensembles, especially for long peptides that can form secondary structures (> 4 residues) (Ciemny et al., 2018; Forli et al., 2016; Irwin et al., 2022; Y. Li et al., 2022).

1.3.3 Model development

Model development is the process to connect representative vectors of peptides and their bioactivity/properties. Recent progress in machine learning has benefited from numerous biochemistry studies that have increased model robustness and accuracy (Bo et al., 2021; Du et al., 2020; Du, Tian, et al., 2022; Du, Wang, et al., 2022). Generally speaking, model development includes three steps: feature/variable selection, model building, and performance evaluation. Feature/variable selection as well as model building and performance evaluation have no specific prerequisites. A previous review from our lab reviewed in detail popular feature selection methods and model building and performance evaluation methods to compare their advantages and disadvantages (Du, Tian, et al., 2022). The review here only focuses on the models used in Table 4 (recent QSAR studies on FBPs) and Table 3 (web servers).

All the studies in Table 4 created a regression model, and most of them adopted AADs or global descriptors plus PLSR for model building. Because AADs have gone through a round of feature selection from original properties, it is not necessary to conduct one more feature

selection before model building (Y. Fu et al., 2016; Nongonierma & FitzGerald, 2016; Zhou et al., 2021). In some cases, researchers integrated different AADs to have as much information as possible for model building. In those cases, feature selection would be necessary since there would be a lot of redundant features among different descriptors, which will undermine the model (Deng et al., 2017, 2019; Xu & Chung, 2019). For researchers who characterized peptides from the individual properties of amino acids, feature selection would be indispensable to reduce feature dimension and remove redundant features (N. Chen et al., 2018; Du, Wang, et al., 2022). For global descriptors, the criteria for whether to adopt feature selection were the same for AAD descriptors.

As for model building methods, traditional machine learning methods are the mainstream because they work well with limited dataset sizes compared to emerging deep learning approaches. Linear regression methods (e.g., PLSR) can calculate feature importance by variable importance in projection (VIP) values and illustrate which peptide residues or whole peptide properties are more influential (Du, Tian, et al., 2022; Sun et al., 2017; Vukic et al., 2017; Y.-T. Wang et al., 2020). Non-linear regression methods generally perform better than linear regression, but some of them require complex model processes, and sometimes it is difficult to figure out the contribution of each feature to the predicted bioactivity because of their poor explainability (e.g., artificial neural network) (Du, Tian, et al., 2022). The regression models in Table 4 can give the specific bioactivity values of a peptide, which equips the model with more potential for precise screening. In terms of model performance parameters in antioxidant activity prediction, an improved regression model among available QSAR regression models was achieved by our lab, where the RP2 was 0.847 using a non-linear random forest regression (RFR) model (Du, Wang, et al., 2022).

Table 3 collects the latest QSAR web servers developed by bioinformatics researchers. Their feature selection and modeling methods are diverse, and some of them adopt the latest machine learning methods (e.g., CNN) (J. Chen et al., 2021; Olsen et al., 2020; Thi Phan et al., 2022). Unlike the regression models in Table 4, these QSAR web servers (except AHTpin for dipeptides and tripeptides) are classification models, whose responses (Y) in the model development are labels (e.g., active or not active) instead of the specific activity (e.g., IC50 = 1 μM). Such models can only predict which group an unknown peptide belongs to (e.g., high activity (IC50 < 10 μM), medium activity (10 μM < IC50 <100 μM), low activity (100 μM < IC50 < 1000 μM), or non-activity (IC50 >1000 μM) and, therefore, are much more suitable for rough screening. In addition, these classification models are all based on larger datasets by unifying the feature dimensions of different peptide lengths or by creating new global descriptors (Mooney et al., 2012; Olsen et al., 2020). For example, PeptideRanker is the most popular binary classification web server among FBP studies. Developed in 2012, it used 1330 short peptides (4–20 amino acids) and 4731 long peptides (>20 amino acids) for its neural network classification model. Its method of peptide representation was not clearly explained. One thing that should be mentioned is that PeptideRanker's dataset for model development was composed of bioactive peptides retrieved from peptide databases (i.e., BIOPEP, PeptideDB, APD2, and CAMP), including peptide hormones, antimicrobial peptides, toxin/venom peptides, antifreeze proteins, antibacterial peptides, antibiotic peptides, and anticancer peptides (Mooney et al., 2012). It is obvious that the number of peptides with each type of bioactivity was not equally distributed in PeptideRanker's original dataset, which caused bias when it was used to predict the potential for bioactivity. Several web servers have been recently developed to predict specific types of bioactivity and, therefore, are likely to be more accurate when a specific type of bioactivity is of

interest (Table 3). Among recent FBP studies, only two studies (L. Wen et al. and Sansi et al.) employed the latest immunomodulatory activity prediction sever NetMHCpan 4.0 (released in 2020) and antimicrobial bioactivity prediction server CAMPR3 (released in 2014) for virtual screening (Sansi et al., 2022; Waghu et al., 2014; L. Wen et al., 2021). This is expected to change in the future as the collaboration between biochemistry and bioinformatics studies increases.

1.3.4 Current status of QSAR application in FBP virtual screening

When employed in practical wet chemistry studies to screen potentially highly bioactive peptides, QSAR models incur a lower cost than traditional methods. The latest QSAR applications in FBP studies are summarized in Table 4, and advanced QSAR web servers are reviewed in Table 3.

There is usually a time lag between the release of advanced QSAR models in the bioinformatics field and their wide application to biochemistry studies (Coscueta et al., 2022; Mooney et al., 2012; Pearman et al., 2020). As shown in Table 1, many database-driven FBP studies adopted the PeptideRanker tool to evaluate the possibility of bioactivity for previously unreported peptides generated by in silico proteolytic simulation. Peptides predicted to have a high probability of bioactivity were then synthesized and tested by in vitro or in vivo experiments. Compared to blind experimental searches, these approaches can accelerate FBP exploration, but the nonspecificity for general bioactivity prediction (e.g., PeptideRanker in Section 3.3) also to some extent hinders their efficiency (Amigo et al., 2020; Barati et al., 2020; Coscueta et al., 2022; Garg et al., 2018; Iwaniak et al., 2020; Pearman et al., 2020; Pooja et al., 2017; Sansi et al., 2022; Tu, Liu, et al., 2018).

Some researchers with food science backgrounds have participated in QSAR model development and applied the developed models in FBP screening studies. For example, a multi-bioactive tripeptide, IRW, was found to have antihypertensive, anti-inflammatory, and antioxidant properties in both in vitro and in vivo studies (Liao et al., 2018, 2019; Majumder et al., 2013). The initial finding of this valuable tripeptide was based on the AAD-based ACE inhibitory activity prediction QSAR model for tripeptides developed in 2006 (J. Wu et al., 2006). Four years later, the model was further applied to ACE inhibitory prediction of 76 theoretical peptides released from egg proteins, and IRW was successfully screened because of its lowest IC50 value (Majumder & Wu, 2010; J. Wu et al., 2006). Another example involves an AAD-based model for predicting DPP IV inhibitory activity developed in 2016 by Nongonierma & FitzGeral. It is used for FBPs liberated from milk protein under the hydrolysis of enzymes in the intestinal tract. With the additional in vitro validation for the synthesized peptides, some high-activity DPP IV inhibitory FBPs (e.g., IPM and LPVPQ) were identified (Nongonierma & FitzGerald, 2016). In 2018, a study based on the model was conducted to predict the theoretically generated peptides from camel milk proteins, and the predicted high-activity peptides were synthesized to validate the results of LC-MS/MS from the experimental protein hydrolysates (Nongonierma et al., 2018). Both models successfully obtained some valuable FBPs using their QSAR models, but it should be noted that both models used the determination of the coefficient in cross-validation of the final performance evaluation, which strictly speaking is not objective in model development (X. Li et al., 2022; K. Lin et al., 2017; Nongonierma & FitzGerald, 2016; Sun et al., 2017; Vukic et al., 2017; J. Wu et al., 2006; S. Wu et al., 2014; Xu & Chung, 2019). Most biochemistry researchers tend to use wet chemistry experiments as an additional validation approach, which to some extent are a better way to demonstrate a model's value in predicting

high activity peptides. However, synthesizing peptides with a large range of bioactivity (as predicted by the QSAR model) instead of only those peptides predicted to have high activity can be a better way to demonstrate the overall performance of the model (N. Chen et al., 2018; Du, Wang, et al., 2022). There is no best solution given the cost of synthesis, so the best way is to choose peptide synthesis based on the practical needs for high activity FBPs or FBPs with a specific bioactivity.

For linear and tree-based regression methods (e.g., PLSR, multiple linear regression (MLR), random forest regression, and XGBoost), the feature importance used in peptide representation for model development is available and can be used to reveal deeper structure-activity relationships (N. Chen et al., 2018; Du, Wang, et al., 2022). In the study of Fu et al., AADs (5z-scale) were used to build PLSR models for ACE inhibitory activity, and the results revealed that the hydrophobicity and side chain bulk of residues at the C-terminus contributed more to the bioactivity for dipeptides, while for tripeptides, the hydrophobicity and electronic properties of residues at C-terminus were most important (Y. Fu et al., 2016). Another study selected features for peptide representation from 195 physicochemical properties, which resulted in a better explainability (N. Chen et al., 2018). In the 3D QSAR study of Vukic et al., the favorable effects of steric interactions and electronegativity at the C-terminus in ACE inhibitory activity were highlighted by the CoMFA method (Vukic et al., 2017). Sometimes, it is difficult to achieve such great explainability when the features used for peptide representation are complex intrinsically. In our lab, 566 physicochemical properties and biochemical properties were included for feature selection, and the features retained in the final list for model development (such as “optimized propensity to form a reverse turn”) were difficult to understand intuitively and do not provide an obvious mechanism for the bioactivity (Du, Wang, et al., 2022).

The same dilemma also exists in AAD-based characterization and global descriptor-based QSAR modeling. For example, the 5 features in the T-scale were extracted from 67 structural and topological variables by PCA, but no information was provided to explain each feature. For the QSAR model developed based on the T-scale, it would be very difficult to track back which properties of the peptides contribute to the bioactivity (F. Tian et al., 2007).

1.3.5 Application of QSAR models in evaluating other properties of FBPs

There is a long way from FBPs to commercial products (e.g., functional foods, nutraceuticals, or pharmaceuticals). It is not wise to synthesize all the FBPs predicted to have high activity for in vitro or in vivo experiments after a single round of in silico rough and refined screening. This is because some of the FBPs might not meet the basic requirements for commercial products (Liang et al., 2021). The most common strategy is to conduct a second round of filtration for all the theoretical FBPs based on commercial feasibility (e.g., easy to synthesize, allergenicity, toxicity) before wet chemistry validation (X. Li et al., 2022; Liang et al., 2021). Then, the final candidates would be limited to peptides that have a high potential for both the desired bioactivity and economical commercial production. Important properties for commercial product development based on FBPs include bioaccessibility, bioavailability, metabolism, toxicity, allergenicity, excretion, bitterness, etc. (Iwaniak et al., 2020; Karaś, 2019; L. Li et al., 2017; Udenigwe, 2014; Xiao et al., 2022).

Evaluation of some of these properties has been achieved by QSAR models using large datasets and has been established as user-friendly web servers (Daina et al., 2017; G. Xiong et al., 2021). In Table 3, the latest web servers for the final round of screening, with their model development methods and model performance in the benchmark dataset, are summarized. These servers examine taste, cell penetration, plasma stability, metabolism prediction, ADMET,

allergenicity, toxicity, and drug-likeness (Xiao et al., 2022). It should be noted that some of these models were first developed for general chemicals but can be used for peptides since peptides are chemicals in the broad sense. Among these web servers, ADMET evaluation, which is an integrated one-stop evaluation for the absorption, distribution, metabolism, excretion, and toxicity properties of peptides, is the most popular in FBP screening (Baskaran et al., 2021; N. Chen et al., 2018; Iwaniak et al., 2020; Kartal et al., 2020; X. Li et al., 2022; Liang et al., 2021; Pooja et al., 2017). Among these web servers, SwissADME was released in 2017 and evaluates absorption, distribution, metabolism, and excretion properties, which were the top priorities in FBP studies (Daina et al., 2017). A newly developed tool, ADMETlab 2.0, was released in 2021, which integrated prediction of more properties. However, given its relative newness, so far few FBP studies have adopted it (G. Xiong et al., 2021). The same situation is also observed among other latest prediction servers (e.g., those on plasma stability and bitterness).

Given the current state of FBP studies, it is still highly recommended to include external evaluation in FBP screening to validate accuracy. This additional work will also allow researchers who have expertise in other bioactivity studies to use the data for further experiments.

1.4. Molecular docking approaches and their application in FBP virtual screening

Molecular docking is a structure-based method for virtual screening where the structure information of both peptide ligands and targets are needed. This method overcomes the limitation of bioactivity value availability in QSAR approaches (Ferreira et al., 2015; Meng et al., 2011). It aims to find the best matching binding mode between a ligand and a targeted receptor using conformational sampling and a binding affinity scoring function (Salmaso & Moro, 2018; Tao et al., 2020). The three-dimensional structures of proteins and other

biomacromolecules have been determined by X-ray crystallography, NMR, and increasingly by cryo-electron microscopy. Nearly 200 000 such structures are freely available and include pharmaceutical targets related to various common health issues (e.g., hypertension, diabetes, aging, etc.) (Berman et al., 2003). These structures, which include many small proteins and peptides, have also provided training data to make prediction of peptide ligand conformations more readily available and accurate. Both of these factors have stimulated the application of molecular docking in FBP discoveries (Ferreira et al., 2015; Salmaso & Moro, 2018).

Molecular docking can be used to visualize the interaction mechanism between ligands and receptors at the atomic level, and this ability makes it the most popular bioinformatics method for elucidating the mechanism of action of FBPs (Majid et al., 2022; Tao et al., 2020; C. Wen et al., 2020; W. Zhao et al., 2023). In addition, the scores associated with the best binding mode of ligands can be used for virtual screening by ranking these scores for different ligands. However, these scores are at best a rough estimation of the standard binding free energy, which is directly related to the association (Ka) and dissociation constants (Kd), but is not necessarily correlated with specific bioactivity values (e.g., IC50) (Guedes et al., 2018; X. Li et al., 2010; Tao et al., 2020). Molecular docking has demonstrated its great potential in drug discovery (since 1980s) and has also recently employed for FBP virtual screening (L. Feng et al., 2018; Panyayai et al., 2019; Salmaso & Moro, 2018; Tonolo et al., 2020; Tu et al., 2017). Given the availability of a 3D structure for the receptor and reasonable coordinates for the ligand, molecular docking includes two critical parts. The first is sampling of different locations, orientations, and conformations of the ligand relative to the receptor. In the context of docking, each configuration (location/orientation/conformation) is referred to as a “pose”. The second part is assigning a “score” to each pose that represents its thermodynamic favorability. These scores are the quantity

that the optimization algorithm of the docking program seeks to optimize. Well-designed scoring functions can also be used to compare the best-scoring poses of distinct ligands, enabling virtual screening. Table 5 summarizes popular molecular docking software and web servers with detailed information on sampling algorithms, scoring functions, availability, flexibility, maintenance, and more.

1.4.1. Conformational sampling in molecular docking

Conformational sampling (also known as conformation search) attempts different conformations of the ligand around the whole surface/cavities of the receptor (global docking) or at a designated binding site (local docking) (Ciemny et al., 2018; Ferreira et al., 2015; Maia, Assis, et al., 2020). The number of accessible conformers for peptides of more than a few amino acids is astronomical; hence, it is not feasible to computationally generate all the possible conformations for scoring. Furthermore, for each conformation of the ligand, many orientations and positions relative to the receptor must be considered. In addition, different conformations of the receptor may be considered. A major difference among docking tools and protocols is which portions of the ligand and receptor are regarded as flexible, meaning they are subjected to the conformational sampling algorithm, and which are treated as rigid, meaning that they retain the conformation originally provided by the user throughout the docking process. Some docking tools, such as those intended for protein–protein docking, treat both the ligand and receptor as rigid and "conformational" sampling entails only overall rotations and displacements of the provided ligand structure (Ciemny et al., 2018). Such an approach is usually unsuitable for docking of peptides, which typically have many internal degrees of freedom and a diverse ensemble of thermodynamically accessible structures. The most commonly used approach in FBP studies is semi-flexible docking where the ligand is flexible, while the receptor is rigid (Tao

et al., 2020). Many docking programs also support treating chosen side chains of the receptor (typically those near the binding site) as flexible groups. However, side chain flexibility is not sufficient to permit sampling of distinct receptor conformations that involve changes in protein backbone positions, such as inward-facing and outward-facing states of transporters. Hence, in all cases it is essential to choose an initial receptor conformation that is relevant for the bioactivity of interest. For some applications, it may be necessary to perform docking of a ligand over multiple distinct conformational states of the receptor. Indeed, given the inherent flexibility of proteins, improved performance has been found with ensemble docking, where the ligand is docked to multiple snapshots of the receptor protein obtained from molecular dynamics simulation (Amaro et al., 2018).

The commonly used semi-flexible docking approach is a compromise between computational effort and exhaustive sampling of accessible conformations (Forli et al., 2016; Salmaso & Moro, 2018). It should be noted that more flexibility does not guarantee a significant improvement in docking performance, especially since greater flexibility usually comes at the cost of less exhaustive sampling. A performance assessment by Huang demonstrated that semi-flexible docking generated bound poses nearer to experimentally derived structures than rigid docking, although ranking of ligands for virtual screening showed no clear difference in performance between rigid and flexible approaches (S.-Y. Huang, 2018). However, the study of Huang considered drug-like molecules, which are typically more rigid than many-residue peptides. For peptides of more than a few amino acids, it is likely that consideration of multiple conformations of the peptide is necessary unless it is known to have well-defined folded structure.

Various algorithms are used for sampling ligand and receptor conformations with the available computational power (Meng et al., 2011). The algorithms can be generally divided into three categories based on their searching strategies: systematic search, stochastic search, and deterministic search (dynamics simulation search), and some docking software combines different strategies into a multi-phase approach (e.g., GLIDE, CDOCKER, and DOCK 6) (Allen et al., 2015; Friesner et al., 2004; Gagnon et al., 2016; Maia, Assis, et al., 2020). Briefly, the systematic method is more time consuming than the stochastic method (e.g., the Monte-Carlo methods or genetic algorithm) because of it includes an exhaustive search of possible rotamer states for subsets of the molecule. The computational demand of deterministic search is the largest and proportional to the dynamics simulation runtime. This approach is highly sensitive to the initial conformation and spatial position because they can change very slowly in dynamics simulation. Therefore, a deterministic search can easily be trapped in local minima unless long simulation times are used to cross the barriers or tempering strategies to accelerate this crossing are applied, which consume more time (Ferreira et al., 2015; Ugur et al., 2019). Therefore, deterministic search is usually integrated with stochastic search or systematic search for the final conformation optimization (e.g., CDOCKER, GLOD, and Glide) (Table 5).

1.4.2 Scoring function in molecular docking

The scoring function is the conformation selector for the results from the sampling engine. It selects by estimating the difference in energy (or free energy) between the ligand–receptor complex and the isolated unbound molecules (Gagnon et al., 2016; Maia, Assis, et al., 2020; Salmaso & Moro, 2018). There are three types of scoring functions: force field-based scoring functions, empirical scoring functions, and knowledge-based scoring functions (Guedes et al., 2018). Force field-based scoring functions are composed of a series of energy terms from a

classical force field, including bonded (intramolecular) and nonbonded (intermolecular) components, and in some cases, the solvation terms are also included, but they are more computationally expensive (Meng et al., 2011). Empirical scoring functions have relatively simple energy terms for calculation by assigning weights to different empirical energy terms (e.g., hydrogen bonding, ionic bonding, non-polar interactions, desolvation, and entropic effects), and the weights are obtained by training models for the experimental data (Maia, Assis, et al., 2020; Salmaso & Moro, 2018; Verdonk et al., 2003). Knowledge-based scoring functions are computationally simple and based on statistical analysis of interaction atom pairs from complexes with experimentally derived 3D structures; the frequency of the ligand-receptor atom pairs is computed for scoring (Salmaso & Moro, 2018; Trott & Olson, 2009). There are also some integrated strategies (also called consensus scoring) that attempt to take advantage of different scoring functions. An example is AutoDock Vina, which combines knowledge-based and empirical scoring functions for scoring and ranking.

1.4.3 Other considerations in molecular docking

Docking studies can also be characterized as local, where a known active site of the receptor is targeted, or global (also known as blind docking), where the active site is unknown. Some docking servers (such as CBDock) will first perform an analysis to locate possible binding cavities and then apply local docking in turn to each cavity (Liu et al., 2020). Most of the targeted proteins for FBP studies have known active pockets, so there is no need to conduct global docking, but some web servers (e.g., Pepsite2, CABS-dock, HADDOCK, etc.) adopt global docking by default (Garzón et al., 2022; Martini et al., 2022; Mudgil et al., 2019; Sun et al., 2017). For all docking software, users can choose whether to conduct docking over an

important part of the receptor (local) or over the entire receptor (global), although the latter requires more computational effort or reduced exhaustiveness.

1.4.4 Application of molecular docking in FBP virtual screening

In FBP studies, the most common molecular docking technique is to identify potential bioactive peptides by LC-MS from the highest activity fractionation or proteolysis simulation and then conduct molecular docking of these peptides on targeted proteins (e.g., ACE, renin, DPPIV, etc.) (Table 6). The peptides with the most favorable docking scores can then be synthesized for in vitro or in vivo bioactivity determination, and the docked conformations can help elucidate the interaction mechanism, highlighting particular hydrogen bonds, hydrophobic contacts, π–π stacking, or other such interactions between protein receptors and peptide ligands (L. Feng et al., 2018; Liang et al., 2021; Tonolo et al., 2020; Tu et al., 2017; M. Zhang et al., 2022). This approach does not lend itself to large-scale virtual screening because of the difficulty of identifying and synthesizing large numbers of peptides (Forli et al., 2016). The study conducted by L. Li et al. on Keap1–Nrf2 interaction inhibitory FBPs was a great example for future molecular docking-driven FBP screening. In the study, a total of 400 dipeptides and 6138 tripeptides were screened by CDOCKER, and six dipeptides and ten tripeptides with stronger binding affinity were synthesized for in vitro and in vivo experiments. Finally, two tripeptides (DKK and DDW) with strong inhibitory activity were successfully obtained (L. Li et al., 2017). Another such attempt employed GOLD to screen ACE inhibitory activity from 8000 theoretical tripeptides, of which the five with the highest binding affinity were synthesized for experimental validation (Panyayai et al., 2018).

Molecular docking involves many approximations and, even for exhaustive sampling, the resulting poses are not guaranteed to be correct nor is there a guarantee that the docking score

closely corresponds to experimental inhibition activity, as shown by the inconsistency between binding affinity rank and experimental activity results in the studies of Panyayai et al., and X. Li et al. (Gunalan et al., 2020; L. Li et al., 2017; Panyayai et al., 2018). Some researchers proposed a new docking strategy called consensus docking, which has shown a performance improvement compared to single docking protocols in the reliability of pose prediction for interaction mechanism determination and virtual screening (Houston & Walkinshaw, 2013; Ren et al., 2018). In consensus docking, different molecular docking protocols are used for the same docking task, and then the binding poses from different docking protocols are compared to calculate the root-mean-square deviation (RMSD) value. A molecular docking prediction will be accepted when the RMSD value between different protocols is below 2 angstroms (Ren et al., 2018). There is a well-designed one-stop python package (dockbox, available at https://pypi.org/project/dockbox/) for consensus docking or docking by different protocols (Preto & Gentile, 2019). Another more straightforward strategy for consensus docking relies on summation of normalized docking scores from different docking protocols, and corresponding tools have been released for academic free usage (e.g., MolAr software and the DockingPie plugin for PyMol) (Maia, Medaglia, et al., 2020; Rosignoli & Paiardini, 2022).In the studies of Tonolo et al. and Nardo et al., a similar idea was adopted where two docking protocols were employed consecutively to refine the docking results, but neither study adopted RMSD threshold values or normalized summation of docking score for refinement (Nardo et al., 2020; Tonolo et al., 2020). By combining large-scale screening for FBPs and consensus docking, we can hope to build molecular docking binding affinity databases for small peptides.

Although consensus docking can improve the reliability of docking results for unknown peptides, we also need to assess the correlation between identified FBPs with bioactivity values and the binding affinity results from molecular docking. The only one such study was conducted in 2007 by Pripp, where the coefficient is only 0.29 between the docking score of 29 ACE inhibitory FBPs with ACE and the experimentally observed log(1/IC50), which means the prediction power of molecular docking in the virtual screening of ACE inhibitory peptides was poor (Pripp, 2007). With the increasing availability of FBPs and improvements in molecular docking strategies in the past decades, especially the introduction of consensus docking, we may hope to see performance improvement in the near future.

Finally, there are two areas of concern in molecular docking. First, unlike conventional drug-like molecules that may include a wide variety of chemical groups but typically have a limited number of accessible conformations, most peptides are composed of only 20 different amino acids but have high conformational flexibility. As such, more information in the representative descriptors or strings may be needed to describe their conformation flexibility, especially for long peptides that form secondary structures (> 4 residues) (Forli et al., 2016; Irwin et al., 2022; Yu et al., 2018). Table 1 includes some secondary structure prediction tools for molecular docking of long peptides and protein receptors when a secondary structure is needed (e.g., CABS-dock). Second, unless an allosteric binding site is known and considered as part of the docking region, molecular docking is not suitable for virtual screening of non-competitive inhibitors, which may be another important factor that causes inconsistency between in vitro or in vivo studies and molecular docking (Nong & Hsu, 2022).

1.5. Molecular dynamics simulation and its application in FBP virtual screening

Molecular dynamics simulation is also a structure-based approach for virtual screening and can explore interaction mechanisms at the atomic level, but is typically much more computationally demanding than molecular docking (Ferreira et al., 2015; Vidal-Limon et al., 2022). It is typically considered to be more accurate than docking and better able to give physical insight (Mobley & Dill, 2009), compensating for its greater computational cost. In molecular dynamics simulations, the behavior of molecules (not limited to ligands and receptors) over a period of time under the constraints of physical laws (classical or quantum theories) are captured (Salmaso & Moro, 2018; Tao et al., 2020). Quantum chemistry methods, which explicitly treat all or some electrons at the quantum mechanical level, are prized for yielding highly accurate energies and molecular geometries with few or no empirical parameters. However, despite methodological improvements in the past couple decades, the computational expense of quantum mechanics-based methods remains prohibitive for probing the behavior of proteins surrounded with explicit water molecules on time scales relevant for binding processes (often microseconds or more). On the other hand, classical molecular dynamics simulations can be used to simulate complete proteins and ligands in an aqueous environment for time scales currently reaching many microseconds (or even milliseconds with specialized hardware (Shaw et al., 2009)). Classical molecular dynamics is based on numerical solutions of Newton's equations of motion (or similar equations for constant temperature or pressure thermodynamic ensembles) for collections of atoms. The downside of this method is that the accuracy of the results is dependent on the accuracy of the empirical potential energy functions, termed "force fields", that describe interactions between atoms (or particles representing groups of atoms in united atom or coarse-grained models). Fortunately for the study of peptide–protein binding, force fields describing

polypeptides consisting of the 20 proteinogenic amino acids in water have been developed and constantly improved over the past three decades (Best et al., 2012; Cornell et al., 1995; J. Huang et al., 2017; Jorgensen et al., 1996; MacKerell et al., 1998; Ploetz et al., 2021; C. Tian et al., 2020).

The most practical consideration for molecular dynamics simulation is the selection of the force field. Commonly used force fields and molecular dynamics simulation software are summarized in Table 5. The most common force fields for protein and peptide simulations are the CHARMM and Amber force fields, which are based on models where each atom is represented as a point particle carrying a partial charge and the network of covalent bonds is fixed at the beginning of the simulation (J. Huang et al., 2017; C. Tian et al., 2020). Therefore, while conventional molecular dynamics simulations can represent interactions that give rise to physical bonding, including hydrogen bonds, salt bridges, the hydrophobic effect, and other solvation-dependent interactions, they are unable to directly describe chemical reactions and changes in protonation state and may not accurately represent changes in electrical polarization between different environments. Techniques and alternate force fields to overcome these limitations have been developed; however, they often incur increased computational cost and, for typical peptide–protein simulations, improved accuracy is not guaranteed. There can be trade-offs between systematic errors, which are reduced by more accurate models, and statistical errors due to insufficient sampling, which can be increased by using more accurate models since these models are often slower and the real time available to researchers is fixed.

1.5.1 Molecular dynamics simulations for protein–ligand binding

Conventional molecular dynamics simulations (with fixed atomic charges, protonation states, and covalent bonds) have become mainstream for studying peptide-protein binding,

including for studying FBPs (T. Feng et al., 2015; Hollingsworth & Dror, 2018). By putting peptide ligands and protein receptors into a simulation box, typically along with explicit water molecules and dissolved ions, one can simulate protein-ligand interaction for anywhere between femtoseconds to milliseconds (Shaw et al., 2009) at a given temperature and pressure. If the kinetics of binding between the ligand and receptor is sufficiently fast, a protein-ligand complex will be observed after a period of dynamics simulation. In the case of very fast kinetics and low binding affinity, multiple binding and unbinding events can be observed, allowing for direct estimation of the equilibrium constants. However, for peptide ligands, the kinetics will be almost always much too slow to estimate equilibrium constants from brute force simulation. Including multiple ligands in the simulation box can help to observe spontaneous binding, although there is no guarantee that any of the binding poses found are the lowest energy (Comer et al., 2021; Ishiguro et al., 2022). Docking is often a more efficient way to search for and find putative bound conformations of protein–ligand complexes. Hence, a typical molecular dynamics approach to virtual screening is to perform docking to obtain multiple poses of the bound ligand in the protein binding site. Each of the poses can be solvated in explicit water and ions and subjected to molecular dynamics simulation. Restraints are usually applied during the energy minimization and equilibration stages of the simulation to avoid prematurely disrupting the complex while water molecules, ions, and some parts of the protein may be far from equilibrium. Such molecular dynamics simulations can help to identify which poses predicted by docking are truly stable (Gunalan et al., 2020; Kalyan, Junghare, Bhattacharya, et al., 2021). For example, if for a given pose, the peptide dissociates from the protein a few nanoseconds after equilibration for several different initial conditions (different initial atomic velocities or water molecule positions), then it can be assumed that the binding affinity for this pose is marginal.

Numerous methodologies have been used to estimate free energies of protein-ligand binding for virtual screening and the field continues to develop rapidly (Mobley & Dill, 2009). The SAMPL challenges provide a snapshot of methods currently used for estimating binding free energies from molecular dynamics simulations (or by other means such as machine learning) and some indication of which methods may work better than others (Amezcua et al., 2021, 2022; Rizzi et al., 2018; Yin et al., 2017). While these SAMPL challenges focus on host-guest systems, which are likely easier than protein-ligand systems due to the relative symmetry and rigidity of the hosts compared to proteins, they reveal aspects of the calculation that need to be considered for accurate results such as the details of the computational protocol, conformational sampling, multiple bound conformations, minor protonation states, and atomic polarizability.

At the present time, molecular models with fixed atomic charges, protonation states, and covalent networks and explicit solvent are the most commonly used for simulating peptide–protein interactions; however, for specific cases, more sophisticated methods and force fields might be needed. For instance, peptide therapeutics with a covalent mechanism of action have recently attracted increased interest (Berdan et al., 2021) and can be studied by hybrid methods combining classical molecular dynamics and quantum chemistry methods (Jagger et al., 2020). Protein force fields that explicitly represent atomic polarizability have also been developed (Shi et al., 2013) and are continually being improved (F.-Y. Lin et al., 2020; Rackers et al., 2016). However, models with atomic polarizability incur a greater computational cost and have less history of development than force fields with fixed atomic charges and, so far, there is no guarantee of improved accuracy over fixed-charge force fields for all systems. Nonetheless, recent submissions to the SAMPL challenges using a polarizable force field appear have been among the best performers (Amezcua et al., 2021, 2022). Furthermore, for systems where highly

polarizable ions play important roles, force fields including atomic polarizability can be crucial for reasonable results (Vergara-Jaque et al., 2017).

1.5.2 Binding free energy estimates with MM-PBSA and MM-GBSA methods

Even if association or dissociation is observed on a time scale accessible to molecular dynamics simulations, a single simulation is not sufficient accurately characterize the equilibrium constant for protein–ligand binding or, equivalently, the binding free energy. As end-point methods, the MM-PBSA (molecular mechanics Poisson–Boltzmann surface area) and MM-GBSA (molecular mechanics generalized Born surface area) methods provide a convenient way to estimate binding free energies from molecular dynamics simulations without the need for observing association or dissociation events. Tools for performing these calculations with the GROMACS and Amber molecular dynamics packages have been developed (Miller et al., 2012; Valdés-Tresanco et al., 2021). The basic idea of the methods is use explicit–solvent molecular dynamics simulation to obtain an ensemble of different conformations of bound complex, and separately, ensembles of conformations of the free receptor and ligand (Kollman et al., 2000). The binding free energy is estimated by post-processing these simulation trajectories using the equation $\Delta G_{bind} = G_{complex} - G_{free\ ligand} - G_{free\ receptor}$. Calculating these Gibbs free energy terms directly from explicit-solvent simulation is challenging due to the noisiness of the solvent contribution to the enthalpy (water molecules and ions sample many different positions and orientations during the simulation) and the difficulty of calculating changes in the entropy of the solvent (Azhagiya Singam et al., 2019). The MM-GBSA and MM-PBSA methods overcome these difficulties by discarding the explicit water molecules and dissolved ions from the trajectory and estimating the solvation contributions to the free energy using continuum models. The polar solvent/ion contribution to free energy is calculated using the continuum Poisson–

Boltzmann (PB) or generalized Born (GB) techniques and the nonpolar contribution with a term proportional to the solvent-accessible surface area (SA). Hence, Gibbs free energy terms for each of the three systems (complex, free ligand, free receptor) are calculated by G = <EMM> + <Gsolv> + γ<A> – TSconf, where <...> denotes an average over the simulation trajectory, EMM is the potential energy from the force field for the ligand and receptor (neglecting the explicit water molecules and dissolved ions), Gsolv is the polar portion of the solvation free energy calculated by Poisson–Boltzmann or generalized Born method and averaged over the trajectory, γA is the nonpolar contribution to the solvation free energy, A is the solvent-accessible surface area of the ligand and receptor molecules, γ is a parameter linearly relating the solvent-accessible surface area and free energy, T is the temperature, Sconf is the conformational entropy of the ligand and receptor. The conformation entropy is typically estimated by a normal mode analysis or the quasi-harmonic method, which are both implemented in the GROMACS package (Abraham et al., 2015; Genheden et al., 2012; Karplus & Kushick, 1981). It should be noted that these methods of computing conformational entropy are costly and the inclusion of the conformational entropy term not guaranteed to improve agreement with experiments (Sun et al., 2018).

1.5.3 Rigorous binding free energy calculations

MM-PBSA and MM-GBSA methods are not rigorous in the sense that the simulations are typically performed with explicit water, yet the solvation free energy is obtained with a continuum model, and the conformational entropy is not directly calculated, but approximated using normal mode or quasi-harmonic methods. While MM-PBSA and MM-GBSA are more efficient than rigorous methods, they can yield poor predictions for protein-peptide binding free energies where rigorous methods obtain good agreement with experiment (H. Fu et al., 2017).

Rigorous methods instead seek to calculate the free energy under a given molecular dynamics model and force field without approximations (although this free energy value may differ from the true experimental value). These methods are not end-point methods and require comprehensive sampling along a path from the bound to unbound states, which requires considerable computational resources for even a single calculation. The path taken between the bound and unbound states may be entirely unphysical, such as in alchemical methods where the atoms of the ligand are gradually deleted or inserted, or the ligand and the receptor may be physically separated along geometric coordinates, which may still not be the most likely path taken in reality (Chodera et al., 2011; Gumbart et al., 2013). However, because free energy is a state function, the path taken between the two states should not affect the change in free energy between the states and, therefore, one can choose an unphysical path that makes the calculation most efficient (or merely feasible). Notably, changes in the conformation and orientation of the ligand relative to the receptor between the bound and unbound states can make convergence of the free energy unacceptably slow. Hence, for both alchemical and geometric pathways, it is often necessary to apply conformational and orientational restraints to the ligand (and possibly parts of the receptor as well) (Woo & Roux, 2005). However, since these restraints are unphysical, their effect must be removed. Therefore, conceptually, one takes the following approach. First, the free energy of applying conformational and orientational restraints to the free ligand in solution is calculated. Next, the free energy of binding to the receptor is calculated under these restraints, which allow relatively rapid convergence. Finally, the free energy of releasing the conformational and orientational restraints is calculated, so that the final free energy is calculated as $\Delta G_{bind} = \Delta G_{apply\ restraints} + \Delta G_{bind\text{-}restrained} + \Delta G_{release\ restraints}$. While this approach is quite complex, recently several tools have been released to make simplify

performing rigorous free binding calculations, including BFEE (H. Fu et al., 2018), BFEE2 (H. Fu et al., 2021), YANK, BRIDGE (Senapathi et al., 2020), and pAPRika (Velez-Vega & Gilson, 2013).

It should be noted that this rigorous approach requires an initial pose, used as the reference for the conformational and orientational restraints, that is assumed to be an accurate representation of the bound state of the complex. The MM-PBSA/MM-GBSA methods also require an initial pose. Obtaining such an accurate pose can be difficult in the absence of an experimental structure, but can be found by docking and brute-force molecular dynamics simulations. Identifying the lowest free energy bound state may require performing the rigorous or MM-PBSA/MM-GBSA method with different poses. It is also possible that more than one distinct pose contributes significantly to the bound free energy. Hence, it is often beneficial to use a hierarchy of methods for screening. For example, we used FlexPepDock to generate initial poses and docking scores for 17 different peptide sequences, screened them by performing conventional molecular dynamics simulations for up to 2 μs and calculating the MM-GBSA free energies, and, finally, performed rigorous free energy calculations using the BFEE tool for two selected peptides (Thakkar et al., 2023).

1.5.4 Application of molecular dynamics simulation in FBP studies

In FBP studies, the application of molecular dynamics simulation is not as popular as that of molecular docking, which may be caused by the tremendous computational demands and the complexity of the operation (Arantes et al., 2021; Shaw et al., 2009). It is more likely to be used to supplement molecular docking results by optimizing the conformations of peptide–protein complexes for interaction mechanism elucidation, or it may be used to validate the stability of peptide–protein complexes predicted through RMSD values (Aguilar-Toalá et al., 2022; Y. Li et

al., 2022; Liang et al., 2021; Panyayai et al., 2019). For example, two dual-functional peptides (RALP and WYT) were studied by molecular docking and molecular dynamics simulation to understand their difference in inhibitory potency. Both peptide–ACE complexes were intact after 200 ns of simulation, and the two peptides moved out of the active sites of peptide–renin complexes, which demonstrated the instability of the peptide–protein complexes predicted from molecular docking. The instability of this complex was further supported by the low bioactivity (Gunalan et al., 2020). Similar applications were also seen in the studies of Liang et al. and Panyayai et al., where molecular dynamics simulation was used to check the stability of the complex in salt solution under room temperature (300 K) and one bar pressure for 30 ns and 25 ns, respectively (Liang et al., 2021; Panyayai et al., 2018). Such additional work can help explain the inconsistency between predicted and experimental results in FBP studies.

Recently, a user-friendly front-end was proposed to run molecular dynamics simulation protocols in Google Colaboratory. It successfully simplified the running procedures and provided the needed computation power using cloud computing (Arantes et al., 2021). Such efforts from bioinformatics could enhance the availability of molecular simulations and benefit biochemistry researchers who struggle with limited in-house computing facilities, programming environments, simulation operations, and result analysis.

1.6 Progress of integrated strategies in FBP screening

The integration of QSAR models, molecular docking, and molecular dynamics simulation has grown popular among biochemistry researchers for FBP virtual screening (Table 7). The most popular strategy is to employ the QSAR model for the first round of screening of peptide sequences from in silico proteolysis or LC-MS identification; further screening is then conducted by molecular docking, and sometimes molecular dynamics simulation is used to optimize

conformation for the interaction mechanism elucidation (Baba, Baby, et al., 2021; Baba, Mudgil, et al., 2021; Kalyan, Junghare, Khan, et al., 2021; X. Li et al., 2022; Marcet et al., 2022; F. Wang & Zhou, 2020). The combination of different virtual screening methods is expected to possess higher accuracy in FBP identification. These methods were established on different theories to describe and characterize peptides for screening, so double screening would be more efficient than refinement under the same theory (e.g., double screening by different molecular docking methods) (K. Lin et al., 2017; Mudgil et al., 2019; Nardo et al., 2020).

Most of the integrated studies focus on ACE inhibitory activity and DPP IV inhibitory activity (Aguilar-Toalá et al., 2022; Baba, Baby, et al., 2021; Garzón et al., 2022; Kalyan, Junghare, Khan, et al., 2021; X. Li et al., 2022; Y. Li et al., 2022; K. Lin et al., 2017, 2017; Marcet et al., 2022; Martini et al., 2021, 2022; Mudgil et al., 2019; Sun et al., 2017; Tahir et al., 2020; Y.-T. Wang et al., 2020, 2020; Wenhui et al., 2022; Yu et al., 2018; T. Zhang et al., 2015). FBPs with these two inhibitory activities have the potential to relieve two chronic diseases: hypertension and diabetes. As such, they deserve intensive attention from researchers. On the other hand, the bioinformatics approach is a kind of knowledge-based study, which means the more information that is available for one type of bioactivity, the better future bioinformatics studies will be for that type of bioactivity, especially for QSAR model development. Therefore, for rarely studied types of bioactivity, to initiate bioinformatics-aided studies, we need to conduct large-scale molecular docking screening as well as wet chemistry studies to provide foundational knowledge on structure-activity relationships for QSAR model development.

Besides the general evaluation model (PeptideRanker), some QSAR classification web servers for specific types of activity (e.g., PreAIP, AntiAngioPred, and PlifePred) are employed in FBP studies. In addition, some QSAR models built by biochemistry researchers have

outperformed in dataset collection and curation, characterization methods, and model analysis, compared to those models created bioinformatics background laboratories perhaps due to biochemistry researchers' better understanding of biological properties (Kalyan, Junghare, Khan, et al., 2021; Y. Li et al., 2022; K. Lin et al., 2017; Sun et al., 2017; F. Wang & Zhou, 2020; Y.-T. Wang et al., 2020; T. Zhang et al., 2015). For example, Kalyan et al. used data mining to retrieve 1687 peptides from two databases, which was significantly larger than most self-built models. In addition, non-linear regression methods such as regression decision tree and back-propagation neural network (BPNN) were employed by some biochemistry researchers (Kalyan, Junghare, Khan, et al., 2021; T. Zhang et al., 2015). However, most of these models suffer from one or several issues mentioned in Section 3, such as small datasets and poor peptide descriptors (Kalyan, Junghare, Khan, et al., 2021; Y. Li et al., 2022; K. Lin et al., 2017; Sun et al., 2017; F. Wang & Zhou, 2020; Y.-T. Wang et al., 2020; T. Zhang et al., 2015).

Molecular docking tools used in integrated studies are diverse, and some studies even adopted two docking tools for better refinement and docking conformation optimization (Baba, Baby, et al., 2021; Garzón et al., 2022; Martini et al., 2022; Mudgil et al., 2019; Nardo et al., 2020; Xue et al., 2022; Yu et al., 2018). It is difficult to differentiate the performance of different molecular docking protocols in specific protein receptor cases, but generally, for the same docking task, performance is proportional to time consumption (Arantes et al., 2021; Attique et al., 2019). Compared to the use of two docking tools for refinement, consensus docking would be a better alternative by combining different molecular docking protocols with different sampling algorithms or scoring functions (Ren et al., 2018). The use of molecular docking web servers is popular among FBP studies, and brief introductions of these web servers are given in Table 5.

A novel strategy (ensemble docking) was introduced by Aguilar-Toalá et al. for FBP screening, where molecular dynamics simulations of 300 ns were conducted to generate protein receptor conformations. Three representative conformations were selected for further molecular docking tasks, and the average score of the ligand with the three representative receptor conformations was used for virtual screening (Aguilar-Toalá et al., 2022). Ensemble docking is also an alternative approach to increase molecular docking accuracy. It can help locate high-activity peptides by changing the protein receptor conformation from a single source conformation (experimental or modeled) to a number of conformations and thus enhance docking performance (Evangelista Falcon et al., 2019; Korb et al., 2012).

1.7 Conclusion and directions for future research

This review provided an overview of using bioinformatics to accelerate FBP screening and interaction mechanism exploration. Database-driven virtual screening with proteolysis simulation has been widely used to identify FBPs, to compare them with reported FBPs, and to differentiate unknown peptides for further screening. QSAR, molecular docking, and molecular dynamics simulation are now commonly used for virtual screening of unknown peptides and elucidating interaction mechanisms. This review discussed in detail the limitations of database-driven studies and bioactivity potency evaluation system; the role of the dataset, peptide representation, and model development in QSAR studies; as well as sampling algorithms, scoring functions, force fields, and free energy estimation methods. Although a lot of progress has been made using bioinformatics in FBP studies, there are still many challenges and specific technical expertise can be required to obtain accurate results.

Bioinformatics has the potential to guide the production of value-added products from agricultural byproducts, to promote sustainable agriculture, to evaluate the potency of food

bioactivity for precise nutrition, and to rationally design peptides derived from food for nutraceutical, cosmeceutical, and pharmaceutical industries. The collaboration between biochemistry and bioinformatics researchers is essential to achieve this potential.

To advance the application of bioinformatics in FBP studies, we recommend the following areas:

1). It is expected to build government-led non-profit databases for bioactive peptides, similar to protein databases (e.g., NCBI, RCSB PDB, etc.). Such large and high quality databases are critical to QSAR model development and efficient information retrieval of bioactive peptides.

2). Dataset cleaning is a challenging task when collecting FBPs data from different literature sources, where some bioactivity results are affected by manual operations, experimental conditions, etc., and should be removed for high quality dataset construction. An example is our lab's web server for ACE inhibitory peptide prediction, where a confident learning theory based tool, CleanLab, is employed to clean real-world datasets and generate a high quality ACE inhibitory peptide dataset (Du et al., 2024).

3). Limited negative sample datasets are available for virtual screening of FBPs, since most reported FBPs were identified with high activity. A solution is to manually create negative samples. For example, DUD-E server (http://dude.docking.org/) can provide challenging decoys for checking molecular docking or QSAR model performance, which have been used in umami dipeptide screening (Y. Xiong et al., 2022).

4). Protein language models are expected to gradually become the mainstream peptide representation for bioactivity classification model development. Though it exhibited better performance in peptide information representation, it suffers from high feature dimension

problem and might undermine its application in small datasets. For QSAR model development, besides the employment of advanced feature selection methods and modeling methods, an stacking framework is another alternative for modeling strategy, where different modeling methods are combined together for decision-making.

5). When the 3D structure of a targeted protein is not available from X-ray crystallography or NMR, de novo structure prediction by Alpha-Fold2 can be a great alternative to create the 3D structure (Jumper et al., 2021). In addition, homology modeling and threading modeling methods can be considered. In the last five years, Electron Microscopy Data Bank (EMDB, https://www.ebi.ac.uk/emdb/search/) has rapidly added new entries and cryo-electron microscopy has rapidly become a major source of experimental structures, overcoming many difficulties with traditional X-ray crystallography and NMR methods (Alnabati et al., 2022; Callaway, 2015, 2020).

6). Molecular docking with advanced docking strategies, such as consensus docking and ensemble docking, are expected to be applied in FBPs discovery and improve the docking accuracy. In addition, large-scale virtual screening based on open-source programs have not yet been widely used in FBPs studies. With the rapidly developing cloud computing platforms such as Amazon Web Services (AWS), a user-friendly and easy-operated servers is in great demand and will significantly promote the popularization of molecular docking and molecular dynamics simulation in FBPs studies.

7). Graphic processing unit (GPU) parallel acceleration and high-performance computation clusters have been introduced into molecular docking and molecular dynamics simulation. They can decrease computer time by more than an order of magnitude and enable even expensive molecular dynamics simulations to be performed with commodity computer

hardware (H. Chen et al., 2020; Eastman et al., 2017; Kondratyuk et al., 2021; Kutzner et al., 2019; Páll et al., 2020; Phillips et al., 2020).

8). Current virtual screening studies that use molecular docking and molecular dynamics simulation focus on the interaction between different ligands and the same receptor. It would be meaningful to conduct molecular docking with identified FBPs that have specific bioactivities and proteins with other desired bioactivities for the screening of multi-bioactivity FBPs. This strategy has been used in QSAR studies (e.g., SwissTargetPrediction) where various 2D and 3D molecular fingerprints were proposed to encode molecules with unknown bioactivity in digital formats for molecular similarity calculations (Manhattan distance-based similarity). The most matchable molecules among the 376,342 molecules in the dataset were used to predict the potential activity corresponding to 3,068 macromolecular targets (Daina et al., 2019).

9). Biochemistry researchers are encouraged to keep up with the latest progress in bioinformatics fields and its corresponding outcomes, and bioinformatic researchers could make their findings more accessible to biochemistry researchers, such as developing and providing web servers.

Declaration of interest

The authors declare that there are no known conflicts of interest.

Acknowledgements

This is contribution No. 23-155-J from the Kansas Agricultural Experimental Station. This work was supported in part by the Agriculture and Food Research Initiative Competitive Grant no. 2020-68008-31408 and no. 2021-67021-34495 from the USDA National Institute of Food and Agriculture, and a seed grant from the Global Food Systems initiative of Kansas State University.

References

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, *1–2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001

Agirbasli, Z., & Cavas, L. (2017). In silico evaluation of bioactive peptides from the green algae Caulerpa. *Journal of Applied Phycology*, *29*(3), 1635–1646. https://doi.org/10.1007/s10811-016-1045-7

Aguilar-Toalá, J. E., Vidal-Limon, A., & Liceaga, A. M. (2022). Multifunctional Analysis of Chia Seed (Salvia hispanica L.) Bioactive Peptides Using Peptidomics and Molecular Dynamics Simulations Approaches. *International Journal of Molecular Sciences*, *23*(13), Article 13. https://doi.org/10.3390/ijms23137288

Agyei, D., Tsopmo, A., & Udenigwe, C. C. (2018). Bioinformatics and peptidomics approaches to the discovery and analysis of food-derived bioactive peptides. *Analytical and Bioanalytical Chemistry*, *410*(15), 3463–3472. https://doi.org/10.1007/s00216-018-0974-1

Allen, W. J., Balius, T. E., Mukherjee, S., Brozell, S. R., Moustakas, D. T., Lang, P. T., Case, D. A., Kuntz, I. D., & Rizzo, R. C. (2015). DOCK 6: Impact of new features and current docking performance. *Journal of Computational Chemistry*, *36*(15), 1132–1156. https://doi.org/10.1002/jcc.23905

Alnabati, E., Terashi, G., & Kihara, D. (2022). Protein Structural Modeling for Electron Microscopy Maps Using VESPER and MAINMAST. *Current Protocols*, *2*(7), e494. https://doi.org/10.1002/cpz1.494

Amaro, R. E., Baudry, J., Chodera, J., Demir, Ö., McCammon, J. A., Miao, Y., & Smith, J. C. (2018). Ensemble Docking in Drug Discovery. *Biophysical Journal*, *114*(10), 2271–2278. https://doi.org/10.1016/j.bpj.2018.02.038

Amezcua, M., El Khoury, L., & Mobley, D. L. (2021). SAMPL7 Host–Guest Challenge Overview: Assessing the reliability of polarizable and non-polarizable methods for binding free energy calculations. *Journal of Computer-Aided Molecular Design*, *35*(1), 1–35. https://doi.org/10.1007/s10822-020-00363-5

Amezcua, M., Setiadi, J., Ge, Y., & Mobley, D. L. (2022). An overview of the SAMPL8 host–guest binding challenge. *Journal of Computer-Aided Molecular Design*, *36*(10), 707–734. https://doi.org/10.1007/s10822-022-00462-5

Amigo, L., Martínez-Maqueda, D., & Hernández-Ledesma, B. (2020). In Silico and In Vitro Analysis of Multifunctionality of Animal Food-Derived Peptides. *Foods*, *9*(8), 991. https://doi.org/10.3390/foods9080991

Arámburo-Gálvez, J. G., Arvizu-Flores, A. A., Cárdenas-Torres, F. I., Cabrera-Chávez, F., Ramírez-Torres, G. I., Flores-Mendoza, L. K., Gastelum-Acosta, P. E., Figueroa-Salcido, O. G., & Ontiveros, N. (2022). Prediction of ACE-I Inhibitory Peptides Derived from Chickpea (Cicer arietinum L.): In Silico Assessments Using Simulated Enzymatic Hydrolysis, Molecular Docking and ADMET Evaluation. *Foods*, *11*(11), Article 11. https://doi.org/10.3390/foods11111576

Arantes, P. R., Polêto, M. D., Pedebos, C., & Ligabue-Braun, R. (2021). Making it Rain: Cloud-Based Molecular Simulations for Everyone. *Journal of Chemical Information and Modeling*, *61*(10), 4852–4856. https://doi.org/10.1021/acs.jcim.1c00998

Attique, S., Hassan, M., Usman, M., Atif, R., Mahboob, S., Al-Ghanim, K., Bilal, M., & Nawaz, M. (2019). A Molecular Docking Approach to Evaluate the Pharmacological Properties of Natural and Synthetic Treatment Candidates for Use against Hypertension. *International Journal of Environmental Research and Public Health*, *16*(6), 923. https://doi.org/10.3390/ijerph16060923

Azhagiya Singam, E. R., Zhang, Y., Magnin, G., Miranda-Carvajal, I., Coates, L., Thakkar, R., Poblete, H., & Comer, J. (2019). Thermodynamics of Adsorption on Graphenic Surfaces from Aqueous Solution. *Journal of Chemical Theory and Computation*, *15*(2), 1302–1316. https://doi.org/10.1021/acs.jctc.8b00830

Baba, W. N., Baby, B., Mudgil, P., Gan, C.-Y., Vijayan, R., & Maqsood, S. (2021). Pepsin generated camel whey protein hydrolysates with potential antihypertensive properties: Identification and molecular docking of antihypertensive peptides. *LWT*, *143*, 111135. https://doi.org/10.1016/j.lwt.2021.111135

Baba, W. N., Mudgil, P., Baby, B., Vijayan, R., Gan, C.-Y., & Maqsood, S. (2021). New insights into the cholesterol esterase- and lipase-inhibiting potential of bioactive peptides from camel whey hydrolysates: Identification, characterization, and molecular interaction. *Journal of Dairy Science*, *104*(7), 7393–7405. https://doi.org/10.3168/jds.2020-19868

Barati, M., Javanmardi, F., Jabbari, M., Mokari-Yamchi, A., Farahmand, F., Eş, I., Farhadnejad, H., Davoodi, S. H., & Mousavi Khaneghah, A. (2020). An in silico model to predict and estimate digestion-resistant and bioactive peptide content of dairy products: A primarily study of a time-saving and affordable method for practical research purposes. *LWT*, *130*, 109616. https://doi.org/10.1016/j.lwt.2020.109616

Baskaran, R., Chauhan, S. S., Parthasarathi, R., & Mogili, N. S. (2021). In silico investigation and assessment of plausible novel tyrosinase inhibitory peptides from sesame seeds. *LWT*, *147*, 111619. https://doi.org/10.1016/j.lwt.2021.111619

Bechaux, J., Ferraro, V., Sayd, T., Chambon, C., Le Page, J. F., Drillet, Y., Gatellier, P., & Santé-Lhoutellier, V. (2020). Workflow towards the generation of bioactive hydrolysates from porcine products by combining in silico and in vitro approaches. *Food Research International*, *132*, 109123. https://doi.org/10.1016/j.foodres.2020.109123

Berdan, V. Y., Klauser, P. C., & Wang, L. (2021). Covalent peptides and proteins for therapeutics. *Bioorganic & Medicinal Chemistry*, *29*, 115896. https://doi.org/10.1016/j.bmc.2020.115896

Berman, H., Henrick, K., & Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. *Nature Structural & Molecular Biology*, *10*(12), Article 12. https://doi.org/10.1038/nsb1203-980

Best, R. B., Zhu, X., Shim, J., Lopes, P. E. M., Mittal, J., Feig, M., & MacKerell, A. D. Jr. (2012). Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ1 and χ2 Dihedral Angles. *Journal of Chemical Theory and Computation*, *8*(9), 3257–3273. https://doi.org/10.1021/ct300400x

Bo, W., Chen, L., Qin, D., Geng, S., Li, J., Mei, H., Li, B., & Liang, G. (2021). Application of quantitative structure-activity relationship to food-derived peptides: Methods, situations, challenges and prospects. *Trends in Food Science & Technology*, *114*, 176–188. https://doi.org/10.1016/j.tifs.2021.05.031

Boachie, R. T., Okoro, F. L., Imai, K., Sun, L., Elom, S. O., Nwankwo, J. O., Ejike, C. E. C. C., & Udenigwe, C. C. (2019). Enzymatic release of dipeptidyl peptidase-4 inhibitors (gliptins) from pigeon pea (*Cajanus cajan*) nutrient reservoir proteins: In silico and in vitro assessments. *Journal of Food Biochemistry*, *43*(12). https://doi.org/10.1111/jfbc.13071

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., & Linial, M. (2022). ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, *38*(8), 2102–2110. https://doi.org/10.1093/bioinformatics/btac020

Callaway, E. (2015). The revolution will not be crystallized: A new method sweeps through structural biology. *Nature*, *525*(7568), 172–174. https://doi.org/10.1038/525172a

Callaway, E. (2020). Revolutionary cryo-EM is taking over structural biology. *Nature*, *578*(7794), 201–201. https://doi.org/10.1038/d41586-020-00341-9

Charoenkwan, P., Nantasenamat, C., Hasan, M. M., Manavalan, B., & Shoombuatong, W. (2021). BERT4Bitter: A bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides. *Bioinformatics*, *37*(17), 2556–2562. https://doi.org/10.1093/bioinformatics/btab133

Chatterjee, A., Kanawjia, S. K., Khetra, Y., & Saini, P. (2015). Discordance between in silico & in vitro analyses of ACE inhibitory & antioxidative peptides from mixed milk tryptic whey protein hydrolysate. *Journal of Food Science and Technology*, *52*(9), 5621–5630. https://doi.org/10.1007/s13197-014-1669-z

Chen, H., Maia, J. D. C., Radak, B. K., Hardy, D. J., Cai, W., Chipot, C., & Tajkhorshid, E. (2020). Boosting Free-Energy Perturbation Calculations with GPU-Accelerated NAMD.

Journal of Chemical Information and Modeling, *60*(11), 5301–5307. https://doi.org/10.1021/acs.jcim.0c00745

Chen, J., Cheong, H. H., & Siu, S. W. I. (2021). xDeep-AcPEP: Deep Learning Method for Anticancer Peptide Activity Prediction Based on Convolutional Neural Network and Multitask Learning. *Journal of Chemical Information and Modeling*, *61*(8), 3789–3803. https://doi.org/10.1021/acs.jcim.1c00181

Chen, N., Chen, J., Yao, B., & Li, Z. (2018). QSAR Study on Antioxidant Tripeptides and the Antioxidant Activity of the Designed Tripeptides in Free Radical Systems. *Molecules*, *23*(6), 1407. https://doi.org/10/gdttrq

Chodera, J. D., Mobley, D. L., Shirts, M. R., Dixon, R. W., Branson, K., & Pande, V. S. (2011). Alchemical free energy methods for drug discovery: Progress and challenges. *Current Opinion in Structural Biology*, *21*(2), 150–160. https://doi.org/10.1016/j.sbi.2011.01.011

Ciemny, M., Kurcinski, M., Kamel, K., Kolinski, A., Alam, N., Schueler-Furman, O., & Kmiecik, S. (2018). Protein–peptide docking: Opportunities and challenges. *Drug Discovery Today*, *23*(8), 1530–1537. https://doi.org/10.1016/j.drudis.2018.05.006

Comer, J., Bassette, M., Burghart, R., Loyd, M., Ishiguro, S., Azhagiya Singam, E. R., Vergara-Jaque, A., Nakashima, A., Suzuki, K., Geisbrecht, B. V., & Tamura, M. (2021). Beta-1,3 Oligoglucans Specifically Bind to Immune Receptor CD28 and May Enhance T Cell Activation. *International Journal of Molecular Sciences*, *22*(6), Article 6. https://doi.org/10.3390/ijms22063124

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., & Kollman, P. A. (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*, *117*(19), 5179–5197. https://doi.org/10.1021/ja00124a002

Coscueta, E. R., Batista, P., Gomes, J. E. G., da Silva, R., & Pintado, M. M. (2022). Screening of Novel Bioactive Peptides from Goat Casein: In Silico to In Vitro Validation. *International Journal of Molecular Sciences*, *23*(5), Article 5. https://doi.org/10.3390/ijms23052439

Dai, Z., Wang, L., Chen, Y., Wang, H., Bai, L., & Yuan, Z. (2014). A pipeline for improved QSAR analysis of peptides: Physiochemical property parameter selection via BMSF, near-neighbor sample selection via semivariogram, and weighted SVR regression and prediction. *Amino Acids*, *46*(4), 1105–1119. https://doi.org/10/f5vxh2

Daina, A., Michielin, O., & Zoete, V. (2017). SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Scientific Reports*, *7*(1), Article 1. https://doi.org/10.1038/srep42717

Daina, A., Michielin, O., & Zoete, V. (2019). SwissTargetPrediction: Updated data and new features for efficient prediction of protein targets of small molecules. *Nucleic Acids Research*, *47*(W1), W357–W364. https://doi.org/10.1093/nar/gkz382

Deng, B., Long, H., Tang, T., Ni, X., Chen, J., Yang, G., Zhang, F., Cao, R., Cao, D., Zeng, M., & Yi, L. (2019). Quantitative Structure-Activity Relationship Study of Antioxidant Tripeptides Based on Model Population Analysis. *International Journal of Molecular Sciences*, *20*(4), 995. https://doi.org/10/gnmdb7

Deng, B., Ni, X., Zhai, Z., Tang, T., Tan, C., Yan, Y., Deng, J., & Yin, Y. (2017). New Quantitative Structure–Activity Relationship Model for Angiotensin-Converting Enzyme Inhibitory Dipeptides Based on Integrated Descriptors. *Journal of Agricultural and Food Chemistry*, *65*(44), 9774–9781. https://doi.org/10.1021/acs.jafc.7b03367

Devita, L., Lioe, H. N., Nurilmala, M., & Suhartono, M. T. (2021). The Bioactivity Prediction of Peptides from Tuna Skin Collagen Using Integrated Method Combining In Vitro and In Silico. *Foods*, *10*(11), Article 11. https://doi.org/10.3390/foods10112739

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*. https://doi.org/10.48550/arXiv.1810.04805

Du, Z., Ding, X., Hsu, W., Munir, A., Xu, Y., & Li, Y. (2024). pLM4ACE: A protein language model based predictor for antihypertensive peptide screening. *Food Chemistry*, *431*, 137162. https://doi.org/10.1016/j.foodchem.2023.137162

Du, Z., Ding, X., Xu, Y., & Li, Y. (2023). UniDL4BioPep: A universal deep learning architecture for binary classification in peptide bioactivity. *Briefings in Bioinformatics*, 1–10. https://doi.org/10.1093/bib/bbad135

Du, Z., & Li, Y. (2022a). Computer-Aided Approaches for Screening Antioxidative Dipeptides and Application to Sorghum Proteins. *ACS Food Science & Technology*. https://doi.org/10.1021/acsfoodscitech.2c00286

Du, Z., & Li, Y. (2022b). Review and perspective on bioactive peptides: A roadmap for research, development, and future opportunities. *Journal of Agriculture and Food Research*, *9*, 100353. https://doi.org/10.1016/j.jafr.2022.100353

Du, Z., Tian, W., Tilley, M., Wang, D., Zhang, G., & Li, Y. (2022). Quantitative assessment of wheat quality using near-infrared spectroscopy: A comprehensive review. *Comprehensive Reviews in Food Science and Food Safety*, *21*(3), 2956–3009. https://doi.org/10.1111/1541-4337.12958

Du, Z., Wang, D., & Li, Y. (2022). Comprehensive Evaluation and Comparison of Machine Learning Methods in QSAR Modeling of Antioxidant Tripeptides. *ACS Omega*, acsomega.2c03062. https://doi.org/10.1021/acsomega.2c03062

Du, Z., Zeng, X., Li, X., Ding, X., Cao, J., & Jiang, W. (2020). Recent advances in imaging techniques for bruise detection in fruits and vegetables. *Trends in Food Science & Technology*, *99*, 133–141. https://doi.org/10.1016/j.tifs.2020.02.024

Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L.-P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., Wiewiora, R. P., Brooks, B. R., & Pande, V. S. (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology*, *13*(7), e1005659. https://doi.org/10.1371/journal.pcbi.1005659

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., & Rost, B. (2022). ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381

Evangelista Falcon, W., Ellingson, S. R., Smith, J. C., & Baudry, J. (2019). Ensemble Docking in Drug Discovery: How Many Protein Configurations from Molecular Dynamics Simulations are Needed To Reproduce Known Ligand Binding? *The Journal of Physical Chemistry B*, *123*(25), 5189–5195. https://doi.org/10.1021/acs.jpcb.8b11491

Feng, L., Tu, M., Qiao, M., Fan, F., Chen, H., Song, W., & Du, M. (2018). Thrombin inhibitory peptides derived from Mytilus edulis proteins: Identification, molecular docking and in silico prediction of toxicity. *European Food Research and Technology*, *244*(2), 207–217. https://doi.org/10.1007/s00217-017-2946-7

Feng, T., Li, M., Zhou, J., Zhuang, H., Chen, F., Ye, R., Campanella, O., & Fang, Z. (2015). Application of molecular dynamics simulation in food carbohydrate research—A review. *Innovative Food Science & Emerging Technologies*, *31*, 1–13. https://doi.org/10.1016/j.ifset.2015.06.015

Feng, Y., Wang, Z., Chen, J., Li, H., Wang, Y., Ren, D.-F., & Lu, J. (2021). Separation, identification, and molecular docking of tyrosinase inhibitory peptides from the hydrolysates of defatted walnut (Juglans regia L.) meal. *Food Chemistry*, *353*, 129471. https://doi.org/10.1016/j.foodchem.2021.129471

Ferreira, L., dos Santos, R., Oliva, G., & Andricopulo, A. (2015). Molecular Docking and Structure-Based Drug Design Strategies. *Molecules*, *20*(7), 13384–13421. https://doi.org/10.3390/molecules200713384

FitzGerald, R. J., Cermeño, M., Khalesi, M., Kleekayai, T., & Amigo-Benavent, M. (2020). Application of in silico approaches for the generation of milk protein-derived bioactive peptides. *Journal of Functional Foods*, *64*, 103636. https://doi.org/10.1016/j.jff.2019.103636

Forli, S., Huey, R., Pique, M. E., Sanner, M. F., Goodsell, D. S., & Olson, A. J. (2016). Computational protein–ligand docking and virtual drug screening with the AutoDock suite. *Nature Protocols*, *11*(5), 905–919. https://doi.org/10.1038/nprot.2016.051

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., & Shenkin, P. S. (2004). Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of Medicinal Chemistry*, *47*(7), 1739–1749. https://doi.org/10.1021/jm0306430

Fu, H., Cai, W., Hénin, J., Roux, B., & Chipot, C. (2017). New Coarse Variables for the Accurate Determination of Standard Binding Free Energies. *Journal of Chemical Theory and Computation*, *13*(11), 5173–5178. https://doi.org/10.1021/acs.jctc.7b00791

Fu, H., Chen, H., Cai, W., Shao, X., & Chipot, C. (2021). BFEE2: Automated, Streamlined, and Accurate Absolute Binding Free-Energy Calculations. *Journal of Chemical Information and Modeling*, *61*(5), 2116–2123. https://doi.org/10.1021/acs.jcim.1c00269

Fu, H., Gumbart, J. C., Chen, H., Shao, X., Cai, W., & Chipot, C. (2018). BFEE: A User-Friendly Graphical Interface Facilitating Absolute Binding Free-Energy Calculations. *Journal of Chemical Information and Modeling*, *58*(3), 556–560. https://doi.org/10.1021/acs.jcim.7b00695

Fu, Y., Young, J. F., Løkke, M. M., Lametsch, R., Aluko, R. E., & Therkildsen, M. (2016). Revalorisation of bovine collagen as a potential precursor of angiotensin I-converting enzyme (ACE) inhibitory peptides based on in silico and in vitro protein digestions. *Journal of Functional Foods*, *24*, 196–206. https://doi.org/10/f8sgc4

Gagnon, J. K., Law, S. M., & Brooks, C. L. (2016). Flexible CDOCKER: Development and application of a pseudo-explicit structure-based docking method within CHARMM: Adding Receptor Flexibility Improves Protein-Ligand Docking Within CDOCKER. *Journal of Computational Chemistry*, *37*(8), 753–762. https://doi.org/10.1002/jcc.24259

Garg, S., Apostolopoulos, V., Nurgali, K., & Mishra, V. K. (2018). Evaluation of in silico approach for prediction of presence of opioid peptides in wheat. *Journal of Functional Foods*, *41*, 34–40. https://doi.org/10.1016/j.jff.2017.12.022

Garzón, A. G., Veras, F. F., Brandelli, A., & Drago, S. R. (2022). Purification, identification and in silico studies of antioxidant, antidiabetogenic and antibacterial peptides obtained from sorghum spent grain hydrolysate. *LWT*, *153*, 112414. https://doi.org/10.1016/j.lwt.2021.112414

Genheden, S., Kuhn, O., Mikulskis, P., Hoffmann, D., & Ryde, U. (2012). The Normal-Mode Entropy in the MM/GBSA Method: Effect of System Truncation, Buffer Region, and Dielectric Constant. *Journal of Chemical Information and Modeling*, *52*(8), 2079–2088. https://doi.org/10.1021/ci3001919

Gomez, H. L. R., Peralta, J. P., Tejano, L. A., & Chang, Y.-W. (2019). In Silico and In Vitro Assessment of Portuguese Oyster (Crassostrea angulata) Proteins as Precursor of Bioactive Peptides. *International Journal of Molecular Sciences*, *20*(20), Article 20. https://doi.org/10.3390/ijms20205191

Gu, Y., Majumder, K., & Wu, J. (2011). QSAR-aided in silico approach in evaluation of food proteins as precursors of ACE inhibitory peptides. *Food Research International*, *44*(8), 2465–2474. https://doi.org/10.1016/j.foodres.2011.01.051

Guan, X., & Liu, J. (2019). QSAR Study of Angiotensin I-Converting Enzyme Inhibitory Peptides Using SVHEHS Descriptor and OSC-SVM. *International Journal of Peptide Research and Therapeutics*, *25*(1), 247–256. https://doi.org/10/gnmdct

Guedes, I. A., Pereira, F. S. S., & Dardenne, L. E. (2018). Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges. *Frontiers in Pharmacology*, *9*, 1089. https://doi.org/10.3389/fphar.2018.01089

Gumbart, J. C., Roux, B., & Chipot, C. (2013). Standard Binding Free Energies from Computer Simulations: What Is the Best Strategy? *Journal of Chemical Theory and Computation*, *9*(1), 794–802. https://doi.org/10.1021/ct3008099

Gunalan, S., Somarathinam, K., Bhattacharya, J., Srinivasan, S., Jaimohan, S. M., Manoharan, R., Ramachandran, S., Kanagaraj, S., & Kothandan, G. (2020). Understanding the dual mechanism of bioactive peptides targeting the enzymes involved in Renin Angiotensin System (RAS): An *in-silico* approach. *Journal of Biomolecular Structure and Dynamics*, *38*(17), 5044–5061. https://doi.org/10.1080/07391102.2019.1695668

Hollingsworth, S. A., & Dror, R. O. (2018). Molecular Dynamics Simulation for All. *Neuron*, *99*(6), 1129–1143. https://doi.org/10.1016/j.neuron.2018.08.011

Houston, D. R., & Walkinshaw, M. D. (2013). Consensus Docking: Improving the Reliability of Docking in a Virtual Screening Context. *Journal of Chemical Information and Modeling*, *53*(2), 384–390. https://doi.org/10.1021/ci300399w

Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B. L., Grubmüller, H., & MacKerell, A. D. (2017). CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nature Methods*, *14*(1), Article 1. https://doi.org/10.1038/nmeth.4067

Huang, S.-Y. (2018). Comprehensive assessment of flexible-ligand docking algorithms: Current effectiveness and challenges. *Briefings in Bioinformatics*, *19*(5), 982–994. https://doi.org/10.1093/bib/bbx030

Iram, D., Sansi, M. S., Zanab, S., Vij, S., Ashutosh, & Meena, S. (2022). In silico identification of antidiabetic and hypotensive potential bioactive peptides from the sheep milk proteins—A molecular docking study. *Journal of Food Biochemistry*. https://doi.org/10.1111/jfbc.14137

Irwin, R., Dimitriadis, S., He, J., & Bjerrum, E. J. (2022). Chemformer: A pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, *3*(1), 015022. https://doi.org/10.1088/2632-2153/ac3ffb

Ishiguro, S., Upreti, D., Bassette, M., Singam, E. R. A., Thakkar, R., Loyd, M., Inui, M., Comer, J., & Tamura, M. (2022). Local immune checkpoint blockade therapy by an adenovirus encoding a novel PD-L1 inhibitory peptide inhibits the growth of colon carcinoma in immunocompetent mice. *Translational Oncology*, *16*, 101337. https://doi.org/10.1016/j.tranon.2021.101337

Iwaniak, A., Darewicz, M., Mogut, D., & Minkiewicz, P. (2019). Elucidation of the role of in silico methodologies in approaches to studying bioactive peptides derived from foods. *Journal of Functional Foods*, *61*, 103486. https://doi.org/10.1016/j.jff.2019.103486

Iwaniak, A., Minkiewicz, P., Pliszka, M., Mogut, D., & Darewicz, M. (2020). Characteristics of Biopeptides Released In Silico from Collagens Using Quantitative Parameters. *Foods*, *9*(7), 965. https://doi.org/10/gnmdcc

Jagger, B. R., Kochanek, S. E., Haldar, S., Amaro, R. E., & Mulholland, A. J. (2020). Multiscale simulation approaches to modeling drug–protein binding. *Current Opinion in Structural Biology*, *61*, 213–221. https://doi.org/10.1016/j.sbi.2020.01.014

Ji, D., Udenigwe, C. C., & Agyei, D. (2019). Antioxidant peptides encrypted in flaxseed proteome: An in silico assessment. *Food Science and Human Wellness*, *8*(3), 306–314. https://doi.org/10.1016/j.fshw.2019.08.002

Ji, D., Xu, M., Udenigwe, C. C., & Agyei, D. (2020). Physicochemical characterisation, molecular docking, and drug-likeness evaluation of hypotensive peptides encrypted in flaxseed proteome. *Current Research in Food Science*, *3*, 41–50. https://doi.org/10.1016/j.crfs.2020.03.001

Jorgensen, W. L., Maxwell, D. S., & Tirado-Rives, J. (1996). Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society*, *118*(45), 11225–11236. https://doi.org/10.1021/ja9621760

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), Article 7873. https://doi.org/10.1038/s41586-021-03819-2

Kalyan, G., Junghare, V., Bhattacharya, S., & Hazra, S. (2021). Understanding structure-based dynamic interactions of antihypertensive peptides extracted from food sources. *Journal of Biomolecular Structure and Dynamics*, *39*(2), 635–649. https://doi.org/10.1080/07391102.2020.1715836

Kalyan, G., Junghare, V., Khan, M. F., Pal, S., Bhattacharya, S., Guha, S., Majumder, K., Chakrabarty, S., & Hazra, S. (2021). Anti-hypertensive Peptide Predictor: A Machine Learning-Empowered Web Server for Prediction of Food-Derived Peptides with Potential Angiotensin-Converting Enzyme-I Inhibitory Activity. *Journal of Agricultural and Food Chemistry*, *69*(49), 14995–15004. https://doi.org/10.1021/acs.jafc.1c04555

Karaś, M. (2019). Influence of physiological and chemical factors on the absorption of bioactive peptides. *International Journal of Food Science & Technology*, *54*(5), 1486–1496. https://doi.org/10.1111/ijfs.14054

Karplus, M., & Kushick, J. N. (1981). Method for estimating the configurational entropy of macromolecules. *Macromolecules*, *14*(2), 325–332. https://doi.org/10.1021/ma50003a019

Kartal, C., Kaplan Türköz, B., & Otles, S. (2020). Prediction, identification and evaluation of bioactive peptides from tomato seed proteins using in silico approach. *Journal of Food Measurement and Characterization*, *14*(4), 1865–1883. https://doi.org/10.1007/s11694-020-00434-z

Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D. A., & Cheatham, T. E. (2000). Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Accounts of Chemical Research*, *33*(12), 889–897. https://doi.org/10.1021/ar000033j

Kondratyuk, N., Nikolskiy, V., Pavlov, D., & Stegailov, V. (2021). GPU-accelerated molecular dynamics: State-of-art software performance and porting from Nvidia CUDA to AMD HIP. *The International Journal of High Performance Computing Applications*, *35*(4), 312–324. https://doi.org/10.1177/10943420211008288

Korb, O., Olsson, T. S. G., Bowden, S. J., Hall, R. J., Verdonk, M. L., Liebeschuetz, J. W., & Cole, J. C. (2012). Potential and Limitations of Ensemble Docking. *Journal of Chemical Information and Modeling*, *52*(5), 1262–1274. https://doi.org/10.1021/ci2005934

Kutzner, C., Páll, S., Fechner, M., Esztermann, A., de Groot, B. L., & Grubmüller, H. (2019). More bang for your buck: Improved use of GPU nodes for GROMACS 2018. *Journal of Computational Chemistry*, *40*(27), 2418–2431. https://doi.org/10.1002/jcc.26011

Li, L., Liu, J., Nie, S., Ding, L., Wang, L., Liu, J., Liu, W., & Zhang, T. (2017). Direct inhibition of Keap1–Nrf2 interaction by egg-derived peptides DKK and DDW revealed by molecular docking and fluorescence polarization. *RSC Advances*, *7*(56), 34963–34971. https://doi.org/10.1039/C7RA04352J

Li, X., Li, Y., Cheng, T., Liu, Z., & Wang, R. (2010). Evaluation of the performance of four molecular docking programs on a diverse set of protein-ligand complexes. *Journal of Computational Chemistry*, *31*(11), 2109–2125. https://doi.org/10.1002/jcc.21498

Li, X., Pan, F., Yang, Z., Gao, F., Li, J., Zhang, F., & Wang, T. (2022). Construction of QSAR model based on cysteine-containing dipeptides and screening of natural tyrosinase inhibitors. *Journal of Food Biochemistry*, e14338. https://doi.org/10.1111/jfbc.14338

Li, Y., Zhang, S., Bao, Z., Sun, N., & Lin, S. (2022). Exploring the activation mechanism of alcalase activity with pulsed electric field treatment: Effects on enzyme activity, spatial conformation, molecular dynamics simulation and molecular docking parameters.

Innovative Food Science & Emerging Technologies, *76*, 102918. https://doi.org/10.1016/j.ifset.2022.102918

Liang, F., Shi, Y., Shi, J., Zhang, T., & Zhang, R. (2021). A novel Angiotensin-I-converting enzyme (ACE) inhibitory peptide IAF (Ile-Ala-Phe) from pumpkin seed proteins: In silico screening, inhibitory activity, and molecular mechanisms. *European Food Research and Technology*, *247*(9), 2227–2237. https://doi.org/10.1007/s00217-021-03783-1

Liao, W., Bhullar, K. S., Chakrabarti, S., Davidge, S. T., & Wu, J. (2018). Egg White-Derived Tripeptide IRW (Ile-Arg-Trp) Is an Activator of Angiotensin Converting Enzyme 2. *Journal of Agricultural and Food Chemistry*, *66*(43), 11330–11336. https://doi.org/10.1021/acs.jafc.8b03501

Liao, W., Fan, H., Davidge, S. T., & Wu, J. (2019). Egg White–Derived Antihypertensive Peptide IRW (Ile-Arg-Trp) Reduces Blood Pressure in Spontaneously Hypertensive Rats via the ACE2/Ang (1-7)/Mas Receptor Axis. *Molecular Nutrition & Food Research*, *63*(9), 1900063. https://doi.org/10.1002/mnfr.201900063

Lin, F.-Y., Huang, J., Pandey, P., Rupakheti, C., Li, J., Roux, B., & MacKerell, A. D. Jr. (2020). Further Optimization and Validation of the Classical Drude Polarizable Protein Force Field. *Journal of Chemical Theory and Computation*, *16*(5), 3221–3239. https://doi.org/10.1021/acs.jctc.0c00057

Lin, K., Zhang, L., Han, X., & Cheng, D. (2017). Novel angiotensin I-converting enzyme inhibitory peptides from protease hydrolysates of Qula casein: Quantitative structure-activity relationship modeling and molecular docking study. *Journal of Functional Foods*, *32*, 266–277. https://doi.org/10.1016/j.jff.2017.03.008

Lin, K., Zhang, L., Han, X., Xin, L., Meng, Z., Gong, P., & Cheng, D. (2018). Yak milk casein as potential precursor of angiotensin I-converting enzyme inhibitory peptides based on in silico proteolysis. *Food Chemistry*, *254*, 340–347. https://doi.org/10.1016/j.foodchem.2018.02.051

Liu, Y., Grimm, M., Dai, W., Hou, M., Xiao, Z.-X., & Cao, Y. (2020). CB-Dock: A web server for cavity detection-guided protein–ligand blind docking. *Acta Pharmacologica Sinica*, *41*(1), Article 1. https://doi.org/10.1038/s41401-019-0228-6

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, *379*(6637), 1123–1130. https://doi.org/10.1126/science.ade2574

Luo, F., Fu, Y., Ma, L., Dai, H., Wang, H., Chen, H., Zhu, H., Yu, Y., Hou, Y., & Zhang, Y. (2022). Exploration of Dipeptidyl Peptidase-IV (DPP-IV) Inhibitory Peptides from Silkworm Pupae (*Bombyx mori*) Proteins Based on *In Silico* and *In Vitro* Assessments. *Journal of Agricultural and Food Chemistry*, *70*(12), 3862–3871. https://doi.org/10.1021/acs.jafc.1c08225

MacKerell, A. D. Jr., Bashford, D., Bellott, M., Dunbrack, R. L. Jr., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., … Karplus, M. (1998). All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *The Journal of Physical Chemistry B*, *102*(18), 3586–3616. https://doi.org/10.1021/jp973084f

Mahmoodi-Reihani, M., Abbasitabar, F., & Zare-Shahabadi, V. (2020). In Silico Rational Design and Virtual Screening of Bioactive Peptides Based on QSAR Modeling. *ACS Omega*, *5*(11), 5951–5958. https://doi.org/10/gnmdcr

Maia, E. H. B., Assis, L. C., de Oliveira, T. A., da Silva, A. M., & Taranto, A. G. (2020). Structure-Based Virtual Screening: From Classical to Artificial Intelligence. *Frontiers in Chemistry*, *8*, 343. https://doi.org/10.3389/fchem.2020.00343

Maia, E. H. B., Medaglia, L. R., da Silva, A. M., & Taranto, A. G. (2020). Molecular Architect: A User-Friendly Workflow for Virtual Screening. *ACS Omega*, *5*(12), 6628–6640. https://doi.org/10.1021/acsomega.9b04403

Majid, A., Lakshmikanth, M., Lokanath, N. K., & Poornima Priyadarshini, C. G. (2022). Generation, characterization and molecular binding mechanism of novel dipeptidyl peptidase-4 inhibitory peptides from sorghum bicolor seed protein. *Food Chemistry*, *369*, 130888. https://doi.org/10.1016/j.foodchem.2021.130888

Majumder, K., Chakrabarti, S., Davidge, S. T., & Wu, J. (2013). Structure and Activity Study of Egg Protein Ovotransferrin Derived Peptides (IRW and IQW) on Endothelial Inflammatory Response and Oxidative Stress. *Journal of Agricultural and Food Chemistry*, *61*(9), 2120–2129. https://doi.org/10.1021/jf3046076

Majumder, K., & Wu, J. (2010). A new approach for identification of novel antihypertensive peptides from egg proteins by QSAR and bioinformatics. *Food Research International*, *43*(5), 1371–1378. https://doi.org/10.1016/j.foodres.2010.04.027

Marcet, I., Delgado, J., Díaz, N., Rendueles, M., & Díaz, M. (2022). Peptides recovery from egg yolk lipovitellins by ultrafiltration and their in silico bioactivity analysis. *Food Chemistry*, *379*, 132145. https://doi.org/10.1016/j.foodchem.2022.132145

Martini, S., Cattivelli, A., Conte, A., & Tagliazucchi, D. (2022). Application of a Combined Peptidomics and In Silico Approach for the Identification of Novel Dipeptidyl Peptidase-IV-Inhibitory Peptides in In Vitro Digested Pinto Bean Protein Extract. *Current Issues in Molecular Biology*, *44*(1), Article 1. https://doi.org/10.3390/cimb44010011

Martini, S., Solieri, L., Cattivelli, A., Pizzamiglio, V., & Tagliazucchi, D. (2021). An Integrated Peptidomics and In Silico Approach to Identify Novel Anti-Diabetic Peptides in Parmigiano-Reggiano Cheese. *Biology*, *10*, 563. https://doi.org/10.3390/biology10060563

Meng, X.-Y., Zhang, H.-X., Mezei, M., & Cui, M. (2011). Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Current Computer Aided-Drug Design*, *7*(2), 146–157. https://doi.org/10.2174/157340911795677602

Miller, B. R. I., McGee, T. D. Jr., Swails, J. M., Homeyer, N., Gohlke, H., & Roitberg, A. E. (2012). MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. *Journal of Chemical Theory and Computation*, *8*(9), 3314–3321. https://doi.org/10.1021/ct300418h

Minkiewicz, Iwaniak, & Darewicz. (2019). BIOPEP-UWM Database of Bioactive Peptides: Current Opportunities. *International Journal of Molecular Sciences*, *20*(23), 5978. https://doi.org/10/gnmt2w

Mobley, D. L., & Dill, K. A. (2009). Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get." *Structure*, *17*(4), 489–498. https://doi.org/10.1016/j.str.2009.02.010

Mooney, C., Haslam, N. J., Pollastri, G., & Shields, D. C. (2012). Towards the Improved Discovery and Design of Functional Peptides: Common Features of Diverse Classes Permit Generalized Prediction of Bioactivity. *PLOS ONE*, *7*(10), e45012. https://doi.org/10.1371/journal.pone.0045012

Mudgil, P., Baby, B., Ngoh, Y.-Y., Kamal, H., Vijayan, R., Gan, C.-Y., & Maqsood, S. (2019). Molecular binding mechanism and identification of novel anti-hypertensive and anti-inflammatory bioactive peptides from camel milk protein hydrolysates. *LWT*, *112*, 108193. https://doi.org/10.1016/j.lwt.2019.05.091

Nardo, A. E., Añón, M. C., & Quiroga, A. V. (2020). Identification of renin inhibitors peptides from amaranth proteins by docking protocols. *Journal of Functional Foods*, *64*, 103683. https://doi.org/10.1016/j.jff.2019.103683

Nong, N. T. P., & Hsu, J.-L. (2022). Bioactive Peptides: An Understanding from Current Screening Methodology. *Processes*, *10*(6), Article 6. https://doi.org/10.3390/pr10061114

Nongonierma, A. B., & FitzGerald, R. J. (2016). Structure activity relationship modelling of milk protein-derived peptides with dipeptidyl peptidase IV (DPP-IV) inhibitory activity. *Peptides*, *79*, 1–7. https://doi.org/10.1016/j.peptides.2016.03.005

Nongonierma, A. B., & FitzGerald, R. J. (2018). Enhancing bioactive peptide release and identification using targeted enzymatic hydrolysis of milk proteins. *Analytical and Bioanalytical Chemistry*, *410*(15), 3407–3423. https://doi.org/10.1007/s00216-017-0793-9

Nongonierma, A. B., Paolella, S., Mudgil, P., Maqsood, S., & FitzGerald, R. J. (2018). Identification of novel dipeptidyl peptidase IV (DPP-IV) inhibitory peptides in camel milk protein hydrolysates. *Food Chemistry*, *244*, 340–348. https://doi.org/10.1016/j.foodchem.2017.10.033

Olsen, T. H., Yesiltas, B., Marin, F. I., Pertseva, M., García-Moreno, P. J., Gregersen, S., Overgaard, M. T., Jacobsen, C., Lund, O., Hansen, E. B., & Marcatili, P. (2020). AnOxPePred: Using deep learning for the prediction of antioxidative properties of peptides. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-78319-w

Páll, S., Zhmurov, A., Bauer, P., Abraham, M., Lundborg, M., Gray, A., Hess, B., & Lindahl, E. (2020). Heterogeneous parallelization and acceleration of molecular dynamics simulations in GROMACS. *The Journal of Chemical Physics*, *153*(13), 134110. https://doi.org/10.1063/5.0018516

Panjaitan, F. C. A., Gomez, H. L. R., & Chang, Y.-W. (2018). In Silico Analysis of Bioactive Peptides Released from Giant Grouper (Epinephelus lanceolatus) Roe Proteins Identified by Proteomics Approach. *Molecules*, *23*(11), Article 11. https://doi.org/10.3390/molecules23112910

Panyayai, T., Ngamphiw, C., Tongsima, S., Mhuantong, W., Limsripraphan, W., Choowongkomon, K., & Sawatdichaikul, O. (2019). FeptideDB: A web application for new bioactive peptides from food protein. *Heliyon*, *5*(7), e02076. https://doi.org/10.1016/j.heliyon.2019.e02076

Panyayai, T., Sangsawad, P., Pacharawongsakda, E., Sawatdichaikul, O., Tongsima, S., & Choowongkomon, K. (2018). The potential peptides against angiotensin-I converting enzyme through a virtual tripeptide-constructing library. *Computational Biology and Chemistry*, *77*, 207–213. https://doi.org/10.1016/j.compbiolchem.2018.10.001

Patel, H. M., Noolvi, M. N., Sharma, P., Jaiswal, V., Bansal, S., Lohan, S., Kumar, S. S., Abbot, V., Dhiman, S., & Bhardwaj, V. (2014). Quantitative structure–activity relationship (QSAR) studies as strategic approach in drug discovery. *Medicinal Chemistry Research*, *23*(12), 4991–5007. https://doi.org/10.1007/s00044-014-1072-3

Pearman, N. A., Ronander, E., Smith, A. M., & Morris, G. A. (2020). The identification and characterisation of novel bioactive peptides derived from porcine liver. *Current Research in Food Science*, *3*, 314–321. https://doi.org/10.1016/j.crfs.2020.11.002

Phillips, J. C., Hardy, D. J., Maia, J. D. C., Stone, J. E., Ribeiro, J. V., Bernardi, R. C., Buch, R., Fiorin, G., Hénin, J., Jiang, W., McGreevy, R., Melo, M. C. R., Radak, B. K., Skeel, R. D., Singharoy, A., Wang, Y., Roux, B., Aksimentiev, A., Luthey-Schulten, Z., … Tajkhorshid, E. (2020). Scalable molecular dynamics on CPU and GPU architectures with NAMD. *The Journal of Chemical Physics*, *153*(4), 044130. https://doi.org/10.1063/5.0014475

Pissurlenkar, R. R. S., Shaikh, M. S., & Coutinho, E. C. (2007). 3D-QSAR studies of Dipeptidyl peptidase IV inhibitors using a docking based alignment. *Journal of Molecular Modeling*, *13*(10), 1047–1071. https://doi.org/10.1007/s00894-007-0227-2

Ploetz, E. A., Karunaweera, S., & Smith, P. E. (2021). Kirkwood–Buff-Derived Force Field for Peptides and Proteins: Applications of KBFF20. *Journal of Chemical Theory and Computation*, *17*(5), 2991–3009. https://doi.org/10.1021/acs.jctc.1c00076

Pooja, K., Rani, S., & Prakash, B. (2017). In silico approaches towards the exploration of rice bran proteins-derived angiotensin-I-converting enzyme inhibitory peptides. *International Journal of Food Properties*, *20*(sup2), 2178–2191. https://doi.org/10.1080/10942912.2017.1368552

Preto, J., & Gentile, F. (2019). Assessing and improving the performance of consensus docking strategies using the DockBox package. *Journal of Computer-Aided Molecular Design*, *33*(9), 817–829. https://doi.org/10.1007/s10822-019-00227-7

Pripp, A. H. (2007). Docking and virtual screening of ACE inhibitory dipeptides. *European Food Research and Technology*, *225*(3–4), 589–592. https://doi.org/10.1007/s00217-006-0450-6

Qian, Y., Liang, Y., Liu, W., & Liang, G. (2017). Comprehensive comparison of twenty structural characterization scales applied as QSAM of antimicrobial dodecapeptides derived from Bac2A against P. aeruginosa. *Journal of Molecular Graphics and Modelling*, *71*, 88–95. https://doi.org/10.1016/j.jmgm.2016.11.003

Rackers, J. A., Wang, Q., Liu, C., Piquemal, J.-P., Ren, P., & Ponder, J. W. (2016). An optimized charge penetration model for use with the AMOEBA force field. *Physical Chemistry Chemical Physics*, *19*(1), 276–291. https://doi.org/10.1039/C6CP06017J

Ren, X., Shi, Y.-S., Zhang, Y., Liu, B., Zhang, L.-H., Peng, Y.-B., & Zeng, R. (2018). Novel Consensus Docking Strategy to Improve Ligand Pose Prediction. *Journal of Chemical Information and Modeling*, *58*(8), 1662–1668. https://doi.org/10.1021/acs.jcim.8b00329

Rizzi, A., Murkli, S., McNeill, J. N., Yao, W., Sullivan, M., Gilson, M. K., Chiu, M. W., Isaacs, L., Gibb, B. C., Mobley, D. L., & Chodera, J. D. (2018). Overview of the SAMPL6 host–guest binding affinity prediction challenge. *Journal of Computer-Aided Molecular Design*, *32*(10), 937–963. https://doi.org/10.1007/s10822-018-0170-6

Rosignoli, S., & Paiardini, A. (2022). DockingPie: A consensus docking plugin for PyMOL. *Bioinformatics*, *38*(17), 4233–4234. https://doi.org/10.1093/bioinformatics/btac452

Salmaso, V., & Moro, S. (2018). Bridging Molecular Docking to Molecular Dynamics in Exploring Ligand-Protein Recognition Process: An Overview. *Frontiers in Pharmacology*, *9*. https://doi.org/10.3389/fphar.2018.00923

Sansi, M. S., Iram, D., Zanab, S., Vij, S., Puniya, A. K., Singh, A., Ashutosh, & Meena, S. (2022). Antimicrobial bioactive peptides from goat Milk proteins: In silico prediction and analysis. *Journal of Food Biochemistry*. https://doi.org/10.1111/jfbc.14311

Sayd, T., Dufour, C., Chambon, C., Buffière, C., Remond, D., & Santé-Lhoutellier, V. (2018). Combined in vivo and in silico approaches for predicting the release of bioactive peptides

from meat digestion. *Food Chemistry*, *249*, 111–118. https://doi.org/10.1016/j.foodchem.2018.01.013

Senapathi, T., Suruzhon, M., Barnett, C. B., Essex, J., & Naidoo, K. J. (2020). BRIDGE: An Open Platform for Reproducible High-Throughput Free Energy Simulations. *Journal of Chemical Information and Modeling*, *60*(11), 5290–5295. https://doi.org/10.1021/acs.jcim.0c00206

Shaw, D. E., Dror, R. O., Salmon, J. K., Grossman, J. P., Mackenzie, K. M., Bank, J. A., Young, C., Deneroff, M. M., Batson, B., Bowers, K. J., Chow, E., Eastwood, M. P., Ierardi, D. J., Klepeis, J. L., Kuskin, J. S., Larson, R. H., Lindorff-Larsen, K., Maragakis, P., Moraes, M. A., … Towles, B. (2009). Millisecond-scale molecular dynamics simulations on Anton. *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, 1–11. https://doi.org/10.1145/1654059.1654126

Shen, Y., Hong, S., Singh, G., Koppel, K., & Li, Y. (2022). Improving functional properties of pea protein through "green" modifications using enzymes and polysaccharides. *Food Chemistry*, *385*, 132687. https://doi.org/10.1016/j.foodchem.2022.132687

Shen, Y., & Li, Y. (2021). Acylation modification and/or guar gum conjugation enhanced functional properties of pea protein isolate. *Food Hydrocolloids*, *117*, 106686.

Shi, Y., Xia, Z., Zhang, J., Best, R., Wu, C., Ponder, J. W., & Ren, P. (2013). Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. *Journal of Chemical Theory and Computation*, *9*(9), 4046–4063. https://doi.org/10.1021/ct4003702

Sun, H., Chang, Q., Liu, L., Chai, K., Lin, G., Huo, Q., Zhao, Z., & Zhao, Z. (2017). High-Throughput and Rapid Screening of Novel ACE Inhibitory Peptides from Sericin Source and Inhibition Mechanism by Using in Silico and in Vitro Prescriptions. *Journal of Agricultural and Food Chemistry*, *65*(46), 10020–10028. https://doi.org/10.1021/acs.jafc.7b04043

Sun, H., Duan, L., Chen, F., Liu, H., Wang, Z., Pan, P., Zhu, F., Zhang, J. Z. H., & Hou, T. (2018). Assessing the performance of MM/PBSA and MM/GBSA methods. 7. Entropy effects on the performance of end-point binding free energy calculation approaches. *Physical Chemistry Chemical Physics*, *20*(21), 14450–14460. https://doi.org/10.1039/C7CP07623A

Tahir, R. A., Bashir, A., Yousaf, M. N., Ahmed, A., Dali, Y., Khan, S., & Sehgal, S. A. (2020). In Silico identification of angiotensin-converting enzyme inhibitory peptides from MRJP1. *PLOS ONE*, *15*(2), e0228265. https://doi.org/10.1371/journal.pone.0228265

Tao, X., Huang, Y., Wang, C., Chen, F., Yang, L., Ling, L., Che, Z., & Chen, X. (2020). Recent developments in molecular docking technology applied in food science: A review. *International Journal of Food Science & Technology*, *55*(1), 33–45. https://doi.org/10.1111/ijfs.14325

Thakkar, R., Upreti, D., Ishiguro, S., Tamura, M., & Comer, J. (2023). Computational design of a cyclic peptide that inhibits the CTLA4 immune checkpoint. *RSC Medicinal Chemistry*, 10.1039.D2MD00409G. https://doi.org/10.1039/D2MD00409G

Thi Phan, L., Woo Park, H., Pitti, T., Madhavan, T., Jeon, Y.-J., & Manavalan, B. (2022). MLACP 2.0: An updated machine learning tool for anticancer peptide prediction. *Computational and Structural Biotechnology Journal*, *20*, 4473–4480. https://doi.org/10.1016/j.csbj.2022.07.043

Tian, C., Kasavajhala, K., Belfon, K. A. A., Raguette, L., Huang, H., Migues, A. N., Bickel, J., Wang, Y., Pincay, J., Wu, Q., & Simmerling, C. (2020). ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *Journal of Chemical Theory and Computation*, *16*(1), 528–552. https://doi.org/10.1021/acs.jctc.9b00591

Tian, F., Zhou, P., & Li, Z. (2007). T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides. *Journal of Molecular Structure*, *830*(1–3), 106–115. https://doi.org/10/dqr84q

Tonolo, F., Folda, A., Cesaro, L., Scalcon, V., Marin, O., Ferro, S., Bindoli, A., & Rigobello, M. P. (2020). Milk-derived bioactive peptides exhibit antioxidant activity through the Keap1-Nrf2 signaling pathway. *Journal of Functional Foods*, *64*, 103696. https://doi.org/10.1016/j.jff.2019.103696

Trott, O., & Olson, A. J. (2009). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, NA-NA. https://doi.org/10.1002/jcc.21334

Tu, M., Cheng, S., Lu, W., & Du, M. (2018). Advancement and prospects of bioinformatics analysis for studying bioactive peptides from food-derived protein: Sequence, structure, and functions. *TrAC Trends in Analytical Chemistry*, *105*, 7–17. https://doi.org/10.1016/j.trac.2018.04.005

Tu, M., Feng, L., Wang, Z., Qiao, M., Shahidi, F., Lu, W., & Du, M. (2017). Sequence analysis and molecular docking of antithrombotic peptides from casein hydrolysate by trypsin digestion. *Journal of Functional Foods*, *32*, 313–323. https://doi.org/10.1016/j.jff.2017.03.015

Tu, M., Liu, H., Zhang, R., Chen, H., Fan, F., Shi, P., Xu, X., Lu, W., & Du, M. (2018). Bioactive hydrolysates from casein: Generation, identification, and in silico toxicity and allergenicity prediction of peptides. *Journal of the Science of Food and Agriculture*, *98*(9), 3416–3426. https://doi.org/10.1002/jsfa.8854

Udenigwe, C. C. (2014). Bioinformatics approaches, prospects and challenges of food bioactive peptide research. *Trends in Food Science & Technology*, *36*(2), 137–143. https://doi.org/10.1016/j.tifs.2014.02.004

Ugur, I., Schroft, M., Marion, A., Glaser, M., & Antes, I. (2019). Predicting the bioactive conformations of macrocycles: A molecular dynamics-based docking procedure with DynaDock. *Journal of Molecular Modeling*, *25*(7), 197. https://doi.org/10.1007/s00894-019-4077-5

Uno, S., Kodama, D., Yukawa, H., Shidara, H., & Akamatsu, M. (2020). Quantitative analysis of the relationship between structure and antioxidant activity of tripeptides. *Journal of Peptide Science*, *26*(3). https://doi.org/10/gnmdb5

Valdés-Tresanco, M. S., Valdés-Tresanco, M. E., Valiente, P. A., & Moreno, E. (2021). gmx_MMPBSA: A New Tool to Perform End-State Free Energy Calculations with GROMACS. *Journal of Chemical Theory and Computation*, *17*(10), 6281–6291. https://doi.org/10.1021/acs.jctc.1c00645

Velez-Vega, C., & Gilson, M. K. (2013). Overcoming dissipation in the calculation of standard binding free energies by ligand extraction. *Journal of Computational Chemistry*, *34*(27), 2360–2371. https://doi.org/10.1002/jcc.23398

Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., & Taylor, R. D. (2003). Improved protein-ligand docking using GOLD. *Proteins: Structure, Function, and Bioinformatics*, *52*(4), 609–623. https://doi.org/10.1002/prot.10465

Vergara-Jaque, A., Fong, P., & Comer, J. (2017). Iodide Binding in Sodium-Coupled Cotransporters. *Journal of Chemical Information and Modeling*, *57*(12), 3043–3055. https://doi.org/10.1021/acs.jcim.7b00521

Vidal-Limon, A., Aguilar-Toalá, J. E., & Liceaga, A. M. (2022). Integration of Molecular Docking Analysis and Molecular Dynamics Simulations for Studying Food Proteins and Bioactive Peptides. *Journal of Agricultural and Food Chemistry*, *70*(4), 934–943. https://doi.org/10.1021/acs.jafc.1c06110

Vukic, V. R., Vukic, D. V., Milanovic, S. D., Ilicic, M. D., Kanuric, K. G., & Johnson, M. S. (2017). In silico identification of milk antihypertensive di- and tripeptides involved in angiotensin I–converting enzyme inhibitory activity. *Nutrition Research*, *46*, 22–30. https://doi.org/10.1016/j.nutres.2017.07.009

Waghu, F. H., Gopi, L., Barai, R. S., Ramteke, P., Nizami, B., & Idicula-Thomas, S. (2014). CAMP: Collection of sequences and structures of antimicrobial peptides. *Nucleic Acids Research*, *42*(Database issue), D1154–D1158. https://doi.org/10.1093/nar/gkt1157

Wang, F., & Zhou, B. (2020). Investigation of angiotensin-I-converting enzyme (ACE) inhibitory tri-peptides: A combination of 3D-QSAR and molecular docking simulations. *RSC Advances*, *10*(59), 35811–35819. https://doi.org/10.1039/D0RA05119E

Wang, Y.-T., Russo, D. P., Liu, C., Zhou, Q., Zhu, H., & Zhang, Y.-H. (2020). Predictive Modeling of Angiotensin I-Converting Enzyme Inhibitory Peptides Using Various Machine Learning Approaches. *Journal of Agricultural and Food Chemistry*, *68*(43), 12132–12140. https://doi.org/10/gnmdcn

Wen, C., Zhang, J., Zhang, H., Duan, Y., & Ma, H. (2020). Plant protein-derived antioxidant peptides: Isolation, identification, mechanism of action and application in food systems: A review. *Trends in Food Science & Technology*, *105*, 308–322. https://doi.org/10.1016/j.tifs.2020.09.019

Wen, L., Huang, L., Li, Y., Feng, Y., Zhang, Z., Xu, Z., Chen, M.-L., & Cheng, Y. (2021). New peptides with immunomodulatory activity identified from rice proteins through peptidomic and in silico analysis. *Food Chemistry*, *364*, 130357. https://doi.org/10.1016/j.foodchem.2021.130357

Wenhui, T., Shumin, H., Yongliang, Z., Liping, S., & Hua, Y. (2022). Identification of in vitro angiotensin-converting enzyme and dipeptidyl peptidase IV inhibitory peptides from draft beer by virtual screening and molecular docking. *Journal of the Science of Food and Agriculture*, *102*(3), 1085–1094. https://doi.org/10.1002/jsfa.11445

Woo, H.-J., & Roux, B. (2005). Calculation of absolute protein–ligand binding free energy from computer simulations. *Proceedings of the National Academy of Sciences*, *102*(19), 6825–6830. https://doi.org/10.1073/pnas.0409005102

Wu, J., Aluko, R. E., & Nakai, S. (2006). Structural Requirements of Angiotensin I-Converting Enzyme Inhibitory Peptides: Quantitative Structure−Activity Relationship Study of Di- and Tripeptides. *Journal of Agricultural and Food Chemistry*, *54*(3), 732–738. https://doi.org/10.1021/jf0512631

Wu, S., Qi, W., Su, R., Li, T., Lu, D., & He, Z. (2014). CoMFA and CoMSIA analysis of ACE-inhibitory, antimicrobial and bitter-tasting peptides. *European Journal of Medicinal Chemistry*, *84*, 100–106. https://doi.org/10/f6g45c

Xiao, C., Toldrá, F., Zhao, M., Zhou, F., Luo, D., Jia, R., & Mora, L. (2022). In vitro and in silico analysis of potential antioxidant peptides obtained from chicken hydrolysate produced using Alcalase. *Food Research International*, *157*, 111253. https://doi.org/10.1016/j.foodres.2022.111253

Xiong, G., Wu, Z., Yi, J., Fu, L., Yang, Z., Hsieh, C., Yin, M., Zeng, X., Wu, C., Lu, A., Chen, X., Hou, T., & Cao, D. (2021). ADMETlab 2.0: An integrated online platform for accurate and comprehensive predictions of ADMET properties. *Nucleic Acids Research*, *49*(W1), W5–W14. https://doi.org/10.1093/nar/gkab255

Xiong, Y., Gao, X., Pan, D., Zhang, T., Qi, L., Wang, N., Zhao, Y., & Dang, Y. (2022). A strategy for screening novel umami dipeptides based on common feature pharmacophore and molecular docking. *Biomaterials*, 121697. https://doi.org/10.1016/j.biomaterials.2022.121697

Xu & Chung. (2019). Quantitative Structure–Activity Relationship Study of Bitter Di-, Tri- and Tetrapeptides Using Integrated Descriptors. *Molecules*, *24*(15), 2846. https://doi.org/10/gnmdcq

Xue, W., Liu, X., Zhao, W., & Yu, Z. (2022). Identification and molecular mechanism of novel tyrosinase inhibitory peptides from collagen. *Journal of Food Science*, *87*(6), 2744–2756. https://doi.org/10.1111/1750-3841.16160

Yin, J., Henriksen, N. M., Slochower, D. R., Shirts, M. R., Chiu, M. W., Mobley, D. L., & Gilson, M. K. (2017). Overview of the SAMPL5 host–guest challenge: Are we doing better? *Journal of Computer-Aided Molecular Design*, *31*(1), 1–19. https://doi.org/10.1007/s10822-016-9974-4

Yu, Z., Chen, Y., Zhao, W., Li, J., Liu, J., & Chen, F. (2018). Identification and molecular docking study of novel angiotensin-converting enzyme inhibitory peptides from Salmo salar using in silico methods. *Journal of the Science of Food and Agriculture*, *98*(10), 3907–3914. https://doi.org/10.1002/jsfa.8908

Zhang, M., Zhu, L., Wu, G., Liu, T., Qi, X., & Zhang, H. (2022). Rapid Screening of Novel Dipeptidyl Peptidase-4 Inhibitory Peptides from Pea (*Pisum sativum L.*) Protein Using Peptidomics and Molecular Docking. *Journal of Agricultural and Food Chemistry*, *70*(33), 10221–10228. https://doi.org/10.1021/acs.jafc.2c03949

Zhang, T., Nie, S., Liu, B., Yu, Y., Zhang, Y., & Liu, J. (2015). Activity Prediction and Molecular Mechanism of Bovine Blood Derived Angiotensin I-Converting Enzyme Inhibitory Peptides. *PLOS ONE*, *10*(3), e0119598. https://doi.org/10.1371/journal.pone.0119598

Zhao, W., Su, L., Huo, S., Yu, Z., Li, J., & Liu, J. (2023). Virtual screening, molecular docking and identification of umami peptides derived from Oncorhynchus mykiss. *Food Science and Human Wellness*, *12*(1), 89–93. https://doi.org/10.1016/j.fshw.2022.07.026

Zhao, X., Chen, M., Huang, B., Ji, H., & Yuan, M. (2011). Comparative Molecular Field Analysis (CoMFA) and Comparative Molecular Similarity Indices Analysis (CoMSIA) Studies on α1A-Adrenergic Receptor Antagonists Based on Pharmacophore Molecular Alignment. *International Journal of Molecular Sciences*, *12*(10), 7022–7037. https://doi.org/10.3390/ijms12107022

Zheng, L., Zhao, Y., Dong, H., Su, G., & Zhao, M. (2016). Structure–activity relationship of antioxidant dipeptides: Dominant role of Tyr, Trp, Cys and Met residues. *Journal of Functional Foods*, *21*, 485–496. https://doi.org/10/f8cqgs

Zhou, P., Liu, Q., Wu, T., Miao, Q., Shang, S., Wang, H., Chen, Z., Wang, S., & Wang, H. (2021). Systematic Comparison and Comprehensive Evaluation of 80 Amino Acid Descriptors in Peptide QSAR Modeling. *Journal of Chemical Information and Modeling*, *61*(4), 1718–1731. https://doi.org/10.1021/acs.jcim.0c01370

Zhu, Z., Chen, Y., Jia, N., Zhang, W., Hou, H., Xue, C., & Wang, Y. (2022). Identification of three novel antioxidative peptides from Auxenochlorella pyrenoidosa protein hydrolysates based on a peptidomics strategy. *Food Chemistry*, *375*, 131849. https://doi.org/10.1016/j.foodchem.2021.131849

Figures and Tables:

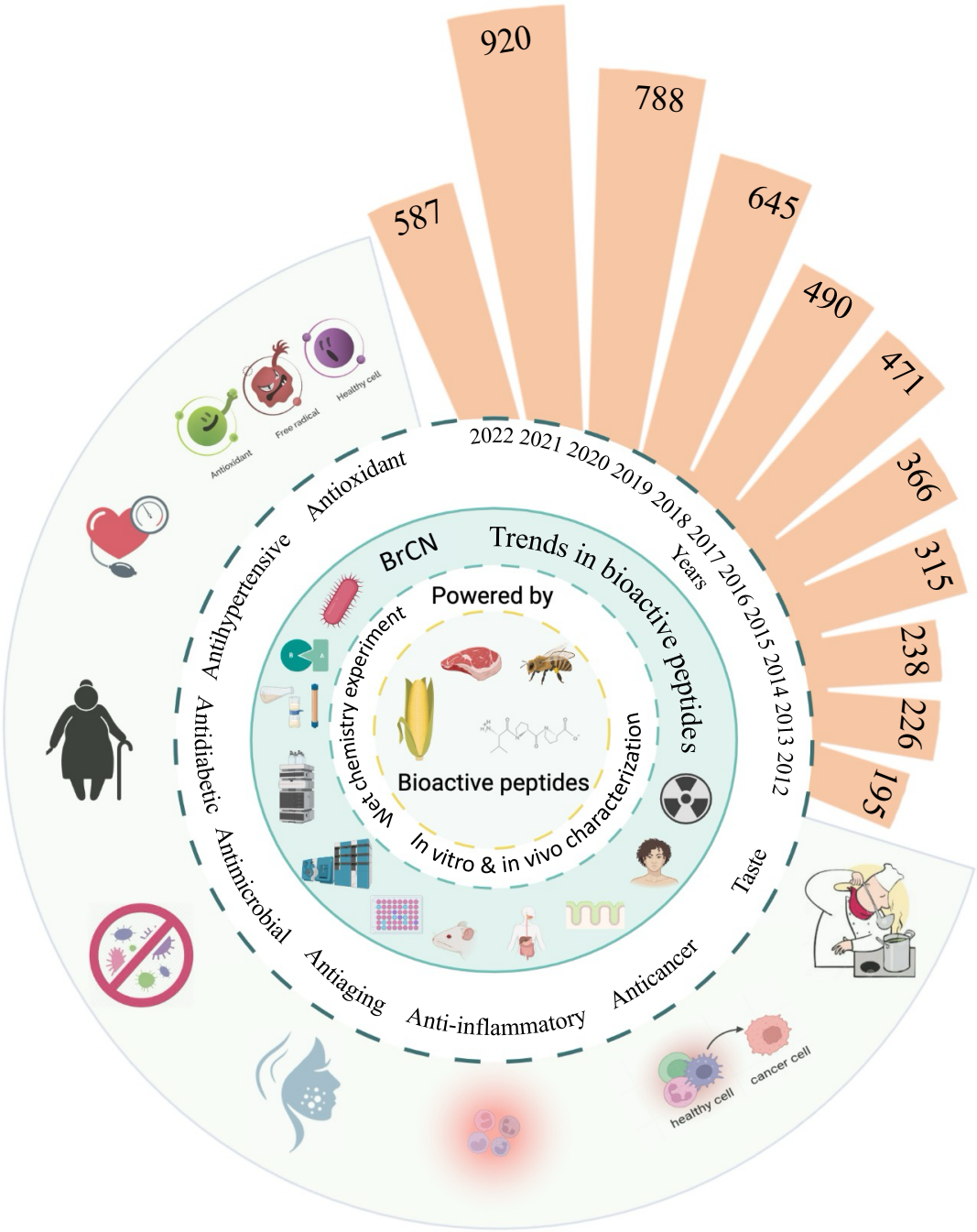


Figure 1-1 Status quo of food protein-derived bioactive peptide research.

Note: Data on the number of publications per year were obtained from Web of Science in August 2022 using “food & bioactive peptide” as keywords; BrCN: cyanogen bromide.

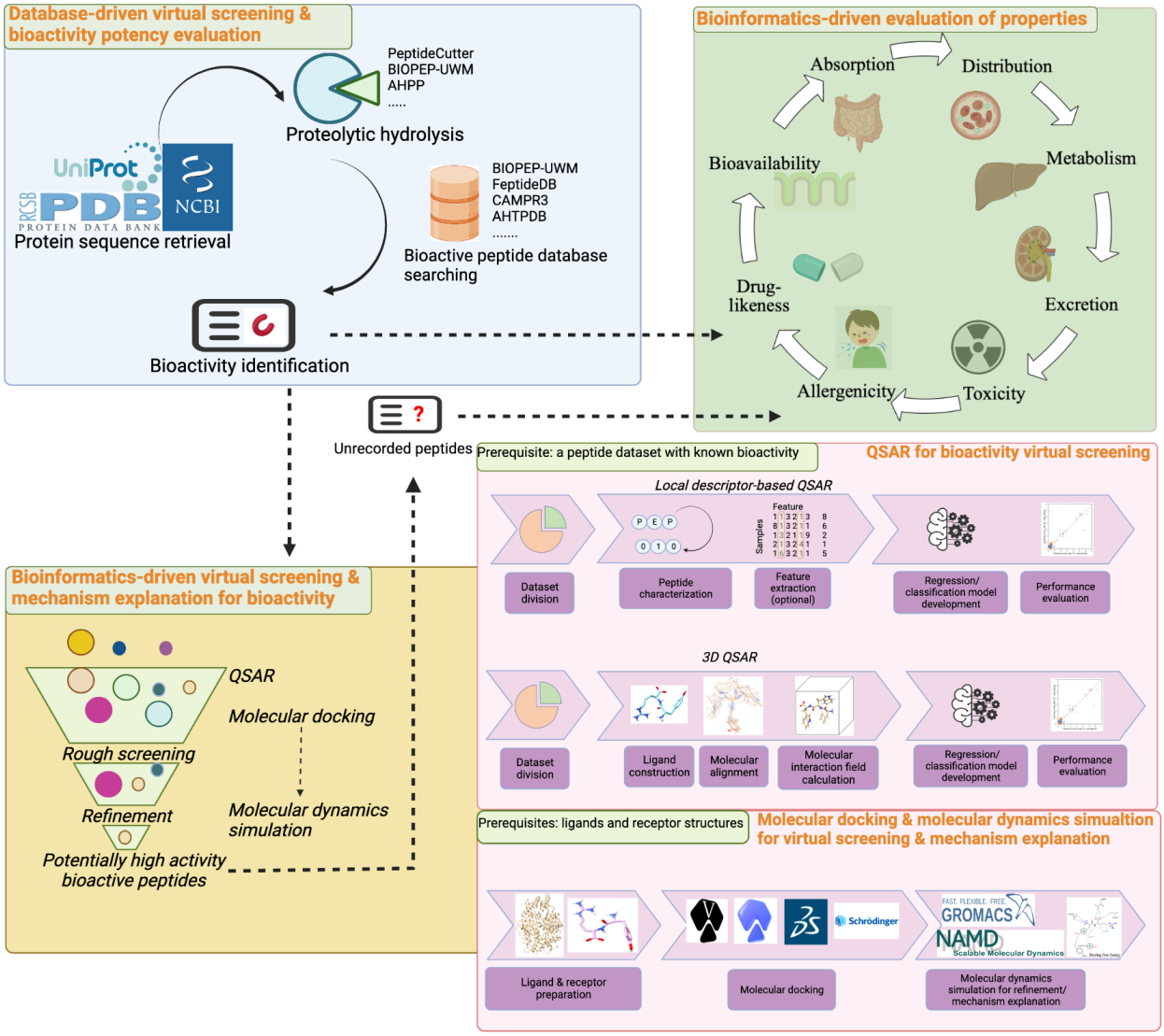


Figure 1-2 Overall workflow of bioinformatics application in food protein-derived bioactive peptide studies.

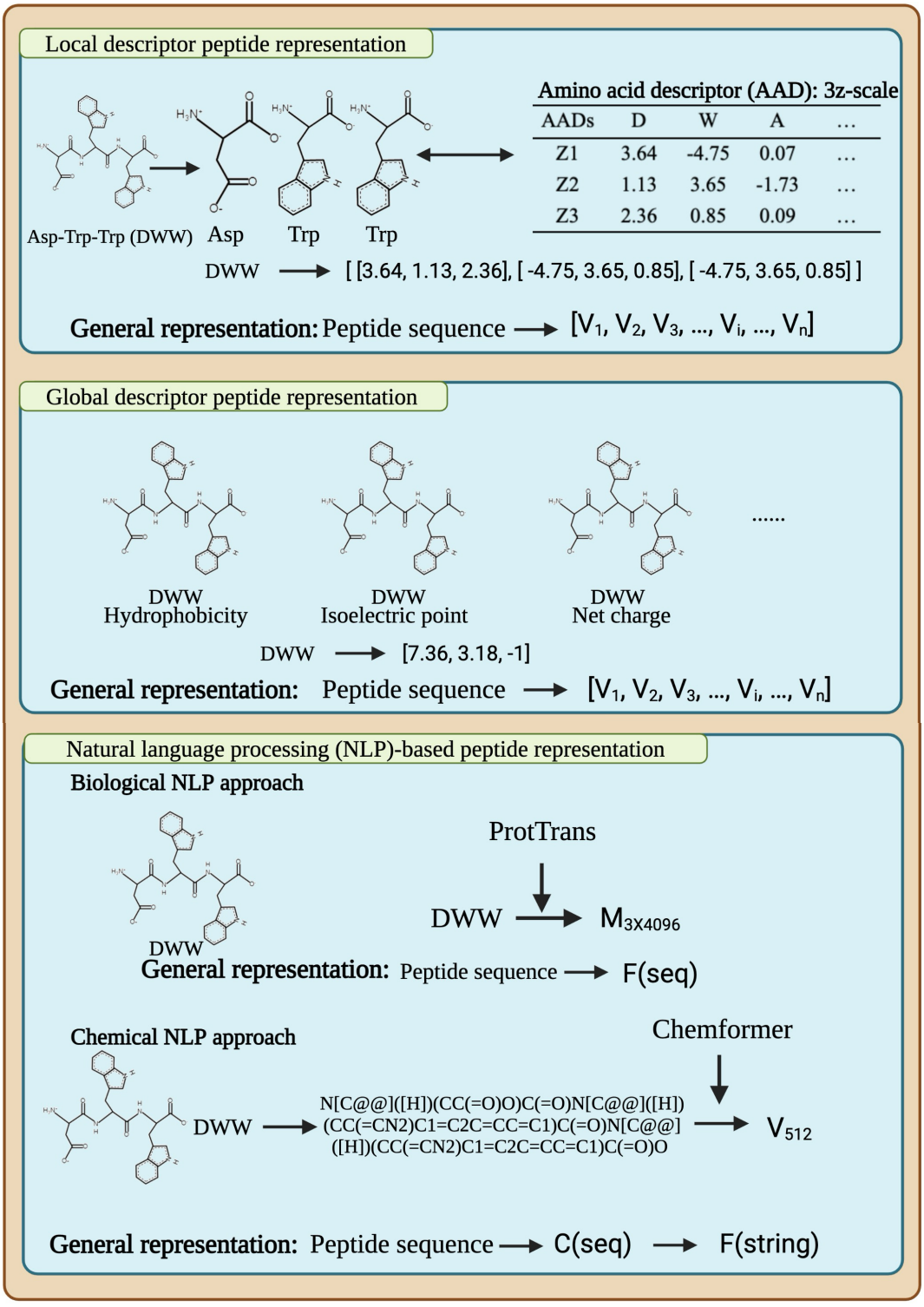


Figure 1-3 Overall workflow of peptide representation.

Note: Local descriptor peptide representation: V_i is the amino acid descriptor of the i^{th} residue in the peptide sequence, and it is a one-by-k vector; k is the feature dimension of AAD; n is the length of the peptides; the total dimension of one peptide is k*n.

Global descriptor peptide representation: V_i is the property of the whole peptide; n is the number of the properties used for peptide encoding.

NLP -based peptide representation: the biological NLP approach uses the pre-trained transformer (e.g., ProtTrans) from protein sequences as the dataset for the peptide encoding; the chemical NLP approach uses the pre-trained transformer (e.g., Chemformer) from molecules as the dataset for the peptide encoding. Before the transformation, peptide need to be converted into chemicals strings (e.g., SMILE format).

Table 1-1 Summary of protein information databases, proteolysis simulation web servers, bioactive peptide databases, mass spectroscopy data processing software, physiochemical property prediction web servers, peptide structure prediction web servers, and software/webserver for protein and peptide structure building, modification, and visualization.

Protein information database		
Name	*Website*	*Description*[*]
UniProt	https://www.uniprot.org/	The most commonly used protein database. Contains Swiss-Prot with 568,002 manually reviewed protein sequences and TrEMBL with 226,771,948 unreviewed protein sequences
NCBI Protein	https://www.ncbi.nlm.nih.gov/protein/	Collection of protein sequences from 7 public databases (including sequences translated from annotated coding regions from genes)
RCSB	https://www.rcsb.org/	Most commonly used protein structure database for molecular docking and molecular dynamics simulation. Contains 194,259 protein three-dimensional structures from X-ray crystallography, NMR, and electron microsopy experiments.
EMDB	https://www.ebi.ac.uk/emdb/	Contains 21,807 entries of electron cryo-microscopy maps and tomograms of macromolecular complexes and subcellular structures
Proteolysis simulation web server		
Name	*Website*	*Description*
PeptideCutter	https://web.expasy.org/peptide_cutter/	The most commonly used tool. Has 34 different enzymes and chemicals available for proteolytic simulation
FeptideDB	http://www4g.biotec.or.th/FeptideDB/enzyme_digestion.php	Has the same enzymes as PeptideCutter
BIOPEP-UWM	https://biochemia.uwm.edu.pl/biopep-uwm/	Has 34 different enzymes from microbes, vegetables, fruits, and humans
AHPP	http://hazralab.iitr.ac.in/ahpp/index.php	Has 10 enzymes from the gastrointestinal tract and 7 enzymes from vegetables and fruits
SpirPep	http://spirpepapp.sbi.kmutt.ac.th/UserGuide.html	Has 10 enzymes with 1-3 miss cleavage options
Bioactive peptide database		
Name	*Website*	*Description*[**]

DFBP	http://www.cqudfbp.net/	Contains a total of 6,276 peptide entries in 31 types from different food sources (last updated in 2022)
BIOPEP-UWM	https://biochemia.uwm.edu.pl/	Most commonly used database in FBP studies. Contains 4,485 bioactive peptides with 58 bioactivity categories from literature and 533 sensory peptides and amino acids (last updated in 2019)
FeptideDB	http://www4g.biotec.or.th/FeptideDB/index.php	Combination of 12 public bioactive peptide databases and peptides manually extracted from literature (last updated in 2019)
SpirPep	http://spirpepapp.sbi.kmutt.ac.th/BioactivePeptideDB.html	Combination of 13 public peptide databases (28,334 bioactive peptides)
$CAMP_{R3}$	www.camp3.bicnirrh.res.in	Contains 10,247 antimicrobial peptide sequences captured by analysis of 1,386 antimicrobial peptide sequences from experiments (last updated in 2019)
BaAMPs	http://www.baamps.it/	Antimicrobial peptides (AMPs) specifically tested against microbial biofilms (last updated in 2015)
YADAMP	http://yadamp.unisa.it/about.aspx	Contains 252 antimicrobial peptides (last updated in 2018)
BioPepDB	http://bis.zju.edu.cn/biopepdbr/index.php	Contains 4,807 bioactive peptides (51.07% are antimicrobial peptides, 34.9% are antihypertensive peptides, and 13.21% are anticancer peptides.) (last updated in 2018)
AHTPDB	http://crdd.osdd.net/raghava/ahtpdb/	Contains about 1,700 antihypertensive peptides (last updated in 2015)
BERT4Bitter	http://pmlab.pythonanywhere.com/dataset	Contained 256 bitter peptides (last updated in 2021)
DBAASP	https://dbaasp.org/home	Contains 19,902 antimicrobial/cytotoxic peptides with detailed 3D structure information (last updated in 2021)
AVPdb	http://crdd.osdd.net/servers/avpdb/	Contains 2,683 antiviral peptides (last updated in 2014)
TumorHoPe	http://crdd.osdd.net/raghava/tumorhope/	Contains 744 tumor homing peptides (last updated in 2012)
CancerPPD	http://crdd.osdd.net/raghava/cancerppd/index.php	Contains 3,491 anticancer peptides (last updated in 2015)
CPP site2.0	http://crdd.osdd.net/raghava/cppsite/	Contains 1,855 cell penetrating peptides (last updated in 2015)
BRAINPEP	https://brainpeps.ugent.be/	Blood-brain barrier peptide database (last updated in 2012)

NeuroPep	http://isyslab.info/NeuroPep/	Contains 5,949 neuropeptides (last updated in 2015)
Hemolytik	http://crdd.osdd.net/raghava/hemolytik/	Contains about 3,000 hemolytic and 2,000 non-hemolytic peptides (last updated in 2013)
MBPDB	http://mbpdb.nws.oregonstate.edu/	Contains 994 milk-derived bioactive peptides (last updated in 2021)
FermFooDb	https://webs.iiitd.edu.in/raghava/fermfoodb/	Contains 2,325 bioactive peptides from fermented food (last updated in 2021)
THPdb	http://crdd.osdd.net/raghava/thpdb/index.html	Contains 1,238 FDA-approved therapeutic peptides (last updated in 2017)
StraPep	http://isyslab.info/StraPep/index.php	Contains 3,791 bioactive peptide 3D-structures (last updated in 2018)

Mass spectroscopy data processing software

Name	*Website*	*Description*
Mascot	https://www.matrixscience.com/search_form_select.html	Searching software for peptide sequence identification using MS data and precursor proteins
PEAKS X	https://www.bioinfor.com/peaks-studio/	Searching software for peptide sequence identification using MS data and precursor proteins

Physiochemical property prediction web server

Name	*Website*	*Description*
PepDraw	http://www2.tulane.edu/~biochem/WW/PepDraw/	The most commonly used tool. Predicts net charges, isoelectric points, hydrophobicity, molar extinction coefficient, and molecular weight
AhtPom	http://crdd.osdd.net/raghava/ahtpin/allinone.php	Predicts net charges, isoelectric points, hydrophobicity, hydrophilicity, steric hindrance, solvation, net hydrogen, charge, and molecular weight
Compute pI/Mw	https://web.expasy.org/compute_pi/	Predicts isoelectric point and molecular weight
ProtParam	https://web.expasy.org/protparam/	Predicts molar extinction coefficient, estimated half-life, instability index, aliphatic index, and grand average of hydropathicity
PepCalc	https://pepcalc.com/	Predicts molar extinction coefficient, water solubility, net charges at neural pH, isoelectric points, and molecular weight

Peptide solubility calculator	https://pepcalc.com/peptide-solubility-calculator.php	Predicts solubility

Peptide structure prediction web server

Name	*Website*	*Description*
DEBT	https://comp.chem.nottingham.ac.uk/debt/	Secondary structure prediction
PPIIPred	http://bioware.ucd.ie/~compass/biowareweb/Server_pages/ppIIpredSEQUENCE.php	Secondary structure prediction
PEPstrMOD	http://osddlinux.osdd.net/raghava/pepstrmod/nat_ss.php	Tertiary structure prediction
APPTEST	https://research.timmons.eu/apptes	Tertiary structure prediction
PEP-FOLD3	https://bioserv.rpbs.univ-paris-diderot.fr/servixes/PEP-FOLD3	*De novo* structure prediction
PEP-FOLD	https://bioserv.rpbs.univ-paris-diderot.fr/services/PEP-FOLD/	*De novo* structure prediction (less optimized than PEP-FOLD3 but includes options such as disulfide bridges unavailable in the newer version)

Software/web server for protein and peptide structure building, modification, and visualization

Name	*Availability*	*Website*
MolView	Free	https://molview.org/
VMD	Free for academic use	http://www.ks.uiuc.edu/Research/vmd/
DiscoveryStudio	Free for academic use	https://discover.3ds.com/discovery-studio-visualizer-download
Chimera	Free for academic use	https://www.cgl.ucsf.edu/chimera/
Avogadro2	Free and open source	https://two.avogadro.cc/
PyMol	Free for academic use	https://pymol.org/2/

Note: * The number of sequences was obtained in August 2022; ** The “last updated” time refers to the time of the latest publication of the databases or notice of updates on the database websites

Table 1-2 Database-driven FBP virtual screening and bioactivity potency evaluation studies in the last 5 years.

Pure database-driven approach				
Protein source	*Bioactivity*	*Bioactivity potency evaluation*[*]	*Additional comments*	*Reference*
Chickpea	ACE inhibitory activity	A, B, A_E, B_E	Conducted ADMET evaluation and molecular docking for interaction mechanism explanation	(Arámburo-Gálvez et al., 2022)
Sheep milk protein	ACE inhibitory activity DPP-IV inhibitory activity	A, A_E	Used PeptideRanker for bioactivity evaluation; conducted physicochemical property and toxicity evaluation; used molecular docking for virtual screening and mechanism explanation.	(Iram et al., 2022)
Mammal milk proteins	ACE inhibitory activity	Molar concentration of released peptides	Also searched for DPP-III inhibitory activity, antioxidant activity, hypo-cholesterolemic activity, immunomodulatory activity, and antithrombotic activity in FBPs with ACE inhibitory activity	(Parastouei, 2022)
Goat casein	ACE inhibitory activity	A, B	Used PeptideRanker and AHTpin for bioactivity evaluation. Conducted *in vitro* and *in vivo* validation	(Coscueta et al., 2022)
Pumpkin seed	ACE inhibitory activity	A, A_E, W	Used molecular docking for virtual screening and molecular dynamics simulation for refinement. Conducted ADMET evaluation and *in vitro* validation	(Liang et al., 2021)
Porcine liver	Antioxidant activity	Not mentioned	Used PeptideRanker for bioactivity evaluation. Conducted *in vitro* and *in vivo* validation	(Pearman et al., 2020)

Flaxseed	ACE inhibitory activity Renin inhibitory activity	Not mentioned	Conducted physicochemical property, ADMET, and drug-likeness evaluation; used molecular docking for mechanism exploration and affinity comparison with common drugs (aliskiren and captopril)	(Ji et al., 2020)
Tomato seed	ACE inhibitory activity DPP-IV inhibitory activity Antioxidant activity	A, A_E, sum of A_E	Conducted allergenicity and toxicity evaluation	(Kartal et al., 2020)
Animal and fish collagens	ACE inhibitory activity DPP-IV inhibitory activity	A_E, DH_t, W	Conducted ADMET evaluation; used PeptideRanker for bioactivity evaluation; used SwissTargetPrediction to predict potential interaction between selected FBPs and other enzymes and proteins	(Iwaniak et al., 2020)
Bovine milk protein	ACE inhibitory activity DPP-IV inhibitory activity Antioxidant activity	Molar concentration of released peptides	Used PeptideRanker for bioactivity evaluation and comparison of released FBPs with digestion-resistant peptides	(Barati et al., 2020)
Milk and meat derived protein	ACE inhibitory activity Antioxidant activity Opioid activity	Not mentioned	Conducted physicochemical property and toxicity evaluation for FBPs and peptides; used PeptideRanker for bioactivity evaluation; conducted *in vitro* and *in vivo* validation	(Amigo et al., 2020)
Tilapia	ACE inhibitory activity	DH_t	Conducted physicochemical property and toxicity evaluation and *in vitro* and *in vivo* validation	(J. Chen et al., 2020)
Flaxseed	ACE inhibitory activity DPP-IV inhibitory activity Antioxidant activity	A, A_E, DH_t, W	Conducted physicochemical property and toxicity evaluation	(Ji et al., 2019)
Oyster	ACE inhibitory activity DPP-IV inhibitory activity	Not mentioned	Validated DH_t of enzymes by *in vitro* hydrolysis; used proteolytic simulation to guide enzyme selection	(Gomez et al., 2019)
Yak milk	ACE inhibitory activity	A	Conducted toxicity evaluation	(K. Lin et al., 2018)

Salmo salar	ACE inhibitory activity	Not mentioned	Conducted physicochemical property and toxicity evaluation; used molecular docking screening and *in vitro* study for validation	(Yu et al., 2018)
Giant grouper roe	ACE inhibitory activity DPP-IV inhibitory activity	A	Conducted *in vitro* study on trypsin in-gel digestion	(Panjaitan et al., 2018)
Bovine casein	Anticancer activity Antithrombotic activity Anti-inflammatory activity Immunomodulating activity	Not mentioned	Used PeptideRanker for bioactivity, toxicity, and allergenicity evaluation	(Tu, Liu, et al., 2018)
Caulerpa RuBisCO	ACE inhibitory activity DPP-IV inhibitory activity Antioxidant activity Neuroprotective activity Antithrombotic activity	A, A_E	Proteins only had 49 % sequence homology among 28 species	(Agirbasli & Cavas, 2017)
Rice bran	ACE inhibitory activity DPP-IV inhibitory activity	A, B, A_E, W	Used PeptideRanker for bioactivity, physicochemical property, allergenicity, and toxicity evaluation	(Pooja et al., 2017)

Partial database-driven approach

Protein source	*Chemistry experiments and other analysis*	*Database-driven analysis*	*Reference*
Tuna skin collagen	FBP identification by LC-MS; physicochemical property evaluation	All the available bioactivities of identified FBPs were searched in BIOPEP	(Devita et al., 2021)
Cheese	FBP identification by LC-MS; FBPs with anti-diabetic activity were synthesized and confirmed by *in vitro* experiments	Identified FBPs were searched in BIOPEP and MBPDB (Table 1)	(Martini et al., 2021)
Food matrix	FBP identification by LC-MS; FBP conformation prediction	All the available bioactivities of identified FBPs were searched in BIOPEP	(Sayd et al., 2018)
Bean protein	FBP identification by LC-MS; physicochemical property evaluation	Bioactivities of identified FBPs were searched in BIOPEP	(Martini et al., 2022)

Porcine products	*In vitro* hydrolysis and bioactivity evaluation	*In silico* hydrolysis simulation and potential bioactivity evaluation were conducted to guide optimal enzyme selection	(Bechaux et al., 2020)
Wheat	Peptideranker was used for FBP screening, and opioid activity of FBPs was confirmed by *in vitro* experiment	Used *in silico* hydrolysis simulation	(Garg et al., 2018)

Notes: *: $A = \frac{a}{N}$ or $A_E = \frac{d}{N}$ where A is the occurrence frequency of peptides in a protein, A_E is the occurrence frequency of released peptides from a protein under a specific enzyme, a is the number of peptides with bioactivity encrypted in the selected protein chain, d is the number of released peptides with bioactivity, and N is the number of amino acid residues in the protein. $DH_t = \frac{d}{N}$ where DH_t is the theoretical degree of hydrolysis, d is the number of hydrolyzed peptide bonds, and D is the total number of peptide bonds in a protein chain. $W = \frac{A_E}{A}$ where W is the relative frequency of the peptide released by given activity by selected enzymes or chemicals, A is the occurrence frequency of peptides, and A_E is the occurrence frequency of released peptides. $B = \frac{\sum_{i=1}^{k} \frac{a_i}{EC_{50i}}}{N}$ or $B_E = \frac{\sum_{i=1}^{k} \frac{a_{Ei}}{EC_{50i}}}{N}$ where B is the potential bioactivity of a protein, B_E is the potential bioactivity of released peptides from a protein under a specific enzyme, a_i is the number of repetition of the i-th bioactive fragment in the protein sequence, a_{Ei} is the number of repetition of the i-th bioactive peptide released from the protein sequence, EC_{50i} is the concentration of the i-th bioactive peptide corresponding to its half-maximal activity (μM), k is the number of different fragments with a given activity, and N is the number of amino acid residues.

Abbreviations: ACE: angiotensin-I-converting enzyme; ADMET: absorption, distribution, metabolism, excretion, and toxicity; DPP-IV: dipeptidyl peptidase IV; RuBisCO: ribulose-1,5-bisphosphate carboxylase; UbMP: activate ubiquitin-mediated proteolysis.

Table 1-3 Selected virtual screening strategies using QSAR, molecular docking, and molecular dynamics simulation.

Local descriptor-based QSAR model application						
Bioactivity	*Dataset size*	*Amino acid descriptors (AADs)*	*Model*	*Performance*	*Additional comments**	*Reference*
Antioxidant activity (ABTS assay)	133 tripeptides	566 amino acid properties were selected by 6 ML methods as AADs or directly used as AADs	14 ML regression models	R^2_P = 0.847 (best model by RFR for feature selection and XGB for regression model)	QAY, PHC, YPQ, VYV, GPE, and YSQ; performance comparison between 98 models	(Du, Wang, et al., 2022)
ACE inhibitory activity	58 dipeptides	80 different AADs	SVMR and PLSR	R^2_P varied from 0.6 to 0.82 for SVMR and from 0.48 to 0.7 for PLSR	N/A	(Zhou et al., 2021)
Bitterness	48 dipeptides, 52 tripeptides, and 23 tetrapeptides	16 AADs were integrated by BOSS	PLSR	R^2_{CV} = 0.941 for dipeptides R^2_{CV} = 0.742 for tripeptides R^2_{CV} = 0.956 for tetrapeptides	MPA for outlier detection; performance comparison between the descriptor selector by BOSS and 16 basic AADs	(Xu & Chung, 2019)
Global descriptor-based QSAR (3D-QSAR) model application						
Bioactivity	*Dataset size*	*Molecular alignment & MIF calculation*	*Model*	*Performance*	*Additional comments**	*Reference*
ACE inhibitory activity	53 di/tripeptides	Docking-based alignment (Suflex-Dock); CoMFA and CoMSIA for MIF calculation	PLSR	R^2_{CV} = 0.773 for CoMFA R^2_P = 0.664 for CoMSIA	GEF, VEF, VRF, and VKF were synthesized for validation	(Qi et al., 2017)
ACE inhibitory activity	40 dipeptides, 32 tripeptides, and 41	Template ligand-based alignment; CoMFA	PLSR	R^2_{CV} = 0.862 for dipeptides (CoMSIA)	N/A	(S. Wu et al., 2014)

	tetra/penta/hexa peptides	and CoMSIA for MIF calculation		R^2_{CV} = 0.848 for tripeptides (CoMFA) R^2_{CV} = 0.656 for longer peptides (CoMFA)		
Molecular docking & molecular dynamics simulation for virtual screening						
Protein source	*Bioactivity*	*Molecular docking*	*Molecular dynamics simulation*	*Additional comments*	*In vitro or in vivo experiments*	
Fermented soy	Keap1–Nrf2 interaction inhibitory activity	CABS-Dock, GalaxyPepDock, and HADDOCK	-	28 peptides identified by LC-MS; CABS-Dock and GalaxyPepDock were used for screening, and HADDOCK was used for conformation refinement	Thirteen peptides were selected for *in vitro* and *in vivo* antioxidant experiments	(Tonolo, Moretto, et al., 2020)
Pumpkin seed	ACE inhibitory activity	MOE	GROMACS for 30 ns of simulation; Force field: Amber99SB-ILDN	Identified 47 di/tripeptides virtually screened by MOE; ADMET evaluation to select for peptide synthesis	Tripeptide IFA was synthesized for *in vitro* ACE inhibitory activity assay	(Liang et al., 2021)
-	ACE inhibitory activity	GOLD	AMBER12 for 25 ns of simulation; Force field: AMBER FF03	8000 theoretical tripeptides were ranked by GOLD, and MDS was used to elucidate interaction mechanism	Five peptides were synthesized for *in vitro* ACE inhibitory activity assay	(Panyayai et al., 2018)
Walnut meal	Tyrosinase inhibitory activity	AutoDock 4 and CDOCKER	-	606 peptides identified from LC-MS were screened by AutoDock 4; CDOCKER was used for conformation refinement	FPY was synthesized for *in vitro* experiment	(Y. Feng et al., 2021)
Integrated strategy						

Protein source	*Bioactivity*	*QSAR model*	*Molecular docking*	*Molecular dynamics simulation*	*Additional comments*	*Reference*
-	DPP-IV inhibitory activity	Self-built regression model	AutoDock Vina	GROMACS for 100 ns of simulation Forcefiled: AMBER14SB and General AMBER	PaDEL descriptors for peptide representation, GA for feature selection, and MLR for regression model	(X. Li et al., 2022)
Rubing cheese	ACE, a-glucosidase, and Keap1–Nrf2 interaction inhibitory activity	PeptideRanker	Pepsite2 and AutoDock Vina		PeptideRanker for rough virtual screening and PepSite2 for refinement and selection; AutoDock Vina for interaction mechanism	(Wei et al., 2022)
Chia Seed	ACE inhibitory activity, anti-inflammatory activity, and plasma stability	PeptideRanker, AHTpin, PreAIP, AntiAngioPred, and PlifePred	AutoDock Vina	PMEMD for 300 ns of simulation Force fields: amberff14 and GAFF	PMEMD for conformation generation of protein receptors and AutoDock Vina for molecular scoring	(Aguilar-Toalá et al., 2022)
	ACE inhibitory activity	Self-built regression model	AutoDock4	-	3D QSAR models (PLSR) based on ligand template-based molecular alignment and MIF calculation; AutoDock4 for interaction mechanism	(F. Wang & Zhou, 2020)
Donkey collagen	Tyrosinase inhibitory activity	PeptideRanker and prediction for water solubility	CDOCKER	GROMACS for 60 ns of simulation; Force field: CHARMM36	PeptideRanker, physicochemical properties, and CDOCKER combined for virtual screening	(Xue et al., 2022)

	and toxicity	

Notes: Here, only the representative strategies are summarized. The comprehensive review of the latest virtual screening strategies in FBPs is available at Supplementary Document S1-Table S1.

*: Peptides were synthesized by chemical approach for validation experiments; N/A: not available.

Abbreviation: AADs: amino acid descriptors; BOSS: bootstrapping soft shrinkage; CoMFA: comparative molecular field analysis; CoMSIA: comparative molecular similarity indices analysis; GA: genetic algorithm; MPA: model population analysis; MIF: molecular interaction field; MLR: multiple linear regression; PLSR: partial least squares regression; R^2_P: determination of coefficient in test dataset; R^2_{CV}: determination of coefficient of cross-validation; SVMR: support vector machine regression; XGB: extreme gradient boost.

Table 1-4 Summary of QSAR prediction webservers.

Common bioactivity prediction					
Bioactivity	*Web server*	*Website*	*Model development*	*Model performance*	*Release time*
18 different properties (20 datasets)*	UniDL4BioPep	https://nepc2pvmzy.us-east-1.awsapprunner.com/	A protein language model based CNN model	Better performances than the respective state-of-the-art models for 15 out of 20 different bioactivity dataset prediction tasks)	2023
Antioxidant	AnOxPePred	http://services.bioinformatics.dtu.dk/service.php?AnOxPePred-1.0	CNN model	ACC is around 80% MCC is around 0.7	2020
Antioxidant	IDAod	http://antioxidant.weka.cc	Neural network for feature extraction; t-SNE for feature reduction; binary SVM classifier model	ACC=97.05% MCC= 0.7409	2018
ACE inhibition	pLM4ACE	https://sqzujiduce.us-east-1.awsapprunner.com/	A protein language model based CNN model; confident learning theory for data cleaning	BACC=88.3% MCC=0.77	2023
ACE inhibition	MAT-AHT	http://hazralab.iitr.ac.in/ahpp/index.php	No extra feature selection method; regression decision tree model	r = 0.9513	2021
ACE inhibition	PAAP	http://codes.bio/paap/	No extra feature selection method; random forest binary classifier	ACC = 84.73%	2018
ACE inhibition	AHTpin	http://crdd.osdd.net/raghava/ahtpin/di_mat.php	No extra feature selection method; SVM regression model for di/tripeptides; SVM classification	r = 0.701 and 0.543 for dipeptides and tripeptides, respectively; ACC = 76.67%, 72.04%, 77.39, 82.61%, and 84.21% for tetrapeptide, pentapeptide, hexapeptides, medium peptides (7-13 residues), and large peptides	2015

				(above 13 residues), respectively	
DPPIV inhibition	StackDPPIV	http://pmlabstack.pythonanywhere.com/StackDPPIV	GA-SAR for feature selection; random forest binary classifier	ACC = 89.1% MCC = 0.784 AUC = 96.1%	2022
DPPIV inhibition	iDPPIV-SCM	http://camt.pythonanywhere.com/iDPPIV-SCM	No extra feature selection method; Scoring card method for binary classifier	ACC = 69.7% MCC = 0.594 AUC = 84.7	2020
Antimicrobial	ClassAMP	http://www.bicnirrh.res.in/classamp/	SVM or RF classifier	MCC = 0.92, 0.83, and 0.96 for antibacterial, antifungal, and antiviral peptides, respectively	2020
Antimicrobial	MAT-AMP	http://hazralab.iitr.ac.in/ampgp.php	Not available	Not available	2021
Antimicrobial	iAMGpred	http://hazralab.iitr.ac.in/ampgp.php	SVM classifier	AUC = 94% MCC =0.88	2017
Antimicrobial	ADAM	https://bioinformatics.cs.ntou.edu.tw/ADAM/tool.html	SVM or profile hidden Markov models	Not available	2015
Anticancer	xDeep-AcPEP	https://app.cbbio.online/acpep/home	CNN model	r = 0.8073, 0.8322, 0.7289, 0.8179, 0.8370, and 0.8285 for breast, cervix, skin, prostate, lung, and colon cancer, respectively	2021
Anticancer	MLACP 2.0	https://balalab-skku.org/mlacp2	CNN model	ACC = 76.5% MCC = 0.513 AUC = 0.773	2022
Anticancer	AntiCP 2.0	https://webs.iiitd.edu.in/raghava/anticp2/index.html	Extra trees classifier	ACC = 92.1% MCC = 0.84	2021
Anti-inflammatory	InflamNat	http://www.inflamnat.com/	Multi-tokenization transformer model	AUC = 84.2%	2022

Anti-inflammatory	PreAIP	http://kurata14.bio.kyutech.ac.jp/PreAIP/	RF model	ACC = 77% MCC = 0.512 AUC = 84%	2019
Anti-angiogenic	AntiAngioPred	http://webs.iiitd.edu.in/raghava/antiangiopred/	SVM classifier	ACC = 80.9% MCC = 0.62	2015
Hemolytic	HAPPENN	https://research.timmons.eu/happenn	Neural network	ACC = 85.7% MCC = 0.71	2020
Hemolytic	HemoPred	http://codes.bio/hemopred/	RF model	ACC = 95% MCC = 0.91	2017
Hemolytic	HemoPI	http://crdd.osdd.net/raghava/hemopi/	SVM model	ACC = 96.4% or 75.7% MCC = 0.93 or 0.51 for two different datasets	2016
Anti-tubercular	AtbPpred	http://thegleelab.org/AtbPpred	Neural network model	AAC = 87.3%	2019
Immunomodulatory	NetMHCpan 4.0	https://services.healthtech.dtu.dk/service.php?NetMHCpan-4.0	Neural network model	r = 0.790 AUC = 93.4%	2020

Additional property prediction

Function	*Web server*	*Website*	*Model development*	*Model performance*	*Release time*
Whether a is bioactive	PeptideRanker	http://distilldeep.ucd.ie/PeptideRanker/	Neural network model	With 0.8 as threshold, FPR = 0.02 and 0.06, MCC = 0.54 and 0.74, and AUC = 0.04 and 0.932 for short peptides (4-20 residues) and long peptides (> 20 residues), respectively	2012
Peptide intestinal stability	HLP	http://crdd.osdd.net/raghava/hlp/	SVM regression model	r = 0.70 and 0.98 for two different datasets	2014

Peptide penetrate cell	CPPpred	http://distilldeep.ucd.ie/CPPpred/	Neural network model	MCC = 0.69 FPR = 2.13	2013
Plasma stability	PlifePred	http://webs.iiitd.edu.in//raghava/plifepred/	SVM regression model	r = 0.743	2018
Metabolism prediction	BioTransformer 3.0	https://biotransformer.ca/	A knowledge-based prediction tool and a set of random forest and ensemble models	Jaccard score range from 0.380 to 0.452 for 9 metabolic transformations	2022
ADMET evaluation	ADMETlab 2.0	https://admetmesh.scbdd.com/	40 classification models and 13 regression models were built on MGA framework	For the regression models, R^2 ranges from 0.678 to 0.957, and average R^2 is 0.783; AUC ranges from 0.707 to 0.983, and average AUC is 0.863	2021
ADME evaluation	SwissADME	http://www.swissadme.ch/	19 classification models and 15 regression models were built on existing models	Best model R^2 = 0.75 for water solubility; R^2 = 0.67 for skin permeability coefficient; ACC ranges from 0.78 to 90.8 for pharmacokinetic parameters, and average ACC is 80.5%; r = 0.91 or 0.62 (depending on datasets)	2017
Protein allergenicity	AllerCatPro	https://allercatpro.bii.a-star.edu.sg	Combination of protein clustering program (cd-hit) and BLAST search tool; homology detection and molecular dynamics simulation for 3D structure correction and comparison	ACC = 84% MCC = 0.727	2022
Peptide allergenicity	SORTALLER	http://sortaller.gzhmc.edu.cn/	A substantially optimized SVM binary classification model	AAC = 98.5% MCC =0.97	2012

Chemical allergenicity	ChAlPred	https://webs.iiitd.edu.in/raghava/chalpred/	Pearson correlation and support vector classifier for feature selection; random forest for classifier	AAC = 83.39% AUC = 0.93	2021
Chemical cytochrome activity	SuperCYPsPred	http://insilico-cyp.charite.de/SuperCYPsPred/	Five RF classifiers for CYP1A2, CYP2C9, CYP2C19, CYP2D6, and CYP3A4	All the ACC > 93% All the AUC >0.93	2020
Peptide toxicity	ProTox-II	http://tox.charite.de/protox_II	31 models were developed by RF, ensembles of SVM and RF, or Bernoulli–Naive Bayes models for different toxicities	Average ACC = 85 % Average AUC = 0.83	2018
Peptide toxicity	ToxinPred	https://webs.iiitd.edu.in/raghava/toxinpred/index.html	SVM model	ACC = 96.01% MCC = 0.89	2013
Taste	VirtualTaste	http://virtualtaste.charite.de/VirtualTaste/	RF classifier	ACC = 90% AUC = 98%	2021
Bitterness	BERT4Bitter	http://pmlab.pythonanywhere.com/BERT4Bitter	A BERT-based binary classification model	92.2% accuracy in test dataset; the BERT-model outperformed models built on CNN and LSTM neural network	2021
Bitterness	iBitter-SCM	http://camt.pythonanywhere.com/iBitter-SCM	A SCM-based binary predictor	AAC = 84.38%	2020
Umami	iUmami-SCM	http://camt.pythonanywhere.com/iUmami-SCM	A SCM-based binary predictor	ACC= 86.5% MCC=0.679	2020

*: The 18 bioactivities include angiotensin- converting enzyme (ACE) inhibitory activity (anti-hypertension), dipeptidyl peptidase IV (DPPIV) inhibitory activity (antidiabetes), bitter, umami, antimicrobial activity, antimalarial activity, quorum-sensing (QS) activity,

anticancer activity, anti-methicillin-resistant S. aureus (MRSA) strains activity, tumor T cell antigens (TTCA), blood-brain barrier, antiparasitic activity, neuropeptide, antibacterial activity, antifungal activity, antiviral activity, toxicity and antioxidant activity

Abbreviation: ACC: accuracy; ADMET: absorption, distribution, metabolism, excretion, and toxicity; AUC: area under the curve of a receiver operating characteristic curve plots; BERT: bidirectional encoder representation from transformer, BLAST: Basic Local Alignment Search Tool; CNN: convolutional neural network; FPR: false positive rate; GA-SAR: genetic algorithm based on self-assessment report; LSTM: long short-term memory; MCC: Matthews correlation coefficient; MGA: multi-task graph attention; r: Pearson correlation coefficient; R^2: determination of coefficient; RF: random forest; SCM: scoring card method; SVM: support vector machine; t-SNE: t-distributed Stochastic Neighbor Embedding.

Table 1-5 Summary of molecular docking software/web servers, dynamics simulation software, and Force fields.

Molecular docking software						
Name	*Sampling algorithm*	*Scoring function*	*Additional comments*	*Successl rates*[*]	*Availability*	*Websites*
AutoDock 4	Stochastic algorithm (Lamarckian genetic algorithm)	Force field-based and empirical scoring function (AD4)	The basic version uses flexible ligands and a rigid receptor, but it also has specific modified solutions for hydrated docking, zinc metalloprotein docking, flexible docking (flexible residues for receptors), and multiple ligand docking. AutoDock4 can be up to 100× slower than AutoDock Vina. AutoDock Vina has a batch mode for a large number of ligands and also has modified versions (e.g., QuickVina2, Vinardo, and InstaDock). Both AutoDock4 and AutoDock Vina are continually maintained and updated.	53%	Free for academic use	https://autodock.scripps.edu/download-autodock4/
AutoDock Vina	Stochastic search (Monte-Carlo/BFGS searching)	Knowledge-based and empirical energy scoring function (Vina)		80%	Free for academic use	https://vina.scripps.edu/
AutoDockFR	Stochastic algorithm (genetic algorithm and Solis-Wets local search)	Force field-based and empirical scoring function (AD4)	It uses receptors with flexible side chains and flexible ligands.	74%	Free for academic use	https://ccsb.scripps.edu/adfr/

AutoDock CrankPep	Stochastic search (Monte-Carlo searching)	Force field-based and empirical scoring function (AD4)	It uses rigid receptors and flexible ligands. It is specially designed for protein-ligand docking.	85.7%	Free for academic use	https://ccsb.scripps.edu/adcp/documentation/
DOCK 6	Systematic conformational search (anchor-and-grow search algorithm/BFGS)	Empirical energy scoring function	It uses rigid receptors and flexible ligands. There are also six more scoring functions and deterministic search options.	73.3%	Free for academic use	https://dock.compbio.ucsf.edu/DOCK_6/index.htm
CDOCKER	Stochastic search (simulated annealing) and deterministic search (MDS+energy minimization)	Force field-based scoring function (soft-core potential)	It uses rigid receptors and flexible ligands. The last update was made in 2016 and called Flexible CDOCKER (62.7 % successful rate), which can set flexibility for receptors and was accelerated by SGLD. Flexible CDOCKER can remove the simulated annealing procedure.	74%	Need to purchase; built in Discovery Studio	https://discover.3ds.com/discovery-studio-visualizer-download
LibDock	Systematic conformational search (geometric hashing algorithm/BFGS)	Empirical energy scoring function (Ligscore)	It uses rigid receptors and flexible ligands. No update was reported after 2007.	46%	Need to purchase; built in Discovery Studio	https://discover.3ds.com/discovery-studio-visualizer-download
LigandFit	Stochastic search (Monte-Carlo searching)	Empirical energy scoring function (Ligscore)	It uses rigid receptors and flexible ligands. No update was reported after 2003.	-	Need to purchase; built in Discovery Studio	https://discover.3ds.com/discovery-studio-visualizer-download

GOLD	Stochastic search (genetic algorithm)	Empirical energy scoring function (Chemscore) or knowledge-based scoring function (GOLD)	It uses partial flexibility for receptors and flexible ligands.	81%	Need to purchase	https://www.ccdc.cam.ac.uk/solutions/csd-discovery/
PLANT	Stochastic search (ant colony algorithm)	Empirical energy scoring function ($LANTS_{CHEMPLP}$ and $PLANTS_{PLP}$)	It uses partial flexibility for receptors and flexible ligands.	72%	Free for academic use	http://www.tcd.uni-konstanz.de/research/plants.php
Glide	Systematic conformational search (hierarchical searching) and deterministic search algorithm	Empirical energy scoring function (GlideScore) or empirical energy scoring function (Emodel)	It uses partial flexibility for receptors and flexible ligands. It uses GlideScore to rank ligands and Emodel to find the best conformation. It is continually maintained and updated.	82%	Need to purchase	https://www.schrodinger.com/products/glide
Surflex-Dock	Systematic conformational search (incremental construction search)	Force field-based and empirical scoring function	It uses a rigid receptor and flexible ligands. It is continually maintained and updated.	78.6%	Need to purchase	https://www.biopharmics.com/downloads/
MOE	Stochastic search (triangle matcher algorithm)	Force field-based scoring function (S value)	It uses a rigid receptor and flexible ligands.	61.2%	Need to purchase	https://www.chemcomp.com/Products.htm

Molecular docking web servers

	Website	*Type*	*Additional comments*	*Availability*
CB-Dock	http://clab.labshare.cn/cb-dock/php/index.php	Global docking	Cavity-detection guided global docking based on Autodock Vina	Free

PepSite2	http://pepsite2.russelllab.org/	Global docking	Stochastic search (spatial position specific scoring matrix) and knowledge-based scoring function (hot spot score)	Free
SwissDock/EADock DSS	http://www.swissdock.ch/	Local/global docking	Stochastic search (EADock dihedral space sampling) and Force field-based and empirical energy scoring function (EADock2)	Free
HPEPDOCK	http://huanglab.phys.hust.edu.cn/hpepdock/	Global docking	Systematic conformational search (hierarchical searching) and knowledge-based and empirical energy scoring function	Free
HADDOCK	https://wenmr.science.uu.nl/haddock2.4/	Global docking	Systematic conformational search (an experimental knowledge-driven searching method) and energy scoring function (HADDOCK score)	Free
CABS-dock	http://biocomp.chem.uw.edu.pl/CABSdock	Global docking	Fully flexible receptors and ligands; deterministic search (CABS coarse-grained protein model) and empirical scoring function (energy scoring and structural clustering)	Free
PIPER-FlexPepDock	http://piperfpd.furmanlab.cs.huji.ac.il/	Global docking	Systematic conformational search (fragment-based searching) and Force field-based and empirical scoring function (FFT docking algorithm)	Free
FlexPepDock	http://flexpepdock.furmanlab.cs.huji.ac.il/	Local docking	Fully flexible refinement; stochastic search (Monte-Carlo sampling with energy minimization) and Force field-based scoring function (Rosetta score12 and Rosetta centroid score4)	Free
InterEvDock3	http://bioserv.rpbs.univ-paris-diderot.fr/services/InterEvDock3/	Template-based docking	Mainly used for protein-protein docking and can also be used for protein and long peptide docking. Systematic conformational search (FRODOCK algorithm) and Force field energy scoring function (InterEvScore)	Free

Molecular dynamics simulation software

Name	*Website*	*Availability*
NAMD	https://www.ks.uiuc.edu/Research/namd/	Free and open source for academic users
GROMACS	https://www.gromacs.org/	Free and open source

OpenMM	https://openmm.org	Free and open source
CHARMM	http://www.charmm.org/	Free for academic users
LAMMPS	https://lammps.org	Free and open source
AMBER	http://ambermd.org/	Need to purchase
	Force fields	
Name	*Website*	*Availability*
CHARMM	http://mackerell.umaryland.edu/charmm_ff.shtml	Free
AMBER	http://amber.scripps.edu/	Free
GROMOS	https://www.igc.ethz.ch/gromos.html	Free
OPLS	http://zarbi.chem.yale.edu/oplsaam.html	Free
KBFF20	https://kbff.chem.k-state.edu/	Free
	Free energy calculation tools	
Name	*Website*	*Availability*
YANK	https://getyank.org	Free
BFEE2	https://github.com/fhh2626/BFEE2	Free
pAPRika	https://paprika.readthedocs.io	Free
BRIDGE	https://github.com/scientificomputing/bridge	Free

Note: * means the root-mean-square deviation of the difference between the predicted molecular conformation and the experimental conformation is lower than 2.0 angstrom. The success rate can only be used to compare the performance of different docking tools when they have the same benchmark.

Abbreviations: AIRs: Ambiguous Interaction Restraints; BFGS: Broyden–Fletcher–Goldfarb–Shanno algorithm; FFT: Fast Fourier transform; MDS: molecular dynamics simulation; SGLD: self-guided Langevin dynamics.

Chapter 2 - Comprehensive evaluation and comparison of machine learning methods in QSAR modeling of antioxidant tripeptides*

Abstract

Due to their multiple beneficial effects, antioxidant peptides have attracted increasing interest. Currently, the screening and identification of bioactive peptides, including antioxidative peptides based on wet-chemistry methods are time-consuming and highly rely on many advanced instruments and trained personnel. Quantitative structure-activity relationship (QSAR) analysis as an *in silico* method can be more efficient and cost-effective. However, model performance of QSAR studies on antioxidant peptides was still poor due to limited attempts in model development approaches. The objective of this study was to compare popular machine learning methods for antioxidant activity modeling and screening of tripeptides and identify the critical amino acid features that determine the antioxidant activity. 533 numerical indices of amino acids were adopted to characterize 130 tripeptides with known antioxidant activity from published literatures, and then 7 feature selection strategies plus pairwise correlation were used to screen the most important indices for antioxidant activity and model building. 14 machine learning methods were used to build models based on the feature selection strategies, respectively. Among the 98 models, non-linear regression methods tended to perform better, and the best model with R^2_{Test} of 0.847 and $RMSE_{Test}$ of 0.627 for tripeptide antioxidants was obtained by combining random forest for feature selection and tree-based extreme gradient boost regression for model development. Based on the predicted antioxidant values of 7870 unknown tripeptides, potentially high antioxidant activity tripeptides all have a tyrosine, tryptophan, or cysteine residue at the C-terminal position. Furthermore, the predicted antioxidant activity of 6 synthesized tripeptides was confirmed through experimental determination, and for the first time,

cysteine or tyrosine residue at the C-terminal was found to be critical to the antioxidant activity based on both QSAR models and experimental observations.

Keywords: Quantitative structure-activity relationship, feature selection, machine learning, antioxidants, bioactive peptides, peptide synthesis

---

* Du, Z., Wang, D., & Li, Y. (2022). Comprehensive evaluation and comparison of machine learning methods in QSAR modeling of antioxidant tripeptides. ACS omega, 7(29), 25760-25771. https://doi.org/10.1021/acsomega.2c03062

2.1 Introduction

Antioxidants are useful in reducing and preventing the harmful effect of *in vivo* free radicals by donating electrons to neutralize them which induce cardiovascular diseases, cancers, and aging-related disorders (Lobo et al., 2010; Phongthai & Rawdkuen, 2020; Zhuang et al., 2013). Due to their multiple benefits, food protein-derived antioxidative peptides have gained increasing attention from today's consumers and researchers (Baakdah & Tsopmo, 2016; Girgih et al., 2013; Nimalaratne et al., 2015). Various *in vitro* antioxidant assays have been developed to evaluate antioxidant capacity, which are approximately divided into 2 categories, i.e., single-electron-transfer (SET) reaction and hydrogen-atom-transfer (HAT) reaction, respectively (Y.-Z. Zheng et al., 2017). *In vitro* assays based on SET reaction are generally preferred due to their convenience and accuracy (N. Chen et al., 2018).

Conventional ways for screening peptides with high antioxidant activity are based on sequential and rigorous wet chemistry steps, such as enzymatic hydrolysis and/or microbial fermentation to release or produce peptides, *in vitro* antioxidant assays to determine antioxidant activity, and advanced chromatography and spectroscopy (e.g., high performance liquid chromatograph-mass spectrometer) to purify and identify potential peptides (Ulug et al., 2021a). There are also studies that directly synthesized multiple peptides for screening on the basis of the theoretical knowledge (e.g., literature information of antioxidative peptides, important amino acids in peptides that contribute to antioxidant activity) (Gu et al., 2012; Saito et al., 2003; M. Tian et al., 2015; Ulug et al., 2021b; L. Zheng et al., 2016). Up to now, some high activity peptides have been found, such as Cys-Gln-Cys and Pro-His-His (H.-M. Chen et al., 1996; M. Tian et al., 2015). However, these conventional wet-chemistry methods for the preparation, fractionation, purification, and identification of antioxidative peptides and for screening

potentially high activity peptides are time-consuming and highly rely on many advanced instruments and trained personnel (Girgih et al., 2013; Phongthai & Rawdkuen, 2020; Ulug et al., 2021a).

Quantitative structure-activity relationship (QSAR) is a computational modeling method for revealing relationships between chemical structures of molecules and their bioactivity (Pripp & Ardo, 2007). In QSAR analysis, peptides are encoded by a series of numerical values, including properties of the amino acid residues (hydrophobicity, polarity, topological information, etc.) comprising the peptides and properties of the entire peptides (electronegativity, sequence information, solubility, molecular weight, topological information etc.) (Dai et al., 2014; Kawashima et al., 2008). Then, feature selection and modeling methods are combined to connect the structure information and bioactivity (Dai et al., 2014; Zhou et al., 2021). More than 80 amino acid descriptors (AADs) extracted from properties of amino acids by principal component analysis (PCA) were presented to characterize peptide structures and encode the peptides (Sandberg et al., 1998; F. Tian, Zhou, & Li, 2007; F. Tian, Zhou, Lv, et al., 2007; Yousefinejad et al., 2012). However, directly using these AADs usually led to undesirable model performance, since most of them were not intended for the antioxidant activity modeling (e.g., T-scale for angiotensin-converting enzyme inhibitory activity) (N. Chen et al., 2018; Cocchi & Johansson, 1993; Li et al., 2011b; Mahmoodi-Reihani et al., 2020; Shu et al., 2009; F. Tian, Zhou, & Li, 2007; Udenigwe & Aluko, 2011; Yang et al., 2010; Zaliani & Gancia, 1999; L. Zheng et al., 2016; Zhou et al., 2021).

Machine learning methods have been successfully applied for feature selection and model development in QSAR studies on peptide bioactivity (e.g., angiotensin converting enzyme inhibitory activity) (N. Chen et al., 2018; Dai et al., 2014; Deng et al., 2019; Kalyan et al., 2021;

F. Tian, Zhou, & Li, 2007; F. Tian, Zhou, Lv, et al., 2007; Uno et al., 2020). A total of 566 numerical values of amino acid, including physicochemical properties and biochemical properties of amino acids and pairs of amino acids have been available in AAIndex database (Kawashima et al., 2008). This makes it possible to use feature selection to find the important variables for bioactivity prediction compared with using AADs from PCA where the principal components were composed of various original variables. In addition, increasing studies on antioxidant peptides allowed compilation of datasets on their structures and activities (H.-M. Chen et al., 1996; N. Chen et al., 2018; Saito et al., 2003; M. Tian et al., 2015). Most previous studies on antioxidant peptides were still focused on the linear regression models, which would limit the model fitting to some extent due to synergic effect among the residues in peptides (N. Chen et al., 2018; Li et al., 2011b, 2011a; Li & Li, 2013; Udenigwe & Aluko, 2011; L. Zheng et al., 2016). Dataset division was another issue in most studies where the samples were sorted in a descending or ascending order by their activity, and training and test datasets were evenly selected from the samples based on the sorted sequence (e.g., first 3 for training dataset and the following 1 as the test dataset). The over-even dataset division strategy would undermine the model robustness since the bias in the test dataset could lead to poor model performance in cross validation than that in the test dataset (N. Chen et al., 2018; Deng et al., 2019; M. Tian et al., 2015; L. Zheng et al., 2016).

Previous studies reported that tripeptides exhibited higher antioxidant activity and better bioavailability than other oligopeptides and have diverse structural variations (H.-M. Chen et al., 1996, p. 199; M. Tian et al., 2015). The objective of this study was to compare popular machine learning methods for antioxidant activity modeling and screening of tripeptides and identify the critical amino acid features that determine the antioxidant activity (Figure 1). A total of 130

tripeptides with Trolox-equivalent antioxidant capacity (TEAC) values (SET reaction based) was manually collected from published studies for QSAR model development. Further, 553 numerical indices were firstly screened by pairwise correlation, followed by comparative evaluation using 7 different feature selection strategies. Description of the important feature variables from 553 numerical indices were developed. A total of 14 different advanced regression methods including both linear and non-linear methods were firstly used to develop models based on the extracted important variables, and the best model was used to predict tripeptides with high antioxidant potential for future study. Model performance of the 14 regression methods was compared and discussed. 6 tripeptides were synthesized and characterized for antioxidant activity to further evaluate the model performance for practical applications. Generalizability of these models were further tested by 20 times random dataset splitting and introduction of leave-one-group-our cross validation. This study provides a useful approach to screen the key factors influencing the antioxidant activity of tripeptides and a guideline for future application of various machine learning methods in QSAR modeling.

2.2 Materials and methods

2.2.1 Dataset collection

A total of 566 numerical indices of amino acids was collected using Beautiful Soup (4.5.3) from AAIndex (Kawashima et al., 2008; Richardson, 2007), and detailed definition and description of each index are available online (https://www.genome.jp/aaindex/). The indices with missing values for amino acids were deleted resulting in a total of 553 remaining indices (Supplementary Document S1-Table S1). Next, 130 antioxidant tripeptides were manually collected from BIOPEP-UWM (Minkiewicz et al., 2019), and their activities analyzed by 2,2'-azino-bis(3-ethylbenzothiazoline-6-sulfonic acid) (ABTS) radical scavenging activity assay were obtained

from the published literatures and expressed as the TEAC values (μmol TE / μmol peptide) (Table 1) (N. Chen et al., 2018; Ma et al., 2010; Saito et al., 2003). The tripeptides with no antioxidant activity (i.e., 0 μmol TE/μmol peptide) were all retained in the dataset for further QSAR model development.

2.2.2 Data processing

2.2.2.1 Pre-processing of numerical indices of amino acids

Pairwise correlation method was used to pre-screen collinearity of the 553 numerical indices (Supplementary Document S1-Table S2 & Figure S1). If an absolute value of Pearson's correlation coefficient between 2 indices was greater than 0.95, one of them was removed randomly due to the strong correlation (Sabilla et al., 2019). The remaining numerical indices were standardized for further feature selection.

2.2.2.2 Tripeptide encoding and feature selection

The pre-screened numerical indices of amino acids were used to encode tripeptides. Briefly, if the 'n' numerical indices were selected after the pre-processing, each amino acid was encoded as a '1 × n' vector (tripeptide was encoded as a '3 ×n' matrix). The matrix was transformed into a '1 ×3n' matrix, where the 1 to the n elements in the vector belonged to N-terminal residue, n+1 to 2n elements were referred to the central amino acid, and the 2n+1 to 3n elements belonged to the C-terminal residue (N. Chen et al., 2018).

After the encoding, each tripeptide is represented by 3n variables which were further screened by feature selection methods to identify the key variables for antioxidant activity prediction as amino acid descriptors. 6 representative feature selection methods were evaluated, namely linear regression based recursive feature elimination (RFE), named RFE-LR (Pedregosa et al., 2011); support vector machine regression (SVR) based RFE, named RFE-SVR (Pedregosa

et al., 2011); random forest regression (RFR) based RFE, named RFE-RFR (Pedregosa et al., 2011); feature coefficient (FC) based on lasso regression, named FC-LR (Pedregosa et al., 2011); feature importance based on RFR, named FI-RFR (Pedregosa et al., 2011); and feature importance based on extreme gradient boosting (XGB) regression, named FI-XGB (T. Chen & Guestrin, 2016). The detailed mathematical methodologies of these selection methods are available in scikit-learn (https://scikit-learn.org/).

After the feature selection, all the encoded tripeptides in the entire dataset were transformed into the new feature encoded version as X-matrix (variables) for further model development with tripeptides' activity values as responses (Y-vector). Furthermore, those encoded tripeptides without feature selection were also directly used as the X-matrix (variables) for model development and compared with the models developed by feature selection methods.

2.2.3 QSAR model development

2.2.3.1 Dataset division

Totally 130 samples were used for model development, cross validation, and model evaluation. The transformed X-matrix and Y-vector were shuffled and randomly split into training dataset and test dataset at a ratio of 3:1. 98 samples were used for training dataset to build models. The remaining 32 samples were used as test dataset to evaluate the performance of the models. The leave-one-out cross-validation (LOOCV) was utilized for the validation dataset split from training dataset.

2.2.3.2 Regression models

Fourteen popular regression models available through scikit-learn (https://scikit-learn.org/) and XGBoost (https://xgboost.readthedocs.io/en/stable/) were comparatively evaluated, namely tree based XGBoost regression (tree-XGB) (T. Chen & Guestrin, 2016), linear based XGBoost

regression (linear-XGB) (T. Chen & Guestrin, 2016), random forest regression (RFR) (Pedregosa et al., 2011), gradient boosting decision trees regression (GBDT) (J. H. Friedman, 2002), decision tree based bagging regression (Bagging) (Breiman, 1996), multi-layer perceptron regression (MLP) (Rumelhart et al., 1986), nearest neighbors regression (KNN) (Pedregosa et al., 2011), radial basis function kernel based support vector machine regression (rbf-SVR) (Pedregosa et al., 2011), linear kernel based support vector machine regression (linear-SVR) (Pedregosa et al., 2011), linear regression with L1 regularization (Lasso) (J. Friedman et al., 2010), linear regression with L2 regularization (Ridge) (Pedregosa et al., 2011), linear regression by minimizing a regularized empirical loss with stochastic gradient descent (SGD) (Pedregosa et al., 2011), ridge regression with kernel trick (KernelRidge) (Murphy, 2012), and Huber regression (Huber) (Huber, 1973).

2.2.3.3 QSAR model building and optimization

The model building was conducted using Python 3.8.8 with a computer (MacOS Monterey 12.0.1, CPU intel Core-i5 2.3 GHz). Models were imported from scikit-learn and XGBoost package (T. Chen & Guestrin, 2016; Pedregosa et al., 2011). LOOCV was used to avoid overfitting and tune the hyperparameters because of our small dataset size (Bo et al., 2021; Kartal et al., 2020; Nardo et al., 2020). The hyperparameters with the best performance from LOOCV were used as the final model for performance evaluation with the test dataset.

2.2.3.4 Model performance evaluation

Determination of coefficient (R^2) and root mean square error (RMSE) were used to evaluate the model performance. The R^2 and RMSE from the training dataset, LOOCV, and test dataset were named as R^2_{Train} and $RMSE_{Train}$, R^2_{CV} and $RMSE_{CV}$, and R^2_{Test} and $RMSE_{Test}$, respectively. To further evaluate the model generalizability, the developed models with the tunned

hyperparameters were rebuild by 20 times with different random dataset splitting and evaluated by R^2 and RMSE from the training dataset, LOOCV, Leave-one-group-out cross validation (LOGOCV) and test dataset. The result of the extra evaluation was available in Supplementary Document S1-Table S3.

2.2.4 Prediction of unpublished tripeptides with antioxidant activity from the models

A dataset containing 7870 potential tripeptides were built, and the published 130 tripeptides used for model building and validation were not included. After obtaining the model with best performance, the 7870 tripeptides were encoded by the selected features and used for the antioxidant activity prediction based on the selected model.

2.2.5 Model application for antioxidant tripeptides selection and tripeptides synthesis

The prediction results for the antioxidant activity of the 7870 unknown tripeptides showed that tyrosine, tryptophan and cystine at C-terminal residue were favorable to the antioxidant capacity. Considering the diversity of peptides, some unfavorable residues were also selected when designing tripeptide sequences for model validation. 6 tripeptides, namely QAY, PHC, YPQ, VYV, GPE, and YSQ were synthesized by Genscript Corp (Piscataway, NJ, USA) or purchased from Sigma Aldrich (St. Louis, MO, USA). The purity of the tripeptides was above 95% and the sequences were validated by liquid chromatography–mass spectrometry.

2.2.6 Characterization of antioxidant activity of the synthesized tripeptides

The ABTS radical scavenging capacity assay was based on the method described in the studies of Phongthai et al. and Chen et al. (N. Chen et al., 2018; Phongthai et al., 2018) with a few modifications. Briefly, stock solution was prepared by mixing 7.4 mM ABTS and 2.6 mM potassium persulfate in deionized (DI) water and incubating at room temperature for 12 h. The

working solution was made by diluting the stock solution till the absorbance of the mixture of ABTS•$^{+}$ solution and DI water at 734 nm was at 0.70 ± 0.02. Then, 150 μL ABTS•$^{+}$ solution was mixed with 50 μL tripeptide solution (20 μM) and allowed for 30 minutes incubation at 30 °C, and subsequently the absorbance was measured at 734 nm using a Biotek Synergy H1 Hybrid Microplate Reader (Winooski, VT, USA). Trolox (TE) was used as a standard antioxidant, and results were expressed as μmol TE / μmol peptide. All the chemicals and reagents used were of analytical grade and purchased from Sigma-Aldrich (St. Louis, MO, USA).

2.3. Results

2.3.1 Model development based on variables selected by FI-XGB

Five variables were selected by FI-XGB with a feature importance threshold of 0.01 (Table 2) and then used to encode the 130 tripeptides as the X-matrix (i.e., 130 × 5). Based on the variable importance results, C-terminal residues contributed the most to the antioxidant activity (Y-vector), while the central amino acids contributed the least to the activity. Among the 14 QSAR models (Table 3), tree-XGB achieved the best performance with R^2_{Test} and $RMSE_{test}$ of 0.814 and 0.692, respectively. The next satisfactory model was based on RFR (R^2_{Test} = 0.807 and $RMSE_{test}$ =0.698), while the R^2_{Test} of the remaining models were all below 0.8 which was less desirable. In general, the non-linear regression methods, including GBDT, MLP, Bagging, KNN, and rbf-SVR, achieved better performance than the linear regression methods such as linear-XGB, linear-SVR, Lasso, Ridge, SGD, and Huber.

2.3.2 Model development based on variables selected by FI-RFR

Fifteen variables were selected by FI-RFR with a threshold (feature importance = 0.01) (Table 2) and then used to encode the 130 tripeptides as the X-matrix (130 × 15). Based on the variable importance, C-terminal residues also contributed the most to the antioxidant activity (Y-

vector), while there was little contribution from the central amino acids based on these selected variables.

Among the 14 QSAR models (Table 3), Tree-XGB gained the best performance for test dataset, and the followings were RFR, rbf-SVR, bagging, and KNN respectively, while the model performance of RFR and rbf-SVR in LOOCV was not as good as that in the test dataset. For the remaining models where R^2_{Test} was below 0.8, GBDT as the only non-linear regression methods still gained better performance compared with these linear regression methods.

2.3.3 Model development based on variables selected by FC-LR

Eight variables were selected by FC-LR with a threshold (feature coefficient= 0.01) (Table 2) and then used to encode the 130 tripeptides as the X-matrix (130 × 8). Based on the variable importance, C-terminal residues contributed the most to the antioxidant activity (Y-vector), while the central amino acids contributed the least to the activity. Model performances of the 14 different regression methods are shown in Table 3. The KNN gained the best performance in test dataset (R^2_{Test} = 0.813 and $RMSE_{test}$ =0.693), while the R^2 of the remaining models was all less than 0.7 (Table 3). For the remaining models, linear regression methods (linear-XGB, linear SVR, lasso, Ridge, SGD, and Huber) achieved better performance than the nonlinear regression methods.

2.3.4 Model development based on variables selected by RFE-LR

Recursive feature elimination (RFE) eliminates one variable with the least feature importance or feature coefficient in one iteration, and the procedure is recursively repeated on the pruned dataset until achieving desired number of features. 17 variables were selected from RFE-LR (Table 2) and then used to encode the 130 tripeptides as the X-matrix (130 × 17). Based on the variable importance, N-terminal residues contributed the most to the antioxidant activity (Y-vector), while the central amino acids contributed the least to the activity. Model performances of

the 14 different regression methods are shown in Table 3. The MLP gained the best performance in test dataset (R^2_{Test} = 0.824 and $RMSE_{test}$ =0.672), followed by Huber (R^2_{Test} = 0.819 and $RMSE_{test}$ =0.681). The GBDT also provided a good result with R^2_{Test} larger than 0.8. From RFE-LR, linear-XGB, Ridge, and SGD as linear methods gained competitive performance compared with the nonlinear regression methods like KNN, rbf-SVR and RFR.

2.3.5 Model development based on variables selected by RFE-SVR

Seventeen variables were selected by RFE-SVR (Table 2) and then used to encode the 130 tripeptides as the X-matrix (130 × 17). Based on the variable importance, C-terminal residues and N-terminal residues contributed almost equally to the antioxidant activity (Y-vector), while the central amino acids contributed less to the activity. Model performances of the 14 different regression methods are shown in Table 3. Linear regression method, SGD, gained the best performance in test dataset (R^2_{Test} = 0.886 and $RMSE_{test}$ =0.541), while its performance in LOOCV was lower. The MLP and Huber were the next acceptable models with R^2_{Test} larger than 0.84. The KNN and linear-SVR also gained ideal performance. Based on the variables selected by RFE-SVR, there was no obvious difference between the linear and non-linear regression methods.

2.3.6 Model development based on variables selected by RFE-RFR

Fifteen variables were selected by RFE-SVR (Table 2) and then used to encode the 130 tripeptides as the X-matrix (130 × 15). Based on the variable importance, C-terminal residues contributed the most to the antioxidant activity (Y-vector), while the central amino acids contributed the least to the activity. Model performances of the 14 different regression methods are shown in Table 3. Tree-XGB achieved the best performance where R^2_{Test} and $RMSE_{test}$ were 0.828 and 0.665, respectively. For the remaining models, non-linear regression methods, even the worst one, KNN outperformed the linear regression methods.

2.3.7 Model development without feature selection

A total of 1026 variables were used to encode the 130 tripeptides as the X-matrix (130 × 1026). Model performances of the 14 different regression methods are shown in Table 3. RFR gained the best model performance where R^2_{Test} and $RMSE_{test}$ were 0.749 and 0.803, respectively. A significant overfitting was observed in MLP model, where the R^2_{Train} was 0.929, but the R^2_{Test} was only 0.404.

2.3.8 Prediction of unpublished tripeptides with antioxidant activity from the models

Based on the optimal values of R^2_{cv} and R^2_{test}, tree-XGB based on FI-RFR was used to predict the antioxidant activity of the 7870 unpublished tripeptides (Supplementary Document S1-Table S4). A total of 178 tripeptides with a C-terminal tyrosine was predicted to possess the highest antioxidant activity of 6.1672 μmol TE / μmol peptides, and the followings were the 167 tripeptides with a C-terminal tryptophan (6.1147 μmol TE / μmol peptides). Tripeptides with a C-terminal cysteine were predicted to have antioxidant activity of 6.0230 μmol TE / μmol peptides. As for the remaining tripeptides, there was no such obvious preferable amino acid residue at specific positions.

2.3.9 Application of QSAR model in synthetic tripeptide activity prediction

The experimental antioxidant activity of the synthetic tripeptides is summarized with their corresponding predicted activity in Table 4. QAY was predicted to be the most powerful antioxidant peptides (6.167 μmol TE / μmol peptide), while its observed activity was ranked second (4.270 μmol TE / μmol peptide) among the 6 synthesized tripeptides. PHC was also predicted to exhibit strong antioxidant activity (6.023 μmol TE / μmol peptide), and its observed activity (5.013 μmol TE / μmol peptides) was even stronger than QAY. Overall, the QSAR model has been very useful for the selection of potentially high antioxidant activity tripeptides, although

the antioxidant activity from the model was a little bit overestimated compared to the experimental results.

2.4 Discussion

Various numerical indices were screened and selected by the 6 different feature selection methods. Based on the variable importance values, almost all the feature selection methods showed that the C-terminal residues played the most important role in antioxidant activity, while the central amino acid contributed the least to the activity, which was partly consistent with previous results from wet chemistry and QSAR studies, where there was no comparison between N- and C-terminals (Li et al., 2011b; Saito et al., 2003; Uno et al., 2020). Previous studies were confined to amino acid physicochemical properties (with about 195 indices) or the AADs which could not take full advantage of all the amino acid indices to identify the most representative indices to characterize tripeptides (N. Chen et al., 2018; Deng et al., 2019; L. Zheng et al., 2016). Although some of the selected features, especially non-physicochemical properties (e.g., LIFS790103 standing for "Conformational preference for antiparallel beta-strands") might be difficult to be understood and explained, these selected features are much targeted and less redundant (N. Chen et al., 2018). For those AADs derived from PCA analysis, each principal component was composed of multiple original properties, and there are usually several principal components adopted in the model development, which can only be roughly explained (e.g., the first component was related to hydrophobicity), but impeded the further explanation of the feature importance and distracted the application of these features for peptide design and modification (F. Tian, Zhou, & Li, 2007; Yang et al., 2010). Even though some features are difficult to explain here, they all have the standard protocols to be determined and that would be easier when applying in the structure design and modification of bioactive peptides.

Among these selected features, some of them, such as CHOP780215, LIFS790103, OOBM850102, and WEBA780101, were selected multiple times for the characterization of C-terminal residues by different feature selection strategies, which showed their importance in antioxidant activity prediction. CHOP780215 was not only selected for encoding C-terminal residues by FC-LR, RFE-LR and RFE-SVR, but also selected by RFE-RFR and FI-XGB to characterize N-terminal residues. Some features, such as GEOR030105 and PALJ810116 selected by REF-SVR and GEIM800110 selected by REF-LR, were used to encode both the central and N-terminal residues, and central and C-terminal residues, respectively. This implies that some features of amino acids can contribute to antioxidant activity at any position, even though their importance varied with positions. The theoretical conclusion derived from the selected features was also support by the study of Uno et al (Uno et al., 2020).

For the models without feature selection, inferior performance was observed as shown in Table 3. The main reason of the poor performance under this preprocessing method was mainly because of the high dimensionality on the features and small sample size. Therefore, the significant improvement in model performance was achieved by feature selection because plenty of irrelevant features were eliminated (Dai et al., 2014).

For the 14 different regression methods, nonlinear regression methods overall achieved better model performance based on the 6 feature selection methods, which proved existing of non-linearity in antioxidant activity prediction. This also explained the poor model performance in most previous studies which were based on linear regression methods (Deng et al., 2019; Uno et al., 2020; L. Zheng et al., 2016). In addition, some studies subjectively removed the non-active tripeptides from the dataset in order to improve the model fitting, which resulted in misleading models (N. Chen et al., 2018; Deng et al., 2019). Further, improper dataset division between

training dataset and test dataset increased the bias in the model and undermined the robustness of the models (N. Chen et al., 2018), while model evaluation without test dataset was not complete because the performance in cross-validation could not represent the real performance of the model in unknown dataset (Deng et al., 2019). In this study, these biases were all overcome, and the performance was greatly improved compared with the most recent study on tripeptides (N. Chen et al., 2018; Deng et al., 2019). In fact, we also adopted processing methods using the bias-existing data from literature to develop these models during preliminary study, and the R^2_{test} could be larger than 0.9559, which also proved the bias in the previous studies. Abnormal phenomena were observed in some models (e.g., rbf-SVR regression based on FI-RFR) where performance in LOOCV was poorer than that in the test dataset. That implied the overfitting of these models, and the same situation was difficult to avoid in bioactivity prediction, since LOOCV was a optimistic cross validation method (N. Chen et al., 2018). The main reason that the n-fold cross validation was not used in our study was primarily due to the relatively small dataset. In order to further evaluate the generalizability of these models, we introduced the more challenging cross validation (LOGOCV) as well as 20 times of random dataset splitting for the model evaluation (Supplementary Document S1-Table S3). Performance of those overfitting strategies suffered more in generalizability evaluation. It also can be seen that XGBoost regression method with random forest regression for feature selection was the most powerful and robust strategies for antioxidant activity prediction.

From the prediction of tripeptides with potential high antioxidant activity, the 525 unpublished tripeptides with activity higher than 6 μmol TE / μmol peptides all had a tyrosine or tryptophan or cysteine residue at C-terminal position, which was consistent with previous studies

(Li et al., 2011b; Saito et al., 2003; Uno et al., 2020). Compared with previous studies, our model clearly specifies the tripeptides with most promising antioxidant activity.

The preferable attributes of strong antioxidant tripeptides concluded from the model development were supported by the antioxidant activity determination from the synthetic tripeptides. It was observed that tripeptides with tyrosine and cysteine residues at C-terminal exhibited the highest antioxidant activity compared to that with tyrosine residue at the N-terminal, which also showed lower contribution to antioxidant activity in the feature importance analysis. In addition, the model successfully predicted that the tripeptide (PHC) with a cysteine residue at C-terminal had strong antioxidant activity (Table 1), which had not been reported previously. In addition, there was no tripeptide with tyrosine at C-terminal showing high antioxidant activity (e.g., above 4 μmol TE / μmol peptide). Our results supported the hypothesis of the model development that these amino acid indices had the capacity to represent the residues in tripeptide for unknown antioxidant tripeptide activity prediction. The deviation between observed and predicted activity was inevitable, but it is overall acceptable (N. Chen et al., 2018; Uno et al., 2020).

2.5 Conclusion

In this study, we collected 553 latest amino acid numerical indices and 130 published tripeptides with available TEAC values for QSAR analysis. 7 feature selection strategies and 14 regression methods were combined to build QSAR models and used to comprehensively evaluate the performance of the application of machine learning methods in predicting antioxidant tripeptides. Results showed that C-terminal residues played a more important role in antioxidant activity and non-linear regression methods were more suitable for QSAR study on antioxidant activity. The best model based on FI-RFR for feature selection plus tree-based XGB for model

building was used to predict the antioxidant activities of the unknown 7870 tripeptides, and the high activity tripeptides have tyrosine, tryptophan, or cysteine residue at the C-terminal position. Furthermore, 6 unpublished tripeptides were synthesized and characterized to evaluate the practical application of the best model. The predicted activity can reflect the rank of the potential activity of these tripeptides and their approximate activity, although there was an overestimation. This study also for the first time demonstrates through both *in silico* and wet chemistry experiment that cysteine and tyrosine residues at C-terminal are highly corresponding to antioxidant activity for tripeptide. In addition, this study also provides critical reference for antioxidant tripeptide screening and a useful model development template for future QSAR studies on bioactive peptides.

Declaration of interest

The authors declare that there is no known conflict of interest.

Acknowledgements

This is contribution No. 22-159-J from the Kansas Agricultural Experimental Station. This research was supported by the Agriculture and Food Research Initiative Competitive Grant no. 2020-68008-31408 and no. 2021-67021-34495 from the USDA National Institute of Food and Agriculture.

References

Baakdah, M. M., & Tsopmo, A. (2016). Identification of peptides, metal binding and lipid peroxidation activities of HPLC fractions of hydrolyzed oat bran proteins. *Journal of Food Science and Technology*, *53*(9), 3593–3601. https://doi.org/10/f89dmq

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10/fm3fwj

Chen, H.-M., Muramoto, K., Yamauchi, F., & Nokihara, K. (1996). Antioxidant Activity of Designed Peptides Based on the Antioxidative Peptide Isolated from Digests of a Soybean Protein. *Journal of Agricultural and Food Chemistry*, *44*(9), 2619–2623. https://doi.org/10.1021/jf950833m

Chen, N., Chen, J., Yao, B., & Li, Z. (2018). QSAR Study on Antioxidant Tripeptides and the Antioxidant Activity of the Designed Tripeptides in Free Radical Systems. *Molecules*, *23*(6), 1407. https://doi.org/10/gdttrq

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10/gdp84q

Cocchi, M., & Johansson, E. (1993). Amino Acids Characterization by GRID and Multivariate Data Analysis. *Quantitative Structure-Activity Relationships*, *12*(1), 1–8. https://doi.org/10/ft5vbj

Dai, Z., Wang, L., Chen, Y., Wang, H., Bai, L., & Yuan, Z. (2014). A pipeline for improved QSAR analysis of peptides: Physiochemical property parameter selection via BMSF, near-neighbor sample selection via semivariogram, and weighted SVR regression and prediction. *Amino Acids*, *46*(4), 1105–1119. https://doi.org/10/f5vxh2

Deng, B., Long, H., Tang, T., Ni, X., Chen, J., Yang, G., Zhang, F., Cao, R., Cao, D., Zeng, M., & Yi, L. (2019). Quantitative Structure-Activity Relationship Study of Antioxidant Tripeptides Based on Model Population Analysis. *International Journal of Molecular Sciences*, *20*(4), 995. https://doi.org/10/gnmdb7

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. https://doi.org/10/fxb956

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1. https://doi.org/10/bb3d

Girgih, A. T., Udenigwe, C. C., Hasan, F. M., Gill, T. A., & Aluko, R. E. (2013). Antioxidant properties of Salmon (Salmo salar) protein hydrolysate and peptide fractions isolated by reverse-phase HPLC. *Food Research International*, *52*(1), 315–322. https://doi.org/10/f4zm2b

Gu, L., Zhao, M., Li, W., You, L., Wang, J., Wang, H., & Ren, J. (2012). Chemical and cellular antioxidant activity of two novel peptides designed based on glutathione structure. *Food and Chemical Toxicology*, *50*(11), 4085–4091. https://doi.org/10.1016/j.fct.2012.08.028

Huber, P. J. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics*, *1*(5), 799–821. https://doi.org/10/b7kv6r

Kalyan, G., Junghare, V., Khan, M. F., Pal, S., Bhattacharya, S., Guha, S., Majumder, K., Chakrabarty, S., & Hazra, S. (2021). Anti-hypertensive Peptide Predictor: A Machine Learning-Empowered Web Server for Prediction of Food-Derived Peptides with Potential Angiotensin-Converting Enzyme-I Inhibitory Activity. *Journal of Agricultural and Food Chemistry*, *69*(49), 14995–15004. https://doi.org/10.1021/acs.jafc.1c04555

Kawashima, S., & Kanehisa, M. (n.d.). *AAindex: Amino Acid index database*. 1.

Li, Y.-W., & Li, B. (2013). Characterization of structure–antioxidant activity relationship of peptides in free radical systems using QSAR models: Key sequence positions and their amino acid properties. *Journal of Theoretical Biology*, *318*, 29–43. https://doi.org/10/f4hh8h

Li, Y.-W., Li, B., He, J., & Qian, P. (2011a). Quantitative structure–activity relationship study of antioxidative peptide by using different sets of amino acids descriptors. *Journal of Molecular Structure*, *998*(1–3), 53–61. https://doi.org/10.1016/j.molstruc.2011.05.011

Li, Y.-W., Li, B., He, J., & Qian, P. (2011b). Structure-activity relationship study of antioxidative peptides by QSAR modeling: The amino acid next to *C* -terminus affects the activity: QSAR STUDY OF ANTIOXIDATIVE PEPTIDES. *Journal of Peptide Science*, *17*(6), 454–462. https://doi.org/10/bz9smv

Lobo, V., Patil, A., Phatak, A., & Chandra, N. (2010). Free radicals, antioxidants and functional foods: Impact on human health. *Pharmacognosy Reviews*, *4*(8), 118–126. https://doi.org/10/fh6cc4

Ma, Y., Xiong, Y. L., Zhai, J., Zhu, H., & Dziubla, T. (2010). Fractionation and evaluation of radical scavenging peptides from in vitro digests of buckwheat protein. *Food Chemistry*, *118*(3), 582–588. https://doi.org/10/dr2jtv

Mahmoodi-Reihani, M., Abbasitabar, F., & Zare-Shahabadi, V. (2020). In Silico Rational Design and Virtual Screening of Bioactive Peptides Based on QSAR Modeling. *ACS Omega*, *5*(11), 5951–5958. https://doi.org/10/gnmdcr

Minkiewicz, Iwaniak, & Darewicz. (2019). BIOPEP-UWM Database of Bioactive Peptides: Current Opportunities. *International Journal of Molecular Sciences*, *20*(23), 5978. https://doi.org/10/gnmt2w

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press.

Nimalaratne, C., Bandara, N., & Wu, J. (2015). Purification and characterization of antioxidant peptides from enzymatically hydrolyzed chicken egg white. *Food Chemistry*, *188*, 467–472. https://doi.org/10/f79msm

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (n.d.). Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*, 6.

Phongthai, S., D'Amico, S., Schoenlechner, R., Homthawornchoo, W., & Rawdkuen, S. (2018). Fractionation and antioxidant properties of rice bran protein hydrolysates stimulated by in vitro gastrointestinal digestion. *Food Chemistry*, *240*, 156–164. https://doi.org/10/gngrk9

Phongthai, S., & Rawdkuen, S. (2020). Fractionation and characterization of antioxidant peptides from rice bran protein hydrolysates stimulated by in vitro gastrointestinal digestion. *Cereal Chemistry*, *97*(2), 316–325. https://doi.org/10/gngp85

Pripp, A., & Ardo, Y. (2007). Modelling relationship between angiotensin-(I)-converting enzyme inhibition and the bitter taste of peptides. *Food Chemistry*, *102*(3), 880–888. https://doi.org/10/b7sxf6

Richardson, L. (2007). *Beautiful soup documentation*.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10/cvjdpk

Sabilla, S., Sarno, R., Institut Teknologi Sepuluh Nopember, Triyana, K., & Universitas Gadjah Mada Sekip Utara. (2019). Optimizing Threshold using Pearson Correlation for Selecting Features of Electronic Nose Signals. *International Journal of Intelligent Engineering and Systems*, *12*(6), 81–90. https://doi.org/10/gw3m

Saito, K., Jin, D.-H., Ogawa, T., Muramoto, K., Hatakeyama, E., Yasuhara, T., & Nokihara, K. (2003). Antioxidative Properties of Tripeptide Libraries Prepared by the Combinatorial Chemistry. *Journal of Agricultural and Food Chemistry*, *51*(12), 3668–3674. https://doi.org/10/c2f6wr

Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., & Wold, S. (1998). New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *Journal of Medicinal Chemistry*, *41*(14), 2481–2491. https://doi.org/10/d88tzn

Shu, M., Mei, H., Yang, S., Liao, L., & Li, Z. (2009). Structural Parameter Characterization and Bioactivity Simulation Based on Peptide Sequence. *QSAR & Combinatorial Science*, *28*(1), 27–35. https://doi.org/10.1002/qsar.200710169

Tian, F., Zhou, P., & Li, Z. (2007). T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides. *Journal of Molecular Structure*, *830*(1–3), 106–115. https://doi.org/10/dqr84q

Tian, F., Zhou, P., Lv, F., Song, R., & Li, Z. (2007). Three-dimensional holograph vector of atomic interaction field (3D-HoVAIF): A novel rotation-translation invariant 3D structure descriptor and its applications to peptides. *Journal of Peptide Science*, *13*(8), 549–566. https://doi.org/10/c775td

Tian, M., Fang, B., Jiang, L., Guo, H., Cui, J., & Ren, F. (2015). Structure-activity relationship of a series of antioxidant tripeptides derived from β-Lactoglobulin using QSAR modeling. *Dairy Science & Technology*, *95*(4), 451–463. https://doi.org/10/f7gdk9

Udenigwe, C. C., & Aluko, R. E. (2011). Chemometric Analysis of the Amino Acid Requirements of Antioxidant Food Protein Hydrolysates. *International Journal of Molecular Sciences*, *12*(5), 3148–3161. https://doi.org/10/fc8829

Ulug, S. K., Jahandideh, F., & Wu, J. (2021a). Novel technologies for the production of bioactive peptides. *Trends in Food Science & Technology*, *108*, 27–39. https://doi.org/10.1016/j.tifs.2020.12.002

Ulug, S. K., Jahandideh, F., & Wu, J. (2021b). Novel technologies for the production of bioactive peptides. *Trends in Food Science & Technology*, *108*, 27–39. https://doi.org/10/gnnsx5

Uno, S., Kodama, D., Yukawa, H., Shidara, H., & Akamatsu, M. (2020). Quantitative analysis of the relationship between structure and antioxidant activity of tripeptides. *Journal of Peptide Science*, *26*(3). https://doi.org/10/gnmdb5

Yang, L., Shu, M., Ma, K., Mei, H., Jiang, Y., & 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. https://doi.org/10.1007/s00726-009-0287-y

Yousefinejad, S., Hemmateenejad, B., & Mehdipour, A. R. (2012). New autocorrelation QTMS-based descriptors for use in QSAM of peptides. *Journal of the Iranian Chemical Society*, *9*(4), 569–577. https://doi.org/10/gnmdcs

Zaliani, A., & Gancia, E. (1999). MS-WHIM Scores for Amino Acids: A New 3D-Description for Peptide QSAR and QSPR Studies. *Journal of Chemical Information and Computer Sciences*, *39*(3), 525–533. https://doi.org/10/fxb9dw

Zheng, L., Zhao, Y., Dong, H., Su, G., & Zhao, M. (2016). Structure–activity relationship of antioxidant dipeptides: Dominant role of Tyr, Trp, Cys and Met residues. *Journal of Functional Foods*, *21*, 485–496. https://doi.org/10/f8cqgs

Zheng, Y.-Z., Deng, G., Liang, Q., Chen, D.-F., Guo, R., & Lai, R.-C. (2017). Antioxidant Activity of Quercetin and Its Glucosides from Propolis: A Theoretical Study. *Scientific Reports*, *7*(1), 7543. https://doi.org/10/gbsvpb

Zhou, P., Liu, Q., Wu, T., Miao, Q., Shang, S., Wang, H., Chen, Z., Wang, S., & Wang, H. (2021). Systematic Comparison and Comprehensive Evaluation of 80 Amino Acid

Descriptors in Peptide QSAR Modeling. *Journal of Chemical Information and Modeling*, *61*(4), 1718–1731. https://doi.org/10.1021/acs.jcim.0c01370

Zhuang, H., Tang, N., & Yuan, Y. (2013). Purification and identification of antioxidant peptides from corn gluten meal. *Journal of Functional Foods*, *5*(4), 1810–1821. https://doi.org/10/f6dzjp

Figures and tables

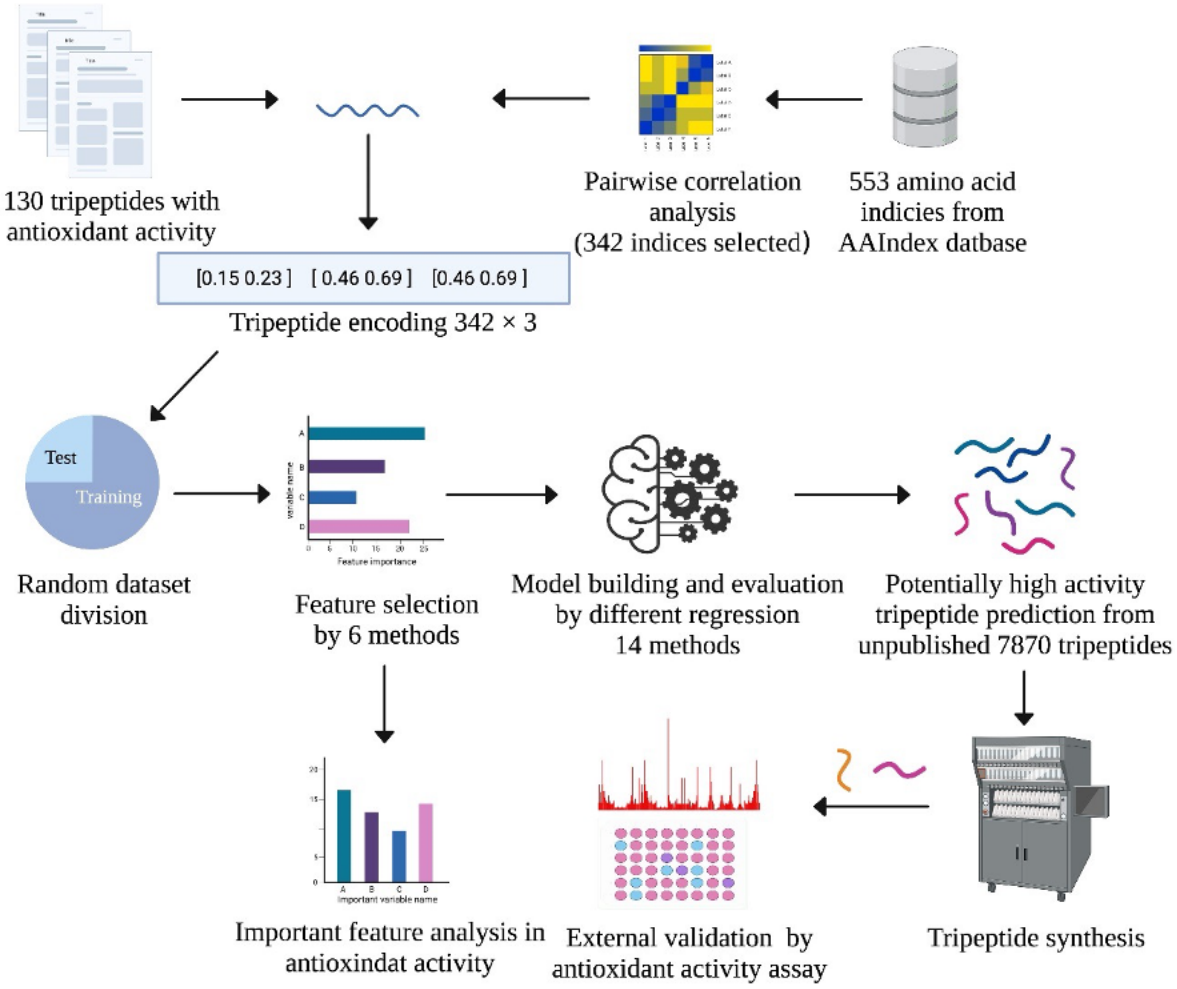


Figure 2-1 Flowchart for quantitative structure-activity relationship (QSAR) modeling and validation of antioxidant tripeptides.

Table 2-1 Sequence and Trolox-equivalent antioxidant capacity (TEAC) of tripeptides (μmol TE / μmol peptide) dataset from literature.

No.	Sequence	Activity	No.	Sequence	Activity	No.	Sequence	Activity
1	LHA	0	47	PHN	0.24	93	PWR	0.822
2	LHD	0	48	LWF	0.25	94	PWI	0.832
3	LHE	0	49	PWD	0.262	95	RWG	0.842
4	LHF	0	50	LVG	0.266	96	LWN	0.866
5	LHG	0	51	PHH	0.266	97	LWR	0.869
6	LHH	0	52	PWE	0.339	98	PWL	0.88
7	LHQ	0	53	PHI	0.344	99	PWT	0.9
8	PHA	0	54	PHQ	0.348	100	PWN	0.943
9	PHD	0	55	GHG	0.365	101	RWH	0.995
10	PHE	0	56	LWD	0.402	102	RWQ	0.995
11	PHF	0	57	LWG	0.406	103	KHP	1.143
12	PHM	0	58	RHS	0.409	104	GVR	1.157
13	RHA	0	59	PWA	0.414	105	ECG	1.413
14	RHD	0	60	GHP	0.426	106	PHW	1.768
15	RHE	0	61	PWS	0.44	107	PWW	1.774
16	RHH	0	62	PWV	0.457	108	RWW	1.837
17	RHK	0	63	RWD	0.485	109	LHW	1.84
18	RHQ	0	64	LWM	0.49	110	LWW	1.931
19	RHT	0	65	PHG	0.496	111	WPL	1.972
20	PHT	0.028	66	RWA	0.497	112	VPW	1.972
21	LHM	0.031	67	PWM	0.498	113	RHW	2.203
22	LHN	0.046	68	LWV	0.499	114	LWY	2.332
23	GVT	0.047	69	RWV	0.51	115	RWY	2.334
24	PHS	0.058	70	LWL	0.515	116	RHY	2.464
25	KHR	0.067	71	LWQ	0.519	117	PHY	2.707
26	GHT	0.079	72	LWS	0.522	118	LHY	2.753
27	LWH	0.098	73	LWA	0.594	119	PWY	2.785
28	LHK	0.108	74	RWS	0.6	120	GVW	4.365
29	LHR	0.108	75	RHF	0.6	121	GKW	4.687
30	LHT	0.108	76	LWT	0.627	122	GHW	4.745
31	LHV	0.108	77	LWI	0.628	123	QVW	5.161
32	RHR	0.118	78	LWK	0.629	124	KVW	5.218
33	PHK	0.176	79	PWH	0.632	125	NKW	5.349
34	LHL	0.186	80	PWK	0.634	126	NHW	5.368
35	RHI	0.189	81	PWQ	0.637	127	QHW	5.524

36	PHV	0.198	82	RWR	0.651	128	KHW	5.566
37	PWF	0.202	83	RWT	0.651	129	PYW	5.683
38	PWG	0.203	84	RWE	0.663	130	YHW	6.169
39	RHG	0.203	85	LHS	0.68			
40	RHL	0.206	86	RWF	0.689			
41	RHM	0.207	87	RWL	0.689			
42	RHN	0.208	88	RWI	0.702			
43	PHR	0.211	89	RWM	0.702			
44	RHV	0.212	90	RWN	0.702			
45	LHI	0.217	91	RWK	0.753			
46	PHL	0.238	92	LWE	0.777			

Table 2-2 Amino acid positions, variable importance, and description of selected variables from 7 feature selection strategies.

AAindex accession number	Amino acid position	Variable importance	Description	Note
Selected variables by FI-XGB				
BURA740101	N-terminal	0.0199	Normalized frequency of alpha-helix	
CHOP780215	N-terminal	0.161	Frequency of the 4th residue in turn	A
BEGF750102	Central	0.036	Conformational parameter of beta-structure	
KANM800103	C-terminal	0.0138	Average relative probability of inner helix	
LIFS790103	C-terminal	0.7049	Conformational preference for antiparallel beta-strands	B
Selected variables by FI-RFR				
PALJ810113	N-terminal	0.025	Normalized frequency of turn in all-alpha class	
ONEK900102	N-terminal	0.0108	Helix formation parameters (delta delta G)	
FUKS010101	N-terminal	0.015	Surface composition of amino acids in intracellular proteins of thermophiles (percent)	
JOND750102	C-terminal	0.0171	pK (-COOH)	
LIFS790103	C-terminal	0.0518	Conformational preference for antiparallel beta-strands	B
MCMT640101	C-terminal	0.0286	Refractivity	
NAKH920102	C-terminal	0.0688	AA composition of CYT2 of single-spanning proteins	
OOBM850102	C-terminal	0.037	Optimized propensity to form reverse turn	C
WEBA780101	C-terminal	0.0371	RF value in high salt chromatography	D
VINM940102	C-terminal	0.051	Normalized flexibility parameters (B-values) for each residue surrounded by none rigid neighbors	
PARS000101	C-terminal	0.0367	p-Values of mesophilic proteins based on the distributions of B values	N
PARS000102	C-terminal	0.0768	p-Values of thermophilic proteins based on the distributions of B values	K
FODM020101	C-terminal	0.0416	Free energy change of epsilon(i) to alpha(Rh)	E
MITS020101	C-terminal	0.1532	Amphiphilicity index	F
DIGM050101	C-terminal	0.0563	Hydrostatic pressure asymmetry index, PAI	G

Selected variables by FC-LR				
MAXF760103	N-terminal	0.025	Normalized frequency of zeta R	
NAKH900102	N-terminal	0.0371	SD of AA composition of total proteins	
QIAN880114	N-terminal	0.051	Weights for beta-sheet at the window position of -6	
KHAG800101	Central	0.0367	The Kerr-constant increments	
CHOP780215	C-terminal	0.0768	Frequency of the 4th residue in turn	A
OOBM850102	C-terminal	0.0416	Optimized propensity to form reverse turn	C
WEBA780101	C-terminal	0.0153	RF value in high salt chromatography	D
MITS020101	C-terminal	0.0563	Amphiphilicity index	F
Selected variables by RFE-LR				
WERD780102	N-terminal	0.3136	Free energy change of epsilon(i) to epsilon(ex)	
AURR980107	N-terminal	0.9357	Normalized positional residue frequency at helix termini N2	
AURR980111	N-terminal	2.0325	Normalized positional residue frequency at helix termini C5	H
AURR980116	N-terminal	1.5792	Normalized positional residue frequency at helix termini Cc	
CEDJ970105	N-terminal	0.2991	Composition of amino acids in nuclear proteins (percent)	I
KARS160120	N-terminal	0.5644	Weighted minimum eigenvalue based on the atomic numbers	
CHOC760104	Central	0.8074	Proportion of residues 100% buried	
GEIM800110	Central	0.6058	Aperiodic indices for beta-proteins	J
QIAN880136	Central	0.9265	Weights for coil at the window position of 3	
KARS160113	Central	0.7146	Weighted domination number using the atomic number	
CHOP780215	C-terminal	0.8292	Frequency of the 4th residue in turn	A
GEIM800110	C-terminal	0.0872	Aperiodic indices for beta-proteins	J
HUTJ700101	C-terminal	0.271	Heat capacity	
HUTJ700103	C-terminal	0.3975	Entropy of formation	
KARP850102	C-terminal	0.6556	Flexibility parameter for one rigid neighbor	
NAKH900110	C-terminal	1.1114	Normalized composition of membrane proteins	
WILM950102	C-terminal	0.66043	Hydrophobicity coefficient in RP-HPLC, C8 with 0.1%TFA/MeCN/H2O	
Selected variables by RFE-SVR				
CHAM820102	N-terminal	0.303	Free energy of solution in water, kcal/mole	
NAKH920101	N-terminal	0.4642	AA composition of CYT of single-spanning proteins	
RICJ880114	N-terminal	0.1628	Relative preference value at C1	

PARS000102	N-terminal	0.2303	p-Values of thermophilic proteins based on the distributions of B values	K
CEDJ970105	N-terminal	0.1578	Composition of amino acids in nuclear proteins (percent)	I
GEOR030105	N-terminal	0.0656	Linker propensity from small dataset (linker length is less than six residues)	L
GEIM800106	Central	0.2531	Beta-strand indices for beta-proteins	
NAKH900108	Central	0.1859	Normalized composition from fungi and plant	
PALJ810116	Central	0.1345	Normalized frequency of turn in alpha/beta class	M
GEOR030105	Central	0.1183	Linker propensity from small dataset (linker length is less than six residues)	L
CHOP780215	C-terminal	0.1984	Frequency of the 4th residue in turn	A
OOBM850102	C-terminal	0.3586	Optimized propensity to form reverse turn	C
PALJ810116	C-terminal	0.1983	Normalized frequency of turn in alpha/beta class	M
WERD780104	C-terminal	0.2361	Free energy change of epsilon(i) to alpha (Rh)	
PARS000101	C-terminal	0.3056	p-Values of mesophilic proteins based on the distributions of B values	N
MITS020101	C-terminal	0.0605	Amphiphilicity index	F
DIGM050101	C-terminal	0.1109	Hydrostatic pressure asymmetry index, PAI	G
Selected variables by RFE-RFR				
CHOP780215	N-terminal	0.0336	Frequency of the 4th residue in turn	A
ISOY800108	N-terminal	0.0294	Normalized relative frequency of coil	
MAXF760104	N-terminal	0.0341	Normalized frequency of left-handed alpha-helix	
GEOR030105	N-terminal	0.0486	Linker propensity from small dataset (linker length is less than six residues)	L
KARS160122	N-terminal	0.0362	Weighted second smallest eigenvalue of the weighted Laplacian matrix	
QIAN880127	Central	0.0362	Weights for coil at the window position of -6	
AURR980111	Central	0.0291	Normalized positional residue frequency at helix termini C5	H
LIFS790103	C-terminal	0.1127	Conformational preference for antiparallel beta-strands	B
MCMT640101	C-terminal	0.0969	Refractivity	
OOBM850102	C-terminal	0.0462	Optimized propensity to form reverse turn	C
WEBA780101	C-terminal	0.0245	Normalized frequency of turn in all-alpha class	D
PARS000102	C-terminal	0.0745	p-Values of mesophilic proteins based on the distributions of B values	K

FODM020101	C-terminal	0.1246	Free energy change of epsilon(i) to alpha(Rh)	E
MITS020101	C-terminal	0.2131	Amphiphilicity index	F
DIGM050101	C-terminal	0.0603	Hydrostatic pressure asymmetry index, PAI	G

Note: Detailed information of these selected variables are available at https://www.genome.jp/aaindex/ . The same capitalized letter in the last column indicates same amino acid features.

Table 2-3 Performance of 14 QSAR models based on the 7 feature selection strategies.

Model	Training dataset				Test dataset		Note
	R^2_{Train}	$RMSE_{Train}$	R^2_{CV}	$RMSE_{CV}$	R^2_{Test}	$RMSE_{Test}$	
QSAR models based on FI-XGB							
tree-XGB	0.955	0.295	0.911	0.416	0.814	0.692	***
linear-XGB	0.566	0.921	0.478	1.01	0.558	1.067	
RFR	0.956	0.295	0.924	0.386	0.807	0.698	**
GBDT	0.976	0.219	0.904	0.434	0.78	0.752	*
Bagging	0.974	0.226	0.904	0.434	0.769	0.77	
MLP	0.961	0.276	0.847	0.548	0.77	0.769	
KNN	0.84	0.559	0.598	0.887	0.555	1.069	
rbf-SVR	0.965	0.263	0.831	0.574	0.726	0.84	
linear-SVR	0.387	1.095	0.355	1.123	0.345	1.298	
Lasso	0.575	0.912	0.473	1.015	0.59	1.027	
Ridge	0.566	0.921	0.478	1.01	0.557	1.068	
SGD	0.516	0.973	0.424	1.062	0.49	1.146	
KernelRidge	0.074	1.346	-0.073	1.448	0.206	1.429	
Huber	0.567	0.92	0.478	1.01	0.559	1.064	
QSAR models based on FI-RFR							
tree-XGB	0.954	0.3	0.872	0.5	0.847	0.627	***
linear-XGB	0.789	0.643	0.722	0.738	0.681	0.906	
RFR	0.928	0.375	0.842	0.556	0.854	0.613	**
GBDT	0.978	0.207	0.866	0.512	0.781	0.75	
Bagging	0.962	0.274	0.833	0.571	0.822	0.677	
MLP	0.976	0.219	0.82	0.592	0.773	0.764	
KNN	0.933	0.362	0.832	0.573	0.814	0.691	

rbf-SVR	0.954	0.3	0.832	0.574	0.844	0.632	*
linear-SVR	0.78	0.655	0.709	0.755	0.623	0.984	
Lasso	0.796	0.632	0.714	0.748	0.685	0.901	
Ridge	0.779	0.657	0.721	0.739	0.679	0.909	
SGD	0.789	0.642	0.724	0.735	0.674	0.916	
KernelRidge	0.279	1.187	-0.118	1.479	0.295	1.346	
Huber	0.792	0.637	0.719	0.742	0.682	0.904	
QSAR models based on FC-LR							
tree-XGB	0.977	0.214	0.883	0.477	0.707	0.868	
linear-XGB	0.827	0.582	0.775	0.663	0.783	0.748	**
RFR	0.983	0.182	0.923	0.389	0.652	0.946	
GBDT	0.991	0.134	0.928	0.377	0.626	0.981	
Bagging	0.989	0.145	0.92	0.396	0.681	0.905	
MLP	0.975	0.221	0.771	0.669	0.763	0.781	
KNN	0.94	0.342	0.835	0.568	0.813	0.693	***
rbf-SVR	0.988	0.155	0.741	0.711	0.716	0.855	
linear-SVR	0.815	0.602	0.756	0.691	0.782	0.75	
Lasso	0.817	0.598	0.759	0.686	0.739	0.819	
Ridge	0.821	0.592	0.779	0.658	0.777	0.757	
SGD	0.826	0.584	0.771	0.669	0.785	0.743	
KernelRidge	0.319	1.155	0.034	1.375	0.37	1.272	
Huber	0.829	0.579	0.759	0.687	0.786	0.741	*
QSAR models based on RFE-LR							
tree-XGB	0.951	0.31	0.801	0.624	0.773	0.764	
linear-XGB	0.849	0.542	0.752	0.697	0.78	0.752	
RFR	0.939	0.345	0.793	0.636	0.737	0.823	
GBDT	0.986	0.164	0.821	0.592	0.8	0.718	*

Bagging	0.976	0.217	0.815	0.601	0.766	0.775	
MLP	0.979	0.202	0.868	0.509	0.824	0.672	***
KNN	0.859	0.526	0.749	0.701	0.627	0.98	
rbf-SVR	0.993	0.118	0.774	0.666	0.569	1.053	
linear-SVR	0.887	0.47	0.781	0.654	0.634	0.97	
Lasso	0.89	0.464	0.774	0.664	0.653	0.945	
Ridge	0.908	0.425	0.814	0.602	0.77	0.769	
SGD	0.787	0.645	0.684	0.786	0.768	0.773	
KernelRidge	0.24	1.219	0.004	1.395	0.328	1.315	
Huber	0.915	0.407	0.831	0.575	0.819	0.681	**
QSAR models based on RFE-SVR							
tree-XGB	0.945	0.329	0.893	0.457	0.772	0.766	
linear-XGB	0.844	0.553	0.756	0.691	0.759	0.787	
RFR	0.955	0.295	0.939	0.346	0.758	0.788	
GBDT	0.982	0.187	0.891	0.462	0.811	0.696	
Bagging	0.992	0.126	0.947	0.321	0.778	0.756	
MLP	0.979	0.202	0.882	0.48	0.846	0.628	***
KNN	0.924	0.386	0.897	0.449	0.839	0.643	*
rbf-SVR	0.996	0.095	0.835	0.568	0.666	0.927	
linear-SVR	0.922	0.39	0.829	0.579	0.809	0.701	
Lasso	0.844	0.552	0.756	0.691	0.759	0.787	
Ridge	0.859	0.525	0.806	0.616	0.83	0.662	
SGD	0.916	0.405	0.834	0.569	0.886	0.541	
KernelRidge	0.329	1.145	0.156	1.285	0.448	1.191	
Huber	0.926	0.381	0.84	0.559	0.84	0.642	**
QSAR models based on RFE-RFR							
tree-XGB	0.978	0.205	0.931	0.367	0.828	0.665	***

linear-XGB	0.852	0.539	0.786	0.647	0.704	0.872	
RFR	0.976	0.219	0.937	0.349	0.808	0.703	
GBDT	0.989	0.145	0.935	0.358	0.815	0.689	*
Bagging	0.992	0.122	0.939	0.345	0.799	0.719	
MLP	0.98	0.197	0.89	0.465	0.817	0.686	**
KNN	0.966	0.259	0.915	0.409	0.791	0.734	
rbf-SVR	0.996	0.089	0.924	0.385	0.801	0.716	
linear-SVR	0.852	0.539	0.761	0.684	0.699	0.88	
Lasso	0.861	0.521	0.778	0.66	0.706	0.869	
Ridge	0.867	0.51	0.783	0.651	0.702	0.875	
SGD	0.863	0.518	0.787	0.647	0.703	0.874	
KernelRidge	0.382	1.099	0.105	1.324	0.144	1.484	
Huber	0.86	0.523	0.788	0.643	0.706	0.869	
QSAR models without feature selection							
tree-XGB	0.987	0.161	0.860	0.523	0.705	0.87	
linear-XGB	0.927	0.378	0.786	0.647	0.746	0.807	*
RFR	0.946	0.324	0.892	0.459	0.749	0.803	***
GBDT	0.992	0.126	0.898	0.447	0.744	0.811	**
Bagging	0.929	0.374	0.419	1.066	0.404	1.238	
MLP	0.773	0.666	0.621	0.861	0.619	0.99	
KNN	0.996	0.089	0.752	0.697	0.628	0.978	
rbf-SVR	0.926	0.381	0.709	0.754	0.734	0.827	
linear-SVR	0.893	0.457	0.765	0.678	0.731	0.831	
Lasso	0.936	0.355	0.758	0.688	0.744	0.811	
Ridge	0.463	1.025	0.160	1.281	0.263	1.377	
SGD	0.411	1.073	-0.081	1.454	0.073	1.544	
KernelRidge	0.936	0.355	0.757	0.69	0.743	0.812	
Huber	0.981	0.195	0.858	0.526	0.743	0.812	

Model	Training dataset				Test dataset		Note
	R^2_{Train}	$RMSE_{Train}$	R^2_{CV}	$RMSE_{CV}$	R^2_{Test}	$RMSE_{Test}$	
QSAR models based on FI-XGB							
tree-XGB	0.955	0.295	0.911	0.416	0.814	0.692	***
linear-XGB	0.566	0.921	0.478	1.01	0.558	1.067	
RFR	0.956	0.295	0.924	0.386	0.807	0.698	**
GBDT	0.976	0.219	0.904	0.434	0.78	0.752	*
Bagging	0.974	0.226	0.904	0.434	0.769	0.77	
MLP	0.961	0.276	0.847	0.548	0.77	0.769	
KNN	0.84	0.559	0.598	0.887	0.555	1.069	
rbf-SVR	0.965	0.263	0.831	0.574	0.726	0.84	
linear-SVR	0.387	1.095	0.355	1.123	0.345	1.298	
Lasso	0.575	0.912	0.473	1.015	0.59	1.027	
Ridge	0.566	0.921	0.478	1.01	0.557	1.068	
SGD	0.516	0.973	0.424	1.062	0.49	1.146	
KernelRidge	0.074	1.346	-0.073	1.448	0.206	1.429	
Huber	0.567	0.92	0.478	1.01	0.559	1.064	
QSAR models based on FI-RFR							
tree-XGB	0.954	0.3	0.872	0.5	0.847	0.627	***
linear-XGB	0.789	0.643	0.722	0.738	0.681	0.906	
RFR	0.928	0.375	0.842	0.556	0.854	0.613	**
GBDT	0.978	0.207	0.866	0.512	0.781	0.75	
Bagging	0.962	0.274	0.833	0.571	0.822	0.677	
MLP	0.976	0.219	0.82	0.592	0.773	0.764	
KNN	0.933	0.362	0.832	0.573	0.814	0.691	
rbf-SVR	0.954	0.3	0.832	0.574	0.844	0.632	*

linear-SVR	0.78	0.655	0.709	0.755	0.623	0.984	
Lasso	0.796	0.632	0.714	0.748	0.685	0.901	
Ridge	0.779	0.657	0.721	0.739	0.679	0.909	
SGD	0.789	0.642	0.724	0.735	0.674	0.916	
KernelRidge	0.279	1.187	-0.118	1.479	0.295	1.346	
Huber	0.792	0.637	0.719	0.742	0.682	0.904	
QSAR models based on FC-LR							
tree-XGB	0.977	0.214	0.883	0.477	0.707	0.868	
linear-XGB	0.827	0.582	0.775	0.663	0.783	0.748	**
RFR	0.983	0.182	0.923	0.389	0.652	0.946	
GBDT	0.991	0.134	0.928	0.377	0.626	0.981	
Bagging	0.989	0.145	0.92	0.396	0.681	0.905	
MLP	0.975	0.221	0.771	0.669	0.763	0.781	
KNN	0.94	0.342	0.835	0.568	0.813	0.693	***
rbf-SVR	0.988	0.155	0.741	0.711	0.716	0.855	
linear-SVR	0.815	0.602	0.756	0.691	0.782	0.75	
Lasso	0.817	0.598	0.759	0.686	0.739	0.819	
Ridge	0.821	0.592	0.779	0.658	0.777	0.757	
SGD	0.826	0.584	0.771	0.669	0.785	0.743	
KernelRidge	0.319	1.155	0.034	1.375	0.37	1.272	
Huber	0.829	0.579	0.759	0.687	0.786	0.741	*
QSAR models based on RFE-LR							
tree-XGB	0.951	0.31	0.801	0.624	0.773	0.764	
linear-XGB	0.849	0.542	0.752	0.697	0.78	0.752	
RFR	0.939	0.345	0.793	0.636	0.737	0.823	

GBDT	0.986	0.164	0.821	0.592	0.8	0.718	*
Bagging	0.976	0.217	0.815	0.601	0.766	0.775	
MLP	0.979	0.202	0.868	0.509	0.824	0.672	***
KNN	0.859	0.526	0.749	0.701	0.627	0.98	
rbf-SVR	0.993	0.118	0.774	0.666	0.569	1.053	
linear-SVR	0.887	0.47	0.781	0.654	0.634	0.97	
Lasso	0.89	0.464	0.774	0.664	0.653	0.945	
Ridge	0.908	0.425	0.814	0.602	0.77	0.769	
SGD	0.787	0.645	0.684	0.786	0.768	0.773	
KernelRidge	0.24	1.219	0.004	1.395	0.328	1.315	
Huber	0.915	0.407	0.831	0.575	0.819	0.681	**
QSAR models based on RFE-SVR							
tree-XGB	0.945	0.329	0.893	0.457	0.772	0.766	
linear-XGB	0.844	0.553	0.756	0.691	0.759	0.787	
RFR	0.955	0.295	0.939	0.346	0.758	0.788	
GBDT	0.982	0.187	0.891	0.462	0.811	0.696	
Bagging	0.992	0.126	0.947	0.321	0.778	0.756	
MLP	0.979	0.202	0.882	0.48	0.846	0.628	***
KNN	0.924	0.386	0.897	0.449	0.839	0.643	*
rbf-SVR	0.996	0.095	0.835	0.568	0.666	0.927	
linear-SVR	0.922	0.39	0.829	0.579	0.809	0.701	
Lasso	0.844	0.552	0.756	0.691	0.759	0.787	
Ridge	0.859	0.525	0.806	0.616	0.83	0.662	
SGD	0.916	0.405	0.834	0.569	0.886	0.541	
KernelRidge	0.329	1.145	0.156	1.285	0.448	1.191	

Huber	0.926	0.381	0.84	0.559	0.84	0.642	**
QSAR models based on RFE-RFR							
tree-XGB	0.978	0.205	0.931	0.367	0.828	0.665	***
linear-XGB	0.852	0.539	0.786	0.647	0.704	0.872	
RFR	0.976	0.219	0.937	0.349	0.808	0.703	
GBDT	0.989	0.145	0.935	0.358	0.815	0.689	*
Bagging	0.992	0.122	0.939	0.345	0.799	0.719	
MLP	0.98	0.197	0.89	0.465	0.817	0.686	**
KNN	0.966	0.259	0.915	0.409	0.791	0.734	
rbf-SVR	0.996	0.089	0.924	0.385	0.801	0.716	
linear-SVR	0.852	0.539	0.761	0.684	0.699	0.88	
Lasso	0.861	0.521	0.778	0.66	0.706	0.869	
Ridge	0.867	0.51	0.783	0.651	0.702	0.875	
SGD	0.863	0.518	0.787	0.647	0.703	0.874	
KernelRidge	0.382	1.099	0.105	1.324	0.144	1.484	
Huber	0.86	0.523	0.788	0.643	0.706	0.869	
QSAR models without feature selection							
tree-XGB	0.987	0.161	0.860	0.523	0.705	0.87	
linear-XGB	0.927	0.378	0.786	0.647	0.749	0.803	***
RFR	0.946	0.324	0.892	0.459	0.746	0.807	**
GBDT	0.992	0.126	0.898	0.447	0.744	0.811	*
Bagging	0.929	0.374	0.419	1.066	0.404	1.238	
MLP	0.773	0.666	0.621	0.861	0.619	0.99	
KNN	0.996	0.089	0.752	0.697	0.628	0.978	
rbf-SVR	0.926	0.381	0.709	0.754	0.734	0.827	
linear-SVR	0.893	0.457	0.765	0.678	0.731	0.831	

Lasso	0.936	0.355	0.758	0.688	0.744	0.811
Ridge	0.463	1.025	0.160	1.281	0.263	1.377
SGD	0.411	1.073	-0.081	1.454	0.073	1.544
KernelRidge	0.936	0.355	0.757	0.69	0.743	0.812
Huber	0.981	0.195	0.858	0.526	0.743	0.812

Note: Detailed description of these models are available at https://scikit-learn.org/stable/ and https://xgboost.readthedocs.io/en/stable/.

The models with more stars in the last column indicate better performance from the same feature selection method.

Table 2-4 Antioxidant activity of synthesized tripeptides.

Synthetic tripeptide	Observed activity (μmol TE / μmol peptide)	Predicted activity (μmol TE / μmol peptide)
QAY	4.270±0.124	6.167
PHC	5.013±0.184	6.023
YSQ	3.736±0.024	5.696
YPQ	3.028±0.173	5.696
VYV	3.601±0.039	4.837
GPE	0.598±0.099	2.741

Chapter 3 - Computer-aided approaches for screening antioxidative dipeptides and application to sorghum proteins

Abstract

There has been a growing interest in extracting antioxidant peptides from food proteins. This study aimed to develop efficient computer-aided approaches to accelerate the screening efficiency of antioxidative dipeptides. An newly developed quantitative structure-activity relationship model and an improved hydrolysis simulation tool, R-PeptideCutter, were applied to screen high-activity dipeptides in sorghum kafirin. The R^2_{Test} and MSE_{Test} were 0.6082 and 0.6764, and 0.5302 and 0.5467, respectively, for 2,2'-azinobis-(3-ethylbenzothiazoline-6-sulfonate) (ABTS) radical scavenging capacity and oxygen radical absorbance capacity (ORAC) models. N-terminus residues dominated the antioxidant activity, especially in the ABTS assay, and Y and W at the N-terminus strongly corresponded to higher activity in both assays. A dipeptide YR was predicted as the strongest antioxidant in kafirin (3.352/2.099 μmol Trolox/μmol peptide for ABTS/ORAC activity). Eight kafirin-derived dipeptides were synthesized for model validation. The ORAC corresponding model achieved greater prediction performance, while the ABTS radical scavenging capacity model showed an underestimation in prediction. The new tools and knowledge can be applied to other proteins and benefit the research and development on antioxidant peptides.

Keywords: Antioxidant peptides, artificial intelligence, hydrolysis simulation, sorghum protein, *in silico*

---

* Du, Z., & Li, Y. (2022). Computer-aided approaches for screening Antioxidative dipeptides and application to sorghum proteins. ACS Food Science & Technology, 2(11), 1781-1788. https://doi.org/10.1021/acsfoodscitech.2c00286

3.1 Introduction

Developing natural antioxidants has gained expanding interest because of the increasing health concerns for synthetic antioxidants and strict regulation on their usage (Du & Li, 2022; Lorenzo et al., 2018; Phongthai et al., 2018). Antioxidant peptides derived from food proteins have captured worldwide attention due to their advantages such as naturally sourced, better sustainability, and no or low toxic effects (Du, Wang, et al., 2022; Du & Li, 2022; Phongthai et al., 2018). Generally, antioxidant peptides are released from parent proteins by enzymatic hydrolysis, fermentation, chemical hydrolysis, germination, and/or ripening, and then screened by laborious chemistry methods (e.g., fractionation, isolation, purification, identification, and characterization) (Du & Li, 2022; Trinidad-Calderón et al., 2021). However, the conventional wet-chemistry methods are time-consuming and highly rely on many advanced instruments and equipment (Du & Li, 2022; Lorenzo et al., 2018; Trinidad-Calderón et al., 2021). Chemical synthesis of peptides is an alternative approach for producing and screening potentially highly active peptides (Saito et al., 2003; Trinidad-Calderón et al., 2021; Zheng et al., 2016). Nonetheless, it is practically impossible to synthesize all the peptides for antioxidative peptides screening, considering the cost of synthesis and a large number of theoretically possible peptides, i.e., 400 dipeptides, 8000 tripeptides, 160,000 tetrapeptides, etc. (N. Chen et al., 2018; Deng et al., 2019; Du, Wang, et al., 2022).

Tremendous bioactive peptides have been identified, making it possible to use these accumulated activity data for modeling quantitative structure-activity relationship. The models can also provide more efficient and cost-effective guidance for the exploration of new bioactive peptides (Daroit & Brandelli, 2021; Du & Li, 2022; Li et al., 2019; Lorenzo et al., 2018). Such *in-silico* approaches have been successfully applied to predict angiotensin I-converting enzyme

inhibitory (ACE-I) activity, dipeptidyl peptidase IV (DPP-IV) inhibitory activity, and antimicrobial activity (FitzGerald et al., 2020; G. Liang & Li, 2007; Wu et al., 2014). Besides, some peptide cutting simulation tools (e.g., PeptideCutter) were developed for protein *in-silico* hydrolysis, which can be combined with more than 156 million identified protein sequences from the Uniprot database to obtain potentially bioactive peptide sequences generated by specific enzymes or multiple enzymes combination (Iwaniak et al., 2020; Kalyan et al., 2021; The UniProt Consortium, 2021; Walker, 2005). There were few studies and limited models on antioxidant peptides, especially for dipeptides, which demonstrate ideal absorption ability and bioavailability in the intestine compared to larger peptides (N. Chen et al., 2018; Deng et al., 2019; M. Tian et al., 2015; Zheng et al., 2016). In addition, the order of *in-silico* hydrolysis with multiple enzymes was not specified in the previous peptide cutter tools, which is critical in practical experiments and should be developed (Iwaniak et al., 2020; Kalyan et al., 2021; The UniProt Consortium, 2021; Walker, 2005).

Therefore, the objectives of this study were to: 1) develop highly predictive models that could guide the discovery of peptides with high antioxidant activity and shed light on critical amino acid features that determine the antioxidant activity; 2) build a improved protein cutting simulation tool with consideration of hydrolysis order and enlarge the inclusion of published enzymes or chemicals for protein hydrolysis; 3) employ the prediction models and protein cutter tool to screen antioxidant dipeptides in an underutilized protein, sorghum kafirin; and 4) design and synthesize antioxidant dipeptides encrypted in kafirin sequence and evaluate their antioxidant activity for model performance validation. The overall technical workflow of the study is summarized in Figure 1.

3.2 Materials and Methods

3.2.1 Dataset collection

Data mining (Beautiful Soup, 4.5.3) was used to collect 566 numerical indices of amino acids from AAIndex, and detailed definition and description of each index are available online (https://www.genome.jp/aaindex/) (Kawashima et al., 2008). The indices with missing values for amino acids were manually deleted, resulting in a total of 553 remaining indices (Supplementary Document S1). Antioxidant activities of dipeptides based on *in vitro* antioxidant assays including both 2,2'-azinobis-(3-ethylbenzothiazoline-6-sulfonate) (ABTS) radical scavenging capacity assay and oxygen radical absorbance capacity (ORAC) assay, were manually collected from published literatures for antioxidant activity prediction model development. Sixty-seven antioxidant dipeptides characterized by ABTS radical scavenging capacity assay, and seventy-three dipeptides characterized by ORAC assay were collected as two separate datasets (Supplementary Document S2), and the antioxidant activity values are expressed as Trolox-equivalent antioxidant capacity (μmol TE / μmol peptide) (Amigo et al., 2020; Anna et al., 2016; Hernández-Ledesma et al., 2007; Huang et al., 2010; Je et al., 2015; Suetsuna et al., 2000; Zheng et al., 2016).

3.2.2 Data processing

3.2.2.1 Pre-processing of numerical indices of amino acids

Among the 553 numerical indices, some of them are highly correlated. It is necessary to remove those features since it only provided limited information but increases the model complexity. Collinearity of the 553 numerical indices was pre-screened by pairwise correlation methods (Supplementary Documents S3 & S4). If an absolute value of Pearson's correlation coefficient between two indices was above 0.95, one of them was removed randomly due to the

strong correlation (Sabilla et al., 2019). The remaining numerical indices were standardized for further feature selection.

3.2.2.2 Dipeptide encoding and feature selection

The pre-screened numerical indices of amino acids were used to encode dipeptides. Briefly, if the 'n' numerical indices were selected after the pre-processing, each amino acid was encoded as a '1 × n' vector. Since each dipeptide has two amino acid residues, it was encoded as a '2 × n' matrix, and then transformed into a '1 × 2n' matrix, where the 1 to the n elements in the vector belonged to the N-terminus residue and n+1 to 2n elements were referred to the C-terminus residue. All the dipeptides were encoded again by the new features as X-matrix (variables) and modeled with its corresponding antioxidant activity values (Y-vector). After the encoding, each dipeptide was represented by 2n variables.

Feature selection is used to further screen the important feature particularly for antioxidant activity prediction, and therefore significantly simplify the model complexity. In addition, it can also point out the important features to enhance the understanding in features contributing to antioxidant activity.[4] These samples with 2n variables and corresponding Y-vector in each activity dataset were modeled by extreme gradient boosting (XGboost) regression. Feature importances were calculated and used to identify the key variables for antioxidant activity prediction.[28]

3.2.3 Model development

3.2.3.1 Dataset division

Datasets were shuffled and randomly split into a training dataset and a test dataset at a ratio of three to one. For the ABTS radical scavenging capacity dataset, 51 samples were used in the training dataset for model building, and the remaining 16 samples were used in the test dataset to

evaluate the performance of the model. For the ORAC dataset, the number of samples in the training dataset and test dataset were 55 and 18, respectively.

3.2.3.2 Model building, optimization, and evaluation

XGBoost was employed to build regression model based on the training dataset and the parameters was tuned by leave-one-out cross-validation (LOOCV). The hyperparameters with the best performance in LOOCV were used as the final model for performance evaluation with the test dataset. The detailed codes for model development are available in Supplementary Document S5. Model development was performed using Python 3.8.8 (MacOS Monterey 12.0.1, CPU intel Core-i5 2.3 GHz), and the related functions are available through scikit-learn (https://scikit-learn.org/) and XGBoost (https://xgboost.readthedocs.io/en/stable/) (T. Chen & Guestrin, 2016; Pedregosa et al., 2011).

Determination of coefficient (R^2) and mean square error (MSE) were used to evaluate the model performance. The R^2 and MSE from the training dataset, LOOCV, and test dataset were labelled as R^2_{Train} and MSE_{Train}, R^2_{CV} and MSE_{CV}, and R^2_{Test} and MSE_{Test}, respectively.

3.2.4 Prediction of dipeptides with high antioxidant activity

A dataset containing 400 possible dipeptides with ABTS radical scavenging capacity and ORAC was built, and the published 67 ABTS-associated dipeptides and 73 ORAC-associated dipeptides in model building and validation were also included. After obtaining the two prediction models, the possible 400 dipeptides were encoded by the selected features, and then their antioxidant activities were predicted (Supplementary Document S6).

3.2.5 Protein cutting simulation tool development

The cleavage sites of fifty-one enzymes and chemicals were collected from published literatures and publicly available web servers such as PeptideCutter (Kalyan et al., 2021;

Minkiewicz et al., 2019; Walker, 2005). Functions for a total of 51 enzymes and chemicals in the simulation tool were designed to obtain specific cleavage sites in a given protein sequence, including Arg-C proteinase, Asp-N endopeptidase, Asp-N endopeptidase + N-terminus Glu, BNPS-Skatole, caspase1, caspase2, caspase3, caspase4, caspase5, caspase6, caspase7, caspase8, caspase9, caspase10, chymotrypsin-high specificity (C-term to [F Y W], not before P), chymotrypsin-low specificity (C-term to [F Y W M L], not before P), clostripain (clostridiopeptidase B), CNBr, enterokinase, factor Xa, formic acid, glutamyl endopeptidase, granzymeB, hydroxylamine, iodosobenzoic acid, LysC, neutrophil elastase, NTCB (2-nitro-5-thiocyanobenzoic acid), Pepsin (pH=1.3), Pepsin (pH>2), proline-endopeptidase, proteinase K, staphylococcal peptidase I, thermolysin, thrombin, trypsin, elastase 1, elastase 2, chymotrypsinogen B1, chymotrypsinogen C, pancreatic enteropeptidase E enteropeptidase, prostasin, gastricsin, fruit bromelain, stem bromelain, ananain, papaya proteinase 4, chymopapain, chymosin, and caricain (Kalyan et al., 2021; Walker, 2005). All the detailed cleavage sites of enzymes and chemicals are available in Supplementary Document S7-Table S3.

Based on these designed functions, we developed a user-friendly and well-annotated protein hydrolysis simulation tool, named Refining-PeptideCutter (R-PeptideCutter), which takes the adding sequence of enzymes or chemicals for hydrolysis into consideration. Users will only need to download the fasta format files containing the parent protein sequence from UniProt, set enzymes or chemicals and desired peptide length in R-PeptideCutter, and run the scripts. It will automatically generate all the possible peptides encrypted in the protein sequence as well as the number of peptides that can be released. The detailed codes were written in python and are available in Supplementary Document S7. In addition, the R-PeptideCutter tool was further tailored to link with the predicted antioxidant results from the newly developed models, so that the

antioxidant activity values of the *in-silico* released peptides can be assigned. These codes can be modified and used for studies of other bioactive peptides and are available in Supplementary Document S7.

3.2.6 Application of simulation tool and activity prediction models in kafirin proteins

Eight available kafirin sequences were selected from UniProt (https://www.uniprot.org/) and used for hydrolysis simulation, including three alpha-kafirin sequences, one beta-kafirin sequence, two gamma-kafirin sequences, and two delta-kafirin sequences (Supplementary Document S7-Table S1). All the downloaded kafirin sequences were pretreated by removing the signal peptide from the peptide sequences before further hydrolysis simulation. All these sequences were simulated to be hydrolyzed by one or a combination of two enzymes or chemicals in order for all the 51 enzymes or chemicals. The predicted antioxidant activity values (ABTS radical scavenging capacity & ORAC) of the generated peptides were assigned, respectively (Supplementary Document S7).

3.2.7 Antioxidant peptide synthesis

Eight dipeptides present in kafirin sequences were purchased from Sigma-Aldrich (St. Louis, MO, USA). The purity of these dipeptides was all above 95%. Considering the diversity of peptides, kafirin protein hydrolysate simulation results, and availability form Sigma-Aldrich , some peptides with low antioxidant activity based on model prediction were also selected (N. Chen et al., 2018). These peptides (CG, AY, YA, YF, GW, GG, GT, and GV) were used to validate the performance of the constructed models.

3.2.8 Antioxidant activity assays

1,1-Diphenyl-2-picrylhydrazyl (DPPH), ABTS, fluorescein disodium (FL), and 2,2′-azobis (2-methylpropionamide)-dihydrochloride (AAPH) were purchased from Sigma-Aldrich (St.

Louis, MO, USA). All the chemicals and reagents used were of analytical grade. Besides the ABTS and ORAC assays, we also conducted the DPPH, which can be used for database building in future antioxidant activity-related models. All the tests were conducted in triplicate.

3.2.8.1 ABTS radical scavenging capacity assay

The ABTS radical scavenging capacity assay was conducted according to the study of Zheng et al.(Zheng et al., 2016) Briefly, 150 μL $ABTS^{\bullet+}$ solution was mixed with 50 μL dipeptide solution (10 μM) in a 96-well microplate. After 30 minutes incubation at 30 °C, the absorbance was measured at 734 nm using a Biotek Synergy H1 Hybrid Microplate Reader (Winooski, VT, USA). An equivalent volume of 50 mM phosphate buffer (PBS) at pH 7.4 was used as the control, and the initial absorbance at 734 nm was controlled at 0.70 ± 0.02. Dipeptide solution was prepared in 75 mM PBS buffer (pH 7.4). Trolox (TE) was used as a standard antioxidant, and results were expressed as μmol TE / μmol peptide.

3.2.8.2 ORAC assay

The ORAC assay was performed according to the study of Zheng et al.(Zheng et al., 2016). Twenty microliter dipeptide solution (20 μM) and 60 μL FL solution (5 nM) were transferred in a well of a 96-well microplate, which was allowed for 15 minutes incubation at 37 °C. Then 120 μL AAPH solution (80 mM) was mixed with the incubated mixture in the plate for 30 s. A BioTek Synergy H1 Hybrid Microplate Reader was used to record the fluorescence for 100 min at 485 nm for excitation and 520 nm for emission, respectively. All the dipeptide solutions, FL and AAPH solutions were prepared in 75 mM PBS buffer (pH 7.4). Trolox was used as a standard antioxidant, and the ORAC values were also expressed as μmol TE / μmol peptide.

3.2.8.3 DPPH radical scavenging capacity assay

DPPH radical scavenging capacity assay was performed according to the study of Chen et al., with some modifications (N. Chen et al., 2018). Briefly, 100 μL dipeptide solution (20 μM) was mixed with 100 μL DPPH solution (0.2mM in 95% ethanol) in a 96-well microplate and then incubated for 30 min at room temperature in dark. The absorbance was measured at 517 nm using a Biotek Synergy H1 Hybrid Microplate Reader (Winooski, VT, USA). Dipeptide solution was prepared in 75 mM PBS buffer (pH 7.4). Trolox was used as the standard antioxidant and the results were expressed as μmol TE / μmol peptide.

3.3 Results and discussion

3.3.1 Feature importance analysis

Eleven variables were selected by XGBoost regression with a feature importance threshold of 0.01 (Table 1) and then used to encode the 67 dipeptides with known ABTS radical scavenging capacity values as the X-matrix (i.e., 67 × 11). Among them, 7 variables were used to encode N-terminus residues, accounting for 0.8109 in feature importance for ABTS radical scavenging capacity prediction, while the sum of the feature importance of the remaining 4 variables representing C-terminus residues was only 0.1087. LIFS790103 and CHAM820102 standing for ‘conformational preference for antiparallel beta-strands’ and ‘free energy of solution in water’, respectively, were the two most important variables for ABTS radical scavenging capacity prediction, and both were associated with N-terminus residues.

Similarly, 14 important variables for the 73 dipeptides with known ORAC values were selected by the same feature selection method employed for ABTS radical scavenging capacity values (Table 2). The N-terminus residue was represented by 7 variables, accounting for 0.6437 in feature importance for ORAC prediction, while only 0.2449 in feature importance was stood for

C-terminus with the same number of representative variables as N-terminus. CHOP780215 (N-terminus), PALJ810108 (N-terminus), and DESM900101(C-terminus) standing for 'frequency of the 4th residue in turn', 'normalized frequency of alpha-helix in alpha + beta class', and 'membrane preference for cytochrome b: MPH89' were the three most important variables for ORAC activity prediction.

Feature importance of these variables in ABTS radical scavenging capacity model development and ORAC model development showed that N-terminus residues played a more important role in antioxidant activity. In the studyof Chen et al., the variable importance in projection (VIP) values revealed that the C-terminus contributed a lot to the ABTS radical scavenging capacity of tripeptides, and such observation was also confirmed in the study of Du et al.[4,8] It as quite interesting results that antioxidant dipeptides showed a different pattern where the importance of N-terminus was higher than that in C-terminus. It was also confirmed in the VIP values of the partial least square regression (PLSR) models in Zheng et al., although the authors instead stressed the importance of tyrosine and tryptophan in the N-terminus (Zheng et al., 2016). Compared to ABTS radical scavenging capacity prediction, C-terminus residues contributed more to ORAC prediction. Among the selected variables, CHAM820102 and YUTK870102 for N-terminus residues encoding in ABTS radical scavenging capacity prediction were also selected in the study of Chen et al. for tripeptide's ABTS radical scavenging capacity prediction, which showed their generality in antioxidant activity prediction (N. Chen et al., 2018). Previous studies on peptides were mainly focused on the amino acid composition of antioxidant peptides for the explanation of the antioxidant activity (e.g., residue content of tyrosine) (N. Chen et al., 2018; Deng et al., 2019; M. Tian et al., 2015; Udenigwe & Aluko, 2011; Zheng et al., 2016; Zou et al., 2016).

Some variables, such as LIFS790103, were selected in both models. The feature importance of LIFS790103 was 0.4321 and 0.0819 for the ABTS radical scavenging capacity corresponding model and ORAC corresponding model, respectively. That might be due to the difference in mechanisms between single electron transfer (SET) mechanism and hydrogen atom transfer (HAT) mechanism, since ABTS assay and ORAC assay belong to SET mechanism and HAT mechanism, respectively (N. Liang & Kitts, 2014). However, it should be mentioned that some of the variables, such as LIFS790103 standing for conformational preference for antiparallel beta-strands, were not directly related to the antioxidant activity of dipeptides, since there was no secondary structure in dipeptides. This issue complicated the interpretation of models, which was also indicated by other researchers for machine learning applications (N. Chen et al., 2018; Deng et al., 2019). Compared to previous studies based on components from principal components analysis (PCA) as the variables for model development, our selected variables further clarified specific variables associated with the antioxidant activity instead of groups of variables and also enlarged the feature source for peptide encoding (Mahmoodi-Reihani et al., 2020; F. Tian et al., 2007; Zhou et al., 2021). In addition, the selected variables from our study were more targeted and less redundant compared with previous models adopting amino acid descriptors (e.g., z scale in the study of Zheng et al.) composed of a combination of various weighted variables since there is no any amino acid descriptors specifically designed for antioxidant activity.[13,35,36] These findings provide alternative ways for further theoretical and mechanism studies on antioxidant peptides (Deng et al., 2019; M. Tian et al., 2015).

3.3.2 Model performance evaluation

The relationships between the observed activity and the predicted activity of ABTS and ORAC models are shown in Figure 2 (a) & (b), respectively. The R^2_{CV} and R^2_{Test} for the ABTS

and ORAC prediction models were 0.7102/0.6082 and 0.5698/0.5302, respectively. For the ABTS model, the predicted activity was inclined to be lower than the observed activity. In particular, the observed activity of around 2 to 4 μmol TE / μmol peptide was predicted to be 1 to 2 μmol TE / μmol peptides. In contrast, the predicted activity was much closer to the observed activity in the ORAC model except for several samples which showed a larger difference between the predicted activity and the observed one (e.g., the highest observed activity sample).

Our model performances were substantial breakthroughs compared to the previous study conducted by Zheng et al., where the best R^2_{CV} values, instead of R^2_{test} for ABTS radical scavenging capacity and ORAC prediction were only 0.38 and 0.26 relying on a linear regression modelling method (PLSR) (Zheng et al., 2016). Technically, R^2_{CV} could not describe the model's performance in unknown dataset because the data used in R^2_{CV} obtaining were used in model building (Deng et al., 2019). In this study, larger datasets were collected from the latest publication, and feature selection ad non-linear regression modeling approaches were employed to connect the dataset and antioxidant activity, which finally contributed a better performance in test datasets . (Du, Tian, et al., 2022, Du, Wang, et al., 2022; Du, Zeng, et al., 2020; Kalyan et al., 2021; N. Chen et al., 2018).

3.3.3 Prediction of dipeptides with potentially high antioxidant activity from the models

For ABTS radical scavenging capacity prediction, dipeptides with Y residue at N-terminus showed higher activity (3.37 μmol TE / μmol peptide), followed by those with W residue at N-terminus (Supplementary Document S6). For ORAC, dipeptides with W residue at N-terminus showed higher activity (4.81 μmol TE / μmol peptide), and there was no obvious trend of preferred residue in the N or C-terminus for the remaining high activity dipeptides. The highlighted residues (W and Y) agreed with other studies where they were reported to be strongly corresponding to

high antioxidant activity of peptides (N. Chen et al., 2018; Deng et al., 2019; M. Tian et al., 2015; Udenigwe & Aluko, 2011; Zheng et al., 2016; Zou et al., 2016).

3.3.4 Application of simulation tool and antioxidant activity prediction models in kafirin proteins

The kafirin protein sequences were cleaved *in-silico* by R-PeptideCutter tool, and then the antioxidant activity of the generated dipeptides was predicted by the ABTS radical scavenging capacity and ORAC models, respectively. Among these enzymes or chemicals, the three enzymes, namely chymotrypsin C, Proteinase K, and thermolysin, generated more diverse dipeptides (more than ten different dipeptides among all these kafirin proteins) (Supplementary Document S7-Table S2). However, all these generated dipeptides were predicted to have low antioxidant activity in both ABTS and ORAC assays. The most diverse results were obtained from A9XEC1 (alpha kafirin B3), where 18 different dipeptides were generated by Thermolysin, and 17 and 14 different dipeptides were from the hydrolysis with chymotrypsin C and Proteinase K, respectively (Supplementary Document S7-Table S2). Besides, both pepsin (pH = 1.3) and pepsin (pH > 2) released six QQ (dipeptide) from a single alpha kafirin sequence A9XEC1. The theoretically generated dipeptide varied among the enzymes due to the variation in enzyme cleavage sites and protein sequences (Kalyan et al., 2021). For example, Thermolysin can cleave protein sequences when P1 position is not D or E and the P1' position is A, F, I, L, M, or V. Both proteinase K (A, E, F, I, L, T, V, W or Y at P1) and the chymotrypsin C (F, M, Y, W, L, N or Q at P1) have broad cleavage sites and therefore can generate various dipeptides. Thermolysin cannot generate dipeptides with a tyrosine, tryptophan or cysteine at N-terminus which actually limits its application in antioxidant dipeptide production. Both chymotrypsin C and proteinase K could generate dipeptide with tyrosine or tryptophan at C-terminus, but these preferable residues at N-

terminus depend on sequence variation. Other enzymes, e.g., pepsin (pH = 1.3 or pH> 2), trypsin, chymotrypsin B1, and chymotrypsin (low or high specificity) also had broad cleavage sites, but the diversity was not comparable to the three enzymes which resulted from the sequences. Also, these enzymes might suffer from the broad cleavage sites, since this might lead to more residues released as free amino acids instead of peptides.

When two enzymes or chemicals were combined to generate dipeptides, there are 2601 combinations from the 51 available enzymes or chemicals simulation. Overall, there are many additional dipeptides generated from two enzymes/chemicals compared to the hydrolysis with single enzyme or chemical (Supplementary Document S7-Table S2). The best way to increase the diversity of the dipeptides was to combine one enzyme with broad cleavage sites and one with highly specific sites. The dipeptide diversity (12 dipeptides) generated by the combination of pepsin (pH = 1.3) and neutrophil elastase in Q9XE78 (alpha kafirin) was at least twice of the number of dipeptides from pepsin (pH =1.3) (2 dipeptides) or neutrophil elastase (5 dipeptides) alone (Supplementary Document S7-Table S2). This combination also released a highly potent antioxidant dipeptide, YR where the ABTS radical scavenging activity and ORAC were 3.352 and 2.099 μmol TE / μmol peptide, respectively. Furthermore, this dipeptide cannot be produced by single enzymes process.

It should be noted that there was no enzyme that can recognize cysteine residue occurring at the P1 or P1' position except chemicals, like NTCB which can break the peptide bonds with cystine at the P1' position. Therefore, obtaining a cysteine-containing dipeptide is quite hard to be produced from available proteins and only happens occasionally. For example, papaya proteinase can break the peptide bonds with glycine residue at P1 position. Therefore, when there is a cysteine residue at P1' position, it is possible to generate cysteine-contained dipeptides. Among the selected

kafirin sequences, G3FMW5 and Q6Q299 were observed to generate dipeptide (CG) under papaya proteinase.

The R-PeptideCutter has the generality to be used in any protein sequences for peptide cutting simulation and can generate all possible peptides with any length (Supplementary Document S7). Overall, the diversity of the total possible dipeptides generated by two enzymes or chemicals was significantly higher, and some of the dipeptides (e.g., YR) were exclusively generated when combining two enzymes, which could not be achieved when using an individual enzyme or chemical (Kalyan et al., 2021). The predicted antioxidant activity of the generated peptides hydrolyzed by different enzymes or chemicals could be used to guide wet chemistry for other studies on sorghum protein-derived antioxidant dipeptides.

3.3.5 Antioxidant activity model validation using synthesized dipeptides

The ABTS radical scavenging capacity, ORAC, and DPPH radical scavenging capacity of eight synthetic dipeptides are shown in Figure 3. For ABTS radical scavenging capacity, YA exhibited the highest capacity (6.50 μmol TE / μmol peptide), while AY (2.19 μmol TE / μmol peptide) with the same amino acid residue composition only showed one-third capacity of YA. Besides, YF as the second strongest dipeptide also exhibited competitive activity (5.62 μmol TE / μmol peptide). For the ORAC, GW showed the highest activity (3.49 μmol TE / μmol peptide), while only AY, YA and YF exhibited some ORAC activity among the remaining synthesized dipeptides. Regarding DPPH radical scavenging capacity, the trend was comparable to that in ORAC. The only exception was that the antioxidant capacity of YA (1.46 μmol TE / μmol peptide) was similar to that of GW (1.49 μmol TE / μmol peptide).

The ABTS radical scavenging capacity of the synthetic dipeptides (CG, AY, YA, YF, GW) was underestimated, which was observed in the ABTS corresponding model development and was

also consistent with an antioxidant tripeptide modeling study (N. Chen et al., 2018). The ABTS radical scavenging capacity of YA was significantly higher than that of AY, even though they had the same amino acid composition and length. This is consistent with the feature importance results which showed that the N-terminus residue was more important and Y had been proven to be more favorable at N-terminus compared to A (Udenigwe & Aluko, 2011; Zheng et al., 2016). The ORAC model achieved better activity prediction in YA, AY and GW, where the observed/prediction activity were 1.11/1.16, 1.23/1.13, and 3.48/3.49 μmol TE / μmol peptide, respectively, which showed great prediction performance. In the ORAC prediction, the capacity of YA was only predicted to be slightly higher than that of AY, while the GW exhibited approximately triple the activity of YA. The result implied that the amino acid residue Y was not as important as in ABTS radical scavenging capacity model, while W residue was more favorable to ORAC, even at C-terminus (Zheng et al., 2016). However, dipeptides GG, GT, and GV barely showed any antioxidant activity, although they were predicted to have weak antioxidant capacity. The findings indicated that the models were not suitable for predicting the non-antioxidative peptides and clarified the non-linearity attribute in antioxidant activity prediction. Although the mechanism of DPPH radical scavenging capacity was based on both SET mechanism and HAT mechanism, high DPPH radical scavenging capacity was more corresponding to HAT mechanism, where W played a more important role in antioxidant activity (Figure 3) (N. Liang & Kitts, 2014).

3.4 Conclusion

In summary, we successfully developed prediction models for dipeptide antioxidant activity by XGboost regression method plus other machine learning methods with the latest dataset of antioxidant dipeptides with ABTS radical scavenging capacity and ORAC values. The results showed that N-terminus residues played more important roles in antioxidant activity of dipeptides.

Specifically, Y residue at N-terminus and W at N-terminus were strongly corresponding to high activity in ABTS radical scavenging capacity and ORAC, respectively. Model performance was significantly improved compared to the previous studies on peptide antioxidant activity prediction. A well designed and user-friendly protein hydrolysis simulation tool, R-PepttideCutter, was developed and released. Application of R-PeptideCutter and the models on sorghum protein revealed the dipeptide (YR) encrypted in kafirin protein sequence (Q9XE78) with potentially the highest antioxidant activity, and the enzymes that could be used to release the dipeptide were also targeted. Eight dipeptides derived from the kafirin protein cutting simulation were synthesized and evaluated for their antioxidant activity, and they were used to validate the model performance. The ORAC corresponding model achieved great prediction performance, while the ABTS radical scavenging capacity corresponding model showed underestimation in activity prediction, although both models exhibited inadaptability for dipeptides with low activity or non-activity prediction. The developed models and R-PeptideCutter are capable of being applied to all proteins and identifying bioactive peptides in natural proteins. In addition, the selected variables in model development offer alternative ways to explain the key features that determine bioactivity.

Challenges have remained and are expected to be further studied in the future. Even though there is an improvement in the newly developed hydrolysis simulation tool, the gap between *in silico* simulation and the experimental results exists, and the optimization work is in great demand for the industrialization of protein hydrolysate production. More features for encoding of peptides or amino acids, especially the less complex ones, are also needed for a better QSAR model development and understanding of the key factors contributing the antioxidant activity. In addition, the peptide feature dimension in this study is based on peptide length, which limits the peptide data in antioxidant dipeptides instead of all the reported peptides. Global descriptors, which treat

peptides as a whole for the encoding process, are an important alternative approach to enlarge datasets.

Declaration of interest

The authors declare that there is no known conflict of interest.

Acknowledgements

This is contribution No. 22-279-J from the Kansas Agricultural Experimental Station. This research was supported by the Agriculture and Food Research Initiative Competitive Grant no. 2020-68008-31408 and no. 2021-67021-34495 from the USDA National Institute of Food and Agriculture.

References

Amigo, L., Martínez-Maqueda, D., & Hernández-Ledesma, B. (2020). In Silico and In Vitro Analysis of Multifunctionality of Animal Food-Derived Peptides. *Foods*, *9*(8), Article 8. https://doi.org/10.3390/foods9080991

Anna, T., Alexey, K., Anna, B., Vyacheslav, K., Mikhail, T., & Ulia, M. (2016). Effect of in Vitro Gastrointestinal Digestion on Bioactivity of Poultry Protein Hydrolysate. *Current Research in Nutrition and Food Science Journal*, *4*(Special Issue Nutrition in Conference October 2016), 77–86.

Chen, N., Chen, J., Yao, B., & Li, Z. (2018). QSAR Study on Antioxidant Tripeptides and the Antioxidant Activity of the Designed Tripeptides in Free Radical Systems. *Molecules*, *23*(6), 1407. https://doi.org/10/gdttrq

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10/gdp84q

Daroit, D. J., & Brandelli, A. (2021). In vivo bioactivities of food protein-derived peptides – a current review. *Current Opinion in Food Science*, *39*, 120–129. https://doi.org/10/gngqbj

Deng, B., Long, H., Tang, T., Ni, X., Chen, J., Yang, G., Zhang, F., Cao, R., Cao, D., Zeng, M., & Yi, L. (2019). Quantitative Structure-Activity Relationship Study of Antioxidant Tripeptides Based on Model Population Analysis. *International Journal of Molecular Sciences*, *20*(4), 995. https://doi.org/10/gnmdb7

Du, Z., & Li, Y. (2022). Review and perspective on bioactive peptides: A roadmap for research, development, and future opportunities. *Journal of Agriculture and Food Research*, *9*, 100353. https://doi.org/10.1016/j.jafr.2022.100353

Du, Z., Tian, W., Tilley, M., Wang, D., Zhang, G., & Li, Y. (2022). Quantitative assessment of wheat quality using near-infrared spectroscopy: A comprehensive review. *Comprehensive Reviews in Food Science and Food Safety*, *21*(3), 2956–3009. https://doi.org/10.1111/1541-4337.12958

Du, Z., Wang, D., & Li, Y. (2022). Comprehensive Evaluation and Comparison of Machine Learning Methods in QSAR Modeling of Antioxidant Tripeptides. *ACS Omega*, acsomega.2c03062. https://doi.org/10.1021/acsomega.2c03062

Du, Z., Zeng, X., Li, X., Ding, X., Cao, J., & Jiang, W. (2020). Recent advances in imaging techniques for bruise detection in fruits and vegetables. *Trends in Food Science & Technology*, *99*, 133–141. https://doi.org/10.1016/j.tifs.2020.02.024

FitzGerald, R. J., Cermeño, M., Khalesi, M., Kleekayai, T., & Amigo-Benavent, M. (2020). Application of in silico approaches for the generation of milk protein-derived bioactive peptides. *Journal of Functional Foods*, *64*, 103636. https://doi.org/10.1016/j.jff.2019.103636

Hernández-Ledesma, B., Amigo, L., Recio, I., & Bartolomé, B. (2007). ACE-Inhibitory and Radical-Scavenging Activity of Peptides Derived from β-Lactoglobulin f(19−25). Interactions with Ascorbic Acid. *Journal of Agricultural and Food Chemistry*, *55*(9), 3392–3397. https://doi.org/10.1021/jf063427j

Huang, W.-Y., Majumder, K., & Wu, J. (2010). Oxygen radical absorbance capacity of peptides from egg white protein ovotransferrin and their interaction with phytochemicals. *Food Chemistry*, *123*(3), 635–641. https://doi.org/10.1016/j.foodchem.2010.04.083

Iwaniak, A., Minkiewicz, P., Pliszka, M., Mogut, D., & Darewicz, M. (2020). Characteristics of Biopeptides Released In Silico from Collagens Using Quantitative Parameters. *Foods*, *9*(7), 965. https://doi.org/10/gnmdcc

Je, J.-Y., Cho, Y.-S., Gong, M., & Udenigwe, C. C. (2015). Dipeptide Phe-Cys derived from in silico thermolysin-hydrolysed RuBisCO large subunit suppresses oxidative stress in cultured human hepatocytes. *Food Chemistry*, *171*, 287–291. https://doi.org/10.1016/j.foodchem.2014.09.022

Kalyan, G., Junghare, V., Khan, M. F., Pal, S., Bhattacharya, S., Guha, S., Majumder, K., Chakrabarty, S., & Hazra, S. (2021). Anti-hypertensive Peptide Predictor: A Machine Learning-Empowered Web Server for Prediction of Food-Derived Peptides with Potential Angiotensin-Converting Enzyme-I Inhibitory Activity. *Journal of Agricultural and Food Chemistry*, *69*(49), 14995–15004. https://doi.org/10/gnxx7w

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., & Kanehisa, M. (2008). AAindex: Amino acid index database, progress report 2008. *Nucleic Acids Research*, *36*(Database issue), D202-205. https://doi.org/10.1093/nar/gkm998

Li, S., Bu, T., Zheng, J., Liu, L., He, G., & Wu, J. (2019). Preparation, Bioavailability, and Mechanism of Emerging Activities of Ile-Pro-Pro and Val-Pro-Pro. *Comprehensive Reviews in Food Science and Food Safety*, *18*(4), 1097–1110. https://doi.org/10.1111/1541-4337.12457

Liang, G., & 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 & Combinatorial Science*, *26*(6), 754–763. https://doi.org/10/cfzhpz

Liang, N., & Kitts, D. D. (2014). Antioxidant Property of Coffee Components: Assessment of Methods that Define Mechanisms of Action. *Molecules*, *19*(11), Article 11. https://doi.org/10.3390/molecules191119180

Lorenzo, J. M., Munekata, P. E. S., Gómez, B., Barba, F. J., Mora, L., Pérez-Santaescolástica, C., & Toldrá, F. (2018). Bioactive peptides as natural antioxidants in food products – A review. *Trends in Food Science & Technology*, *79*, 136–147. https://doi.org/10.1016/j.tifs.2018.07.003

Mahmoodi-Reihani, M., Abbasitabar, F., & Zare-Shahabadi, V. (2020). In Silico Rational Design and Virtual Screening of Bioactive Peptides Based on QSAR Modeling. *ACS Omega*, *5*(11), 5951–5958. https://doi.org/10/gnmdcr

Minkiewicz, Iwaniak, & Darewicz. (2019). BIOPEP-UWM Database of Bioactive Peptides: Current Opportunities. *International Journal of Molecular Sciences*, *20*(23), 5978. https://doi.org/10/gnmt2w

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Phongthai, S., D'Amico, S., Schoenlechner, R., Homthawornchoo, W., & Rawdkuen, S. (2018). Fractionation and antioxidant properties of rice bran protein hydrolysates stimulated by in vitro gastrointestinal digestion. *Food Chemistry*, *240*, 156–164. https://doi.org/10/gngrk9

Sabilla, S., Sarno, R., Institut Teknologi Sepuluh Nopember, Triyana, K., & Universitas Gadjah Mada Sekip Utara. (2019). Optimizing Threshold using Pearson Correlation for Selecting Features of Electronic Nose Signals. *International Journal of Intelligent Engineering and Systems*, *12*(6), 81–90. https://doi.org/10/gw3m

Saito, K., Jin, D.-H., Ogawa, T., Muramoto, K., Hatakeyama, E., Yasuhara, T., & Nokihara, K. (2003). Antioxidative Properties of Tripeptide Libraries Prepared by the Combinatorial Chemistry. *Journal of Agricultural and Food Chemistry*, *51*(12), 3668–3674. https://doi.org/10/c2f6wr

Suetsuna, K., Ukeda, H., & Ochi, H. (2000). Isolation and characterization of free radical scavenging activities peptides derived from casein. *The Journal of Nutritional Biochemistry*, *11*(3), 128–131. https://doi.org/10.1016/s0955-2863(99)00083-2

The UniProt Consortium. (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*(D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100

Tian, F., Zhou, P., & Li, Z. (2007). T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides. *Journal of Molecular Structure*, *830*(1–3), 106–115. https://doi.org/10/dqr84q

Tian, M., Fang, B., Jiang, L., Guo, H., Cui, J., & Ren, F. (2015). Structure-activity relationship of a series of antioxidant tripeptides derived from β-Lactoglobulin using QSAR modeling. *Dairy Science & Technology*, *95*(4), 451–463. https://doi.org/10/f7gdk9

Trinidad-Calderón, P. A., Acosta-Cruz, E., Rivero-Masante, M. N., Díaz-Gómez, J. L., García-Lara, S., & López-Castillo, L. M. (2021). Maize bioactive peptides: From structure to human health. *Journal of Cereal Science*, *100*, 103232. https://doi.org/10.1016/j.jcs.2021.103232

Udenigwe, C. C., & Aluko, R. E. (2011). Chemometric Analysis of the Amino Acid Requirements of Antioxidant Food Protein Hydrolysates. *International Journal of Molecular Sciences*, *12*(5), 3148–3161. https://doi.org/10/fc8829

Walker, J. M. (2005). *The proteomics protocols handbook*. Springer.

Wu, S., Qi, W., Su, R., Li, T., Lu, D., & He, Z. (2014). CoMFA and CoMSIA analysis of ACE-inhibitory, antimicrobial and bitter-tasting peptides. *European Journal of Medicinal Chemistry*, *84*, 100–106. https://doi.org/10/f6g45c

Zheng, L., Zhao, Y., Dong, H., Su, G., & Zhao, M. (2016). Structure–activity relationship of antioxidant dipeptides: Dominant role of Tyr, Trp, Cys and Met residues. *Journal of Functional Foods*, *21*, 485–496. https://doi.org/10/f8cqgs

Zhou, P., Liu, Q., Wu, T., Miao, Q., Shang, S., Wang, H., Chen, Z., Wang, S., & Wang, H. (2021). Systematic Comparison and Comprehensive Evaluation of 80 Amino Acid Descriptors in Peptide QSAR Modeling. *Journal of Chemical Information and Modeling*, *61*(4), 1718–1731. https://doi.org/10.1021/acs.jcim.0c01370

Zou, T.-B., He, T.-P., Li, H.-B., Tang, H.-W., & Xia, E.-Q. (2016). The Structure-Activity Relationship of the Antioxidant Peptides from Natural Proteins. *Molecules*, *21*(1), Article 1. https://doi.org/10.3390/molecules21010072

Figures and tables

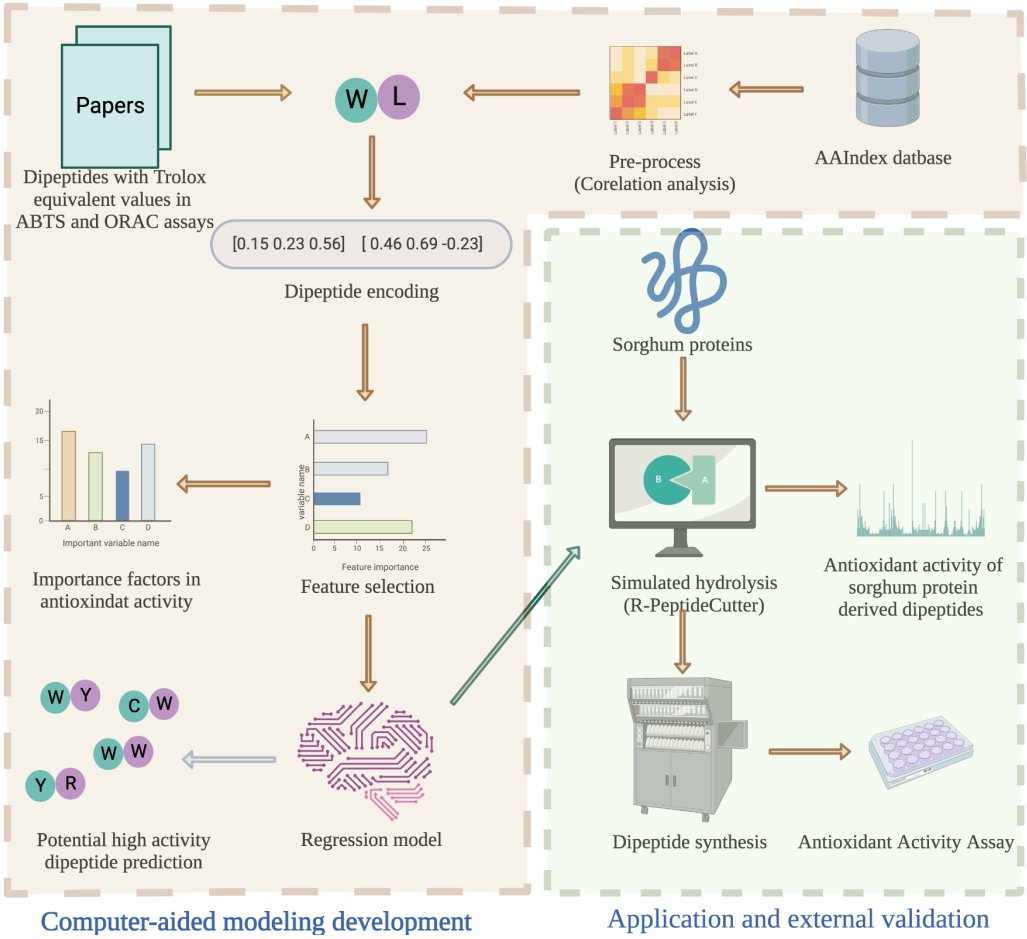


Figure 3-1 Technical route for computer-aided approaches for screening antioxidative dipeptides and application to sorghum proteins.

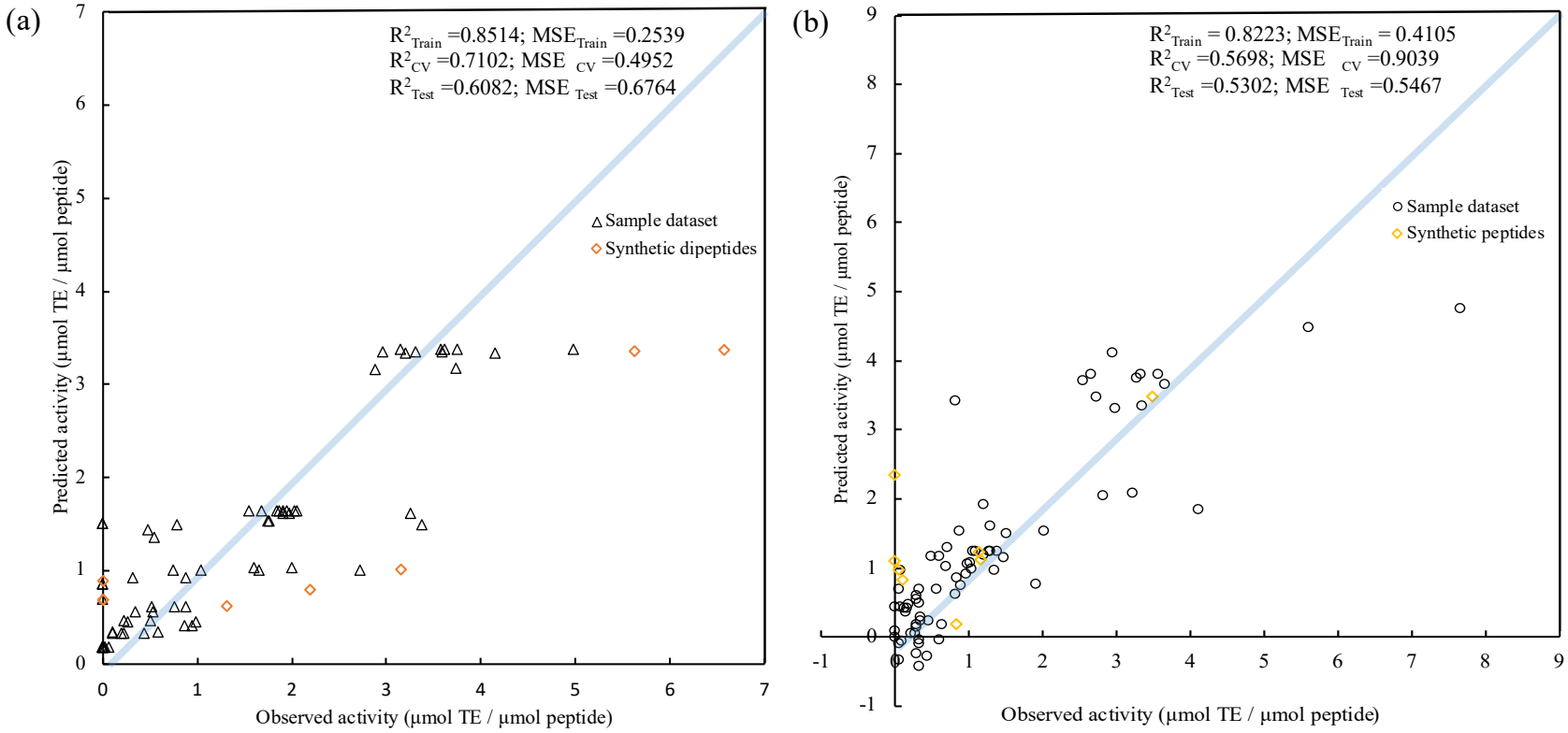


Figure 3-2 Relationship between observed and predicted antioxidant activities: (a) ABTS scavenging activity; (b) ORAC.

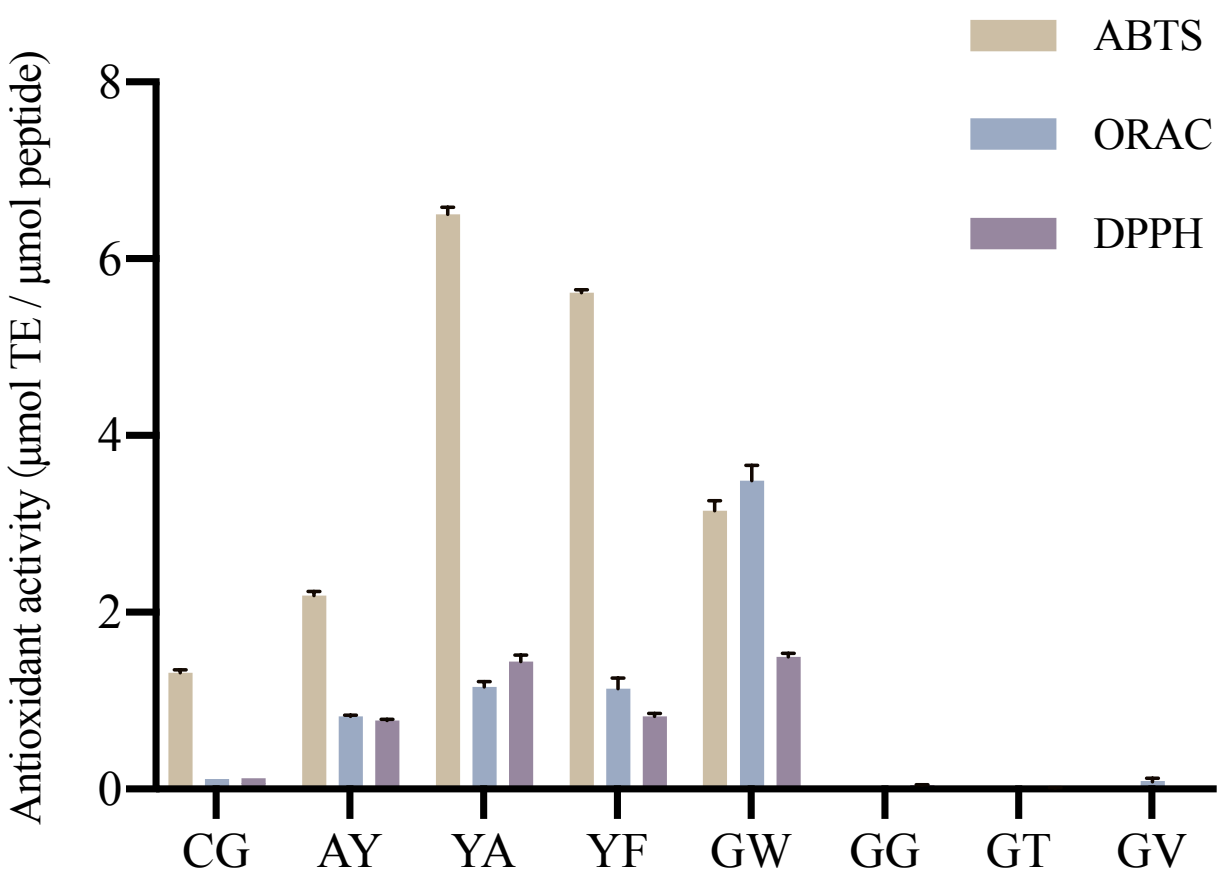


Figure 3-3 Antioxidant activities of different synthetic dipeptides determined by ABTS, ORAC and reducing power assays.

Note: CG, AY, YA, YF, GW, GG, GT, GV refer to dipeptide Cys-Gly, Ala-Tyr, Tyr-Ala, Tyr-Phe, Gly-Trp, Gly-Gly, Gly-Thr, and Gly-Val, respectively.

Table 3-1 Amino acid positions, variable importance, and description of selected variables by XGBoost regression for ABTS radical scavenging capacity dataset.

Accession number	Amino acid position	Variable importance	Description
LIFS790103	N-terminus	0.4321	Conformational preference for antiparallel beta-strands
CHAM820102	N-terminus	0.1647	Free energy of solution in water
PALJ810108	N-terminus	0.0834	Normalized frequency of alpha-helix in alpha+beta class
ARGP820102	N-terminus	0.0732	Signal sequence helical potential
KARS160122	N-terminus	0.0268	Weighted second smallest eigenvalue of the weighted Laplacian matrix
OOBM850103	N-terminus	0.0167	Optimized transfer energy parameter
YUTK870102	N-terminus	0.0139	Unfolding Gibbs energy in water, pH9.0
CHAM830103	C-terminus	0.0419	The number of atoms in the side chain labelled 1+1
MAXF760102	C-terminus	0.0281	Normalized frequency of extended structure
QIAN880119	C-terminus	0.0272	Weights for beta-sheet at the window position of -1
MCMT640101	C-terminus	0.0116	Refractivity

Note: Variable importance is presented as absolute values. The detailed information of these selected variables is available at https://www.genome.jp/aaindex/.

Table 3-2 Amino acid positions, variable importance, and description of selected variables by XGBoost regression for ORAC dataset.

Accession number	Amino acid position	Variable importance	Description
CHOP780215	N-terminus	0.2067	Frequency of the 4th residue in turn
PALJ810108	N-terminus	0.2046	Normalized frequency of alpha-helix in alpha+beta class
CHOP780205	N-terminus	0.0918	Normalized frequency of C-terminus helix
LIFS790103	N-terminus	0.0819	Conformational preference for antiparallel beta-strands
BIOV880102	N-terminus	0.0323	Information value for accessibility; average fraction 23%
FODM020101	N-terminus	0.0153	Propensity of amino acids within pi-helices
BROC820102	N-terminus	0.0112	Retention coefficient in HFBA
CHOP780211	C-terminus	0.0557	Normalized frequency of C-terminus non beta region
BUNA790103	C-terminus	0.0244	Spin-spin coupling constants 3JHalpha-NH
QIAN880136	C-terminus	0.0181	Weights for coil at the window position of 3
PALJ810111	C-terminus	0.0122	Normalized frequency of beta-sheet in alpha+beta class
CHAM830105	C-terminus	0.0121	The number of atoms in the side chain labelled 3+1
KARS160112	C-terminus	0.0119	Second smallest eigenvalue of the Laplacian matrix of the graph
DESM900101	C-terminus	0.1105	Membrane preference for cytochrome b: MPH89

Note: Variable importance is presented as absolute values. The detailed information of these selected variables is available at https://www.genome.jp/aaindex/.

Chapter 4 - UniDL4BioPep: a universal deep learning architecture for binary classification in peptide bioactivity*

Abstract

Identification of potent peptides through model prediction can reduce benchwork in wet experiments. However, the conventional process of model buildings can be complex and time-consuming due to challenges such as peptide representation, feature selection, model selection, hyperparameter tuning, etc. Recently, advanced pre-trained deep learning-based language models (LM) have been released for protein sequence embedding and applied to structure and function prediction. Based on these developments, we have developed UniDL4BioPep, a universal deep-learning model architecture for transfer learning in bioactive peptide binary classification modeling. It can directly assist users in training a high-performance deep-learning model with a fixed architecture and achieve cutting-edge performance to meet the demands in efficiently novel bioactive peptide discovery. To the best of our best knowledge, this is the first time that a pretrained biological language model is utilized for peptide embeddings and successfully predicts peptide bioactivities through large-scale evaluations of those peptide embeddings. The model was also validated through uniform manifold approximation and projection analysis. By combining the LM with a convolutional neural network, UniDL4BioPep achieved greater performances than the respective state-of-the-art models for 15 out of 20 different bioactivity dataset prediction tasks. The accuracy, Mathews correlation coefficient, and area under the curve were 0.7-7%, 1.23-26.7%, and 0.3-25.6% higher, respectively. A user-friendly web server of UniDL4BioPep for the tested bioactivities is established and freely accessible at https://nepc2pvmzy.us-east-1.awsapprunner.com. The source codes, datasets, and templates of UniDL4BioPep for other bioactivity fitting and prediction tasks are available at https://github.com/dzjxzyd/UniDL4BioPep.

Keywords:

Protein language model; bioactive peptide; protein sequence classification; deep learning; universal architecture

---

* Du, Z., Ding, X., Xu, Y., & Li, Y. (2023). UniDL4BioPep: a universal deep learning architecture for binary classification in peptide bioactivity. Briefings in Bioinformatics, 24(3), bbad135.

4.1 Introduction

Bioactive peptides (BPs) are protein fragments with positive biological effects, which can be produced from sustainable food protein sources through enzymatic hydrolysis or fermentation (Du & Li, 2022b; Ulug et al., 2021). BPs have gained tremendous attention from both researchers and consumers, which is driven by the increasing demand for natural nutraceuticals, concerns of synthetic products, sustainability, and most importantly, the diverse bioactivities exhibited by BPs and their potential to relieve the health burdens (FitzGerald et al., 2020; Iwaniak et al., 2019). The global BPs market, excluding peptide drugs, is expected to double from 48.6 billion USD in 2020 to 95.7 billion USD by 2028 (https://www.verifiedmarketresearch.com/product/bioactive-peptides-market/). Besides, BPs related research publications were tripled in the past ten years, and thousands of BPs have been identified and deposited in free publicly-accessible databases (e.g., DFBP, BIOPEP-UWM, $CAMP_{R3}$, AHTPDB, SpirPep, etc.) (Anekthanakul et al., 2018; Du et al., 2023; Du & Li, 2022b; Kumar et al., 2015; Minkiewicz et al., 2019; Pang et al., 2022; Qin et al., 2022; Waghu et al., 2016). So far, some peptides have been commercialized for applications such as drugs, nutraceuticals, and cosmeceuticals (Du & Li, 2022b).

Conventionally, biochemistry approaches including protein pretreatments, protein hydrolysis (or microbial fermentation), hydrolysate fractionation & purification, and *in vitro* or *in vivo* evaluation, are the mainstream in the novel BPs identification (Barati et al., 2020; Duffuler et al., 2022; Perez Espitia et al., 2012; Tu et al., 2018; Ulug et al., 2021; Wen et al., 2020). However, these technical routes are hindered by their low efficiency, high cost, and heavy reliance on advanced instruments and skilled personnel. To overcome these limitations, some researchers have turned to employ bioinformatics tools, particularly quantitative structure-activity relationships (QSAR) modeling, to fully utilize accumulated BPs data and conduct QSAR models for predicting

BPs and decision-making before wet bench research (J. Chen et al., 2021; Du & Li, 2022a; Olsen et al., 2020; Tu et al., 2018).

There are three essential steps in QSAR modeling: BPs data collection, peptide representation/encoding/feature extraction, and model development. The most challenging part is the peptide representation, where peptide sequences are converted into numerical vectors or matrix by different descriptors for further model development (Du et al., 2023). Up to now, various encoding approaches have been proposed, including the physicochemical properties and biochemical properties of amino acids (e.g., AAIndex), amino acid descriptors (e.g., z-scale), amino acid composition, molecular descriptors and fingerprints of the peptide (e.g., three-dimensional (3D) structure information), composition-based descriptors (e.g., amino acid composition and dipeptide composition), binary profile (one-hot encoding), etc.(Charoenkwan, Kanthawong, et al., 2020; Charoenkwan, Nantasenamat, et al., 2020; J. Chen et al., 2021; Kalyan et al., 2021; Olsen et al., 2020). There are some disadvantages in the application of these encoding methods. Firstly, the simple combination of these descriptors leads to high dimensional features which include redundant information, hence undermining the predictive performance. Even though many feature selection methods are available to solve the high-dimension problem, the process is time-consuming and tedious because of enormous trial-and-error tests (Charoenkwan, Nantasenamat, Hasan, Manavalan, et al., 2021; Du, Tian, et al., 2022; Du, Wang, et al., 2022). Additionally, the descriptors may not accurately represent peptides. For example, composition-based descriptors lack information about the sequential order, and physicochemical properties can only provide information that can be determined.

In natural language processing (NLP) tasks, word embeddings have been prevalently used to capture semantic properties and linguistic relationships between words and numerical output

representations of raw text data for further machine learning (ML) model development (Wang et al., 2018). The protein sequence is highly similar to human beings' natural languages, where different amino acids compose a 'language of life' (Elnaggar et al., 2022). Most recently, such pre-trained deep learning-based language models (LMs) for protein sequence embeddings have been released, such as evolutionary scale modeling (ESM), unified representation (UniRep), ProtTrans, etc. (Alley et al., 2019; Dallago et al., 2021; Elnaggar et al., 2022; Z. Lin et al., 2023; Rao et al., 2020; Rives et al., 2021). These LMs are trained on large datasets (billions of sequences) to internalize the information encrypted in protein sequences and have exhibited enormous potential in descriptive representations (embeddings) of proteins only relying on their sequences for comparable or improved predictive power in downstream tasks, such as subcellular location, structure prediction, function prediction, etc. (Dallago et al., 2021; Elnaggar et al., 2022; Z. Lin et al., 2023). Very recently, Based on ESM-2 LM, the Meta Fundamental AI Research Protein Team (FAIR) developed a protein 3D structure predictor (ESMFold). Using the similar fold blocks to AlphaFold2, ESMFold was up to 60 times faster than AlphaFold2 and had lower template modeling score (TM-score) in CAMEO and CASP14 datasets (Z. Lin et al., 2023).

Given that the building blocks of peptides (i.e., amino acids) are the same as those in proteins, and that their bioactivity is comparable to the functionality of proteins, peptide sequences can also be embedded using LMs for bioactivity prediction. To the best of our knowledge, there are few attempts to employ these cutting-edge LMs for peptide embeddings in bioactivity prediction. Transfer learning relying on deep learning has been used to simplify efforts in model architecture design when new tasks can be solved by the previous knowledge (Elnaggar et al., 2022; Tammina, 2019). Convolutional neural network (CNN) exhibited promising performance in BPs prediction (Charoenkwan, Nantasenamat, Hasan, Manavalan, et al., 2021; J. Chen et al., 2021;

Olsen et al., 2020; Veltri et al., 2018). Since each element in the peptide embeddings vector/matrix is unique under the specific peptide sequence and the output feature dimension of ESM is a fixed length, it is not necessary to conduct additional feature extraction and processing to unify the feature dimension after embeddings. The fixed feature dimension makes it more suitable for the designing of feature extraction layers in CNN and applying the CNN model to other BP datasets. Therefore, the combination of those LMs and deep learning models may build a universal modeling architecture for BPs prediction among different BP datasets.

The objective of this study was to build a model architecture that relies on a pretrained-LM and a CNN model, and to investigate the potential of this architecture for transfer learning in different BPs prediction tasks. A state-of-the-art (SOTA) LM, ESM-2, was first used for peptide embeddings, and then a CNN model was proposed to be compatible with the peptide embeddings for BPs prediction (Figure 1). The universal architecture UniDL4BioPep was used to fit 20 different BPs' benchmark datasets, and the performance of the model was compared with the SOTA performance. The UniDL4BioPep has the potential to be used to fit various BP datasets with different bioactivities and generate high-performance models for prediction tasks. It could also inspire future prediction model development for bioactivity prediction.

4.2 Materials and methods

4.2.1 Benchmark datasets

All the benchmark datasets were retrieved from the reported SOTA models to conduct a fair and unbiased performance evaluation and comparison. Twenty BPs datasets for eighteen different bioactivities were collected, including angiotensin-converting enzyme (ACE) inhibitory activity (anti-hypertension) (Manavalan et al., 2019), dipeptidyl peptidase IV (DPPIV) inhibitory activity (anti-diabetes) (Charoenkwan, Kanthawong, et al., 2020), bitter (Charoenkwan,

Nantasenamat, Hasan, Manavalan, et al., 2021), umami (Charoenkwan, Yana, Nantasenamat, et al., 2020), antimicrobial activity (Pang et al., 2022), antimalarial activity (Charoenkwan, Schaduangrat, et al., 2022), quorum-sensing (QS) activity (Wei et al., 2020), anticancer activity (Agrawal et al., 2021), anti-methicillin-resistant *S. aureus* (MRSA) strains activity (Charoenkwan, Kanthawong, et al., 2022), tumor T cell antigens (TTCA) (Charoenkwan, Nantasenamat, et al., 2020), blood-brain barrier (Dai et al., 2021), anti-parasitic activity (Zhang et al., 2022), neuropeptide (Bin et al., 2020; S. Chen et al., 2022), antibacterial activity (Pinacho-Castellanos et al., 2021), antifungal activity (Pinacho-Castellanos et al., 2021), antiviral activity (Pinacho-Castellanos et al., 2021), toxicity (Wei et al., 2021), and antioxidant activity (Olsen et al., 2020). The dataset information is summarized in Table 1, and further details on the benchmark datasets can be found in previous studies (also accessible at https://github.com/dzjxzyd/UniDL4BioPep).

4.2.2 Language model for peptide embeddings and data processing

ESM is a LM project initiated by FAIR in 2019 (https://github.com/facebookresearch/esm). Most recently, an updated version of ESM was released as ESM-2, which was trained on the UR50/D 2021_04 dataset and outperformed across a range of structure prediction tasks (Z. Lin et al., 2023). ESM-2 contained 6 LMs varying from 48 layers and 15 billion parameters for 5120 output embeddings dimensions to 6 layers and 8 million parameters for 320 output embeddings dimensions. Due to the relatively small size of BPs datasets, the LM (esm2_t6_8M_UR50D) with the lowest output embeddings dimension was selected for peptide embeddings to simplify the CNN model architecture and avoid the curse of dimensionality in model development.

The dataset splitting followed the previous reports where the training and test datasets samples were the same as those used in the SOTA models. In each bioactivity benchmark dataset

test, each peptide sequence was first input into the pretrained ESM model to generate a 1*320 dimension numerical vector (Figure 1). Before the CNN model development, a min-max normalization was conducted relying on a training dataset to scale the features in range (0, 1). Normalization in the test dataset is based on the maximum values and minimum from the training dataset. To understand the effectiveness of the ESM-2 in peptide embeddings for bioactivity prediction, uniform manifold approximation and projection (UMAP) was used to visualize the high dimension embeddings in a two-dimension graph (McInnes et al., 2020). Besides, visualization based on t-distributed stochastic neighbor embedding (t-SNE) was also conducted and provided (Maaten & Hinton, 2008).

4.2.3 CNN model architecture

The CNN model was built with Keras framework (http://www.keras.io). For this study, there are totally eight layers (Figure 1) in the CNN models. The first layer was the input layer, where the peptide sequence information was represented by a numerical vector generated by the LM model. Then, the next layer is a 1D convolutional layer with filter sizes of 32, kernel size of 3, and ReLU activation, and then a 1D max pooling layer was added to reduce the dimensionality. In addition, batch normalization and dropout (rate=0.15) were added to avoid overfitting. Subsequently, the output of the convolutional layer was flattened and followed by a dense layer containing 64 hidden neurons with ReLU activation function and a dropout rate of 0.15. The last layer is also a dense layer with two neurons with a SoftMax activation function for binary classification. Stochastic gradient descent (SGD) was chosen as the optimizer to accelerate the model fitting and maximize its performance, and step decay was set for the learning rate during the epoch. Besides, an early stop was also set, and the best weight would be stored under the

consideration of validation accuracy. The largest epoch time is 200, but in practice, it usually stops around 100 epochs, which takes around two to three minutes on the Google Colab platform.

4.2.4 Model evaluation

Accuracy (ACC), balanced accuracy (BACC), sensitivity (Sn), specificity (Sp), Matthews correlation coefficient (MCC), and area under the curve (AUC) were adopted to evaluate the model performance. Those parameters were calculated based on the number of true positive (TP), false positive (FP), false negative (FN), and true negative (TN). They are calculated by the following equations:

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{4.1}$$

$$Sn = \frac{TP}{TP+FN} \tag{4.2}$$

$$Sp = \frac{TN}{TN+FP} \tag{4.3}$$

$$BACC = 0.5 * Sn + 0.5 * Sp \tag{4.4}$$

$$MCC = \frac{(TP*TN)-(FN*FP)}{\sqrt{(TP+FN)*(TN+FP)*(TP+FP)*(TN*FN)}} \tag{4.5}$$

The AUC is calculated relying on the sklearn 'roc_auc_score' function through the portability distribution of the model prediction in the test dataset.

4.2.5 UniDL4BioPep-FL for imbalanced datasets

In some cases, the bioactive peptide dataset may not be well balanced, resulting in one class being under-represented, which can cause the model in learning less from the minority class (Lemaıtre & Nogueira, 2017). Therefore, the focal loss (FL) function was introduced to tackle the challenge in the imbalanced dataset relying on its ability to down-weight easy examples and thus focus training on hard negatives (T.-Y. Lin et al., 2020). Besides, the FL function allows for adjustment of the weighting factor to balance the importance of positive/negative samples. The

modified version is named UniDL4BioPep-FL, which contains two hyperparameters, the focusing parameter (gamma) and weighting factor (alpha), which need to be tuned and specified compared to the original UniDL4BioPep model. To be compatible with FL function, the last layer was changed to a one neuron with a sigmoid activation function for binary classification. The equation for FL calculation is below:

$$Focal\ loss = \begin{cases} -\alpha(1-p)^{\gamma}\log(p) & if\ y = 1 \\ -(1-\alpha)p^{\gamma}\log(1-p) & otherwise \end{cases} \quad (4.6)$$

Where p is the model's estimated probability for the class with label y=1 (positive); γ is the focal parameter ($\gamma \geqslant 0$); α is the weighting factor.

4.3 Results and discussion

4.3.1 Peptide embeddings analysis

The ESM-2 LMs were trained with the masked language objective, which enabled the LMs to learn dependencies between residues and internalize sequence patterns during millions of repetitions of evolutionarily diverse protein sequences (Z. Lin et al., 2023). The architecture of EMS-2 LMs is based on bidirectional encoder representations from transformers (BERT) style encoder, which is known for its ability to return different representations for the same words depending on the contexts (Devlin et al., 2019). The residues in a peptide sequence contain both information about the residues and information about their position/sequential information. One-hot encoding is a solution to capture both types of information encoded in the peptide sequence. However, the feature matrix is a sparse matrix, and the matrix shape is dependent on the length of the peptide (Olsen et al., 2020). This can limit the performance of the model since capturing the feature in a spare matrix is difficult. ESM-2 LMs, on the other hand, can generate fixed-length peptide embeddings and dense matrices.

UMAP is a general dimension reduction technique for visualization purpose. Similar to t-SNE, it prioritizes the preservation of local distances over global distances and has the ability to handle outliers and non-linear relationships, resulting in better performance than principal component analysis in biological data processing (McInnes et al., 2020). In this study, UMAP distribution of positive and negative samples in both training and testing datasets were plotted in two-dimensional feature space by using 320-dimensional embeddings. As shown in Figure 2, most positive and negative samples are clearly distributed in two clusters. Such distinct differences in the dimensional feature space are generally an indication of the great representation of peptide sequence information for further model development and correspond to optimal models (Manavalan et al., 2019; Wei et al., 2020). This is also consistent with UniDL4BioPep's performance in these datasets (Table 2). There is an apparent discrepancy in the number of positive and negative samples in several datasets (Figure 2 e, f, k), which is due to an imbalance in the original datasets. Similar observations were also noticed in the t-SNE distribution graphs (Supplementary Document S1-Figure S1). However, the clustering boundary in t-SNE is more ambiguous compared to UMAP. This may be because of the preservation of more global structures in t-SNE, or it may not be suitable for visualizing clusters in these cases, as has been reported in real-world datasets (McInnes et al., 2020; Yang et al., 2021).

4.3.2 Performance evaluation of UniDL4BioPep with SOTA models on the independent test datasets of BPs

To evaluate the robustness and accuracy of our proposed UniDL4BioPep in different BPs datasets, the performance of the SOTA models using the same datasets for training and performance evaluation was collected and compared with the performance of UniDL4BioPep (Table 2). There was a total of twenty datasets for the eighteen bioactivities, where both

antimalarial activity and anticancer activity had two datasets. Among the twenty datasets (Table 2), UniDL4BioPep achieved better performance than the SOTA models in fifteen datasets, where the ACC, MCC, and AUC were 0.7-7%, 1.23-26.7%, and 0.3-25.6% higher than those in the SOTA models. For the remaining datasets, the performance of UniDL4BioPep was also comparable to that achieved by the SOTA models. It is worth noting that during model development in all these datasets, UniDL4BioPep only performed data loading and waited for the trained model for performance evaluation in independent test datasets, while each SOTA model had different model architectures, feature selection strategy, and hyperparameter tuning. Such performance comparisons demonstrate the great generalization ability of UniDL4BioPep, and its tremendous potential to adapt to new target (i.e., bioactivity or other functions) predictions. The only requirement for the new bioactivity prediction model development is a high-quality and well-labeled dataset.

A total of twenty SOTA model architectures for the respective bioactivities are briefly reviewed and compared with UniDL4BioPep in Table 3. Most SOTA models for the twenty bioactivities adopted similar peptide representation methods, including composition-based methods (e.g., amino acid composition and dipeptide composition), binary profile (one-hot encoding), and physicochemical properties of amino acids (e.g., Amino Acid Index databases). As previously mentioned, these embedding approaches suffer from the loss of sequential information, insufficient description of peptides, and the curse of dimensionality. Most researchers still tend to choose traditional machine learning methods (e.g., random forest) due to the limited data and better explainability compared to neural networks. Therefore, feature selection is essential for simplifying the model complexity and increasing prediction power when the feature dimension is high or too many features are combined. Most recently, Charoenkwan et al. conducted a series of

experiments demonstrating better performance and set new SOTA records based on the scoring card method (SCM), which also had a built-in feature selection procedure based on a genetic algorithm (Charoenkwan, Kanthawong, et al., 2020; Charoenkwan, Yana, Nantasenamat, et al., 2020, 2020; Charoenkwan, Kanthawong, et al., 2022; Charoenkwan, Schaduangrat, et al., 2022). There were also studies employing the idea of word embedding from NLP. In the study of Pang et al., 31 million amino acid sequences from the Pfam dataset were used to train a language model by self-learning for peptide embeddings (Pang et al., 2022). Other NLP-based approaches, including Word2Vec, TFIDF, Pep2Vec, and FastText were also used in peptide embeddings (Charoenkwan, Nantasenamat, Hasan, Manavalan, et al., 2021; S. Chen et al., 2022). This method achieved comparable performance compared to the SOTA prediction performance in bitterness, neuropeptide, and toxicity from previous studies. Specifically, the better performance in bitterness prediction to some extent supports the idea that with appropriate feature dimensions, deep learning can also achieve great performance in a small dataset (e.g., a total of 320 positive and 320 negative samples) (Charoenkwan, Nantasenamat, Hasan, Moni, et al., 2021; Charoenkwan, Yana, Schaduangrat, et al., 2020).

In UniDL4BioPep, peptide representation is generated by LM (ESM-2), which takes into account each residue and sequential information to generate fixed-length peptide embeddings. There are also other ESM-2 based LMs with longer fixed-length output dimensions (Z. Lin et al., 2023). However, lower dimensions correspond to fewer parameters in the following deep learning model, and thus we chose the available pretrained ESM-2 with the lowest output dimension (320 dimensions). ESM-2 uses bidirectional encoder representations from transformers (BERT)-based architecture with modifications. The output is the hidden state of the ESM-2 model, which is tuned by millions of protein sequences. Hence, both the value of each element in the output vector and

the relationship between each element is important in representing the peptide sequence and the loss of each element might undermine the information contained in the peptide embeddings. As a result, feature selection was not conducted. The CNN architecture is inspired by the CNN architectures in literature and also practical application of CNN in the handwritten digits recognition MNIST database with a few modifications based on the dimension of peptide embeddings from ESM-2 (Charoenkwan, Nantasenamat, Hasan, Manavalan, et al., 2021; Olsen et al., 2020). The shallow CNN model in UniDL4BioPep had fewer parameters and was compatible with embedding dimensions from the ESM-2. BPs with small datasets (e.g., bitter, umami, Quorum sensing, Tumor T cell antigens, Anti-parasitic activity, etc.) had improved performance in test dataset prediction with UniDL4BioPep, compared to the SOTA models. Besides, a modified universal model version, UniDL4BioPep-FL, is introduced to meet the needs of imbalanced dataset modeling relying on the FL function. Improvement in the balance of Sn and Sp were observed in six imbalanced benchmark datasets (umami, antimicrobial activity, antimalarial activity (main and alternative datasets), anti-MRSA strains activity, Tumor T cell antigens), which support its validity. Overall, UniDL4BioPep is a universal architecture in BPs prediction and overcomes the challenges in dataset size (compatible), peptide representation (repeatable and fixed approach), feature selection (not needed), machine learning methods selection (not needed), and hyperparameter tuning (universal to the fixed-length feature dimension, only needed for two hyperparameters in UniDL4BioPep-FL).

4.4 Conclusion

Novel bioactive peptide exploration boosted by bioinformatics can significantly reduce experimental periods and cost. As such, fast and accurate prediction models are highly desirable. In this study, we first introduced the latest pretrained language model based on millions of protein

sequences for peptide embeddings and further application in bioactivity prediction. Besides, relying on the fixed-length output of the embeddings, corresponding convolutional layers and dense layers were designed for feature extraction of the embeddings and bioactivity prediction. Because of the self-learned characteristics of CNN models, it is possible to build a universal architecture for general bioactivity prediction, and thus UniDL4BioPep was proposed. The model was trained and evaluated on eighteen different bioactivities with a total of twenty datasets and was compared with the state-of-the-art models. The UMAP results showed that the pretrained language model (esm2_t6_8M_UR50D) is suitable for peptide embeddings. The CNN model architecture was compatible with the embeddings and exhibited better or comparable performance than the state-of-the-art models, which further confirmed the feasibility and superiority of using a language model in peptide embeddings over conventional approaches.

UniDL4BioPep has great potential to be applied to other bioactive peptide predictions and generate prediction models with comparable performance. Users would only need to prepare their BP dataset in MS Excel in a required format (the first column for sequence and the second column for label). Then, UniDL4BioPep can automatically read the dataset, embed the peptide sequences, split datasets (8:2 ratio as train and test datasets), generate a prediction model, evaluate the model performance, and save the model for further prediction needs. Besides, we also provide a modified version (UniDL4BioPep-FL) for advanced usage of the architecture for imbalanced datasets, where two hyperparameters are needed to be tuned. The templates for UniDL4BioPep-based model training for other BP datasets and the employment of generated model for prediction are available at https://github.com/dzjxzyd/UniDL4BioPep. Besides, UniDL4BioPep/ UniDL4BioPep-FL can also be used for multiclass classification model development, which

further expands its application scenario. All the scripts were written and tested on Google Colab, which allows users to execute them through browsers.

With the great predictive performance of UniDL4BioPep, a user-friendly web server for accelerating peptide screening and practical applications has been deployed and is freely available online at https://nepc2pvmzy.us-east-1.awsapprunner.com/. Users can choose a prediction model and submit a peptide sequence, a batch of sequences, or a file in FASTA, txt, Microsoft Excel formats to the webserver and receive the predicted results in real time. We anticipate the proposed UniDL4BioPep architectures will be an efficient and powerful tool for BP discovery and inspire future model design.

Declaration of interest

The authors declare that there are no known conflicts of interest.

Acknowledgements

This is contribution No. 23-192-J from the Kansas Agricultural Experimental Station. This work was supported in part by the Agriculture and Food Research Initiative Competitive Grant no. 2020-68008-31408 and no. 2021-67021-34495 from the USDA National Institute of Food and Agriculture, and a seed grant from the Global Food Systems initiative of Kansas State University.

References

Agrawal, P., Bhagat, D., Mahalwal, M., Sharma, N., & Raghava, G. P. S. (2021). AntiCP 2.0: An updated model for predicting anticancer peptides. *Briefings in Bioinformatics*, *22*(3), bbaa153. https://doi.org/10.1093/bib/bbaa153

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., & Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, *16*(12), Article 12. https://doi.org/10.1038/s41592-019-0598-1

Anekthanakul, K., Hongsthong, A., Senachak, J., & Ruengjitchatchawalya, M. (2018). SpirPep: An in silico digestion-based platform to assist bioactive peptides discovery from a genome-wide database. *BMC Bioinformatics*, *19*(1), 149. https://doi.org/10.1186/s12859-018-2143-0

Barati, M., Javanmardi, F., Mousavi Jazayeri, S. M. H., Jabbari, M., Rahmani, J., Barati, F., Nickho, H., Davoodi, S. H., Roshanravan, N., & Mousavi Khaneghah, A. (2020). Techniques, perspectives, and challenges of bioactive peptide generation: A comprehensive systematic review. *Comprehensive Reviews in Food Science and Food Safety*, *19*(4), 1488–1520. https://doi.org/10.1111/1541-4337.12578

Bin, Y., Zhang, W., Tang, W., Dai, R., Li, M., Zhu, Q., & Xia, J. (2020). Prediction of Neuropeptides from Sequence Information Using Ensemble Classifier and Hybrid Features. *Journal of Proteome Research*, *19*(9), 3732–3740. https://doi.org/10.1021/acs.jproteome.0c00276

Charoenkwan, P., Kanthawong, S., Nantasenamat, C., Hasan, Md. M., & Shoombuatong, W. (2020). iDPPIV-SCM: A Sequence-Based Predictor for Identifying and Analyzing Dipeptidyl Peptidase IV (DPP-IV) Inhibitory Peptides Using a Scoring Card Method. *Journal of Proteome Research*, *19*(10), 4125–4136. https://doi.org/10.1021/acs.jproteome.0c00590

Charoenkwan, P., Kanthawong, S., Schaduangrat, N., Li', P., Moni, M. A., & Shoombuatong, W. (2022). SCMRSA: A New Approach for Identifying and Analyzing Anti-MRSA Peptides Using Estimated Propensity Scores of Dipeptides. *ACS Omega*, *7*(36), 32653–32664. https://doi.org/10.1021/acsomega.2c04305

Charoenkwan, P., Nantasenamat, C., Hasan, M. M., Manavalan, B., & Shoombuatong, W. (2021). BERT4Bitter: A bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides. *Bioinformatics*, *37*(17), 2556–2562. https://doi.org/10.1093/bioinformatics/btab133

Charoenkwan, P., Nantasenamat, C., Hasan, M. M., & Shoombuatong, W. (2020). iTTCA-Hybrid: Improved and robust identification of tumor T cell antigens by utilizing hybrid feature representation. *Analytical Biochemistry*, *599*, 113747. https://doi.org/10.1016/j.ab.2020.113747

Charoenkwan, P., Nantasenamat, C., Hasan, Md. M., Moni, M. A., Lio', P., & Shoombuatong, W. (2021). iBitter-Fuse: A Novel Sequence-Based Bitter Peptide Predictor by Fusing Multi-View Features. *International Journal of Molecular Sciences*, *22*(16), 8958. https://doi.org/10.3390/ijms22168958

Charoenkwan, P., Schaduangrat, N., Lio, P., Moni, M. A., Chumnanpuen, P., & Shoombuatong, W. (2022). iAMAP-SCM: A Novel Computational Tool for Large-Scale Identification of Antimalarial Peptides Using Estimated Propensity Scores of Dipeptides. *ACS Omega*, *7*(45), 41082–41095. https://doi.org/10.1021/acsomega.2c04465

Charoenkwan, P., Yana, J., Nantasenamat, C., Hasan, Md. M., & Shoombuatong, W. (2020). iUmami-SCM: A Novel Sequence-Based Predictor for Prediction and Analysis of Umami Peptides Using a Scoring Card Method with Propensity Scores of Dipeptides. *Journal of Chemical Information and Modeling*, *60*(12), 6666–6678. https://doi.org/10.1021/acs.jcim.0c00707

Charoenkwan, P., Yana, J., Schaduangrat, N., Nantasenamat, C., Hasan, Md. M., & Shoombuatong, W. (2020). iBitter-SCM: Identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides. *Genomics*, *112*(4), 2813–2822. https://doi.org/10.1016/j.ygeno.2020.03.019

Chen, J., Cheong, H. H., & Siu, S. W. I. (2021). xDeep-AcPEP: Deep Learning Method for Anticancer Peptide Activity Prediction Based on Convolutional Neural Network and Multitask Learning. *Journal of Chemical Information and Modeling*, *61*(8), 3789–3803. https://doi.org/10.1021/acs.jcim.1c00181

Chen, S., Li, Q., Zhao, J., Bin, Y., & Zheng, C. (2022). NeuroPred-CLQ: Incorporating deep temporal convolutional networks and multi-head attention mechanism to predict neuropeptides. *Briefings in Bioinformatics*, *23*(5), bbac319. https://doi.org/10.1093/bib/bbac319

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

Dallago, C., Schütze, K., Heinzinger, M., Olenyi, T., Littmann, M., Lu, A. X., Yang, K. K., Min, S., Yoon, S., Morton, J. T., & Rost, B. (2021). Learned Embeddings from Deep Learning to Visualize and Predict Protein Sets. *Current Protocols*, *1*(5), e113. https://doi.org/10.1002/cpz1.113

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*. https://doi.org/10.48550/arXiv.1810.04805

Du, Z., Comer, J., & Li, Y. (2023). Bioinformatics approaches to discovering food-derived bioactive peptides: Reviews and perspectives. *TrAC Trends in Analytical Chemistry*, 117051. https://doi.org/10.1016/j.trac.2023.117051

Du, Z., & Li, Y. (2022a). Computer-Aided Approaches for Screening Antioxidative Dipeptides and Application to Sorghum Proteins. *ACS Food Science & Technology*. https://doi.org/10.1021/acsfoodscitech.2c00286

Du, Z., & Li, Y. (2022b). Review and perspective on bioactive peptides: A roadmap for research, development, and future opportunities. *Journal of Agriculture and Food Research*, *9*, 100353. https://doi.org/10.1016/j.jafr.2022.100353

Du, Z., Tian, W., Tilley, M., Wang, D., Zhang, G., & Li, Y. (2022). Quantitative assessment of wheat quality using near-infrared spectroscopy: A comprehensive review. *Comprehensive Reviews in Food Science and Food Safety*, *21*(3), 2956–3009. https://doi.org/10.1111/1541-4337.12958

Du, Z., Wang, D., & Li, Y. (2022). Comprehensive Evaluation and Comparison of Machine Learning Methods in QSAR Modeling of Antioxidant Tripeptides. *ACS Omega*, acsomega.2c03062. https://doi.org/10.1021/acsomega.2c03062

Duffuler, P., Bhullar, K. S., de Campos Zani, S. C., & Wu, J. (2022). Bioactive Peptides: From Basic Research to Clinical Trials and Commercialization. *Journal of Agricultural and Food Chemistry*, *70*(12), 3585–3595. https://doi.org/10.1021/acs.jafc.1c06289

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., & Rost, B. (2022). ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381

FitzGerald, R. J., Cermeño, M., Khalesi, M., Kleekayai, T., & Amigo-Benavent, M. (2020). Application of in silico approaches for the generation of milk protein-derived bioactive peptides. *Journal of Functional Foods*, *64*, 103636. https://doi.org/10.1016/j.jff.2019.103636

Iwaniak, A., Darewicz, M., Mogut, D., & Minkiewicz, P. (2019). Elucidation of the role of in silico methodologies in approaches to studying bioactive peptides derived from foods. *Journal of Functional Foods*, *61*, 103486. https://doi.org/10.1016/j.jff.2019.103486

Kalyan, G., Junghare, V., Khan, M. F., Pal, S., Bhattacharya, S., Guha, S., Majumder, K., Chakrabarty, S., & Hazra, S. (2021). Anti-hypertensive Peptide Predictor: A Machine Learning-Empowered Web Server for Prediction of Food-Derived Peptides with Potential Angiotensin-Converting Enzyme-I Inhibitory Activity. *Journal of Agricultural and Food Chemistry*, *69*(49), 14995–15004. https://doi.org/10.1021/acs.jafc.1c04555

Kumar, R., Chaudhary, K., Sharma, M., Nagpal, G., Chauhan, J. S., Singh, S., Gautam, A., & Raghava, G. P. S. (2015). AHTPDB: A comprehensive platform for analysis and presentation of antihypertensive peptides. *Nucleic Acids Research*, *43*(Database issue), D956-962. https://doi.org/10.1093/nar/gku1141

Lemaıtre, G., & Nogueira, F. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, *18*, 1–5.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2020). Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 318–327. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2018.2858826

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, *379*(6637), 1123–1130. https://doi.org/10.1126/science.ade2574

Maaten, L. van der, & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605.

Manavalan, B., Basith, S., Shin, T. H., Wei, L., & 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. https://doi.org/10.1093/bioinformatics/bty1047

McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *bioRxiv*. https://doi.org/10.48550/arXiv.1802.03426

Minkiewicz, Iwaniak, & Darewicz. (2019). BIOPEP-UWM Database of Bioactive Peptides: Current Opportunities. *International Journal of Molecular Sciences*, *20*(23), 5978. https://doi.org/10/gnmt2w

Olsen, T. H., Yesiltas, B., Marin, F. I., Pertseva, M., García-Moreno, P. J., Gregersen, S., Overgaard, M. T., Jacobsen, C., Lund, O., Hansen, E. B., & Marcatili, P. (2020). AnOxPePred: Using deep learning for the prediction of antioxidative properties of peptides. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-78319-w

Pang, Y., Yao, L., Xu, J., Wang, Z., & Lee, T.-Y. (2022). Integrating transformer and imbalanced multi-label learning to identify antimicrobial peptides and their functional activities. *Bioinformatics*, *38*(24), 5368–5374. https://doi.org/10.1093/bioinformatics/btac711

Perez Espitia, P. J., de Fátima Ferreira Soares, N., dos Reis Coimbra, J. S., de Andrade, N. J., Souza Cruz, R., & Alves Medeiros, E. A. (2012). Bioactive Peptides: Synthesis, Properties, and Applications in the Packaging and Preservation of Food. *Comprehensive Reviews in Food Science and Food Safety*, *11*(2), 187–204. https://doi.org/10.1111/j.1541-4337.2011.00179.x

Pinacho-Castellanos, S. A., García-Jacas, C. R., Gilson, M. K., & Brizuela, C. A. (2021). Alignment-Free Antimicrobial Peptide Predictors: Improving Performance by a

Thorough Analysis of the Largest Available Data Set. *Journal of Chemical Information and Modeling*, *61*(6), 3141–3157. https://doi.org/10.1021/acs.jcim.1c00251

Qin, D., Bo, W., Zheng, X., Hao, Y., Li, B., Zheng, J., & Liang, G. (2022). DFBP: A comprehensive database of food-derived bioactive peptides for peptidomics research. *Bioinformatics*, *38*(12), 3275–3280. https://doi.org/10.1093/bioinformatics/btac323

Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., & Rives, A. (2020). Transformer protein language models are unsupervised structure learners. *bioRxiv*, 2020.12.15.422761. https://doi.org/10.1101/2020.12.15.422761

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, *118*(15), e2016239118. https://doi.org/10.1073/pnas.2016239118

Tammina, S. (2019). Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images. *International Journal of Scientific and Research Publications (IJSRP)*, *9*(10), p9420. https://doi.org/10.29322/IJSRP.9.10.2019.p9420

Tu, M., Cheng, S., Lu, W., & Du, M. (2018). Advancement and prospects of bioinformatics analysis for studying bioactive peptides from food-derived protein: Sequence, structure, and functions. *TrAC Trends in Analytical Chemistry*, *105*, 7–17. https://doi.org/10.1016/j.trac.2018.04.005

Ulug, S. K., Jahandideh, F., & Wu, J. (2021). Novel technologies for the production of bioactive peptides. *Trends in Food Science & Technology*, *108*, 27–39. https://doi.org/10.1016/j.tifs.2020.12.002

Veltri, D., Kamath, U., & Shehu, A. (2018). Deep learning improves antimicrobial peptide recognition. *Bioinformatics*, *34*(16), 2740–2747. https://doi.org/10.1093/bioinformatics/bty179

Waghu, F. H., Barai, R. S., Gurung, P., & Idicula-Thomas, S. (2016). CAMPR3: A database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Research*, *44*(D1), D1094-1097. https://doi.org/10.1093/nar/gkv1051

Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., Kingsbury, P., & Liu, H. (2018). A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, *87*, 12–20. https://doi.org/10.1016/j.jbi.2018.09.008

Wei, L., Hu, J., Li, F., Song, J., Su, R., & Zou, Q. (2020). Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. *Briefings in Bioinformatics*, *21*(1), 106–119. https://doi.org/10.1093/bib/bby107

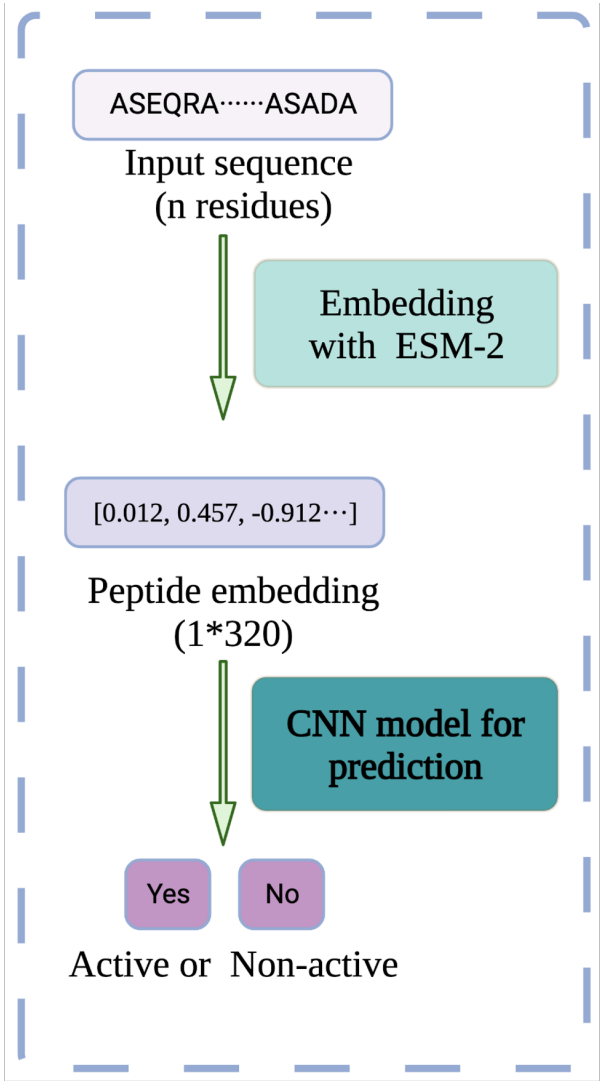
Wei, L., Ye, X., Xue, Y., Sakurai, T., & Wei, L. (2021). ATSE: A peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. *Briefings in Bioinformatics*, *22*(5), bbab041. https://doi.org/10.1093/bib/bbab041

Wen, C., Zhang, J., Zhang, H., Duan, Y., & Ma, H. (2020). Plant protein-derived antioxidant peptides: Isolation, identification, mechanism of action and application in food systems: A review. *Trends in Food Science & Technology*, *105*, 308–322. https://doi.org/10.1016/j.tifs.2020.09.019

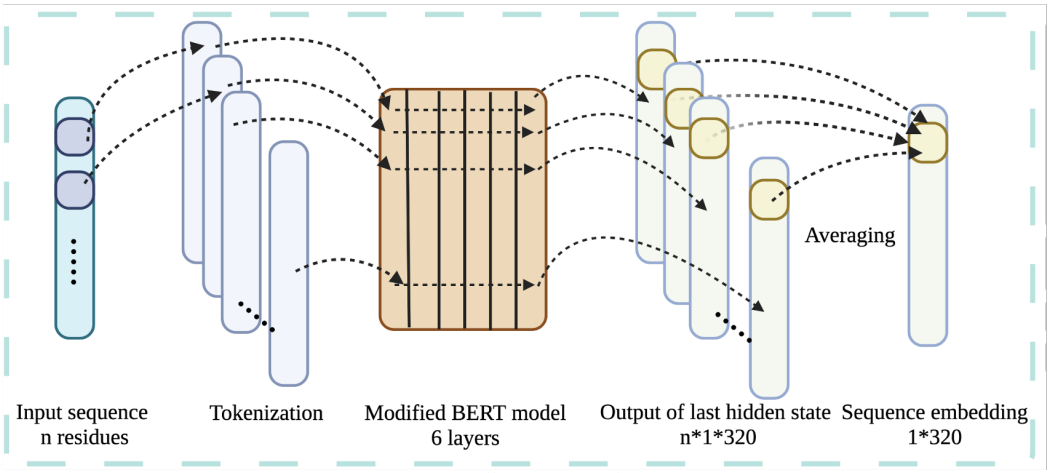
Yang, Z., Chen, Y., & Corander, J. (2021). T-SNE Is Not Optimized to Reveal Clusters in Data. *arXiv*. http://arxiv.org/abs/2110.02573

Zhang, W., Xia, E., Dai, R., Tang, W., Bin, Y., & Xia, J. (2022). PredAPP: Predicting Anti-Parasitic Peptides with Undersampling and Ensemble Approaches. *Interdisciplinary Sciences: Computational Life Sciences*, *14*(1), 258–268. https://doi.org/10.1007/s12539-021-00484-x

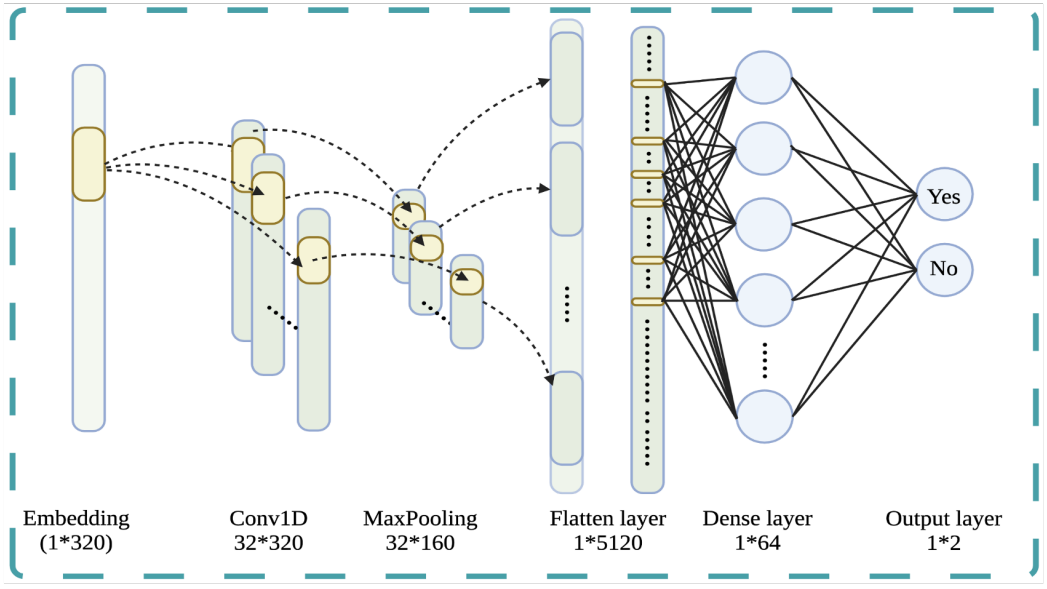
Figures and Tables:



UniDL4BioPep architecture



ESM-2 architecture



CNN architecture

Figure 4-1 Schematic framework of UniDL4BioPep for peptide bioactivity prediction by integrating ESM-2 and convolutional neural network (CNN).

Note: Peptide of any length is encoded as 320 dimensions embedding by ESM-2, and the embedding is fed into the CNN model. The first layer has 32 filters, and each of these filters undergoes 1D convolution with kernel size 3, stride 1 and ReLU activation function and further down-sampled via a max pooling layer with stride 2. The output 32*160 feature matrix is transformed as a 1D vector in the flatten layer and loaded into a dense layer (fully-connected layer) with one hidden layer (64 neurons and ReLU as activation function). The final output layer has two neurons with SoftMax function for prediction.

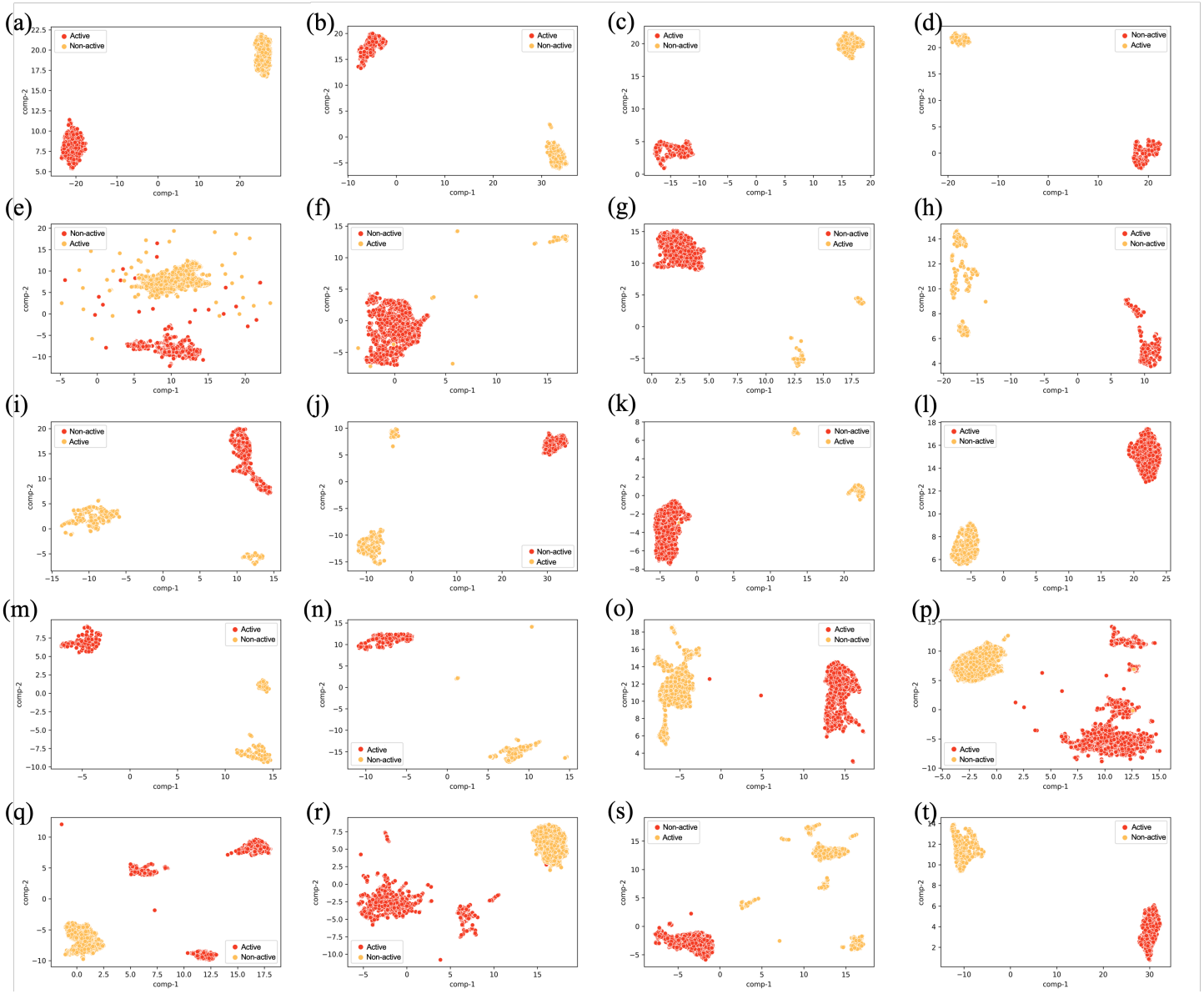

(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
(l)
(m)
(n)
(o)
(p)
(q)
(r)
(s)
(t)
Active
Non-active
comp-1
comp-2

Figure 4-2 Uniform manifold approximation and projection (UMAP) of positive and negative samples in different bioactive peptide datasets.

Note: (a) ACE inhibitory activity, (b) DPP IV inhibitory activity, (c) Bitter, (d) Umami, (e) Antimicrobial activity, (f) Antimalarial activity (main dataset), (g) Antimalarial activity (alternative dataset), (h) Quorum sensing activity, (i) Anticancer activity (main dataset), (j) Anticancer activity (alternative dataset), (k) Anti-MRSA strains activity, (l) Tumor T cell antigens, (m) blood-brain barrier, (n) anti-parasitic activity, (o) Neuropeptide; (p)Antibacterial activity, (q) antifungal activity, (r) antiviral activity, (s) toxicity, and (t) antioxidant activity. Graphs were generated using the whole dataset with n_neighbors parameters = 20.

Table 4-1 Benchmark datasets collected from publications with state-of-the-art models.

Bioactivity	Training dataset	Test dataset	Reference
ACE inhibitory activity	913 positives and 913 negatives	386 positives and 386 negatives	(Manavalan et al., 2019)
DPP IV inhibitory activity	532 positives and 532 negatives	133 positives and 133 negatives	(Charoenkwan, Kanthawong, et al., 2020)
Bitter	256 positives and 256 negatives	64 positives and 64 negatives	(Charoenkwan, Nantasenamat, Hasan, Manavalan, et al., 2021)
Umami	112 positives and 241 negatives	28 positives and 61 negatives	(Charoenkwan, Yana, Nantasenamat, et al., 2020)
Antimicrobial activity	3876 positives and 9552 negatives	2584 positives and 6369 negatives	(Pang et al., 2022)
Antimalarial activity	Main dataset (111 positives and 1708 negatives); alternative dataset (111 positives and 542 negatives)	Main dataset (28 positives and 427 negatives); alternative dataset (28 positives and 135 negatives)	(Charoenkwan, Schaduangrat, et al., 2022)
Quorum sensing activity	200 positives and 200 negatives	20 positives and 20 negatives	(Wei et al., 2020)
Anticancer activity	Main dataset (689 positives and 689 negatives); alternative dataset (776 positives and 776 negatives)	Main dataset (172 positives and 172 negatives); alternative dataset (194 positives and 194 negatives)	(Agrawal et al., 2021; Charoenkwan, Chiangjong, et al., 2021)
Anti-MRSA strains activity	118 positives and 678 negatives	30 positives and 169 negatives	(Charoenkwan, Kanthawong, et al., 2022)
Tumor T cell antigens	470 positives and 318 negatives	122 positives and 75 negatives	(Charoenkwan, Nantasenamat, et al., 2020)
Blood-Brain Barrier	100 positives and 100 negatives	19 positives and 19 negatives	(Dai et al., 2021)
Anti-parasitic activity	255 positives and 255 negatives	46 positives and 46 negatives	(Zhang et al., 2022)
Neuropeptide	1940 positives and 1940 negatives	485 positives and 485 negatives	(Bin et al., 2020; S. Chen et al., 2022)
Antibacterial activity	6583 positives and 6583 negatives	1695 positives and 1695 negatives	(Pinacho-Castellanos et al., 2021)

Antifungal activity	778 positives and 778 negatives	215 positives and 215 negatives	(Pinacho-Castellanos et al., 2021)
Antiviral activity	2321 positives and 2321 negatives	623 positives and 623 negatives	(Pinacho-Castellanos et al., 2021)
Toxicity	1642 positives and 1642 negatives	290 positives and 290 negatives	(Wei et al., 2021)
Antioxidant activity	582 positives and 541 negatives	146 positives and 135 negatives	(Olsen et al., 2020)

Table 4-2 Comparison between the performance of UniDL4BioPep and state-of-the-art models from the same benchmark datasets

Bioactivity*	Model name	ACC	BACC	Sn	Sp	MCC	AUC
ACE inhibitory activity	UniDL4BioPep	0.851	0.851	0.836	0.852	0.703	0.901
	mAHTPred(Manavalan et al., 2019)	0.883	N/A	0.894	0.873	0.767	0.951
DPP IV inhibitory activity	UniDL4BioPep	**0.853**	**0.853**	**0.861**	**0.846**	**0.707**	**0.938**
	iDPPIV-SCM(Charoenkwan, Kanthawong, et al., 2020)	0.797	N/A	0.789	0.805	0.594	0.847
Bitter	UniDL4BioPep	**0.938**	**0.938**	**0.924**	**0.952**	**0.875**	**0.982**
	BERT4Bitter(Charoenkwan, Nantasenamat, Hasan, Manavalan, et al., 2021)	0.922	N/A	0.938	0.906	0.844	0.964
	iBitter-Fuse[54]	0.93	N/A	0.938	0.922	0.859	0.933
Umami	UniDL4BioPep	**0.888**	**0.875**	**0.846**	**0.905**	**0.735**	**0.948**
	UniDL4BioPep-FL (gamma=3) **	0.888	0.883	0.892	0.875	0.733	0.943
	iUmami-SCM(Charoenkwan, Yana, Nantasenamat, et al., 2020)	0.865	N/A	0.714	0.934	0.679	0.898
Antimicrobial activity	UniDL4BioPep	0.962	0.958	0.968	0.948	0.908	0.991
	UniDL4BioPep-FL (gamma=3.5) **	0.96	0.961	0.96	0.963	0.903	0.991
	TransImbAMP(Pang et al., 2022)	N/A	0.969	0.963	0.974	N/A	N/A
	UniDL4BioPep	**0.98**	**0.989**	**1**	**0.979**	**0.815**	**0.921**

Antimalarial activity (main dataset)	UniDL4BioPep-FL (gamma=4)**	0.978	0.965	0.979	0.95	0.793	0.898
	iAMAP-SCM(Charoenkwan, Schaduangrat, et al., 2022)	0.978	0.826	0.654	0.998	0.776	0.82
Antimalarial activity (alternative dataset)	UniDL4BioPep	**0.975**	**0.97**	**0.962**	**0.978**	**0.912**	**0.987**
	UniDL4BioPep-FL (gamma=1)**	0.989	0.993	0.985	1.0	0.9570	0.987
	iAMAP-SCM(Charoenkwan, Schaduangrat, et al., 2022)	0.957	0.896	0.808	0.985	0.834	0.903
Quorum sensing activity	UniDL4BioPep	**0.95**	**0.955**	**0.909**	**1**	**0.905**	**0.99**
	iQSP[55]	0.93	N/A	N/A	N/A	0.86	0.96
	QSPred-FL(Wei et al., 2020)	0.943	N/A	0.935	0.95	0.885	0.945
Anticancer activity (main dataset)	UniDL4BioPep	0.735	0.735	0.734	0.737	0.471	0.805
	iACP-FSCM[56]	0.825	0.825	0.726	0.903	0.646	0.812
	AntiCP 2.0(Agrawal et al., 2021)	0.754	0.754	0.774	0.734	0.51	N/A
Anticancer activity (alternative dataset)	UniDL4BioPep	**0.946**	**0.948**	**0.978**	**0.918**	**0.894**	**0.971**
	iACP-FSCM[56]	0.889	N/A	0.876	0.902	0.779	0.93
	AntiCP 2.0(Agrawal et al., 2021)	0.92	N/A	0.923	0.918	0.84	N/A
Anti-MRSA strains activity	UniDL4BioPep	**0.99**	**0.994**	**1**	**0.988**	**0.96**	**0.999**
	UniDL4BioPep-FL (gamma=3)**	0.994	0.997	0.994	1	0.98	0.999
	SCMRSA(Charoenkwan, Kanthawong, et al., 2022)	0.96	0.935	0.9	0.97	0.848	0.986
Tumor T cell antigens	UniDL4BioPep	**0.751**	**0.743**	**0.763**	**0.724**	**0.457**	**0.788**

	UniDL4BioPep-FL (gamma=2)**	0.746	0.762	0.734	0.791	0.446	0.796
	iTTCA-Hybrid(Charoenkwan, Nantasenamat, et al., 2020)	0.71	N/A	0.844	0.493	0.363	0.756
Blood-Brain Barrier	UniDL4BioPep	**0.842**	**0.846**	**0.882**	**0.809**	**0.688**	**0.992**
	BBPpred(Dai et al., 2021)	0.7895	N/A	0.6316	0.9474	0.6102	0.7895
Anti-parasitic activity	UniDL4BioPep	**0.891**	**0.911**	0.821	**1**	**0.8**	**0.94**
	PredAPP(Zhang et al., 2022)	0.88	N/A	0.978	0.783	0.775	0.922
Neuropeptide	UniDL4BioPep	0.892	0.892	0.875	0.909	0.784	0.953
	PredNeuroP(Bin et al., 2020)	0.897	N/A	0.886	0.907	0.794	0.954
	NeuroPred-CLQ(S. Chen et al., 2022)	0.936	N/A	0.897	0.975	0.875	0.988
Antibacterial activity	UniDL4BioPep	**0.94**	**0.941**	**0.966**	0.915	**0.881**	**0.978**
	ABPDiscover(Pinacho-Castellanos et al., 2021)	0.935	N/A	0.912	0.957	0.87	0.975
Antifungal activity	UniDL4BioPep	**0.948**	**0.952**	**1**	0.904	**0.902**	**0.994**
	ABPDiscover (Pinacho-Castellanos et al., 2021)	0.942	N/A	0.921	0.963	0.884	0.988
Antiviral activity	UniDL4BioPep	**0.842**	**0.853**	**0.916**	**0.79**	**0.694**	**0.907**
	ABPDiscover(Pinacho-Castellanos et al., 2021)	0.828	N/A	0.764	0.892	0.662	0.896
Toxicity	UniDL4BioPep	0.941	0.941	0.923	0.959	0.883	0.978
	ATSE(Wei et al., 2021)	0.952	N/A	0.965	0.94	0.903	0.976
Antioxidant activity	UniDL4BioPep	**0.804**	**0.804**	**0.81**	**0.799**	**0.608**	**0.872**

	AnOxPePred-FRS (Olsen et al., 2020)	N/A	N/A	N/A	N/A	0.48	0.79

Note:

* The corresponding datasets are summarized in Table 1; ** Only gamma is specified for UniDL4BioPep-FL; Performance parameters are marked in bold when UniDL4BioPep achieved better performance; N/A: not available in the original paper.

Abbreviations: ACC: Accuracy; AUC: area under the curve; BACC: balanced accuracy; Sn: sensitivity; Sp: specificity; MCC: Matthews correlation coefficient.

Table 4-3 Comparison and overview of UniDL4BioPep and the state-of-the-art models for binary classification in peptide bioactivity.

Bioactivity*	Model name	Peptide representation**	Feature selection methods	Model development***	References
Various bioactivities, universal	UniDL4BioPep	Language model, evolutionary scale modeling (ESM-2, esm2_t6_8M_UR50D with 320 fixed length dimension output for any length peptide input)	None	CNN model with six layers	This work
ACE inhibitory activity	mAHTPred	AAC, AAI, BPF, CTD, DPC, OVP, Hybrid features, BPF.	Feature importance scores and sequential forward search (SFS) based on RF	An ensemble model with ERT, RF, SVM, GB	(Manavalan et al., 2019)
DPP IV inhibitory activity	iDPPIV-SCM	AAC, DPC	None	SCM	(Charoenkwan, Kanthawong, et al., 2020)
Bitter	iBitter-Fuse	AAC, DPC, PAAC, APAAC, AAI	Genetic algorithm utilizing self-assessment-report	SVM	[54]
Bitter	BERT4Bitter	TFIDF, Pep2Vec, FastText	None	CNN, LSTM neural network, **BERT-based BiLSTM**	(Charoenkwan, Nantasenamat, Hasan, Manavalan, et al., 2021)
Umami	iUmami-SCM	DPC	None	DT, kNN, MLP, NB, SVM, and RF, **SCM**	(Charoenkwan, Yana,

					Nantasenamat, et al., 2020)
Antimicrobial activity	TransImbAMP	Embedded by a self-designed language model	None	MLP	(Pang et al., 2022)
Antimalarial activity (main dataset and alternative dataset)	iAMAP-SCM	AAC, **DPC**, TPC, CTD, CTDC, AAI, CTD, CTDD, CTDC, DPS	None	kNN, MLP, SVM, and RF, **SCM**	(Charoenkwan, Schaduangrat, et al., 2022)
Quorum sensing activity	QSPred-FL	AAC, CTD, 188-bit features, GDC, ASDC, IT, BPF, N- and C-terminus approach, 21-bit features, OVP	Minimum redundancy feature selection and SFS based on RF	**RF**, NB, SVM, LR, and J48	(Wei et al., 2020)
Quorum sensing activity	iQSP	AAI	Genetic algorithm utilizing self-assessment-report	SVM	[55]
Anticancer activity (main dataset and alternative dataset)	iACP-FSCM	AAC, **DPC**, composition on terminal region	None	Flexible SCM	[56]
Anticancer activity (main dataset and alternative dataset)	AntiCP 2.0	AAC, **DPC**, BPF, split composition	None	ANN, SVM, KNN, **ETree**, RF, Ridge classifier	(Agrawal et al., 2021)
Anti-MRSA strains activity	SCMRSA	AAC, AAI, APS, **DPC**, CTD	None	DT, KNN, LR, NB, PLS, SVM, **SCM**	(Charoenkwan, Kanthawong, et al., 2022)

Tumor T cell antigens	iTTCA-Hybrid	AAC, DPC, **PACC**, **CTDD**, AAI, hybrid feature	None	SVM, **RF**	(Charoenkwan, Nantasenamat, et al., 2020)
Anti-parasitic	PredAPP	AAC, AAI, CT5, CTD, DPC, GAAC, GDPC, NT5 and PAAC	Nine feature was used to build model for prediction with six modeling methods (KNN, LR, MLP, RF, SVM and XGB). The output of the best model for each feature was used as final selected feature.	KNN, **LR**, MLP, RF, SVM and XGB	(Zhang et al., 2022)
Blood-Brain Barrier	BPPred	AAC, CTD, BIT12, BIT118, GDPC, GTPC, GGAP, BIT21, IT, OLP, DPC, ASDC, GGAP, PAAC, CTF, BIT20, QSO	Combination of F1 score obtained from ERT, Spearman correlation coefficient, and SFS based selection relying on ERT, XGB, KNN, LR, MLP, RF, and SVM	ERT, XGB, KNN, **LR**, MLP, RF, and SVM	(Dai et al., 2021)
Neuropeptide	PredNeuroP	AAC, DPC, BPF, AAE, AAAI, GAAC, GDPC, GTPC, CTD	Five classifiers (ANN, ERT, XGB, KNN, and LR) were used to build 45 base-learners based on the nine features. Eight of them were selected relying on Pearson correlation coefficients and accuracy.	A stacking model (eight optimal base-learner as the first layer and LR as the second layer)	(Bin et al., 2020)

Neuropeptide	NeuroPred-CLQ	Word2Vec for sequence embeddings	None	CNN, TCNN and a multi-head attention layer were connected sequentially.	(S. Chen et al., 2022)
Antibacterial activity/ Antifungal activity/ Antiviral activity	ABPDiscover ****	Generated by ProtDCal	Six feature selection methods (correlation subset, Relief-F, information gain, gain ratio, and symmetrical uncertainty) were used separately and jointly to generated subset features. Wrapper feature selection using genetic algorithm with RF as learning method.	RF	(Pinacho-Castellanos et al., 2021)
Toxicity	ATSE	molecular graphs and position-specific scoring matrix (PSSM) of sequences	None	GNN and CNN_BiLSTM were used to extract information from molecular graphs and PSSM, respectively. Their output was flattened and loaded into a attention module and output module (fully connected layer).	(Wei et al., 2021)
Antioxidant activity	AnOxPePred	BF	None	CNN	(Olsen et al., 2020)

Note:

[*] The corresponding datasets are summarized in Table 1; [**] and [***]: Feature names and model methods in bold indicate that they were combined to achieve the state-of-the-art performance in Table 3, while in studies with feature selection approach relying on the entire feature pool, all the features are not marked in bold. [****]: For each bioactivity, feature selection needs to be redone.

Abbreviation in peptide representation: AAC: Amino acid composition; AAE: Amino acid entropy; AAI: Amino acid index database; BPF: Binary profiles (one hot encoding); APAAC: Amphiphilic pseudo-amino acid composition; ASDC: integration of the DPC and GDC; BIT12: combination of SAAC, GAAC, molecular weight, and peptide length; BIT 118: combination of AAC and CTD; BIT20: same as BPF; BIT 21: same as TOB; CTF: conjoint triad feature; CTD: composition-transition-distribution; CTDC: composition-transition-distribution composition; CTDD: composition-transition-distribution distribution; CTDT: composition-transition-distribution transition; CKSAAGP: composition of k-spaced amino acid group pairs; DPC: Dipeptide composition; APS: amino acid propensities; DPS: propensities of 400 dipeptides; Estate: Electrotopological state atom types; FP4: Presence of SMARTS patterns for functional groups; GAAC: grouped amino acid composition; GDC: the correlation of nonadjacent residue pairs; GDPC: grouped dipeptide composition; GTPC: grouped tripeptide composition; GGAP: g-gap dipeptide composition; IT: the input peptide is encoded as 3-dimensional vector (Shannon Entropy, Relative Shannon Entropy and information gain score); MACCS: Binary representation of chemical features defined by MACCS key; OVP: Overlapping property features; OF: other features (absolute charge per residue, aliphatic index, a fraction of transforming residue, molecular weight, and sequence length); OLP: overlapping property feature; PAAC: Pseudo amino acid composition; Pubchem: Binary representation of substructures defined by PubChem; QSO: quasi-sequence

order feature; SAAC: selected amino acid composition ; TFIDF: term frequency-inverse document frequency; TOB: Twenty-one-bit features; TPC: tripeptide composition

Abbreviation in model development:

ANN: artificial neural network; BERT: bidirectional encoder representations from transformers; BiLSTM: bidirectional long-short term memory; CNN: convolutional neural network; DT: decision tree; ERT: extremely randomized tree; ETree: extra trees; GB: gradient boosting; GNN: graph neural networks; kNN: k nearest neighbors; LR: logistic regression; LSTM: long-short term memory; MLP: multilayer perceptron; NB, naïve Bayes, PLS: partial least squares regression; RF: random forest; SCM: scoring card method; SVM: support vector machine; TCNN: temporal convolutional neural network; XGB: eXtreme gradient boosting

Chapter 5 - pLM4ACE: a protein language model based predictor for antihypertensive peptide screening*

Abstract

Angiotensin-I converting enzyme (ACE) regulates the renin-angiotensin system and is a drug target in clinical treatment for hypertension. This study aims to develop a protein language model (pLM) with evolutionary scale modeling (ESM-2) embeddings that is trained on experimental data to screen peptides with strong ACE inhibitory activity. Twelve conventional peptide embedding approaches and five machine learning (ML) modeling methods were also tested for performance comparison. Among the 65 classifiers tested, logistic regression with ESM-2 embeddings showed the best performance, with balanced accuracy (BACC), Matthews correlation coefficient (MCC), and area under the curve of 0.883±0.017, 0.77±0.032, and 0.96±0.009, respectively. Multilayer perceptron and support vector machine also exhibited great compatibility with ESM-2 embeddings. These models utilize simpler ML methods with pLM but resulted in better prediction performance compared to more complicated models. A user-friendly webserver (https://sqzujiduce.us-east-1.awsapprunner.com) with the top three models is now freely available.

Keywords: ACE inhibitory peptide; antihypertension; bioactive peptide; protein language model; machine learning

---

* Du, Z., Ding, X., Hsu, W., Munir, A., Xu, Y., & Li, Y. (2024). pLM4ACE: A protein language model based predictor for antihypertensive peptide screening. Food Chemistry, 431, 137162. https://doi.org/10.1016/j.foodchem.2023.137162

5.1 Introduction

Hypertension affects approximately 1 billion people worldwide and is a critical risk factor for cardiovascular diseases (Majumder & Wu, 2014; Mudgil et al., 2019). One of the most common reasons for hypertension is the disorder of the renin-angiotensin system (RAS), which involves the renin-mediated conversion of angiotensinogen to angiotensin I and finally to angiotensin II, resulting in vasoconstriction and elevated blood pressure (Aluko, 2015). Angiotensin-I converting enzyme (ACE) (EC 3.4.15.1), a dipeptidyl carboxypeptidase from the zinc proteases class, plays a vital role in the RAS and the conversion of angiotensin I to angiotensin II. It also participates in the kallikrein-kinin system (KKS) where it inactivates bradykinin (Mudgil et al., 2019). Therefore, it is a promising drug target for the treatment of hypertension. Since the first ACE inhibitor, captopril, was synthesized in 1977, a large number of ACE inhibitors have been identified and widely used in the clinical treatment for hypertension (F. Wang & Zhou, 2020). Recently, natural compound based alternatives, particularly bioactive peptides, have attracted more attention because of the increasing health concern about the side effects of synthesized drugs (Aluko, 2015; Du & Li, 2022b).

At this time, more than 1,000 ACE inhibitory peptides have been identified with inhibitory activity reported from *in vitro* or *in vivo* experiments (Minkiewicz et al., 2019). However, most of them were screened and identified by conventional chemistry approaches, where bioactive peptides were isolated from protein hydrolysates or fermented products or directly obtained through chemical synthesis (Du & Li, 2022b; Majumder & Wu, 2014). Given the high cost, low efficiency, and reliance on advanced equipment and experienced personnel in wet experiments, researchers are turning to bioinformatics to reverse the traditional workflow and guide efficient peptide screening. Various regression models and classification models have been reported for

bioactive peptide prediction (Bin et al., 2020; Du, Wang, et al., 2022; FitzGerald et al., 2020; Kalyan et al., 2021). During model development, there are two critical steps that can hinder model performance, namely model selection and peptide representation (Du, Comer, et al., 2023). The accumulated dataset size of ACE inhibitory peptides makes it possible to employ deep learning for bioactivity prediction, which too some extent reduce the effort in model selection and has demonstrated great potential in accuracy (Olsen et al., 2020). However, current peptide representation mostly relies on the properties of single amino acid or amino acid composition, or the estimation of overall physicochemical properties (Bin et al., 2020; Du, Ding, et al., 2023; Du & Li, 2022a; Kalyan et al., 2021; Lertampaiporn et al., 2022; Olsen et al., 2020).

Natural language processing (NLP) is a branch of artificial intelligence (AI) that enables machines to understand natural language. Vector representation of words (i.e., word embeddings) is the critical point affecting the final model performance since it makes text data understandable to machine learning (ML) models (Santos et al., 2017; Sharma et al., 2017). Using millions of available protein sequences, bioinformatic researchers have employed self-supervised learning approaches (e.g., BERT (bidirectional encoder representations from transformers)) for protein language model (pLM) development. Several pLMs have been released and achieved great performance in downstream predicting tasks (e.g., protein structure, functionality, subcellular location, etc.) (Alley et al., 2019; Elnaggar et al., 2022, 2022; Lin et al., 2023; Lu et al., 2020; Rao et al., 2020). Evolutionary scale modeling (ESM-2) is the latest language model released in 2022 by the Fundamental AI Research (FAIR) Protein Team at Meta. ESM-2 is up to 60x faster than AlphaFold2 and lower template modeling score (TM-score) in CAMEO and CASP14 datasets (Lin et al., 2023). To our knowledge, prior to our study, no pLM with ESM-2 embeddings has been employed in peptide sequence representation for bioactive peptide prediction.

Since there are few peptides that are experimentally proven to be non-ACE inhibitory, in previous model development studies, the general solution is to extract random peptides from the UniProt knowledgebase (https://www.uniprot.org/). However, some of these extracted peptides may have ACE inhibitory activity that has not yet been tested, which would mislead subsequent model development (Kalyan et al., 2021; Lertampaiporn et al., 2022; Manavalan et al., 2019; Y.-T. Wang et al., 2020; Y. Zhang et al., 2023). In addition, there is no clear bioactivity threshold between positive and negative peptides in previous training datasets. The objective of this study was to build a pLM-based classification model to screen peptides with strong ACE inhibitory activity and to train the model entirely on experimental data with a clear bioactivity threshold for labeling. The workflow of this study is shown in Figure 1. Our main contributions in this work are as follows:

1). Manually curate the latest ACE inhibitory peptide dataset from 267 published papers reporting the half maximal inhibitory concentration (IC_{50}) and clean the data using CleanLab based on confident learning;

2). Develop a pLM-based classification model with ESM-2 embeddings that is entirely trained on experimental data with a clear bioactivity threshold to screen peptides with strong ACE inhibitory activity;

3). Compare 12 conventional peptide embedding approaches with the ESM-2-based embeddings in combination with five different ML modeling methods;

4). Experimentally reveal that our proposed logistic regression with ESM-2 embeddings attains an accuracy of 88.3% for the classification of ACE inhibitory peptides;

5). Deploy a user-friendly web server with the top three models supporting multiple upload file formats and large-scale prediction;

6). Provide a valuable reference for peptide discovery and potential application concerning other bioactivities.

5.2 Materials and methods

5.2.1 Dataset collection and curation

The ACE inhibitory peptide sequence items were initially collected from three different sources: BIOPEP-UWM (https://biochemia.uwm.edu.pl/biopep/peptide_data.php), AHTPDB (https://webs.iiitd.edu.in/raghava/ahtpdb/), and a published paper (Kalyan et al., 2021). Specifically, a total of 1,066 entries with sequences, references, and IC_{50} values were retrieved from BIOPEP-UWM using the keyword search "ACE inhibitor" under the activity category. For entries without IC_{50} values, we manually checked the original references and extracted the activity information if available. From AHTBPD, we obtained a dataset (peptide with IC_{50}) contianing 3,364 experimentally validated ACE inhibitory peptides. In Kalyan et al. (2021), a total of 1,648 ACE inhibnitory peptide sequences were summarized and made available for download.

In addition, we conducted a literature search using the Web of Sciencce platform with keywords "Angiotension converting enzyme inhibitory peptide" and "ACE inhibitory peptide" and manually reviewed the search results to identify newly synthesized or purified peptide sequencces with IC_{50} values for ACE inhibitory activity.

All the collected peptide entries were compiled, including their sequences, IC_{50} values, and original references. IC_{50} values were converted into μM if they were in a different unit. Duplicated entries, entries without references, and entries whose IC_{50} values were not reported in the original references were removed. In cases where different literature reported different IC_{50} values for the same peptide sequence, we retained the highest activity data if the references employed the same *in vitro* experiment protocol. For peptides whose inhibitory activities were reported as "non-

activity," the record was entered as such in our dataset, even though the peptides might not be truly inactive and might have activities that were too low and out of the determination range.

In total, we collected 1,385 unique peptide sequences from 267 different scientific papers, which are presented in Supplementary Document S1. The dataset was divided into two groups based on IC_{50}: a high activity (0-50 μM) group and a low & non-activity (>50 μM) group. The IC_{50} threshold was decided based on two criteria: making the dataset balanced and setting a relatively low IC_{50} value for high-activity peptide screening (Figure 2).

5.2.2 Dataset cleaning by CleanLab

CleanLab was utilized for the first time to clean noisy samples and generate a high-quality, brand-new dataset of ACE inhibitory peptides. CleanLab is a data-centric AI package that facilitates ML work with messy, real-world data by providing clean labels for robust training and error flagging. It can adapt to any existing scikit-learn classification model. The theoretical foundation of CleanLab is confident learning, which directly estimates the joint distribution between noisy (given) labels and true (unknown) labels and identifies the noisy data (Northcutt et al., 2021). It has been successfully used to find label errors in popular computer vision and NLP benchmark datasets (Northcutt et al., 2021). In this study, even though experimental data from different papers may share the same experimental protocol, there could still be errors in manual operations, experimental conditions, etc. Specifically, peptides with IC_{50} values close to the threshold were likely to be classified into the wrong class.

We combined CleanLab with logistic regression for noisy label detection and removal. During the cleaning process, peptide embeddings for logistic regression were generated by ESM-2. After cleaning, there were 1,020 samples remaining for further model development, of which 394 peptides were in the high activity group and 626 peptides were in the low & non-activity group

(Figure 2). The peptide sequences and corresponding labels are presented in Supplementary Document S1. The cleaned dataset was divided into a training dataset and a test dataset at a ratio of 8:2. Stratified sampling was used in dataset division based on the labels. To understand the effectiveness of the CleanLab in dataset cleaning for bioactivity prediction, uniform manifold approximation and projection (UMAP) was used for visualization in a two-dimensional graph (McInnes et al., 2020).

5.2.3 Peptide embeddings

ESM was a pLM project initiated by the Fundamental AI Research (FAIR) Protein Team at Meta in 2019. The latest, pre-trained pLM version is ESM-2, which was trained on the UR50/D 2021_04 dataset and achieved great performance in a range of structure prediction tasks (Lin et al., 2023). ESM-2 contains 6 pLMs varying from 48 layers and 15 billion parameters for 5,120 output embeddings to 6 layers and 8 million parameters for 320 output embeddings. The pLM (esm2_t6_8M_UR50D) with 320 output embeddings was selected in this study (Figure 1), because of the limited size of the ACE inhibitory peptide dataset and the potential curse of dimension. The pre-trained pLM and implementation procedures are available at https://github.com/facebookresearch/esm.

There is no available model for comparison with the ESM-2-based peptide embeddings. Hence, 12 other peptide embedding methods used in previous studies were also used on the cleaned dataset for comparison. The 12 methods are amino acid composition (AAC), dipeptide composition (DPC), pseudo amino acid composition (PseAAC), amino acid index (AAI), overlapping property features (OPF), binary profile/one-hot encoding (BP), adaptive skip dipeptide composition (ASDC), composition-transition-distribution composition (CTDC), composition-transition-distribution transition (CTDT), composition-transition-distribution

distribution (CTDD), grouped amino acid composition (GAAC), and grouped dipeptide composition (GDPC) (Chen et al., 2022; Lertampaiporn et al., 2022; Manavalan et al., 2019; L. Wang et al., 2021). All embeddings were generated by iFeatureOmega (https://github.com/Superzchen/iFeatureOmega-CLI) (Chen et al., 2022). Detailed explanations about the 12 methods are available at https://github.com/dzjxzyd/pLM4ACE.

5.2.4 Model development

5.2.4.1 Classification model selection

Five popular ML algorithms with different attributes for classification model development available at scikit-learn (https://scikit-learn.org/) were comparatively evaluated. The five algorithms are logistic regression (LR), random forest (RF), support vector machine (SVM), k-nearest neighboring (KNN), and multilayer perceptron (MLP) (Pedregosa et al., 2011). Logistic regression is the simplest binary classification algorithm, which assumes that there is a linear relationship between features, and the linear summation of each feature with different weights goes through a sigmoid function for classification (James et al., 2021). RF is an ensemble learning method that combines multiple decision trees trained on different subsets of the original datasets and features and aggregates their predictions (Du, Tian, et al., 2022). SVM finds the best-separating hyperplane in high-dimensional feature spaces based on different kernel functions (Du, Tian, et al., 2022; James et al., 2021). KNN operates by finding the k closest data points based on distance calculation between data points and giving prediction based on the majority class of its k nearest neighbors (James et al., 2021). MLP is a feedforward neural network that learns from labeled data by back-propagation of the loss and by tuning the weights and biases of its neurons (James et al., 2021).

5.2.4.2 Hyperparameter tuning and model performance evaluation

With 13 peptide embedding methods and five ML algorithms, a total of 65 combinations were used for model development. To tune the hyperparameters of each combination, 10-fold cross-validation (10-Fold-CV) was employed. The process involved splitting the training dataset into 10 subsets and testing the model ten times. In each iteration, nine subsets were used for training, and one subset was used for evaluation. To find the best hyperparameters, a grid search strategy was combined with 10-Fold-CV. The evaluation method used in 10-Fold-CV was balanced accuracy (BACC), which served as the criterion for hyperparameter selection. BACC was chosen to avoid models being misled by imbalanced data. The performance of the 65 models, together with their corresponding best hyperparameter settings that achieved the highest BACC, was listed in Supplementary Document S2 (Table S1-S13).

The 65 models were further evaluated on an independent test dataset. To avoid any bias from dataset splitting, each model was trained and evaluated ten times on both the training and test datasets. Random dataset splitting was employed to ensure the robustness of model performance. BACC, sensitivity (Sn), specificity (Sp), Matthews correlation coefficient (MCC), and the area under the curve (AUC) were adopted to evaluate model performance. Parameters were calculated based on the number of true positive (TP), false positive (FP), false negative (FN), and true negative (TN), according to the following equations:

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{5.1}$$

$$Sn = \frac{TP}{TP+FN} \tag{5.2}$$

$$Sp = \frac{TN}{TN+FP} \tag{5.3}$$

$$BACC = 0.5 * Sn + 0.5 * Sp \tag{5.4}$$

$$MCC = \frac{(TP*TN)-(FN*FP)}{\sqrt{(TP+FN)*(TN+FP)*(TP+FP)*(TN*FN)}} \quad (5.5)$$

The AUC is calculated using the sklearn 'roc_auc_score' function through the portability distribution of the model prediction in the test dataset.

5.2.4.3 Model implementation

The entire process, including peptide embedding, data cleaning, model development, model evaluation, and graph generation, was conducted on the Google Colab platform with Python 3.8, and the scripts are available at https://github.com/dzjxzyd/pLM4ACE.

5.2.5 Web server development

A user-friendly web server (available at https://sqzujiduce.us-east-1.awsapprunner.com) was deployed at Amazon Web Services (AWS) App Runner. The website was designed with html and css scripts, and the model deployment was achieved with Flask (2.2.2). Three models (LR, SVM, and MLP) with best performance in ACE inhibitory peptide prediction are deployed. The web server supports large-scale processing, which allows users to upload their peptide information through xls, xlsx, fasta, and txt formats for ACE inhibitory activity prediction. The detailed scripts for web server development are available at https://github.com/dzjxzyd/LM4ACE_webserver.

5.3 Results and discussion

5.3.1 Dataset cleaning

A high-quality dataset is crucial for model development and significantly influences model performance in practical applications. During cleaning, the dataset was divided into two groups based on IC_{50} values (i.e., high activity (0-50 μM) and low & non-activity (>50 μM)) and loaded into a logistic regression model for fitting. The predicted probability of each peptide sequence, combined with its label, was used as the input for noisy data cleaning. Therefore, peptides that were in the same group but with different IC_{50} values were treated based on their probability

prediction results. A 2 by 2 confident joint matrix was computed to partition and count label errors as well as to estimate the joint distribution matrix. Off-diagonal samples were considered noise labels and removed. Finally, a total of 1,020 data points were retained without any label issues detected by CleanLab (Northcutt et al., 2021). This cleaning process remarkably improved model performance compared to the model developed based on the original dataset in our preliminary experiments.

Figure 2(b) shows the changes in peptide length distribution before and after cleaning. It should be noted that CleanLab has been proven highly robust for imbalanced datasets. Given the two types of input (predicted probability and the original label) and the robustness of CleanLab, the variation in the number of peptide sequences removed from the two groups can primarily be attributed to original dataset quality. The selected modeling methods also accounted for data point removal. In our preliminary experiments, we tested CleanLab with other modeling methods (such as RF and SVM), and LR showed the best fitting performance during the modeling process, hence its selection.

UMAP is a general dimension reduction technique for visualization purposes. It prioritizes the preservation of local distances over global distances and has the ability to handle outliers and non-linear relationships, resulting in better performance than principal component analysis (PCA) in biological data processing (McInnes et al., 2020). Figure 3 (a) presents the visualization of peptide embeddings through ESM-2 prior to cleaning. The two groups are separated along the first component direction (comp-1), but not well separated along the second component direction (comp-2). After noise was removed by CleanLab, the remaining data points are well separated along both component directions (Figure 3 (b)). This further demonstrates the effectiveness of CleanLab for cleaning the ACE inhibitory peptide dataset. In addition, the UMAP results confirm

that ESM-2-based peptide embeddings exhibit great performance in peptide representation for ACE inhibitory activity prediction (Figure 3).

5.3.2 Peptide embedding analysis

The most challenging aspect of model development lies in peptide representation, and a key novelty of our study is the application of the latest and highly advanced protein language model, ESM-2. In this section, we elucidate the mechanism of ESM-2 for peptide embedding and compare it with other conventional peptide embedding approaches. ESM-2 is a modified version of the Bidirectional Encoder Representations from Transformers (BERT) architecture. The output of ESM-2 is the last hidden state of the ESM-2 model, which has been fine-tuned using millions of protein sequences (Figure 1). The dimensions of generated peptide embeddings typically vary based on the length of the input peptides, as each residue is represented by the same dimension vector, taking into account its neighboring residues and the whole sequence context. For example, ESM-2 generates three 1*320 dimension vectors for three residues in a tripeptide, while generating four 1*320 dimension vectors for a tetrapeptide (Elnaggar et al., 2022; Rives et al., 2021). The final output embeddings are obtained by averaging the representation of each residue, therefore allowing peptides of any length to be unified to the same length and loaded into the model for prediction tasks (Lin et al., 2023). As such, ESM-2-based peptide embeddings can be considered a global descriptor strategy, ensuring a fixed feature dimension regardless of the input peptide length.

Similarly, some conventional global descriptors (e.g., physicochemical properties, AAC, and DPC) also offer the advantage of fixed peptide embedding dimensions. Kalyan et al. (2021) proposed an ACE inhibitory peptide model based on physicochemical properties. In the model, the three-dimensional (3D) structure of the peptide was constructed to extract overall properties

(hydrophobicity, volume, charge, and molecular weight) for peptide representation in ACE inhibitory activity prediction (Kalyan et al., 2021). However, the 3D structure reconstruction was based on calculation and so were the overall physicochemical properties, which may introduce additional noise and uncertainty into the features used for peptide representation. As for composition-based global descriptors (e.g., AAC and DPC), the lack of sequential information can potentially limit model performance, and combining multiple composition-based descriptors may result in feature redundancy and increase the risk of the curse of dimensionality (Du, Comer, et al., 2023).

Local descriptors represent peptides at the residue level and play a significant role in peptide representation; however, they cannot be directly applied to peptides with different lengths for model development. For example, the AHTpin webserver utilized five different models with five different lengths for ACE inhibitory peptide prediction (Kumar et al., 2015). For longer peptides, the solution was to employ several residues' information at the N and C terminus for peptide representations. The research group later upgraded their AHTpin web server to the mAHTPred web server, which used conventional global descriptors for peptide representation and gained better performance by relying on a single model for peptides with different lengths (Manavalan et al., 2019). Although long short-term memory (LSTM) networks can handle sequential data with different lengths, embedding through global descriptors remains the mainstream approach for bioactive peptide classification tasks, while local descriptors are primarily used in regression model development (Du, Comer, et al., 2023). Overall, both conventional global descriptors and local descriptors face issues including the loss of sequential information, inadequate peptide descriptions, and the curse of dimensionality.

5.3.3 Model performance comparison on benchmark features

The representation power of different peptide descriptors can be assessed through their performance in model development. Five popular ML methods (LR, RF, SVM, KNN, and MLP) were employed to build ACE inhibitory activity prediction models. Thirteen peptide embedding strategies as previously described were used in the model building process. The cross-validation results of a total of 65 models (13 embedding strategies * 5 ML methods) are presented in Supplementary Document S2-Table S1-S13. Model performance was evaluated on the test dataset, with the model development process repeated ten times, and the results are shown in Figure 4 and Supplementary Document S2-Table S14-S26.

In this study, we conducted a comprehensive hyperparameter optimization process to identify the most suitable model types and corresponding hyperparameters for each peptide embedding method. The best combinations were as follows: ESM-LR (BACC=0.883 ± 0.017), AAC-RF (BACC= 0.744 ± 0.023), DPC-RF (BACC=0.768 ± 0.053), PseAAC-MLP (BACC=0.792 ± 0.035), AAI-SVM (BACC=0.685 ± 0.041), OFP-RF (BACC=0.696 ± 0.049), BP-MLP (BACC=0.777 ± 0.034), ASDC-RF (BACC=0.764 ± 0.034), CTDC-SVM (BACC=0.766 ± 0.019), CTDT-SVM (BACC=0.724 ± 0.033), CTDD-RF (BACC=0.775 ± 0.026), GAAC-RF (BACC=0.713 ± 0.028), and GDPC-KNN (BACC=0.718 ± 0.032). ESM-LR achieved the best performance, with a BACC that is 11.4% higher than that of the best model using a non-ESM embedding method (PseAAC-MLP). In addition, peptide embeddings methods, including PseAAC, AAC, DPC, ASDC, CTDC, CTDD, and CTDT also showed a feasible performance in prediction tasks and those individual models developed in this study (e.g., PseAAC-MLP, AAI-SVM, OFP-RF, etc.) have the potential to be integrated into an ensemble model, which can leverage the strength of each model and further enhance performance. This

concept was explored in a previous study, where six machine learning methods and 51 feature descriptors were combined to find the optimal combination, resulting in the selection of four models for the ensemble model construction (Manavalan et al., 2019).

Traditional descriptors for ACE inhibitory prediction have reached a bottleneck in performance improvement. Our study achieved a significant improvement in ACE inhibitory prediction accuracy, advancing it by approximately 6% compared to previous studies (Manavalan et al., 2019). Besides, although the previous model application scenario was limited to the positive and negative peptide classification task, our study can differentiate the high activity peptides. Out of the 65 models developed, three models, ESM-LR, ESM-SVM, and ESM-MLP, which were all based on ESM-2 embeddings, had BACC above 0.8. Table 1 shows the detailed performance, including AUC, of the top three models, and the standard deviation for each performance evaluator was very low. The three models' performance in terms of BACC and MCC was 7.95-11.49% and 65.60-67.63% higher than that of the fourth-highest performing model. It was observed that during model development, the penalties assigned to features in ESM-LR and ESM-SVM were quite low, indicating that no further pre-processing was needed for the ESM-2 embeddings (Supplementary Document S2-Table S1). These findings showed that EMS-2-based peptide embedding was superior to the other twelve embedding methods when using LR, SVM, and MLP model methods. The results were consistent with model performance on the test dataset (Figure 4).

Two models, ASDC-RF and DPC-RF, had comparable performance regarding Sn, but their overall prediction accuracy (BACC) was inferior to the three ESM-based models. There were no model performance data for BP-RF because, in several iterations of model development, the BACC of BP-RF was zero. It was assumed that BP-RF might not be capable of handling ACE inhibitory prediction. This may be because RF was not compatible with this type of feature, where

the peptide was represented by an extremely sparse feature matrix of BP features (e.g., for a dipeptide, only two elements are 1, and the remaining elements are all 0 in a vector with 600 elements).

The performance of RF and KNN models using ESM-2-based peptide embeddings did not reach the same level as that of the LR, SVM, and MLP models. This suggests that the fundamental theories of RF and KNN may not be suitable for peptide embeddings generated by ESM-2. Specifically, RF uses a subset of the entire 320 dimension features for multiple decision trees by bootstrapping, and the information contained in a single element in the 320 dimension vectors is meaningful only when the remaining 319 elements exist (James et al., 2021; Lin et al., 2023). This could mislead the learning processing of decision trees and cause inferior performance. As for the failure of KNN, it is mainly attributed to the unsuitability of the distance calculation method (Euclidean distance) for ESM-2-generated features. In this case, each element in the peptide embedding vectors independently contributes to the distance calculation, causing ineffective results.

pLM researchers were previously recommended to employ MLP for downstream tasks (Elnaggar et al., 2022; Rives et al., 2021). However, LR outperformed MLP in our study, possibly due to the limited dataset size. In LR model development, peptide embeddings are based on directed multiplication with a set of learnable parameters, which are then regularized using a penalty parameter. Then, a sigmoid function is applied for probability calculation, and the classification of the input sequence is determined based on this probability. In MLP model development, there are hidden layers with varying numbers of nodes with learnable parameters. An activation function extracts patterns and features from the peptide embeddings before feeding them into the output layer, which essentially functions as a logistic regression. All the learnable

parameters in both MLP and LR are tuned based on the cross-entropy loss function. However, due to the limited amount of available data, it is possible that the MLP models were not sufficiently tuned, resulting in inferior performance. Surprisingly, the SVM model was found to be effective for pLMs, which, to the best of our knowledge, has not been previously reported. Similar to MLP and LR, SVM simultaneously considers all elements in the peptide embedding vector for decision making, rather than treating them as separate features as in KNN or feature subsets. This characteristic enables SVM to fully leverage the potential of peptide embeddings generated by pLMs for classification model development. A recent study comparing embedding methods for protein phosphoglycerylation prediction also reported the remarkable compatibility of ESM models with SVM and LR, while the model based on RF showed inferior performance (Chandra et al., 2023).

5.3.4 Model performance comparison with state-of-the-art (SOTA) models

Many prediction models were previously developed with hybrid features or employed an ensemble classifier to achieve better performance (Bin et al., 2020; Charoenkwan et al., 2021; Dai et al., 2021; Lertampaiporn et al., 2022; Manavalan et al., 2019; W. Zhang et al., 2022). In the study by Manavalan et al. (2019), six ML approaches (adaboost, extremely randomized tree (ERT), gradient boosting (GB), KNN, RF, and SVM) were employed to build models using 51 different peptide embedding methods. The four models with the best prediction results were combined for final ACE inhibitory prediction, increasing model accuracy and MCC (based on AAC) from 0.800 to 0.883 and from 0.601 to 0.767, respectively. In another study conducted by Qin et al. (2022), six amino acid descriptors (local descriptors) were selected from twenty-two descriptors for peptide encoding, and an LSTM network was used to overcome feature dimension inconsistency in peptide embeddings (Qin et al., 2022). While the model achieved great performance in its own

dataset, its performance was inferior in the dataset of mAHTPred (Qin et al., 2022). In these studies, performance improvement was due mainly to increased complexity in model development (feature selection methods and modeling strategies). However, feature selection is a tedious and time-consuming trial-and-error process. It is difficult to exhaust all possible feature combinations in a single study. Therefore, in our model development, each ML model was combined with only a single type of peptide embedding approach as benchmark models for performance comparison. This is also where the benefit of ESM-2 peptide embedding lies. Even though the mechanism of ESM-2 is quite complex, once the pLM is available, no further pre-processing is needed before modeling. Therefore, the application of ESM-2-based methods can simplify model development.

Because the datasets used in previous studies differ, we cannot compare the numerical performance of the ESM-2-based models with that of SOTA models. However, model performance improvement reported in previous studies can be used for limited comparison. In the study of Manavalan et al. (2019), the mAHTPred model had 27.62% and 10.37% increase in MCC and accuracy, respectively, compared to the previous model (AHTpin_AAC). In our study, the ESM-LR model showed 14.08 – 25.78% and 50.09 – 86.44% increase in MCC and BACC, respectively, compared to the five AAC-based models (Figure 4). These results showed that performance improvements in ESM-LR, ESM-SVM, and ESM-MLP models were much greater than those of the two SOTA models. In addition, it is worth noting that the dataset used in this study was retrieved entirely from experimental data, making the developed models more attractive to researchers in the biochemistry fields.

5.4 Conclusions

Bioinformatics can significantly reduce the experiment time and cost of novel bioactive peptide exploration, and fast and accurate prediction models are highly desirable. In this study, we

introduced the latest pretrained language model for ACE inhibitory peptide embeddings and employed confident learning theory for bioactive peptide dataset cleaning. To our best knowledge, this is the first high ACE inhibitory activity peptide classification model that was built fully on experimental datasets. UMAP results confirmed the validity of confident learning for data cleaning and ESM-2 for peptide embedding. Five machine learning methods were employed to build models based on the peptide embeddings generated from ESM-2. Twelve conventional peptide embedding approaches combined with the same machine learning methods were also tested for performance comparison. Results showed that LR, SVM, and MLP performed very well with ESM-2-generated embeddings and achieved significantly higher model performance than other feature-based models, which demonstrated the superiority of ESM-2 for peptide representation. Model performance improvement compared to that of SOTA models supports this conclusion. It is worth noting that, in our study, CleanLab was employed to clean the entire dataset instead of the training dataset merely, leading to the removal of some hard data points in the test set. Future work should avoid this approach when using CleanLab for data cleaning.

The original scripts are provided for the reproduction and usage of this method for other peptide bioactivity prediction tasks. A user-friendly web server (pLM4ACE) for antihypertensive peptide screening and practical applications was deployed and is freely available online at https://sqzujiduce.us-east-1.awsapprunner.com/. Users can select from the three prediction models; submit a peptide sequence, a batch of sequences, or a file in FASTA, txt, or Microsoft Excel format; and receive the predicted results in real time. We anticipate that the proposed pLM4ACE architecture will be an efficient and powerful tool for ACE inhibitory peptide discovery and will inspire future model design.

Declaration of interest

The authors declare that there are no known conflicts of interest.

Acknowledgements

This is contribution No. 23-259-J from the Kansas Agricultural Experimental Station. This work was supported in part by the Agriculture and Food Research Initiative Competitive Grant no. 2020-68008-31408 and no. 2021-67021-34495 from the USDA National Institute of Food and Agriculture and a seed grant from the Global Food Systems initiative of Kansas State University.

References

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., & Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, *16*(12), Article 12. https://doi.org/10.1038/s41592-019-0598-1

Aluko, R. E. (2015). Antihypertensive Peptides from Food Proteins. *Annual Review of Food Science and Technology*, *6*(1), 235–262. https://doi.org/10.1146/annurev-food-022814-015520

Bin, Y., Zhang, W., Tang, W., Dai, R., Li, M., Zhu, Q., & Xia, J. (2020). Prediction of Neuropeptides from Sequence Information Using Ensemble Classifier and Hybrid Features. *Journal of Proteome Research*, *19*(9), 3732–3740. https://doi.org/10.1021/acs.jproteome.0c00276

Charoenkwan, P., Nantasenamat, C., Hasan, Md. M., Moni, M. A., Lio', P., & Shoombuatong, W. (2021). iBitter-Fuse: A Novel Sequence-Based Bitter Peptide Predictor by Fusing Multi-View Features. *International Journal of Molecular Sciences*, *22*(16), 8958. https://doi.org/10.3390/ijms22168958

Chen, Z., Liu, X., Zhao, P., Li, C., Wang, Y., Li, F., Akutsu, T., Bain, C., Gasser, R. B., Li, J., Yang, Z., Gao, X., Kurgan, L., & Song, J. (2022). iFeatureOmega: An integrative platform for engineering, visualization and analysis of features from molecular sequences, structural and ligand data sets. *Nucleic Acids Research*, *50*(W1), W434–W447. https://doi.org/10.1093/nar/gkac351

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

Du, Z., Comer, J., & Li, Y. (2023). Bioinformatics approaches to discovering food-derived bioactive peptides: Reviews and perspectives. *TrAC Trends in Analytical Chemistry*, 117051. https://doi.org/10.1016/j.trac.2023.117051

Du, Z., Ding, X., Xu, Y., & Li, Y. (2023). UniDL4BioPep: A universal deep learning architecture for binary classification in peptide bioactivity. *Briefings in Bioinformatics*, 1–10. https://doi.org/10.1093/bib/bbad135

Du, Z., & Li, Y. (2022a). Computer-Aided Approaches for Screening Antioxidative Dipeptides and Application to Sorghum Proteins. *ACS Food Science & Technology*. https://doi.org/10.1021/acsfoodscitech.2c00286

Du, Z., & Li, Y. (2022b). Review and perspective on bioactive peptides: A roadmap for research, development, and future opportunities. *Journal of Agriculture and Food Research*, *9*, 100353. https://doi.org/10.1016/j.jafr.2022.100353

Du, Z., Tian, W., Tilley, M., Wang, D., Zhang, G., & Li, Y. (2022). Quantitative assessment of wheat quality using near-infrared spectroscopy: A comprehensive review. *Comprehensive Reviews in Food Science and Food Safety*, *21*(3), 2956–3009. https://doi.org/10.1111/1541-4337.12958

Du, Z., Wang, D., & Li, Y. (2022). Comprehensive Evaluation and Comparison of Machine Learning Methods in QSAR Modeling of Antioxidant Tripeptides. *ACS Omega*, acsomega.2c03062. https://doi.org/10.1021/acsomega.2c03062

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., & Rost, B. (2022). ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381

FitzGerald, R. J., Cermeño, M., Khalesi, M., Kleekayai, T., & Amigo-Benavent, M. (2020). Application of in silico approaches for the generation of milk protein-derived bioactive peptides. *Journal of Functional Foods*, *64*, 103636. https://doi.org/10.1016/j.jff.2019.103636

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1

Kalyan, G., Junghare, V., Khan, M. F., Pal, S., Bhattacharya, S., Guha, S., Majumder, K., Chakrabarty, S., & Hazra, S. (2021). Anti-hypertensive Peptide Predictor: A Machine Learning-Empowered Web Server for Prediction of Food-Derived Peptides with Potential Angiotensin-Converting Enzyme-I Inhibitory Activity. *Journal of Agricultural and Food Chemistry*, *69*(49), 14995–15004. https://doi.org/10.1021/acs.jafc.1c04555

Kumar, R., Chaudhary, K., Singh Chauhan, J., Nagpal, G., Kumar, R., Sharma, M., & Raghava, G. P. S. (2015). An in silico platform for predicting, screening and designing of antihypertensive peptides. *Scientific Reports*, *5*(1), Article 1. https://doi.org/10.1038/srep12512

Lertampaiporn, S., Hongsthong, A., Wattanapornprom, W., & Thammarongtham, C. (2022). Ensemble-AHTPpred: A Robust Ensemble Machine Learning Model Integrated With a New Composite Feature for Identifying Antihypertensive Peptides. *Frontiers in Genetics*, *13*, 883766. https://doi.org/10.3389/fgene.2022.883766

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, *379*(6637), 1123–1130. https://doi.org/10.1126/science.ade2574

Lu, A. X., Zhang, H., Ghassemi, M., & Moses, A. (2020). Self-Supervised Contrastive Learning of Protein Representations By Mutual Information Maximization. *bioRxiv*, 2020.09.04.283929. https://doi.org/10.1101/2020.09.04.283929

Majumder, K., & Wu, J. (2014). Molecular Targets of Antihypertensive Peptides: Understanding the Mechanisms of Action Based on the Pathophysiology of Hypertension. *International Journal of Molecular Sciences*, *16*(1), 256–283. https://doi.org/10.3390/ijms16010256

Manavalan, B., Basith, S., Shin, T. H., Wei, L., & 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. https://doi.org/10.1093/bioinformatics/bty1047

McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *bioRxiv*. https://doi.org/10.48550/arXiv.1802.03426

Minkiewicz, Iwaniak, & Darewicz. (2019). BIOPEP-UWM Database of Bioactive Peptides: Current Opportunities. *International Journal of Molecular Sciences*, *20*(23), 5978. https://doi.org/10/gnmt2w

Mudgil, P., Baby, B., Ngoh, Y.-Y., Kamal, H., Vijayan, R., Gan, C.-Y., & Maqsood, S. (2019). Molecular binding mechanism and identification of novel anti-hypertensive and anti-inflammatory bioactive peptides from camel milk protein hydrolysates. *LWT*, *112*, 108193. https://doi.org/10.1016/j.lwt.2019.05.091

Northcutt, C., Jiang, L., & Chuang, I. (2021). Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, *70*, 1373–1411.

Olsen, T. H., Yesiltas, B., Marin, F. I., Pertseva, M., García-Moreno, P. J., Gregersen, S., Overgaard, M. T., Jacobsen, C., Lund, O., Hansen, E. B., & Marcatili, P. (2020). AnOxPePred: Using deep learning for the prediction of antioxidative properties of peptides. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-78319-w

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Qin, D., Wang, R., Jiao, L., Li, B., Wang, G., & Liang, G. (2022). ACEiPP: A Deep Learning-Based Framework to Predict Angiotensin-Converting Enzyme (ACE)-Inhibitory Peptides Using High-Efficiency Amino Acid Descriptors. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4177978

Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., & Rives, A. (2020). Transformer protein language models are unsupervised structure learners. *bioRxiv*, 2020.12.15.422761. https://doi.org/10.1101/2020.12.15.422761

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, *118*(15), e2016239118. https://doi.org/10.1073/pnas.2016239118

Santos, I., Nedjah, N., & de Macedo Mourelle, L. (2017). Sentiment analysis using convolutional neural network with fastText embeddings. *2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, 1–5. https://doi.org/10.1109/LA-CCI.2017.8285683

Sharma, Y., Agrawal, G., Jain, P., & Kumar, T. (2017). Vector representation of words for sentiment analysis using GloVe. *2017 International Conference on Intelligent Communication and Computational Techniques (ICCT)*, 279–284. https://doi.org/10.1109/INTELCCT.2017.8324059

Wang, F., & Zhou, B. (2020). Investigation of angiotensin-I-converting enzyme (ACE) inhibitory tri-peptides: A combination of 3D-QSAR and molecular docking simulations. *RSC Advances*, *10*(59), 35811–35819. https://doi.org/10.1039/D0RA05119E

Wang, L., Niu, D., Wang, X., Khan, J., Shen, Q., & Xue, Y. (2021). A Novel Machine Learning Strategy for the Prediction of Antihypertensive Peptides Derived from Food with High Efficiency. *Foods*, *10*(3), 550. https://doi.org/10.3390/foods10030550

Wang, Y.-T., Russo, D. P., Liu, C., Zhou, Q., Zhu, H., & Zhang, Y.-H. (2020). Predictive Modeling of Angiotensin I-Converting Enzyme Inhibitory Peptides Using Various Machine Learning Approaches. *Journal of Agricultural and Food Chemistry*, *68*(43), 12132–12140. https://doi.org/10/gnmdcn

Zhang, W., Xia, E., Dai, R., Tang, W., Bin, Y., & Xia, J. (2022). PredAPP: Predicting Anti-Parasitic Peptides with Undersampling and Ensemble Approaches. *Interdisciplinary Sciences: Computational Life Sciences*, *14*(1), 258–268. https://doi.org/10.1007/s12539-021-00484-x

Zhang, Y., Dai, Z., Zhao, X., Chen, C., Li, S., Meng, Y., Suonan, Z., Sun, Y., Shen, Q., Wang, L., & Xue, Y. (2023). Deep learning drives efficient discovery of novel antihypertensive peptides from soybean protein isolate. *Food Chemistry*, *404*, 134690. https://doi.org/10.1016/j.foodchem.2022.134690

Figures and tables:

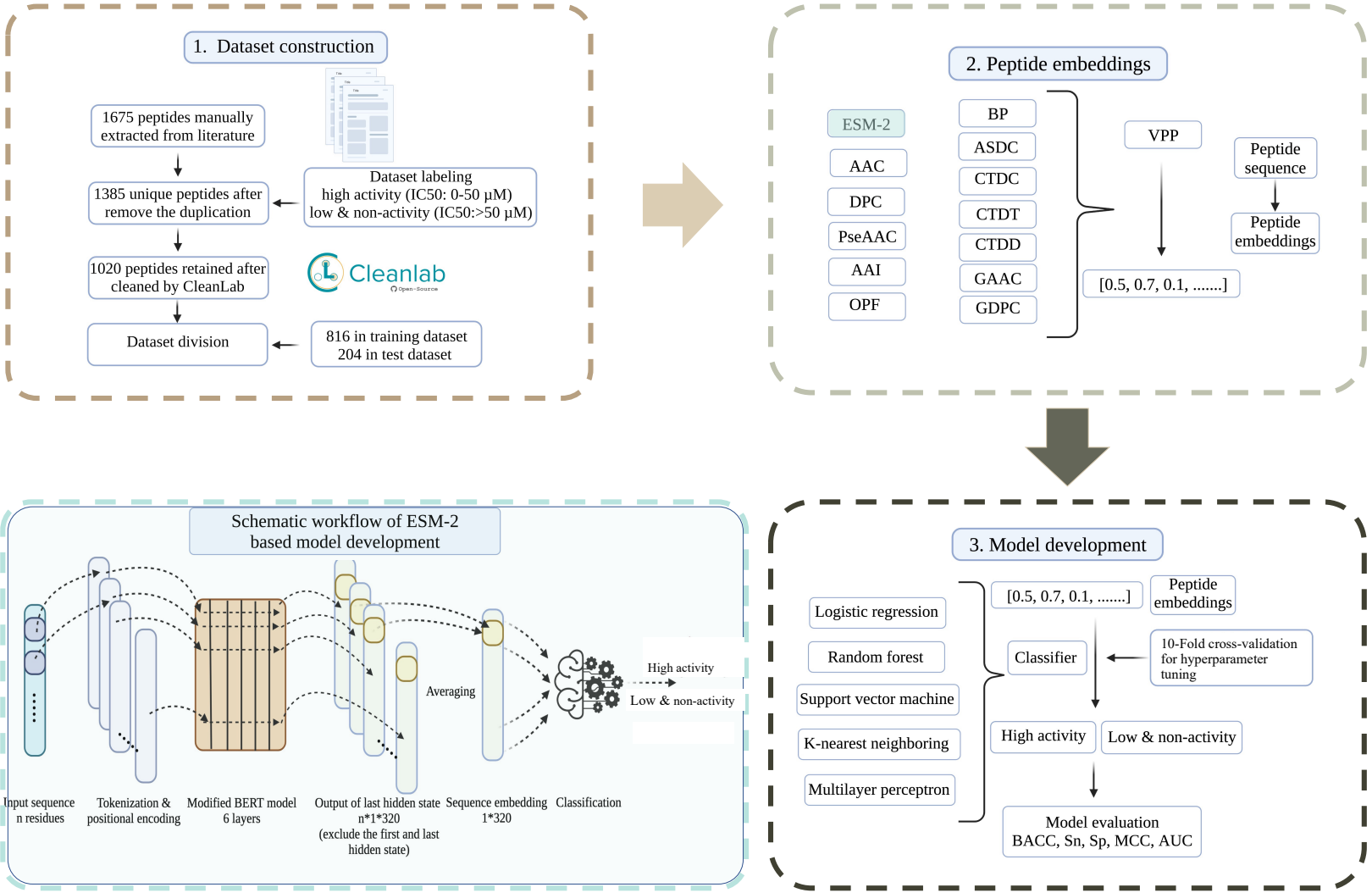


Figure 5-1 Workflow of the study for screening peptides with high antihypertensive activity.

AAC: amino acid composition; AAI: amino acid index; ASDC: adaptive skip dipeptide composition; BP: binary profile/one-hot encoding; CTDC: composition-transition-distribution composition; CTDT: composition-transition-distribution transition; CTDD: composition-transition-distribution distribution; DPC: dipeptide composition; GAAC: grouped amino acid composition; GDPC: grouped dipeptide composition; OPF: overlapping property features; PseAAC: pseudo amino acid composition.

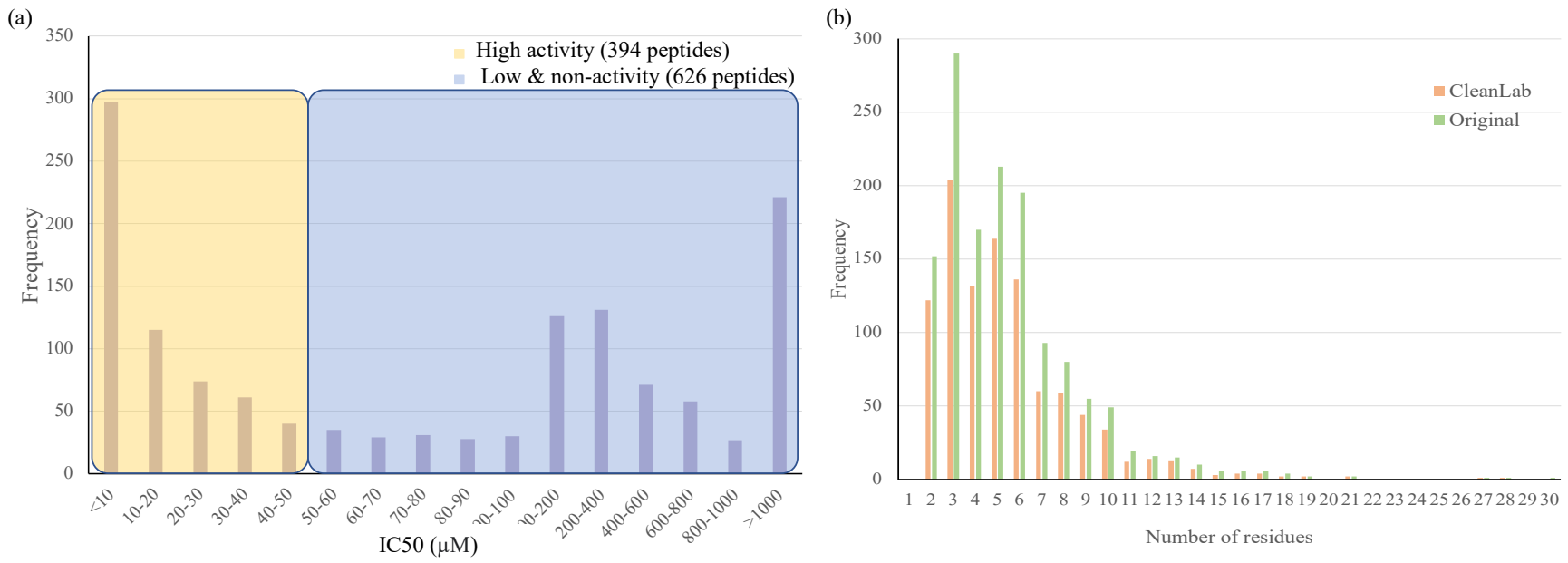


Figure 5-2 Dataset profiles: (a) Antihypertensive activity distribution and labeling, (b) Peptide distribution by length before and after data cleaning. Peptides with low & non-activity are grouped together in the binary classification model development and treated as one same category.

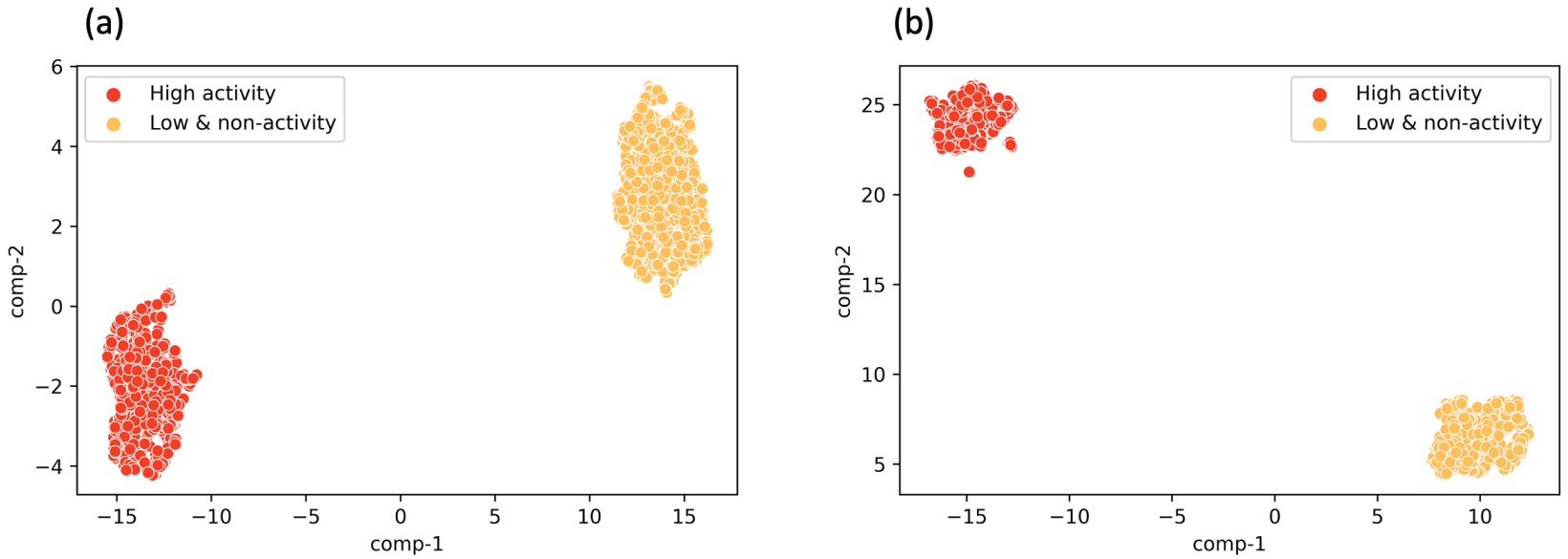


Figure 5-3 Uniform manifold approximation and projection (UMAP) of positive and negative samples in the peptide dataset before data cleaning (a) and after data cleaning (b). Peptides with low & non-activity are grouped together in the binary classification model development and treated as one same category.


(a)
ESM-MLP
ESM-SVM
ESM-KNN
ESM-RF
ESM-LR
GDPC-MLP
GDPC-SVM
GDPC-KNN
GDPC-RF
GDPC-LR
GAAC-MLP
GAAC-SVM
GAAC-KNN
GAAC-RF
GAAC-LR
CTFF-MLP
CTDD-SVM
CTDD-KNN
CTDD_RF
CTDD-LR
CTDT-MLP
CTDT-SVM
CTDT-KNN
CTDT-RF
CTDT-LR
CTDC-MLP
CTDC-SVM
CTDC-KNN
CTDC-RF
CTDC-LR
ASDC-MLP
ASDC-SVM
ASDC-SVM
ASDC-RF
ASDC-LR
BP-MLP
BP-SVM
BP-KNN
BP-RF
BP-LR
OPF-MLP
OPF-SVM
OPF-KNN
OFP-RF
OPF-LR
AAI-MLP
AAI-SVM
AAI-SVM
AAI-RF
AAI-LR
PseAAC-MLP
PseAAC-SVM
PseAAC-KNN
PseAAC-RF
PseAAC-LR
DPC-MLP
DPC-SVM
DPC-KNN
DPC-RF
DPC-LR
AAC-MLP
AAC-SVM
AAC-KNN
AAC-RF
AAC-LF
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
(b)
ESM-MLP
ESM-SVM
ESM-KNN
ESM-RF
ESM-LR
GDPC-MLP
GDPC-SVM
GDPC-KNN
GDPC-RF
GDPC-LR
GAAC-MLP
GAAC-SVM
GAAC-KNN
GAAC-RF
GAAC-LR
CTFF-MLP
CTDD-SVM
CTDD-KNN
CTDD_RF
CTDD-LR
CTDT-MLP
CTDT-SVM
CTDT-KNN
CTDT-RF
CTDT-LR
CTDC-MLP
CTDC-SVM
CTDC-KNN
CTDC-RF
CTDC-LR
ASDC-MLP
ASDC-SVM
ASDC-SVM
ASDC-RF
ASDC-LR
BP-MLP
BP-SVM
BP-KNN
BP-RF
BP-LR
OPF-MLP
OPF-SVM
OPF-KNN
OFP-RF
OPF-LR
AAI-MLP
AAI-SVM
AAI-SVM
AAI-RF
AAI-LR
PseAAC-MLP
PseAAC-SVM
PseAAC-KNN
PseAAC-RF
PseAAC-LR
DPC-MLP
DPC-SVM
DPC-KNN
DPC-RF
DPC-LR
AAC-MLP
AAC-SVM
AAC-KNN
AAC-RF
AAC-LR
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
(c)
ESM-MLP
ESM-SVM
ESM-KNN
ESM-RF
ESM-LR
GDPC-MLP
GDPC-SVM
GDPC-KNN
GDPC-RF
GDPC-LR
GAAC-MLP
GAAC-SVM
GAAC-KNN
GAAC-RF
GAAC-LR
CTFF-MLP
CTDD-SVM
CTDD-KNN
CTDD_RF
CTDD-LR
CTDT-MLP
CTDT-SVM
CTDT-KNN
CTDT-RF
CTDT-LR
CTDC-MLP
CTDC-SVM
CTDC-KNN
CTDC-RF
CTDC-LR
ASDC-MLP
ASDC-SVM
ASDC-SVM
ASDC-RF
ASDC-LR
BP-MLP
BP-SVM
BP-KNN
BP-RF
BP-LR
OPF-MLP
OPF-SVM
OPF-KNN
OFP-RF
OPF-LR
AAI-MLP
AAI-SVM
AAI-SVM
AAI-RF
AAI-LR
PseAAC-MLP
PseAAC-SVM
PseAAC-KNN
PseAAC-RF
PseAAC-LR
DPC-MLP
DPC-SVM
DPC-KNN
DPC-RF
DPC-LR
AAC-MLP
AAC-SVM
AAC-KNN
AAC-RF
AAC-LR
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
(d)
ESM-MLP
ESM-SVM
ESM-KNN
ESM-RF
ESM-LR
GDPC-MLP
GDPC-SVM
GDPC-KNN
GDPC-RF
GDPC-LR
GAAC-MLP
GAAC-SVM
GAAC-KNN
GAAC-RF
GAAC-LR
CTFF-MLP
CTDD-SVM
CTDD-KNN
CTDD_RF
CTDD-LR
CTDT-MLP
CTDT-SVM
CTDT-KNN
CTDT-RF
CTDT-LR
CTDC-MLP
CTDC-SVM
CTDC-KNN
CTDC-RF
CTDC-LR
ASDC-MLP
ASDC-SVM
ASDC-SVM
ASDC-RF
ASDC-LR
BP-MLP
BP-SVM
BP-KNN
BP-RF
BP-LR
OPF-MLP
OPF-SVM
OPF-KNN
OFP-RF
OPF-LR
AAI-MLP
AAI-SVM
AAI-SVM
AAI-RF
AAI-LR
PseAAC-MLP
PseAAC-SVM
PseAAC-KNN
PseAAC-RF
PseAAC-LR
DPC-MLP
DPC-SVM
DPC-KNN
DPC-RF
DPC-LR
AAC-MLP
AAC-SVM
AAC-KNN
AAC-RF
AAC-LR
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Figure 5-4 Model performance of the five machine learning methods combined with ESM-2 and 12 other peptide embedding approaches. (a) Balanced accuracy (BACC); (b) Matthews correlation coefficient (MCC); (c) Sensitivity (Sn); (d) Specificity (Sp)

Peptide embedding methods are AAC: amino acid composition; AAI: amino acid index; ASDC: adaptive skip dipeptide composition; BP: binary profile/one-hot encoding; CTDC: composition-transition-distribution composition; CTDT: composition-transition-distribution transition; CTDD: composition-transition-distribution distribution; DPC: dipeptide composition; GAAC: grouped amino acid composition; GDPC: grouped dipeptide composition; OPF: overlapping property features; PseAAC: pseudo amino acid composition. Machine learning methods are LR: logistic regression; RF: random forest; SVM: support vector machine; KNN: k-nearest neighboring; MLP: multilayer perceptron.

Table 5-1 Top model performance in test dataset with ESM-2-based embeddings for peptide ACE inhibitory activity prediction

Model	**Balanced accuracy**	**Sensitivity**	**Specificity**	**Matthews correlation coefficient**	**Area under the curve**
Logistic regression	0.883 ± 0.017	0.845 ± 0.041	0.92 ± 0.025	0.77 ± 0.032	0.96 ± 0.009
Support vector machine	0.867 ± 0.02	0.825 ± 0.041	0.91 ± 0.027	0.74 ± 0.038	0.955 ± 0.01
Multilayer perceptron	0.855 ± 0.024	0.815 ± 0.036	0.895 ± 0.037	0.711 ± 0.054	0.951 ± 0.017

Chapter 6 - pLM4Alg: protein language model-based predictors for allergenic proteins and peptides*

Abstract

The rising prevalence of allergy demands efficient and accurate bioinformatic tools to expedite allergen identification and risk assessment while also reducing wet experiment expenses and time. Recently, pre-trained protein language models (pLMs) have successfully predicted protein structure and function. However, to our best knowledge, they have not been used for predicting allergenic proteins/peptides. Therefore, this study aims to develop robust models for allergenic protein/peptide prediction using five pLMs of varying sizes and systematically assess their performance through fine-tuning with a convolutional neural network. The developed pLM4Alg models have achieved state-of-the-art performance with accuracy, Matthews correlation coefficient, and area under the curve scoring 94.1-95.5%, 0.882-0.911, and 98.3-99%, respectively. Moreover, pLM4Alg is the first model capable of handling prediction tasks involving residue-missed sequences and sequences containing non-standard amino acid residues. To facilitate easy access, a user-friendly web server (https://f6wxpfd3sh.us-east-1.awsapprunner.com/) has been established. pLM4Alg is expected to become the leading machine learning-based prediction model for allergenic peptides and proteins. Its collaboration with other predictors holds a great promise in accelerating allergy research.

Keywords: Allergenicity; allergens; proteins and peptides; protein language model; convolutional neural network; deep learning

---

* Du, Z., Xu, Y., Liu, C., & Li, Y. (2023). pLM4Alg: Protein Language Model-Based Predictors for Allergenic Proteins and Peptides. Journal of Agricultural and Food Chemistry, 72(1), 752-760. https://doi.org/10.1021/acs.jafc.3c07143

Abstract

The rising prevalence of allergy demands efficient and accurate bioinformatic tools to expedite allergen identification and risk assessment while also reducing wet experiment expenses and time. Recently, pre-trained protein language models (pLMs) have successfully predicted protein structure and function. However, to our best knowledge, they have not been used for predicting allergenic proteins/peptides. Therefore, this study aims to develop robust models for allergenic protein/peptide prediction using five pLMs of varying sizes and systematically assess their performance through fine-tuning with a convolutional neural network. The developed pLM4Alg models have achieved state-of-the-art performance with accuracy, Matthews correlation coefficient, and area under the curve scoring 93.4-95.1%, 0.869-0.902, and 0.981-0.990, respectively. Moreover, pLM4Alg is the first model capable of handling prediction tasks involving residue-missed sequences and sequences containing non-standard amino acid residues. To facilitate easy access, a user-friendly web server (https://f6wxpfd3sh.us-east-1.awsapprunner.com/) has been established. pLM4Alg is expected to become the leading machine learning-based prediction model for allergenic peptides and proteins. Its collaboration with other predictors holds a great promise in accelerating allergy research.

Keywords: Allergenicity; allergens; proteins and peptides; protein language model; convolutional neural network; deep learning

6.1 Introduction

Allergic diseases encompass a range of disorders driven by both innate and adaptive immune responses, in collaboration with epithelial cells (Celebi Sozener et al., 2022). These interactions lead to hypersensitivity against otherwise harmless foreign substances (e.g., foods, dust mites, pollens, chemicals, etc.) (Ogulur et al., 2021). Most allergic diseases involve a type 2 immune response, including various cells like Th2 cells, type 2 B cells, type 2 innate lymphoid cells, type 2 macrophages, basophils, eosinophils, mast cells, as well as specific NK cells and NK-T cells that secrete interleukin (IL)-4, IL-5, IL-9, IL-13, IL-25, and IL-33 cytokines (Breiteneder et al., 2019; Han et al., 2020). Upon exposure, allergens cross-link immunoglobulin E (IgE) antibodies on the surface of granulocytes, which leads to the release of mediators and inflammatory molecules (e.g., histamine and leukotriene, etc.), triggering various symptoms, including allergic asthma, rhinitis, skin reactions, and even death (Sharma, Patiyal, Dhall, Pande, et al., 2021). Over 30 % of the U.S. population has been affected by these adverse reactions, and the prevalence continues to rise (Goodman et al., 2016; Sharma, Patiyal, Dhall, Pande, et al., 2021). Numerous theories have been proposed to explain the growing allergy epidemics including the hygiene hypothesis and its extensions (e.g., the old friends hypothesis and biodiversity hypothesis) (Lambrecht & Hammad, 2017), the nutritional immunomodulation hypothesis (e.g., the vitamin D hypothesis) (Bozzetto et al., 2012; Vlieg-Boerstra et al., 2023), the epithelial barrier hypothesis (Akdis, 2006; Celebi Sozener et al., 2022), and the climate change hypothesis (Rothenberg, 2022).

For a substance to be allergenic, it must induce IgE production during sensitization, involving dendritic cells, T cells, and B cells, and trigger clinical reactions upon subsequent exposures, encompassing immediate- and late-phase responses. However, this simplistic view overlooks complexities during initial allergen exposure, such as allergen biochemistry,

surrounding innate immune stimulants, allergen stability in tissues and mucosa, and dosage and duration of interaction with the immune system (Akdis, 2006). The majority of allergens are protein/peptides, which are widely present in foods, personal care products, and the environment (Sharma, Patiyal, Dhall, Devi, et al., 2021; J. Wang et al., 2013; Yu et al., 2023). Notably, many allergens, especially in plants and foods, belong to a few protein families (Radauer et al., 2008). The reason certain proteins act as allergens remains unclear, but they often share characteristics contributing to allergenicity, including small size, water solubility, glycosylation, repetitive structures, numerous disulfide bonds, resistance to heat, acid, and proteolysis, ligand-binding property, ability to interact with membranes and lipids, and intrinsic adjuvant property (Akdis, 2006; Sathe et al., 2016; Scheurer et al., 2015).

To improve efficiency in allergen identification and accelerate allergenic risk evaluation, computational approaches have been continuously proposed and have made significant progress in the past decades. The first guideline was proposed by the consultation group of FAO/WHO in 2001, where the allergenic potential of a protein was defined by an exact match of a stretch of six or more consecutive identical amino acids (rule 1), or more than 35% identity within any window of 80 amino acids in comparison to any known allergen (rule 2)(FAO/WHO, 2001). Subsequently, various approaches are proposed from different perspectives for allergen prediction, such as motif-based approaches, similarity searches, machine learning-based modeling, etc. Prediction tools, such as AllerCatPro 2.0, AlgPred 2.0, have been used extensively for preliminary allergenic risk assessment and identification of new allergens (Behbahani et al., 2020; Dang & Lawrence, 2014; Dimitrov, Bangov, et al., 2014; Dimitrov, Naneva, et al., 2014; Goto et al., 2023; Maurer-Stroh et al., 2019; Nedyalkova et al., 2023; Nguyen et al., 2022; Sharma, Patiyal, Dhall, Pande, et al., 2021; L. Wang et al., 2021; Westerhout et al., 2019; Yu et al., 2023). One representative project derived

from the FAO/WHO guidelines was AllerCatPro/AllerCatPro 2.0, which implemented a hierarchical workflow employing five criteria with improved similarity-checking methods for both sequences and 3D structures (Maurer-Stroh et al., 2019; Nguyen et al., 2022). The criteria used in AllerCatPro were based on statistical analysis of existing allergens, representing a form of experience/knowledge-based decision-making. For instance, if a protein sequence shares >35% identity with known allergens over 90 windows, it would be classified into the group with strong evidence for allergenic potential (Maurer-Stroh et al., 2019; Nguyen et al., 2022). So far, this approach has achieved a great performance in the allergen benchmark datasets (Nguyen et al., 2022). However, the thresholds set in these criteria were based on previous studies, common sense, and accumulated knowledge. There was no independent dataset for validating these criteria. Additionally, the reliance on allergens in the search databases for similarity checking significantly limited model robustness when encountering unseen allergens.

Another representative technique is the machine learning (ML) approach where the classification criterion is determined by the ML model by learning from a separate training dataset (Behbahani et al., 2020; Dang & Lawrence, 2014; Dimitrov, Bangov, et al., 2014; Dimitrov, Naneva, et al., 2014; Goto et al., 2023; Nedyalkova et al., 2023; Sharma, Patiyal, Dhall, Pande, et al., 2021; L. Wang et al., 2021; Westerhout et al., 2019; Yu et al., 2023). Besides the selection of modeling method, the most challenging task in the ML approach is the representation or encoding of protein/peptide sequences into a numerical vector/matrix. Various encoding methods, including amino acid descriptors, amino acid composition (AAC), pseudo-amino acid composition (PseAAC), dipeptide composition (DPC), amino acid descriptors (AAD), position-specific scoring matrix (PSSM), physicochemical descriptors, biomedical properties, k-mer dictionary-based binary representation, etc., have been widely used in predicting allergenicity and other

properties/bioactivities (Dang & Lawrence, 2014; Dimitrov, Naneva, et al., 2014; Du et al., 2022; Du, Comer, et al., 2023; Du & Li, 2022; Islam et al., 2021; Naneva et al., 2019; Nedyalkova et al., 2023; Nedyalkova & Simeonov, 2020; Sharma, Patiyal, Dhall, Pande, et al., 2021; L. Wang et al., 2021; Yu et al., 2023). However, these features may not always accurately represent protein sequences and simple combinations can cause high-dimensional problems as well as the feature redundancy (Du, Ding, et al., 2023).

Recently, inspired by the extraordinary performance of transformer-based models in natural language processing (NLP) tasks, a series of attempts have been made in pretrained protein language models (pLMs) (e.g., ProtTrans, ESM, UniRep, etc.) for downstream protein sequence tasks (Alley et al., 2019; Du, Ding, et al., 2023; Du et al., 2024; Elnaggar et al., 2022; Lin et al., 2023; Rives et al., 2021; L. Wang et al., 2021). The central concept lies in incorporating the attention mechanism during the self-supervised learning process in these pLMs. This mechanism enables the model to observe the entire sequence simultaneously to learn the representation of each residue as well as their relationship, rather than locally (as in convolutional neural network (CNN)) or sequentially (as in long short-term networks (LSTM)). As a result, the model can capture the overall contextual information of the entire sequence more effectively and efficiently (Vaswani et al., 2017).

Evolutionary scale modeling (ESM), one of the pretrained pLMs, was updated into ESM-2 in 2022 by the Meta Fundamental AI Research Protein Team (FAIR) for protein representation (Lin et al., 2023). In our previous study, it exhibited a great potential in developing a universal model architecture and achieved state-of-the-art (SOTA) performance in 15 out of 20 peptide bioactivity/function datasets by fitting the same model architecture in different peptide datasets (Du, Ding, et al., 2023). Due to the limited data availability of bioactive peptides and the universal

application scenarios, a small size ESM-2 pLM (320 output dimension) and a simplified model architecture were selected (Du, Ding, et al., 2023). ESM-2 includes five additional pretrained pLMs varying from 48 layers with 15 billion parameters for 5120 output embeddings dimensions to 6 layers with 8 million parameters for 320 output embeddings dimensions. The higher output dimensions corresponded to more information extracted and retained from the protein sequences, resulting in better performance in the downstream tasks (Lin et al., 2023; Rives et al., 2021). However, the high dimension feature space can sometimes cause the curse of dimensionality and hinder the improvement in performance when the dataset size is small (Du, Comer, et al., 2023). Up to now, more than 8000 allergens have been identified and archived in public databases (Goodman et al., 2016; Mari et al., 2009; Pomés et al., 2018; The UniProt Consortium, 2021; van Ree et al., 2021), which allows us to fully exploit the potential of various ESM-2 pLMs with different sizes in allergen prediction.

To the best of our knowledge, ESM-2 LMs have not been used for the numerical representation of allergenic proteins/peptides for prediction model development. This study aims to develop pLM-based high-quality models for assessing the allergenic potential of proteins/peptides. A series of pLMs of different sizes were compared regarding their performance in allergenicity prediction when combined with CNN (Figure 1). In addition, the study established a new allergenic sequence dataset that includes non-standard residues ([BOUZ]) and employed it in the model development. Based on the final performance, we deployed a user-friendly web server, accessible online at no cost, to meet the great demands for allergenicity risk assessment and allergen identification. The comparison between different pretrained pLMs will provide valuable insight into the relationship between pLM size, fine-tuning with CNN, and prediction performance in downstream protein/peptide tasks. Moreover, the study may inspire future

development of pLM-based models for other physicochemical/bioactive property predictions of proteins and peptides, particularly in the areas of architecture design and hyperparameter tuning.

6.2 Materials and methods

6.2.1 Dataset construction

The dataset used in this study was compiled from five public databases: ALLERGOME, COMPARE, AllergenOnline, WHO/IUIS, and Uniprot (Mari et al., 2009; Pomés et al., 2018; The UniProt Consortium, 2021; van Ree et al., 2021). The sequences in these databases were all peer-reviewed by expert panels and/or sourced from peer-reviewed journals and were treated as high-quality data sources. Specifically, a total of 2631 allergens were downloaded from COMPARE at https://comparedatabase.org/ (the latest updates on January 26, 2023). Additionally, 2233, 1090, and 4981 sequences were retrieved from AllergenOnline (http://www.allergenonline.org/) (released on February 14, 2021), WHO/IUIS (http://www.allergen.org/index.php) (the latest updates on April 3, 2023), and ALLERGOME (https://www.allergome.org/index.php) (the latest updates on February 28, 2023), respectively. These sequences were downloaded from the NCBI database (https://www.ncbi.nlm.nih.gov/) by data mining using Selenium with python. The AllergenOnline database also contains 1041 peptides and 75 proteins related to celiac disease, released on April 7, 2022. Furthermore, 1279 allergenic sequences were extracted from UniProt (https://www.uniprot.org/) using the keyword “allergen AND reviewed: yes”. After removing duplicate sequences and a very limited number of long sequences (>900 residues), 8029 allergenic sequences as the positive dataset from a total of 22330 collected allergenic sequences are finally obtained (Supplementary Document S1). Sequences containing non-standard amino acids ([BOUZ]) were retained in the datasets, since ESM-2 pLMs can recognize and encode them.

Additionally, some sequences from ALLERGOME containing X ('unknown') residues were retained since they can also be encoded by ESM-2 pLMs.

Constructing a non-allergens dataset is challenging since there are no ready-to-use databases, specifically dedicated to non-allergenic sequences. A common approach is to retrieve random sequences through the exclusion of keywords (Du, Comer, et al., 2023). In this study, we followed a previous study by using the query "NOT allergen NOT cancer NOT allergenic AND reviewed: yes" to retrieve the dataset from UniProt (Sharma, Patiyal, Dhall, Pande, et al., 2021). This process yielded 557,590 sequences (retrieved on May 12, 2023). Out of these, 8029 sequences were selected based on the same length distribution as observed in the positive dataset. The length distribution of the allergenic and nonallergenic sequences is shown in Figure 2. Finally, the dataset contains 8029 allergenic and 8029 non-allergenic sequences.

6.2.2 Protein language model for embeddings and dataset preprocessing

ESM is a pLM project initiated by FAIR in 2019 (https://github.com/facebookresearch/esm). The SOTA version (ESM-2) was released in 2022 and trained on UR50/D 2021_04 datasets (Lin et al., 2023). Out of the 6 available pLMs, five were selected for sequence embeddings due to their more manageable RAM memory consumption (Table 1).

6.2.3 CNN model architecture

Inspired by our previous study, a convolutional neural network (CNN) was used to build the connection between the sequence embeddings and allergenicity prediction (Du, Ding, et al., 2023). The overall architecture in our previous study was retained, where a convolutional layer, a dropout layer (rate = 0.15), a flatten layer, and two dense layers with a dropout layer (rate = 0.15) were connected sequentially. Four hyperparameters (filter size, kernel size, and stride in the

convolutional layer, and units in the first dense layer) were optimized through customized grid search. Filter size was selected from [16,32,64,128,256], kernel size from [32,64,128,256], stride from [1,2,4,8], and units from [3,6,9,12,32,64,128,256,512,1024,2056,4089,8192]), respectively.

Stochastic gradient descent (SGD) was chosen as the optimizer to accelerate the model fitting and a customized learning rate scheduler was designed to update the learning rate at every epoch. Additionally, an early stop and model checkpoint were set, and the model with best performance in accuracy was stored under the consideration of validation accuracy. The model was built with Keras framework (http://www.keras.io) and tested on a Google Colab platform with GPU for acceleration. The detailed codes are available at https://github.com/dzjxzyd/pLM4Alg .

The dataset was randomly divided into a training dataset and a test dataset at a ratio of 8:2. The test dataset is independent from training process and only used in the for the final model performance evaluation. A grid search will be conducted initially to roughly find out the potential hyperparameters for better model performance. Then 5 fold cross validation (CV) was used in hyperparameter tuning. The optimal hyperparameters selected for each pLM will be used for the CNN training on the whole training dataset without early stop and check point. The training process was stopped after it reach the epoch limit, which was observed in the CV process. Performance evaluation was conducted in the test dataset ten times with a random dataset division to ensure the robustness of the model performance.

6.2.4 Model evaluation

Accuracy (ACC), sensitivity (Sn), specificity (Sp), Matthews correlation coefficient (MCC), and area under the curve (AUC) were adopted to evaluate the model performance. These parameters were calculated based on the number of true positive (TP), false positive (FP), false negative (FN), and true negative (TN), according to the following equations:

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (6.1)$$

$$Sn = \frac{TP}{TP+FN} \quad (6.2)$$

$$Sp = \frac{TN}{TN+FP} \quad (6.3)$$

$$MCC = \frac{(TP*TN)-(FN*FP)}{\sqrt{(TP+FN)*(TN+FP)*(TP+FP)*(TN*FN)}} \quad (6.4)$$

The AUC is calculated using the sklearn 'roc_auc_score' function through the portability distribution of the model prediction in the test dataset.

6.2.5 Web server development

A user-friendly web server (available at https://f6wxpfd3sh.us-east-1.awsapprunner.com) was deployed on Amazon Web Services (AWS) App Runner based on AWS Elastic Container Registry (ECR). The website was designed with html and css scripts, and the model deployment was achieved with Flask (2.2.2) and Docker (24.0.5).

Considering computation needs in pLMs for sequence embeddings, available resources, and financial constraints, three pLM-based models, excluding esm2_t36_3B_UR50D (2560 output dimension) and esm2_t33_650M_UR50D (1280 output dimension), were deployed for the perdition tasks on the web server. The web server provides a large-scale processing option, which allows users to upload their peptides information through xls, xlsx, fasta, and txt formats for allergenic sequence prediction. The detailed scripts and the deployed models for web server development are provided and available at https://github.com/dzjxzyd/pLM4Alg and https://drive.google.com/drive/folders/1veD40uj8R7gpIo8y1niXtoKij_vZ3eiD?usp=sharing.

Besides. We also provide a user-friendly script for web server setup based on Google Colab to support the usage of all the five models in allergenic peptides/proteins prediction. Users can download the script at

https://drive.google.com/file/d/1xwlSCZhuTvbRo54CzEtqMEWwIEV9xkyf/view?usp=sharing

and open it with Google Colab to run the web server locally on Colab cloud platform freely.

6.3 Results and discussion

6.3.1 Construction of a comprehensive database of known allergens

A high-quality dataset is vital for the development of a machine learning model as it directly determines the application scenarios of the final model. In this study, we only focused on datasets that have been regularly maintained and continuously updated over the past decades and derived from reliable data sources (peer-reviewed papers and/or reviewed by experts). The datasets used in previous studies were not used since some of them were outdated or unavailable (Dimitrov, Bangov, et al., 2014; Sharma, Patiyal, Dhall, Pande, et al., 2021). The detailed allergenic sequences extracted from each database are available in Supplementary Document S1.

Though there are some overlaps in the allergenic sequences, each database has some unique features. ALLERGENOME is the largest database and fully based on allergens causing IgE-mediated reactions collected from peer-reviewed papers. It excludes low molecular weight compounds causing non-IgE-mediated allergic diseases (e.g. contact dermatitis, drug allergy) and molecules causing cell-mediated diseases (e.g. celiac disease) (Mari et al., 2009). COMPARE and AllergenOnline are general allergenic protein databases. Recently, AllergenOnline included a separate celiac disease database containing a large number of short peptides (Goodman et al., 2016; van Ree et al., 2021). The WHO/IUIS database is more stringent, requiring allergens archived to meet the WHO/IUIS criteria and demonstrate protein-specific IgE-reactivity from at least five patients (Maurer-Stroh et al., 2019; Pomés et al., 2018). The Uniprot database is not solely dedicated to allergens, but is one of the largest databases for general protein annotations. The Swiss-Prot (reviewed) database is a high-quality, manually annotated, and non-redundant protein

sequence database, which consolidates experimental results, computed features, and scientific conclusions (The UniProt Consortium, 2021).

The allergenic sequence length distribution is shown in Figure 2. The generalizability of machine learning models is highly dependent on their original datasets. Our datasets cover a wide range of lengths, spanning from 0 to 900 residues, with most sequences having lengths below 400 residues. Notably, there are 1816 sequences (22.5%) shorter than 50 residues, which assures the model's ability in predicting small allergenic proteins/peptides. In contrast, in the study of Yu et al., a total of 1000 sequences between 5-50 residues were used for developing their classification model (Yu et al., 2023). The dataset constructed in our study is the largest and the most up-to-date, except for the study of Sharma et al., where 10075 sequences were collected by compiling public databases and datasets from previous studies (Dimitrov, Bangov, et al., 2014; Nedyalkova et al., 2023; Sharma, Patiyal, Dhall, Pande, et al., 2021; Yu et al., 2023). It should be noticed, it this study, CD-HIT was not used for the non-redundancy construction since the superiority of pLMs in effective representation/embeddings of peptide/protein sequence. The pLMs based sequence embeddings can avoid the similarity in feature space between similar sequences to a large extent, compared to traditional peptide representation methods (see details in the next subsection). Besides, we also conduct some preliminary experiments with the dataset cleaned by CD-HIT under different thresholds (0.4 and 0.9). The model performance was undermined by the decrease in dataset size, but not improved by the less redundant dataset generated from CD-HIT (Supplementary Document S2).

6.3.2 Pretrained pLMs in sequence embeddings

The ESM-2 pLMs are a modified version of the bidirectional encoder representations from transformers (BERT)-based architecture trained on a dataset (UR50/D 2021_04 at

https://huggingface.co/datasets/nferruz/UR50_2021_04) containing 48 million sequences. The dataset (UR50/D 2021_04) used for developing ESM-2 pLMs also contained a proportion (around 2%) of sequences with lengths below 50 residues (https://www.uniprot.org/uniref?query=%2A). This assured the universality of ESM-2 pLMs in handling our allergenic protein/peptide datasets. Moreover, since the dataset contains non-standard amino acids ([BOUZ]), the resulting ESM-2 pLMs have the ability to encode protein sequences with these residues, whereas the sequences with non-standard residues were generally removed during model development in previous studies (Nedyalkova et al., 2023; Nguyen et al., 2022; Sharma, Patiyal, Dhall, Pande, et al., 2021). Furthermore, even if a given sequence has absent residues at a few positions, ESM-2 pLMs can still encode the sequence because of its inherent capacity from the masked language model strategy (Lin et al., 2023; Rives et al., 2021).

The output embedding matrix (n*d) for the input sequence is the last hidden state of ESM-2 models, where n represents the length of the sequences and d depends on the selected pretrained model (e.g., the d of esm2_t6_8M_UR50D model is 320) (Figure 1b). Each row in the embedding matrix represents the corresponding residues, and the generation of the embedding considers both the residues and the context of the entire sequence. Subsequently, an average pooling operation is conducted to generate a 1*dim feature to represent the original sequences, thus allowing sequences with varying lengths to be unified to the same length and loaded into the model for prediction tasks (Lin et al., 2023). The representation of a sequence generated by ESM-2 pLM can be considered a global descriptor of the sequence.

In previous studies where global descriptors (e.g., AAC, DPC, PSSM, etc.) were used for sequence embeddings, the sequential information of proteins/peptides was lost, thereby hindering further improvement in prediction model performance (Nedyalkova et al., 2023; Sharma, Patiyal,

Dhall, Pande, et al., 2021; J. Wang et al., 2013). For example, a tripeptide "ALL" encoded by AAC would be indistinguishable from the tripeptide "LLA". It can be observed in the study of Sharma et al., where the pure machine learning models only achieve 84% accuracy with AAC as the descriptor. On the other hand, local descriptors (e.g., amino acid descriptor, physicochemical properties, etc.) faced challenges in unifying the feature dimension of the embeddings. In a study by Yu et al., each property in AAindex of a peptide was the average of all residues, resulting in a less promising performance for allergenic peptide prediction (accuracy = 72.2%) (Yu et al., 2023). In addition, amino acid descriptors such as DPPS, 3Z-scale, 5-Zscale, HESH, FASGAI, ISA-ECI, and E-scale, yielded inferior performance than AAindex (Yu et al., 2023). These encoding approaches also suffered from the lack of sequential information present in global descriptors (AAC, DPC, etc.). Another feature dimension unifying strategy is to select m N-terminus residues and n C-terminus residues (m and n are numbers). However, this is also an expedient solution since the information of the middle residues is lost (Du, Comer, et al., 2023).

Since pLMs could capture more feature information at both the residue-wise and the contextual levels, it is more sensitive compared to the abovementioned traditional encoding methods. Consequently, pLMs-based embeddings can effectively differentiate proteins/peptides. For example, similar sequences (e.g., "ALL" and "LLA", "AAL" and "ALLA", "ALLERGY" and "ALLERGG") can be encoded distinctly, avoiding highly similar or identical feature matrixes/vectors between them. This distinction is critical for subsequent machine learning models to identify specific patterns with a high-resolution. Herein, ESM-2 pLMs possess such a capacity and thus do not require the popular redundancy removal approach (CD-HIT) to clean the dataset. It is worth noting that in our previous study on the angiotensin-converting enzyme inhibitory peptide prediction model, ESM-based embeddings exhibited a significant superiority over

traditional descriptors (local descriptors such as AAIndex, 3Z-scale, DPPS, etc., and global descriptors such as AAC, DPC, etc.) (Du et al., 2024). Such a superiority is also observed in the study of Wang et al. where a LM, protBERT, was used for the allergenic sequence representation and exhibited better performance compared to PseAAC (L. Wang et al., 2021).

6.3.3 Performance of pLMs-based models

The overall architecture used in this study was based on our previous universal architecture design using pLMs (Du, Ding, et al., 2023). Four key hyperparameters (filter size, kernel size, and stride in the convolutional layer, and units in the first dense layer) were selected in the CNN model development to fit the training dataset after embedding with different ESM-2 pLMs. In our customized grid searching, units in the first dense layer played a dominate role in improving the final model performance as shown in Supplementary Document S2. Due to limited computation resources, random dataset divisions were not performed during the grid search. Instead, the hyperparameter combinations that had the higher performance with a specific ESM-2 pLM were further evaluated in 5 fold CV. The hyperparameters achieved promising performance in CV would be used for the final model development on the whole training dataset and evaluated on the independent test dataset ten times with random dataset divisions. As shown in Table 2, pLM4Alg-2560 achieved the best performance with ACC = 95.1±0.4%, Sn = 94.2±0.7%, Sp = 96.0±0.4%, MCC = 0.902±0.008, and AUC = 0.99±0.001.

A positive correlation between overall performance and output dimension was observed, which is consistent with previous studies on downstream tasks of pLMs (Elnaggar et al., 2022; Lin et al., 2023). This can be attributed to the higher output dimension and the greater amount of contained information in the larger pretrained pLMs. The original feature dimension of a sequence is high when encoded to contain the complete sequence information (e.g., one hot encoding). As

an encoder, the pLM can map the full input space to a lower dimension, while striving to retain as much meaningful information as possible. Hence, a higher output dimension means that the encoder preserved more information after the encoding process. The final model's performance was also dependent on the dataset size, as enough data points are needed to allow the neural network to tune the parameters with backpropagation. Although a higher output dimension contains more information, it also requires more parameters to fit the feature space. Consequently, a sufficiently large dataset size is necessary to accommodate the increase in feature dimensions. Otherwise, it would paralyze further improvement in model performance because of the curse of dimensionality (Du et al., 2024). This might explain why there was a logarithmic increase in the model performance with the increase in output dimension.

6.3.4 Comparison with existing methods

The performances of the existing ML-based predictors are shown in Table 2. All pLM-based models outperformed the state-of-the-art (SOTA) models regarding the evaluation matrics. The performance of a ML model is influenced by various factors, such as its datasets, features, and modeling methods. Considering the different datasets used in the listed models, it would be unfair to simply state that one performs better than others. To ensure a fair comparison, we employed pLM-based models to fit the three original datasets of existing models (Table 2) that are available online, using the selected hyperparameters based on our datasets. The pLM-based models exhibited superior performance (Supplementary Document S3) on the AllerTOPv2, AllergenFP, and AllerStat datasets than the existing models. ACC of our models ranged from 91.2 to 92.8%, higher than those of AllerTOPv2 (88.7%) and of AllerTOPv (87.9%), using the same dataset. Furthermore, our models had an AUC from 96.8 to 97.7%, comparing to that (87.8%) of the AllerStat model (Dimitrov, Bangov, et al., 2014; Dimitrov, Naneva, et al., 2014; Goto et al., 2023).

These results highlight the superiority of pLM-based models over the traditional modeling methods in allergenicity prediction.

The improvement in the model performance for the pLM4Alg architecture also partially resulted from the dataset size. Our dataset is larger compared to the AllerTOPv2 and AllergenFP dataset, which was only two-thirds of ours, and the AllerStat dataset was a highly imbalanced dataset with only 2248 allergenic sequences. For AlgPred 2.0, our dataset source is the same as that used in its machine learning model development regarding the databases. The best model in the AlgPred 2.0 papers is a hybrid model, integrating a k-nearest neighbors (KNN) model based on traditional embeddings (AAC and PSSM), along with two approaches for mapping IgE epitopes for allergen prediction: basic local alignment search tool (BLAST)-based searching and motif-emerging and with classes-identification (MERCI) (Sharma, Patiyal, Dhall, Pande, et al., 2021). With an additional boost from epitope mapping for classification, the performance in ACC increased from 84.1% under the solely AAC-based random forest model to 92.8% and achieved the SOTA performance at that time. In this study, we used a similar dataset as that in AlgPred 2.0 and achieved better performance in pLM4Alg models without the help of epitope mapping.

The primary distinction between pLM4Alg and previous approaches lies in the utilization of pLM. As discussed in the previous section, pLM demonstrated the highest prediction power based on our previous study comparing different peptide descriptors (Du et al., 2024). Furthermore, to the best of our knowledge, deep learning methods such as CNN have not been used in allergenicity prediction until now, which contributed to the observed improvement in performance. In the original pLM papers, a multilayer perceptron (MLP) was suggested and used in downstream tasks (Elnaggar et al., 2022; Lin et al., 2023). In this study, we introduced a convolutional layer before the dense layers by employing CNN, which allows to extract more

patterns from the embeddings, as this operation yielded a promising performance in our previous attempts (Du, Ding, et al., 2023). Traditional ML methods (e.g., random forest, support vector machine, k-nearest neighboring, Naïve Bayes, etc.) have been the mainstream approaches in previous studies on allergenicity prediction (Dimitrov, Bangov, et al., 2014; Dimitrov, Naneva, et al., 2014, 2014; Goto et al., 2023; Nedyalkova et al., 2023; Sharma, Patiyal, Dhall, Pande, et al., 2021; L. Wang et al., 2021; Yu et al., 2023). Deep learning methods generally have higher model complexity than traditional ML modeling methods. As allergy research gains increasing attention, it is anticipated that the dataset sizes will continuously grow in the future. Consequently, deep learning methods are expected to be extensively exploited in predicting allergenic sequences and achieve higher performance in the future.

In addition to its superior model performance, pLM4Alg stands out as an integrated prediction model capable of handling both peptides and proteins, including sequences containing absent residues or non-standard amino acids, a feature unavailable in existing predictors to date. However, a limitation of pLM4Alg is the length restriction of input protein sequences. When protein sequences exceed 900 residues, the computational requirements for embeddings become enormous, surpassing the capabilities of our test setup with 80 GB RAM and 40 GB GPU memory. Meanwhile, due to the limited sequence availability, sequences larger than 900 residues were excluded in our study. Therefore, pLM4Alg is not recommended for prediction tasks which involve the long sequence (>900 residues) proteins. Since the traditional sequence embedding methods have significantly lower computational demand than pLMs, they are not limited by this constraint (Dimitrov, Bangov, et al., 2014; Dimitrov, Naneva, et al., 2014; Nedyalkova et al., 2023; Sharma, Patiyal, Dhall, Pande, et al., 2021; Yu et al., 2023).

AllerCatPro 2.0 is another representative allergenicity prediction tool relying on the similarity in protein sequences and 3D structures between input sequences and its self-constructed allergenic sequence datasets (Nguyen et al., 2022). AllerCatPro 2.0 could predict weak, strong, or no evidence of allergenic potential, while pLM4Alg are binary classification models. However, it should be noted that there was no machine learning involved in the prediction-making process of AllerCatPro 2.0. Most of the benchmark datasets in AllerCarPro 2.0 were also included in our dataset construction and used for model training or evaluation. Thus, directly comparing the performance of our models with AllerCatPro 2.0 may not be entirely fair or necessary. According to the dataset construction of AllerCatPro 2.0, it may not be suitable for sequence prediction shorter than 50 residues or containing non-standard amino acids. As both tools possess unique capabilities, pLM4Alg and AllerCatPro 2.0 can be employed jointly to meet the diverse needs for identification of an unknown allergen and risk assessment.

6.4 Conclusions

The growing demand for identifying new allergens and allergenic risk assessment drives the development of efficient and accurate *in silico* prediction models. In this study, the latest pre-trained protein language models are introduced to develop an allergenic sequence prediction model. A systematic comparison was made on the model performance based on different sizes of pLMs and fine-tuning with CNN. The proposed models, pLM4Alg, achieved state-of-the-art performance in the field of machine learning-based allergenic sequence prediction. A positive relationship between the prediction performance and the size of pLMs was observed. Additionally, we constructed the latest allergenic sequence dataset (Supplementary Document S1) by collecting from well-known and peer-reviewed databases. To make our newly developed pLM-based models

easily accessible to the scientific and industrial communities, we deployed them on a user-friendly and free-access web server (https://f6wxpfd3sh.us-east-1.awsapprunner.com/).

Declaration of interest

The authors declare that there are no known conflicts of interest.

Acknowledgements

This is contribution No.24-055-J from the Kansas Agricultural Experimental Station. This work was supported in part by the USDA Pulse Crop Health Initiative projects (Grant Accession No. 0439200), the USDA Agricultural Research Service Non-Assistance Cooperative Project (Grant Accession No. 443993), and the USDA National Institute of Food and Agriculture Hatch project (Grant Accession No. 7003330). Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

References

Akdis, C. A. (2006). Allergy and hypersensitivity: Mechanisms of allergic disease. *Current Opinion in Immunology*, *18*(6), 718–726. https://doi.org/10.1016/j.coi.2006.09.016

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., & Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, *16*(12), Article 12. https://doi.org/10.1038/s41592-019-0598-1

Behbahani, M., Rabiei, P., & Mohabatkar, H. (2020). A Comparative Analysis of Allergen Proteins between Plants and Animals Using Several Computational Tools and Chou's PseAAC Concept. *International Archives of Allergy and Immunology*, *181*(11), 813–821. https://doi.org/10.1159/000509084

Bozzetto, S., Carraro, S., Giordano, G., Boner, A., & Baraldi, E. (2012). Asthma, allergy and respiratory infections: The vitamin D hypothesis. *Allergy*, *67*(1), 10–17. https://doi.org/10.1111/j.1398-9995.2011.02711.x

Breiteneder, H., Diamant, Z., Eiwegger, T., Fokkens, W. J., Traidl-Hoffmann, C., Nadeau, K., O'Hehir, R. E., O'Mahony, L., Pfaar, O., Torres, M. J., Wang, D. Y., Zhang, L., & Akdis, C. A. (2019). Future research trends in understanding the mechanisms underlying allergic diseases for improved patient care. *Allergy*, *74*(12), 2293–2311. https://doi.org/10.1111/all.13851

Celebi Sozener, Z., Ozdel Ozturk, B., Cerci, P., Turk, M., Gorgulu Akin, B., Akdis, M., Altiner, S., Ozbey, U., Ogulur, I., Mitamura, Y., Yilmaz, I., Nadeau, K., Ozdemir, C., Mungan, D., & Akdis, C. A. (2022). Epithelial barrier hypothesis: Effect of the external exposome on the microbiome and epithelial barriers in allergic disease. *Allergy*, *77*(5), 1418–1449. https://doi.org/10.1111/all.15240

Dang, H. X., & Lawrence, C. B. (2014). Allerdictor: Fast allergen prediction using text classification techniques. *Bioinformatics (Oxford, England)*, *30*(8), 1120–1128. https://doi.org/10.1093/bioinformatics/btu004

Dimitrov, I., Bangov, I., Flower, D. R., & Doytchinova, I. (2014). AllerTOP v.2—A server for in silico prediction of allergens. *Journal of Molecular Modeling*, *20*(6), 2278. https://doi.org/10/gm593m

Dimitrov, I., Naneva, L., Doytchinova, I., & Bangov, I. (2014). AllergenFP: Allergenicity prediction by descriptor fingerprints. *Bioinformatics*, *30*(6), 846–851. https://doi.org/10.1093/bioinformatics/btt619

Du, Z., Comer, J., & Li, Y. (2023). Bioinformatics approaches to discovering food-derived bioactive peptides: Reviews and perspectives. *TrAC Trends in Analytical Chemistry*, 117051. https://doi.org/10.1016/j.trac.2023.117051

Du, Z., Ding, X., Hsu, W., Munir, A., Xu, Y., & Li, Y. (2024). pLM4ACE: A protein language model based predictor for antihypertensive peptide screening. *Food Chemistry*, *431*, 137162. https://doi.org/10.1016/j.foodchem.2023.137162

Du, Z., Ding, X., Xu, Y., & Li, Y. (2023). UniDL4BioPep: A universal deep learning architecture for binary classification in peptide bioactivity. *Briefings in Bioinformatics*, 1–10. https://doi.org/10.1093/bib/bbad135

Du, Z., & Li, Y. (2022). Review and perspective on bioactive peptides: A roadmap for research, development, and future opportunities. *Journal of Agriculture and Food Research*, *9*, 100353. https://doi.org/10.1016/j.jafr.2022.100353

Du, Z., Wang, D., & Li, Y. (2022). Comprehensive Evaluation and Comparison of Machine Learning Methods in QSAR Modeling of Antioxidant Tripeptides. *ACS Omega*, acsomega.2c03062. https://doi.org/10.1021/acsomega.2c03062

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., & Rost, B. (2022). ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381

FAO/WHO. (2001). Evaluation of Allergenicity of Genetically Modified Foods: Report of a Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology. *Rome, Italy,* 22-25.

Goodman, R. E., Ebisawa, M., Ferreira, F., Sampson, H. A., Van Ree, R., Vieths, S., Baumert, J. L., Bohle, B., Lalithambika, S., Wise, J., & Taylor, S. L. (2016). AllergenOnline: A peer-reviewed, curated allergen database to assess novel food proteins for potential cross-reactivity. *Molecular Nutrition & Food Research*, *60*(5), 1183–1198. https://doi.org/10.1002/mnfr.201500769

Goto, K., Tamehiro, N., Yoshida, T., Hanada, H., Sakuma, T., Adachi, R., Kondo, K., & Takeuchi, I. (2023). Novel Machine Learning Method AllerStat Identifies Statistically Significant Allergen-Specific Patterns in Protein Sequences. *Journal of Biological Chemistry*, 104733. https://doi.org/10.1016/j.jbc.2023.104733

Han, X., Krempski, J. W., & Nadeau, K. (2020). Advances and novel developments in mechanisms of allergic inflammation. *Allergy*, *75*(12), 3100–3111. https://doi.org/10.1111/all.14632

Islam, M. R., Kashem, M. A., & Mia, L. (2021). AllerHybrid: A Hybrid System to Predict the Allergen Using K-mer and Physicochemical Properties. *2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, 1–6. https://doi.org/10.1109/ICEEICT53905.2021.9667943

Lambrecht, B. N., & Hammad, H. (2017). The immunology of the allergy epidemic and the hygiene hypothesis. *Nature Immunology*, *18*(10), 1076–1083. https://doi.org/10.1038/ni.3829

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, *379*(6637), 1123–1130. https://doi.org/10.1126/science.ade2574

Mari, A., Rasi, C., Palazzo, P., & Scala, E. (2009). Allergen databases: Current status and perspectives. *Current Allergy and Asthma Reports*, *9*(5), 376–383. https://doi.org/10.1007/s11882-009-0055-9

Maurer-Stroh, S., Krutz, N. L., Kern, P. S., Gunalan, V., Nguyen, M. N., Limviphuvadh, V., Eisenhaber, F., & Gerberick, G. F. (2019). AllerCatPro—Prediction of protein allergenicity potential from the protein sequence. *Bioinformatics*, *35*(17), 3020–3027. https://doi.org/10.1093/bioinformatics/btz029

Naneva, L., Nedyalkova, M., Madurga, S., Mas, F., & Simeonov, V. (2019). Applying Discriminant and Cluster Analyses to Separate Allergenic from Non-allergenic Proteins. *Open Chemistry*, *17*(1), 401–407. https://doi.org/10.1515/chem-2019-0045

Nedyalkova, M., & Simeonov, V. (2020). Multivariate Chemometrics as a Strategy to Predict the Allergenic Nature of Food Proteins. *Symmetry*, *12*(10), Article 10. https://doi.org/10.3390/sym12101616

Nedyalkova, M., Vasighi, M., Azmoon, A., Naneva, L., & Simeonov, V. (2023). Sequence-Based Prediction of Plant Allergenic Proteins: Machine Learning Classification Approach. *ACS Omega*, *8*(4), 3698–3704. https://doi.org/10.1021/acsomega.2c02842

Nguyen, M. N., Krutz, N. L., Limviphuvadh, V., Lopata, A. L., Gerberick, G. F., & Maurer-Stroh, S. (2022). AllerCatPro 2.0: A web server for predicting protein allergenicity potential. *Nucleic Acids Research*, *50*(W1), W36–W43. https://doi.org/10.1093/nar/gkac446

Ogulur, I., Pat, Y., Ardicli, O., Barletta, E., Cevhertas, L., Fernandez-Santamaria, R., Huang, M., Bel Imam, M., Koch, J., Ma, S., Maurer, D. J., Mitamura, Y., Peng, Y., Radzikowska, U., Rinaldi, A. O., Rodriguez-Coira, J., Satitsuksanoa, P., Schneider, S. R., Wallimann, A., … Akdis, C. A. (2021). Advances and highlights in biomarkers of allergic diseases. *Allergy*, *76*(12), 3659–3686. https://doi.org/10.1111/all.15089

Pomés, A., Davies, J. M., Gadermaier, G., Hilger, C., Holzhauser, T., Lidholm, J., Lopata, A., Mueller, G. A., Nandy, A., Radauer, C., Chan, S. K., Jappe, U., Kleine-Tebbe, J., Thomas, W. R., Chapman, M. D., van Hage, M., van Ree, R., Vieths, S., Raulf, M., & Goodman, R. E. (2018). WHO/IUIS Allergen Nomenclature: Providing a common language. *Molecular Immunology*, *100*, 3–13. https://doi.org/10.1016/j.molimm.2018.03.003

Radauer, C., Bublin, M., Wagner, S., Mari, A., & Breiteneder, H. (2008). Allergens are distributed into few protein families and possess a restricted number of biochemical functions. *Journal of Allergy and Clinical Immunology*, *121*(4), 847-852.e7. https://doi.org/10.1016/j.jaci.2008.01.025

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, *118*(15), e2016239118. https://doi.org/10.1073/pnas.2016239118

Rothenberg, M. E. (2022). The climate change hypothesis for the allergy epidemic. *Journal of Allergy and Clinical Immunology*, *149*(5), 1522–1524. https://doi.org/10.1016/j.jaci.2022.02.006

Sathe, S. K., Liu, C., & Zaffran, V. D. (2016). Food Allergy. *Annual Review of Food Science and Technology*, *7*, 191–220. https://doi.org/10.1146/annurev-food-041715-033308

Scheurer, S., Toda, M., & Vieths, S. (2015). What makes an allergen? *Clinical and Experimental Allergy: Journal of the British Society for Allergy and Clinical Immunology*, *45*(7), 1150–1161. https://doi.org/10.1111/cea.12571

Sharma, N., Patiyal, S., Dhall, A., Devi, N. L., & Raghava, G. P. S. (2021). ChAlPred: A web server for prediction of allergenicity of chemical compounds. *Computers in Biology and Medicine*, *136*, 104746. https://doi.org/10.1016/j.compbiomed.2021.104746

Sharma, N., Patiyal, S., Dhall, A., Pande, A., Arora, C., & Raghava, G. P. S. (2021). AlgPred 2.0: An improved method for predicting allergenic proteins and mapping of IgE epitopes. *Briefings in Bioinformatics*, *22*(4), bbaa294. https://doi.org/10.1093/bib/bbaa294

The UniProt Consortium. (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*(D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100

van Ree, R., Sapiter Ballerda, D., Berin, M. C., Beuf, L., Chang, A., Gadermaier, G., Guevera, P. A., Hoffmann-Sommergruber, K., Islamovic, E., Koski, L., Kough, J., Ladics, G. S., McClain, S., McKillop, K. A., Mitchell-Ryan, S., Narrod, C. A., Pereira Mouriès, L., Pettit, S., Poulsen, L. K., … Bowman, C. (2021). The COMPARE Database: A Public Resource for Allergen Identification, Adapted for Continuous Improvement. *Frontiers in Allergy*, *2*. https://doi.org/10.3389/falgy.2021.700533

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Vlieg-Boerstra, B., Groetch, M., Vassilopoulou, E., Meyer, R., Laitinen, K., Swain, A., Durban, R., Benjamin, O., Bottse, R., Grimshaw, K., Netting, M., O'Mahony, L., de Jong, N., & Skypala, I. J. (2023). The immune-supportive diet in allergy management: A narrative review and proposal. *Allergy*, *78*(6), 1441–1458. https://doi.org/10.1111/all.15687

Wang, J., Yu, Y., Zhao, Y., Zhang, D., & Li, J. (2013). Evaluation and integration of existing methods for computational prediction of allergens. *BMC Bioinformatics*, *14*(4), S1. https://doi.org/10.1186/1471-2105-14-S4-S1

Wang, L., Niu, D., Zhao, X., Wang, X., Hao, M., & Che, H. (2021). A Comparative Analysis of Novel Deep Learning and Ensemble Learning Models to Predict the Allergenicity of Food Proteins. *Foods*, *10*(4), Article 4. https://doi.org/10.3390/foods10040809

Westerhout, J., Krone, T., Snippe, A., Babé, L., McClain, S., Ladics, G. S., Houben, G. F., & Verhoeckx, K. CM. (2019). Allergenicity prediction of novel and modified proteins: Not a mission impossible! Development of a Random Forest allergenicity prediction model. *Regulatory Toxicology and Pharmacology*, *107*, 104422. https://doi.org/10.1016/j.yrtph.2019.104422

Yu, X.-X., Liu, M.-Q., Li, X.-Y., Zhang, Y.-H., & Tao, B.-J. (2023). Qualitative and quantitative prediction of food allergen epitopes based on machine learning combined with in vitro experimental validation. *Food Chemistry*, *405*, 134796. https://doi.org/10.1016/j.foodchem.2022.134796

Figures and Tables:

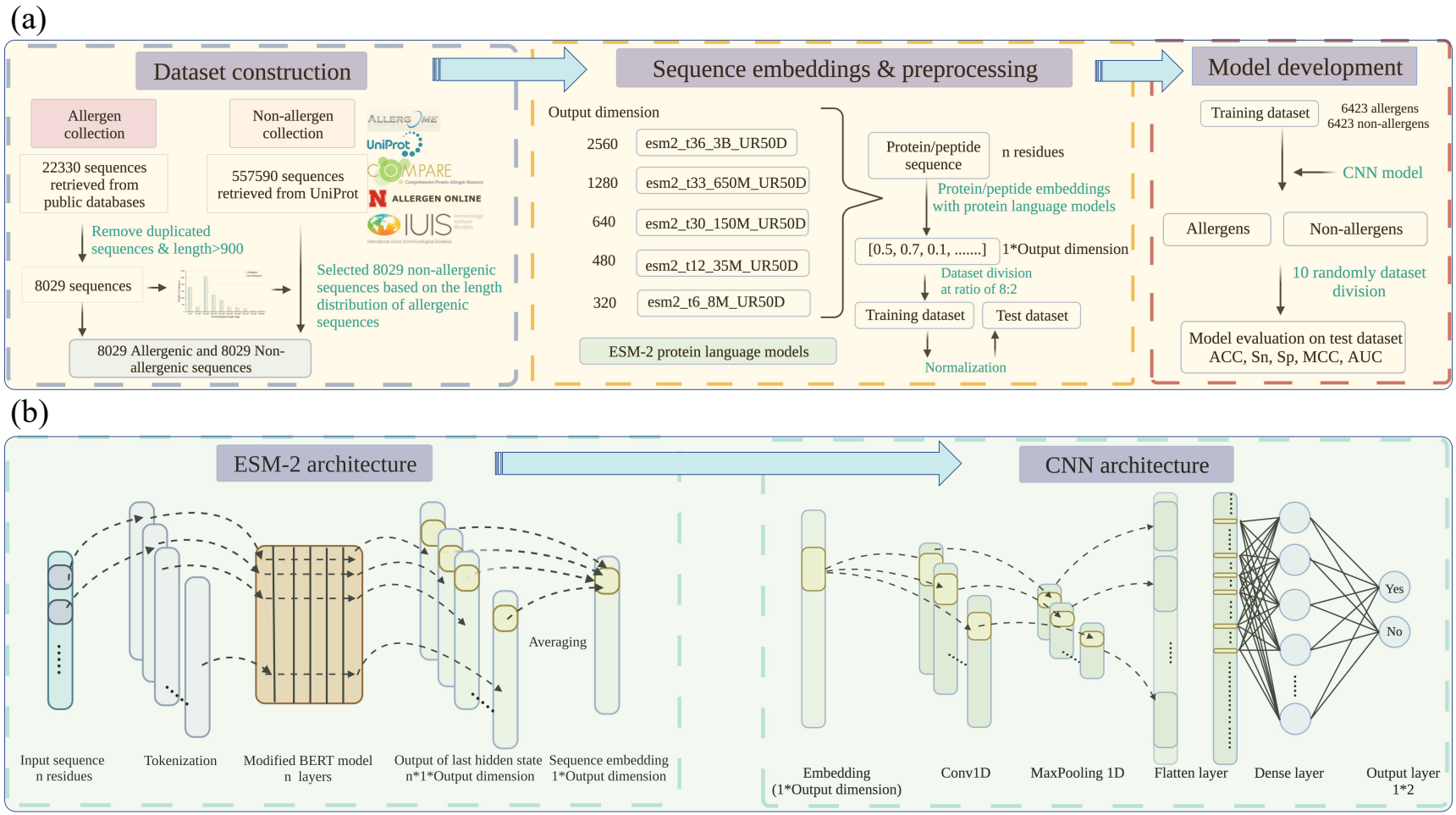


Figure 6-1 Schematic framework of pLM4Alg model development (a) and detailed architecture of ESM-2 protein language models and convolutional neural network (CNN) models (b).

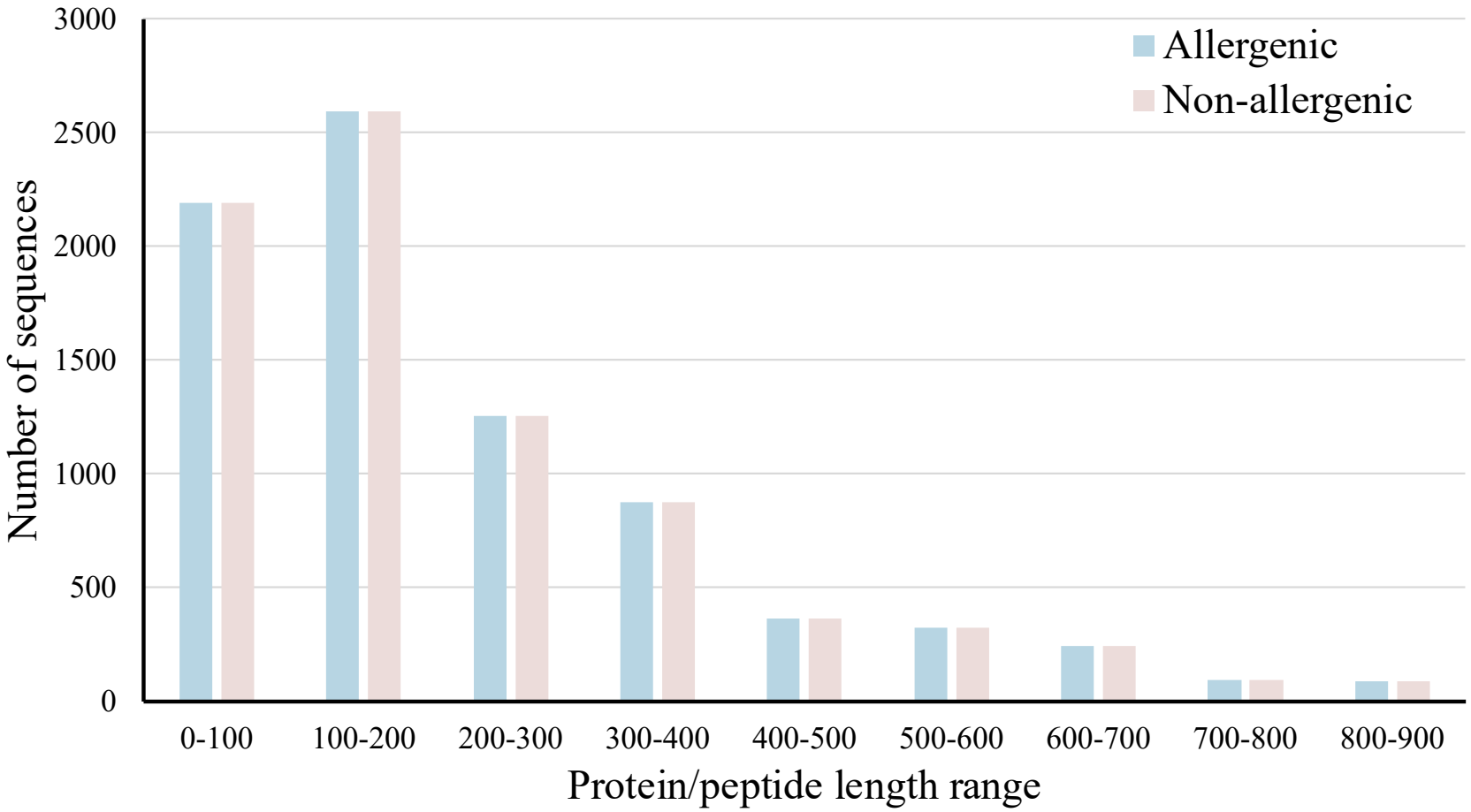


Figure 6-2 Dataset profile of protein/peptide length distribution.

Note: There are 1816 sequences with less than 50 residues in the 0-100 range.

Table 6-1 Available ESM-2 pretrained models.

Pretrained model name	Number of layers	Number of parameters	Output dimensions
esm2_t48_15B_UR50D*	48	15 billion	5120
esm2_t36_3B_UR50D	36	3 billion	2560
esm2_t33_650M_UR50D	33	650 million	1280
esm2_t30_150M_UR50D	30	150 million	640
esm2_t12_35M_UR50D	12	35 million	480
esm2_t6_8M_UR50D	6	8 million	320

Note: * This pretrained model was not selected because of the excessive computation demands in sequence embeddings.

Table 6-2 Comparison between pLM4Alg and machine learning-based allergenic sequence prediction models

Project name*	Dataset	ACC (%)	Sn (%)	Sp (%)	MCC	AUC	Web server working	Year release	Ref.
pLM4Alg -2560	16058	95.1±0.4	94.2±0.7	96.0±0.4	0.902±0.008	0.99±0.001	Yes	2023	Our model
pLM4Alg -1280	16058	94.7±0.4	93.5±0.6	95.9±0.4	0.894±0.009	0.988±0.001	Yes	2023	Our model
pLM4Alg -640	16058	94.4±0.4	93.3±0.6	95.4±0.4	0.888±0.008	0.986±0.002	Yes	2023	Our model
pLM4Alg -480	16058	94.0±0.4	93.0±0.8	95.1±0.5	0.881±0.008	0.985±0.002	Yes	2023	Our model
pLM4Alg -320	16058	93.4±0.3	91.9±0.6	94.9±0.6	0.869±0.006	0.981±0.001	Yes	2023	Our model
AlgPred 2.0	20150	92.7	94.04	91.46	0.86	0.97	No	2021	(Sharma, Patiyal, Dhall, Pande, et al., 2021)
AllerStat**	21154	N/A	N/A	N/A	0.4495	0.878	No	2023	(Goto et al., 2023)
Not specified-1	1183	93.1±1.5	N/A	N/A	N/A	0.9578	No	2021	(L. Wang et al., 2021)
Not specified-2	525745	88	91	87	0.32	N/A	No	2019	(Westerhout et al., 2019)
Not specified-3	954	93	90	95	0.86	0.98	No	2023	(Nedyalkova et al., 2023)
Not specified-4	2000	69.66	N/A	N/A	N/A	0.764	No	2023	(Yu et al., 2023)

Not specified-5***	178	88.24	83.82	90.91	0.74	N/A	No	2020	(Behbahani et al., 2020)
Allerdictor**	42977	N/A	N/A	N/A	N/A	0.969±0.005	No	2014	(Dang & Lawrence, 2014)
AllerTOP v2	4854	88.7	86.7	90.7	0.775	N/A	Yes	2014	(Dimitrov, Bangov, et al., 2014)
AllergenFP	4854	87.9	86.8	89.1	0.759	N/A	Yes	2014	(Dimitrov, Naneva, et al., 2014)

Note: N/A: not available; ACC: accuracy; Sn: sensitivity; Sp: specificity; MCC: Matthews correlation coefficient; AUC: Area under the curve; *The number after the hyphen sign represents output dimension of the specific pretrained ESM-2 pLMs (e.g., pLM4Alg-320* represents the model that was developed based on pretrained pLM esm2_t6_8M_UR50D where the output dimension after embeddings is 320.); **Imbalanced dataset (3907 allergenic sequences and 39070 nonallergenic sequences for Allerdictor; 2248 allergenic sequences and 18906 nonallergenic sequences for AllerStat; 583 allergenic sequences and 600 nonallergenic sequences for the Not specified-1; 1673 allergenic sequences and 524072 nonallergenic sequences for the Not specified-2); This model is designed to predict the animal or plant source allergens.

Chapter 7 - FusionESP: improved enzyme-substrate pair prediction by fusing protein and chemical knowledge*

Abstract

To reduce the cost of experimental characterization of the potential substrates for enzymes, machine learning prediction model offers an alternative solution. Pretrained language models, as powerful approaches for protein and molecule representation, have been employed in the development of enzyme-substrate prediction models, achieving promising performance. In addition to continuing improvements in language models, effectively fusing encoders to handle multimodal prediction tasks is critical for further enhancing model performance using available representation methods. Here, we present FusionESP, a multimodal architecture that integrates protein and chemistry language models with a newly designed contrastive learning strategy for predicting enzyme-substrate pairs. Our best model achieved state-of-the-art performance with an accuracy of 94.77% on independent test data and exhibited better generalization capacity while requiring fewer computational resources and training data, compared to previous studies of finetuned encoder or employing more encoders. It also confirmed our hypothesis that embeddings of positive pairs are closer to each other in high-dimension space, while negative pairs exhibit the opposite trend. The proposed architecture is expected to be further applied to enhance performance in additional multimodality prediction tasks in biology. A user-friendly web server of FusionESP is established and freely accessible at https://rqkjkgpsyu.us-east-1.awsapprunner.com/.

Keywords: Large language model; multimodal model; protein-compound interaction; contrastive learning

---

* Du, Z., Fu, W., Guo, X., Caragea, D., & Li, Y. (2024). FusionESP: Improved enzyme-substrate pair prediction by fusing protein and chemical knowledge. bioRxiv, 2024-08. https://doi.org/10.1101/2024.08.13.607829

7.1 Introduction

Most enzymes are proteins capable of catalyzing a wide range of reactions within living organisms or under mild conditions *in vitro* with up to over a million-fold compared to spontaneous rates (Kim et al., 2023; Kroll, Ranjan, Engqvist, et al., 2023). Moreover, enzymes typically exhibit promiscuity, facilitating multiple reactions that may include physiologically irrelevant or potentially harmful processes (Copley, 2017; Nobeli et al., 2009). A comprehensive mapping of enzyme-substrate relationships will provide crucial guidance for future research in medicine, pharmaceuticals, bioengineering, and agriculture (Kroll, Ranjan, & Lercher, 2023; D. Zhang et al., 2023, 2024). However, it is prohibitively expensive to experimentally determine the catalytic interactions between molecules and enzymes. According to the UniProt Knowledgebase, approximately 10.8 million entries are related to enzymes, yet only 0.6% of these sequences have high-quality annotations of catalyzed reactions that are manually curated (The UniProt Consortium, 2021). There is a pressing need for high-throughput methods to address the scarcity of experimental validation in enzyme-substrate relationships.

Since the introduction of transformer architecture (Vaswani et al., 2017), various machine learning tasks have been revolutionized by leveraging pretrained large language models (LLMs) for molecular and protein representation, achieving significant advancements (Ahmad et al., 2022; Akdis, 2021; Du, Ding, et al., 2023; Lin et al., 2023; Ross et al., 2022, 2022; Q. Zhang et al., 2024). Molecules and proteins can be represented as strings similar to natural language, using vocabularies such as the simplified molecular-input line-entry system (SMILES) for molecules and single-letter amino acid codes for proteins. Besides, protein and molecular databases (e.g., UniProt and ZINC) contained billions of protein sequences and molecular data, allowing to pretrain protein/chemistry LLMs. Based on their architecture, LLMs can be classified into

encoder-only models (e.g., BERT - bidirectional encoder representations from transformers), decoder-only models (e.g., GPT - generative pre-trained transformers), and encoder-decoder models (e.g., T5 - text-to-text translation transformer). Among these, encoder-only architectures have gained popularity for remarkable performance in various downstream classification and regression tasks through transfer learning or fine-tuning (Ahmad et al., 2022; Du, Ding, et al., 2023; Kroll, Ranjan, & Lercher, 2023; Ross et al., 2022; Thumuluri et al., 2022; Q. Zhang et al., 2024). Recently, machine learning approaches for predicting compound-protein interaction (CPI) have increasingly adopted these encoder-only models (Nguyen et al., 2023, 2024; Song et al., 2023; Wang et al., 2022), and enzyme-substrate pair prediction task is a subset of the CPI prediction task (Kroll, Ranjan, Engqvist, et al., 2023; Kroll, Ranjan, & Lercher, 2023; Mou et al., 2021; Yang et al., 2018). Compared to traditional approaches using physicochemical parameters of molecules, enzyme sequences, or enzyme active site descriptors, employing pretrained BERT style encoders allows for better representation of proteins and molecules due to the domain knowledge learned via self-supervised learning from enormous datasets (Du, Ding, et al., 2023; Kroll, Ranjan, Engqvist, et al., 2023; Lim et al., 2021; Mou et al., 2021; Yang et al., 2018; Q. Zhang et al., 2024).

Enzyme substrate pairs prediction model is a multimodal prediction task, where two different “language” systems (i.e., amino acid sequences and SMILES) are involved during the embedding process. How to effectively leverage the enzyme embeddings and molecular embeddings generated from corresponding encoders becomes a crucial point in model performance enhancement. The most popular strategy is to simply concatenate the protein and molecule embeddings and this strategy and its modifications has been a common practice in the CPI discovery community (Lee et al., 2019; Nguyen et al., 2024; Wang et al., 2022). So far, the

SOTA performance in enzyme-substrate pair prediction was achieved with the simple concatenation strategy in the study of Kroll et al., where a task-specific fine-tuned protein language model (PLM) (i.e., ESM-1b) and a pretrained domain-specific graph neural network (GNN) was used for embeddings and obtained an accuracy of 91.5% (Kroll, Ranjan, Engqvist, et al., 2023). Similar strategy was also used in the study of Song et al., where Kronecker products of protein and molecule embeddings as additional features were concatenated with the original embeddings to enhance CPI prediction performance (Song et al., 2023). Beyond simple concatenation, cross-attention blocks have been employed to integrate local and high-level information from multiple protein and molecule encoders, capturing the relationship between protein and molecule representations for fused representation in CPI prediction (Nguyen et al., 2023, 2024). Recently, Kroll et al. introduced a multimodal BERT model for embedding protein-molecule complex, where the protein sequences and SMILES were input to a multimodal BERT model. They achieved SOTA performance across four datasets, including drug-target interactions, protein-small molecule interactions, enzyme-substrate Michaelis constants (KM), and substrate identification for enzymes, by concatenating PLM-generated protein embeddings, chemically learned molecule (CLM)-generated molecule embeddings, and multimodal encoder-generated protein-molecule embeddings (Kroll, Ranjan, & Lercher, 2023). To maximize the prediction performance, these models either employed more encoders to enrich embedding protein and molecule or fused embeddings to enhance enzyme-substrate complex representations. However, these methods often require either complex fusion architecture design and substantial computational resources or additional datasets for finetuning, which impeded the deployment and application to different scenarios.

Inspired by the success of Contrastive Language-Image Pre-training (CLIP), where two independent encoders for images and texts were jointly trained to predict 400 million correct image-text pairs (Radford et al., 2021), we hypothesized that enzyme and substrate in a correct pair should be closer in a high dimensional space after projection, while unrelated components should be distinctly separated. This concept has been explored in single modality settings. To predict peptide-receptor complex pair, ESM-1b and ESM-MSA-1b were connected with independent encoder for peptide and protein receptor embeddings, respectively. These additional peptide and protein receptor encoders were jointly trained to capture inter-embedding similarities using a CLIP style model architecture (Bhat et al., 2023; Palepu et al., 2022). In contrast, Yu et al., (2023) took a different approach by designing a project head to refine embeddings without training additional encoders, thus reducing computation resource consumption. This strategy was employed in the single modality task of enzyme commission (EC) number prediction, where a frozen ESM-1b was used for enzyme embeddings, and a simple projection head was integrated with a contrastive loss function to refine embeddings (Yu et al., 2023). This approach achieved state-of-the-art (SOTA) performance across various benchmark datasets, particularly excelling in rare EC number predictions.

To the best of our knowledge, our model, FusionESP, is the first attempt to employ contrastive learning concept to address multimodality prediction tasks in biological science domain. Our objective was to develop an advanced machine learning module tailored for empowering existing LLMs to predict enzyme-substrate relationships across a wide range of proteins and molecules. This module aims to serve as a practical tool to streamline experimental processes and boost laboratory efficiency. In this study, we designed a simplified yet highly effective model architecture that employs a contrastive learning strategy. The architecture

demonstrated remarkable performance enhancement in knowledge fusion within a multimodal context, even with limited data. Specifically, we utilized ESM-2 for enzyme embeddings and MolFormer for molecule embeddings. Rather than fine-tuning these encoders, adding additional encoders, or incorporating more features (e.g., extended-connectivity fingerprints), we designed projection layers for both encoders to refine and align the embeddings in the same high-dimensional space (Fig. 1). Our model outperforms previous approaches with a simplified architecture and reduced computational demands during both training and inference phases.

7.2 Methods

7.2.1 Dataset

To ensure a fair comparison with previous studies, the datasets used in this study were sourced from existing studies (Kroll, Ranjan, Engqvist, et al., 2023; Kroll, Ranjan, & Lercher, 2023). The positive enzyme-substrate pairs were from Gene Ontology (GO) annotation database, where entries have different levels of evidence: experimental, phylogenetically-inferred, computational analysis, author statement, curator statement, and electronic evidence. We utilized two datasets: one based on experimental evidence and the other on phylogenetic evidence. To challenge the model to distinguish true from false substrates, negative pairs were generated by randomly sampling three small molecules highly similar to the true substrates for the same enzyme sequences (Kroll, Ranjan, Engqvist, et al., 2023).

The experimental evidence-based dataset, originally split into training, validation, and test sets in the original study (Kroll, Ranjan, Engqvist, et al., 2023; Kroll, Ranjan, & Lercher, 2023), contained 50,093 enzyme-substrate pairs in the training set, 5,422 pairs in the validation set, and 13,336 pairs in the test set. In contrast, the phylogenetic evidence-based dataset comprised a total of 765,635 pairs. These two datasets were downloaded from

https://github.com/AlexanderKroll/ESP and https://github.com/AlexanderKroll/ProSmith, respectively.

Because of the computational demands of processing long protein sequences, we excluded some positive or negative enzyme-substrate pairs whose sequence embeddings exceeded hardware capacity (NVIDIA A100 GPU). Specifically, for ESM-2 models with output dimensions of 480 and 1280, sequences longer than 8000 were removed. Sixteen pairs were removed from the training dataset and no pair was removed from the test dataset. For the ESM-2 model with a 2560-dimensional output, sequences longer than 5500 were excluded. Correspondingly, 212 pairs were removed from the training dataset. Notably, no sequence in the validation and test datasets was removed.

7.2.2 Calculating enzyme representation

In this study, enzymes were represented numerically using ESM2 models (Lin et al., 2023; Rives et al., 2021). ESM is a PLM project, initiated by Meta Fundamental AI Research (FAIR) in 2019 (https://github.com/facebookresearch/esm), which includes 19 pretrained PLMs with various dimensional output embeddings based on a modified Bidirectional Encoder Representations from Transformers (BERT) architecture. Those PLMs were trained on large protein sequence datasets (e.g., UR50/D 2021_04) through self-supervised learning. In this study, we employed three ESM-2 models, including the 480-dimensional output, 1280-dimensional output, and 2560-dimensional output, denoted as ESM-2-480, ESM-2-1280, and ESM-2-2560, respectively. The details about each PLM are listed in Supplementary Document S1-Table S1. The amino acid one letter code of each enzyme was loaded into the PLM for embeddings. For an enzyme with L amino acids, the output of the ESM-2-n model for the enzymes is a $(L+1)*n$ matrix, where n represents the output dimension of the ESM-2 model. The

first row is the representation of the [sos] token, which is the indicator of "start of sentence" during training process. After that, each row in the matrix represented the corresponding amino acid from N-terminal to C-terminal with the consideration of the amino acid itself and the context of the entire sequence. We used an average pooling operation to unify the output dimension of enzymes with different lengths. Finally, any enzyme sequence loaded into the ESM-2-*n* for embeddings would result in a 1***n* dimensional vector as its representation.

7.2.3 Calculating molecule representation

MoLFormer was selected for numerical representation generation of molecules (Ross et al., 2022). MoLFormer was developed by IBM Research in 2022 and trained on canonical SMILES sequences of 1 billion molecules from ZINC database and 111 million molecules from PubChem with employment of rotary positional embeddings and linear attention mechanism. There was only one model available, called "MoLFormer-XL-both-10pct" trained with 10% of both datasets, at https://huggingface.co/ibm/MoLFormer-XL-both-10pct. Although the dataset size was smaller, this model demonstrated performance comparable to the full-size model. Therefore, we used this model for the molecule embeddings. We will use "MoLFormer" to refer this model in this paper. MoLFormer has a BERT-like architecture, similar to that of ESM models. To encode a canonical SMILES, the SMILES sequence was first converted into a numerical matrix, and then an average pooling operation was conducted to unify the feature dimension of SMILES sequences of different length into a 1*768 dimensional vector.

7.2.4 Encoding models

ESM-2 models represent an upgrade over ESM-1b used in previous studies (Kroll, Ranjan, Engqvist, et al., 2023; Kroll, Ranjan, & Lercher, 2023). They were trained on a new dataset (UR50/D 2021_04), while ESM-1b model was trained on UR50/S 2018_03 (Lin et al.,

2023; Rives et al., 2021). Besides, the ESM-2 models employed rotary position embedding to replace the learned position embedding, which allowed it to embed any length of protein sequences, while ESM-1b can only embed sequences shorter than 1023 amino acids (Lin et al., 2023; Rives et al., 2021). As for MoLFormer, it was trained on a larger dataset than the one used in previous study (i.e., ChemBERTa-2) and employed more advanced algorithms (i.e., rotary position embedding and linear attention mechanism) (Kroll, Ranjan, & Lercher, 2023). It exhibited better performance in one regression task (Lipophilicity dataset) and three classification tasks (BACE, BBBP, and ClinTox datasets) during downstream task performance evaluation (Ahmad et al., 2022; Ross et al., 2022).

7.2.5 Model architecture design

The model architecture comprised two independent modules with similar inner architecture design for enzyme sequence input and molecule SMILES input, respectively (Fig. 1). The enzyme module utilized an ESM model as the encoder for enzyme sequence embeddings, followed by average pooling to unify the embedding dimensions. These embeddings were then processed through a projection head to refine the dimension from 480/1280/2560 to 128. Similarly, the molecule module employed MoLFormer as the encoder, followed by average pooling and a projection head to refine the dimension from 768 to 128. Every enzyme substrate pair was finally transformed into two 128 dimensional vectors, where one represented the enzymes and the other represented the molecule. Cosine similarity between the two refined embeddings was calculated finally as the interaction portability of the pair. The value of the positive enzyme-substrate pairs was 1 and the true value of the negative pair was 0. A mean squared error (MSE) loss function was employed for the loss calculation as following:

$$\boldsymbol{Loss} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{Sim_i} - \boldsymbol{Label_i})^{\mathbf{2}} \tag{7.1}$$

where n is the number of data points; Sim_i is the cosine similarity of the 128 dimensional vectors of enzyme and substrate in the single pair; the $Label_i$ is the true label.

The projection head mainly referred to the projection heads in contrastive learning studies with a few modifications (Chen et al., 2020; Grill et al., 2020; Richemond et al., 2020). The two projection heads for enzyme embeddings and molecular embeddings did not share weights and were trained simultaneously and independently through the MSE loss function except for the second batch norm layer. While the contrastive learning idea was inspired from CLIP (Radford et al., 2021), there was no huge dataset available for training, similar to the one that CLIP used to train the image and text encoders from scratch. Thus, in this study, we leveraged two pretrained encoders for embeddings to save the computational resources required for LLMs' training and bypass the need for a huge enzyme-substrate pair dataset. At the same time, we employed the knowledge/expert-based negative dataset to address the challenges of scarcity in negative data points. The encoders were frozen and two independent projection heads were trained jointly as the learnable adaptor to maximize the model performance.

For each projection head, the number of neurons in the first fully connected layer corresponded to the encoder's output dimension after average pooling. A batch normalization layer followed the fully connected layer before the ReLU activation function. Subsequently, the second fully connected layer (bottleneck layer) projected the input into a 128-dimensional vector, employing batch normalization and L2 normalization. The 128 dimension was selected mainly referring to SimCLR (Chen et al., 2020). The neurons in the first fully connected layer varied based on the encoder selected for model development. For instance, with the optimal combination of ESM-2-2560 and MoLFormer, the first fully connected layers had 2560 and 768 neurons for enzyme embeddings and molecule embeddings, respectively. Following projection

into a 128-dimensional vector, both enzyme and molecule vectors shared the same batch normalization layer, assuming that they were at the same high-dimensional space.

7.2.6 Model training

To expedite training and reduce computation needs, enzyme and molecule embeddings were pre-generated and saved for projection head training, bypassing the iterative generation of embeddings during training process. Adam optimizer was employed with default parameters. The batch size was 16 for the training on the experimental evidence-based datasets and 512 for the training on the phylogenetic evidence-based dataset. A relatively small batch size for experimental evidence-based datasets resulted from the small size of the dataset and the better capability of capturing the nuances, while a larger batch size was selected for the larger phylogenetic evidence-based dataset and captured the general features among positive/negative pairs.

The model trained on the experimental evidence-based dataset, denoted as FusionESP-exp, underwent 500 epochs, with the best-performing model checkpoint saved based on validation dataset performance. Training typically took 2-3 hours using a Tesla T4 GPU on Google Colab. The model trained on both the phylogenetic and experimental evidence-based datasets denoted FusionESP-XL, underwent similar processes, beginning with training on the phylogenetic evidence-based dataset for 500 epochs, followed by an additional 30 epochs on the experimental evidence-based dataset. The epoch numbers were selected based on learning curves (Supplementary Document S2-Figure S1). The model, after training on a phylogenetic evidence-based dataset, was noted as FusionESP-phylo. At each state, best model checkpoint was saved based on validation dataset performance during this extended training process.

7.2.7 Software

The neural network models were all implemented with Python and trained using the PyTorch library. The datasets and codes used to generate the results of this paper are available from https://github.com/dzjxzyd/FusionESP.

A user-friendly web server was deployed at Amazon Web Services (AWS) App Runner. The website was designed with html and css scripts, and the model deployment was achieved with Flask (2.2.2). Due to the constraint of computational resources, the FusionESP-XL with ESM-2-1280 and MolFormer was deployed for enzyme-substrate prediction. The web server supports large-scale processing, which allows users to upload their peptide information through xls or xlsx formats. The detailed scripts for web server development are available at https://github.com/dzjxzyd/FusionESP_server_1280.

7.3 Results and discussion

7.3.1 Contrastive learning strategy exhibits superiority over simple concatenation strategy

The dataset was divided into training, validation, and test sets as in the original paper (Kroll, Ranjan, & Lercher, 2023). Three different ESM-2 models and MoLFormer were employed as encoders for enzyme and molecule representations, respectively. FusionESP-exp was exclusively trained on the provided experimental evidence-based datasets. Although our training dataset was slightly smaller compared to those of previous models, all models, including ours, were evaluated on the same test dataset, ensuring a fair final performance comparison.

To assess the superiority of our contrastive learning strategy in enhancing model performance, we also implemented three counterpart baseline models with the popular concatenation strategy. We employed a simple two-layer neural network as the classifier to learn from the training data, details of which can be found in our GitHub repository. The results from all 6 models are shown in Table 1. All the models based on our contrastive learning strategy

outperformed their counterparts, especially the FusionESP-exp with ESM-2-480 and MoLFormer, where the accuracy was 6.7% higher than that of the model based on simple concatenation strategy with ESM-2-480 and MoLFormer for embeddings. This underscores the significant advantage of our proposed contrastive learning strategy over the widely used concatenation strategy. A frozen BERT-based encoder paired with learnable neural networks as a projection head and contrastive loss function has previously shown success in single-modality conditions for EC number prediction (Yu et al., 2023). In this study, we employed two independent projection heads and an MSE loss function referring to CLIP model for multimodal prediction task. The promising results validated our hypothesis that positive enzyme-substrate pairs tend to be closer in high-dimensional space, while negative pairs are more distant. Moreover, due to the small size of the projection heads, our approach required a relatively modest dataset size, making it well-suited for biological applications with limited data availability.

Additionally, a positive correlation between model performance and the size of ESM models/output dimensions was observed in both the contrastive learning and concatenation strategy groups. This trend aligns with findings from our previous studies (Du, Xu, et al., 2023), where larger model sizes/output dimensions encoded more information into the embeddings, resulting in improved overall performance.

7.3.2 Model performance comparison with SOTA performance trained only on experimental evidence-based dataset

The FusionESP-exp models achieved superior performance compared to existing SOTA models, with accuracy, area under the curve, and Matthews correlation coefficient (MCC) ranging from 92.21% to 93.57%, 0.9468 to 0.9594, and 0.7945 to 0.8314, respectively. In Kroll

et al.'s study, models combining ESM-1b or fine-tuned ESM-1b with a pretrained GNN on the same datasets outperformed models using ESM-1b and traditional extended-connectivity fingerprints (ECFP) for enzyme and substrate representation, which exhibited the superiority of pretrained model GNN over ECFP-based features (Kroll, Ranjan, Engqvist, et al., 2023). The performance of ESM-1b and GNN model achieved 88.8% accuracy, which was quite close to the one (ACC = 88.29%) of the FusionESP-exp composed of ESM-2-1280 & MoLFormer base models with simple concatenation strategy. The enzyme encoders in both models had the same model size (650 million parameters) and output dimension (1280 dimensions), with some modifications in ESM-2 regarding the training set and model architectures. As for the molecule encoders, MoLFormer was pretrained for general molecule representation, while GNN was pertained on a domain specific prediction task (i.e., production of Michaelis constants K_M of enzyme-substrate pairs) which was closer to our application scenario (i.e., enzyme-substrate pair prediction) (Kroll, Ranjan, Engqvist, et al., 2023). Besides, the pretrained GNN was still learnable during the training task on the enzyme-substrate pair dataset, and additional hyperparameter optimization for the gradient-boosting classifier was used to enhance the performance. Those reasons together contributed to slightly better performance of the ESM-1b & GNN based model over our simple concatenation models (ESM-2-1280 & MoLFormer), though we employed more advanced encoders for enzymes. When employing the contrastive learning strategy, FusionESP-exp with ESM-2-1280 & MoLFormer achieved 92.94% accuracy which was 5.27% higher than that of the model with ESM-1b & GNN. The second model from Kroll et al., employed a task specific fine-tuned ESM-1b model which was fine-tuned on 200,634 enzyme-substrate pairs with phylogenetically inferred evidence (Kroll, Ranjan, Engqvist, et al., 2023). The additional fine-tuning process enhanced the representation power of ESM-1b and

increased the accuracy from 88.8 % to 91.5 %. However, compared to our models with contrastive learning strategy, the performance was inferior to ours by approximately 0.78- 2.26 %, where no additional dataset, ESM-b fine-tuning and GNN fine-tuning were needed, highlighting the effectiveness of our proposed architectures for enzyme-substrate prediction tasks.

7.3.3 Ablation study on the best-performed model

To further explore the function of each layer in the projection head, we conducted an ablation study on the best-performing FusionESP-exp model using ESM-2-2560 and MoLFormer with accuracy of 93.57% (as discussed in the previous section). We systematically investigated the effects of the first dense layer, ReLU activation function, bottleneck layer dimension, batch normalization layer, L2 normalization layer, and loss function. Table 2 summarized the results, indicating that the removal of either the first dense layer or entropy loss function had a detrimental effect on model performance. Given the original dataset's 1:3 ratio of positive to negative pairs, these models were not trainable. Additionally, the bottleneck size, ReLU activation, and batch normalization layers contributed marginally to performance improvement, whereas the L2 normalization layer had negligible impact.

7.3.4 Training the model with including phylogenetic evidence-based data further enhances its performance

In this section, we further enhanced model performance by training on a larger dataset combining phylogenetic evidence-based and experimental evidence-based datasets. Assuming the phylogenetic evidence-based dataset contained rich relationship information between enzymes and substrates, we directly trained models on this dataset for 500 epochs and subsequently the model underwent further training on the experimental evidence-based dataset

for 30 epochs to optimize performance. Following the approach from a reference (Kroll, Ranjan, & Lercher, 2023), we trained and evaluated the model performance on the same datasets for fair performance comparison. The results are shown in Table 3.

Notably, all the models trained on phylogenetic evidence-based dataset only (FusionESP-phylo) outperformed all the models trained on experimental evidence-based dataset from Table 1 (FusionESP-exp), including models employing ESM-1b and ChemBERTa-2. The phylogenetic evidence-based dataset contained 764,449 data points, which was around 15 times larger of the experimental evidence-based dataset. We can infer that the number of learnable parameters was not yet saturated under current datasets and can be expected to further enhance performance with a larger dataset.

Similarly, larger encoder-based models exhibited better performance, with FusionESP-phylo and FusionESP-XL using ESM-2-2560, achieving accuracies of 93.98% and 94.77%, respectively, compared to 93.93% and 94.56% from models using ESM-2-1280. Besides, an improvement was also observed after the models were further trained on the experimental evidence-based dataset. This is not surprising, as the data points in training dataset from experimental evidence-based dataset were much closer to the data points in test dataset, and thus allowing the model to perform better in test dataset. There was also a performance improvement between the models with ESM-2-1280 and MoLFormer and the models with ESM-1b and ChemBERTa-2 under the same contrastive learning strategy. It mainly resulted from the two encoders. ESM-2-1280 and MoLFormer were trained with larger or updated datasets and advanced architectures (e.g., rotary position embeddings, linear attention mechanism) and achieved better performance than that of ESM-1b and ChemBERTa-2 in downstream prediction tasks (Ahmad et al., 2022; Lin et al., 2023; Ross et al., 2022). It is worth noting that ESM-1b can

only embed enzyme sequences shorter than 1024 amino acids or truncated enzymes sequences with the first 1023 amino acids, which caused some information loss and inferior performance.

7.3.5 The projection heads outperform an additional multimodal BERT

Compared to the previous SOTA model (ProSmith designed by Kroll et al.), our model trained and evaluated on the same datasets achieved superior performance (Kroll, Ranjan, & Lercher, 2023). In the study of ProSmith, besides employing ESM1b and ChemBERTa-2 for enzyme and molecule embeddings, they introduced an additional multimodal BERT model trained on the combined text of the enzyme sequence and SMILES. The BERT model was pretrained on 1,039,565 data points from ligand target affinity dataset (Kroll, Ranjan, & Lercher, 2023). Therefore, ProSmith employed three encoders for embeddings and connected them with an optimized gradient boost model for prediction, achieving an accuracy of 94.2% (Kroll, Ranjan, & Lercher, 2023). In contrast, our model (FusionESP-XL) achieved 94.77% accuracy using only two pretrained encoders without additional pretraining of a new multimodal BERT model. With our contrastive learning strategy, FusionESP-XL with ESM-1b and ChemBERTa-2 only employed two encoders same as that of ProtSmith without the multimodal BERT achieved slightly higher accuracy.

In order to explore the contribution to the performance of our proposed architecture, we employed ESM1b and ChemBERTa-2 for enzyme and molecule embeddings and trained the model with our architecture with modifications according to the input dimensions. The performance was slightly better than that of ProSmith. Our promising performance was attributed to our simple projection heads, effectively fusing embeddings from two encoders for prediction tasks, marginally superior to the effect of an additional multimodal BERT encoder and well-optimized gradient boost classifier. At the same time, the ESM1b model was based on

learned position embeddings, and it cannot embed the enzymes whose sequence length is longer than 1023 amino acids (Rives et al., 2021). Therefore, removed those enzyme-substrate pairs whose sequence length was longer than 1023. Though our training dataset was smaller than the one used in ProSmith, our performance was still comparable. Such a comparison further demonstrated the superiority of our proposed architecture.

7.3.6 Evaluating FusionESP-XL performance in unseen enzymes and small molecule

In order to further characterize the model performance in rarely seen enzymes, we first split the test dataset into three subgroups: datapoints with enzymes with a maximal sequence identity to training data between 0 and 40%, between 40% and 60%, and between 60% and 80% based on CD-HIT (Li & Godzik, 2006). The results are presented in Table 4. It was observed that FusionESP-XL model with ESM-2-2560 & MolFormer achieved higher performance in enzyme-substrate pairs with higher sequence identity. Specifically, the accuracy, AUC, and MCC were 96.85%, 0.9871, 0.9198, respectively for enzymes with 60-80% sequence identity. The model still performed well for enzymes with 0-40% sequence identity with an accuracy of 93.07%, AUC of 0.9476, and MCC of 0.8185. It is worth noting that our model also exhibited better performance in the three subsets compared to the performance of ESP by approximately 1.95 - 4.57% in accuracy, respectively (Kroll, Ranjan, Engqvist, et al., 2023). Though trained on the same experimental evidence based dataset, FusionESP-exp with ESM-2-2560 & MolFormer still outperformed the ESP model across all three sub-datasets. When compared with ProtSmith, our model (FusionESP-XL with ESM-2-2560 & MolFormer) outperformed in most challenging sub-dataset (0-40 % maximal sequence identity) with the MCC increasing from 0.78 to 0.8185. This model also surpassed ProtSmith in the 40–60% maximal sequence identity sub-dataset in terms

of MCC, while the MCC was slightly lower than ProtSmith in the 60–80% maximal sequence identity sub-dataset.

Furthermore, we investigated the predictive capabilities of our models for small molecules with different frequency in training dataset (Table 5). Similar to that in maximal sequence identity, FusionESP-XL achieved better performance compared to FusionESP-exp across the 12 sub-datasets. Besides, the rarely seen small molecules tend to be wrongly predicted by the model, especially in unseen small molecules with a ACC of 80.93%. Compared to the performance in unseen small molecules in ProtSmith (MCC = 0.29) and ESP (MCC = 0.00), our model (FusionESP-XL with ESM-2-2560 & MolFormer) achieved slightly higher performance with a MCC of 0.2966. When it came to rarely seen small molecules (only present once in training dataset), our model's performance was increased from 80.93 to 92.57%. Compared to ESP (MCC = 0.28) and ProtSmith (MCC= 0.69), our model achieved much higher performance (MCC = 0.7654).

The performance of our model in rarely seen and unseen enzyme and small molecules indicated that our model has better generalization capability. The projection heads can not only contribute to the increase in overall performance, but also enable the model to perform well in unseen data points.

To further explore the model performance under different maximal sequence identity and frequency of small molecules in training dataset, we further explore the model's performance under different frequencies of small molecules in each sequence similarity subgroups (Fig. 2). It was observed that within the same maximal sequence identity subgroup, performance improved as the frequency of small molecules in the training dataset increased. For example, in the 0–40% maximal sequence identity subgroup, the accuracy increased from 76.13% for unseen small

molecules to 90.72% for those with a frequency ranging from 1 to 10, and further to 94.26% for frequencies greater than 10. Similarly, the prediction performance for unseen and rarely seen small molecules also improved as the maximal sequence identity increased. Particularly, the accuracy for those unseen small molecules was increased from 76.13% with 0 – 40 % maximal sequence identity to 85.09% for enzymes with 40 – 60 % maximal sequence identity and 93.07% for enzymes with 60 – 80 % maximal sequence identity. Predicting enzyme-substrate pairs was particularly challenging when the enzyme exhibited 0 – 40% sequence identity and the small molecule was unseen. However, when the enzyme had a higher maximal sequence identity, or the small molecules were present in the training dataset the model was able to make significantly more reliable predictions.

7.3.7 FusionESP-XL models can express uncertainty

Internally, our model can also provide the cosine similarity score instead of only the positive or negative prediction results to interpret how confident the model is regarding its prediction. In this paper, we set 0.5 as the threshold for output results, where a cosine similarity score between an enzyme and a small molecule was predicted as a positive pair if the score is higher than 0.5, and otherwise it is a negative pair. The cosine similarity score can be also provided as output at our webserver at https://78k6imn5wp.us-east-1.awsapprunner.com for single reaction prediction or large-scale prediction.

To provide a more detailed assessment of prediction accuracies, Fig. 3 displays the distributions of true (blue) and false (orange) predictions within our test dataset across various prediction scores. Most correct predictions had scores either close to 0 or close to 1, indicating that FusionESP-XL made predictions with high confidence. In contrast, false predictions were distributed more evenly across the prediction score range. The predictions for the data points

with scores between 0.4 and 0.6 were more likely to be incorrectly predicted. Therefore, for practical usage of FusionESP-XL, input pairs with cosine similarity score within the 0.4 to 0.6 range should be treated as uncertain and used cautiously in decision making.

7.4 Conclusion

Overall, our proposed contrastive learning strategy (FusionESP) achieved SOTA performance, leveraging two frozen encoders and two simple projection heads with an MSE loss function. The best model, FusionESP-XL, which utilized ESM-2-2560 and MoLFormer, achieved SOTA performance with accuracy, AUC, and MCC of 94.77%, 0.9653 and 0.8628, respectively. Even under limited data availability (using an experimental evidence-based dataset only), our model FusionESP-exp with ESM-2-2560 and MoLFormer also achieved SOTA performance with accuracy, AUC, and MCC of 93.57%, 0.9594 and 0.8314, respectively. Furthermore, our models can handle proteins of any length, unlike ProSmith, which truncates all enzymes into 1023 amino acids due to constraints from the ESM-1b models. This truncation can lead to confusion when two different enzymes share the same first 1023 residues from N- to C-terminus. Moreover, our strategy does not require extra pretrained encoders or datasets for pretraining, but exhibited better generalization capability, whereas ProSmith's development relied on a highly correlated dataset (ligand-target affinity) to pre-train the multimodal BERT encoders before applying them to enzyme-substrate complex embedding. Our model (FusionESP-XL with ESM-2-2560 and MoLFormer) shows great potential for future enzyme-substrate pair prediction, mapping reliable relationship between enzymes and their substrates. The contrastive learning strategy proposed in this study holds significant potential for other multimodal prediction tasks.

Declaration of interest

The authors declare that there are no known conflicts of interest.

Acknowledgements

This is contribution No. 25-059-J from the Kansas Agricultural Experimental Station. This work was supported in part by the National Science Foundation (2419880) and the Plant Protein Innovation Center.

References

Ahmad, W., Simon, E., Chithrananda, S., Grand, G., & Ramsundar, B. (2022). *ChemBERTa-2: Towards Chemical Foundation Models* (arXiv:2209.01712). arXiv. https://doi.org/10.48550/arXiv.2209.01712

Akdis, C. A. (2021). Does the epithelial barrier hypothesis explain the increase in allergy, autoimmunity and other chronic conditions? *Nature Reviews Immunology*, *21*(11), Article 11. https://doi.org/10.1038/s41577-021-00538-7

Bhat, S., Palepu, K., Yudistyra, V., Hong, L., Kavirayuni, V. S., Chen, T., Zhao, L., Wang, T., Vincoff, S., & Chatterjee, P. (2023). De Novo *Generation and Prioritization of Target-Binding Peptide Motifs from Sequence Alone*. https://doi.org/10.1101/2023.06.26.546591

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). *A Simple Framework for Contrastive Learning of Visual Representations* (arXiv:2002.05709). arXiv. https://doi.org/10.48550/arXiv.2002.05709

Copley, S. D. (2017). Shining a light on enzyme promiscuity. *Current Opinion in Structural Biology*, *47*, 167–175. https://doi.org/10.1016/j.sbi.2017.11.001

Du, Z., Ding, X., Xu, Y., & Li, Y. (2023). UniDL4BioPep: A universal deep learning architecture for binary classification in peptide bioactivity. *Briefings in Bioinformatics*, 1–10. https://doi.org/10.1093/bib/bbad135

Du, Z., Xu, Y., Liu, C., & Li, Y. (2023). pLM4Alg: Protein Language Model-Based Predictors for Allergenic Proteins and Peptides. *Journal of Agricultural and Food Chemistry*, acs.jafc.3c07143. https://doi.org/10.1021/acs.jafc.3c07143

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., & Valko, M. (2020). *Bootstrap your own latent: A new approach to self-supervised Learning* (arXiv:2006.07733). arXiv. https://doi.org/10.48550/arXiv.2006.07733

Kim, G. B., Kim, J. Y., Lee, J. A., Norsigian, C. J., Palsson, B. O., & Lee, S. Y. (2023). Functional annotation of enzyme-encoding genes using deep learning with transformer layers. *Nature Communications*, *14*(1), 7370. https://doi.org/10.1038/s41467-023-43216-z

Kroll, A., Ranjan, S., Engqvist, M. K. M., & Lercher, M. J. (2023). A general model to predict small molecule substrates of enzymes based on machine and deep learning. *Nature Communications*, *14*(1), 2787. https://doi.org/10.1038/s41467-023-38347-2

Kroll, A., Ranjan, S., & Lercher, M. J. (2023). *Drug-target interaction prediction using a multi-modal transformer network demonstrates high generalizability to unseen proteins* (p. 2023.08.21.554147). bioRxiv. https://doi.org/10.1101/2023.08.21.554147

Lee, I., Keum, J., & Nam, H. (2019). DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLOS Computational Biology*, *15*(6), e1007129. https://doi.org/10.1371/journal.pcbi.1007129

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Lim, S., Lu, Y., Cho, C. Y., Sung, I., Kim, J., Kim, Y., Park, S., & Kim, S. (2021). A review on compound-protein interaction prediction methods: Data, format, representation and model. *Computational and Structural Biotechnology Journal*, *19*, 1541–1556. https://doi.org/10.1016/j.csbj.2021.03.004

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, *379*(6637), 1123–1130. https://doi.org/10.1126/science.ade2574

Mou, Z., Eakes, J., Cooper, C. J., Foster, C. M., Standaert, R. F., Podar, M., Doktycz, M. J., & Parks, J. M. (2021). Machine learning-based prediction of enzyme substrate scope: Application to bacterial nitrilases. *Proteins: Structure, Function, and Bioinformatics*, *89*(3), 336–347. https://doi.org/10.1002/prot.26019

Nguyen, N.-Q., Jang, G., Kim, H., & Kang, J. (2023). Perceiver CPI: A nested cross-attention network for compound–protein interaction prediction. *Bioinformatics*, *39*(1), btac731. https://doi.org/10.1093/bioinformatics/btac731

Nguyen, N.-Q., Park, S., Gim, M., & Kang, J. (2024). MulinforCPI: Enhancing precision of compound–protein interaction prediction through novel perspectives on multi-level information integration. *Briefings in Bioinformatics*, *25*(1), bbad484. https://doi.org/10.1093/bib/bbad484

Nobeli, I., Favia, A. D., & Thornton, J. M. (2009). Protein promiscuity and its implications for biotechnology. *Nature Biotechnology*, *27*(2), 157–167. https://doi.org/10.1038/nbt1519

Palepu, K., Ponnapati, M., Bhat, S., Tysinger, E., Stan, T., Brixi, G., Koseki, S. R. T., & Chatterjee, P. (2022). *Design of Peptide-Based Protein Degraders via Contrastive Deep Learning*. https://doi.org/10.1101/2022.05.23.493169

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). *Learning Transferable Visual Models From Natural Language Supervision* (arXiv:2103.00020). arXiv. https://doi.org/10.48550/arXiv.2103.00020

Richemond, P. H., Grill, J.-B., Altché, F., Tallec, C., Strub, F., Brock, A., Smith, S., De, S., Pascanu, R., Piot, B., & Valko, M. (2020). *BYOL works even without batch statistics* (arXiv:2010.10241). arXiv. https://doi.org/10.48550/arXiv.2010.10241

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, *118*(15), e2016239118. https://doi.org/10.1073/pnas.2016239118

Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., & Das, P. (2022). Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, *4*(12), Article 12. https://doi.org/10.1038/s42256-022-00580-7

Song, N., Dong, R., Pu, Y., Wang, E., Xu, J., & Guo, F. (2023). PMF-CPI: Assessing drug selectivity with a pretrained multi-functional model for compound–protein interactions. *Journal of Cheminformatics*, *15*(1), 97. https://doi.org/10.1186/s13321-023-00767-z

The UniProt Consortium. (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*(D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100

Thumuluri, V., Martiny, H.-M., Almagro Armenteros, J. J., Salomon, J., Nielsen, H., & Johansen, A. R. (2022). NetSolP: Predicting protein solubility in *Escherichia coli* using language models. *Bioinformatics*, *38*(4), 941–946. https://doi.org/10.1093/bioinformatics/btab801

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Wang, X., Liu, J., Zhang, C., & Wang, S. (2022). SSGraphCPI: A Novel Model for Predicting Compound-Protein Interactions Based on Deep Learning. *International Journal of Molecular Sciences*, *23*(7), Article 7. https://doi.org/10.3390/ijms23073780

Yang, M., Fehl, C., Lees, K. V., Lim, E.-K., Offen, W. A., Davies, G. J., Bowles, D. J., Davidson, M. G., Roberts, S. J., & Davis, B. G. (2018). Functional and informatics analysis enables glycosyltransferase activity prediction. *Nature Chemical Biology*, *14*(12), 1109–1117. https://doi.org/10.1038/s41589-018-0154-9

Yu, T., Cui, H., Li, J. C., Luo, Y., Jiang, G., & Zhao, H. (2023). Enzyme function prediction using contrastive learning. *Science*, *379*(6639), 1358–1363. https://doi.org/10.1126/science.adf2465

Zhang, D., Jia, C., Sun, D., Gao, C., Fu, D., Cai, P., & Hu, Q.-N. (2023). Data-Driven Prediction of Molecular Biotransformations in Food Fermentation. *Journal of Agricultural and Food Chemistry*, *71*(22), 8488–8496. https://doi.org/10.1021/acs.jafc.3c01172

Zhang, D., Xing, H., Liu, D., Han, M., Cai, P., Lin, H., Tian, Y., Guo, Y., Sun, B., Le, Y., Tian, Y., Wu, A., & Hu, Q.-N. (2024). Discovery of Toxin-Degrading Enzymes with Positive Unlabeled Deep Learning. *ACS Catalysis*, 3336–3348. https://doi.org/10.1021/acscatal.3c04461

Zhang, Q., Ding, K., Lyv, T., Wang, X., Yin, Q., Zhang, Y., Yu, J., Wang, Y., Li, X., Xiang, Z., Zhuang, X., Wang, Z., Qin, M., Zhang, M., Zhang, J., Cui, J., Xu, R., Chen, H., Fan, X., … Chen, H. (2024). *Scientific Large Language Models: A Survey on Biological & Chemical Domains* (arXiv:2401.14656). arXiv. https://doi.org/10.48550/arXiv.2401.14656

Figures and Tables:

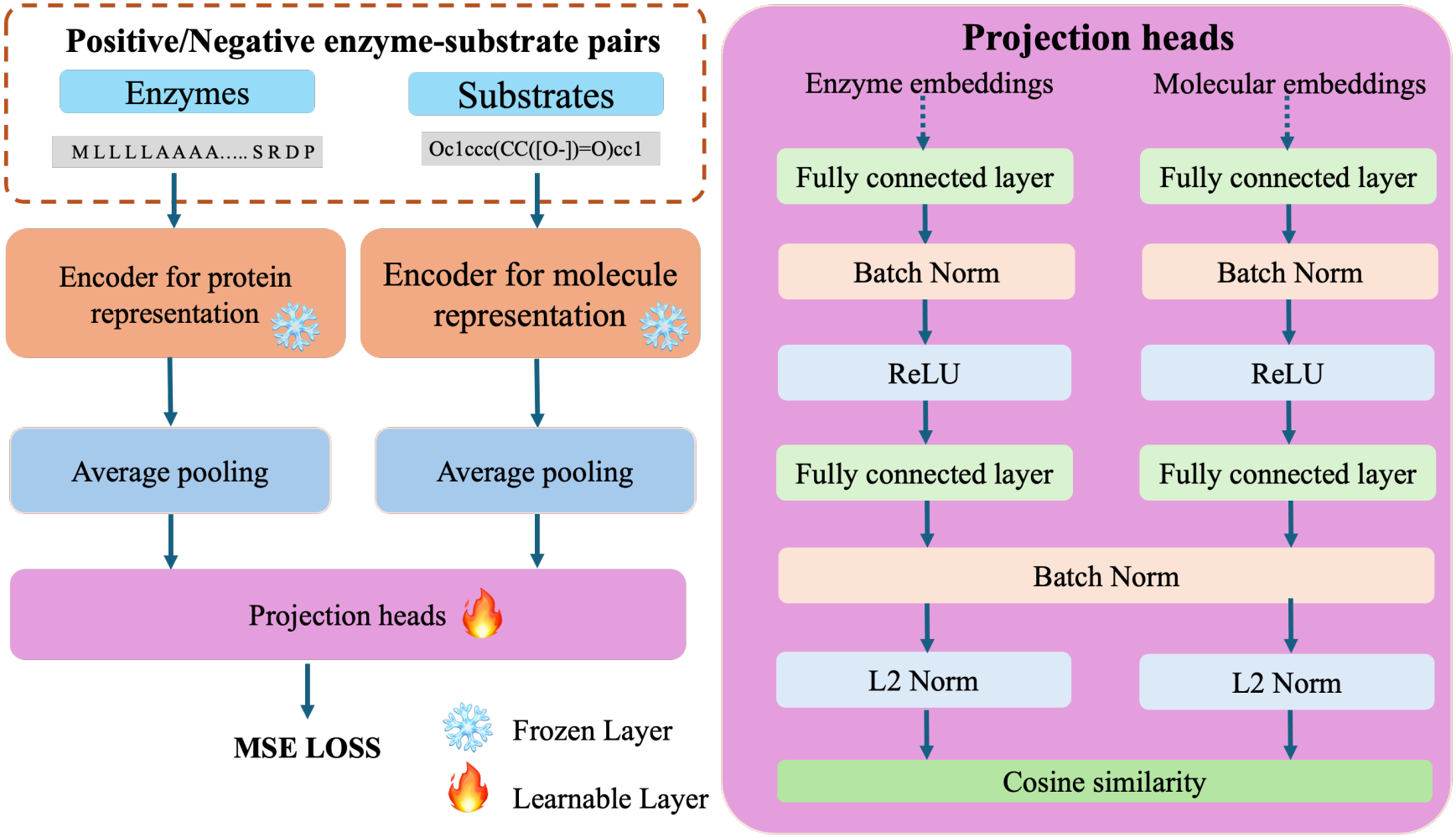


Figure 7-1 Model architecture for enzyme-substrate pair prediction model

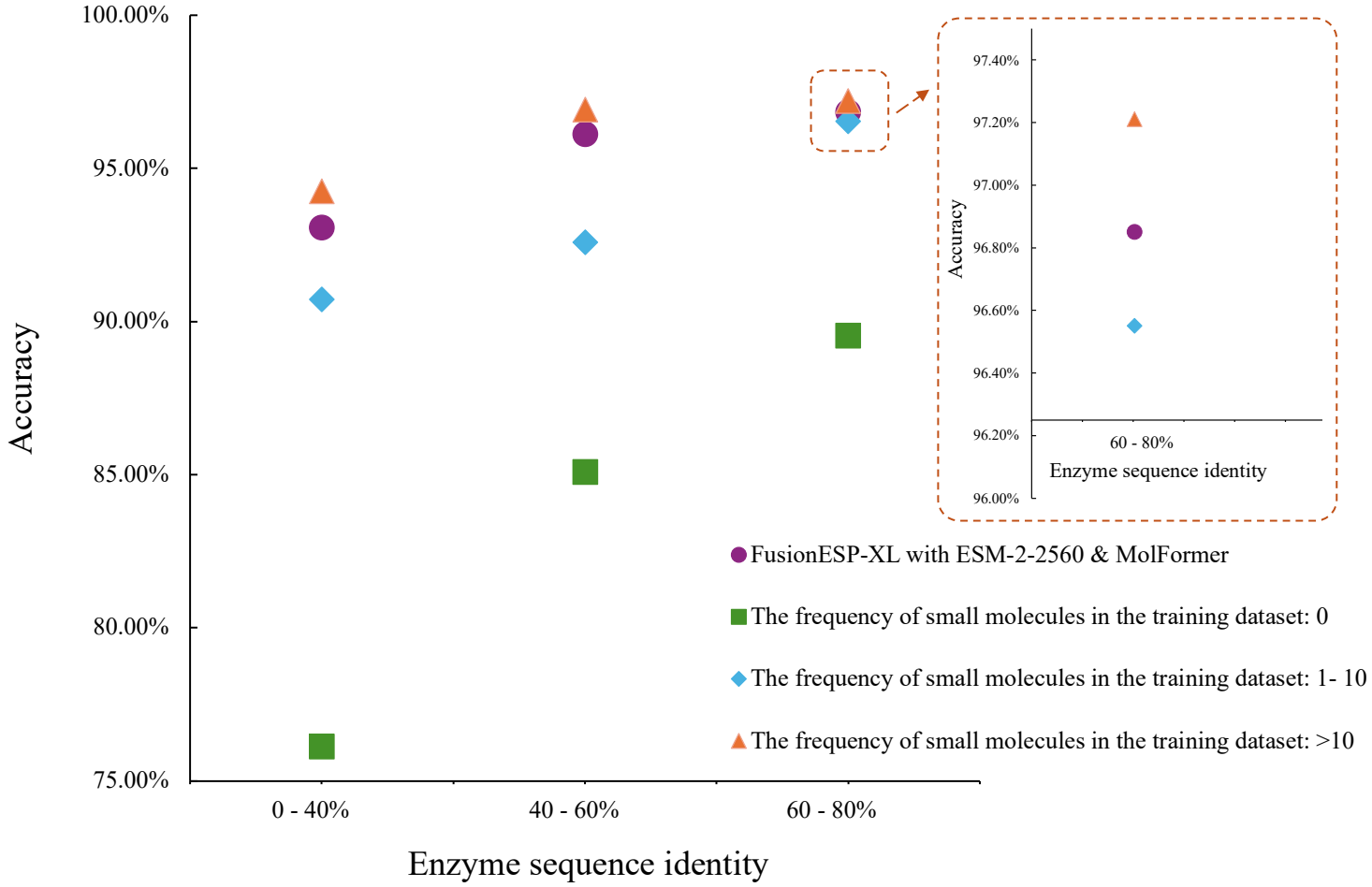


Figure 7-2 Prediction performance of FusionESP-XL with ESM-2-2560 & MolFormer.

Note: we divided the test dataset into subset with different level of enzyme sequence identity and frequency of small molecules in the training dataset. Source data are provided as a Source Data file.

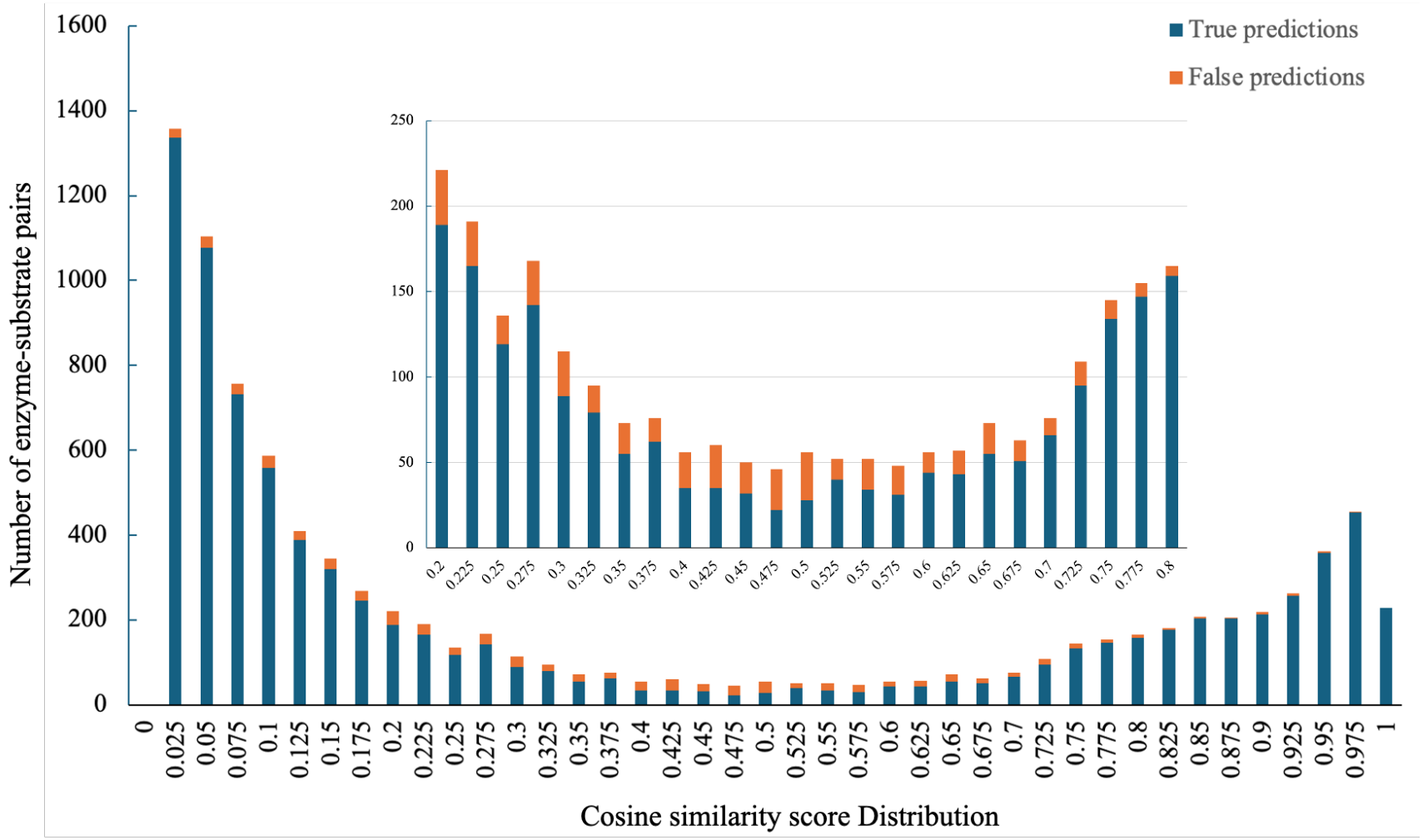


Figure 7-3 Prediction scores around 0.5 indicate model uncertainty.

Note: stacked histogram bars display the prediction score distributions of true predictions (blue) and false predictions (red). The inset shows a blow-up of the interval [0.2, 0.8]. Scores are prediction by FusionESP -XL with ESM-2-2560 and MoLFormer. Source data are provided as a Source Data file.

Table 7-1 Performance of our model (FusionESP-exp) and the previous models trained on the experimental evidence-based dataset only

Model strategy	Encoder for enzymes	Encoder for molecules	Dataset (train, validation, test)	ACC	AUC	MCC	Additional notes
Contrastive learning strategy	ESM-2-480	MoLFormer	50077, 5422, 13336	92.21%	0.9468	0.7945	Our model
	ESM-2-1280	MoLFormer	50077, 5422, 13336	92.94%	0.9558	0.8146	Our model
	ESM-2-2560	MoLFormer	49881, 5422, 13336	**93.57%**	**0.9594**	**0.8314**	Our model
Simple concatenation strategy	ESM-2-480	MoLFormer	50077, 5422, 13336	86.42%	0.9078	0.6420	Our baseline
	ESM-2-1280	MoLFormer	50077, 5422, 13336	88.29%	0.9255	0.6928	Our baseline
	ESM-2-2560	MoLFormer	49881, 5422, 13336	89.64%	0.9357	0.7272	Our baseline
Simple concatenation strategy*	ESM-1b	GNN	50093, 5422, 13336	88.8%	0.94	0.72	ESP[1]
	ESM-1b$_{ts}$	GNN	50093, 5422, 13336	91.5%	0.956	0.78	ESP[1]

Note:

Original training, validation, and test dataset sizes are 50093, 5422, and 13336.

GNN: Graph neural networks.

The GNN was pretrained on predicting the Michaelis constants K_M of enzyme-substrate pairs.

*: Results were retrieved from previous studies by Kroll et al [1].

ESM-1b$_{ts}$: It is a task specific fine-tuned ESM-1b model, where ESM-1b model was fine-tuned on 200634 enzyme substrate pairs with phylogenetically inferred evidence

Table 7-2 Ablation study on FusionESP-exp with ESM-2-2560 and MoLFormer

Bottleneck size	First dense layer	ReLU	BatchNorm	L2 normalization	Loss function	ACC	AUC	MCC
32	✓	✓	✓	✓	MSE	93.09%	0.9541	0.8192
64	✓	✓	✓	✓	MSE	93.02%	0.9563	0.8177
128	✓	✓	✓	✓	MSE	**93.57%**	**0.9594**	**0.8314**
256	✓	✓	✓	✓	MSE	93.38%	0.9574	0.8268
512	✓	✓	✓	✓	MSE	93.16%	0.9562	0.8218
128	×	✓	✓	✓	MSE	74.31%	0.8405	0.1310
128	✓	×	✓	✓	MSE	92.21%	0.9499	0.7939
128	✓	✓	×	✓	MSE	92.01%	0.9448	0.7901
128	✓	✓	✓	×	MSE	93.42%	0.9591	0.8278
128	✓	✓	×	×	MSE	91.97%	0.9447	0.7888
128	×	×	✓	✓	MSE	73.90%	0.6867	0.0856
128	✓	✓	✓	✓	CE*	74.47%	0.8217	0.4467

Note:

MSE: mean square error; CE: cross entropy

* : to calculate the cross entropy, we employed the sigmoid function to get the probability based on similarity valued and then calculated the cross entropy loss between the probability and the label.

Table 7-3 Performance of our models (FusionESP-phylo and FusionESP-XL) and previous model trained on both experimental evidence-based and phylogenetic evidence-based datasets

Encoder for enzymes	Encoder for molecules	ACC	AUC	MCC	Additional notes
ESM-2-2560	MoLFormer	93.98%	0.9567	0.8419	Our model (FusionESP-phylo)*
ESM-2-2560	MoLFormer	**94.77%**	0.9653	**0.8628**	Our model (FusionESP-XL)**
ESM-2-1280	MoLFormer	93.93%	0.9534	0.8404	Our model (FusionESP-phylo)*
ESM-2-1280	MoLFormer	94.56%	0.9635	0.8572	Our model (FusionESP-XL)**
ESM-1b	ChemBERTa-2	93.88%	0.9517	0.8391	Our model (FusionESP-phylo)*
ESM-1b	ChemBERTa-2	94.3%	0.9618	0.8519	Our model (FusionESP-XL)**
ESM-1b	ChemBERTa-2	94.2%	**0.972**	0.85	ProSmith[2]

Note:

* : the model (FusionESP-phylo) was trained on the phylogenetic evidence-based dataset only;

** : the model (FusionESP-XL) was trained on the phylogenetic evidence-based dataset first and continued to be trained on the experimental evidence-based dataset, corresponding model.

Table 7-4 The performance of FusionESP-XL and FusionESP-exp with ESM-2-2560 & MolFormer in rarely seen and unseen enzymes

Maximal sequence identity	ACC	AUC	MCC
FusionESP-XL with ESM-2-2560 & MolFormer			
0-40%	93.07%	0.9476	0.8185
40-60%	96.12%	0.9752	0.9001
60-80%	96.85%	0.9871	0.9198
FusionESP-exp with ESM-2-2560 & MolFormer			
0-40%	91.46%	0.9334	0.7752
40-60%	95.49%	0.9731	0.8842
60-80%	96.31%	0.9864	0.9066
ESP*			
0-40%	89%	0.93	0.72
40-60%	93%	0.97	0.83
60-80%	95%	0.99	0.88

Note: The performance of ProSmith was only exhibited in figure instead of table, and thus we did not have the exact values.

*: The performance of ESP model was based on the task specific fine-tuned ESM-1b model.

Table 7-5 The performance of FusionESP-XL and FusionESP-exp with ESM-2-2560 & MolFormer in rarely seen and unseen small molecules

Number of positive samples with identical metabolite in the training set	Size of Subset	ACC	AUC	MCC
FusionESP-XL with ESM-2-2560 & MolFormer				
0	640	80.93%	0.6942	0.2966
1	498	92.57%	0.9445	0.7654
2	431	93.27%	0.8935	0.7880
3	586	96.07%	0.9657	0.8758
4	351	95.44%	0.9634	0.8631
5	414	93.48%	0.9486	0.7976
6	278	94.96%	0.9597	0.8547
7	313	94.57%	0.9731	0.8460
8	222	92.79%	0.9724	0.7938
9	317	94.32%	0.9575	0.8471
10	157	96.18%	0.9900	0.8954
>10	8876	95.86%	0.9783	0.8995
FusionESP-exp with ESM-2-2560 & MolFormer				
0	640	76.72%	0.6637	0.1343
1	498	87.95%	0.8870	0.6198
2	431	88.86%	0.8738	0.6503

3	586	93.86%	0.9584	0.8080
4	351	92.02%	0.9397	0.7660
5	414	92.03%	0.9432	0.7504
6	278	92.45%	0.9580	0.7833
7	313	94.89%	0.9625	0.8543
8	222	91.44%	0.9432	0.7512
9	317	94.01%	0.9704	0.8371
10	157	95.54%	0.9737	0.8775
>10	8876	95.30%	0.9733	0.8855

Chapter 8 - Overall conclusions and future perspectives

8.1 Overall conclusions

The exploration and utilization of bioactive peptides derived from food proteins have garnered significant attention in recent years due to their potential health benefits and applications in functional foods and nutraceuticals. This dissertation focused on employing machine learning techniques to accelerate the discovery and characterization of these peptides. Collectively, the studies have contributed to enhancing the efficiency of bioactive peptide identification and have extended into knowledge fusion for enzyme-substrate interaction prediction under multimodal conditions. Most of the prediction models developed were deployed on user-friendly web servers to facilitate easier access and usage.

In Chapter 1, a comprehensive review highlighted the critical role of bioinformatics tools—including databases, proteolysis simulations, machine learning-powered quantitative structure-activity relationship (QSAR) models, molecular docking, and molecular dynamics simulations—in food-derived bioactive peptide research. The integration of these tools facilitated the rational design of protein hydrolysates and accelerated the translation of identified peptides into functional food products.

Chapters 2 and 3 investigated the development of QSAR models combined with machine learning techniques for antioxidant activity prediction in dipeptides and tripeptides. In Chapter 2, seven feature selection strategies and 14 regression methods were combined to build QSAR models, comprehensively evaluating the performance of machine learning methods in predicting antioxidant tripeptides. These experiments resulted in a state-of-art (SOTA) regression model with an R^2 of 0.847 on the test dataset. The modeling strategy was then applied to antioxidant

dipeptides in Chapter 3. Feature analysis revealed that C-terminal residues in tripeptides and N-terminal residues in dipeptides contributed more to antioxidant activity. Several dipeptides and tripeptides were synthesized to experimentally validate the model performance. Furthermore, a hydrolysis simulation tool, R-PeptideCutter, was designed to connect the prediction models with food proteins for bioactive peptide production. A case study on sorghum proteins identified a high-antioxidant-activity dipeptide, YR.

To further enhance model performance in peptide discovery, Chapters 4**,** 5**,** and 6 systematically explored the use of protein language models (pLMs) in developing bioactive peptide prediction models. In Chapter 4, we introduced a universal model architecture, UniDL4BioPep, which exhibited better performance on 15 out of 20 bioactive peptide datasets, with accuracy improvements ranging from 0.7% to 7%. A user-friendly protocol was designed for researchers interested in building their own prediction models using the UniDL4BioPep architecture and custom datasets. An advanced version was also developed specifically for extremely imbalanced datasets.

In Chapters 5 and 6, we developed SOTA models for screening peptides with strong angiotensin-converting enzyme (ACE) inhibitory activity for antihypertension management with an accuracy of 88.3% and for predicting allergenic proteins and peptides for allergen identification and risk assessment in food products with an accuracy of 95.1%, respectively. Our studies revealed the superiority of pLMs over traditional peptide descriptors in peptide representation, the strong compatibility of support vector machines with pLMs alongside multilayer perceptron and logistic regression, and a positive correlation between pLM sizes and

prediction model performance. All models developed in these chapters were deployed on user-friendly web servers for academic use without charge.

In Chapter 7**,** moving beyond employing pLMs for peptide discovery in a single-modality context, we presented FusionESP, which integrates protein and chemistry language models using a contrastive learning strategy for multimodal prediction tasks involving enzyme-substrate pairs. With a newly designed projection head to fuse knowledge from protein and chemistry language models, we successfully achieved SOTA performance (accuracy = 94.70%) while requiring fewer computational resources and data points. These prediction models are also deployed on user-friendly web servers to aid scientific discovery in biochemistry.

8.2 Future perspectives

To advance the studies presented in this work, the following areas are recommended for further research:

8.2.1 Enhancing explainability and performance with graph neural networks

Preliminary experiments indicated that graph neural networks (GNNs) exhibited slightly higher performance compared to the models in Chapter 4. Unlike average pooling methods that unify feature dimensions, GNNs can handle data of varying dimensions without information loss, leading to improved model performance. An additional advantage of GNNs is their explainability. Specifically, graph attention networks (GATs) offer direct interpretability by visualizing attention maps, which show the focus the network assigns to each neighboring node during predictions.

8.2.2 Further exploration of contrastive learning strategies

Numerous positive/negative pairs exist in the field of biochemistry, such as in protein-compound, protein-protein, and protein-peptide interactions. Applying contrastive learning strategies for knowledge fusion is expected to enhance model performance in these prediction tasks.

Contrastive learning can also strengthen the representation power of small molecules. By effectively fusing encoders from different modalities using contrastive learning strategies, integrated encoders can generate comprehensive and effective embeddings with lower dimensionality, ultimately improving prediction performance for small molecule properties. Notably, such strategies can be employed in both self-supervised and supervised learning modes.

8.2.3 Development of foundation models for peptide discovery

Developing a peptide-specific language model with a lower output dimension is expected to be more suitable for small peptide datasets and enhance model performance through effective representation.

Current language models have not included modified peptides in their training datasets, despite the prevalence of post-translational and chemical modifications. Utilizing modified SMILES (Simplified Molecular Input Line Entry System) or other chemical notation systems like SELFIES (Self-Referencing Embedded Strings) for peptide representation can facilitate the pretraining of peptide language models in a sequential manner. Representing specific peptide patterns, such as peptide bonds, can minimize sequence length and conserve computational resources.

8.2.4 Establishment of high-quality databases for one-stop data retrieval

In this work, datasets were collected from public databases and previous publications. There is a significant demand for high-quality, well-recognized datasets in peptide discovery—and broader biochemistry—for model developers. Such datasets are crucial for performance comparison among different models and for advancing the field.

8.2.5 Utilizing unlabeled data

A vast number of identified peptides are either unlabeled or only partially labeled concerning certain properties based on experimental results (e.g., solubility, antioxidant activity). Beyond language models, alternative solutions like semi-supervised learning could exploit these unlabeled peptides, enhancing the models' learning capacity and performance.

8.2.6 Development of multi-label models

The multiple bioactivities of a single peptide may have latent relationships. Instead of building separate prediction models for each bioactivity, developing multi-label models could better uncover these relationships. However, addressing the challenge of missing bioactivities remains a significant hurdle.

Beyond those perspectives involved into this work, a few more potential works might be considered related to machine learning in biochemistry.

8.2.7 Enzyme annotation (Enzyme Commission (EC) number prediction)

A non-causal decoder architecture fit very well into this prediction context, where enzyme sequence is the condition, and the EC number has hierarchical properties and good for next work prediction task.

A non-causal decoder architecture is well-suited for this prediction context, where the enzyme sequence serves as the condition, and the Enzyme Commission (EC) number, with its hierarchical properties, is appropriate for next-word prediction tasks.

8.2.8 Regression model development with transformer model

With the success of regression transformer, it is feasible to combine large language models based on natural language processing with scientific prediction tasks, whether classification or regression. This integration could make prediction tools more accessible to the public, fostering broader use and collaboration.

Appendix A - Supplementary documents

Chapter 1- Supplementary Document S1-**Table S1** The latest virtual screening studies using QSAR, molecular docking, and molecular dynamics simulation (Full version).

Local descriptor-based QSAR model application						
Bioactivity	*Dataset size*	*Amino acid descriptors (AADs)*	*Model*	*Performance*	*Additional comments**	*Reference*
Antioxidant activity (ABTS assay)	133 tripeptides	566 amino acid properties were selected by 6 ML methods as AADs or directly used as AADs	14 ML regression models	R^2_P = 0.847 (best model by RFR for feature selection and XGB for regression model)	QAY, PHC, YPQ, VYV, GPE, and YSQ; performance comparison between 98 models	(Du, Wang, et al., 2022)
Antioxidant activity (DPPH assay)	69 peptides	SVRG and SVEEVA were screened by SWR and only for the representation of first five residues at terminus	PLSR	R^2_P = 0.5536 by SVRG R^2_P = 07173 for SVEEVA	AGWACLVG, IDLAY, YPLDL, IPIGP, and EAFDPLG	(Zhu et al., 2022)
Antioxidant activity (ABTS assay)	50 tripeptides	The effective hydrophobicity scale π_α, electronic parameters, and steric parameters of amino acid side chains	MLR	R^2_P = 0.668	All the peptides were synthesized and tested in their own laboratory	(Uno et al., 2020)
Antioxidant activity (FTC and FRAP assays)	214 tripeptides in FTC dataset and 173 tripeptides in FRAP dataset	16 AADs were integrated by BOSS	PLSR	R^2_{CV} =0.7471 for FTC dataset and R^2_{CV} = 0.6088 for FRAP dataset	MPA for outlier detection; performance comparison between the	(Deng et al., 2019)

					descriptor selector by BOSS and 16 basic AADs	
Antioxidant activity	91 tripeptides	195 physiochemical properties of amino acids were screened by SWR	PLSR, SVMR, RFR, and MLR	R^2_{CV} =0.706 for PLSR R^2_{CV} =0.764 for SVMR R^2_{CV} =0.728 for RFR R^2_{CV} =0.798 for MLR	GHG, LVG, GHT, GHG, GHP, KHP, GVR, ECG, GVW, GKW, GHW, QVW, KVW, NKW, NHW, QHW, KHW, PYW, and YHW	(N. Chen et al., 2018)
Antioxidant activity (ORAC and ABTS assays)	172 tripeptides	DPPS	MLR	R^2_{CV} =0.541	N/A	(Deng et al., 2019)
Antioxidant activity (ORAC and ABTS assays)	48 dipeptides	3-z scale, 5-z scale, DPPS, and ISA-ECI	PLSR	All the R^2_{CV} were below 0.5	The bioactivity of peptides used in this studies were determined in the same laboratory	(Zheng et al., 2016)
ACE inhibitory activity	58 dipeptides	80 different AADs	SVMR and PLSR	R^2_P varied from 0.6 to 0.82 for SVMR and from 0.48 to 0.7 for PLSR	N/A	(Zhou et al., 2021)
ACE inhibitory activity	66 tripeptides	553 physicochemical properties selected by PCA	MLR	R^2_P =0.861	Performance comparison with other 9 AAD-based models	(Mahmoodi-Reihani et al., 2020)
ACE inhibitory activity	84 dipeptides, 169 tripeptides, and 15 tetrapeptides	SVHEHS screened by OSC	SVMR	R^2_{CV} in four models were all above 0.995	ACC was employed to unify the feature dimension among	(Guan & Liu, 2019)

					different peptide length datasets	
ACE inhibitory activity	141 dipeptides	16 AADs were integrated by BOSS	PLSR	R^2_{CV} =0.7151	MPA for outlier detection; performance comparison between the descriptor selector by BOSS and 16 basic AADs	(Deng et al., 2017)
ACE inhibitory activity	166 dipeptides and 141 tripeptides	5-z scale	PLSR	R^2_{CV} = 0.756 for dipeptides R^2_{CV} = 0.445 for tripeptides	YW and LRY	(Y. Fu et al., 2016)
DPP IV inhibitory activity	30 peptides	5-z scale and v-scale used for the representation of N- and C-terminus resides	PLSR	R^2_{CV} = 0.775 for 5-z scale R^2_{CV} = 0.754 for v-scale	FP, HP, RP, VP, IPM, LPP, IPPL, IPSK, VPGEIVE, YPFPGP, LPQNIPPLT, IPPLTQT, TPVVVPP, YPVEPF, LPLPLL, QPHQPLPPT, QPLPPT, and LPVPQ	(Nongonierma & FitzGerald, 2016)
Antimicrobial activity	196 dodecapeptides	20 different AADs	PLSR	R^2_{CV} = 0.633 for FASGAI as AADs	AADs were screened by GA-PLSR before model development	(Qian et al., 2017)
Bitterness	48 dipeptides	80 different AADs	SVMR and PLSR	R^2_P varied from 0.6 to 0.82 for SVMR and	N/A	(Zhou et al., 2021)

				from 0.5 to 0.8 for PLSR		
Bitterness	48 dipeptides	553 physicochemical properties selected by PCA	MLR	R^2_P =0.907	Performance comparison with 7 AAD-based models	(Mahmoodi-Reihani et al., 2020)
Bitterness	48 dipeptides, 52 tripeptides, and 23 tetrapeptides	16 AADs were integrated by BOSS	PLSR	R^2_{CV} = 0.941 for dipeptides R^2_{CV} = 0.742 for tripeptides R^2_{CV} = 0.956 for tetrapeptides	MPA for outlier detection; performance comparison between the descriptor selector by BOSS and 16 basic AADs	(Xu & Chung, 2019)

Global descriptor-based QSAR (3D-QSAR) model application

Bioactivity	*Dataset size*	*Molecular alignment & MIF calculation*	*Model*	*Performance*	*Additional comments**	*Reference*
ACE inhibitory activity	53 di/tripeptides	Docking-based alignment (Suflex-Dock); CoMFA and CoMSIA for MIF calculation	PLSR	R^2_{CV} = 0.773 for CoMFA R^2_P = 0.664 for CoMSIA	GEF, VEF, VRF, and VKF were synthesized for validation	(Qi et al., 2017)
ACE inhibitory activity	23 dipeptides and 27 tripeptides	Template ligand-based alignment; CoMFA for MIF calculation	PLSR	R^2_{CV} = 0.542 for dipeptides R^2_{CV} = 0.452 for tripeptides	N/A	(Vukic et al., 2017)
ACE inhibitory activity	40 dipeptides, 32 tripeptides, and 41 tetra/penta/hexa peptides	Template ligand-based alignment; CoMFA and CoMSIA for MIF calculation	PLSR	R^2_{CV} = 0.862 for dipeptides (CoMSIA) R^2_{CV} = 0.848 for tripeptides (CoMFA) R^2_{CV} = 0.656 for longer peptides (CoMFA)	N/A	(S. Wu et al., 2014)

Antimicrobial activity	24 nonapeptides	Template ligand-based alignment; CoMFA and CoMSIA for MIF calculation	PLSR	R^2_{CV} = 0.601	All peptides are rich in R, P, and F residues	(S. Wu et al., 2014)
Bitterness	21 peptides	Template ligand-based alignment; CoMFA and CoMSIA for MIF calculation	PLSR	R^2_{CV} = 0.530	All peptides are rich in W residue	(S. Wu et al., 2014)

Molecular docking & molecular dynamics simulation for virtual screening

Protein source	*Bioactivity*	*Molecular docking*	*Molecular dynamics simulation*	*Additional comments*	*In vitro or in vivo experiments*	
Fermented soy	Keap1–Nrf2 interaction inhibitory activity	CABS-Dock, GalaxyPepDock, and HADDOCK	-	28 peptides identified by LC-MS; CABS-Dock and GalaxyPepDock were used for screening, and HADDOCK was used for conformation refinement	Thirteen peptides were selected for *in vitro* and *in vivo* antioxidant experiments	(Tonolo, Moretto, et al., 2020)
Egg protein	Keap1–Nrf2 interaction inhibitory activity	CDOCKER	-	Fluorescence polarization assay was used to further refine the FBPs selected by CDOCKER	HepG2 cell model for oxidative damage, cytotoxicity, and cytoprotection evaluation	(L. Li et al., 2017)
Pea protein	DPP IV inhibitory activity	AutoDock Vina	-	30 peptides identified by LC-MS and contained P/A at the second position at N-terminus were further screened for peptide synthesis	Eight peptides were synthesized for *in vitro* DPP IV inhibitory activity assay	(M. Zhang et al., 2022)
Pumpkin seed	ACE inhibitory activity	MOE	GROMACS for 30 ns of simulation;	Identified 47 di/tripeptides virtually screened by MOE;	Tripeptide IFA was synthesized for *in*	(Liang et al., 2021)

			Force field: Amber99SB-ILDN	ADMET evaluation to select for peptide synthesis	*vitro* ACE inhibitory activity assay	
-	ACE inhibitory activity	GOLD	AMBER12 for 25 ns of simulation; Force field: AMBER FF03	8000 theoretical tripeptides were ranked by GOLD, and MDS was used to elucidate interaction mechanism	Five peptides (WCW, IWW, WWW, WWI and WLW) were synthesized for *in vitro* ACE inhibitory activity assay	(Panyayai et al., 2018)
Amaranth protein	Renin inhibitory activity	CABS-dock and FlexPepRDock	-	CABS-dock for ligand-receptor docking; FlexPepDock for conformation refinement and peptide-protein interaction energy evaluation (Rosseta score)	Four peptides were synthesized for *in vitro* renin inhibitory activity assay	(Nardo et al., 2020)
Mytilus edulis proteins	Thrombin inhibitory activity	CDOCKER	-	39 peptides identified by LC-MS and further screened by CODCKER	KNAQNQLGEVTVR was synthesized for *in vitro* thrombin inhibitory activity assay	(L. Feng et al., 2018)
Bovine milk casein	Antithrombotic activity	CDOCKER	-	35 peptides identified by UPLC-Q-TOF-MS/MS and further screened by CODCKER	No peptide synthesis for validation	(Tu et al., 2017)
Walnut meal	Tyrosinase inhibitory activity	AutoDock 4 and CDOCKER	-	606 peptides identified from LC-MS were screened by AutoDock 4; CDOCKER was used for conformation refinement	FPY was synthesized for *in vitro* experiment	(Y. Feng et al., 2021)

Sesame seeds	Tyrosinase inhibitory activity	AutoDock 4	-	Eight peptides were selected from 361 peptides reported to exhibit antioxidant activity from *in silico* proteolysis simulation	No peptide synthesis for validation	(Baskaran et al., 2021)
Oncorhynchus mykiss Nebulin	Umami	CDOCKER	-	332 peptides were obtained from *in silico* proteolysis simulation and further screened by CODCKER	20 peptides were synthesized and tested by electronic tongue	(W. Zhao et al., 2023)

Integrated strategy

Protein source	*Bioactivity*	*QSAR model*	*Molecular docking*	*Molecular dynamics simulation*	*Additional comments*	*Reference*
-	DPP-IV inhibitory activity	Self-built regression model	AutoDock Vina	GROMACS for 100 ns of simulation Forcefiled: AMBER14SB and General AMBER	PaDEL descriptors for peptide representation, GA for feature selection, and MLR for regression model	(X. Li et al., 2022)
Sorghum protein	DPP-IV inhibitory activity	PeptideRanker	HPEPDOCK and FlexPepDock	-	PeptideRanker for rough virtual screening, and HPEPDOCK and FlexPepDock for refinement	(Garzón et al., 2022)
Bean protein	DPP-IV inhibitory activity	PeptideRanker	Pepsite2	-	PeptideRanker for rough virtual screening and PepSite2 for refinement and selection for *in vitro* assay peptide synthesis	(Martini et al., 2022)
Draft beer	DPP-IV inhibitory activity and ACE inhibitory activity	PeptideRanker	Sybyl software		PeptideRanker for rough screening and selection for *in vitro* assay peptide synthesis; Sybylfor interaction	(Wenhui et al., 2022)

					mechanism; absorption and toxicity evaluation	
Cheese	Antidiabetic activity	PeptideRanker	Pepsite2 and HPEPDOCK	-	PeptideRanker for rough virtual screening and PepSite2 for refinement and selection for *in vitro* assay peptide synthesis; HPEPDOCK for interaction mechanism	(Martini et al., 2021)
Rubing cheese	ACE, a-glucosidase, and Keap1–Nrf2 interaction inhibitory activity	PeptideRanker	Pepsite2 and AutoDock Vina		PeptideRanker for rough virtual screening and PepSite2 for refinement and selection for *in vitro* assay peptide synthesis; AutoDock Vina for interaction mechanism	(Wei et al., 2022)
Chia Seed	ACE inhibitory activity, anti-inflammatory activity, and plasma stability	PeptideRanker, AHTpin, PreAIP, AntiAngioPred, and PlifePred	AutoDock Vina	PMEMD for 300 ns of simulation Force fields: amberff14 and GAFF	PMEMD for conformation generation of protein receptors and AutoDock Vina for molecular scoring	(Aguilar-Toalá et al., 2022)
Egg yolk protein	ACE inhibitory activity	PeptideRanker and AHTpin	Autodock CrankPep	-	PeptideRanker for bioactivity possibility prediction and AHTpin for virtual screening of ACE inhibitory FBPs	(Marcet et al., 2022)
β-Casein	ACE inhibitory activity	Self-built classification model	AutoDock Vina	-	Eight sequence-based and structure-based features for decision tree model	(Kalyan, Junghare, Khan, et al., 2021)

Camel milk	ACE and renin inhibitory activity	PeptideRanker	PepSite2 and Glide	-	PeptideRanker for rough screening and Pepsite2 for refinement; Glide for interaction mechanism	(Baba, Baby, et al., 2021)
-	ACE inhibitory activity	Self-built regression model	AutoDock4	-	Dragon descriptors, AAindex, and 5-z scale for peptide representation; KNN, RFR, MLP, and SVMR for regression model to further select for peptide synthesis and *in vitro* assays	(Y.-T. Wang et al., 2020)
	ACE inhibitory activity	Self-built regression model	AutoDock4	-	3D QSAR models (PLSR) based on ligand template-based molecular alignment and MIF calculation by CoMFA and CoMSIA (R^2_P = 0.6257 for CoMFA and R^2_P = 0.6969 for CoMSIA); AutoDock4 for interaction mechanism	(F. Wang & Zhou, 2020)
Honey protein	ACE inhibitory activity	AHTpin	PatchDock	-	AHTpin for rough screening, PatchDock for refinement, and FireDock for interaction mechanism	(Tahir et al., 2020)
Camel milk	ACE inhibitory activity	PeptideRanker and AHTpin	PepSite2 and Glide	-	PeptideRanker and AHTpin for rough screening, Pepsite2 for refinement, and Glide for interaction mechanism; toxicity evaluation	(Mudgil, Baby, Ngoh, Kamal, et al., 2019)
Salmo salar collagen	ACE inhibitory activity	PeptideRanker	SwissDock and CDOCKER	-	PeptideRanker for rough screening, SwissDock for refinement, and CDOCKER for interaction mechanism;	(Yu et al., 2018)

					physicochemical property and toxicity evaluation	
Silkworm cocoon	ACE inhibitory activity	Self-built regression model	Surflex-Dock	-	3D QSAR models (PLSR) based on ligand template-based molecular alignment; MIF calculation by CoMFA and CoMSIA; toxicity and digestive stability evaluation to further select peptides for synthesis and *in vitro* assays	(Sun et al., 2017)
Qula casein	ACE inhibitory activity	Self-built regression models	Discovery Studio	-	5z-scale was used to build PLSR models for penta/hexa/hepta/octapeptide; QSAR model for rough screening of peptides identified from LC-MS and Autotock Vina for refinement and further selection for peptide synthesis and *in vitro* assays	(K. Lin et al., 2017)
Bovine blood	ACE inhibitory activity	Self-built regression model	CDOCKER	-	10 descriptors were generated by Discovery Studio and modelled by BPNN for further selection for peptide synthesis and *in vitro* assays; 24 pentapeptides in dataset	(T. Zhang et al., 2015)
Egg	Aminopeptidase N inhibitory peptides	PeptideRanker	CDOCKER	-	PeptideRanker, physicochemical property evaluation, and AMDET evaluation combined for virtual screening; CDOCKER for interaction mechanism	(W. Zhao et al., 2020)

Rice	Immunomodulatory activity	NetMHCpan 4.0	CABS-dock	-	NetMHCpan 4.0 for virtual screening by binding affinity prediction and CABS-dock for interaction mechanism	(L. Wen et al., 2021)
Donkey collagen	Tyrosinase inhibitory activity	PeptideRanker and prediction for water solubility and toxicity	CDOCKER	GROMACS for 60 ns of simulation; Force field: CHARMM36	116 peptides were screened from 140 peptides generated by *in silico* proteolytic simulation. After PeprtideRanker screening and solubility and toxicity prediction, 26 peptides were selected for molecular docking screening. Three peptides were selected by CDOCKER for synthesis and *in vitro* experiments. Molecular dynamics simulation was used for formation refinement and mechanism exploration	(Xue et al., 2022)
Camel milk	Cholesterol esterase inhibitory activity	PeptideRanker	PepSite2 and Glide	-	PeptideRanker and Pepsite2 for rough screening and Glide score and MM-GBSA for refinement; Glide for interaction mechanism	(Mudgil, Baby, Ngoh, Vijayan, et al., 2019)
Camel milk	Esterase- and lipase-inhibitory activity	PeptideRanker	PepSite2 and Glide	-	PeptideRanker for rough screening and Pepsite2 for refinement; Glide for interaction mechanism	(Baba, Mudgil, et al., 2021)

Chapter 2-Supplementary Document S1 A summary of the figures and tables.

The document S1 can be download at here (XLSX). Detailed information of figures and tables are listed as following:

Table S1: 533 numerical indices of the 20 amino acids.

Table S2: Original Pearson correlation coefficients of the 553 numerical indices.

Table S3: Performance of 14 models in 20 times random dataset splitting with leave-one-out cross validation and leave-one-group-our cross validation.

Table S4: The predicted antioxidant activity of the unknown 7790 tripeptides.

Figure S1: Heatmap of the correlation coefficient.

Chapter 2-Supplementary Document S2 Original python scripts

The document S1 can be download at here (ZIP).

Chapter 3-Supplementary Document S1 Five hundred and fifty-three numerical indices of the 20 amino acids

The document S1 can be download at here (XLSX).

Chapter 3-Supplementary Document S2 ABTS radical scavenging capacity and ORAC dataset

The document S2 can be download at here (XLSX).

Chapter 3-Supplementary Document S3 Original Pearson correlation coefficients of the 553 numerical indices

The document S3 can be download at here (XLSX).

Chapter 3-Supplementary Document S4 Heatmap of the correlation coefficient

The document S4 can be download at here (PDF).

Chapter 3-Supplementary Document S5 The detailed python codes for model development

The document S5 can be download at here (ZIP).

Chapter 3-Supplementary Document S6 The predicted antioxidant activity of the possible dipeptides (ABTS radical scavenging capacity & ORAC)

The document S6 can be download at here (XLSX).

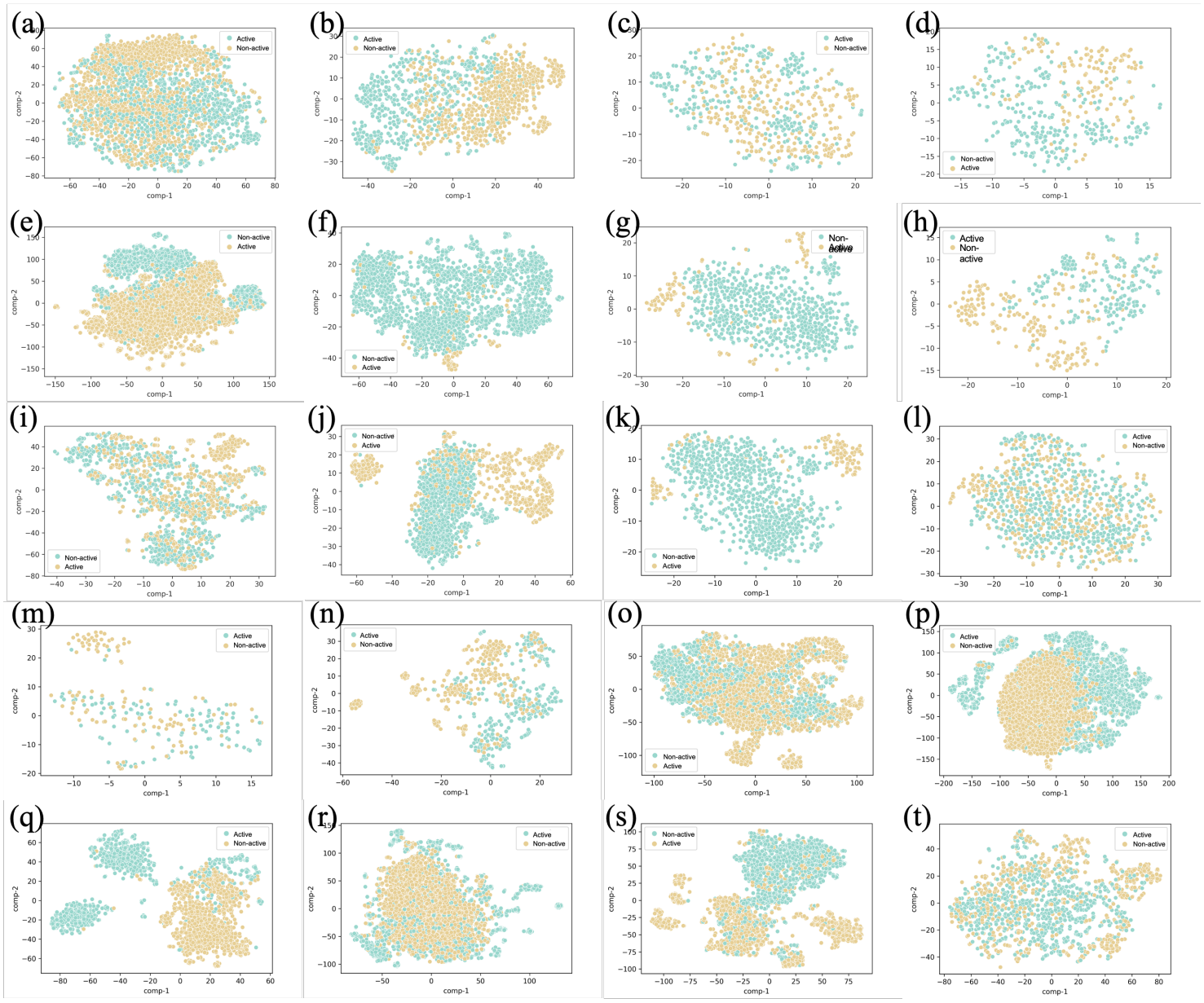
Chapter 3-Supplementary Document S7 Kafirin sequences for hydrolysis simulation with R-PeptideCutter

The document S7 can be download at here (XLSX).

Table S1: The kafirin sequences (fasta format files) and a brief summary

Table S2: all possible kafirin-derived dipeptides with the enzyme information

Table S3: cleavage sites of enzymes and chemicals


(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
(l)
(m)
(n)
(o)
(p)
(q)
(r)
(s)
(t)
Active
Non-active
comp-1
comp-2

Chapter 4- **Supplementary Document S1-**Figure S1 Distributed stochastic neighbor embedding (t-SNE) distribution of positive and negative samples of different bioactive peptide datasets.

Note: (a) ACE inhibitory activity, (b) DPP IV inhibitory activity, (c) Bitter, (d) Umami, (e) Antimicrobial activity, (f) Antimalarial activity (main dataset), (g) Antimalarial activity (alternative dataset), (h) Quorum sensing activity, (i) Anticancer activity (main dataset), (j) Anticancer activity (alternative dataset), (k) Anti-MRSA strains activity, (l) Tumor T cell antigens, (m) blood-brain barrier, (n) anti-parasitic activity, (o) Neuropeptide; (p)Antibacterial activity, (q) antifungal activity, (r) antiviral activity, (s) toxicity, and (t) antioxidant activity. Graphs were generated using the whole dataset with perplexity parameters = 50.

Chapter 5-Supplementary Document S1 Collected dataset from published literatures and preprocessing with CleanLab

The document S1 can be download at here (XLSX).

Chapter 5-Supplementary Document S2 Twenty six prediction model optimization and hyperparameters with 10-fold cross validation

The document S2 can be download at here (WORD). The table captions are listed as following:

Table S1 Model optimization and hyperparameters tuning for ESM-2 embeddings

Table S2 Model optimization and hyperparameters selection for amino acid composition (AAC) embeddings

Table S3 Model optimization and hyperparameters selection for dipeptide composition (DPC) embeddings

Table S4 Model optimization and hyperparameters selection for pseudo amino acid composition (PseAAC) embeddings

Table S5 Model optimization and hyperparameters selection for Amino acid Index (AAI) embeddings

Table S6 Model optimization and hyperparameters selection for overlapping property features (OPF) embeddings

Table S7 Model optimization and hyperparameters selection for binary profile/one-hot encoding (BP) embeddings

Table S8 Model optimization and hyperparameters selection for adaptive skip dipeptide composition (ASDC) embeddings

Table S9 Model optimization and hyperparameters selection for composition-transition-distribution composition (CTDC) embeddings

Chapter 6-Supplementary Document S1 Detailed data set containing the allergenic and nonallergenic sequences

The document S1 can be download at here (XLSX).

Chapter 6-Supplementary Document S2 Hyperparameters tuning of the pLM4Alg models

The document S2 can be download at here (XLSX).

Chapter 6-Supplementary Document S3 Performance comparison between pLM-based models with other existing models

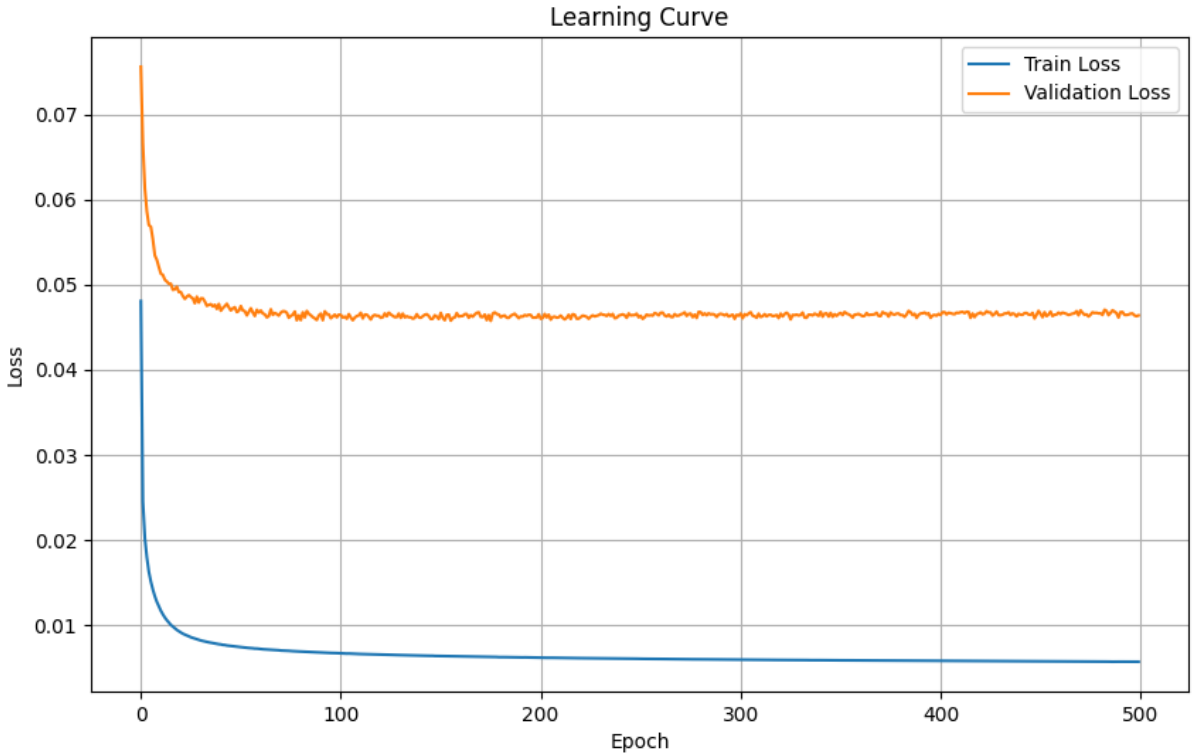
The document S3 can be download at here (XLSX).

Chapter 7- Supplementary Document S1 Table S1 Pretrained protein language model in this study

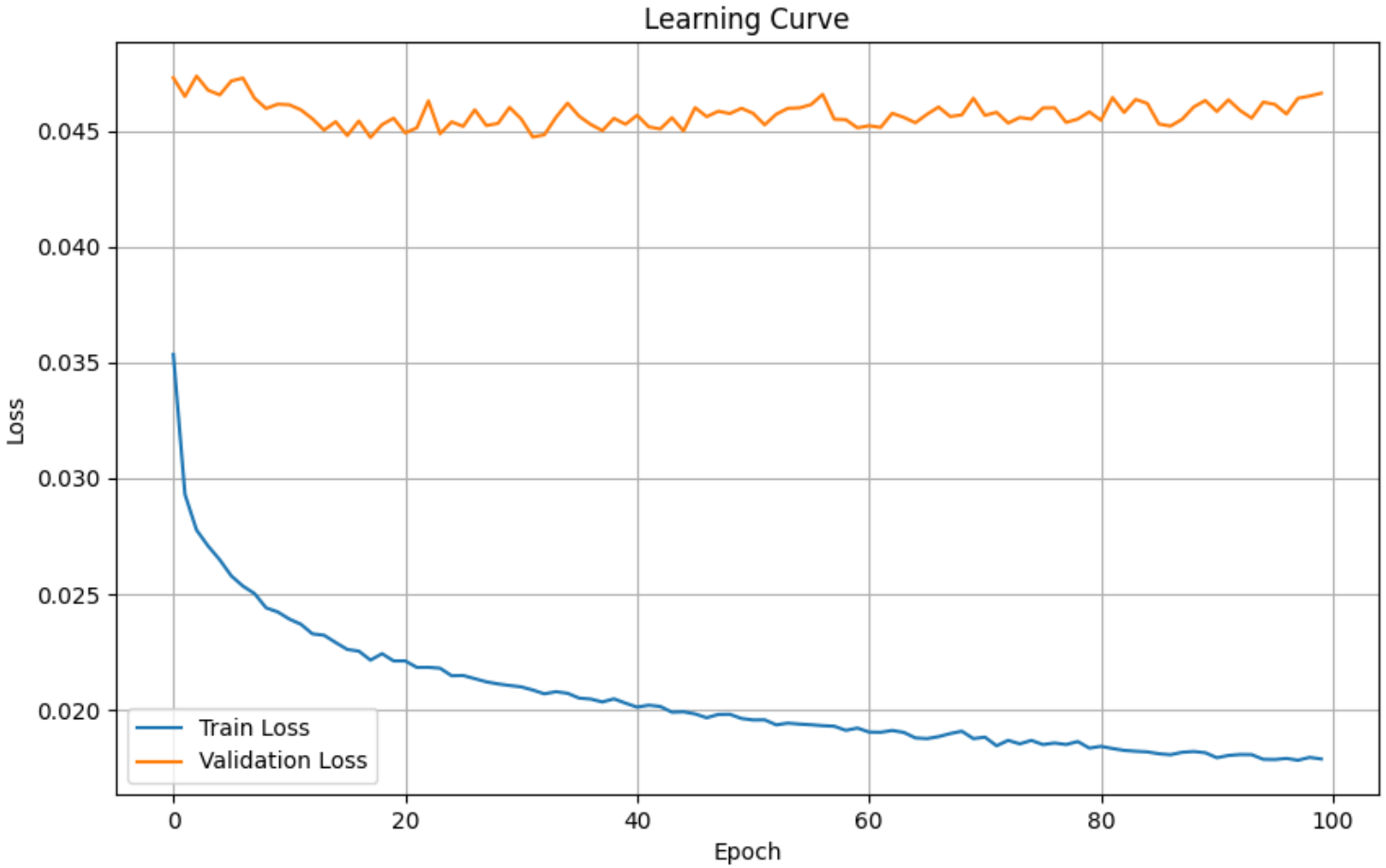
Pretrained model name	Number of layers	Number of parameters	Output dimensions	Dataset
ESM-2-2560	36	3 billion	2560	UR50/D 2021_04
ESM-2-1280	33	650 million	1280	UR50/D 2021_04
ESM-2-480	12	35 million	480	UR50/D 2021_04
ESM-1b	33	650 million	1280	UR50/S 2018_03

Chapter 7- Supplementary Document S2 Figure S1 Learning curves of the model on the phylogenetic evidence-based dataset (a) and the experimental evidence-based dataset

(a)



(b)


Learning Curve
Loss
Epoch
Train Loss
Validation Loss
0.045
0.040
0.035
0.030
0.025
0.020
0
20
40
60
80
100