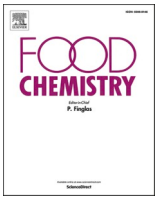
Food Chemistry 431 (2024) 137162

Contents lists available at ScienceDirect

Food Chemistry

journal homepage: www.elsevier.com/locate/foodchem



pLM4ACE: A protein language model based predictor for antihypertensive peptide screening

Zhenjiao Du [a], Xingjian Ding [b], William Hsu [b], Arslan Munir [b], Yixiang Xu [c], Yonghui Li [a,*]

[a] Department of Grain Science and Industry, Kansas State University, Manhattan, KS 66506, USA
[b] Department of Computer Science, Kansas State University, Manhattan, KS 66506, USA
[c] Healthy Processed Foods Research Unit, Western Regional Research Center, USDA-ARS, 800 Buchanan Street, Albany, CA 94710, USA

ARTICLE INFO

Keywords:
ACE inhibitory peptide
Antihypertension
Bioactive peptide
Protein language model
Machine learning

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. The ESM-2 embeddings showed superior performance in enhancing the prediction model compared to the 12 traditional embedding methods. A user-friendly webserver (https://sqzujiduce.us-east-1.awsapprunner.com) with the top three models is now freely available.

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, 2022a).

At this time, >1000 ACE inhibitory peptides have been identified with inhibitory activity reported from *in vitro* or *in vivo* experiments (Minkiewicz and Darewicz, 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, 2022a; 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, & Li, 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, & Li, 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

* Corresponding author.
E-mail address: yonghui@ksu.edu (Y. Li).

https://doi.org/10.1016/j.foodchem.2023.137162
Received 13 April 2023; Received in revised form 9 August 2023; Accepted 13 August 2023
Available online 14 August 2023

properties (Bin et al., 2020; Du, Ding, Xu, & Li, 2023; Du & Li, 2022b; 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., 2021, 2021; Lin et al., 2022; 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., 2022). 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 Fig. 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.

2. Materials and methods

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 1066 entries with sequences, references, and IC_{50} values

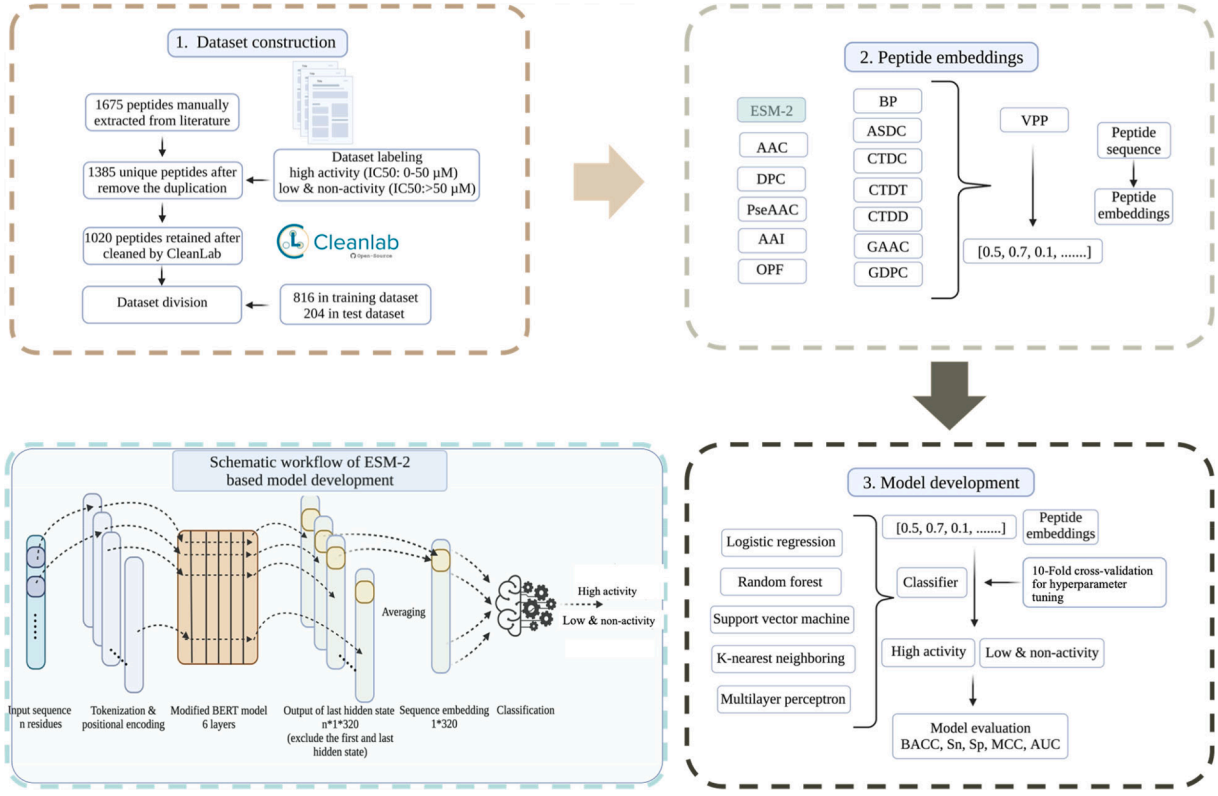


Fig. 1. Workflow of the study for screening peptides with high antihypertensive activity. AAC: amino acid composition; AAI: amino acid index; ASDC: adaptive skip dipeptide composition; BP: binary profile/one-hot encoding; CTDC: composition-transition-distribution composition; CTDT: composition-transition-distribution 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.

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 3364 experimentally validated ACE inhibitory peptides. In Kalyan et al. (2021), a total of 1648 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 1385 unique peptide sequences from 267 different scientific papers, which are presented in Supplementary Document 1. 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 (Fig. 2).

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 1020 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 (Fig. 2). The peptide sequences and corresponding labels are presented in Supplementary Document 1. 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).

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., 2022). 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 (Fig. 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

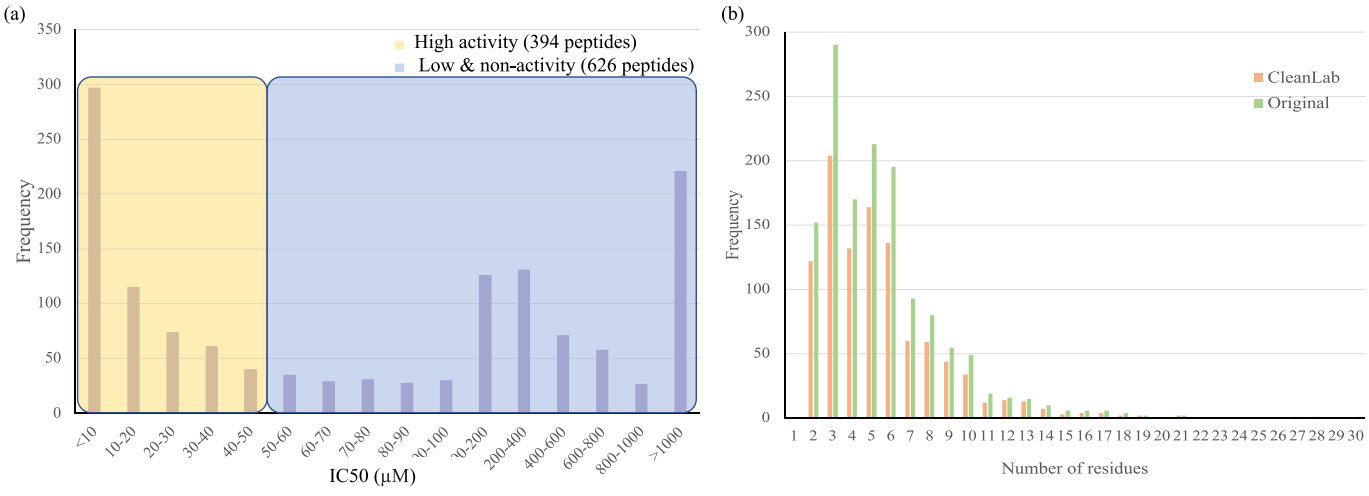


Fig. 2. Dataset profiles: (a) Antihypertensive activity distribution and labeling, (b) Peptide distribution by length before and after data cleaning. Peptides with low & non-activity are grouped together in the binary classification model development and treated as one same category.

(https://github.com/Superzchen/iFeatureOmega-CLI) (Chen et al., 2022). Detailed explanations about the 12 methods are available at https://github.com/dzjxzyd/pLM4ACE.

2.4. Model development

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, Tilley, Wang, Zhang, & Li, 2022). SVM finds the best-separating hyperplane in high-dimensional feature spaces based on different kernel functions (Du, Tian, Tilley, Wang, Zhang, & Li, 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).

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 2 (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:

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

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

$$BACC = 0.5*Sn + 0.5*Sp$$

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

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

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.

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.

3. Results and discussion

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.

Fig. 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). Fig. 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 (Fig. 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 (Fig. 3).

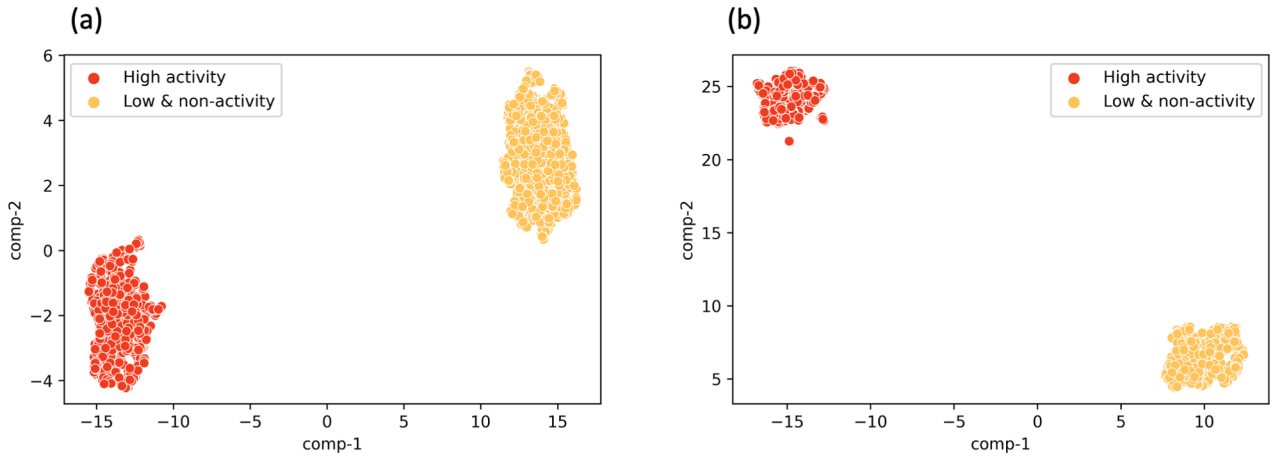


Fig. 3. Uniform manifold approximation and projection (UMAP) of positive and negative samples in the peptide dataset before data cleaning (a) and after data cleaning (b). Peptides with low & non-activity are grouped together in the binary classification model development and treated as one same category.

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 (Fig. 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., 2021; 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., 2022). 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, & Li, 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, & Li, 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.

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 2 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 Fig. 4 and Supplementary Document 2 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

Fig. 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.

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

Table 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

performing model. 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.

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., 2022). 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.

MLP was a popular ML approach for downstream tasks in previous studies (Elnaggar et al., 2021; 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 as in RF. 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).

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 (Fig. 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.

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 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.

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.

CRediT authorship contribution statement

Zhenjiao Du: Methodology, Validation, Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Xingjian Ding:** Investigation, Writing – review & editing. **William Hsu:** Investigation, Writing – review & editing. **Arslan Munir:** Investigation, Writing – review & editing. **Yixiang Xu:** Investigation, Writing – review & editing. **Yonghui Li:** Conceptualization, Methodology, Formal analysis, Investigation, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodchem.2023.137162.

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. Doi: 10.3390/ijms22168958.

Chen, Z., Liu, X., Zhao, P., Li, C., Wang, Y., Li, F., … 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

Chandra, A., Tünnermann, L., Löfstedt, T., & Gratz, R. (2023). Transformer-based deep learning for predicting protein properties in the life sciences. *eLife, 12*, e82819.

Dai, R., Zhang, W., Tang, W., Wynendaele, E., Zhu, Q., Bin, Y., … 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*. Accepted: TrAC Trends in Analytical Chemistry.

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

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

Du, Z., & Li, Y. (2022b). 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., 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., … Rost, B. (2021). ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–1*. 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*, Article 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. Doi: 10.1007/978-1-0716-1418-1.

Kalyan, G., Junghare, V., Khan, M. F., Pal, S., Bhattacharya, S., Guha, 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*, Article 883766. https://doi.org/10.3389/fgene.2022.883766

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., … Rives, A. (2022). 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*. Doi: 10.48550/arXiv.1802.03426.

Minkiewicz, I., & 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*, Article 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., … 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., … 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., … 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). https://doi.org/10.1073/pnas.2016239118. e2016239118.

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. *International Conference on Intelligent Communication and Computational Techniques (ICCT), 2017*, 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/10.1021/acs.jafc.0c04624

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., … Xue, Y. (2023). Deep learning drives efficient discovery of novel antihypertensive peptides from soybean protein isolate. *Food Chemistry, 404*, Article 134690. https://doi.org/10.1016/j.foodchem.2022.134690