Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Prediction of antioxidant peptides using a quantitative structure−activity relationship predictor (AnOxPP) based on bidirectional long short-term memory neural network and interpretable amino acid descriptors

Dongya Qin [a], Linna Jiao [a], Ruihong Wang [a], Yi Zhao [a], Youjin Hao [b], Guizhao Liang [a,*]

[a] *Key Laboratory of Biorheological Science and Technology, Ministry of Education, Bioengineering College, Chongqing University, Chongqing, 400030, China*
[b] *College of Life Sciences, Chongqing Normal University, Chongqing, 401331, China*

ARTICLE INFO

Keywords:
Antioxidant peptide
AnOxPP
Bidirectional long short-term memory (BiLSTM) neural network
Amino acid descriptor (AAD)
Deep learning

ABSTRACT

Antioxidant peptides can protect against free radical-mediated diseases, especially food-derived antioxidant peptides are considered as potential competitors among synthetic antioxidants due to their safety, high activity and abundant sources. However, wet experimental methods can not meet the need for effectively screening and clearly elucidating the structure-activity relationship of antioxidant peptides. Therefore, it is particularly important to build a reliable prediction platform for antioxidant peptides. In this work, we developed a platform, AnOxPP, for prediction of antioxidant peptides using the bidirectional long short-term memory (BiLSTM) neural network. The sequence characteristics of peptides were converted into feature codes based on amino acid descriptors (AADs). Our results showed that the feature conversion ability of the combined-AADs optimized by the forward feature selection method was more accurate than that of the single-AADs. Especially, the model trained by the optimal descriptor SDPZ27 significantly outperformed the existing predictor on two independent test sets (*Accuracy* = 0.967 and 0.819, respectively). The SDPZ27-based AnOxPP learned four key structure-activity features of antioxidant peptides, with the following importance as steric properties > hydrophobic properties > electronic properties > hydrogen bond contributions. AnOxPP is a valuable tool for screening and design of peptide drugs, and the web-server is accessible at http://www.cqudfbp.net/AnOxPP/index.jsp.

1. Introduction

High-level reactive oxygen species (ROS) in human body increases the risks of developing cancers, diabetes, aging, cardiovascular diseases, and neurodegenerative disorders [1]. Antioxidant peptides can effectively eliminate reactive oxygen species (ROS) and block free radical-mediated reactions [2,3]. Moreover, antioxidant peptides have the merits of low-/non-toxicity, abundant food sources, diverse functions [4,5]. These merits confer their great potentials in widely applications.

Massive efforts have been made to develop antioxidant peptides. More than 1000 antioxidant peptides have been isolated from milk, animal- and plant-derived food sources, and seafood using enzymatic digestion, fermentation or autolysis [6,7]. However, the traditional experimental methods are time-consuming and laborious [8]. Currently, computational methods have been used for screening, design and mechanism exploration of antioxidant peptides, including molecular docking, molecular dynamics simulations, bioinformatics modeling, etc. Quantitative structure−activity relationship (QSAR) is a classical ligand-based virtual screening method. QSAR mainly seeks a mathematical relationship between the physicochemical properties of chemical structures and their biological activities [9].

Structural characterization is the core of the QSAR modeling. In QSAR studies of peptides, amino acid descriptors (AADs) are important translators of peptide sequence/structure information. They can effectively describe a variety of property information, such as physicochemical, charged, and geometric properties of peptides [9,10]. Compared with conventional encoding methods, such as one hot residue classification, pseudo amino acid encoding, and k-mer sparse matrix, AADs have the advantages of accurate representation, diverse properties, and interpretability, and can translate the first-level peptide sequence into a high-dimensional vector that contains multidimensional data information [11–13].

Overfitting and feature redundancy in QSAR studies usually make

* Corresponding author. Bioengineering College, Chongqing University, Chongqing, 400030, China.
E-mail address: gzliang@cqu.edu.cn (G. Liang).

https://doi.org/10.1016/j.compbiomed.2023.106591
Received 4 October 2022; Received in revised form 15 January 2023; Accepted 22 January 2023
Available online 24 January 2023

prediction and/or classification unreliable [14]. Recently, machine learning and deep learning techniques effectively overcome these problems, and have successfully identified anti-cancer peptides [15], anti-microbial peptides [16], therapeutic peptides [17], and anti-hypertensive peptides [18]. Especially, deep learning-based QSAR techniques have shown various advantages, e.g. (i) automatically extracting features from raw, high-dimensional, and heterogeneous physicochemical data, (ii) exactly meeting the requirement of modeling for large-scale chemical molecule data, (iii) handling peptide samples with inconsistent sequence lengths, and (iv) effectively avoiding overfitting problems [14,19–21]. However, deep learning-based QSAR has not yet been used for prediction of antioxidant peptides. To the best of our knowledge, AnOxPePred [22] based on deep convolutional neural network (CNN) is currently the only online server for predicting the antioxidant activity of peptides. AnOxPePred displays a prediction performance better than a k-NN sequence identity-based approach on various metrics.

Several issues, in the antioxidant peptide modeling study, including (i) difficulties in feature extraction, (ii) low prediction accuracy, and (iii) poor model transparency and interpretability [14,19,22], still exist. The present work seeks to construct an artificial intelligent (AI)-based QSAR platform for prediction of antioxidant peptides using the bi-directional long short-term memory (BiLSTM) neural network. The BiLSTM neural network models can bi-directionally extract the coding features of more than 1000 residues from peptide sequences [23], which not only effectively avoids the overfitting and sample length inconsistency problems, but also greatly improves the prediction performance [14]. BiLSTM neural network has been applied to the prediction of antifungal peptides [24], antibacterial peptides [25], and human leukocyte antigen I-binding peptides [26]. These models provide important references for the construction of advanced antioxidant peptide predictors.

To solve these issues, we designed a sequence-based interpretable feature representation learning strategy by combining AADs and BiLSTM to explore the structure-activity relationships of antioxidant peptides. We then developed an online server AnOxPP to predict and design antioxidant peptides. This study provides methodological guidance for understanding the structure−activity relationship of peptide sequences.

2. Materials and methods

Fig. 1, shows the construction process of AnOxPP. It includes (i) data collection and preprocessing, (ii) peptide feature generation, (iii) BiLSTM architecture, (iv) model training and evaluation, (v) feature selection and importance score, and (vi) web server designing. Details were described as follows.

2.1. Dataset collection

The sequences of antioxidant peptides were collected from the DFBP (http://www.cqudfbp.net/) and BIOPEP-UWM (https://biochemia.uwm.edu.pl/biopep-uwm/) databases. After removing duplications and deleting samples with inconsistent experimental results, 1060 antioxidant peptides with radical scavenging activities were obtained and used as the positive dataset. For non-antioxidant peptides without experimental evidence, a huge amount of random sequences with different lengths were generated by one in-house Java program. Sequences with more than 90% similarity to the positive samples were removed using CD-HIT [27]. Subsequently, 1060 sequences that have a same length and number distribution with the positive samples were

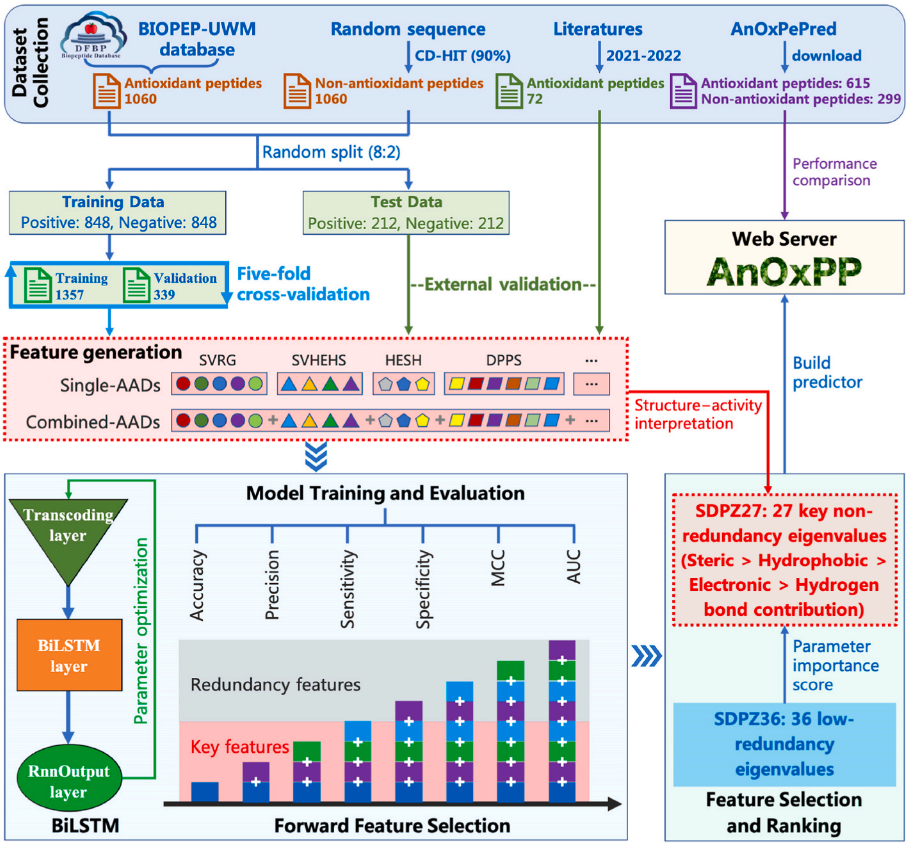

Fig. 1. Construction workflow of the online server AnOxPP. First, to collect the peptides and translate them into feature vectors by single- and combined- AADs. Second, to train BiLSTM models and optimize hyperparameters. Then, to select combined-AADs by using forward feature selection. Next, to interpret the main structure-activity characteristics of antioxidant peptides. Finally, to design and build the web predictor.

randomly selected as the negative samples. The positive and negative sample datasets were randomly split at a ratio of 8:2 to generate the training set and the test set. 72 highly active antioxidant peptides that have no intersection with the above two datasets collected from literatures published in 2021–2022 were used as the second validation set. In addition, the AnOxPePred dataset was downloaded from the AnOxPePred web site to compare the predictive performance of the AADs-based BiLSTM neural network model. We provided download links for all datasets (Table S1).

2.2. Sequence feature encoding

Total 22 AADs that describe the types and physicochemical properties of natural amino acids were collected from literatures (Table S2). These AADs were used to characterize each residue of the peptides, thereby to generate multiple matrixes (number of residues × number of feature descriptors) for a peptide. These matrixes were used as the input parameters of the BiLSTM network. We adopted two coding strategies: (i) Single-AADs: 22 AADs were input as independent codes to train the model; (ii) Combined-AADs: the key feature values of the residues were sequentially extracted by the forward selection method. As a result, the best feature codes were selected from 22 single-AADs based on the prediction accuracy of the trained model on the test set. As shown in Fig. S1A, the selected single-AADs was then linearly spliced with the rest of the single-AADs in turn, and the combined codes that maximizes the model prediction accuracy were screened out. The above iterative processes were repeated until 22 single-AADs were concatenated into a 20 × 179 matrix. Finally, the combination descriptors with the highest prediction accuracy were selected as the optimal codes. Single-AADs and combined-AADs are defined as follows：

$$\text{Single}-\text{AADs}=\begin{bmatrix} V_1^1 & V_2^1 & \cdots & V_n^1 \\ V_1^2 & V_2^2 & \cdots & V_n^2 \\ \vdots & \vdots & \vdots & \vdots \\ V_1^{20} & V_2^{20} & \cdots & V_n^{20} \end{bmatrix}$$

where V denotes the characteristic variables of the 20 natural amino acids, n represents the number of variables, and the dimension of the single-AADs characteristic matrix is 20 × n.

$$\text{Combined}-\text{AADs}=\begin{bmatrix} V(1)_1^1 & \cdots & V(1)_n^1 & \cdots & V(m)_1^1 & \cdots & V(m)_t^1 \\ V(1)_1^2 & \cdots & V(1)_n^2 & \cdots & V(m)_1^2 & \cdots & V(m)_t^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ V(1)_1^{20} & \cdots & V(1)_n^{20} & \cdots & V(m)_1^{20} & \cdots & V(m)_t^{20} \end{bmatrix}$$

where V denotes the characteristic variables of the 20 natural amino acids, n and t represent the number of variables encoded by different single-AADs, and m denotes the numbering of the 22 groups of single-AAD, and its range is 1–22.

2.3. BiLSTM architecture

To build a web server, the BiLSTM network in the Deeplearning 4j framework (DL4J: https://deeplearning4j.org/) was used to train the model. As shown in Fig. 2, three network layers were sequentially set up and grid search was used to optimize the parameters, including the sequence transcoding, BiLSTM, and recurrent neural network ouput (RnnOutput) layers. The sequence transcoding layer was used to convert the peptide sequences into the feature vectors. The BiLSTM layer consisted of a forward layer and a backward layer. The input length was equal to the code length of AADs, and the activation function was set

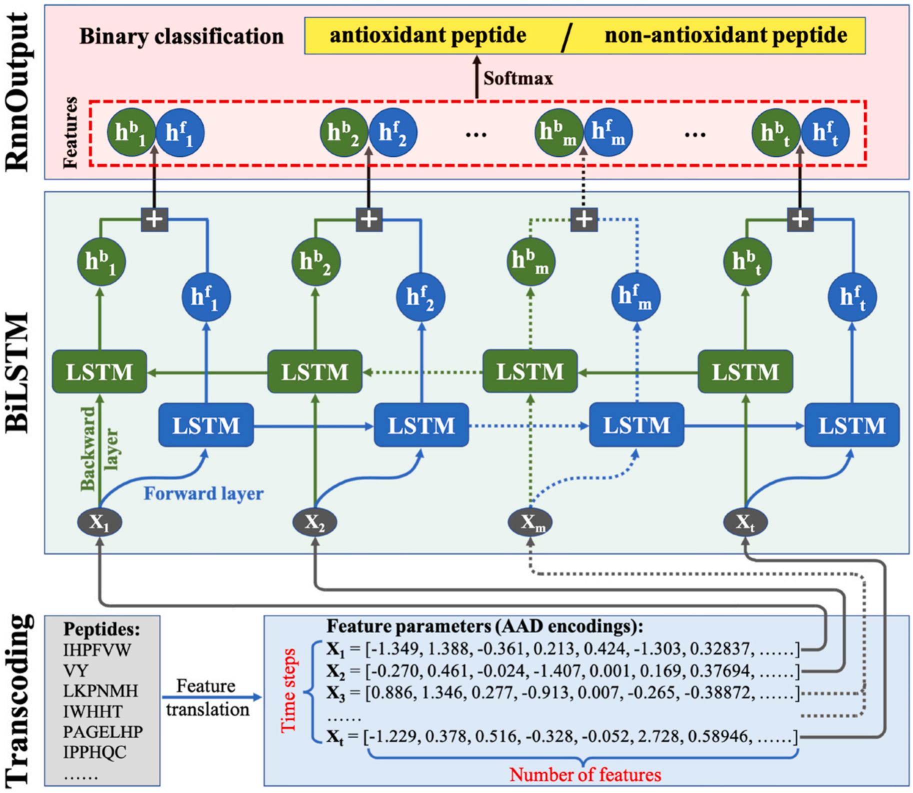
RnnOutput
Binary classification
antioxidant peptide / non-antioxidant peptide
Softmax
Features
h^b_1 h^f_1 h^b_2 h^f_2 … h^b_m h^f_m … h^b_t h^f_t
BiLSTM
LSTM
Backward layer
Forward layer
X_1 X_2 X_m X_t
Transcoding
Peptides:
IHPFVW
VY
LKPNMH
IWHHT
PAGELHP
IPPHQC
……
Feature translation
Feature parameters (AAD encodings):
Time steps
X_1 = [-1.349, 1.388, -0.361, 0.213, 0.424, -1.303, 0.32837, ……]
X_2 = [-0.270, 0.461, -0.024, -1.407, 0.001, 0.169, 0.37694, ……]
X_3 = [0.886, 1.346, 0.277, -0.913, 0.007, -0.265, -0.38872, ……]
……
X_t = [-1.229, 0.378, 0.516, -0.328, -0.052, 2.728, 0.58946, ……]
Number of features

Fig. 2. BiLSTM architecture flowchart. Three network layers were set up, including sequence transcoding layer, BiLSTM layer, and RnnOutput layer. The BiLSTM layer consists of a forward LSTM layer and a backward LSTM layer.

"tanh". The hidden layer outputs of the forward and backward LSTM networks were fed to the output layer through a linear connection (length 128). The activation function of the RnnOutput layer was set to "sigmoid" for binary classification (antioxidant peptides or non-antioxidant peptides). The detailed calculation process of the LSTM unit was performed as described by Hochreiter S. et al. [23,28].

2.4. Evaluation of performance

Six evaluation indexes, including Precision, Sensitivity, Specificity, Accuracy, Matthew's correlation coefficient (MCC), and the area under the receiver operating characteristics curves (AUC), that obtained from five-fold cross validation and external validation were used to evaluate the predictive ability of the model. They were defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

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

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

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

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

where TP, FP, TN, and FN represented the number of correctly predicted antioxidant peptides, incorrectly predicted antioxidant peptides, correctly predicted non-antioxidant peptides, and incorrectly predicted non-antioxidant peptides, respectively.

Currently, there is no standard parameter importance evaluation method for the BiLSTM neural network, thus we masked the single or multiple parameters of the samples with "0" to discuss the importance of single or same attribute parameters (Fig. S2), and the score was calculated by the following formula:

$$Importance\ score = (Accuracy_{nonMask} - Accuracy_{Mask}) \times 100$$

$\text{Accuracy}_{\text{nonMask}}$ and $\text{Accuracy}_{\text{Mask}}$ represent the prediction accuracy of unmasked or masked features using the trained model. $\text{Accuracy}_{\text{nonMask}}$ did not mask the feature parameters, but directly converted all features of the peptide sequences. While, $\text{Accuracy}_{\text{Mask}}$ masked parameters and converted peptide sequences into input parameters containing partial features. An importance score ≤ 0 indicated that the corresponding feature parameter had non or negative contribution to the model (i.e. redundant feature). An importance score >0 indicated that the corresponding parameter had a positive contribution to the model (i.e. key feature).

2.5. Web server implementation

The Web server was built using the HTML, JSP and JavaScript. The MySQL database was used for the data storage, and Java Web technology was applied to deploy AnOxPP to the Tencent Cloud Server (Windows Server 2012 R2, 64-bit). Model calls were implemented using the Java programs, and sequence features with experimental evidence and the predicted sequences can be visualized using the EChart (https://echarts.apache.org/en/index.html). The developed AnOxPP platform can serve to predict, design, and analyze the structure-activity relationship of antioxidant peptides.

3. Results and discussion

3.1. Excellent predictive performance of single-AADs-based BiLSTM models

Total 22 well-defined single-AADs with different dimensions that were collected from 22 literatures were used to describe the key features of the residues of the antioxidant peptides (Table S2). These single-AADs are One hot, DPPS, BLOSUM62, FASGAI, GRID, HESH, ISA-ECI, Lin's scales, MS-WHIM, NNAAIndex, ProtFP, QTMS (ADFQ) indices, ST-scales, SVRG, SVWG, SVHEHS, T-scales, VHSE, VSTV, VSW, Z-scales, and P-scales with the dimensions of 20, 10, 10, 6, 7, 12, 2, 3, 3, 6, 8, 7, 8, 16, 10, 13, 5, 8, 3, 9, 3 and 10, respectively. Each AAD can confer a *NA (number of amino acid residues of a peptide)* × *D* (*dimension of an AAD*) parameter matrix for a peptide. These parameters were used as the input of the BiLSTM model. The training set was first applied to train the single-AADs-based BILSTM models by five-fold cross validation. Then, the test set was used to evaluate the generalization ability of the trained model. The cross-validation results showed that only 17 single-AADs (feature number 5–20) contribute to the high prediction performance of the models, and gave rise of *Accuracy* of 0.828–0.949, *MCC* of 0.656–0.899, and *AUC* of 0.903–0.987 for the models (Table S3). Additionally, the external prediction of the independent test set using the 17 models also yielded *Accuracy* of 0.831–0.942, *MCC* of 0.663–0.885, and *AUC* of 0.899–0.984 (Table 1), indicating their excellent predictive capability. In summary, these results suggested that the single-AADs can effectively characterize the structure-activity information of the peptide sequences, and the BiLSTM models can accurately predict the antioxidant activity of the peptides.

3.2. SDPZ36 coding exhibited better characterization ability than single-AADs

To optimize low-redundancy AADs, the single-AADs were linearly combined into the combined-AADs that have more eigenvalues. The new combined-AADs were further used as the input for BiLSTM model training. Five-fold cross validation was performed to train the benchmark training set, during which the features were iteratively selected using the forward feature selection method (Table S4). Then, the trained models were used to predict the antioxidant activity of the samples of the independent test set. The optimal features were further screened based on the external prediction accuracy (Table 2). Results from the cross and external validations implicated that the models with the input of the combined-AADs had the better predictive capability and were more robust than that with the single-AADs as the input (Table S4 and Table 2). As displayed in Fig. 3, the optimal models using the combined-AADs in each round had *Accuracy* ranging from 0.942 to 0.967, apparently higher than that the single-AADs-based models (*Accuracy*≤0.942), indicating that the combined features were helpful for the model to learn more features of antioxidant peptides. As the number of feature parameters increased from 2 to 36, the prediction accuracy of the model was gradually improved, indicating that the quantity and quality of the effective features were continuously accumulated. The model exhibited the best predictive performance when 36 features were selected, suggesting that the eigenvalues of the combined code have been accumulated to the critical point. Combined_3 consisting of SVHEHS, DPPS, P-scales and Z-scales was named SDPZ36 and demonstrated as the best combined-AADs (Table S5). The specific screening process is shown in Fig. S1A. The average prediction accuracy of the SDPZ36 model on the independent test set was 0.967, 2.46% higher than that of the optimal single-AADs-based model, the SVHEHS model. As the number of features further increased, the prediction accuracy was stabilized at ~0.967. However, when the number reached 179, the accuracy was reduced to 0.942, suggesting that redundant features not only increased the invalid weight of the model, but also increased the training difficulty. Therefore, the combination of SDPZ36 was considered as a set of low-redundancy

Table 1
Performance of models based on 22 single-AADs on the independent test set.

No.	Descriptor	Matrix	Accuracy	Precision	Sensitivity	Specificity	MCC	AUC
1	ISA-ECI	20 × 2	0.6582	0.6792	0.6141	0.7024	0.3230	0.7037
2	MS-WHIM	20 × 3	0.6522	0.6553	0.6438	0.6606	0.3058	0.6834
3	Lin's scales	20 × 3	0.7532	0.7845	0.6997	0.8067	0.5100	0.8328
4	VSTV	20 × 3	0.7236	0.7115	0.7535	0.6936	0.4484	0.8080
5	Z-scales	20 × 3	0.7502	0.7339	0.7886	0.7118	0.5028	0.8382
6	T-scales	20 × 5	0.8586	0.8527	0.8687	0.8485	0.7177	0.9320
7	NNAAIndex	20 × 6	0.8333	0.8502	0.8094	0.8572	0.6675	0.9079
8	FASGAI	20 × 6	0.9034	0.9077	0.8983	0.9084	0.8070	0.9638
9	QTMS	20 × 7	0.8310	0.8301	0.8330	0.8290	0.6626	0.8988
10	GRID	20 × 7	0.9316	0.9410	0.9212	0.9421	0.8636	0.9778
11	ST-scales	20 × 8	0.9226	0.9279	0.9165	0.9286	0.8453	0.9727
12	VHSE	20 × 8	0.8859	0.8840	0.8882	0.8835	0.7720	0.9482
13	ProtFP	20 × 8	0.9091	0.9043	0.9152	0.9030	0.8183	0.9637
14	VSW	20 × 9	0.9407	0.9366	0.9455	0.9360	0.8815	0.9786
15	BLOSUM62	20 × 10	0.8502	0.8451	0.8579	0.8424	0.7007	0.9255
16	P-scales	20 × 10	0.9155	0.9078	0.9253	0.9057	0.8313	0.9657
17	DPPS	20 × 10	0.9256	0.9196	0.9327	0.9185	0.8513	0.9722
18	SVWG	20 × 10	0.9347	0.9391	0.9300	0.9394	0.8698	0.9718
19	HESH	20 × 12	0.9205	0.9150	0.9273	0.9138	0.8412	0.9733
20	**SVHEHS**	**20×13**	**0.9421**	**0.9550**	**0.9279**	**0.9562**	**0.8846**	**0.9842**
21	SVRG	20 × 16	0.9363	0.9491	0.9222	0.9505	0.8732	0.9788
22	One hot	20 × 20	0.9327	0.9530	0.9104	0.9549	0.8664	0.9738

Table 2
Model performance and evaluation indexes based on combined-AADs on the independent test set.

No.	Coding	Added AADs (Coding length)[a]	Matrix	Accuracy	Precision	Sensitivity	Specificity	MCC	AUC
1	SVHEHS	SVHEHS (13)	20 × 13	0.9421	0.9550	0.9279	0.9562	0.8846	0.9842
2	Combined_1	DPPS (10)	20 × 23	0.9626	0.9778	0.9468	0.9785	0.9259	0.9922
3	Combined_2	P-scales (10)	20 × 33	0.9633	0.9720	0.9542	0.9724	0.9268	0.9908
4	**Combined_3 (SDPZ36)**	**Z-scales (3)**	**20×36**	**0.9667**	**0.9813**	**0.9515**	**0.9818**	**0.9338**	**0.9907**
5	Combined_4	VSW (9)	20 × 45	0.9616	0.9793	0.9434	0.9798	0.9241	0.9947
6	Combined_5	QTMS (7)	20 × 52	0.9589	0.9783	0.9387	0.9791	0.9187	0.9928
7	Combined_6	HESH (12)	20 × 64	0.9566	0.9756	0.9367	0.9764	0.9139	0.9938
8	Combined_7	ST-scales (8)	20 × 72	0.9549	0.9774	0.9313	0.9785	0.9108	0.9927
9	Combined_8	ProtFP (8)	20 × 80	0.9532	0.9680	0.9374	0.9690	0.9069	0.9918
10	Combined_9	T-scales (5)	20 × 85	0.9626	0.9771	0.9475	0.9778	0.9258	0.9948
11	Combined_10	Lin's scales (3)	20 × 88	0.9603	0.9797	0.9401	0.9805	0.9213	0.9926
12	Combined_11	VSTV (3)	20 × 91	0.9586	0.9803	0.9360	0.9811	0.9182	0.9936
13	Combined_12	GRID (7)	20 × 98	0.9582	0.9776	0.9380	0.9785	0.9173	0.9925
14	Combined_13	BLOSUM62 (10)	20 × 108	0.9569	0.9742	0.9387	0.9751	0.9145	0.9921
15	Combined_14	NNAAIndex (6)	20 × 114	0.9569	0.9755	0.9374	0.9764	0.9146	0.9920
16	Combined_15	One hot (20)	20 × 134	0.9532	0.9707	0.9347	0.9717	0.9071	0.9917
17	Combined_16	MS-WHIM (3)	20 × 137	0.9559	0.9715	0.9394	0.9724	0.9123	0.9921
18	Combined_17	FASGAI (6)	20 × 143	0.9562	0.9683	0.9434	0.9690	0.9128	0.9935
19	Combined_18	SVRG (16)	20 × 159	0.9549	0.9735	0.9354	0.9744	0.9106	0.9767
20	Combined_19	SVWG (10)	20 × 169	0.9586	0.9763	0.9401	0.9771	0.9179	0.9642
21	Combined_20	ISA-ECI (2)	20 × 171	0.9434	0.9583	0.9273	0.9596	0.8873	0.9804
22	Combined_21	VHSE (8)	20 × 179	0.9424	0.9564	0.9273	0.9576	0.8853	0.9846
23	**SDPZ27**	**SVHEHS (11) + DPPS (8) + GRID (7) + Z-Scales (1)**	**20×27**	**0.9670**	**0.9901**	**0.9434**	**0.9906**	**0.9350**	**0.9949**

[a] Newly added single-AADs during forward feature selection.

codes that can accurately characterize the antioxidant activity of peptides.

Above results showed that the conversion from natural language of peptide sequences to machine language was closely related to the encoded information of AADs, including the number of features, physicochemical significance, and feature redundancy. Therefore, the combined-AADs had more efficient feature transformation capability than the single-AADs, and can do better in comprehensive and accurate characterization of the important characteristics of the peptide residues. The BiLSTM model based on SDPZ36 learned the structure-activity relationships of peptide sequences, eliciting an excellent generalization ability of the model. Therefore, the SDPZ36 was vastly important for further deciphering the structure-activity relationships of antioxidant peptides.

3.3. SDPZ27 coding interpreted the key structure-activity features of antioxidant peptides

SDPZ36 is a 20 × 36 parameter matrix which can characterize 36 attributes of each amino acid. The scores of the 5th, 8th, 17th, 23rd, 26th, 32nd, 33rd, 35th, and 36th parameters characterized by SDPZ27 were less than 0 (Fig. 4A), indicating that these parameters were redundant and had no contribution to the identification of antioxidant peptides. Therefore, these redundant features were removed and a new set of SDPZ27 eigenvalues containing hydrophobic, electronic, hydrogen bond, and steric properties was obtained (The overall screening flowchart of SDPZ36 and SDPZ27 were described in Fig. S1). SDPZ27 was a 20 × 27 parameter matrix characterizing 27 features of each amino acid (Table S5 and Table S6). The SDPZ27-based model did not show the remarkable difference from the SDPZ36-based model in predicting the antioxidant activity of the test set samples (Table 2). The 27 eigenvalues positively contributed to the prediction performance of

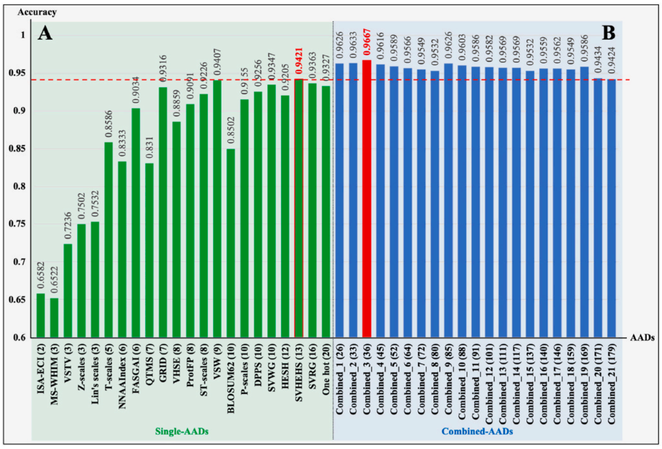

Fig. 3. The average predicted accuracy on the independent test set based on different AADs. (A) 22 single-AADs. (B) 21 combined-AADs. The number in the parentheses represents the number of feature encodings.

the model (Importance score >0). The steric and hydrophobic properties did the most contribution to the model (Fig. 4B). The 27 parameters were divided into four categories (Fig. 4C and Table S7), including steric property, hydrophobic property, electronic property, and hydrogen bond contribution consisting of 9, 7, 9, and 2 parameters, respectively. Their contributions to the developed model showed the decreasing order of steric properties > hydrophobic properties > electronic properties > hydrogen bond contributions (Fig. 4D). These results suggested that the key features encoded by SDPZ27 were suitable for model learning and identification of antioxidant peptides or non-antioxidant peptides.

Previous studies have shown that molecular weight, spatial structure, hydrophobicity, amino acid composition, and distribution of peptides were the main factors affecting their antioxidant activity [29–31]. The peptide samples in the dataset we used indicated that the four properties encoded by SDPZ27 were closely related to the sequence features of antioxidant peptides: (i) antioxidant peptides are mainly composed of 2–10 amino acid residues (Fig. B and S3A); (ii) Two terminals of antioxidant peptides prefer hydrophobic amino acids (Leu, Ala, Pro, and Arg) and aromatic amino acids (Tyr). In addition, its N-terminus also prefers Val and Gly, while its C-terminus prefers Lys and Glu (Fig. S3C); (iii) antioxidant peptides are rich in hydrophobic amino acids (Pro, Leu, Ala, and Val) (6.17%–10.14%), hydrophilic amino acids Gly (8.05%), aromatic amino acids Tyr (6.73%), charged amino acids Glu (6.0%) and His (4.77%) (Fig. S3D). These key residues contain special functional groups (e.g. oxhydryl, guanidyl, imidazole, benzene ring, and ε-amino), which can be used as electron, hydrogen and proton donors to increase the antioxidant activity of peptides [32,33]. Therefore, SDPZ27 accurately characterized the sequence length, N-/C-terminus, amino acid residues, spatial configuration, hydrophobicity, and charged properties of antioxidant peptides.

Our results were consistent with the previous studies that four

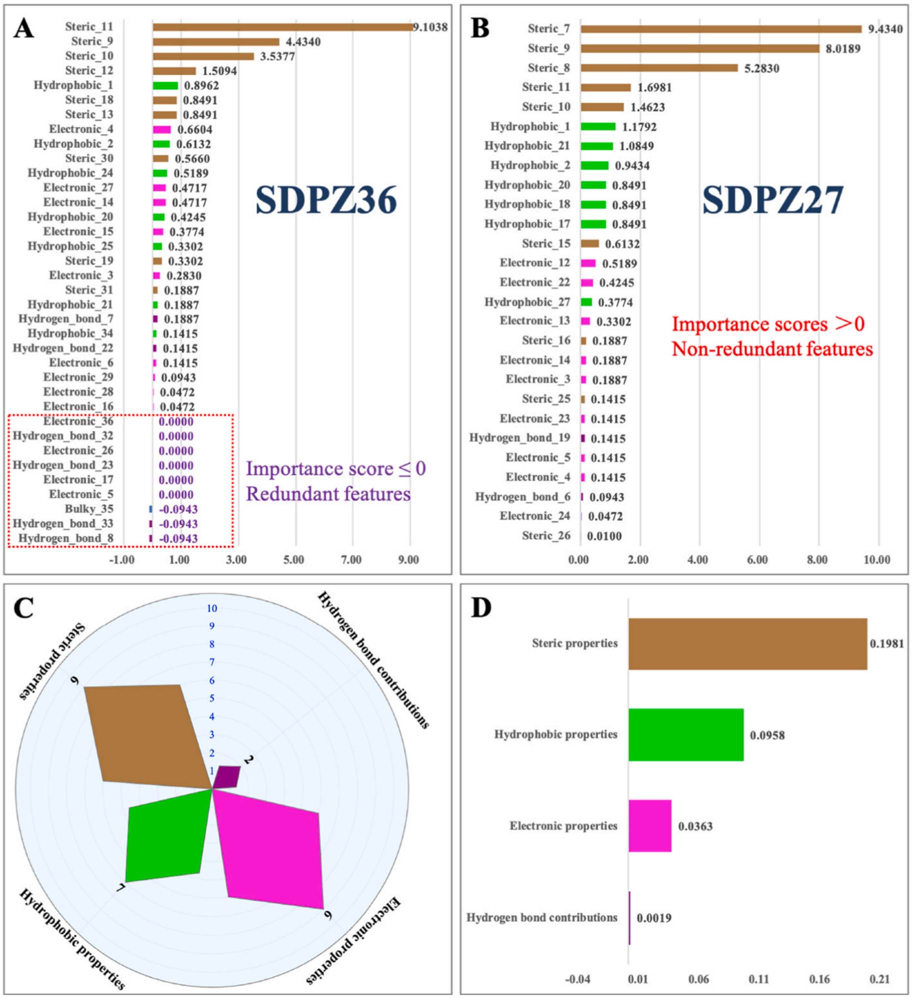

Fig. 4. Feature importance comparison between SDPZ36 and SDPZ27. (A) *Importance scores* for 36 feature parameters encoded in SDPZ36; (B) *Importance scores* for 27 feature parameters encoded in SDPZ27; (C) Four attributes of SDPZ27 feature encoding; (D) *Importance scores* for four attributes encoded in SDPZ27.

pattern attributes encoded by SDPZ27 were key factors affecting the activities of antioxidant peptides. The antioxidant capacities of protein hydrolysates were negatively correlated with the size of their average molecular weights, such as cartilage collagen [34] and marine red algae [35]. Shorter peptides (<1 kDa) were more likely to interact with free radicals and could effectively inhibit the chain reactions of lipid peroxidation and scavenge ROS [36,37]. Hydrophobic amino acid residues of antioxidant peptides, such as Pro, Leu Val and Ala, promoted the interactions of peptides with free radicals by increasing the density of the water-lipid interfaces, which further enhanced their antioxidant activities [38,39]. Aromatic amino acids (Tyr, Trp, and Phe) stabilized reactive oxygen of radicals through transferring electron or proton from the aromatic groups [38,40]. For hydrophilic residues in the peptides, Gly was beneficial for maintaining the strong flexibility of the peptide backbone and could also serve as a hydrogen donor to eliminate ROS [41,42], Glu and Lys had a positive influence on Fe^{2+} chelating ability [43], and His could significantly promote the DPPH radical scavenging activity of peptides due to the high proton donating ability of its imidazole ring [44]. Polar amino acids, such Glu, Asp and Arg, could enhance the ability of chelating metal ion and scavenging HO· of peptides [45–47]. Taken together, the multi-attribute and non-redundancy matrix encoded by SDPZ27 characterized the sequence/structure features that are closely related to the activity of antioxidant peptides. In addition, not only the SDPZ27-based BiLSTM model exhibited robustness and high accuracy in prediction, but also provided a scientific explanation for the key features of antioxidant peptides.

3.4. AnOxPP outperformed the existing predictor

To verify excellent performance of AnOxPP, we compared its prediction ability with the existing predictor, AnOxPePred, using multiple datasets. AnOxPePred is a model based on One hot coding and CNN [22]. For AnOxPePred, the score value of 0.4 was obtained when it predicted its own dataset, generating optimal *Accuracy* (0.753) and *MCC* (0.433), *Sensitivity* (0.810), *Specificity* (0.627), and *Precision* (0.826) (Table S8). Therefore, we used 0.4 as the threshold for performance evaluation. Our BiLSTM model trained with SDPZ27 gave rise of 0.807 for *Accuracy*, 0.558 for *MCC*, 0.861 for *Sensitivity*, 0.695 for *Specificity*, and 0.854 for *Precision* when predicting the AnOxPePred dataset (Table 3), an indicative of better performance for AnOxPP than AnOxPePred. The fact that AnOxPP outperformed AnOxPePred can be further supported by the prediction of the independent test set for external validation and another new dataset containing 72 antioxidant peptides. As shown in Table 3, AnOxPP produced 0.967 for *Accuracy*, 0.935 for *MCC*, 0.943 for *Sensitivity*, 0.991 for *Specificity*, and 0.990 for *Precision* when predicting the independent test set, and 0.819 for *Accuracy* when predicting the new antioxidant peptide dataset. All of these indexes are higher than that of AnOxPePred (*Accuracy* = 0.583, *MCC* = 0.168, *Sensitivity* = 0.570, *Specificity* = 0.600, and *Precision* = 0.670 for prediction of the independent test set, and *Accuracy* = 0.611 for prediction of the new antioxidant peptide dataset). Therefore, above results indicated the BiLSTM model based on SDPZ27 has stronger learning ability than the CNN model based on One hot, and AnOxPP outperformed AnOxPePred.

3.5. Web server implementation

The AnOxPP server (http://www.cqudfbp.net/AnOxPP/index.jsp) was built based on the optimal BiLSTM model trained by SDPZ27. AnOxPP consists of six modules, including Home, Pre-AnOxPs, Seq-Features, Pre-Libraries, Help, and Contact (Fig. S4). The AnOxPP platform allows prediction, residue-based mutation screening, and structure-activity analysis of new antioxidant peptides.

4. Conclusion

In this study, a QSAR model AnOxPP based on the BiLSTM neural network and the optimized SDPZ27 feature matrix was established to predict the antioxidant activity of food-derived peptides. The non-redundant code SDPZ27 optimized by the forward selection method showed efficient conversion of sequence features. The importance of four decisive features of antioxidant peptides showed a decreasing order of steric properties > hydrophobic properties > electronic properties > hydrogen bond contributions, suggesting that they were important factors affecting antioxidant activities of peptides. Importantly, AnOxPP outperformed the existing model AnOxPePred in prediction of three datasets, and accuracy were 0.8066, 0.9670, and 0.8194 for AnOxPePred dataset, the independent test set, and the new antioxidant peptide dataset, respectively. In conclusion, AnOxPP helps to deeply understand the structure−activity relationship of antioxidant peptides and provide a methodological reference for the application of deep learning to study bioactive peptides.

Author contributions

Dongya Qin: Conceptualization, Data curation, Programming, Formal analysis, Writing-original draft.
Linna Jiao: Software, Methodology.
Ruihong Wang: Visualization, Methodology.
Yi Zhao: Data collection, Software.
Youjin Hao: Writing-review & editing.
Guizhao Liang: Writing-review & editing, Supervision, 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

Table 3
Performance comparison of AnOxPP and AnOxPePred.

Dataset	Samples		Predictor	Accuracy	MCC	Sensitivity	Specificity	Precision
	Antioxidant peptides	Non-antioxidant peptides						
AnOxPePred dataset	615	299	BiLSTM[a]	0.8066	0.5581	0.8607	0.6949	0.8537
			AnOxPePred	0.7527	0.4327	0.8102	0.6272	0.8260
Independent test set	212	212	AnOxPP[b]	0.9670	0.9350	0.9434	0.9906	0.9901
			AnOxPePred	0.5825	0.1677	0.5703	0.6000	0.6698
New antioxidant peptide dataset (Reported in 2021 and 2022)	72	0	AnOxPP	0.8194	–	–	–	–
			AnOxPePred	0.6111	–	–	–	–

[a] BiLSTM represents the SDPZ27 encoding-based BiLSTM model trained on the dataset provided by AnOxPePred, and the average evaluation metrics for the five-fold cross-validation are listed.

[b] AnOxPP denotes the optimal model trained on the training set constructed in this study by using SDPZ27 encoding and BiLSTM neural network. The AnOxPePred dataset used was provided by AnOxPePred. Score = 0.4 was considered as the effective classification threshold. The classification results of all thresholds were shown in Supplementary Tables S8–S12.

the work reported in this paper.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 32172196 and 31771975).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2023.106591.

References

[1] F. Sabbatino, V. Conti, L. Liguori, et al., Molecules and mechanisms to overcome oxidative stress inducing cardiovascular disease in cancer patients, Life 11 (2021) 105, https://doi.org/10.3390/life11020105.

[2] B.P. Singh, R.E. Aluko, S. Hati, et al., Bioactive peptides in the management of lifestyle-related diseases: current trends and future perspectives, Crit. Rev. Food Sci. Nutr. (2021) 1–14, https://doi.org/10.1080/10408398.2021.1877109.

[3] C. Lammi, G. Aiello, G. Boschin, et al., Multifunctional peptides for the prevention of cardiovascular disease: a new concept in the area of bioactive food-derived peptides, J. Funct.Foods 55 (2019) 135–145, https://doi.org/10.1016/j.jff.2019.02.016.

[4] C.F. Balthazar, J.F. Guimarães, N.M. Coutinho, et al., The future of functional food: emerging technologies application on prebiotics, probiotics and postbiotics, Compr. Rev. Food Sci. Food Saf. 21 (2022) 2560–2586, https://doi.org/10.1111/1541-4337.12962.

[5] A. Henninot, J.C. Collins, J.M. Nuss, The current state of peptide drug discovery: back to the future? J. Med. Chem. 61 (2018) 1382–1414, https://doi.org/10.1021/acs.jmedchem.7b00318.

[6] D. Qin, W. Bo, X. Zheng, et al., DFBP: a comprehensive database of food-derived bioactive peptides for peptidomics research, Bioinformatics 38 (2022) 3275–3280, https://doi.org/10.1093/bioinformatics/btac323.

[7] G. López-García, O. Dublan-García, D. Arizmendi-Cotero, et al., Antioxidant and antimicrobial peptides derived from food proteins, Molecules 27 (2022), https://doi.org/10.3390/molecules27041343.

[8] Z. Karami, B. Akbari-Adergani, Bioactive food derived peptides: a review on correlation between structure of bioactive peptides and their functional properties, J. Food Sci. Technol. 56 (2019) 535–547, https://doi.org/10.1007/s13197-018-3549-4.

[9] W. Bo, L. Chen, D. Qin, et al., Application of quantitative structure-activity relationship to food-derived peptides: methods, situations, challenges and prospects, Trends Food Sci. Technol. 114 (2021) 176–188, https://doi.org/10.1016/j.tifs.2021.05.031.

[10] A.B. Nongonierma, R.J. FitzGerald, Learnings from quantitative structure-activity relationship (QSAR) studies with respect to food protein-derived bioactive peptides: a review, RSC Adv. 6 (2016) 75400–75413, https://doi.org/10.1039/c6ra12738j.

[11] L. Zheng, Y. Zhao, H. Dong, et al., Structure–activity relationship of antioxidant dipeptides: dominant role of Tyr, Trp, Cys and Met residues, J. Funct.Foods 21 (2016) 485–496, https://doi.org/10.1016/j.jff.2015.12.003.

[12] M. Tian, B. Fang, L. Jiang, et al., Structure-activity relationship of a series of antioxidant tripeptides derived from β-Lactoglobulin using QSAR modeling, Dairy Sci. Technol. 95 (2015) 451–463, https://doi.org/10.1007/s13594-015-0226-5.

[13] Y.W. Li, B. Li, Characterization of structure-antioxidant activity relationship of peptides in free radical systems using QSAR models: key sequence positions and their amino acid properties, J. Theor. Biol. 318 (2013) 29–43, https://doi.org/10.1016/j.jtbi.2012.10.029.

[14] F. Ghasemi, A. Mehridehnavi, A. Pérez-Garrido, et al., Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks, Drug Discov. Today 23 (2018) 1784–1790, https://doi.org/10.1016/j.drudis.2018.06.016.

[15] W. He, Y. Wang, L. Cui, et al., Learning embedding features based on multi-sense-scaled attention architecture to improve the predictive performance of anticancer peptides, Bioinformatics (2021), https://doi.org/10.1093/bioinformatics/btab560.

[16] T.J. Lawrence, D.L. Carper, M.K. Spangler, et al., amPEPpy 1.0: a portable and accurate antimicrobial peptide prediction tool, Bioinformatics 37 (2021) 2058–2060, https://doi.org/10.1093/bioinformatics/btaa917.

[17] Y. Guo, K. Yan, H. Lv, et al., PreTP-EL: prediction of therapeutic peptides based on ensemble learning, Briefings Bioinf. 22 (2021), https://doi.org/10.1093/bib/bbab358.

[18] B. Manavalan, S. Basith, T.H. Shin, et al., mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation, Bioinformatics 35 (2019) 2757–2765, https://doi.org/10.1093/bioinformatics/bty1047.

[19] M. Mann, C. Kumar, W.F. Zeng, et al., Artificial intelligence for proteomics and biomarker discovery, Cell Syst 12 (2021) 759–770, https://doi.org/10.1016/j.cels.2021.06.006.

[20] S. Hu, P. Chen, P. Gu, et al., A deep learning-based chemical system for QSAR prediction, IEEE J Biomed Health Inform 24 (2020) 3020–3028, https://doi.org/10.1109/jbhi.2020.2977009.

[21] Y. Xu, J. Ma, A. Liaw, et al., Demystifying multitask deep neural networks for quantitative structure-activity relationships, J. Chem. Inf. Model. 57 (2017) 2490–2504, https://doi.org/10.1021/acs.jcim.7b00087.

[22] T.H. Olsen, B. Yesiltas, F.I. Marin, et al., AnOxPePred: using deep learning for the prediction of antioxidative properties of peptides, Sci. Rep. 10 (2020), 21471, https://doi.org/10.1038/s41598-020-78319-w.

[23] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, IEEE Trans. Signal Process. 45 (1997) 2673–2681, https://doi.org/10.1109/78.650093.

[24] R. Sharma, S. Shrivastava, S. Kumar Singh, et al., Deep-AFPpred: identifying novel antifungal peptides using pretrained embeddings from seq2vec with 1DCNN-BiLSTM, Briefings Bioinf. 23 (2022), https://doi.org/10.1093/bib/bbab422.

[25] V. Singh, S. Shrivastava, S. Kumar Singh, et al., StaBle-ABPpred: a stacked ensemble predictor based on biLSTM and attention mechanism for accelerated discovery of antibacterial peptides, Briefings Bioinf. 23 (2022), https://doi.org/10.1093/bib/bbab439.

[26] Y. Zhang, G. Zhu, K. Li, et al., HLAB: learning the BiLSTM features from the ProtBert-encoded proteins for the class I HLA-peptide binding prediction, Briefings Bioinf. (2022), https://doi.org/10.1093/bib/bbac173.

[27] L. Fu, B. Niu, Z. Zhu, et al., CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (2012) 3150–3152, https://doi.org/10.1093/bioinformatics/bts565.

[28] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.

[29] T.B. Zou, T.P. He, H.B. Li, et al., The structure-activity relationship of the antioxidant peptides from natural proteins, Molecules 21 (2016) 72, https://doi.org/10.3390/molecules21010072.

[30] Y. Li, J. Yu, Research progress in structure-activity relationship of bioactive peptides, J. Med. Food 18 (2014) 147–156, https://doi.org/10.1089/jmf.2014.0028.

[31] M. Mirzaei, S. Mirdamadi, M. Safavi, et al., The stability of antioxidant and ACE-inhibitory peptides as influenced by peptide sequences, LWT 130 (2020), 109710, https://doi.org/10.1016/j.lwt.2020.109710.

[32] C. Wen, J. Zhang, Y. Feng, et al., Purification and identification of novel antioxidant peptides from watermelon seed protein hydrolysates and their cytoprotective effects on H2O2-induced oxidative stress, Food Chem. 327 (2020), 127059, https://doi.org/10.1016/j.foodchem.2020.127059.

[33] C. Shi, M. Liu, H. Zhao, et al., A novel insight into screening for antioxidant peptides from hazelnut protein: based on the properties of amino acid residues, Antioxidants 11 (2022), https://doi.org/10.3390/antiox11010127.

[34] Z. Li, B. Wang, C. Chi, et al., Influence of average molecular weight on antioxidant and functional properties of cartilage collagen hydrolysates from Sphyrna lewini, Dasyatis akjei and Raja porosa, Food Res. Int. 51 (2013) 283–293, https://doi.org/10.1016/j.foodres.2012.12.031.

[35] K.L. Sun, M. Gao, Y.Z. Wang, et al., Antioxidant peptides from protein hydrolysate of marine red algae eucheuma cottonii: preparation, identification, and cytoprotective mechanisms on H(2)O(2) oxidative damaged HUVECs, Front. Microbiol. 13 (2022), 791248, https://doi.org/10.3389/fmicb.2022.791248.

[36] Y. He, X. Pan, C.-F. Chi, et al., Ten new pentapeptides from protein hydrolysate of miiuy croaker (Miichthys miiuy) muscle: preparation, identification, and antioxidant activity evaluation, LWT 105 (2019) 1–8, https://doi.org/10.1016/j.lwt.2019.01.054.

[37] C. Wen, J. Zhang, H. Zhang, et al., Plant protein-derived antioxidant peptides: isolation, identification, mechanism of action and application in food systems: a review, Trends Food Sci. Technol. 105 (2020) 308–322, https://doi.org/10.1016/j.tifs.2020.09.019.

[38] A. Sila, A. Bougatef, Antioxidant peptides from marine by-products: isolation, identification and application in food systems. A review, J. Funct.Foods 21 (2016) 10–26, https://doi.org/10.1016/j.jff.2015.11.007.

[39] F.-C. Wong, J. Xiao, S. Wang, et al., Advances on the antioxidant peptides from edible plant sources, Trends Food Sci. Technol. 99 (2020) 44–57, https://doi.org/10.1016/j.tifs.2020.02.012.

[40] C.-F. Chi, B. Wang, Y.-M. Wang, et al., Isolation and characterization of three antioxidant peptides from protein hydrolysate of bluefin leatherjacket (Navodon septentrionalis) heads, J. Funct.Foods 12 (2015) 1–10, https://doi.org/10.1016/j.jff.2014.10.027.

[41] L. Zhang, G.X. Zhao, Y.Q. Zhao, et al., Identification and active evaluation of antioxidant peptides from protein hydrolysates of skipjack tuna (Katsuwonus pelamis) head, Antioxidants 8 (2019), https://doi.org/10.3390/antiox8080318.

[42] X. Yang, Q. Jia, F. Duan, et al., Multiwall carbon nanotubes loaded with MoS2 quantum dots and MXene quantum dots: non–Pt bifunctional catalyst for the methanol oxidation and oxygen reduction reactions in alkaline solution, Appl. Surf. Sci. 464 (2019) 78–87, https://doi.org/10.1016/j.apsusc.2018.09.069.

[43] J. Yang, J. Huang, X. Dong, et al., Purification and identification of antioxidant peptides from duck plasma proteins, Food Chem. 319 (2020), 126534, https://doi.org/10.1016/j.foodchem.2020.126534.

[44] R. Abeynayake, S. Zhang, W. Yang, et al., Development of antioxidant peptides from brewers' spent grain proteins, LWT (Lebensm.-Wiss. & Technol.) 158 (2022), https://doi.org/10.1016/j.lwt.2022.113162.

[45] O.K. Chang, G.E. Ha, G.-S. Han, et al., Novel antioxidant peptide derived from the ultrafiltrate of ovomucin hydrolysate, J. Agric. Food Chem. 61 (2013) 7294–7300, https://doi.org/10.1021/jf4013778.

[46] H. Agrawal, R. Joshi, M. Gupta, Purification, identification and characterization of two novel antioxidant peptides from finger millet (Eleusine coracana) protein hydrolysate, Food Res. Int. 120 (2019) 697–707, https://doi.org/10.1016/j.foodres.2018.11.028.

[47] M. Memarpoor-Yazdi, A. Asoodeh, J. Chamani, A novel antioxidant and antimicrobial peptide from hen egg white lysozyme hydrolysates, J. Funct.Foods 4 (2012) 278–286, https://doi.org/10.1016/j.jff.2011.12.004.