ACEiPP: A deep learning-based framework to predict angiotensin-converting enzyme (ACE)-inhibitory peptides using high-efficiency amino acid descriptors

Dongya Qin[1], Ruihong Wang[1], Linna Jiao[1], Bo Li[2], Guixue Wang[1], and Guizhao Liang[1,*]

[1]Key Laboratory of Biorheological Science and Technology, Ministry of Education, College of Bioengineering, Chongqing University, Chongqing 400044, China.

[2]College of Life Sciences, Chongqing Normal University, Chongqing 401331, China.

[*]Corresponding author: Tel, +86-23-65102507; Fax, +86-23-65102508; E-mail: gzliang@cqu.edu.cn.

Biographical note

Dongya Qin[1] is a PhD candidate at the Key Laboratory of Biorheological Science and Technology, College of Bioengineering, Chongqing University, Chongqing, China. He is dedicated to bioinformatics and deep learning research on bioactive peptides.

Ruihong Wang[1] is a PhD candidate at the Key Laboratory of Biorheological Science and Technology, College of Bioengineering, Chongqing University, Chongqing, China. He is interested in virtual screening of bioactive compounds.

Linna Jiao[1] is a PhD candidate at the Key Laboratory of Biorheological Science and Technology, College of Bioengineering, Chongqing University, Chongqing, China. She is interested in computer-aided drug discovery and machine learning.

Bo Li[2] is an associate professor of computational biology and bioinformatics at College of Life Sciences, Chongqing Normal University, China. He is working in the field of systems biology and bioinformatics.

Guixue Wang[1] is a professor of Key Laboratory of Biorheological Science and Technology, College of Bioengineering, Chongqing University, Chongqing, China. His research is in the areas of biomedical engineering and bioinformatics.

Guizhao Liang[1,*] is a professor of bioinformatics and biomedicine in Key Laboratory of Biorheological Science and Technology, College of Bioengineering, Chongqing University, Chongqing, China. He works on application of bioinformatics in food science and biomedical engineering.

Abstract

Food-derived angiotensin-converting enzyme inhibitory peptides (ACEiPs) are potentially safe therapeutic agents against hypertension. Wet experiments used to identify ACEiPs, such as complicated enzymatic purification and mass spectrometry identification, are time-consuming and costly. Here, we build a deep learning-based predictor of ACEiPs (ACEiPP) by using optimized amino acid descriptors (AADs) and long-short term memory (LSTM) neural network. Our results show that the combined-AADs exhibit more efficient feature transformation ability than the single-AADs, especially the training model with the optimal descriptors (named VVSFZL37) as the feature inputs exhibit the highest predictive ability on the independent test (Acc = 0.9479, AUC = 0.9876), with a significant performance improvement compared to the existing three predictors. This high predictive ability is attributed to that LSTM can effectively learn the sequence/structure characteristics closely related to the ACE-inhibitory activity. We then use ACEiPP to screen 1,320,635 potential ACEiPs encoded in 21,249 food-derived proteins, and sequence analysis indicates that these predicted ACEiPs have same/similar characteristics in length, mass, N-/C-terminals and composition. Based on ACEiPP we construct multiple theoretical peptide libraries to provide candidate ACEiPs, especially the theoretical library of multifunctional bioactive peptides, which can theoretically modulate multiple target proteins in cardiovascular diseases and can be potential candidates for therapy and nutritional intervention. Collectively, ACEiPP (http://www.cqudfbp.net/ACEiPP/index.jsp) will be an effective tool to identify, design, and screen ACEiPs and might play a valuable role in health care and drug discovery.

Keywords: *Angiotensin-converting enzyme (ACE), ACE-inhibitory peptide, Multifunctional peptide, Long-short term memory (LSTM) neural network, Amino acid descriptor (AAD)*

1. Introduction

Hypertension is one of the most important controllable risk factors in treating cardiovascular diseases and shows a trend of increasing and younger every year [1, 2]. However, there is no available complete cure method for hypertension today, and patients mainly take drugs to control or relieve high blood pressure. As we know, long-term use of these drugs not only has distinct side effects, such as headaches, heart rate acceleration, dizziness, dry cough, dysgeusia, rashes, etc. but also increases the patient's psychological stress [3, 4]. Therefore, screening for safe and effective natural molecules has currently been the focus of drug development [5-7].

ACE-inhibitory peptides (ACEiPs) are short peptides with about 2-19 amino acid residues [1]. They can block the conversion of angiotensin I to angiotensin II by inhibiting ACE in the renin-angiotensin-aldosterone system, leading to suppression of vasoconstriction and reduction of blood pressure [8, 9]. Many ACEiPs have been identified from milk, animal, plant, seafood, and microorganism source by fermentation, enzymatic hydrolysis, and synthetic methods [10-15]. These ACEiPs can inhibit ACE *in vitro* and even exhibit similar or better blood pressure lowering effects than *Captopril* (an ACE inhibitor for the treatment of hypertension) in animal experiments [16-18]. More importantly, ACEiPs also exhibit multiple activities (antihypertensive, antioxidant, DPP IV-inhibitory, renin-inhibitory, anticancer, etc., as shown in Supplementary Figure S1) according to the Database of Food-derived Bioactive Peptides (DFBP) [19]. Therefore, ACEiPs have the advantages of high activity, multi-function, multi-source, and easy access, especially food-derived multifunctional peptides (MBPs) with ACE-inhibitory activity are expected to be candidates for replacing synthetic drugs in the treatment of cardiovascular diseases [9, 14, 20, 21].

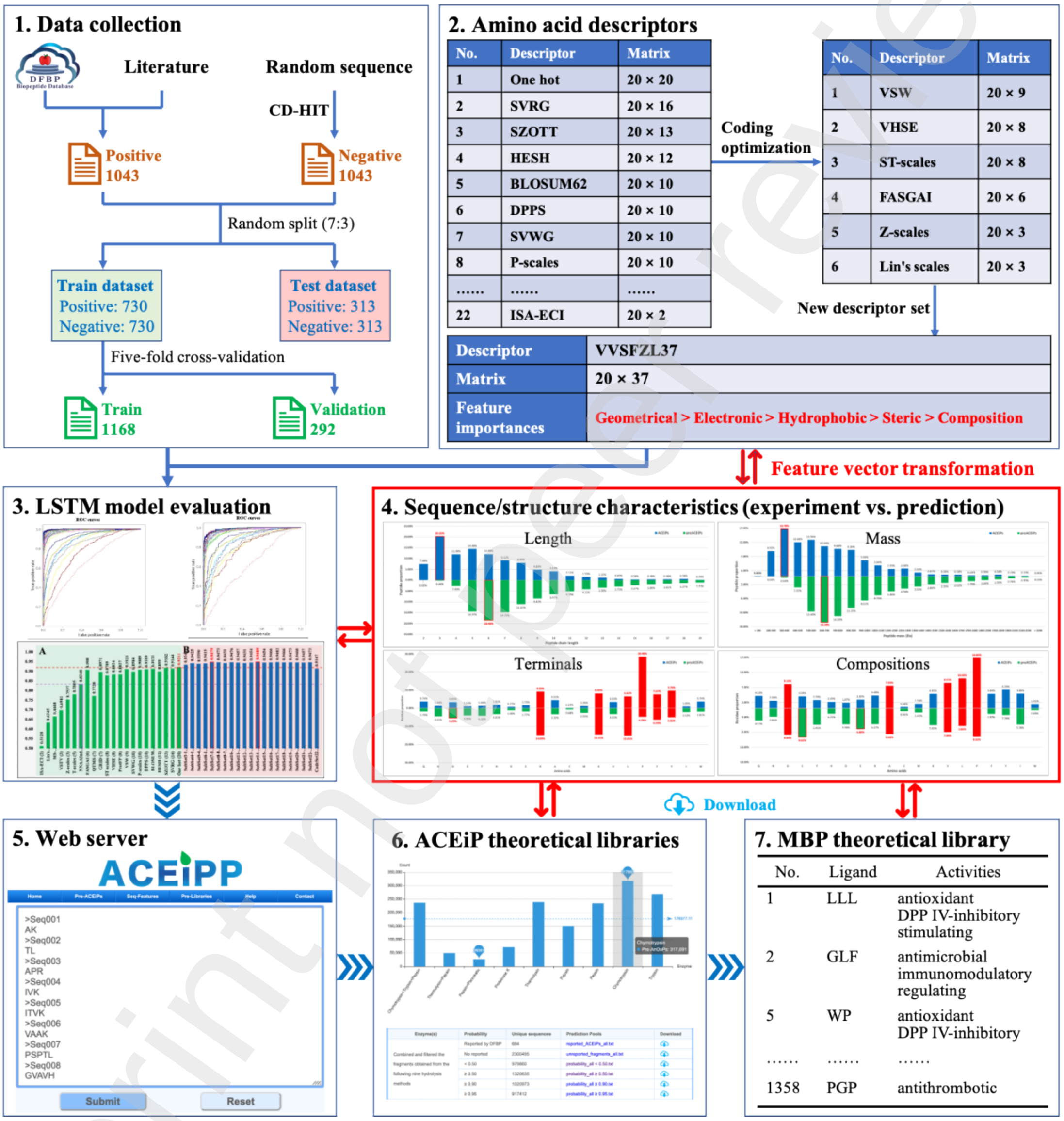
The peptide library consisting of 20 natural amino acids is extensive, so the experimental identification of bioactive peptides (including ACEiPs) is labor-intensive, time-consuming, and costly [22]. Recently, computational methods have emerged as a new strategy for large-scale screening of bioactive peptides, including quantitative structure−activity relationship (QSAR), machine learning (ML), and deep learning (DL) [23, 24]. In the QSAR study, amino acid descriptors (AADs) are commonly used to characterize the sequence/structure characteristics of peptides [23]. Multiple AADs have been proposed to characterize ACEiPs, such as z-scales [25], HESE [26], and FASGAI [27], showing favorable interpretability on the structure-activity relationship of peptides. The ML-based methods have shown advantages in the large-scale prediction of peptides by using physicochemical parameters, structural properties, and experimental data as feature inputs to train models and mine the bio-information in the multi-dimensional data [6]. AHTpin [28] and mAHTPred [29] are currently the only two accessible predictors for antihypertensive peptides (AHTPs). AHTpin used amino acid and atomic composition features to develop multiple classification models for small peptides, medium peptides and large peptides, respectively, among which the model achieved the highest accuracy of 84.21% on large peptides. mAHTPred is a state-of-the-art AHTP classifier that combines the extremely randomized tree and 51 feature descriptors derived from eight different feature encodings, achieving the highest accuracy of 88.3% compared to six different ML algorithms. The DL-based methods can fully use diverse input data, especially for processing images and natural language to accurately capture the sequence/structure characteristics of bioactive peptides. The DL-based techniques have not yet been applied in the study of ACEiPs, but they have been used in the study of several other bioactive peptides, such as cACP-DeepGram (A anticancer peptide predictor based on deep neural network and skip-gram-based word embedding model) [30], sAMP-PFPDeep (An antibacterial peptide predictor based on the convolutional neural network) [31], an MBP predictor based on the convolutional neural network [32]. These computational methods provide important references for building the prediction model of ACEiPs.

The computational methods have made important contributions to the screening of bioactive peptides. However, there is still insufficient research on the prediction of ACEiPs, and the possible reasons are as follows: (i) QSAR is currently mainly used for small-sample regression, while feature extraction is difficult when the lengths of ACEiPs are inconsistent; (ii) The ML-based AHTP predictors can be used to screen ACEiPs, but the prediction accuracy (less than 88.3% [29]) is still expected to be improved, and the ML-based model usually requires specific type and length of data as input parameters and is less interpretable; and (iii) No DL-based prediction server for ACEiPs is available so far. LSTM is a particular recurrent neural network for processing natural sequence data with different lengths, avoiding gradient disappearance and explosion problems during long-sequence training [33, 34]. LSTM can memorize the important transfer information between sequences and forget the non-important information by controlling the transmission state through three gated states (input, forget and output gate). Therefore, combining the learning ability of LSTM neural network and the interpretability of AADs, may be an efficient strategy to improve the prediction accuracy and explain the sequence/structure characteristics of ACEiPs.

To address the above challenges, we combined LSTM neural network and AAD encodings to train an optimal model and explain the main sequence/structure characteristics of ACEiPs, thereby developing a predictor named ACEiPP (http://www.cqudfbp.net/ACEiPP/index.jsp) to predict the potential ACEiPs. ACEiPP achieved higher independent test accuracy as compared to the existing three methods, i.e., mAHTPred [29], AHTpin_AAC, and AHTpin_ATC [28]. We finally performed batch prediction of ACEiPs from food-derived proteins, then constructed theoretical peptide libraries of ACEiPs and MBPs. To our knowledge, ACEiPP is the first DL-based predictor for high-efficiency prediction and screening of ACEiPs.

2. Materials and methods

The construction flowchart of ACEiPP is shown in Figure 1. First, the peptide sequences were collected and translated to feature vectors by single- and combined-AADs. Second, LSTM model optimization and key feature screening of ACEiPs were performed. Third, candidate libraries for ACEiPs and MBPs were constructed by batch prediction of peptides encoded in food-derived proteins. Finally, the optimal model was deployed in the webserver to access the online prediction. Details of this predictor are described in the following sections.



No.	Descriptor	Matrix
1	One hot	20 × 20
2	SVRG	20 × 16
3	SZOTT	20 × 13
4	HESH	20 × 12
5	BLOSUM62	20 × 10
6	DPPS	20 × 10
7	SVWG	20 × 10
8	P-scales	20 × 10
......	......	......
22	ISA-ECI	20 × 2

No.	Descriptor	Matrix
1	VSW	20 × 9
2	VHSE	20 × 8
3	ST-scales	20 × 8
4	FASGAI	20 × 6
5	Z-scales	20 × 3
6	Lin's scales	20 × 3

Descriptor	VVSFZL37
Matrix	20 × 37
Feature importances	Geometrical > Electronic > Hydrophobic > Steric > Composition

No.	Ligand	Activities
1	LLL	antioxidant DPP IV-inhibitory stimulating
2	GLF	antimicrobial immunomodulatory regulating
5	WP	antioxidant DPP IV-inhibitory
......	......	......
1358	PGP	antithrombotic

Figure 1. Workflow of ACEiPP. First, to collect the peptide sequences and translate them into feature vectors by single- and combined-AADs. Second, to train LSTM models and optimize the key features of ACEiPs, followed by comparing the generalization ability of ACEiPP with existing models on the independent dataset. Third, to perform prediction of peptides encoded in food-derived proteins to construct candidate libraries for ACEiPs and MBPs. Finally, to deploy the optimal model in the webserver to access the online prediction.

2.1. Benchmark and independent datasets

One benchmark dataset and three independent datasets (Supplementary Table S1) were used for training the predictor and evaluating its prediction performance. Through deleting reduplicate or controversial sequences with inconsistent experimental results, 1043 ACEiPs with IC_{50} values below 1000 μM were selected as positive samples from three databases DFBP (http://www.cqudfbp.net/), BIOPEP-UWM (https://biochemia.uwm.edu.pl/biopep-uwm/) and AHTPDB (http://crdd.osdd.net/raghava/ahtpdb/), and related literature. In order to maintain the balance of the number and length distribution of positive and negative samples, a manually written java program and CD-HIT [35] were used to generate random sequences and remove the sequences with greater than 90% similarity with positive samples. Total 1043 peptides were randomly selected as negative samples to balance the number of positive samples. Both positive and negative samples were split according to 7:3 to construct a benchmark (benchmark_ACEiPs) dataset and an independent dataset (independent_ACEiPs) [28]. The second independent dataset (independent_newACEiPs) included 48 newly discovered ACEiPs with IC_{50} less than 1000 μM in 2021-2022, ensuring no sequence intersection with the datasets constructed above. The third independent dataset (saved in independent_AHTPs.txt) with sequence length ≥5 was downloaded from mAHTPred (http://thegleelab.org/mAHTPred/), which contained 386 AHTPs and 386 non-AHTPs [28].

2.2. Sequence feature encoding

Here, 22 AADs were used to translate peptide sequences to feature vectors as inputs of neural networks, where One-hot and the other 21 descriptors described the types and physicochemical attributes of natural amino acids, respectively (Supplementary Table S2). Two encoding strategies were used as follows: (i) Single-AADs: Training the models using 22 independent AAD encoding matrices individually; and (ii) Combined-AADs: Combining all the single-AADs into a 20×149 matrix (CodeSet22) and using

"Leave-Group-Out" to sequentially delete a single-AADs from CodeSet22 to optimize the best combination coding.

2.3. LSTM architecture

In this study, LSTM neural network in Deeplearning 4j framework (DL4J: https://deeplearning4j.org/) was used to train the models. LSTM model can effectively learn time or step dependencies in sequence data and enables training on variable-length data. As shown in Figure 2, We first manually set up an input transcoding layer to convert peptide sequences into feature vectors, and then four neural network layers were set up, i.e. three hidden LSTM layers and a Recurrent Neural Network Output (RnnOutput) layer. We used grid search technique to select the best hyperparameters. The input length of the 1st LSTM layer was equal to the encoding length of input descriptors, its activation function was "*tanh*", the dropout value was 0.4, and the output length was 128. The input and output lengths of 2nd and 3rd LSTM layers were all 128, their activation functions were all " *tanh* ", and the dropout values were 0.4 and 0.25, respectively. The input and output lengths of the RnnOutput layer were 128 and 2, respectively, and its activation function was "*sigmoid*" for binary classification (ACEiPs or non-ACEiPs). For the detailed calculation process of the LSTM unit, please refer to the description of the method [33].

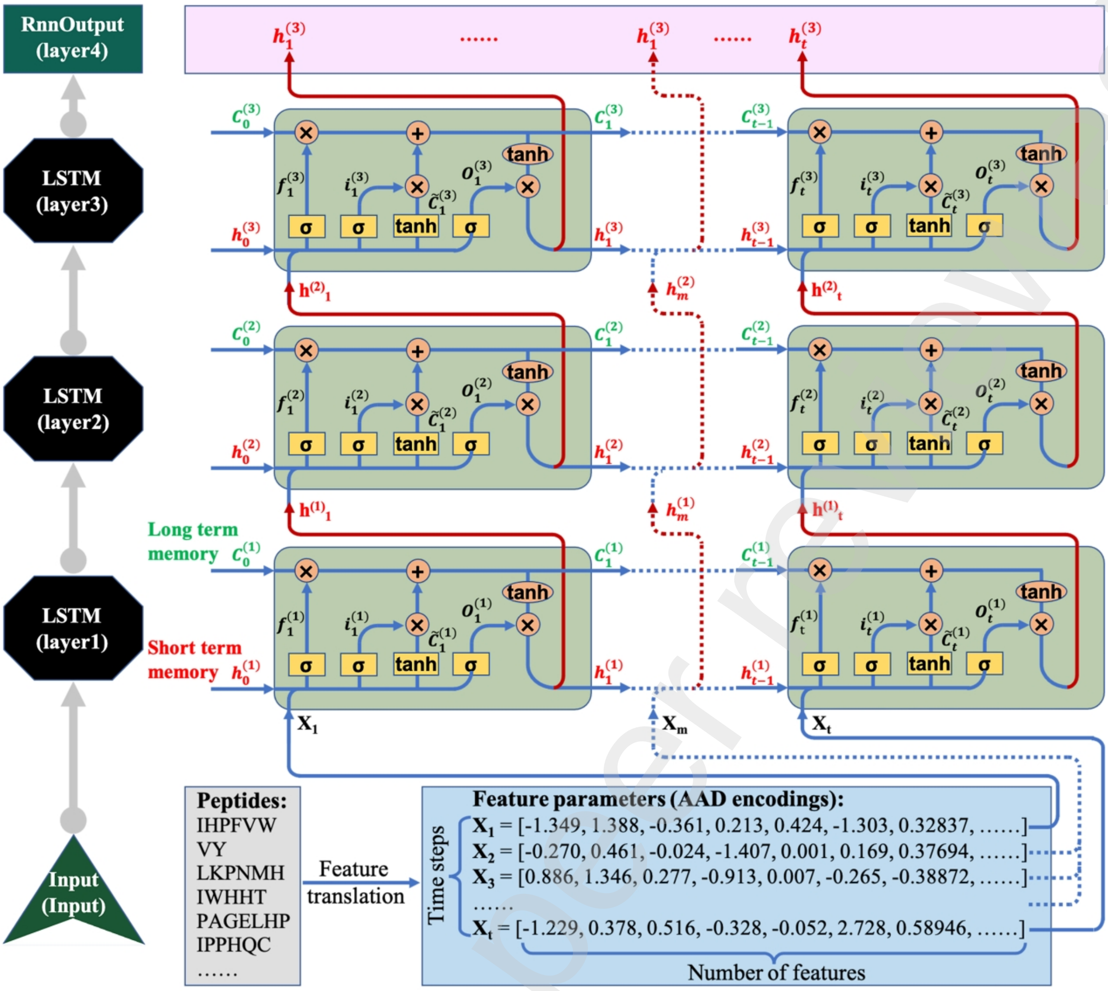

RnnOutput (layer4)
LSTM (layer3)
LSTM (layer2)
LSTM (layer1)
Input (Input)
$h_1^{(3)}$ …… $h_1^{(3)}$ …… $h_t^{(3)}$
$c_0^{(3)}$
$h_0^{(3)}$
$c_1^{(3)}$
$h_1^{(3)}$
$c_{t-1}^{(3)}$
$h_{t-1}^{(3)}$
$f_1^{(3)}$ $i_1^{(3)}$ $o_1^{(3)}$ $\tilde{C}_1^{(3)}$
$f_t^{(3)}$ $i_t^{(3)}$ $o_t^{(3)}$ $\tilde{C}_t^{(3)}$
σ σ tanh σ
tanh
$h^{(2)}_1$
$h_m^{(2)}$
$h^{(2)}_t$
$c_0^{(2)}$
$h_0^{(2)}$
$c_1^{(2)}$
$h_1^{(2)}$
$c_{t-1}^{(2)}$
$h_{t-1}^{(2)}$
$f_1^{(2)}$ $i_1^{(2)}$ $o_1^{(2)}$ $\tilde{C}_1^{(2)}$
$f_t^{(2)}$ $i_t^{(2)}$ $o_t^{(2)}$ $\tilde{C}_t^{(2)}$
$h^{(1)}_1$
$h_m^{(1)}$
$h^{(1)}_t$
Long term memory $c_0^{(1)}$
Short term memory $h_0^{(1)}$
$c_1^{(1)}$
$h_1^{(1)}$
$c_{t-1}^{(1)}$
$h_{t-1}^{(1)}$
$f_1^{(1)}$ $i_1^{(1)}$ $o_1^{(1)}$ $\tilde{C}_1^{(1)}$
$f_t^{(1)}$ $i_t^{(1)}$ $o_t^{(1)}$ $\tilde{C}_t^{(1)}$
$\mathbf{X}_1$
$\mathbf{X}_m$
$\mathbf{X}_t$
Peptides:
IHPFVW
VY
LKPNMH
IWHHT
PAGELHP
IPPHQC
……
Feature translation
Feature parameters (AAD encodings):
Time steps
$\mathbf{X}_1$ = [-1.349, 1.388, -0.361, 0.213, 0.424, -1.303, 0.32837, ……]
$\mathbf{X}_2$ = [-0.270, 0.461, -0.024, -1.407, 0.001, 0.169, 0.37694, ……]
$\mathbf{X}_3$ = [0.886, 1.346, 0.277, -0.913, 0.007, -0.265, -0.38872, ……]
……
$\mathbf{X}_t$ = [-1.229, 0.378, 0.516, -0.328, -0.052, 2.728, 0.58946, ……]
Number of features

Figure 2. LSTM architecture flowchart. An input transcoding layer and four neural network layers were set up, including three hidden LSTM layers and one RnnOutput layer.

2.4. Evaluation of performance

Five evaluation indices were used to evaluate the prediction performance of the proposed models, including sensitivity (Sn), specificity (Sp), accuracy (Acc), Matthews's correlation coefficient (MCC), and area under the receiver operating characteristic curves (AUC). These indices were defined as follows:

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

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

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

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

Additionally, we used 0 to mask single or multiple target parameters for all samples, then evaluated the importance score of features by the following formula:

$$\text{Importance score} = \frac{\mathrm{FP}+\mathrm{FN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$

Where TP, FN, FP, and TN denote the numbers of ACEiPs correctly predicted, non-ACEiPs misclassified, ACEiPs misclassified, and non-ACEiPs correctly predicted, respectively.

2.5. Theoretical screening of food-derived ACEiPs and MBPs

The sequences of 21,249 food-derived proteins were retrieved from DFBP [19], and THP-Tool (http://www.cqudfbp.net/enzymes/hydrolysis_tools/dataInput.jsp) was used for virtual hydrolysis of these proteins. Six typical protein hydrolases, including Chymotrypsin (EC 3.4.21.1), Papain (EC 3.4.22.2), Proteinase K (EC 3.4.21.64), Thermolysin (EC 3.4.24.27), Pepsin (EC 3.4.23.1) and trypsin (EC 3.4.21.4) were used to generate a large number of peptides, and the resulting peptides were predicted by ACEiPP.

2.6. Webserver

ACEiPP was deployed on Tencent Cloud Server (Windows Server 2012 R2 Standard Edition 64-bit) through Java web technology. The web directory was designed using HTML, JSP, and JavaScript. MySQL was used for data storage and invoked by manually written java programs to analyze data. EChart (https://echarts.apache.org/en/index.html) technology was used for data visualization. ACEiPP can provide multiple services such as prediction, screening, and structure-activity exploration of ACEiPs.

3. Results and discussion

3.1. Performance of single coding models

Twenty-two single-AADs were used to convert peptide sequences into matrices containing the sequence/structure characteristics as the input parameters of neural networks to train LSTM models. As shown in the five-fold cross-validation on the benchmark dataset (Supplementary Table S3), eighteen models exhibited higher predictive performance with Acc values of 0.7705-0.9308, MCC values of 0.5458-0.8629, and AUC values of 0.8653-0.9773 than the remaining four models (ISA-ECI, Lin's scales, MS-WHIM, and VSTV) trained with three or two features. Moreover, sixteen models (Number of features: 6-20) exhibited high Acc values (0.8348-0.9211) on the test set (independent_ACEiPs) containing 313 ACEiPs and 313 non-ACEiPs (Supplementary Table S4).

Thus, the LSTM-based models could effectively distinguish the dependencies between residue sites in unknown peptide sequences and accurately identified the key features of ACEiPs and non-ACEiPs. The models that learned ten or more than ten features generally showed higher predictive performance than the models based on less than ten features. Each AADs translated different types and physicochemical attribute characteristics of residues encoded in peptide sequences, so multi-feature inputs could prompt LSTM models to reasonably learn the sequence/structure characteristics of peptides. Therefore, it is feasible to use single-ADDs to extract the features to characterize peptide sequences and feed them into LSTM models. However, we must face a new challenge to optimize and screen new sequence/structure characteristics with high-efficiency but low redundancy.

3.2. Performance of optimized coding models

To further optimize a set of high-efficiency AADs with low-redundancy, we restructured all single-AADs into combined-AADs as inputs of the LSTM model and evaluated the predictive performance of the trained models. First, the twenty-two single-AADs were integrated into CodeSet22 to evaluate the Acc value of the obtained model. Then, "Leave-Group-Out" was used to reduce feature redundancy by gradually removing one group of single-AADs from CodeSet22 in turn, and the optimal feature matrices for each round of training would be saved to perform the next optimization. The five-fold cross-validation scores (Supplementary Table S5) and independent test scores (Supplementary Table S6) indicated that the restructured combined-AADs have more stable feature characterization ability than single-AADs. As shown in Figure 3, the models based on combined-AADs have the following key characteristics relative to the models based on single-AADs: (i) Except for the model based on CodeSet22, the optimal models trained with combined-AADs generated higher Acc values (from 0.9342 to 0.9489) than the model based on single-AADs, even SubSet3-1 (combined with FASGAI and ST-scale, Supplementary Table S6) with only 14 features generated higher Acc than all single-AADs; (ii) As the increasing number of features from 2 to 37, the Acc values gradually improved, indicating that the effective features were continuously accumulated; (iii) When the number of features was added to 37, the SubSet7-5 (named VVSFZL37, composed of FASGAI, Lin's scales, ST-scales, VHSE, VSW, and Z-scales) achieved the critical point where a set of non-redundant features were extracted from the total feature set (CodeSet22 with 179 features). The obtained model exhibited an Acc value of 0.9479, with an enhancement of 2.68% compared to the One-hot model with the highest Acc value in all the models by single-AADs; (iv) With the increasing of features (more than 37), the Acc values basically stabilized above 0.9431, revealing that the redundancy of features began to gradually increase. This not only made variable weights increase ineffectively but also added the difficulty of training, even the Acc values rapidly dropped to 0.9147 when the number of features was accumulated to 179. Therefore, VVSFZL37 was the most efficient coding for translating peptide sequences into high-efficiency features, which were used to train the LSTM model to acquire predictor of ACEiPs.

The effectiveness of the peptide sequence coding was closely related to the characteristics of AAD parameters, containing the number, physicochemical meaning, redundancy of features, etc. The combined-AADs exhibited more efficient feature transformation ability than the single-AADs, and could favorably characterize the features of peptide sequences relative to the single-AADs. However, with the increasing coding features, the Acc values could not continue to be improved or even decline due to the increasing redundant features. Therefore, it is necessary to select high-efficiency AADs or their combinations for LSTM to learn and obtain the parameters closely related to the activity of ACEiPs, thereby improving the prediction performance.

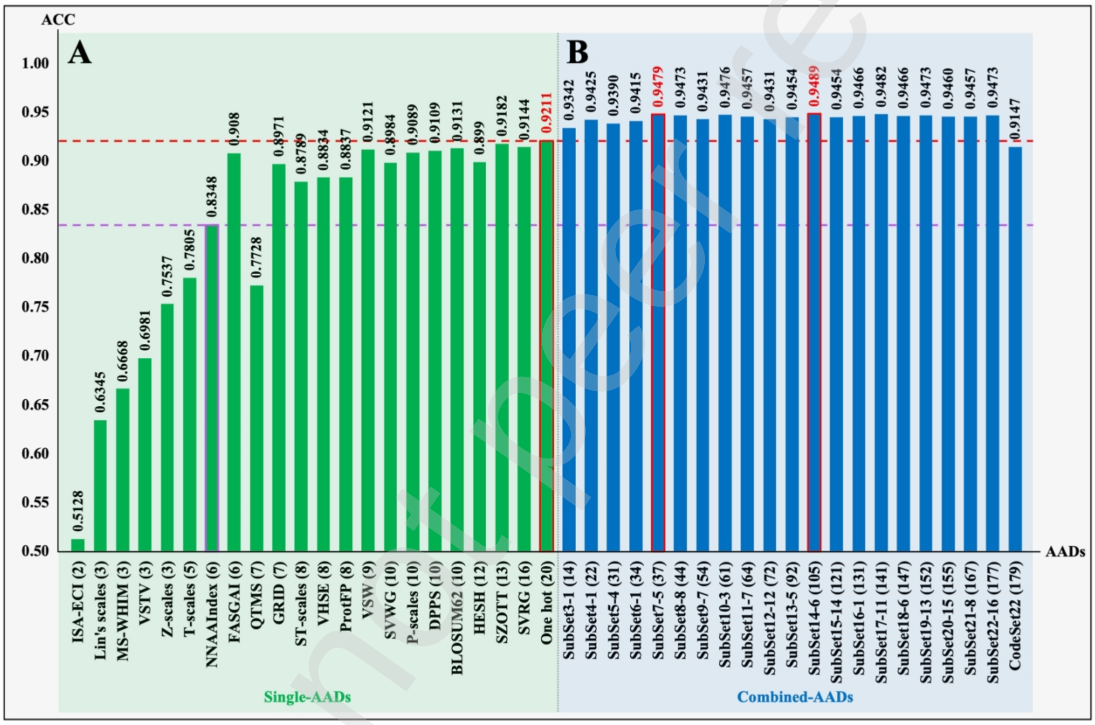


Figure 3. The predicted Acc on the independent dataset (independent_ACEiPs) based on different AADs. (A) 22 single-AADs (Supplementary Table S4). (B) 22 optimized-AADs (Supplementary Table S5). CodeSet22 is a total feature matrix consisting of 22 single-AADs. Each "SubSet" represents the selected combined-AADs gradually screened out from CodeSet22 by the "Leave-Group-Out" method. For example, SubSet22-16 represents the optimized-encodings obtained by removing the 16th single-AADs from CodeSet22 (See Supplementary Table S6 for detailed screening process). The number in the parentheses represents the number of features.

3.3. Comparison with the existing predictors

To evaluate the generalization ability of ACEiPP with the existing predictors (i.e., mAHTPred, AHTpin_AAC, and AHTpin_ATC), we carried out the comprehensive comparisons on three independent datasets. The evaluation indices on the independent_ACEiPs test set (Table 1) indicate that ACEiPP (Acc = 0.9479, MCC = 0.8959) achieved higher prediction performance than the three existing models. To be specific, ACEiPP showed the best prediction ability with an Acc value of 0.9479, while the Acc values of mAHTPred, AHTpin_AAC, AHTpin_ATC were only 0.8492, 0.7706 (average), and 0.756 (average).

We then evaluated the predictive power of each model for AHTPs on the independent_AHTPs test set (Table 1). The Acc values of ACEiPP, mAHTPred, AHTpin_AAC, and AHTpin_ATC were 0.8303, 0.8834, 0.8061 (average), and 0.8204 (average), respectively, showing that ACEiPP has better prediction ability than AHTpin_ACC and AHTpin_ATC, but less than mAHTPred for AHTPs. As we know, AHTPs involve multi-type target protein inhibitors [36] [37], which can be seen in the dataset independent_AHTPs, including ACEiPs, DPP IV-inhibitory, antioxidant, and antihypertensive peptides, etc. [29]. Nevertheless, ACEiPP was trained on the ACEiPs and non-ACEiPs benchmark dataset, that was, ACEiPP only learned the features of ACEiPs, while unable to effectively identify the features of the whole AHTP samples, thereby affecting the prediction for AHTPs.

We further compared the prediction performance of ACEiPP and AHTPs (AHTpin_AAC, AHTpin_ATC, and mAHTPred) on the test set independent_newACEiPs containing all 48 newly reported ACEiPs in 2021-2022. Table 2 shows that ACEiPP generated an Acc value of 81.25%, which was higher than that of the other three predictors, AHTpin_AAC (65.22%), AHTpin_ATC (56.52%), and mAHTPred (50.00%). It is worth emphasizing that ACEiPP exhibited higher Acc in predicting ACEiPs than all AHTPs-based models (AHTpin_AAC, AHTpin_ATC, and mAHTPred), further demonstrating that the training model based on ACEiPs could extract the main features of ACEiPs more accurately than the models trained by multiple different kinds of AHTPs.

Table 1. Performance comparison of ACEiPP and three existing predictors on independent_ACEiPs and independent_AHTPs.

Predictors[a]	Applicable length	Independent_ACEiPs		Independent_AHTPs	
		Acc	MCC	Acc	MCC
AHTpin_AAC	Tetrapeptides	0.7639	0.5168	-	-
	Pentapeptides	0.5429	0.0880	0.7814	0.4121
	Hexapeptides	0.8507	0.7152	0.9045	0.4354
	Medium peptides (7-12)	0.7957	0.5898	0.7622	0.5033
	Large peptides (≥13)	0.9000	0.7980	0.7761	0.2618
AHTpin_ATC	Tetrapeptides	0.7639	0.5168	-	-
	Pentapeptides	0.5810	0.1674	0.7705	0.3292
	Hexapeptides	0.7910	0.5775	0.9045	0.3582
	Medium peptides (7-12)	0.8441	0.7053	0.7744	0.4828
	Large peptides (≥13)	0.8000	0.6162	0.8321	0.3256
mAHTPred	≥5	0.8492	0.7027	0.8834	0.7670
ACEiPP	≥2	0.9479	0.8959	0.8303	0.6614

[a] AHTpin_AAC and AHTpin_ATC are two AHTP models trained with amino acid composition and atomic composition features, respectively. Since their evaluation parameters are determined according to different peptide lengths, the average values are calculated to compare their prediction performance. mAHTPred is an AHTP prediction model based on the extremely randomized tree, which can only predict sequences greater than or equal to five residues.

Table 2. Performance comparison of ACEiPP and the existing models on independent_newACEiPs.

Predictors[a]	Applicable length	Samples	TP[b]	FP[c]	Acc (%)
AHTpin_AAC	Tetrapeptides	7	5	2	71.43
	Pentapeptides	5	2	3	40.00
	Hexapeptides	4	3	1	75.00
	Medium peptides (7-12)	5	3	2	60.00
	Large peptides (≥13)	2	2	0	100.00
AHTpin_ATC	Tetrapeptides	7	5	2	71.43
	Pentapeptides	5	1	4	20.00
	Hexapeptides	4	3	1	75.00
	Medium peptides (7-12)	5	2	3	40.00
	Large peptides (≥13)	2	2	0	100.00
mAHTPred	≥5	16	8	8	50.00
ACEiPP	≥2	48	39	9	81.25

[a] AHTpin_AAC and AHTpin_ATC are two AHTP models trained with amino acid composition and atomic composition features, respectively. Since their evaluation parameters are determined according to different peptide lengths, the average values are calculated to compare their prediction performance. mAHTPred is an AHTP prediction model based on the extremely randomized tree, which can only predict sequences greater than or equal to five residues; [b] TP: The numbers of ACEiPs correctly predicted; [c] FP: The numbers of ACEiPs misclassified.

3.4. Key sequence/structure characteristics learned by ACEiPP

To explore the key characteristics learned by ACEiPP, we interpreted the physicochemical meaning of VVSFZL37 (Supplementary Table S7). As shown in Figure 4A, VVSFZL37 consists of 37 parameters derived from six sets of AADs (VSW, VHSE, ST-scales, Z-scales, FASGAI, and Lin's scales), showing higher predictive power than any single-AADs. All 37 parameters made important contributions to model performance, especially the top ten parameters, involving electronic properties, hydrophobicity, molecular composition and structure, and Van Der Waal's volume, which were important residue characteristics of ACEiPs (Figure 4B). To further compare the overall contribution of the same/similar parameters to the model, we divided these 37 parameters into five patterns according to their physicochemical meanings, namely geometrical, electronic, hydrophobic, steric, and composition characteristics, consisting of 12, 11, 6, 4 and 4 parameters, respectively (Figure 4C, see Supplementary Table S8 for details). Their importance scores are ordered as Geometrical > Electronic > Hydrophobic > Steric > Composition (Figure 4D).

The five patterns of physicochemical attributes encoded by VVSFZL37 are also intuitively observed in the sequence features of ACEiPs (Figure 4E): (i) ACEiPs were mainly oligopeptides with a molecular weight of less than 1000 Da composed of 2-10 residues (Supplementary Figure S2A and S2B), which affects the geometrical and steric features of the whole molecular structure in terms of size, shape, symmetry, and atom distribution; (ii) Their N-terminals tended to be five hydrophobic amino acid residues Leu, Val, Ala, Gly and Ile (proportion: 8.15-13.14%), while the C-terminals favored two hydrophobic residues Pro and Phe (especially a proportion of 28.48% for Pro), and two positively charged amino acids Arg (9.2%) and Lys (8.25%), and one bulky aromatic amino acid Tyr (9.78%) (Supplementary Figure S2C); (iii) The main composition of ACEiPs was composed of hydrophobic residues Pro, Leu, Val, Gly, and Ala (7.53%-16.84%), and aromatic residue Tyr (6.26%) (Supplementary Figure S2D). Therefore, the above features represent the optimal amino acid profiles of ACEiPs.

Furthermore, several reports indicated that hydrophobic, aromatic/aliphatic, C- and N-terminal, residue composition, and charged amino acids are indeed the key factors affecting the activities of ACEiPs. They not only directly determine the molecular geometry, overall hydrophobicity, side-chain functional groups, and charged characteristics of ACEiPs, but also indirectly affect the second structure, steric structure, and molecular compositional characteristics of ACEiPs [9, 10, 24] (see Supplementary Note S1 for details). These studies strongly supported our conclusion that these five patterns of physicochemical attributes are indeed important factors in forming ACEiPs. Therefore, VVSFZL37 constructed a plausible multiple properties and low redundancy matrix to characterize various sequence/structure characteristics closely related to the activity of ACEiPs, thereby improving the robustness and accuracy of the LSTM model than all single-AADs.

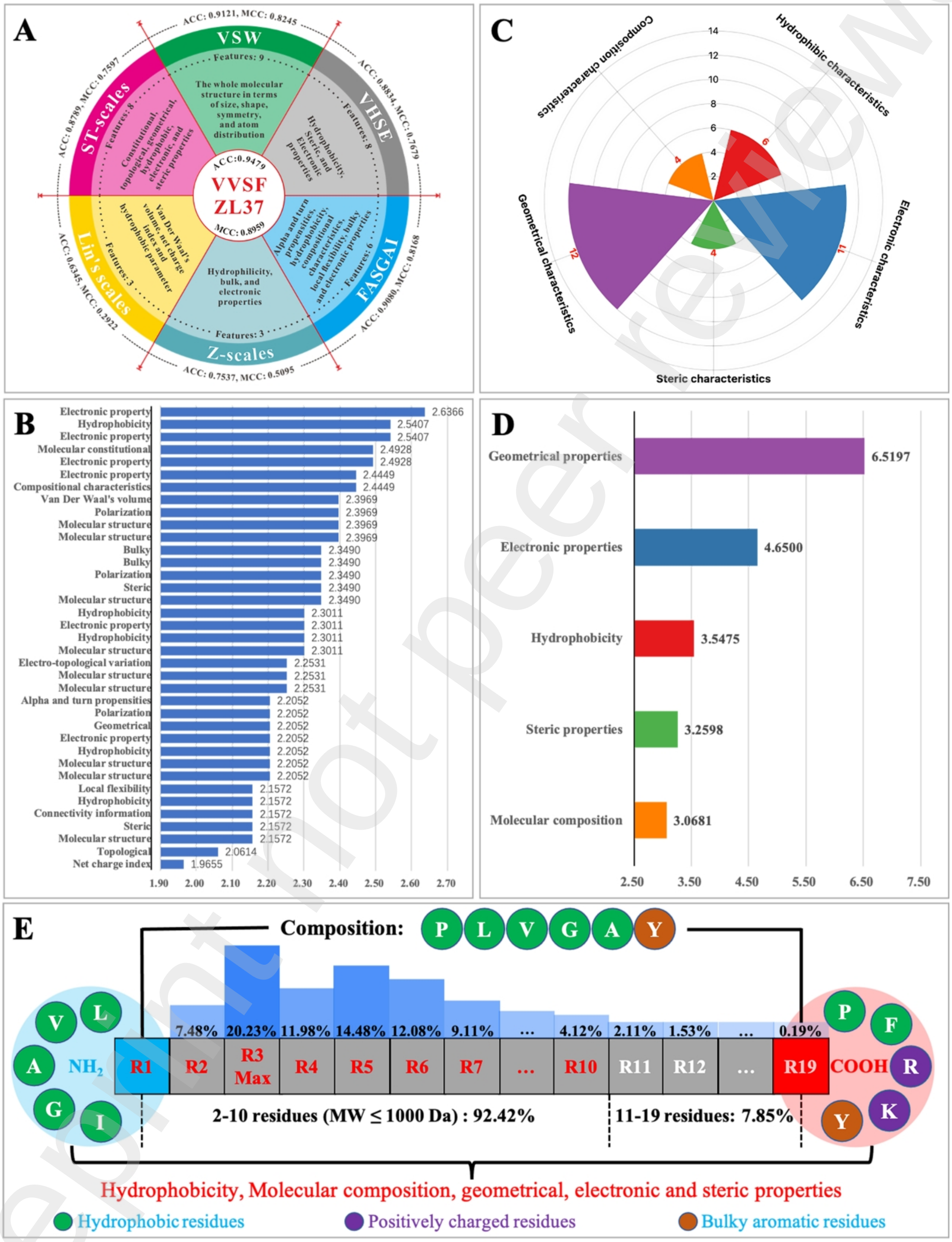


Figure 4. Feature composition and importance scores of VVSFZL37. (A) Six kinds of single AADs contained in VVSFZL37 and their physicochemical properties; (B) *Importance scores* for 37 single features; (C) Five patterns determined by physicochemical meaning; (D) *Importance scores* for five patterns of features; (E)

Sequence characteristics of ACEiPs. ACEiPs are mainly oligopeptides (2-10 residues, molecular weight ≤1000 Da) composed of hydrophobic amino acids (Pro, Leu, Val, Gly, and Ala) and aromatic amino acid (Tyr), the C-terminus tends to be hydrophobic (Pro and Phe), positively charged (Arg and Lys), and bulky aromatic (Tyr) amino acids, and the N-terminus tends to hydrophobic amino acids (Leu, Val, Ala, Gly, and Ile).

3.5. Theoretical screening and sequence characterization of food-derived ACEiPs

We used ACEiPP to predict potential ACEiPs encoded in food-derived proteins and construct theoretical libraries of ACEiPs. First, the "EHP-Tool" tool in DFBP [19] was used to simulate the hydrolysis of 21,249 food-derived proteins according to nine enzymatic hydrolysis methods. As a result, a total of 684 reported ACEiPs (accounting for 41.35% of the reported ACEiPs) and 2,300,495 unreported unique sequences were obtained. Then, ACEiPP was used to predict these unknown sequences, and 1,320,635 predicted ACEiPs with 2-19 residues were screened (probability ≥ 0.5). Nine potential libraries of ACEiPs based on nine hydrolysis methods and one integrated library of ACEiPs were constructed for download (http://www.cqudfbp.net/ACEiPP/prePool/dataPool.jsp).

The number of the predicted ACEiPs according to nine enzymatic hydrolysis methods was quite different. To be specific, *Chymotrypsin* and *Trypsin* generated two top potential ACEiPs with 317,691 and 268,492 entries, respectively (Figure 5A). This is because the cleavage sites of these two enzymes were more in line with the distribution characteristics of ACEiP residues at N- and C-terminals (Arg, Leu, Trp, Tyr, Lys), so they can obtain more candidate ACEiPs.

To verify the effectiveness of model screening, we further compared the sequence features between experimental ACEiPs (IC_{50} < 1000 μM, 1043 unique sequences) and predicted ACEiPs (1,320,635 unique sequences). The experimentally discovered ACEiPs are concentrated in tripeptides to octapeptides (mainly tripeptides and pentapeptides), whereas the predicted ACEiPs were primarily concentrated in pentapeptides to octapeptides, with the most for hexapeptides (Figure 5B and 5C). This difference may be attributed to the limitation of the actual enzymatic hydrolysis conditions and the number of reports. Prospectively, the main distribution area of ACEiPs may be concentrated in the range of pentapeptides to octapeptides, especially hexapeptides, providing an important reference for screening of ACEiPs in the future. This is because small peptides (2-4 residues) can access the active pocket of ACE, but cannot form enough hydrogen bonds, and large peptides (≥9 residues) cannot enter the active pocket easily and be hard to change the catalytic activity of ACE, whereas medium peptides (5-8 residues) might be most suitable, which is partly due to the contribution of hydrogen bonds to their affinity [38, 39]. Moreover, the predicted ACEiPs have similar residue composition and two-terminal features to experimental ACEiPs (Figure 5D-5F), i.e., they are all mainly composed of hydrophobic amino acids (Pro, Leu, Val, Gly, and Ala), the C-terminus tending to hydrophobic (Pro, Phe, and Leu), positively charged (Arg and Lys), and bulky aromatic (Tyr) amino acids, and the N-terminus favoring hydrophobic amino acids (Leu, Val, Ala, Gly, Ile, and Pro). Therefore, we not only demonstrated the feature consistency between predicted and experimental ACEiPs but also answered two open questions in ACEiP research: (i) the relationship between peptide chain length and ACE-inhibitory activity; (ii) the optimal amino acid profiles of ACEiPs derived from food proteins [24].

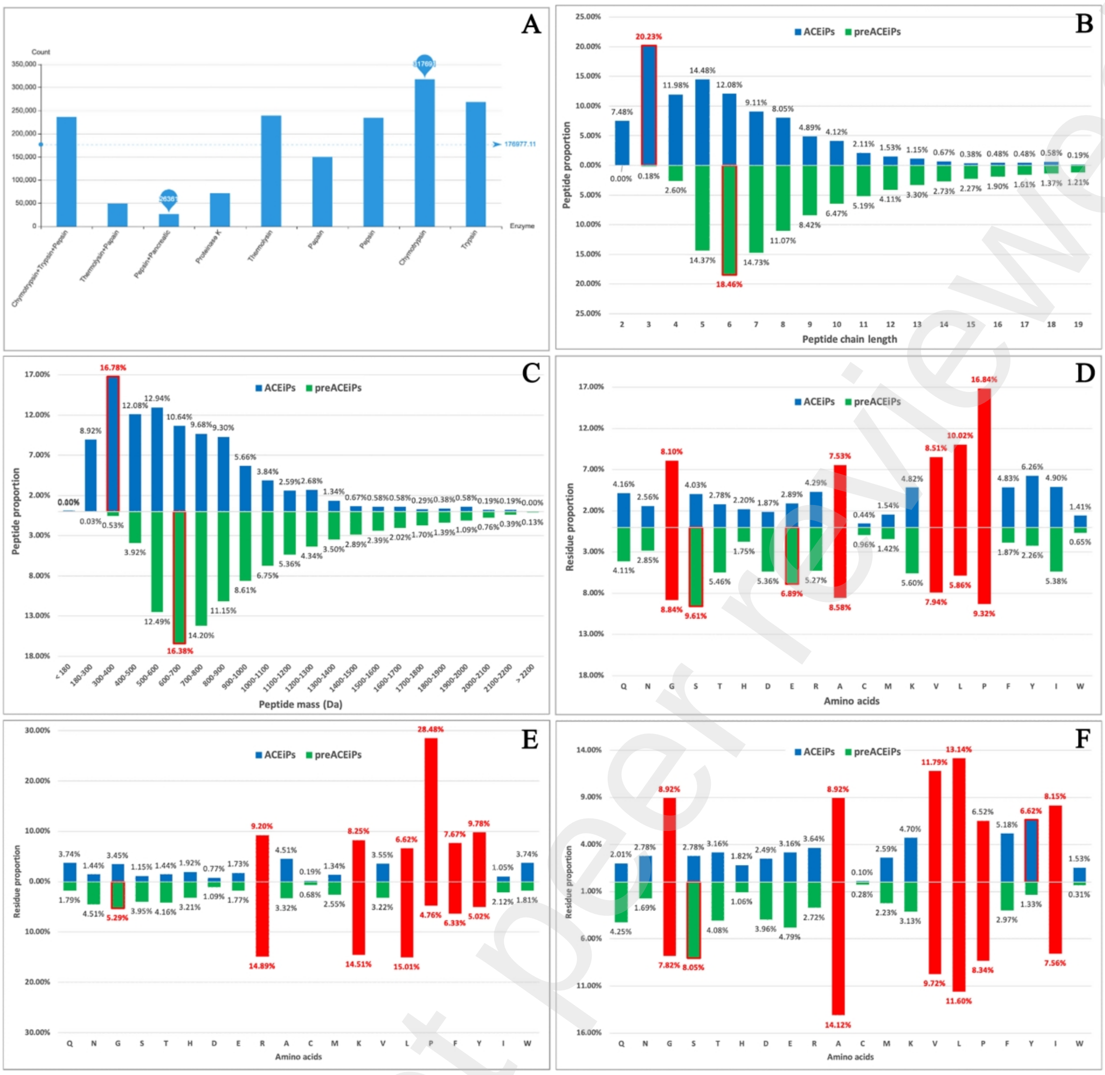


Figure 5. Hydrolyzed peptide prediction (1,320,635 unique sequences, probability ≥ 0.5) and their sequence feature comparison with experimental ACEiPs (1043 unique sequences). (A) Quantitative statistics of the predicted ACEiPs obtained by 9 different enzymatic hydrolysis methods (2 ≤ length ≤ 19); (B) Length distribution: The experimental ACEiPs are concentrated in tripeptides to octapeptides (mainly tripeptides and pentapeptides), whereas the predicted ACEiPs were mainly concentrated in pentapeptides to octapeptides, with the most for hexapeptides; (C) Mass distribution: The experimental ACEiPs and preACEiPs are mainly concentrated in < 1000 Da, and their largest sample sizes are concentrated in 300-400 Da and 600-700 Da, respectively; (D) Residue composition: Experimental and predicted ACEiPs are all mainly composed of hydrophobic amino acids (Pro, Leu, Val, Gly, and Ala); (E) C-terminal residue composition: Experimental and predicted ACEiPs have similar C-terminal residue composition, including hydrophobic (Pro, Phe, and Leu), positively charged (Arg and Lys), and bulky aromatic (Tyr) amino acids; (F) N-terminal residue composition: Experimental and predicted ACEiPs have similar N-terminal hydrophobic amino acid composition (Leu, Val, Ala, Gly, Ile, and Pro).

3.6. Characterization and screening of multifunctional peptides

A total of 653 MBPs was found among the reported 1654 ACEiPs, involving 23 functional activities, especially antihypertensive, antioxidant, DPP IV-inhibitory, renin-inhibitory, and anticancer peptides (Supplementary Figure S1). This high probability (653/1654=39.48%) was mainly due to the fact that the above five types of peptides have similar or same sequence characteristics with ACEiPs (Supplementary Table S9), including (i) main residue distribution in the N-terminus (Leu, Val, and Gly) and C-terminus (Pro, Tyr, Lys, Leu, and Phe), (ii) high residue composition (Pro, Leu, Val, and Gly), and (iii) general length less than ten residues. Therefore, it is meaningful to acquire potential MBPs by screening ACEiPs from bioactive peptides with no reported ACE-inhibitory activity.

Total 30 types of food-derived bioactive peptides (2995 unique sequences) other than ACEiPs from DFBP [19] were collected. Excluding 653 peptides with anti-ACE activity reported in the literature, 2346 peptides were obtained. We then used ACEiPP to predict the 2346 peptides to determine whether they were ACEiPs. As a result, a total of 1358 peptides exhibited potential ACE-inhibitory activity. We finally integrated these peptides into a predicted library of MBPs for researchers to screen peptides with ACE inhibition and relevant activities (such as antioxidant, DPP IV-inhibitory, and antithrombotic, anticancer, etc.). This is the first proposed method to mine MBPs from known bioactive peptides using prediction tools, which provides an important approach for the screening and design of drugs for multifactorial diseases (such as cardiovascular diseases, cancer, etc.).

3.7. Web server implementation

We built the ACEiPP server (http://www.cqudfbp.net/ACEiPP/index.jsp) with the optimal LSTM model trained by VVSFZL37 to predict, design, and screen ACEiPs. ACEiPP comprises six modules, including Home, Pre-ACEiPs, Seq-Features, Pre-Libraries, Help, and Contact (Supplementary Figure S3). As the first DL-based prediction platform for ACEiPs, ACEiPP can perform multiple applications such as prediction, residue-based mutation screening, structure-activity exploration of new ACEiPs, and discovery of potential ACEiPs and MBPs.

4. Conclusions

Overall, we developed the first ACEiP predictor (ACEiPP) to predict, design, and screen ACEiPs using a DL-based model based on natural sequences processing by LSTMs and the interpretability of AADs. ACEiPP effectively learns the key sequence/structure characteristics of ACEiPs transformed by VVSFZL37 and improves the prediction performance compared with the existing three models. We use ACEiPPs to screen 1,320,635 potential ACEiPs encoded in 21,249 food-derived proteins and construct theoretical peptide libraries of ACEiPs according to different hydrolysis modes and predicted probabilities. Moreover, we construct an MBP library containing 1358 entries with potential ACE inhibition and other known bioactivities (Hypertension, diabetes, hyperlipidemia, etc.). Hence, ACEiPP will be considered a leading tool in peptide research and drug development. In addition, our method may be applied to other DL-based studies on peptide sequences for transfer learning, so the next directions, we will try to construct other bioactive peptide classifiers and deeply mine the structure-activity relationship between different bioactive peptides.

Conflict of interest

No conflict of interest

Funding

This work was supported by the National Natural Science Foundation of China [grant numbers 31771975, 31571782].

References

[1] Brouwers S, Sudano I, Kokubo Y, Sulaica EM. Arterial hypertension, Lancet 2021;398:249-61.

[2] Oparil S, Acelajado MC, Bakris GL, Berlowitz DR, Cífková R, Dominiczak AF, Grassi G, Jordan J, Poulter NR, Rodgers A, Whelton PK. Hypertension, Nat Rev Dis Primers 2018;4:18014.

[3] Valenzuela PL, Carrera-Bastos P, Gálvez BG, Ruiz-Hurtado G, Ordovas JM, Ruilope LM, Lucia A. Lifestyle interventions for the prevention and treatment of hypertension, Nature Reviews: Cardiology 2021;18:251-75.

[4] Messerli FH, Williams B, Ritz E. Essential hypertension, Lancet 2007;370:591-603.

[5] Henninot A, Collins JC, Nuss JM. The current state of peptide drug discovery: Back to the future?, Journal of Medicinal Chemistry 2018;61:1382-414.

[6] Rifaioglu AS, Atas H, Martin MJ, Cetin-Atalay R, Atalay V, Dogan T. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases, Briefings in Bioinformatics 2019;20:1878-912.

[7] Balthazar CF, Guimarães JF, Coutinho NM, Pimentel TC, Ranadheera CS, Santillo A, Albenzio M, Cruz AG, Sant'Ana AS. The future of functional food: Emerging technologies application on prebiotics, probiotics and postbiotics, Comprehensive Reviews In Food Science And Food Safety 2022;21:2560-86.

[8] Abachi S, Bazinet L, Beaulieu L. Antihypertensive and angiotensin-I-converting enzyme (ACE)-inhibitory peptides from fish as potential cardioprotective compounds, Marine Drugs 2019;17:613.

[9] Martin M, Deussen A. Effects of natural peptides from food proteins on angiotensin converting enzyme activity and hypertension, Critical Reviews in Food Science and Nutrition 2019;59:1264-83.

[10] Liu YQ, Strappe P, Shang WT, Zhou ZK. Functional peptides derived from rice bran proteins, Critical Reviews in Food Science and Nutrition 2019;59:349-56.

[11] Maleki S, Razavi SH. Pulses' germination and fermentation: Two bioprocessing against hypertension by releasing ACE inhibitory peptides, Critical Reviews in Food Science and Nutrition 2021;61:2876-93.

[12] Lee SY, Hur SJ. Antihypertensive peptides from animal products, marine organisms, and plants, Food Chemistry 2017;228:506-17.

[13] Aluko RE. Antihypertensive peptides from food proteins, Annual Review of Food Science and Technology 2015;6:235-62.

[14] Manikkam V, Vasiljevic T, Donkor ON, Mathai ML. A review of potential marine-derived hypotensive and anti-obesity peptides, Critical Reviews in Food Science and Nutrition 2016;56:92-112.

[15] Morales D, Miguel M, Garcés-Rimón M. Pseudocereals: a novel source of biologically active peptides, Critical Reviews in Food Science and Nutrition 2021;61:1537-44.

[16] Su Y, Chen S, Cai S, Liu S, Pan N, Su J, Qiao K, Xu M, Chen B, Yang S, Liu Z. A novel angiotensin-I-converting enzyme (ACE) inhibitory peptide from takifugu flavidus, Marine Drugs 2021;19:651.

[17] Hyoung Lee D, Ho Kim J, Sik Park J, Jun Choi Y, Soo Lee J. Isolation and characterization of a novel angiotensin I-converting enzyme inhibitory peptide derived from the edible mushroom Tricholoma giganteum, Peptides 2004;25:621-7.

[18] Wang J, Wang G, Zhang Y, Zhang R, Zhang Y. Novel angiotensin-converting enzyme inhibitory peptides identified from walnut glutelin-1 hydrolysates: molecular interaction, stability, and antihypertensive effects, Nutrients 2021;14:151.

[19] Qin D, Bo W, Zheng X, Hao Y, Li B, Zheng J, Liang G. DFBP: A comprehensive database of food-derived bioactive peptides for peptidomics research, Bioinformatics 2022;38:3275-80.

[20] Lammi C, Aiello G, Boschin G, Arnoldi A. Multifunctional peptides for the prevention of cardiovascular disease: A new concept in the area of bioactive food-derived peptides, Journal of Functional Foods 2019;55:135-45.

[21] Wu Q, Luo F, Wang XL, Lin Q, Liu GQ. Angiotensin I-converting enzyme inhibitory peptide: an emerging candidate for vascular dysfunction therapy, Critical Reviews in Biotechnology 2022;42:736-55.

[22] Rifaioglu AS, Atas H, Martin MJ, Cetin-Atalay R, Atalay V, Doğan T. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases, Briefings in Bioinformatics

2019;20:1878-912.

[23] Bo W, Chen L, Qin D, Geng S, Li J, Mei H, Li B, Liang G. Application of quantitative structure-activity relationship to food-derived peptides: Methods, situations, challenges and prospects, Trends In Food Science & Technology 2021;114:176-88.

[24] Xiang L, Qiu Z, Zhao R, Zheng Z, Qiao X. Advancement and prospects of production, transport, functional activity and structure-activity relationship of food-derived angiotensin converting enzyme (ACE) inhibitory peptides, Critical Reviews in Food Science and Nutrition 2021:1-27.

[25] Hellberg S, Sjöström M, Skagerberg B, Wold S. Peptide quantitative structure-activity relationships, a multivariate approach, Journal of Medicinal Chemistry 1987;30:1126-35.

[26] Shu M, Mei H, Yang S, Liao L, Li Z. Structural parameter characterization and bioactivity simulation based on peptide sequence, QSAR & Combinatorial Science 2009;28:27-35.

[27] Liang G, Li Z. Factor analysis scale of generalized amino acid Information as the source of a new set of descriptors for elucidating the structure and activity relationships of cationic antimicrobial peptides, QSAR & Combinatorial Science 2007;26:754-63.

[28] Kumar R, Chaudhary K, Singh Chauhan J, Nagpal G, Kumar R, Sharma M, Raghava GP. An in silico platform for predicting, screening and designing of antihypertensive peptides, Scientific Reports 2015;5:12512.

[29] Manavalan B, Basith S, Shin TH, Wei L, Lee G. mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation, Bioinformatics 2019;35:2757-65.

[30] Akbar S, Hayat M, Tahir M, Khan S, Alarfaj FK. cACP-DeepGram: Classification of anticancer peptides via deep neural network and skip-gram-based word embedding model, Artificial Intelligence in Medicine 2022;131:102349.

[31] Hussain W. sAMP-PFPDeep: Improving accuracy of short antimicrobial peptides prediction using three different sequence encodings and deep neural networks, Briefings in Bioinformatics 2022;23:bbab487.

[32] Tang W, Dai R, Yan W, Zhang W, Bin Y, Xia E, Xia J. Identifying multi-

functional bioactive peptide functions using multi-label deep learning, Briefings in Bioinformatics 2022;23:bbab414.

[33] Hochreiter S, Schmidhuber J. Long short-term memory, Neural Computation 1997;9:1735-80.

[34] Cheng Y, Gong Y, Liu Y, Song B, Zou Q. Molecular design in drug discovery: a comprehensive review of deep generative models, Briefings in Bioinformatics 2021;22:bbab344.

[35] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 2012;28:3150-2.

[36] U GY, Bhat I, Karunasagar I, B SM. Antihypertensive activity of fish protein hydrolysates and its peptides, Critical Reviews in Food Science and Nutrition 2019;59:2363-74.

[37] Kaur A, Kehinde BA, Sharma P, Sharma D, Kaur S. Recently isolated food-derived antihypertensive hydrolysates and peptides: A review, Food Chemistry 2021;346:128719.

[38] Lin K, Zhang L-w, Han X, Cheng D-y. Novel angiotensin I-converting enzyme inhibitory peptides from protease hydrolysates of Qula casein: Quantitative structure-activity relationship modeling and molecular docking study, Journal of Functional Foods 2017;32:266-77.

[39] Lourenço da Costa E, Antonio da Rocha Gontijo J, Netto FM. Effect of heat and enzymatic treatment on the antihypertensive activity of whey protein hydrolysates, International Dairy Journal 2007;17:632-40.