ACS
FOOD
SCIENCE & TECHNOLOGY

pubs.acs.org/acsfoodscitech Article

Computer-Aided Approaches for Screening Antioxidative Dipeptides and Application to Sorghum Proteins

Zhenjiao Du and Yonghui Li*

Cite This: https://doi.org/10.1021/acsfoodscitech.2c00286 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

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

KEYWORDS: *antioxidant peptides, artificial intelligence, hydrolysis simulation, sorghum protein, in silico*

1. INTRODUCTION

Developing natural antioxidants has gained expanding interest because of the increasing health concerns for synthetic antioxidants and strict regulation on their usage.[1−3] Antioxidant peptides derived from food proteins have captured worldwide attention due to their advantages such as naturally sourced, better sustainability, and no or low toxic effects.[1,3,4] Generally, antioxidant peptides are released from parent proteins by enzymatic hydrolysis, fermentation, chemical hydrolysis, germination, and/or ripening and then screened by laborious chemistry methods (e.g., fractionation, isolation, purification, identification, and characterization).[1,5] However, the conventional wet-chemistry methods are time-consuming and rely highly on many advanced instruments and equipment.[1,2,5] The chemical synthesis of peptides is an alternative approach for producing and screening potentially highly active peptides.[5−7] Nonetheless, it is practically impossible to synthesize all the peptides for antioxidative peptide screening, considering the cost of synthesis and a large number of theoretically possible peptides: i.e., 400 dipeptides, 8000 tripeptides, 160000 tetrapeptides, etc.[4,8,9]

Tremendous bioactive peptides have been identified, making it possible to use these accumulated activity data for modeling quantitative structure−activity relationships. The models can also provide more efficient and cost-effective guidance for the exploration of new bioactive peptides.[1,2,10,11] Such *in-silico* approaches have been successfully applied to predict angiotensin I converting enzyme inhibitory (ACE-I) activity, dipeptidyl peptidase IV (DPP-IV) inhibitory activity, and antimicrobial activity.[12−14] Besides, some peptide cutting simulation tools (e.g., PeptideCutter) were developed for protein *in-silico* hydrolysis, which can be combined with more than 156 million identified protein sequences from the Uniprot database to obtain potentially bioactive peptide sequences generated by specific enzymes or a combination of multiple enzymes.[15−18] There have been few studies and limited models on antioxidant peptides, especially for dipeptides, which demonstrate ideal absorption ability and bioavailability in the intestine compared to larger peptides.[7−9,19] In addition, the order of *in-silico* hydrolysis with multiple enzymes was not specified in the previous peptide cutter tools, which is critical in practical experiments and should be developed.[15−18]

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

Received: September 14, 2022
Revised: October 21, 2022
Accepted: October 25, 2022

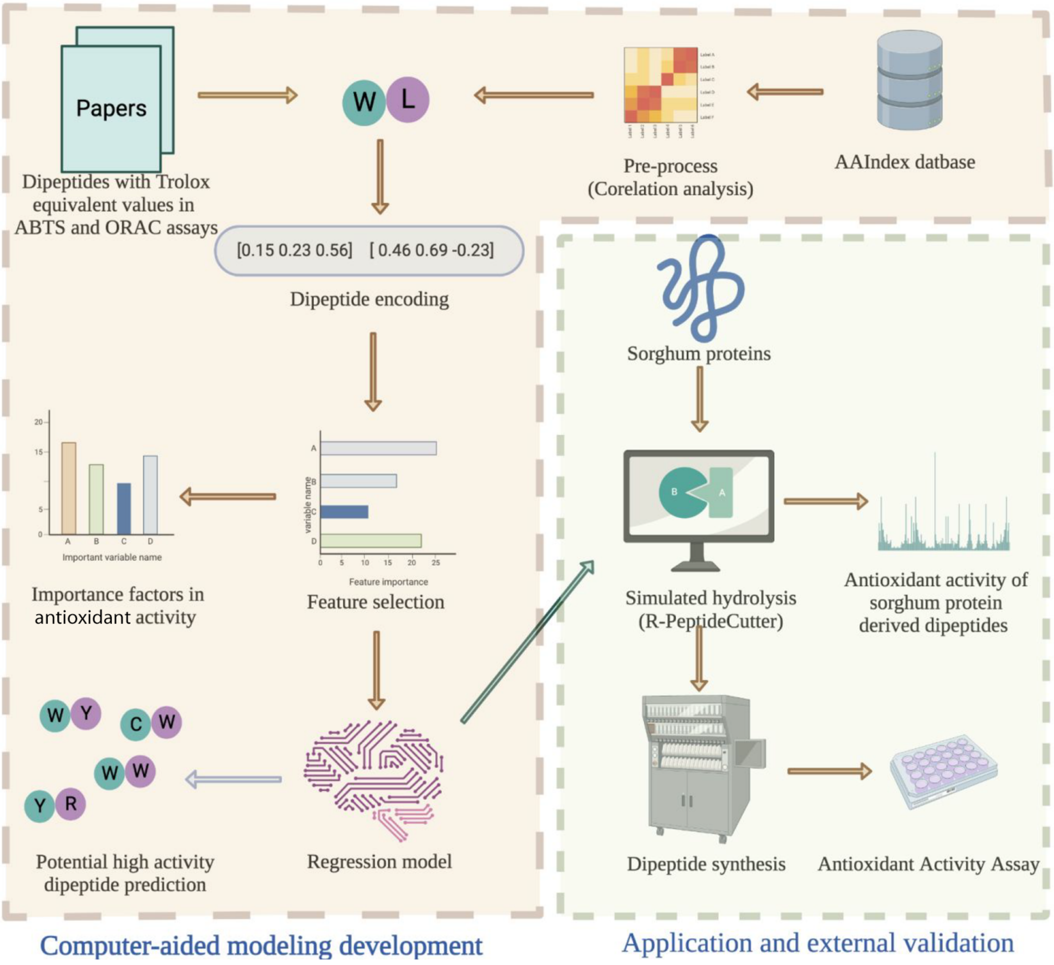
Papers
Dipeptides with Trolox equivalent values in ABTS and ORAC assays
W L
Pre-process (Corelation analysis)
AAIndex dbatase
[0.15 0.23 0.56] [0.46 0.69 -0.23]
Dipeptide encoding
Sorghum proteins
Importance factors in antioxidant activity
Important variable name
Feature selection
Feature importance
variable name
Simulated hydrolysis (R-PeptideCutter)
Antioxidant activity of sorghum protein derived dipeptides
W Y
C W
W W
Y R
Potential high activity dipeptide prediction
Regression model
Dipeptide synthesis
Antioxidant Activity Assay
Computer-aided modeling development
Application and external validation

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

validation. The overall technical workflow of the study is summarized in Figure 1.

2. MATERIALS AND METHODS

2.1. Data Set Collection. Data mining (Beautiful Soup, 4.5.3) was used to collect 566 numerical indices of amino acids from AAIndex, and detailed definitions and descriptions of each index are available online (https://www.genome.jp/aaindex/).[20] The indices with missing values for amino acids were manually deleted, resulting in a total of 553 remaining indices (Supplementary Document 1). Antioxidant activities of dipeptides based on *in vitro* antioxidant assays including both a 2,2′-azinobis(3-ethylbenzothiazoline-6-sulfonate) (ABTS) radical scavenging capacity assay and an oxygen radical absorbance capacity (ORAC) assay, were manually collected from published studies for antioxidant activity prediction model development. Sixty-seven antioxidant dipeptides characterized by an ABTS radical scavenging capacity assay and seventy-three dipeptides characterized by an ORAC assay were collected as two separate data sets (Supplementary Document 2), and the antioxidant activity values are expressed as Trolox-equivalent antioxidant capacity (μmol TE/μmol peptide).[7,21−26]

2.2. Data Processing. *2.2.1. Preprocessing of Numerical Indices of Amino Acids.* Among the 553 numerical indices, some of them are highly correlated. It is necessary to remove those redundant features, as they provide very limited information but increase model complexity. Collinearity of the 553 numerical indices was prescreened by pairwise correlation methods (Supplementary Documents 3 and 4). If an absolute value of Pearson's correlation coefficient between two indices was above 0.95, one of them was removed randomly due to the strong correlation.[27] The remaining numerical indices were standardized for further feature selection.

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

Feature selection was used to further screen the important features particularly for antioxidant activity prediction: therefore, significantly simplifying the model complexity. In addition, it can also shed light on the most important features contributing to antioxidant activity.[4] These samples with $2n$ variables and corresponding Y-vector in each activity data set were modeled by extreme gradient boosting (XGboost) regression. Feature importance was calculated and used to identify the key variables for antioxidant activity prediction.[28]

2.3. Model Development. *2.3.1. Data Set Division.* Data sets were shuffled and randomly split into a training data set and a test data set at a 3:1 ratio. For the ABTS radical scavenging capacity data set, 51 samples were used in the training data set for model building, and the remaining 16 samples were used in the test data set to evaluate the performance of the model. For the ORAC data set, the numbers of samples in the training data set and test data set were 55 and 18, respectively.

2.3.2. Model Building, Optimization, and Evaluation. XGBoost was employed to build a regression model based on the training data set, and the parameters were tuned by leave-one-out cross-validation (LOOCV). The hyperparameters with the best performance in LOOCV were used as the final model for performance evaluation with the test data set. The detailed codes for model development are available in Supplementary Document 5. Model development was performed using Python 3.8.8 (MacOS Monterey 12.0.1, CPU intel Core-i5 2.3 GHz), and the related functions are available through scikit-learn (https://scikit-learn.org/) and XGBoost (https://xgboost.readthedocs.io/en/stable/).[28,29]

Coefficients of determination (R^2) and mean square errors (MSE) were used to evaluate the model performance. The R^2 and MSE values from the training data set, LOOCV, and test data set were labeled as R^2_{Train} and $\text{MSE}_{\text{Train}}$, R^2_{CV} and MSE_{CV}, and R^2_{Test} and MSE_{Test}, respectively.

2.4. Prediction of Dipeptides with High Antioxidant Activity. A data set containing 400 possible dipeptides with ABTS radical scavenging capacity and ORAC was built, and the published 67 ABTS-associated dipeptides and 73 ORAC-associated dipeptides in model building and validation were also included. After obtaining the two prediction models, the 400 possible dipeptides were encoded by the selected features, and then their antioxidant activities were predicted (Supplementary Document 6).

2.5. Protein Cutting Simulation Tool Development. The cleavage sites of 51 enzymes and chemicals were collected from published studies and publicly available web servers such as PeptideCutter.[16,18,30] Functions for a total of 51 enzymes and chemicals in the simulation tool were designed to obtain specific cleavage sites in a given protein sequence, including Arg-C proteinase, Asp-N endopeptidase, Asp-N endopeptidase + N-terminus Glu, BNPS-Skatole, caspase1, caspase2, caspase3, caspase4, caspase5, caspase6, caspase7, caspase8, caspase9, caspase10, chymotrypsin-high specificity (C-term to [F Y W], not before P), chymotrypsin-low specificity (C-term to [F Y W M L], not before P), clostripain (clostridiopeptidase B), CNBr, enterokinase, factor Xa, formic acid, glutamyl endopeptidase, granzymeB, hydroxylamine, iodosobenzoic acid, LysC, neutrophil elastase, NTCB (2-nitro-5-thiocyanobenzoic acid), pepsin (pH = 1.3), pepsin (pH > 2), proline-endopeptidase, proteinase K, staphylococcal peptidase I, thermolysin, thrombin, trypsin, elastase 1, elastase 2, chymotrypsinogen B1, chymotrypsinogen C, pancreatic enteropeptidase E enteropeptidase, prostasin, gastricsin, fruit bromelain, stem bromelain, ananain, papaya proteinase 4, chymopapain, chymosin, and caricain.[16,18] All the detailed cleavage sites of enzymes and chemicals are available in (Table S3 in Supplementary Document 7).

Based on these designed functions, we developed a user-friendly and well-annotated protein hydrolysis simulation tool, named Refining-PeptideCutter (R-PeptideCutter), which takes the adding sequence of enzymes or chemicals for hydrolysis into consideration. Users will only need to download the fasta format files containing the parent protein sequence from Uniprot, set enzymes or chemicals and desired peptide length in R-PeptideCutter, and run the scripts. It will automatically generate all the possible peptides encrypted in the protein sequence as well as the number of peptides that can be released. The detailed codes were written in Python and are available in Supplementary Document 7. In addition, the R-PeptideCutter tool was further tailored to link with the predicted antioxidant results from the newly developed models, so that the antioxidant activity values of the *in-silico* released peptides could be assigned. These codes can be modified and used for studies of other bioactive peptides and are available in Supplementary Document 7.

2.6. Application of Simulation Tool and Activity Prediction Models in Kafirin Proteins. Eight available kafirin sequences were selected from Uniprot (https://www.uniprot.org/) and used for hydrolysis simulation, including three α-kafirin sequences, one β-kafirin sequence, two γ-kafirin sequences, and two δ-kafirin sequences (Table S1 in Supplementary Document 7). All of the downloaded kafirin sequences were pretreated by removing the signal peptide from the peptide sequences before further hydrolysis simulation. All of these sequences were simulated to be hydrolyzed by one or a combination of two enzymes or chemicals in order for all 51 enzymes or chemicals. The predicted antioxidant activity values (ABTS radical scavenging capacity and ORAC) of the generated peptides were assigned, respectively (Supplementary Document 7).

2.7. Antioxidant Peptide Synthesis. Eight dipeptides present in kafirin sequences were purchased from Sigma-Aldrich (St. Louis, MO, USA). The purity of these dipeptides was all above 95%. Taking into consideration both peptide diversity and simulated hydrolysis results with kafirin proteins, some peptides with low antioxidant activity based on model prediction were also selected. These peptides (CG, AY, YA, YF, GW, GG, GT, and GV) were used to validate the performance of the constructed models.

2.8. Antioxidant Activity Assays. 1,1-Diphenyl-2-picrylhydrazyl (DPPH), ABTS, fluorescein disodium (FL), and 2,2′-azobis(2-methylpropionamide) dihydrochloride (AAPH) were purchased from Sigma-Aldrich (St. Louis, MO, USA). All of the chemicals and reagents used were of analytical grade. In addition to the ABTS and ORAC assays, we also conducted the DPPH assay, which can be used for database building in future antioxidant-activity-related models. All tests were conducted in triplicate.

2.8.1. ABTS Radical Scavenging Capacity Assay. The ABTS radical scavenging capacity assay was conducted according to the study of Zheng et al.[7] Briefly, 150 μL of $\text{ABTS}^{\bullet+}$ solution was mixed with 50 μL of dipeptide solution (10 μM) in a 96-well microplate. After 30 min incubation at 30 °C, the absorbance was measured at 734 nm using a Biotek Synergy H1 Hybrid Microplate Reader (Winooski, VT, USA). An equivalent volume of 50 mM phosphate buffer (PBS) at pH 7.4 was used as the control, and the initial absorbance at 734 nm was controlled at 0.70 ± 0.02. The dipeptide solution was prepared in 75 mM PBS buffer (pH 7.4). Trolox (TE) was used as a standard antioxidant, and results were expressed as μmol TE/μmol peptide.

2.8.2. ORAC Assay. The ORAC assay was performed according to the study of Zheng et al.[7] A 20 μl portion of the dipeptide solution (20 μM) and 60 μL of the FL solution (5 nM) were transferred in a well of a 96-well microplate, and incubated for 15 min at 37 °C. Then 120 μL of a AAPH solution (80 mM) was mixed with the incubated mixture in the plate for 30 s. A Biotek Synergy H1 Hybrid Microplate Reader was used to record the fluorescence for 100 min at 485 nm for excitation and 520 nm for emission, respectively. All of the dipeptide solutions and FL and AAPH solutions were prepared in 75 mM PBS buffer (pH 7.4). Trolox was used as a standard antioxidant, and the ORAC values were also expressed as μmol TE/μmol peptide.

2.8.3. DPPH Radical Scavenging Capacity Assay. The DPPH radical scavenging capacity assay was performed according to the study of Chen et al., with some modifications.[8] Briefly, 100 μL of a dipeptide solution (20 μM) was mixed with 100 μL of a DPPH solution (0.2 mM in 95% ethanol) in a 96-well microplate and then incubated for 30 min at room temperature in the dark. The absorbance was measured at 517 nm using a Biotek Synergy H1 Hybrid Microplate Reader (Winooski, VT, USA). The dipeptide solution was prepared in 75 mM PBS buffer (pH 7.4). Trolox was used as the standard antioxidant, and the results were expressed as μmol TE/μmol peptide.

3. RESULTS AND DISCUSSION

3.1. Feature Importance Analysis. Eleven variables were selected by XGBoost regression with a feature importance threshold of 0.01 (Table 1) and then used to encode the 67 dipeptides with known ABTS radical scavenging capacity values as the X-matrix (i.e., 67 × 11). Among them, 7 variables were used to encode N-terminus residues, accounting for 0.8109 in feature importance for ABTS radical scavenging capacity prediction, while the sum of the feature importance of the remaining 4 variables representing C-terminus residues was only 0.1087. LIFS790103 and CHAM820102 standing for "conformational preference for antiparallel β strands" and "free

Table 1. Amino Acid Positions, Variable Importance, and Description of Selected Variables by XGBoost Regression for ABTS Radical Scavenging Capacity Data Set

accession no.	amino acid position	variable importance[a]	description
LIFS790103	N-terminus	0.4321	conformational preference for antiparallel β strands
CHAM820102	N-terminus	0.1647	free energy of solution in water
PALJ810108	N-terminus	0.0834	normalized frequency of α helix in α + β class
ARGP820102	N-terminus	0.0732	signal sequence helical potential
KARS160122	N-terminus	0.0268	weighted second smallest eigenvalue of the weighted Laplacian matrix
OOBM850103	N-terminus	0.0167	optimized transfer energy parameter
YUTK870102	N-terminus	0.0139	unfolding Gibbs energy in water, pH 9.0
CHAM830103	C-terminus	0.0419	no. of atoms in the side chain labeled 1 + 1
MAXF760102	C-terminus	0.0281	normalized frequency of extended structure
QIAN880119	C-terminus	0.0272	weights for β sheet at the window position of −1
MCMT640101	C-terminus	0.0116	refractivity

[a]Variable importance is presented as absolute values. The detailed information on these selected variables is available at https://www.genome.jp/aaindex/.

Table 2. Amino Acid Positions, Variable Importance, and Description of Selected Variables by XGBoost Regression for ORAC Data Set

accession no.	amino acid position	variable importance[a]	description
CHOP780215	N-terminus	0.2067	frequency of the fourth residue in turn
PALJ810108	N-terminus	0.2046	normalized frequency of α helix in α + β class
CHOP780205	N-terminus	0.0918	normalized frequency of C-terminus helix
LIFS790103	N-terminus	0.0819	conformational preference for antiparallel β strands
BIOV880102	N-terminus	0.0323	information value for accessibility; average fraction 23%
FODM020101	N-terminus	0.0153	propensity of amino acids within π helices
BROC820102	N-terminus	0.0112	retention coefficient in HFBA
CHOP780211	C-terminus	0.0557	normalized frequency of C-terminus non-β region
BUNA790103	C-terminus	0.0244	spin−spin coupling constants 3JHalpha-NH
QIAN880136	C-terminus	0.0181	weights for coil at the window position of 3
PALJ810111	C-terminus	0.0122	Normalized frequency of β sheet in α + β class
CHAM830105	C-terminus	0.0121	no. of atoms in the side chain labeled 3 + 1
KARS160112	C-terminus	0.0119	second smallest eigenvalue of the Laplacian matrix of the graph
DESM900101	C-terminus	0.1105	membrane preference for cytochrome *b*: MPH89

[a]Variable importance is presented as absolute values. The detailed information on these selected variables is available at https://www.genome.jp/aaindex/.

energy of solution in water", respectively, were the two most important variables for ABTS radical scavenging capacity prediction, and both were associated with N-terminus residues.

Similarly, 14 important variables for the 73 dipeptides with known ORAC values were selected by the same feature selection method employed for ABTS radical scavenging capacity values (Table 2). The N-terminus residue was represented by 7 variables, accounting for 0.6437 in feature importance for ORAC prediction, while only 0.2449 in feature importance stood for C-terminus with the same number of representative variables as N-terminus. CHOP780215 (N-terminus), PALJ810108 (N-terminus), and DESM900101(C-terminus) standing for "frequency of the fourth residue in turn", "normalized frequency of α helix in α + β class", and "membrane preference for cytochrome *b*: MPH89" were the three most important variables for ORAC activity prediction.

Feature importance of these variables in ABTS radical scavenging capacity model development and ORAC model development showed that N-terminus residues played a more important role in antioxidant activity. In the study of Chen et al.,[8] the variable importance in projection (VIP) values revealed that the C-terminus had greater contribution to the ABTS radical scavenging capacity of tripeptides, and such an observation was also confirmed in the study of Du et al.[4] It is interesting that antioxidant dipeptides showed a different pattern where the importance of the N-terminus was greater than that of the C-terminus. This is consistent with the VIP values from the partial least-squares regression (PLSR) models in Zheng et al., although they instead further emphasized the importance of tyrosine and tryptophan at the N-terminus.[7] Compared to ABTS radical scavenging capacity prediction, C-terminus residues contributed more to ORAC prediction. Among the selected variables, CHAM820102 and YUTK870102 for N-terminus residues encoding in ABTS radical scavenging capacity prediction were also selected in the study of Chen et al. for a tripeptide's ABTS radical scavenging capacity prediction, which showed their generality in antioxidant activity prediction.[8] Previous studies on peptides were mainly focused on the amino acid composition of antioxidant peptides for the explanation of the antioxidant activity (e.g., residue content of tyrosine).[7−9,19,31,32]

Some variables, such as LIFS790103, were selected in both models. The feature importances of LIFS790103 were 0.4321 and 0.0819 for the ABTS radical scavenging capacity corresponding model and ORAC corresponding model, respectively. This might be due to the difference in mechanisms between a single electron transfer (SET) mechanism and hydrogen atom transfer (HAT) mechanism, since ABTS and ORAC assays belong to SET and HAT mechanisms, respectively.[33] However, it should be mentioned that some of the variables, such as LIFS790103 standing for conformational preference for antiparallel β strands, were not directly related to the antioxidant activity of dipeptides, since there was no secondary structure in dipeptides. This issue complicated the interpretation of models, which was also indicated by other researchers for machine-learning applications.[8,9] Compared to previous studies based on components from principal component analysis (PCA) as the variables for model development, our selected variables further clarified specific variables associated with the antioxidant activity instead of groups of variables and also enlarged the feature source for peptide encoding.[34−36] In addition, the selected

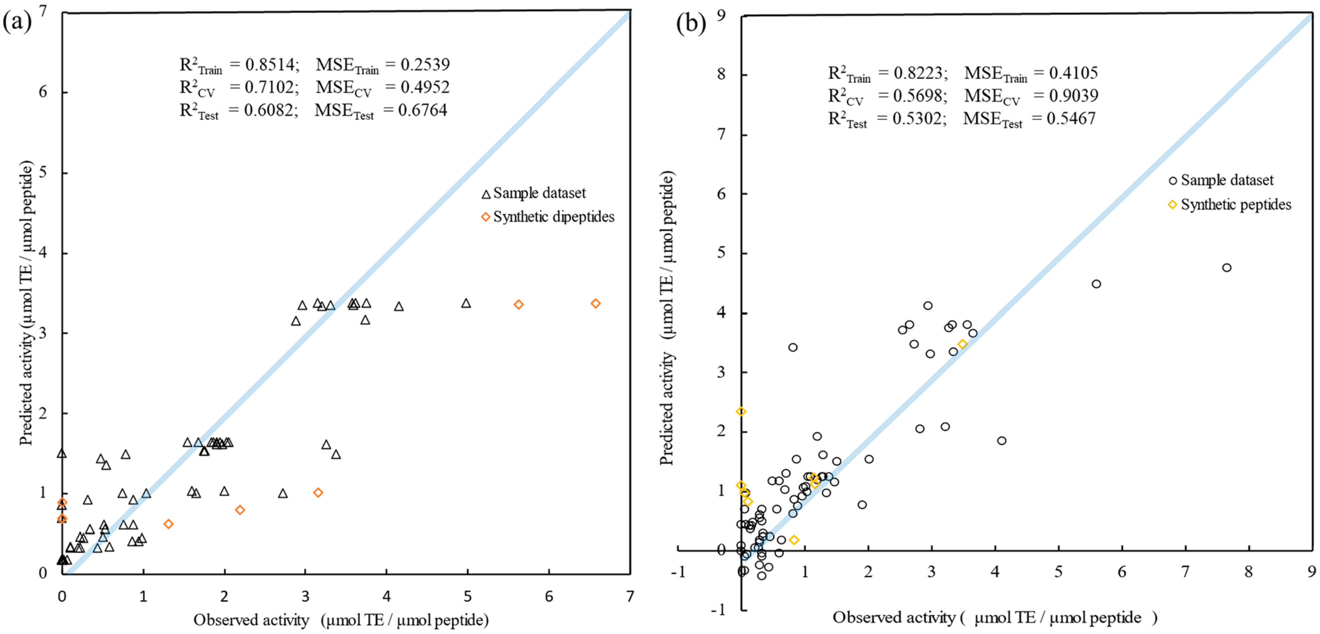

Figure 2. Relationship between observed and predicted antioxidant activities: (a) ABTS scavenging activity; (b) ORAC.

variables from our study were more targeted and less redundant compared with previous models adopting amino acid descriptors (e.g., *z* scale in the study of Zheng et al.) comprised of various weighted variables in combination, since there are no amino acid descriptors specifically designed for antioxidant activity.[13,35,36] These findings provide alternative ways for further theoretical and mechanism studies on antioxidant peptides.[9,19]

3.2. Model Performance Evaluation. The relationships between the observed activity and the predicted activity of ABTS and ORAC models are shown in Figure 2a,b, respectively. The R^2_{CV} and R^2_{Test} values for the ABTS and ORAC prediction models were 0.7102/0.6082 and 0.5698/0.5302, respectively. For the ABTS model, the predicted activity was inclined to be lower than the observed activity. In particular, the observed activity of around 2−4 μmol TE/μmol peptide was predicted to be 1−2 μmol TE/μmol peptides. In contrast, the predicted activity was much closer to the observed activity in the ORAC model except for several samples that showed a larger difference between the predicted activity and the observed activity (e.g., the highest observed activity sample).

Our model performances were substantial breakthroughs compared to the previous study conducted by Zheng et al., where the best R^2_{CV} values, instead of R^2_{Test}, for ABTS radical scavenging capacity and ORAC prediction were only 0.38 and 0.26, respectively, relying on a linear regression modeling method (PLSR).[7] Technically, R^2_{CV} could not describe the model's performance in an unknown data set because the data used to obtain R^2_{CV} were used in model building.[9] In this study, a much larger data set was collected from the latest publications, and feature selection with nonlinear regression modeling approaches was employed to link the data set and antioxidant activity, which all contributed to the better performance in test data sets.[4,8,16,37,38]

3.3. Prediction of Dipeptides with Potentially High Antioxidant Activity from the Models. For ABTS radical scavenging capacity prediction, dipeptides with a Y residue at the N-terminus showed higher activity (3.37 μmol TE/μmol peptide), followed by those with a W residue at the N-terminus (Supplementary Document 6). For ORAC, dipeptides with a W residue at the N-terminus showed higher activity (4.81 μmol TE/μmol peptide), and there was no obvious trend of preferred residue in the N- or C-terminus for the remaining high-activity dipeptides. The highlighted residues (W and Y) agreed with other studies, where they were reported to be strongly corresponding to high antioxidant activity of peptides.[7−9,19,31,32]

3.4. Application of Simulation Tool and Antioxidant Activity Prediction Models in Kafirin Proteins. The kafirin protein sequences were cleaved *in-silico* by R-PeptideCutter tool, and then the antioxidant activities of the generated dipeptides were predicted by the ABTS radical scavenging capacity and ORAC models, respectively. Among these enzymes or chemicals, three enzymes, namely chymotrypsin C, proteinase K, and thermolysin, generated more diverse dipeptides (more than 10 different dipeptides among all these kafirin proteins) (Table S2 in Supplementary Document 7). However, all these generated dipeptides were predicted to have low antioxidant activity in both ABTS and ORAC assays. The most diverse results were obtained from A9XEC1 (α-kafirin B3), where 18 different dipeptides were generated by thermolysin, and 17 and 14 different dipeptides were from the hydrolysis with chymotrypsin C and proteinase K, respectively (Table S2 in Supplementary Document 7). Besides, both pepsin (pH = 1.3) and pepsin (pH > 2) released six QQ dipeptides from the single α-kafirin sequence A9XEC1. The theoretically generated dipeptide varied among the enzymes due to the variation in enzyme cleavage sites and protein sequences.[16] For example, thermolysin can cleave protein sequences when the P1 position is not D or E and the P1′ position is A, F, I, L, M, or V. Both proteinase K (A, E, F, I, L, T, V, W, or Y at P1) and chymotrypsin C (F, M, Y, W, L, N, or Q at P1) have broad cleavage sites and therefore can generate various dipeptides. Thermolysin cannot generate dipeptides with a tyrosine, tryptophan, or cysteine at the N-terminus, which actually limits its application in antioxidant dipeptide production. Both chymotrypsin C and proteinase K could generate a dipeptide with tyrosine or tryptophan at the C-terminus, but these preferable residues at N-terminus depended on the sequence variation. Other enzymes, e.g., pepsin (pH = 1.3 or pH > 2), trypsin, chymotrypsin B1, and chymotrypsin (low or high specificity) also had broad cleavage sites, but the diversity was not comparable to that of the three enzymes which resulted from the sequences. Also, these enzymes might suffer from the broad cleavage sites, since this might lead to more residues released as free amino acids instead of peptides.

When two enzymes or chemicals are combined to generate dipeptides, there are 2601 different combinations from the 51 available enzyme or chemical simulations. Overall, there are many additional dipeptides generated from two enzymes/chemicals compared to the hydrolysis with a single enzyme or chemical (Table S2 in Supplementary Document 7). The best way to increase the diversity of the dipeptides was to combine one enzyme with broad cleavage sites and one with highly specific sites. The dipeptide diversity (12 dipeptides) generated by the combination of pepsin (pH = 1.3) and neutrophil elastase in Q9XE78 (α-kafirin) was at least twice the number of dipeptides from pepsin (pH = 1.3) (2 dipeptides) or neutrophil elastase (5 dipeptides) alone (Table S2 in Supplementary Document 7). This combination also released a highly potent antioxidant dipeptide, YR, where the ABTS radical scavenging activity and ORAC values were 3.352 and 2.099 μmol TE/μmol peptide, respectively. Furthermore, this dipeptide cannot be produced by a single-enzyme process.

It should be noted that there was no enzyme that could recognize a cysteine residue occurring at the P1 or P1′ position except for using chemicals such as NTCB, which can break the peptide bonds with cystine at the P1′ position. Therefore, it would be more challenging to produce cysteine-containing dipeptides from food proteins, and this only occurs occasionally. For example, papaya proteinase can break the peptide bonds with a glycine residue at the P1 position. Therefore, when there is a cysteine residue at the P1′ position, it is possible to generate cysteine-containing dipeptides. Among the selected kafirin sequences, G3FMW5 and Q6Q299 were observed to generate a dipeptide (CG) with papaya proteinase.

The R-PeptideCutter has the generality to be used in any protein sequences for peptide cutting simulation and can generate all possible peptides with any length (Supplementary Document 7). Overall, the diversity of the total possible dipeptides generated by two enzymes or chemicals was significantly higher, and some of the dipeptides (e.g., YR) were exclusively generated when combining two enzymes, which could not be achieved when using an individual enzyme or chemical.[16] The predicted antioxidant activity of the generated peptides hydrolyzed by different enzymes or chemicals could be used to guide wet chemistry for other studies on sorghum protein-derived antioxidant dipeptides.

3.5. Antioxidant Activity Model Validation Using Synthesized Dipeptides.

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

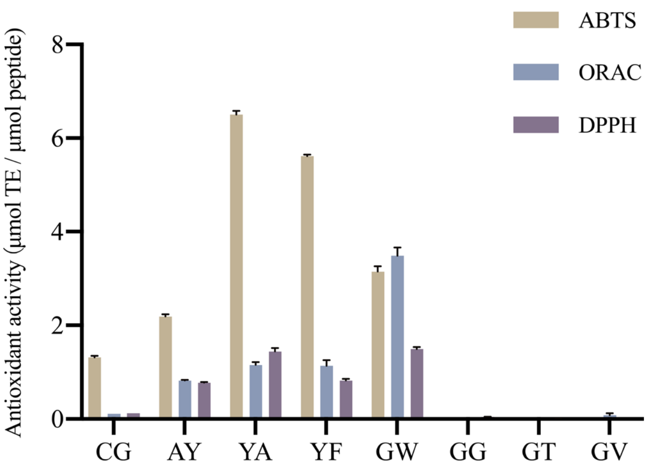

Figure 3. Antioxidant activities of different synthetic dipeptides determined by ABTS, ORAC, and reducing power assays. CG, AY, YA, YF, GW, GG, GT, and GV refer to the dipeptides Cys-Gly, Ala-Tyr, Tyr-Ala, Tyr-Phe, Gly-Trp, Gly-Gly, Gly-Thr, and Gly-Val, respectively.

The ABTS radical scavenging capacity of the synthetic dipeptides (CG, AY, YA, YF, GW) was underestimated, which was observed in the corresponding ABTS model development and was also consistent with an antioxidant tripeptide modeling study.[8] The ABTS radical scavenging capacity of YA was significantly higher than that of AY, even though they had the same amino acid composition and length. This is consistent with the feature importance results, which showed that the N-terminus residue was more important and Y had been proven to be more favorable at the N-terminus compared to A.[7,31] The ORAC model achieved a better activity prediction in YA, AY, and GW, where the observed/predicted activited were 1.11/1.16, 1.23/1.13, and 3.48/3.49 μmol TE/μmol peptide, respectively, which showed great prediction performance. In the ORAC prediction, the capacity of YA was only predicted to be slightly higher than that of AY, while GW exhibited approximately triple the activity of YA. The result implied that the amino acid residue Y was not as important as in the ABTS radical scavenging capacity model, while the W residue was more favorable to ORAC, even at the C-terminus.[7] However, dipeptides GG, GT, and GV barely showed any antioxidant activity, although they were predicted to have weak antioxidant capacity. The findings indicated that the models were not suitable for predicting the nonantioxidative peptides. Although the mechanism of DPPH radical scavenging capacity was based on both SET and HAT mechanisms, a high DPPH radical scavenging capacity was more in agreement with the HAT mechanism, where W played a more important role in antioxidant activity (Figure 3).[33]

In summary, we successfully developed prediction models for dipeptide antioxidant activity by an XGboost regression method plus other machine-learning methods with the latest data set of antioxidant dipeptides with ABTS radical scavenging capacities and ORAC values. The results showed that N-terminus residues played more important roles in the antioxidant activity of dipeptides. Specifically, a Y residue at the N-terminus and a W residue at the N-terminus strongly corresponded to high activity in ABTS radical scavenging capacity and ORAC, respectively. The model performance was significantly improved compared to the previous studies on peptide antioxidant activity prediction. A well-designed and user-friendly protein hydrolysis simulation tool, R-Pepttide-Cutter, was developed and released. Application of R-PeptideCutter and the models on sorghum protein revealed the dipeptide (YR) encrypted in the kafirin protein sequence (Q9XE78) had potentially the highest antioxidant activity, and the enzymes that could be used to release the dipeptide were also targeted. Eight dipeptides derived from the kafirin protein cutting simulation were synthesized and evaluated for their antioxidant activity, and they were used to validate the model performance. The corresponding ORAC model achieved greater prediction performance, while the corresponding ABTS radical scavenging capacity model underestimated the activity prediction, although both models exhibited inadaptability for dipeptides with low activity or nonactivity prediction. The developed models and R-PeptideCutter are capable of being applied to all proteins and identifying bioactive dipeptides released from natural proteins. In addition, the selected variables in model development offer alternative ways to elucidate the key features that determine bioactivity.

There are still some challenges from this study that require further research. Even though there are significant improvements for the newly developed hydrolysis simulation tool, a gap between *in-silico* simulation and experimental results still exists. Bridging the gap will be crucial for the rational design of protein hydrolysates for practical uses. More relevant and straightforward features are needed to encode peptides and amino acids for better QSAR model development and an understanding of the key factors contributing to the antioxidant activity. In addition, the current peptide feature encoding is limited by the peptide length: i.e., dipeptides only. Global descriptors, which characterize peptides as a whole during encoding and can be applied to peptides of any length, are a promising alternative approach to enlarge data sets and build more robust QSAR models.

■ ASSOCIATED CONTENT

Data Availability Statement

All the data and software used are clearly described in Materials and Methods.

SI Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsfoodscitech.2c00286.

Five hundred and fifty-three numerical indices of the 20 amino acids (XLSX)

ABTS radical scavenging capacity and ORAC data set (XLSX)

Original Pearson correlation coefficients of the 553 numerical indices (XLSX)

Heat map of the correlation coefficient (PDF)

Detailed Python codes for model development (ZIP)

Predicted antioxidant activity of the possible dipeptides (ABTS radical scavenging capacity and ORAC) (XLSX)

Kafirin sequences (fasta format files) and a brief summary, the enzymes or chemical information used, all possible kafirin-derived dipeptides with the enzyme information, cleavage sites of enzymes and chemicals, and R-PeptideCutter tool scripts (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

Yonghui Li − *Department of Grain Science and Industry, Kansas State University, Manhattan, Kansas 66506, United States;* orcid.org/0000-0003-4320-0806; Phone: +1 785-532-4061; Email: yonghui@ksu.edu

Author

Zhenjiao Du − *Department of Grain Science and Industry, Kansas State University, Manhattan, Kansas 66506, United States;* orcid.org/0000-0002-8492-4328

Complete contact information is available at: https://pubs.acs.org/10.1021/acsfoodscitech.2c00286

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

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

■ ABBREVIATIONS

ACE-I, angiotensin I converting enzyme inhibitor; DPP-IV, dipeptidyl peptidase IV; ABTS, 2,2′-azinobis(3-ethylbenzothiazoline-6-sulfonate); ORAC, oxygen radical absorbance capacity; XGboost, extreme gradient boosting; R^2, coefficient of determination; MSE, mean square error; LOOCV, leave-one-out cross-validation; DPPH, 1,1-diphenyl-2-picrylhydrazyl; FL, fluorescein disodium; AAPH: 2, 2′-azobis(2-methylpropionamide) dihydrochloride; SET, single electron transfer; HAT, hydrogen atom transfer; PCA, principal component analysis

■ REFERENCES

(1) Du, Z.; Li, Y. Review and Perspective on Bioactive Peptides: A Roadmap for Research, Development, and Future Opportunities. *Journal of Agriculture and Food Research* **2022**, *9*, 100353.

(2) Lorenzo, J. M.; Munekata, P. E. S.; Gómez, B.; Barba, F. J.; Mora, L.; Pérez-Santaescolástica, C.; Toldrá, F. Bioactive Peptides as Natural Antioxidants in Food Products - A Review. *Trends in Food Science & Technology* **2018**, *79*, 136−147.

(3) Phongthai, S.; D'Amico, S.; Schoenlechner, R.; Homthawornchoo, W.; Rawdkuen, S. Fractionation and Antioxidant Properties of Rice Bran Protein Hydrolysates Stimulated by in Vitro Gastrointestinal Digestion. *Food Chem.* **2018**, *240*, 156−164.

(4) Du, Z.; Wang, D.; Li, Y. Comprehensive Evaluation and Comparison of Machine Learning Methods in QSAR Modeling of Antioxidant Tripeptides. *ACS Omega* **2022**, *7*, 25760.

(5) Trinidad-Calderón, P. A.; Acosta-Cruz, E.; Rivero-Masante, M. N.; Díaz-Gómez, J. L.; García-Lara, S.; López-Castillo, L. M. Maize Bioactive Peptides: From Structure to Human Health. *Journal of Cereal Science* **2021**, *100*, 103232.

(6) Saito, K.; Jin, D.-H.; Ogawa, T.; Muramoto, K.; Hatakeyama, E.; Yasuhara, T.; Nokihara, K. Antioxidative Properties of Tripeptide

Libraries Prepared by the Combinatorial Chemistry. *J. Agric. Food Chem.* **2003**, *51* (12), 3668−3674.

(7) Zheng, L.; Zhao, Y.; Dong, H.; Su, G.; Zhao, M. Structure-Activity Relationship of Antioxidant Dipeptides: Dominant Role of Tyr, Trp, Cys and Met Residues. *Journal of Functional Foods* **2016**, *21*, 485−496.

(8) Chen, N.; Chen, J.; Yao, B.; Li, Z. QSAR Study on Antioxidant Tripeptides and the Antioxidant Activity of the Designed Tripeptides in Free Radical Systems. *Molecules* **2018**, *23* (6), 1407.

(9) Deng, B.; Long, H.; Tang, T.; Ni, X.; Chen, J.; Yang, G.; Zhang, F.; Cao, R.; Cao, D.; Zeng, M.; Yi, L. Quantitative Structure-Activity Relationship Study of Antioxidant Tripeptides Based on Model Population Analysis. *IJMS* **2019**, *20* (4), 995.

(10) Daroit, D. J.; Brandelli, A. In Vivo Bioactivities of Food Protein-Derived Peptides - a Current Review. *Current Opinion in Food Science* **2021**, *39*, 120−129.

(11) Li, S.; Bu, T.; Zheng, J.; Liu, L.; He, G.; Wu, J. Preparation, Bioavailability, and Mechanism of Emerging Activities of Ile-Pro-Pro and Val-Pro-Pro. *Comprehensive Reviews in Food Science and Food Safety* **2019**, *18* (4), 1097−1110.

(12) FitzGerald, R. J.; Cermeño, M.; Khalesi, M.; Kleekayai, T.; Amigo-Benavent, M. Application of in Silico Approaches for the Generation of Milk Protein-Derived Bioactive Peptides. *Journal of Functional Foods* **2020**, *64*, 103636.

(13) 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 Comb. Sci.* **2007**, *26* (6), 754−763.

(14) Wu, S.; Qi, W.; Su, R.; Li, T.; Lu, D.; He, Z. CoMFA and CoMSIA Analysis of ACE-Inhibitory, Antimicrobial and Bitter-Tasting Peptides. *Eur. J. Med. Chem.* **2014**, *84*, 100−106.

(15) Iwaniak, A.; Minkiewicz, P.; Pliszka, M.; Mogut, D.; Darewicz, M. Characteristics of Biopeptides Released In Silico from Collagens Using Quantitative Parameters. *Foods* **2020**, *9* (7), 965.

(16) Kalyan, G.; Junghare, V.; Khan, M. F.; Pal, S.; Bhattacharya, S.; Guha, S.; Majumder, K.; Chakrabarty, S.; Hazra, S. Anti-Hypertensive Peptide Predictor: A Machine Learning-Empowered Web Server for Prediction of Food-Derived Peptides with Potential Angiotensin-Converting Enzyme-I Inhibitory Activity. *J. Agric. Food Chem.* **2021**, *69* (49), 14995−15004.

(17) The UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49* (D1), D480−D489.

(18) Walker, J. M. *The Proteomics Protocols Handbook*; Springer: 2005.

(19) Tian, M.; Fang, B.; Jiang, L.; Guo, H.; Cui, J.; Ren, F. Structure-Activity Relationship of a Series of Antioxidant Tripeptides Derived from *β*-Lactoglobulin Using QSAR Modeling. *Dairy Sci. & Technol.* **2015**, *95* (4), 451−463 10/f7gdk9.

(20) Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. AAindex: Amino Acid Index Database, Progress Report 2008. *Nucleic Acids Res.* **2007**, *36*, D202−D205.

(21) Amigo, L.; Martínez-Maqueda, D.; Hernández-Ledesma, B. In Silico and In Vitro Analysis of Multifunctionality of Animal Food-Derived Peptides. *Foods* **2020**, *9* (8), 991.

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

(23) Hernández-Ledesma, B.; Amigo, L.; Recio, I.; Bartolomé, B. ACE-Inhibitory and Radical-Scavenging Activity of Peptides Derived from *β*-Lactoglobulin f(19−25). Interactions with Ascorbic Acid. *J. Agric. Food Chem.* **2007**, *55* (9), 3392−3397.

(24) Huang, W.-Y.; Majumder, K.; Wu, J. Oxygen Radical Absorbance Capacity of Peptides from Egg White Protein Ovotransferrin and Their Interaction with Phytochemicals. *Food Chem.* **2010**, *123* (3), 635−641.

(25) Je, J.-Y.; Cho, Y.-S.; Gong, M.; Udenigwe, C. C. Dipeptide Phe-Cys Derived from in Silico Thermolysin-Hydrolysed RuBisCO Large Subunit Suppresses Oxidative Stress in Cultured Human Hepatocytes. *Food Chem.* **2015**, *171*, 287−291.

(26) Suetsuna, K.; Ukeda, H.; Ochi, H. Isolation and Characterization of Free Radical Scavenging Activities Peptides Derived from Casein. *J. Nutr Biochem* **2000**, *11* (3), 128−131.

(27) Sabilla, S.; Sarno, R.; Triyana, K. Optimizing Threshold Using Pearson Correlation for Selecting Features of Electronic Nose Signals. *IJIES* **2019**, *12* (6), 81−90.

(28) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: 2016; pp 785−794.

(29) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-Learn: Machine Learning in Python. *the Journal of machine Learning research* **2011**, *12*, 2825−2830.

(30) Minkiewicz; Iwaniak; Darewicz. BIOPEP-UWM Database of Bioactive Peptides: Current Opportunities. *IJMS* **2019**, *20* (23), 5978.

(31) Udenigwe, C. C.; Aluko, R. E. Chemometric Analysis of the Amino Acid Requirements of Antioxidant Food Protein Hydrolysates. *IJMS* **2011**, *12* (5), 3148−3161.

(32) Zou, T.-B.; He, T.-P.; Li, H.-B.; Tang, H.-W.; Xia, E.-Q. The Structure-Activity Relationship of the Antioxidant Peptides from Natural Proteins. *Molecules* **2016**, *21* (1), 72.

(33) Liang, N.; Kitts, D. D. Antioxidant Property of Coffee Components: Assessment of Methods That Define Mechanisms of Action. *Molecules* **2014**, *19* (11), 19180−19208.

(34) Mahmoodi-Reihani, M.; Abbasitabar, F.; Zare-Shahabadi, V. In Silico Rational Design and Virtual Screening of Bioactive Peptides Based on QSAR Modeling. *ACS Omega* **2020**, *5* (11), 5951−5958.

(35) Tian, F.; Zhou, P.; Li, Z. T-Scale as a Novel Vector of Topological Descriptors for Amino Acids and Its Application in QSARs of Peptides. *J. Mol. Struct.* **2007**, *830* (1−3), 106−115.

(36) Zhou, P.; Liu, Q.; Wu, T.; Miao, Q.; Shang, S.; Wang, H.; Chen, Z.; Wang, S.; Wang, H. Systematic Comparison and Comprehensive Evaluation of 80 Amino Acid Descriptors in Peptide QSAR Modeling. *J. Chem. Inf. Model.* **2021**, *61* (4), 1718−1731.

(37) Du, Z.; Zeng, X.; Li, X.; Ding, X.; Cao, J.; Jiang, W. Recent Advances in Imaging Techniques for Bruise Detection in Fruits and Vegetables. *Trends in Food Science & Technology* **2020**, *99*, 133−141.

(38) Du, Z.; Tian, W.; Tilley, M.; Wang, D.; Zhang, G.; Li, Y. Quantitative Assessment of Wheat Quality Using Near-infrared Spectroscopy: A Comprehensive Review. *Comp Rev. Food Sci. Food Safe* **2022**, *21* (3), 2956−3009.