ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Comprehensive Evaluation and Comparison of Machine Learning Methods in QSAR Modeling of Antioxidant Tripeptides

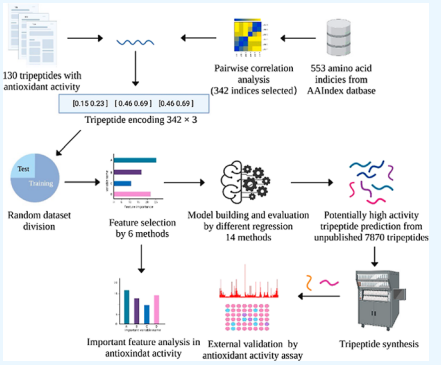
Zhenjiao Du, Donghai Wang, and Yonghui Li*

Cite This: *ACS Omega* 2022, 7, 25760−25771

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

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



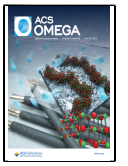
1. INTRODUCTION

Antioxidants are useful in reducing and preventing the harmful effect of *in vivo* free radicals by donating electrons to neutralize them, which induces cardiovascular diseases, cancers, and aging-related disorders.[1−3] Due to their multiple benefits, food protein-derived antioxidative peptides have gained increasing attention from today's consumers and researchers.[4−6] Various *in vitro* antioxidant assays have been developed to evaluate antioxidant capacity, which are approximately divided into two categories, that is, single-electron-transfer (SET) reaction and hydrogen-atom-transfer (HAT) reaction.[7] *In vitro* assays based on the SET reaction are generally preferred due to their convenience and accuracy.[8]

Conventional ways for screening peptides with high antioxidant activity are based on sequential and rigorous wet chemistry steps, such as enzymatic hydrolysis and/or microbial fermentation to release or produce peptides, *in vitro* antioxidant assays to determine antioxidant activity, and advanced chromatography and spectroscopy (e.g., high-performance liquid chromatography-mass spectrometry) to purify and identify potential peptides.[9] There are also studies that directly synthesized multiple peptides for screening on the basis of the theoretical knowledge (e.g., literature information of antioxidative peptides, important amino acids in peptides that contribute to antioxidant activity).[10−14] Up to now, some high-activity peptides have been found, such as Cys-Gln-Cys and Pro-His-His.[14,15] However, these conventional wet-chemistry methods for the preparation, fractionation, purification, and identification or synthesis of antioxidative peptides and for screening potentially high-activity peptides are time-consuming and highly rely on many advanced instruments and trained personnel.[2,5,9]

Quantitative structure−activity relationship (QSAR) is a computational modeling method for revealing relationships between chemical structures of molecules and their bioactivity.[16] In QSAR analysis, peptides are encoded by a series of numerical values, including properties of the amino acid residues (hydrophobicity, polarity, topological information,

Received: May 16, 2022
Accepted: June 30, 2022
Published: July 15, 2022



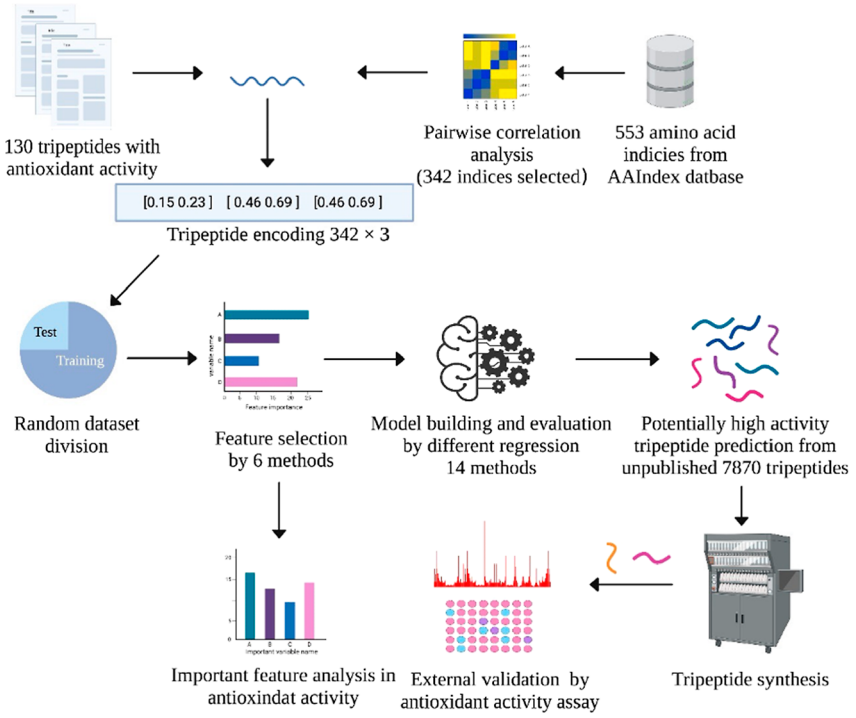


Figure 1. Flowchart for QSAR modeling and validation of antioxidant tripeptides.

etc.) comprising the peptides and properties of the entire peptides (electronegativity, sequence information, solubility, molecular weight, topological information, etc.).[17,18] Then, feature selection and modeling methods are combined to connect the structure information and bioactivity.[17,19] More than 80 amino acid descriptors (AADs) extracted from properties of amino acids by principal component analysis (PCA) were presented to characterize peptide structures and encode the peptides.[20−23] However, directly using these AADs usually led to undesirable model performance since most of them were not intended for the antioxidant activity modeling (e.g., T-scale for angiotensin-converting enzyme inhibitory activity).[8,11,19,22,24−30]

Machine learning methods have been successfully applied for feature selection and model development in QSAR studies on peptide bioactivity (e.g., angiotensin-converting enzyme inhibitory activity).[8,17,22,23,31−33] A total of 566 numerical values of amino acid including physicochemical properties and biochemical properties of amino acids and pairs of amino acids have been available in the AAIndex database.[18] This makes it possible to use feature selection to find the important variables for bioactivity prediction compared with using AADs from PCA where the principal components were composed of various original variables. In addition, increasing studies on antioxidant peptides allowed compilation of data sets on their structures and activities.[8,12,14,15] Most previous studies on antioxidant peptides still focused on the linear regression models, which would limit the model fitting to some extent due to the synergic effect among the residues in peptides.[8,11,24,25,34,35] Data set division was another issue in most studies where the samples were sorted in a descending or ascending order by their activity, and training and test data sets were evenly selected from the samples based on the sorted sequence (e.g., first three for the training data set and the following one as the test data set). The over-even data set division strategy would undermine the model robustness since the bias in the test data set could lead to poor model performance in cross validation compared with that in the test data set.[8,11,14,32]

Previous studies reported that tripeptides exhibited higher antioxidant activity and better bioavailability than other oligopeptides and have diverse structural variations.[14,15] The objective of this study was to compare popular machine learning methods for antioxidant activity modeling and screening of tripeptides and identify the critical amino acid features that determine the antioxidant activity (Figure 1). A total of 130 tripeptides with Trolox-equivalent antioxidant capacity (TEAC) values (SET reaction-based) were manually collected from published studies for QSAR model development. Further, 553 numerical indices were first screened by pairwise correlation, followed by comparative evaluation using 7 different feature selection strategies. Description of the important feature variables from 553 numerical indices was developed. A total of 14 different advanced regression methods including both linear and non-linear methods were first used to develop models based on the extracted important variables, and the best model was used to predict tripeptides with high antioxidant potential for future study. Model performance of the 14 regression methods was compared and discussed. Six tripeptides were synthesized and characterized for antioxidant activity to further evaluate the model performance for practical applications. Generalizability of these models was further tested by 20 times random data set splitting and introduction of leave-one-group-out cross validation. This study provides a useful approach to screen the key factors influencing the antioxidant activity of tripeptides and a guideline for future application of various machine learning methods in QSAR modeling.

2. MATERIALS AND METHODS

2.1. Data Set Collection. A total of 566 numerical indices of amino acids were collected using Beautiful Soup (4.5.3)

https://doi.org/10.1021/acsomega.2c03062
ACS Omega 2022, 7, 25760−25771

Table 1. Sequence and TEAC of Tripeptide (μmol TE/μMol Peptide) Data Set from the Literature.

no.	sequence	activity	no.	sequence	activity	no.	sequence	activity
1	LHA	0	47	PHN	0.24	93	PWR	0.822
2	LHD	0	48	LWF	0.25	94	PWI	0.832
3	LHE	0	49	PWD	0.262	95	RWG	0.842
4	LHF	0	50	LVG	0.266	96	LWN	0.866
5	LHG	0	51	PHH	0.266	97	LWR	0.869
6	LHH	0	52	PWE	0.339	98	PWL	0.88
7	LHQ	0	53	PHI	0.344	99	PWT	0.9
8	PHA	0	54	PHQ	0.348	100	PWN	0.943
9	PHD	0	55	GHG	0.365	101	RWH	0.995
10	PHE	0	56	LWD	0.402	102	RWQ	0.995
11	PHF	0	57	LWG	0.406	103	KHP	1.143
12	PHM	0	58	RHS	0.409	104	GVR	1.157
13	RHA	0	59	PWA	0.414	105	ECG	1.413
14	RHD	0	60	GHP	0.426	106	PHW	1.768
15	RHE	0	61	PWS	0.44	107	PWW	1.774
16	RHH	0	62	PWV	0.457	108	RWW	1.837
17	RHK	0	63	RWD	0.485	109	LHW	1.84
18	RHQ	0	64	LWM	0.49	110	LWW	1.931
19	RHT	0	65	PHG	0.496	111	WPL	1.972
20	PHT	0.028	66	RWA	0.497	112	VPW	1.972
21	LHM	0.031	67	PWM	0.498	113	RHW	2.203
22	LHN	0.046	68	LWV	0.499	114	LWY	2.332
23	GVT	0.047	69	RWV	0.51	115	RWY	2.334
24	PHS	0.058	70	LWL	0.515	116	RHY	2.464
25	KHR	0.067	71	LWQ	0.519	117	PHY	2.707
26	GHT	0.079	72	LWS	0.522	118	LHY	2.753
27	LWH	0.098	73	LWA	0.594	119	PWY	2.785
28	LHK	0.108	74	RWS	0.6	120	GVW	4.365
29	LHR	0.108	75	RHF	0.6	121	GKW	4.687
30	LHT	0.108	76	LWT	0.627	122	GHW	4.745
31	LHV	0.108	77	LWI	0.628	123	QVW	5.161
32	RHR	0.118	78	LWK	0.629	124	KVW	5.218
33	PHK	0.176	79	PWH	0.632	125	NKW	5.349
34	LHL	0.186	80	PWK	0.634	126	NHW	5.368
35	RHI	0.189	81	PWQ	0.637	127	QHW	5.524
36	PHV	0.198	82	RWR	0.651	128	KHW	5.566
37	PWF	0.202	83	RWT	0.651	129	PYW	5.683
38	PWG	0.203	84	RWE	0.663	130	YHW	6.169
39	RHG	0.203	85	LHS	0.68			
40	RHL	0.206	86	RWF	0.689			
41	RHM	0.207	87	RWL	0.689			
42	RHN	0.208	88	RWI	0.702			
43	PHR	0.211	89	RWM	0.702			
44	RHV	0.212	90	RWN	0.702			
45	LHI	0.217	91	RWK	0.753			
46	PHL	0.238	92	LWE	0.777			

from the AAIndex,[18,36] and detailed definition and description of each index are available online (https://www.genome.jp/aaindex/). The indices with missing values for amino acids were deleted, resulting in a total of 553 remaining indices (Table S1). Next, 130 antioxidant tripeptides were manually collected from BIOPEP-UWM,[37] and their activities analyzed by 2,2′-azino-bis(3-ethylbenzothiazoline-6-sulfonic acid) (ABTS) radical scavenging activity assay were obtained from the published literature and expressed as the TEAC values (μmol TE/μmol peptide) (Table 1).[8,12,38] The tripeptides with no antioxidant activity (i.e., 0 μmol TE/μmol peptide) were all retained in the data set for further QSAR model development.

2.2. Data Processing. *2.2.1. Pre-processing of Numerical Indices of Amino Acids.* The pairwise correlation method was used to pre-screen collinearity of the 553 numerical indices (Table S2 and Figure S1). If the absolute value of Pearson's correlation coefficient between 2 indices was greater than 0.95, one of them was removed randomly due to the strong correlation.[39] The remaining numerical indices were standardized for further feature selection.

2.2.2. Tripeptide Encoding and Feature Selection. The pre-screened numerical indices of amino acids were used to encode tripeptides. Briefly, if "n" numerical indices were selected after pre-processing, each amino acid was encoded as a "$1 \times n$" vector (tripeptide was encoded as a "$3 \times n$" matrix).

https://doi.org/10.1021/acsomega.2c03062
ACS Omega 2022, 7, 25760−25771

Table 2. Amino Acid Positions, Variable Importance, and Description of the Selected Variables from Different Feature Selection Strategies[a]

AAindex accession number	amino acid position	variable importance	description	note
selected variables by FI-XGB				
BURA740101	N-terminal	0.0199	normalized frequency of the alpha-helix	
CHOP780215	N-terminal	0.161	frequency of the 4th residue in turn	A
BEGF750102	central	0.036	conformational parameter of the beta-structure	
KANM800103	C-terminal	0.0138	average relative probability of the inner helix	
LIFS790103	C-terminal	0.7049	conformational preference for antiparallel beta-strands	B
selected variables by FI-RFR				
PALJ810113	N-terminal	0.025	normalized frequency of turn in the all-alpha class	
ONEK900102	N-terminal	0.0108	helix formation parameters (delta delta G)	
FUKS010101	N-terminal	0.015	surface composition of amino acids in intracellular proteins of thermophiles (percent)	
JOND750102	C-terminal	0.0171	pK (-COOH)	
LIFS790103	C-terminal	0.0518	conformational preference for antiparallel beta-strands	B
MCMT640101	C-terminal	0.0286	refractivity	
NAKH920102	C-terminal	0.0688	AA composition of CYT2 of single-spanning proteins	
OOBM850102	C-terminal	0.037	optimized propensity to form reverse turn	C
WEBA780101	C-terminal	0.0371	RF value in high-salt chromatography	D
VINM940102	C-terminal	0.051	normalized flexibility parameters (B-values) for each residue surrounded by none rigid neighbors	
PARS000101	C-terminal	0.0367	*p*-Values of mesophilic proteins based on the distributions of B values	N
PARS000102	C-terminal	0.0768	*p*-Values of thermophilic proteins based on the distributions of B values	K
FODM020101	C-terminal	0.0416	free energy change of epsilon(i) to alpha(Rh)	E
MITS020101	C-terminal	0.1532	amphiphilicity index	F
DIGM050101	C-terminal	0.0563	hydrostatic pressure asymmetry index, PAI	G
selected variables by FC-LR				
MAXF760103	N-terminal	0.025	normalized frequency of zeta R	
NAKH900102	N-terminal	0.0371	SD of AA composition of total proteins	
QIAN880114	N-terminal	0.051	weights for beta-sheet at the window position of -6	
KHAG800101	central	0.0367	the Kerr-constant increments	
CHOP780215	C-terminal	0.0768	frequency of the 4th residue in turn	A
OOBM850102	C-terminal	0.0416	optimized propensity to form reverse turn	C
WEBA780101	C-terminal	0.0153	RF value in high salt chromatography	D
MITS020101	C-terminal	0.0563	amphiphilicity index	F
selected variables by RFE-LR				
WERD780102	N-terminal	0.3136	free energy change of epsilon(i) to epsilon(ex)	
AURR980107	N-terminal	0.9357	normalized positional residue frequency at helix termini N2	
AURR980111	N-terminal	2.0325	normalized positional residue frequency at helix termini C5	H
AURR980116	N-terminal	1.5792	normalized positional residue frequency at helix termini Cc	
CEDJ970105	N-terminal	0.2991	composition of amino acids in nuclear proteins (percent)	I
KARS160120	N-terminal	0.5644	weighted minimum eigenvalue based on the atomic numbers	
CHOC760104	Central	0.8074	proportion of residues 100% buried	
GEIM800110	Central	0.6058	aperiodic indices for beta-proteins	J
QIAN880136	Central	0.9265	weights for coil at the window position of 3	
KARS160113	Central	0.7146	weighted domination number using the atomic number	
CHOP780215	C-terminal	0.8292	frequency of the 4th residue in turn	A
GEIM800110	C-terminal	0.0872	aperiodic indices for beta-proteins	J
HUTJ700101	C-terminal	0.271	heat capacity	
HUTJ700103	C-terminal	0.3975	entropy of formation	
KARP850102	C-terminal	0.6556	flexibility parameter for one rigid neighbor	
NAKH900110	C-terminal	1.1114	normalized composition of membrane proteins	
WILM950102	C-terminal	0.66043	hydrophobicity coefficient in RP-HPLC, C8 with 0.1%TFA/MeCN/H2O	
selected variables by RFE-SVR				
CHAM820102	N-terminal	0.303	free energy of solution in water, kcal/mole	
NAKH920101	N-terminal	0.4642	AA composition of CYT of single-spanning proteins	
RICJ880114	N-terminal	0.1628	relative preference value at C1	
PARS000102	N-terminal	0.2303	*p*-Values of thermophilic proteins based on the distributions of B values	K
CEDJ970105	N-terminal	0.1578	composition of amino acids in nuclear proteins (percent)	I
GEOR030105	N-terminal	0.0656	linker propensity from small data set (linker length is less than six residues)	L
GEIM800106	Central	0.2531	beta-strand indices for beta-proteins	
NAKH900108	Central	0.1859	normalized composition from fungi and plant	

https://doi.org/10.1021/acsomega.2c03062
ACS Omega 2022, 7, 25760−25771

Table 2. continued

AAindex accession number	amino acid position	variable importance	description	note
selected variables by RFE-SVR				
PALJ810116	Central	0.1345	normalized frequency of turn in alpha/beta class	M
GEOR030105	Central	0.1183	linker propensity from small data set (linker length is less than six residues)	L
CHOP780215	C-terminal	0.1984	frequency of the 4th residue in turn	A
OOBM850102	C-terminal	0.3586	optimized propensity to form reverse turn	C
PALJ810116	C-terminal	0.1983	normalized frequency of turn in alpha/beta class	M
WERD780104	C-terminal	0.2361	free energy change of epsilon(i) to alpha (Rh)	
PARS000101	C-terminal	0.3056	*p*-Values of mesophilic proteins based on the distributions of B values	N
MITS020101	C-terminal	0.0605	amphiphilicity index	F
DIGM050101	C-terminal	0.1109	hydrostatic pressure asymmetry index, PAI	G
selected variables by RFE-RFR				
CHOP780215	N-terminal	0.0336	frequency of the 4th residue in turn	A
ISOY800108	N-terminal	0.0294	normalized relative frequency of coil	
MAXF760104	N-terminal	0.0341	normalized frequency of left-handed alpha-helix	
GEOR030105	N-terminal	0.0486	linker propensity from small data set (linker length is less than six residues)	L
KARS160122	N-terminal	0.0362	weighted second smallest eigenvalue of the weighted Laplacian matrix	
QIAN880127	central	0.0362	weights for coil at the window position of -6	
AURR980111	central	0.0291	normalized positional residue frequency at helix termini C5	H
LIFS790103	C-terminal	0.1127	conformational preference for antiparallel beta-strands	B
MCMT640101	C-terminal	0.0969	refractivity	
OOBM850102	C-terminal	0.0462	optimized propensity to form reverse turn	C
WEBA780101	C-terminal	0.0245	normalized frequency of turn in all-alpha class	D
PARS000102	C-terminal	0.0745	*p*-Values of mesophilic proteins based on the distributions of B values	K
FODM020101	C-terminal	0.1246	free energy change of epsilon(i) to alpha(Rh)	E
MITS020101	C-terminal	0.2131	amphiphilicity index	F
DIGM050101	C-terminal	0.0603	hydrostatic pressure asymmetry index, PAI	G

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

The matrix was transformed into a "$1 \times 3n$" matrix, where 1 to the n elements in the vector belonged to the N-terminal residue, $n + 1$ to $2n$ elements were referred to the central amino acid, and $2n + 1$ to $3n$ elements belonged to the C-terminal residue.[8]

After encoding, each tripeptide is represented by $3n$ variables which were further screened by feature selection methods to identify the key variables for antioxidant activity prediction as amino acid descriptors. Six representative feature selection methods were evaluated, namely, linear regression-based recursive feature elimination (RFE), named RFE-LR;[40] support vector machine regression (SVR)-based RFE, named RFE-SVR;[40] random forest regression (RFR)-based RFE, named RFE-RFR;[40] feature coefficient (FC) based on lasso regression, named FC-LR;[40] feature importance based on RFR, named FI-RFR;[40] and feature importance based on extreme gradient boosting (XGB) regression, named FI-XGB.[41] The detailed mathematical methodologies of these selection methods are available in scikit-learn (https://scikit-learn.org/).

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

2.3. QSAR Model Development. *2.3.1. Data Set Division.* Totally, 130 samples were used for model development, cross validation, and model evaluation. The transformed *X*-matrix and *Y*-vector were shuffled and randomly split into the training data set and test data set at a ratio of 3:1. 98 samples were used for the training data set to build models. The remaining 32 samples were used as the test data set to evaluate the performance of the models. Leave-one-out cross-validation (LOOCV) was utilized for the validation data set split from THE training data set.

2.3.2. Regression Models. Fourteen popular regression models available through scikit-learn (https://scikit-learn.org/) and XGBoost (https://xgboost.readthedocs.io/en/stable/) were comparatively evaluated, namely, tree-based XGBoost regression (tree-XGB),[41] linear-based XGBoost regression (linear-XGB),[41] random forest regression (RFR), gradient boosting decision tree regression (GBDT),[42] decision tree-based bagging regression (Bagging),[43] multi-layer perceptron regression (MLP),[44] nearest neighbor regression (KNN),[40] radial basis function kernel-based support vector machine regression (rbf-SVR),[40] linear kernel-based support vector machine regression (linear-SVR),[40] linear regression with L1 regularization (Lasso),[45] linear regression with L2 regularization (Ridge),[40] linear regression by minimizing a regularized empirical loss with stochastic gradient descent (SGD),[40] ridge regression with kernel trick (KernelRidge),[46] and Huber regression (Huber).[47]

2.3.3. QSAR Model Building and Optimization. The model building was conducted using Python 3.8.8 with a computer (MacOS Monterey 12.0.1, CPU intel Core-i5 2.3 GHz). Models were imported from scikit-learn and XGBoost package.[40,41] LOOCV was used to avoid overfitting and tune the hyperparameters because of our small data set size.[8,14,33] The hyperparameters with the best performance from LOOCV

https://doi.org/10.1021/acsomega.2c03062
ACS Omega 2022, 7, 25760−25771

Table 3. Performance of 14 QSAR Models Based on the Different Feature Selection Strategies.[a]

model	training data set				test data set		note
	R^2_{Train}	$RMSE_{Train}$	R^2_{CV}	$RMSE_{CV}$	R^2_{Test}	$RMSE_{Test}$	
QSAR models based on FI-XGB							
tree-XGB	0.955	0.295	0.911	0.416	0.814	0.692	***
linear-XGB	0.566	0.921	0.478	1.01	0.558	1.067	
RFR	0.956	0.295	0.924	0.386	0.807	0.698	**
GBDT	0.976	0.219	0.904	0.434	0.78	0.752	*
bagging	0.974	0.226	0.904	0.434	0.769	0.77	
MLP	0.961	0.276	0.847	0.548	0.77	0.769	
KNN	0.84	0.559	0.598	0.887	0.555	1.069	
rbf-SVR	0.965	0.263	0.831	0.574	0.726	0.84	
linear-SVR	0.387	1.095	0.355	1.123	0.345	1.298	
Lasso	0.575	0.912	0.473	1.015	0.59	1.027	
Ridge	0.566	0.921	0.478	1.01	0.557	1.068	
SGD	0.516	0.973	0.424	1.062	0.49	1.146	
KernelRidge	0.074	1.346	−0.073	1.448	0.206	1.429	
Huber	0.567	0.92	0.478	1.01	0.559	1.064	
QSAR models based on FI-RFR							
tree-XGB	0.954	0.3	0.872	0.5	0.847	0.627	***
linear-XGB	0.789	0.643	0.722	0.738	0.681	0.906	
RFR	0.928	0.375	0.842	0.556	0.854	0.613	**
GBDT	0.978	0.207	0.866	0.512	0.781	0.75	
Bagging	0.962	0.274	0.833	0.571	0.822	0.677	
MLP	0.976	0.219	0.82	0.592	0.773	0.764	
KNN	0.933	0.362	0.832	0.573	0.814	0.691	
rbf-SVR	0.954	0.3	0.832	0.574	0.844	0.632	*
linear-SVR	0.78	0.655	0.709	0.755	0.623	0.984	
Lasso	0.796	0.632	0.714	0.748	0.685	0.901	
Ridge	0.779	0.657	0.721	0.739	0.679	0.909	
SGD	0.789	0.642	0.724	0.735	0.674	0.916	
KernelRidge	0.279	1.187	−0.118	1.479	0.295	1.346	
Huber	0.792	0.637	0.719	0.742	0.682	0.904	
QSAR models based on FC-LR							
tree-XGB	0.977	0.214	0.883	0.477	0.707	0.868	
linear-XGB	0.827	0.582	0.775	0.663	0.783	0.748	**
RFR	0.983	0.182	0.923	0.389	0.652	0.946	
GBDT	0.991	0.134	0.928	0.377	0.626	0.981	
Bagging	0.989	0.145	0.92	0.396	0.681	0.905	
MLP	0.975	0.221	0.771	0.669	0.763	0.781	
KNN	0.94	0.342	0.835	0.568	0.813	0.693	***
rbf-SVR	0.988	0.155	0.741	0.711	0.716	0.855	
linear-SVR	0.815	0.602	0.756	0.691	0.782	0.75	
Lasso	0.817	0.598	0.759	0.686	0.739	0.819	
Ridge	0.821	0.592	0.779	0.658	0.777	0.757	
SGD	0.826	0.584	0.771	0.669	0.785	0.743	
KernelRidge	0.319	1.155	0.034	1.375	0.37	1.272	
Huber	0.829	0.579	0.759	0.687	0.786	0.741	*
QSAR models based on RFE-LR							
tree-XGB	0.951	0.31	0.801	0.624	0.773	0.764	
linear-XGB	0.849	0.542	0.752	0.697	0.78	0.752	
RFR	0.939	0.345	0.793	0.636	0.737	0.823	
GBDT	0.986	0.164	0.821	0.592	0.8	0.718	*
Bagging	0.976	0.217	0.815	0.601	0.766	0.775	
MLP	0.979	0.202	0.868	0.509	0.824	0.672	***
KNN	0.859	0.526	0.749	0.701	0.627	0.98	
rbf-SVR	0.993	0.118	0.774	0.666	0.569	1.053	
linear-SVR	0.887	0.47	0.781	0.654	0.634	0.97	
Lasso	0.89	0.464	0.774	0.664	0.653	0.945	
Ridge	0.908	0.425	0.814	0.602	0.77	0.769	
SGD	0.787	0.645	0.684	0.786	0.768	0.773	
KernelRidge	0.24	1.219	0.004	1.395	0.328	1.315	

https://doi.org/10.1021/acsomega.2c03062
ACS Omega 2022, 7, 25760−25771

Table 3. continued

model	training data set				test data set		note
	R^2_{Train}	$RMSE_{Train}$	R^2_{CV}	$RMSE_{CV}$	R^2_{Test}	$RMSE_{Test}$	
QSAR models based on RFE-LR							
Huber	0.915	0.407	0.831	0.575	0.819	0.681	**
QSAR models based on RFE-SVR							
tree-XGB	0.945	0.329	0.893	0.457	0.772	0.766	
linear-XGB	0.844	0.553	0.756	0.691	0.759	0.787	
RFR	0.955	0.295	0.939	0.346	0.758	0.788	
GBDT	0.982	0.187	0.891	0.462	0.811	0.696	
Bagging	0.992	0.126	0.947	0.321	0.778	0.756	
MLP	0.979	0.202	0.882	0.48	0.846	0.628	***
KNN	0.924	0.386	0.897	0.449	0.839	0.643	*
rbf-SVR	0.996	0.095	0.835	0.568	0.666	0.927	
linear-SVR	0.922	0.39	0.829	0.579	0.809	0.701	
Lasso	0.844	0.552	0.756	0.691	0.759	0.787	
Ridge	0.859	0.525	0.806	0.616	0.83	0.662	
SGD	0.916	0.405	0.834	0.569	0.886	0.541	
KernelRidge	0.329	1.145	0.156	1.285	0.448	1.191	
Huber	0.926	0.381	0.84	0.559	0.84	0.642	**
QSAR models based on RFE-RFR							
tree-XGB	0.978	0.205	0.931	0.367	0.828	0.665	***
linear-XGB	0.852	0.539	0.786	0.647	0.704	0.872	
RFR	0.976	0.219	0.937	0.349	0.808	0.703	
GBDT	0.989	0.145	0.935	0.358	0.815	0.689	*
Bagging	0.992	0.122	0.939	0.345	0.799	0.719	
MLP	0.98	0.197	0.89	0.465	0.817	0.686	**
KNN	0.966	0.259	0.915	0.409	0.791	0.734	
rbf-SVR	0.996	0.089	0.924	0.385	0.801	0.716	
linear-SVR	0.852	0.539	0.761	0.684	0.699	0.88	
Lasso	0.861	0.521	0.778	0.66	0.706	0.869	
Ridge	0.867	0.51	0.783	0.651	0.702	0.875	
SGD	0.863	0.518	0.787	0.647	0.703	0.874	
KernelRidge	0.382	1.099	0.105	1.324	0.144	1.484	
Huber	0.86	0.523	0.788	0.643	0.706	0.869	
QSAR models without feature selection							
tree-XGB	0.987	0.161	0.860	0.523	0.705	0.87	
linear-XGB	0.927	0.378	0.786	0.647	0.746	0.807	*
RFR	0.946	0.324	0.892	0.459	0.749	0.803	***
GBDT	0.992	0.126	0.898	0.447	0.744	0.811	**
Bagging	0.929	0.374	0.419	1.066	0.404	1.238	
MLP	0.773	0.666	0.621	0.861	0.619	0.99	
KNN	0.996	0.089	0.752	0.697	0.628	0.978	
rbf-SVR	0.926	0.381	0.709	0.754	0.734	0.827	
linear-SVR	0.893	0.457	0.765	0.678	0.731	0.831	
Lasso	0.936	0.355	0.758	0.688	0.744	0.811	
Ridge	0.463	1.025	0.160	1.281	0.263	1.377	
SGD	0.411	1.073	−0.081	1.454	0.073	1.544	
KernelRidge	0.936	0.355	0.757	0.69	0.743	0.812	
Huber	0.981	0.195	0.858	0.526	0.743	0.812	

[a]Note: Detailed description of these models are available at https://scikit-learn.org/stable/and https://xgboost.readthedocs.io/en/stable/. (*)The models with more stars in the last column indicate better performance from the same feature selection method.

were used as the final model for performance evaluation with the test data set.

2.3.4. Model Performance Evaluation. Determination of the coefficient (R^2) and root mean square error (RMSE) was used to evaluate the model performance. R^2 and RMSE from the training data set, LOOCV, and test data set were named as R^2_{Train} and $RMSE_{Train}$, R^2_{CV} and $RMSE_{CV}$, and R^2_{Test} and $RMSE_{Test}$, respectively. To further evaluate the model generalizability, the developed models with the tuned hyper-parameters were rebuild by 20 times with different random data set splitting and evaluated by using R^2 and RMSE from the training data set, LOOCV, Leave-one-group-out cross validation (LOGOCV), and test data set. The result of the extra evaluation is available in Table S3.

2.4. Prediction of Unpublished Tripeptides with Antioxidant Activity from the Models. A data set containing 7870 potential tripeptides was built, and the published 130 tripeptides used for model building and

https://doi.org/10.1021/acsomega.2c03062
ACS Omega 2022, 7, 25760−25771

validation were not included. After obtaining the model with the best performance, the 7870 tripeptides were encoded by the selected features and used for the antioxidant activity prediction based on the selected model.

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

2.6. Characterization of Antioxidant Activity of the Synthesized Tripeptides. The ABTS radical scavenging capacity assay was based on the method described in the studies of Phongthai et al. and Chen et al.[8,48] with a few modifications. Briefly, stock solution was prepared by mixing 7.4 mM ABTS and 2.6 mM potassium persulfate in deionized (DI) water and incubating at room temperature for 12 h. The working solution was made by diluting the stock solution till the absorbance of the mixture of ABTS$\bullet^+$ solution and DI water at 734 nm was at 0.70 ± 0.02. Then, 150 μL ABTS$\bullet^+$ solution was mixed with 50 μL tripeptide solution (20 μM) and allowed for 30 min incubation at 30 °C, and subsequently, the absorbance was measured at 734 nm using the Biotek Synergy H1 Hybrid Microplate Reader (Winooski, VT, USA). Trolox (TE) was used as a standard antioxidant, and results were expressed as μmol TE/μmol peptide. All the chemicals and reagents used were of analytical grade and purchased from Sigma-Aldrich (St. Louis, MO, USA).

3. RESULTS

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

3.2. Model Development Based on Variables Selected by FI-RFR. Fifteen variables were selected by FI-RFR with a threshold (feature importance = 0.01) (Table 2) and then used to encode the 130 tripeptides as the *X*-matrix (130 × 15). Based on the variable importance, C-terminal residues also contributed the most to the antioxidant activity (*Y*-vector), while there was little contribution from the central amino acids based on these selected variables.

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

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

3.4. Model Development Based on Variables Selected by RFE-LR. Recursive feature elimination (RFE) eliminates one variable with the least feature importance or feature coefficient in one iteration, and the procedure is recursively repeated on the pruned data set until achieving the desired number of features. Seventeen variables were selected from RFE-LR (Table 2) and then used to encode the 130 tripeptides as the *X*-matrix (130 × 17). Based on the variable importance, N-terminal residues contributed the most to the antioxidant activity (*Y*-vector), while the central amino acids contributed the least to the activity. Model performances of the 14 different regression methods are shown in Table 3. MLP gained the best performance in the test data set (R^2_{Test} = 0.824 and $RMSE_{test}$ = 0.672), followed by Huber (R^2_{Test} = 0.819 and $RMSE_{test}$ = 0.681). GBDT also provided a good result with R^2_{Test} larger than 0.8. From RFE-LR, linear-XGB, Ridge, and SGD as linear methods gained competitive performance compared with the non-linear regression methods like KNN, rbf-SVR, and RFR.

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

3.6. Model Development Based on Variables Selected by RFE-RFR. Fifteen variables were selected by RFE-SVR (Table 2) and then used to encode the 130 tripeptides as the *X*-matrix (130 × 15). Based on the variable importance, C-terminal residues contributed the most to the antioxidant activity (*Y*-vector), while the central amino acids contributed

https://doi.org/10.1021/acsomega.2c03062
ACS Omega 2022, 7, 25760−25771

the least to the activity. Model performances of the 14 different regression methods are shown in Table 3. Tree-XGB achieved the best performance where R^2_{Test} and $RMSE_{test}$ were 0.828 and 0.665, respectively. For the remaining models, non-linear regression methods, even the worst one, KNN outperformed the linear regression methods.

3.7. Model Development without Feature Selection. A total of 1026 variables were used to encode the 130 tripeptides as the *X*-matrix (130 × 1026). Model performances of the 14 different regression methods are shown in Table 3. RFR gained the best model performance where R^2_{Test} and $RMSE_{test}$ were 0.749 and 0.803, respectively. A significant overfitting was observed in the MLP model where the R^2_{Train} was 0.929, but the R^2_{Test} was only 0.404.

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

3.9. Application of the QSAR Model in Synthetic Tripeptide Activity Prediction. The experimental antioxidant activity of the synthetic tripeptides is summarized with their corresponding predicted activity in Table 4. QAY was predicted to be the most powerful antioxidant peptides (6.167 μmol TE/μmol peptide), while its observed activity was ranked second (4.270 μmol TE/μmol peptide) among the six synthesized tripeptides. PHC was also predicted to exhibit strong antioxidant activity (6.023 μmol TE/μmol peptide), and its observed activity (5.013 μmol TE/μmol peptides) was even stronger than that of QAY. Overall, the QSAR model has been very useful for the selection of potentially high-antioxidant activity tripeptides, although the antioxidant activity from the model was a little bit overestimated compared to the experimental results.

Table 4. Antioxidant Activity of Synthesized Tripeptides.

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

4. DISCUSSION

Various numerical indices were screened and selected by the six different feature selection methods. Based on the variable importance values, almost all the feature selection methods showed that the C-terminal residues played the most important role in antioxidant activity, while the central amino acid contributed the least to the activity, which was partly consistent with previous results from wet-chemistry and QSAR studies, where there was no comparison between N- and C- terminals.[12,24,31] Previous studies were confined to amino acid physicochemical properties (with about 195 indices) or the AADs which could not take full advantage of all the amino acid indices to identify the most representative indices to characterize tripeptides.[8,11,32] Although some of the selected features, especially non-physicochemical properties (e.g., LIFS790103 stands for "Conformational preference for antiparallel beta-strands"), might be difficult to understand and explain, these selected features are much targeted and less redundant.[8] For these AADs derived from PCA analysis, each principal component was composed of multiple original properties, and there are usually several principal components adopted in the model development, which can only be roughly explained (e.g., the first component was related to hydrophobicity) but impeded the further explanation of the feature importance and distracted the application of these features for peptide design and modification.[22,27] Even though some features are difficult to explain here, they all have the standard protocols to be determined, and this would be easier when applying in the structure design and modification of bioactive peptides.

Among these selected features, some of them, such as CHOP780215, LIFS790103, OOBM850102, and WEBA780101, were selected multiple times for the characterization of C-terminal residues by different feature selection strategies, which showed their importance in antioxidant activity prediction. CHOP780215 was not only selected for encoding C-terminal residues by FC-LR, RFE-LR, and RFE-SVR but also selected by RFE-RFR and FI-XGB to characterize N-terminal residues. Some features, such as GEOR030105 and PALJ810116 selected by REF-SVR and GEIM800110 selected by REF-LR, were used to encode both the central and N-terminal residues, and central and C-terminal residues, respectively. This implies that some features of amino acids can contribute to antioxidant activity at any position, even though their importance varied with positions. The theoretical conclusion derived from the selected features was also supported by the study of Uno et al.[31]

For the models without feature selection, inferior performance was observed, as shown in Table 3. The main reason of the poor performance under this preprocessing method was mainly because of the high dimensionality on the features and small sample size. Therefore, the significant improvement in model performance was achieved by feature selection because plenty of irrelevant features were eliminated.[17]

For the 14 different regression methods, non-linear regression methods overall achieved better model performance based on the 6 feature selection methods, which proved the existence of non-linearity in antioxidant activity prediction. This also explained the poor model performance in most previous studies which were based on linear regression methods.[11,31,32] In addition, some studies subjectively removed the non-active tripeptides from the data set in order to improve the model fitting, which resulted in misleading models.[8,32] Further, improper data set division between the training data set and test data set increased the bias in the model and undermined the robustness of the models,[8] while model evaluation without the test data set was not complete because the performance in cross-validation could not represent the real performance of the model in the unknown data set.[32] In this study, these biases were all overcome, and the performance was greatly improved compared with the most recent study on

https://doi.org/10.1021/acsomega.2c03062
ACS Omega 2022, 7, 25760−25771

tripeptides.[8,32] In fact, we also adopted processing methods using the bias-existing data from the literature to develop these models during the preliminary study, and R^2_{test} could be larger than 0.9559, which also proved the bias in the previous studies. Abnormal phenomena were observed in some models (e.g., rbf-SVR regression based on FI-RFR) where performance in LOOCV was poorer than that in the test data set. This implied the overfitting of these models, and the same situation was difficult to avoid in bioactivity prediction since LOOCV was an optimistic cross validation method.[8] The main reason that the n-fold cross validation was not used in our study was primarily due to the relatively small data set. In order to further evaluate the generalizability of these models, we introduced the more challenging cross validation (LOGOCV) and 20 times of random data set splitting for the model evaluation (Table S3). Performance of these overfitting strategies suffered more in generalizability evaluation. It also can be seen that the XGBoost regression method with random forest regression for feature selection was the most powerful and robust strategy for antioxidant activity prediction.

From the prediction of tripeptides with potentially high antioxidant activity, the 525 unpublished tripeptides with activity higher than 6 μmol TE/μmol peptides all had a tyrosine or tryptophan or cysteine residue at the C-terminal position, which was consistent with previous studies.[12,24,31] Compared with previous studies, our model clearly specifies the tripeptides with the most promising antioxidant activity.

The preferable attributes of strong antioxidant tripeptides concluded from the model development were supported by the antioxidant activity determination from the synthetic tripeptides. It was observed that tripeptides with tyrosine and cysteine residues at the C-terminal exhibited the highest antioxidant activity compared to those with the tyrosine residue at the N-terminal, which also showed lower contribution to antioxidant activity in the feature importance analysis. In addition, the model successfully predicted that the tripeptide (PHC) with a cysteine residue at the C-terminal had strong antioxidant activity (Table 1), which had not been reported previously. In addition, there was no tripeptide with tyrosine at the C-terminal, showing high antioxidant activity (e.g., above 4 μmol TE/μmol peptide). Our results supported the hypothesis of the model development that these amino acid indices had the capacity to represent the residues in tripeptide for unknown antioxidant tripeptide activity prediction. The deviation between the observed and predicted activity was inevitable, but it is overall acceptable.[8,31]

5. CONCLUSIONS

In this study, we collected 553 latest amino acid numerical indices and 130 published tripeptides with available TEAC values for QSAR analysis. Seven feature selection strategies and 14 regression methods were combined to build QSAR models and used to comprehensively evaluate the performance of the application of machine learning methods in predicting antioxidant tripeptides. The results showed that C-terminal residues played a more important role in antioxidant activity, and non-linear regression methods were more suitable for the QSAR study on antioxidant activity. The best model based on FI-RFR for feature selection plus tree-based XGB for model building was used to predict the antioxidant activities of the unknown 7870 tripeptides, and the high-activity tripeptides have the tyrosine, tryptophan, or cysteine residue at the C-terminal position. Furthermore, 6 unpublished tripeptides were synthesized and characterized to evaluate the practical application of the best model. The predicted activity can reflect the rank of the potential activity of these tripeptides and their approximate activity, although there was an overestimation. This study also, for the first time, demonstrates through both the *in silico* and wet-chemistry experiment that cysteine and tyrosine residues at the C-terminal are highly corresponding to antioxidant activity for tripeptides. In addition, this study also provides critical reference for antioxidant tripeptide screening and a useful model development template for future QSAR studies on bioactive peptides.

DATA AND SOFTWARE AVAILABILITY

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

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c03062.

533 numerical indices of the 20 amino acids; original Pearson correlation coefficients of the 553 numerical indices and the heatmap of the correlation coefficient; performance of 14 models in 20 times random data set splitting with leave-one-out cross validation and leave-one-group-our cross validation; and predicted antioxidant activity of the unknown 7790 tripeptides (XLSX)

Original python 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; Email: yonghui@ksu.edu

Authors

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

Donghai Wang − *Department of Biological and Agricultural Engineering, Kansas State University, Manhattan, Kansas 66506, United States;* orcid.org/0000-0001-9293-1387

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c03062

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

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

REFERENCES

(1) Lobo, V.; Patil, A.; Phatak, A.; Chandra, N. Free Radicals, Antioxidants and Functional Foods: Impact on Human Health. *Pharmacogn. Rev.* **2010**, *4*, 118−126.

(2) Phongthai, S.; Rawdkuen, S. Fractionation and Characterization of Antioxidant Peptides from Rice Bran Protein Hydrolysates

Stimulated by in Vitro Gastrointestinal Digestion. *Cereal Chem.* **2020**, *97*, 316−325.
(3) Zhuang, H.; Tang, N.; Yuan, Y. Purification and Identification of Antioxidant Peptides from Corn Gluten Meal. *J. Funct. Foods* **2013**, *5*, 1810−1821.
(4) Baakdah, M. M.; Tsopmo, A. Identification of Peptides, Metal Binding and Lipid Peroxidation Activities of HPLC Fractions of Hydrolyzed Oat Bran Proteins. *J. Food Sci. Technol.* **2016**, *53*, 3593−3601.
(5) Girgih, A. T.; Udenigwe, C. C.; Hasan, F. M.; Gill, T. A.; Aluko, R. E. Antioxidant Properties of Salmon (Salmo Salar) Protein Hydrolysate and Peptide Fractions Isolated by Reverse-Phase HPLC. *Food Res. Int.* **2013**, *52*, 315−322.
(6) Nimalaratne, C.; Bandara, N.; Wu, J. Purification and Characterization of Antioxidant Peptides from Enzymatically Hydrolyzed Chicken Egg White. *Food Chem.* **2015**, *188*, 467−472.
(7) Zheng, Y.-Z.; Deng, G.; Liang, Q.; Chen, D.-F.; Guo, R.; Lai, R.-C. Antioxidant Activity of Quercetin and Its Glucosides from Propolis: A Theoretical Study. *Sci. Rep.* **2017**, *7*, 7543.
(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*, 1407.
(9) Ulug, S. K.; Jahandideh, F.; Wu, J. Novel Technologies for the Production of Bioactive Peptides. *Trends Food Sci. Technol.* **2021**, *108*, 27−39.
(10) Gu, L.; Zhao, M.; Li, W.; You, L.; Wang, J.; Wang, H.; Ren, J. Chemical and Cellular Antioxidant Activity of Two Novel Peptides Designed Based on Glutathione Structure. *Food Chem. Toxicol.* **2012**, *50*, 4085−4091.
(11) 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. *J. Funct. Foods* **2016**, *21*, 485−496.
(12) 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*, 3668−3674.
(13) Ulug, S. K.; Jahandideh, F.; Wu, J. Novel Technologies for the Production of Bioactive Peptides. *Trends Food Sci. Technol.* **2021**, *108*, 27−39.
(14) 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*, 451−463.
(15) Chen, H.-M.; Muramoto, K.; Yamauchi, F.; Nokihara, K. Antioxidant Activity of Designed Peptides Based on the Antioxidative Peptide Isolated from Digests of a Soybean Protein. *J. Agric. Food Chem.* **1996**, *44*, 2619−2623.
(16) Pripp, A.; Ardo, Y. Modelling Relationship between Angiotensin-(I)-Converting Enzyme Inhibition and the Bitter Taste of Peptides. *Food Chem.* **2007**, *102*, 880−888.
(17) Dai, Z.; Wang, L.; Chen, Y.; Wang, H.; Bai, L.; Yuan, Z. A Pipeline for Improved QSAR Analysis of Peptides: Physiochemical Property Parameter Selection via BMSF, near-Neighbor Sample Selection via Semivariogram, and Weighted SVR Regression and Prediction. *Amino Acids* **2014**, *46*, 1105−1119.
(18) Kawashima, S.; Kanehisa, M. AAindex: Amino Acid Index Database. *Nucleic Acids Res.* **2020**, *28*, 374.
(19) 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*, 1718−1731.
(20) Yousefinejad, S.; Hemmateenejad, B.; Mehdipour, A. R. New Autocorrelation QTMS-Based Descriptors for Use in QSAM of Peptides. *J. Iran. Chem. Soc.* **2012**, *9*, 569−577.
(21) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *J. Med. Chem.* **1998**, *41*, 2481−2491.
(22) 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*, 106.
(23) Tian, F.; Zhou, P.; Lv, F.; Song, R.; Li, Z. Three-Dimensional Holograph Vector of Atomic Interaction Field (3D-HoVAIF): A Novel Rotation-Translation Invariant 3D Structure Descriptor and Its Applications to Peptides. *J. Pept. Sci.* **2007**, *13*, 549−566.
(24) Li, Y.-W.; Li, B.; He, J.; Qian, P. Structure-Activity Relationship Study of Antioxidative Peptides by QSAR Modeling: The Amino Acid next to *C* -Terminus Affects the Activity: QSAR Study of Antioxidative Peptides. *J. Pept. Sci.* **2011**, *17*, 454−462.
(25) Udenigwe, C. C.; Aluko, R. E. Chemometric Analysis of the Amino Acid Requirements of Antioxidant Food Protein Hydrolysates. *IJMS* **2011**, *12*, 3148−3161.
(26) Zaliani, A.; Gancia, E. MS-WHIM Scores for Amino Acids: A New 3D-Description for Peptide QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 525−533.
(27) Yang, L.; Shu, M.; Ma, K.; Mei, H.; Jiang, Y.; Li, Z. ST-Scale as a Novel Amino Acid Descriptor and Its Application in QSAM of Peptides and Analogues. *Amino Acids* **2010**, *38*, 805−816.
(28) Shu, M.; Mei, H.; Yang, S.; Liao, L.; Li, Z. Structural Parameter Characterization and Bioactivity Simulation Based on Peptide Sequence. *QSAR Comb. Sci.* **2009**, *28*, 27−35.
(29) Mahmoodi-Reihani, M.; Abbasitabar, F.; Zare-Shahabadi, V. Silico Rational Design and Virtual Screening of Bioactive Peptides Based on QSAR Modeling. *ACS Omega* **2020**, *5*, 5951−5958.
(30) Cocchi, M.; Johansson, E. Amino Acids Characterization by GRID and Multivariate Data Analysis. *Quant. Struct.-Act. Relat.* **1993**, *12*, 1−8.
(31) Uno, S.; Kodama, D.; Yukawa, H.; Shidara, H.; Akamatsu, M. Quantitative Analysis of the Relationship between Structure and Antioxidant Activity of Tripeptides. *J. Pept. Sci.* **2020**, *26*, No. e3238.
(32) 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*, 995.
(33) 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*, 14995−15004.
(34) Li, Y.-W.; Li, B. Characterization of Structure−Antioxidant Activity Relationship of Peptides in Free Radical Systems Using QSAR Models: Key Sequence Positions and Their Amino Acid Properties. *J. Theor. Biol.* **2013**, *318*, 29−43.
(35) Li, Y.-W.; Li, B.; He, J.; Qian, P. Quantitative Structure−Activity Relationship Study of Antioxidative Peptide by Using Different Sets of Amino Acids Descriptors. *J. Mol. Struct.* **2011**, *998*, 53−61.
(36) Richardson, L. *Beautiful Soup Documentation*, 2007.
(37) Minkiewicz; Iwaniak; Darewicz. BIOPEP-UWM Database of Bioactive Peptides: Current Opportunities. *Int. J. Mol. Sci.* **2019**, *20*, 5978.
(38) Ma, Y.; Xiong, Y. L.; Zhai, J.; Zhu, H.; Dziubla, T. Fractionation and Evaluation of Radical Scavenging Peptides from in Vitro Digests of Buckwheat Protein. *Food Chem.* **2010**, *118*, 582−588.
(39) Sabilla, S.; Sarno, R.; Sarno, R.; Triyana, K.; Institut Teknologi Sepuluh Nopember; Universitas Gadjah Mada Sekip Utara. Optimizing Threshold Using Pearson Correlation for Selecting Features of Electronic Nose Signals. *Int. j. eng. innov. technol* **2019**, *12*, 81−90.
(40) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(41) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System *Proceedings of the 22nd ACM SIGKDD International Conference*

https://doi.org/10.1021/acsomega.2c03062
ACS Omega 2022, 7, 25760−25771

on Knowledge Discovery and Data Mining; ACM: San Francisco California USA, 2016; pp 785−794.
(42) Friedman, J. H. Stochastic Gradient Boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367−378.
(43) Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123−140.
(44) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Representations by Back-Propagating Errors. *Nature* **1986**, *323*, 533−536.
(45) Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Software* **2010**, *33*, 1.
(46) Murphy, K. P. *Machine Learning: A Probabilistic Perspective*; MIT press, 2012.
(47) Huber, P. J. Robust Regression: Asymptotics, Conjectures and Monte Carlo. *Ann. Stat.* **1973**, *1*, 799−821.
(48) 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.

https://doi.org/10.1021/acsomega.2c03062
ACS Omega 2022, 7, 25760−25771