Machine Learning for the Cleaner Production of Antioxidant Peptides

Jose Isagani B. Janairo

Biology Department, De La Salle University, 2401 Taft Avenue, Manila 0922, Philippines

Abstract

Antioxidant peptides (AP) are promising functional foods that have the potential to provide multitude health benefits. They are found in a wide variety of sources, but current methods of discovery and extraction dramatically increases the cost of production which hampers the commercial competitiveness of APs. Focusing on the search and development of short AP sequences that can be easily synthesized through synthetic chemical methods may be able to decrease the cost of production and accelerate lead discovery. However, the traditional method of peptide synthesis that relies on solid-phase chemistry adversely impacts the environment. Thus, minimizing trial-and-error will not only shorten AP discovery but can also make the entire process greener and more cost-effective. In this study, the formulation of a machine learning model that can predict the trolox equivalent antioxidant capacity (TEAC) of tripeptides is presented. It was found that the combination of support vector regression with a polynomial kernel and Blosum indices can accurately predict AP TEAC. The optimized regression model was trained, tested, and externally validated on 121 sequences curated from three different publications. The optimized model demonstrates a 7% average percent error based on external validation.

Keywords: QSAR; artificial intelligence; biomolecules; lead discovery; support vector regression

1.0 Introduction

Antioxidant peptides (AP) are promising bioactive components of food that can provide both health and nutritional benefits. Apart from being sources of amino acids that constitute an important part of the human diet, APs can also quench reactive oxygen species (ROS) that are implicated in numerous diseases (Yang et al. 2020). The realization of the full potential of APs is hampered by the difficulty in discovering and extracting antioxidant peptides. APs are found in a broad range of sources, such as plants (Wen et al. 2020), fungi (Nascimento et al. 2021), and animal tissues (Aubry et al. 2020). However, APs require varying and innovative techniques for their efficient extraction and characterization, such as the use of metal-organic frameworks (Chen et al. 2021), omics-tools (López-Pedrouso et al. 2021), ultrasound-microwave-assisted enzymatic methods (Habinshuti et al. 2020), among others. This can increase the cost of production making it a barrier to the wider utilization of APs (Tadesse and Emire 2020). A possible way to simultaneously decrease the cost of production and accelerate lead discovery is to focus on the search and development of short peptide sequences that can be quickly and easily synthesized at a large scale through synthetic chemical methods.

Peptides are usually produced through solid phase peptide synthesis (SPPS), which is also used for the industrial production of bioactive peptides (Verlander 2007). This method involves anchoring a protected amino acid on to a solid support which will be then elongated through a series of deprotection and coupling reactions (Merrifield 1963). Once the desired sequence has been achieved, the elongated peptide chain will be cleaved from the solid matrix for purification and characterization. SPPS can be automated which makes the synthesis of short to medium-length peptides convenient, quick, and straightforward. However, SPPS can have negative environmental consequences. This mode of peptide production requires greater molar equivalence of the reactants to the drive the reaction, as well as its dependence on toxic solvents and reagents (Coin et al. 2007). For example, deprotection of an anchored Fmoc-protected amino acid requires treatment with piperidine in DMF. After the deprotection reaction, the resin needs to be washed thoroughly with DMF several times to ensure that all piperidine has been removed. Thus, SPPS presents an effective yet environmentally impactful process of producing peptides. Efforts have been therefore made to make SPPS more sustainable and environment friendly. These efforts include using greener solvents (Lawrenson 2018), developing a more sustainable deprotection protocol

(Přibylka et al. 2020), among others. One approach in decreasing the environmental impact of chemical processes such as in antioxidant peptide production is minimizing the trial-and-error of the process through computational models (Zhang et al. 2020). Leveraging machine learning (ML) and artificial intelligence (AI) tools for chemical product design can quickly identify promising leads to be developed, thereby reducing the compounds to be synthesized and tested, which can make the entire process more resource-efficient and environment-friendly. In this paper, the creation, training, testing, and external validation of support vector regression models that can predict the antioxidant activity of tripeptides from sequence-based descriptors is presented.

2.0 Method

The data on antioxidant peptides composed of 109 peptide sequences and their corresponding log of the trolox equivalent antioxidant capacity (TEAC) acting as the dependent variable were taken from (Yan et al. 2020) (Uno et al. 2020). The different sequence-based descriptors were then calculated for each antioxidant peptide using the Peptides R package version 2.4 (Osorio et al. 2015). The calculated peptide descriptors were the Blosum indices (Georgiev 2009), Cruciani properties (Cruciani et al. 2004), Factor analysis scale of generalized amino acid information (FASGAI) vectors (Liang and Li 2007), Kidera factors (Kidera et al. 1985), ProtFP (van Westen et al. 2013), ST-scales (Yang et al. 2010), T-scales (Tian et al. 2007), VHSE Scales (Mei et al. 2005), and Z-scales (Sjöström et al. 2002). These peptide descriptors can be generally classified according to what aspect of the peptide they represent. The Blosum indices are under the similarity measures category; the T-scales and ST-scales are topological descriptors; the FASGAI vectors, ProtFP, VHSE scales, and Z-scales describe the physicochemical properties of the peptide (Rifaioglu et al. 2019). These descriptors were then used as variables to predict the antioxidant activity of the peptides. The resulting dataset was used to create support vector regression (SVR) models using different kernels such as polynomial, linear and radial. For all created SVR algorithms, 70% of the dataset was used for training, followed by a 10-fold cross-validation mainly using the caret package (Kuhn et al.

2018) and other dependencies such as the kernlab package (Karatzoglou et al. 2004). The default parameters that yielded the highest r^2 were automatically selected. The selected model was further optimized by systematically removing descriptors in order to balance performance and parsimony of the model. The final and optimized model was then externally validated using data from (Saito et al. 2003). The top two antioxidant peptides for each peptide category reported in the paper that did not appear in the training and testing dataset were used for the external validation. All computations were conducted in R version 3.5.2 (R Core Team 2018) using a Dell Inspiron 15 gaming laptop running on a 64 bit Windows 10 OS, with Intel Core 7th generation 2.80 GHz i7 processor , 16 GB of RAM. The full dataset and relevant R scripts used in this study are available in the supporting information and at www.github.com/jijanairo/AntioxidantPeptides.

3.0 Results and Discussion

The first step in predictive model building involves identification of the peptide descriptor and algorithm pair that can accurately predict AP TEAC. Apart from SVR, other common ML algorithms were also evaluated such as artificial neural networks, multiple linear regression, and random forest. However, the performance of these models was extremely poor which is why only the results for SVR are presented. Figure 1 shows the performance of each peptide descriptor – algorithm pair wherein it was observed that z-scales and SVR with a polynomial kernel exhibited the highest r^2 in the training set. However, this model exhibited overfitting since its performance in the test set was $r^2<0.50$. Overfitting should be avoided because this leads to the creation of a predictive model where the parameters are too tailored to the training data, which leads to poor performance in the testing data and is detrimental to the generalizability of the model. Thus, the SVR-polynomial-Blosum Indices (BI) pair was selected for model refinement. BI is a set of peptide descriptors based on the physicochemical properties of amino acids wherein the calculated ten BI exhibit correlation with a particular property (Georgiev 2009). SVR is a machine learning regression algorithm that builds on the concept of support vector networks formulated by Cortes and Vapnik (Cortes et al. 1995) that relies on the creation of hyperplanes through the kernel functions for data processing (Drucker· et al. 1997).

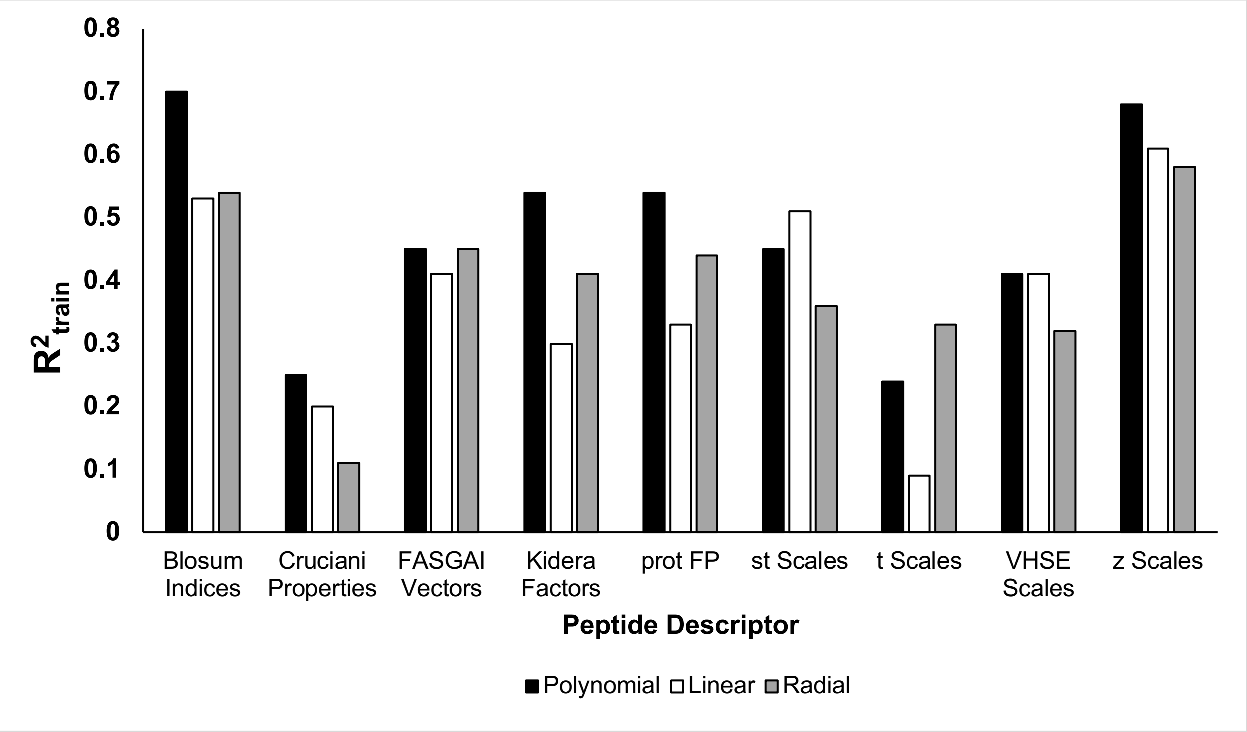

Figure 1. Training performance denoted by r^2 of the peptide descriptor and support vector regression kernel type for the AP TEAC prediction.

After identifying that SVR-polynomial and BI create the best pair for TEAC prediction, the resulting model was refined in order to improve predictive performance and parsimony. Thus, one BI is systematically removed from the model after which the predictive performance in the training and testing sets were monitored. Table 1 summarizes the undertaken model optimization steps which shows that the individual removal of B3, B4 and B10 had the slightest effect on testing performance. These observations provided the rationale to simultaneously remove these three BI, wherein the resulting model (model K) exhibited the best predictive performance. The optimized model, model K, exhibited a r^2_{test} = 0.66 based on the regression values of the observed and predicted values (Figure 2), and a calculated $r^2_{test(1)}$ = 0.64 based on the equation:

$$r^2_{test(1)} = 1 - \frac{\sum(y_{obs} - y_{pred})^2}{\sum(y_{obs} - y_{mean})^2}$$

Where y_{obs} is the observed response variable, y_{pred} is the predicted variable, and y_{mean} is the mean of observed response variables. Since $r^2_{test(1)} > 0.60$, model K can be considered as a valid regression model that exhibits good fit (Alexander et al. 2015).

Table 1. Summary of model optimization wherein a single Blosum Index is systematically removed followed by monitoring the predictive performance of the resulting model.

Model	Removed Descriptor	Training r^2	Testing r^2
A	B1	0.70	0.49
B	B2	<0.40	n.d.
C	B3	0.66	0.57
D	B4	0.61	0.58
E	B5	<0.40	n.d.
F	B6	<0.40	n.d.
G	B7	0.60	0.53
H	B8	0.65	0.52
I	B9	0.69	0.54
J	B10	0.62	0.60
K	B3, B4, B10	0.60	0.66

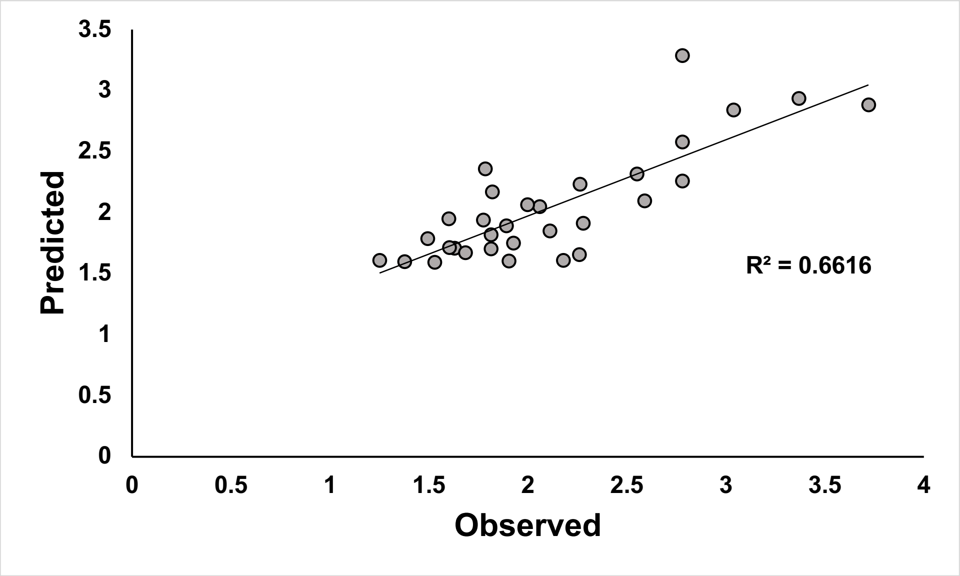

Figure 2. Predicted vs observed log TEAC derived from the optimized model.

The hyperparameters of the optimized SVR model are degree = 3, scale = 0.1, C = 0.25. The algorithm run time for the optimized model for training and testing was determined to be 12.75 seconds, and used approximately 260 megabytes of memory. On the other hand, the optimized descriptors are related to specific properties such as buriability (B1), number of side chain atoms (B2), pKa of the N terminus (B5), extended structure (B6), hydropathy scale (B7), negative charge (B8), and amino acid distribution in the alpha helix (B9) (Georgiev 2009). The identified important descriptors for TEAC prediction are consistent with reports from mechanistic studies that sought to determine factors governing peptide antioxidant activity, such as the importance of electronic properties, structure, flexibility (Wu et al. 2021), and the N-terminus (Yan et al. 2020; Yang et al. 2020). The optimized regression model was further tested through external validation based on 12 tripeptides reported from the study of (Saito et al. 2003). The optimized model performed well as demonstrated by an average error of 7%. A limitation of the presented model that needs to be considered is the non-sequence dependence of the Blosum indices. For instance, the tripeptide Lys-Ser-Val will have identical Blosum indices with any tripeptide that bears these residues in any order. Nonetheless, the optimized regression model still lends itself useful for AP screening and discovery since it can substantially narrow down potential leads to be synthesized and tested.

Table 2. Performance of the optimized model on the external validation.

Sequence	Reported	Predicted	% Error
LHW	2.3	2.0	13.0
LHY	2.4	2.0	16.7
LWN	1.9	1.9	0.0
LWY	2.3	2.1	8.7
PHW	2.2	2.4	9.1
PHY	2.4	2.6	8.3
PWW	2.2	2.2	0.0
PWN	1.9	1.9	0.0
RHW	2.3	2.1	8.7
RHY	2.4	2.2	8.3
RWW	2.3	2.2	4.3
RWY	2.4	2.2	8.3
Average % Error			7.1 ± 5.2

The presented SVR model is a valuable addition to the growing ML and AI tools that are tailored for predicting AP activity. Available ML and AI tools for AP include a deep learning server that predicts the free radical scavenging and chelation scores (Olsen et al. 2020), a classification algorithm based on ensemble learning that uses hybrid peptide predictors (Zhang et al. 2016), a 3D-QSAR model suggesting small and hydrophilic group at the N-terminal region and a bulky and hydrophobic group at the C terminus are important for antioxidant activity (Yan et al. 2020), and multiple linear regression model suggesting the importance of cysteine residues, aromatic residue at the C-terminus, narrow HOMO-LUMO bandgap at the middle residues are important for antioxidant activity (Uno et al. 2020). Thus, the present model has discovered new descriptors that are important in predicting AP activity, which are straightforward to calculate due their sequence-dependent nature. Some of the strengths of the presented model are its real-world applicability since the variable being predicted is a standard measure of antioxidant activity, generalizability since the model was trained, tested and validated on 121 sequences extracted from multiple publications, and robustness as exemplified by the performance in the external validation. The balance between accuracy and parsimony as evidenced by the need for only seven sequence-based descriptors to make a prediction is also one of the main advantages of the presented model not only in terms of simplicity but also in the associated environmental footprint of machine learning since more complex models tend to have higher environmental impacts (Strubell et al. 2019). Overall, the SVR model can potentially accelerate the discovery of three-residue APs which may lower down the costs associated with AP production thereby increasing its market competitiveness and lessen its environmental impact.

4.0 Conclusion

The systematic creation and optimization of a machine learning model that can predict the trolox equivalent antioxidant capacity of tripeptides was reported. It was found that the combination of support vector regression with a polynomial kernel and Blosum indices can accurately predict AP TEAC. The optimized SVR model, which was trained, tested, and externally validated on 121 tripeptide sequences reported an average of 7% error in the external validation. Some of the strengths of the presented model are its real-world applicability, the balance between accuracy and parsimony, generalizability, and robustness. By and large, the presented model has promising potential to accelerate the marketability of antioxidant peptides without compromising the environment and sustainability. It is envisioned for future studies that the presented optimized model may be used for the screening and design of potent AP.

Data Availability

All data generated or analysed during this study are included in this published article [and its supplementary information files].

Declaration of Competing Interest

None.

Funding Information

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

5.0 References

Alexander DLJ, Tropsha A, Winkler DA (2015) Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. J Chem Inf Model 55:1316–1322. https://doi.org/10.1021/acs.jcim.5b00206

Aubry L, De-Oliveira-Ferreira C, Santé-Lhoutellier V, Ferraro V (2020) Redox Potential and Antioxidant Capacity of Bovine Bone Collagen Peptides towards Stable Free Radicals, and Bovine Meat Lipids and Proteins. Effect of Animal Age, Bone Anatomy and Proteases-A Step Forward towards Collagen-Rich Tissue Valorisation. Molecules 25:5422–5422. https://doi.org/10.3390/molecules25225422

Chen ML, Ning P, Jiao Y, et al (2021) Extraction of antioxidant peptides from rice dreg protein hydrolysate via an angling method. Food Chem 337:128069–128069. https://doi.org/10.1016/j.foodchem.2020.128069

Coin I, Beyermann M, Bienert M (2007) Solid-phase peptide synthesis: From standard procedures to the synthesis of difficult sequences. Nat Protoc 2:3247–3256. https://doi.org/10.1038/nprot.2007.454

Cortes C, Vapnik V, Saitta L (1995) Support-Vector Networks. Mach Leaming 20:273–297

Cruciani G, Baroni M, Carosati E, et al (2004) Peptide studies by means of principal properties of amino acids derived from MIF descriptors. J Chemom 18:146–155. https://doi.org/10.1002/cem.856

Drucker· H, Burges CJC, Kaufman L, et al (1997) Support Vector Regression Machines. In: Advances in Neural Information Processing Systems. pp 155–161

Georgiev AG (2009) Interpretable Numerical Descriptors of Amino Acid Space. J Comput Biol 16:703–723. https://doi.org/10.1089/cmb.2008.0173

Habinshuti I, Mu TH, Zhang M (2020) Ultrasound microwave-assisted enzymatic production and characterisation of antioxidant peptides from sweet potato protein. Ultrason Sonochem 69:105262–105262. https://doi.org/10.1016/j.ultsonch.2020.105262

Karatzoglou A, Smola A, Hornik K, Zeileis A (2004) kernlab-An S4 Package for Kernel Methods in R. JSS J Stat Softw 11:1–20

Kidera A, Konish Y, Oka M, et al (1985) Statistical Analysis of the Physical Properties of the 20 Naturally Occurring Amino Acids. J Protein Chem 4:23–55. https://doi.org/10.1007/BF01025492

Kuhn M, Wing J, Weston S, et al (2018) caret: Classification and Regression Training

Lawrenson SB (2018) Greener solvents for solid-phase organic synthesis. Pure Appl Chem 90:157–165. https://doi.org/10.1515/pac-2017-0505

Liang G, Li Z (2007) 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 26:754–763. https://doi.org/10.1002/qsar.200630145

López-Pedrouso M, Borrajo P, Amarowicz R, et al (2021) Peptidomic analysis of antioxidant peptides from porcine liver hydrolysates using SWATH-MS. J Proteomics 232:104037–104037. https://doi.org/10.1016/j.jprot.2020.104037

Mei H, Liao ZH, Zhou Y, Li SZ (2005) A new set of amino acid descriptors and its application in peptide QSARs. Biopolym - Pept Sci Sect 80:775–786. https://doi.org/10.1002/bip.20296

Merrifield RB (1963) Synthesis of a Tetrapeptide. J Am Chem Soc 85:2149–2154

Nascimento TCES, Molino JVD, Donado PRS, et al (2021) Antarctic fungus proteases generate bioactive peptides from caseinate. Food Res Int 139:109944–109944. https://doi.org/10.1016/j.foodres.2020.109944

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

Osorio D, Rondon-Villarreal P, Torres R (2015) Peptides: A Package for Data Mining of Antimicrobial Peptides. R J 7:4–14

Přibylka A, Krchňák V, Schütznerová E (2020) Environmentally Friendly SPPS II: Scope of Green Fmoc Removal Protocol Using NaOH and Its Application for Synthesis of Commercial Drug Triptorelin. J Org Chem 85:8798–8811. https://doi.org/10.1021/acs.joc.0c00599

R Core Team (2018) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Rifaioglu AS, Atas H, Martin MJ, et al (2019) Recent applications of deep learning and machine intelligence on in silico drug discovery: Methods, tools and databases. Brief. Bioinform. 20:1878–1912

Saito K, Jin DH, Ogawa T, et al (2003) Antioxidative properties of tripeptide libraries prepared by the combinatorial chemistry. J Agric Food Chem 51:3668–3674. https://doi.org/10.1021/jf021191n

Sjöström M, Sandberg M, Wold S, et al (2002) New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. J Med Chem 41:2481–2491. https://doi.org/10.1021/jm9700575

Strubell E, Ganesh A, McCallum A (2019) Energy and Policy Considerations for Deep Learning in NLP. arXiv. arXiv.1906.02243

Tadesse SA, Emire SA (2020) Production and processing of antioxidant bioactive peptides: A driving force for the functional food market. Heliyon 6

Tian F, Zhou P, Li Z (2007) T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides. J Mol Struct 830:106–115. https://doi.org/10.1016/j.molstruc.2006.07.004

Uno S, Kodama D, Yukawa H, et al (2020) Quantitative analysis of the relationship between structure and antioxidant activity of tripeptides. J Pept Sci 26:. https://doi.org/10.1002/psc.3238

van Westen GJ, Bender A, Swier RF, et al (2013) Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets. J Cheminform. https://doi.org/10.1186/1758-2946-5-41

Verlander M (2007) Industrial applications of solid-phase peptide synthesis - A status report. Int J Pept Res Ther 13:75–82. https://doi.org/10.1007/s10989-006-9075-7

Wen C, Zhang J, Zhang H, et al (2020) Plant protein-derived antioxidant peptides: Isolation, identification, mechanism of action and application in food systems: A review. Trends Food Sci. Technol. 105:308–322

Wu RB, Huang JF, Huan R, et al (2021) New insights into the structure-activity relationships of antioxidative peptide PMRGGGGYHY. Food Chem 337:. https://doi.org/10.1016/j.foodchem.2020.127678

Yan W, Lin G, Zhang R, et al (2020) Studies on the Bioactivities and Molecular Mechanism of Antioxidant Peptides by 3D-QSAR, in vitro Evaluation and MD Simulations. Food Funct 11:3043–3052. https://doi.org/https://doi.org/10.1039/C9FO03018B

Yang L, Shu M, Ma K, et al (2010) ST-scale as a novel amino acid descriptor and its application in QSAM of peptides and analogues. Amino Acids 38:805–816. https://doi.org/10.1007/s00726-009-0287-y

Yang Q, Cai X, Yan A, et al (2020) A specific antioxidant peptide: Its properties in controlling oxidation and possible action mechanism. Food Chem 327:126984–126984. https://doi.org/10.1016/j.foodchem.2020.126984

Zhang L, Mao H, Liu Q, Gani R (2020) Chemical product design – recent advances and perspectives. Curr Opin Chem Eng 27:22–34. https://doi.org/10.1016/j.coche.2019.10.005

Zhang L, Zhang C, Gao R, et al (2016) Sequence based prediction of antioxidant proteins using a classifier selection strategy. PLoS One 11:. https://doi.org/10.1371/journal.pone.0163274