题目（暂定）：

TastePeptides-Meta: A taste peptide system including a di/tri-peptide taste judgment model Umami\_YYDS, a taste peptide database TastePeptidesDB, and an open source automatic machine learning package Auto\_Taste\_ML

KeyWords:peptides,umami,database,open source,web serve

# Abstract

Taste is one of the important basic attributes of food, which profoundly affects many processes such as production, circulation, and trade. Peptides, as one of the basic components of food, have taste diversity including umami and bitterness. It is necessary to clarify the law of taste peptides presenting the sense of taste. In this study, we use the Django framework to build the first taste peptide database(TastePeptidesDB, published at http:/ /tastepeptides-meta.com:7777/database/son/1) base on reported taste peptide information. 2-3 peptides were selected to generate QSAR descriptors(Quantitative Structure–Activity Relationship), , the number of descriptors is 278. After screening in statistics and modeling, 8 features including BCUT2D\_MWLOW, PEOE\_VSA14, SMR\_VSA1, MinEStateIndex (The Electrotopological State), VSA\_EState5, VSA\_EState6, VSA\_EState7, MolLogP, etc. were obtained as the key feature descriptors. A gradient boosting decision tree model(Umami\_YYDS,) was established by data enhancement, comparison algorithm and model optimization. The model achieved 89.6% accuracy and 98% AUC(Area Under ROC Curve). In order to evaluate the superiority of the model, Umami\_YYDS model was compared with other reported models by a series of indicators including accuracy, recall, precision, F1, Matthews correlation coefficient (MCC). Furthermore, taste peptides with 4-10 amino acids were tested by the model to illustrate the robustness of the model, which proved that Umami\_YYDS had a leading performance. Nowadays, Umami\_YYDS has been published as a prediction web site at <http://tastepeptides-meta.com:7777/cal> for the convenience of researchers. The above binary-model construction process is independently packaged and released as the Auto\_Taste\_ML machine learning package, which is published at https://pypi.org/project/Auto-ML-C/, aiming to help researchers in the field of taste peptide research. In short, the research established the taste peptide system TastePeptides-Meta, including a taste peptide database TastePeptidesDB, a umami/bitter taste prediction model Umami\_YYDS, and an open source machine learning package Auto\_Taste\_ML in the field of taste peptide research.

图形用户界面, 应用程序

描述已自动生成

# 1.Introduction

Food taste is a key attribute of food, which profoundly affects many aspects of food processing, production, trade, and nutrition. Umami and bitterness are one of the five basic tastes. Umami taste is closely related to the pleasant mood, while bitterness represents a perception that is not conducive to eating[1]. Studies have shown that umami peptides inhibit bitter taste perception through bitter taste receptors[2], hence the two flavors are often compared to each other in research. In recent years, increasingly researches on taste peptides, and they have been reported to be found in beef[3], peanuts[4], fermented products[5] and other foods. However, the identification of flavor peptides in the traditional sense is a complex task including the pretreatment, separation, purification, synthesis and characterization, sensory evaluation of flavor substances[6].

With the improvement of computer performance and chemoinformatics, the quantitative structure-activity model (QSAR) has successively【是不是这个词 successfully？？？】 shown great potential in the prediction of material properties[7], successively in the activity judgment of biological peptides[8], ADMET(the property of absorption, distribution, metabolism, excretion, toxicity) [9] and molecular multi-dimensional description[10]. However, in terms of the analysis of the taste regulations, previous studies have been affected by factors such as insufficient amount of flavor peptide data and simple models (linear regression[11], Scoring Card Method[12], support vector machine and ridge regression [13]), etc. As a result, the accuracy and the generalization performance of those models are not ideal. what’s more, those models can only achieve a single taste judgment, such as iBitter-SCM[14] and iUmami-SCM[15]. On the other hand, some researches are pursuing performance and abandon the exploration of the model, and use "black box" algorithms based on random forest [16], neural network[17], XGBoost [18], etc. to predict taste, such as BERT4Bitter [19], etc. This model is unfavorable for the exploration of the taste regluation of peptides. Therefore, it is necessary to build a QSAR model with excellent performance and model interpretability.

So far, most of the model research is still in the method release stage, only include introduction and description of the model construction method. These studies have failed to package the code, methods, or functions of the model to a certain extent for release. Fortunately, there are still a few studies on database construction (such as BIOPEP[20], AroCageDB[7], Toxindb[21]), web prediction services (such as VirtualTaste[16]), software (e-Bitter[22], CBDPS 1.0 [18]) . These three aspects have played an exemplary role. For related model names, algorithms and judgments, please refer to Table 1 of Supplementary Materials.

Due to the huge manpower, material resources and time costs, and the lack of umami peptides mechanism, and there is an urgent need for an effective and professional flavour peptide information summary platform. In this study, we compiled the reported taste peptide information and released it in the form of a database TastePeptidesDB; used the umami bitter peptide data to construct a umami-bitterness judgment model, and encapsulated the modeling process as an automaton to learn package release(see Figure 0 for details). As far as the author knows, there is currently no published research like the TastePeptides-Meta in this study, which builds a systematic taste peptides research universe, consisting of TastePeptidesDB, umami/bitter taste prediction model Umami\_YYDS, and automatic taste model construction package Auto\_Taste\_ML. (see Figure 1 for details). Among them, TastePeptidesDB is today the largest taste peptide database with the most information, and Auto\_Taste\_ML is the first open source machine learning package in the field of taste.

图形用户界面, 应用程序

描述已自动生成

**图1**

**Figure 1: Research framework diagram. Including the main ideas, technical routes and module status of the research**

# 2.Materials and Methods

## 2.1 Benchmark Data Sets

A total of 203 reported umami/bitter peptides wre collected to support model construction. Among them, 99 peptides, of which 31 are umami and 68 are bitter, 104 peptide tripeptides, of which 53 are umami and 61 are bitter. Considering the inhibitory effect of umami substances on bitterness, we labeled umami peptides as positive and bitter peptides as negative[2,23]. 【最后放一个raw 数据地址，github的】

In this study, we also used ‘Tastes’, ‘Sour’, ‘Sweet’, ‘Bitter’, ‘Salty’, ‘Umami’, ‘Kokumi’, ‘Astringent’, and ‘Peptides’ as keywords to search in Web of Science. Finally, we got 483 peptides and shown them on TastepeptidesDB, where is http://tastepeptides-meta.com:7777/database/son/1.

## 2.2 Feature structure

Features selectionwas divided into 4 steps (as shown in the funnel diagram in Figure 2a):

**Step 1**: For each peptide, 208 molecular descriptors were calculated through the chemometrics special toolkit RDKit 2020.9.1[24]. Those descriptors well described the peptide’s water solubility, electrostatic properties, atomic properties[25], etc. This study also added 69 descriptors including the planar properties, cyclic properties[26], aromatic properties[27], electrostatic properties[28] and first and last amino acid properties (Such as the presence or absence of C-terminal hydrophobic amino acids[14]). A total of 278 descriptors features of peptides is were shown in Figure S1

**Step 2**: Use the variance check algorithm of scikit-learn 0.24.2 [29] to discard the features with a variance of 0, leaving 207 features.

**Step 3**: The Kolmogorov-Smirnov test and t test of scipy [30] were used to perform feature screening from a statistical point of view, and 51 features with pvalue=0.0001 significance that Umami and Bitter do not obey the same distribution and the data mean are not equal, see Figure S2.

**Step 4**: Using recursive feature elimination with cross-validation to select the number of features based on Random Forest Model, 51 features are used as features in turn, and the features with the best effect in each iteration are retained. As shown in Figure 2b, when the number of features is 8, the model basically reached the convergence zone of the increase, and it was confirmed that the final 8 features were retained as the final features.

Taking into account the imbalance of the model data, we choose the imblearn0.8.1[29] package to over sampling the umami peptide data, and use the KMeans-SMOTE[31], SMOTE(Synthetic Minority Oversampling Technique)[32], SVM-SMOTE(Support Vector Machine-Synthetic Minority Oversampling Technique)[33] algorithms for data enhancement, which is different from the original The balance data is compared by Random Forest Model(n\_estimators = 1000, max\_depth = 9). The SMOTE algorithm has the best effect, with excellent performance in the four indicators of accuracy and recall, although the precision performance is slightly worse (Figure 2c). The 8 feature values of the enhanced data are scaled to 0-10 for visualization as shown in Figure 2d. People can visually distinguish umami and bitter peptides with the naked eye. In order to observe the changes before and after SMOTE data enhancement, this study reduced the 8-dimensional data to 2-dimensional data through PCA(Principal Component Analysis )to visualize the data as shown in Figure 2e. It can be found that there is a clear distinction between the data, which proves that the SMOTE data enhancement has generalization performance.

For the purpose of independent verification and testing, the training set and the validation set were constructed separately according to the ratio of 4:1 by stratified sampling. In order to better detect the generalization performance of the model, 4-10 peptides constitute a generalization test set, referred to as GTS. All umami and bitter peptides were constructed called ATPD(all taste peptides dataset).

This research is completed in Python3.8.10 , the numerical calculation and transmission adopt Pandas 1.3.3[34] and Numpy 1.2.0[35], and the drawing adopts Matplotlib 3.4.2[36], Seaborn 0.11.2[37] and plotly. express 0.4.1.

## 2.3 Model Selection and Optimization

In order to select the best algorithm to better mine the internal laws of the data, this study selects 19 popular and widely recognized binary classification algorithms for model construction[38], including: LogisticRegression(LR)[31], RidgeClassifier, Perceptron, Stochastic Gradient Descent(SGD), Linear Discriminant Analysis, LinearSVC (LD), Support Vector Classification (SVC)[39], Nu-Support Vector Classification (NuSVC) [39], DecisionTree/ExtraTree(DT)[40], AdaBoost(ABT)[41], Bagging(BT)[42], GradientBoosting(GTBT)[43], RandomForest(RF)[44], XGBoost(XGBT)[44], KNeighbors(KN), NearestNeighbors, Naive Bayes classifier for multivariate Bernoulli models (BNB)[45], Gaussian Naive Bayes (GNB)[46]. The hyperparameter search range of each model is shown in Table S1. The accuracy(ACC) and the AUC(Area under roc curve) were used to evaluated through 5-fold cross-validation, respectively, which shown in Figure2f. The ensemble model has a higher median value of both ACC and AUC, together with more convergent box distribution, showing a more robust effect. Among them, Bagging, GradientBoosting and RandomForest have the best effects. Combined with the ROC trend of each model in Figure S2, the GradientBoosting algorithm with a higher model upper limit (ROC=0.934) is finally selected as the modeling algorithm.

In order to explore the modeling possibilities of the Gradient Boosting algorithm as comprehensively as possible, this study uses different ranges to search for hyperparameters: "n\_estimators", "max\_depth", "min\_samples\_split", "min\_samples\_leaf", and Table 2 of the hyperparameter range. A total of 551, 840 combinations were developed, and accuracy was used as the grid search evaluation index. Each combination was evaluated through 5-fold cross-validation. Figure 2g is a schematic diagram of the results of the first hyperparameter search. "n\_estimator" is the main factor that affects the results, followed by "max\_depth" and "min\_samples\_split", and "min\_samples\_leaf" is not shown due to little influence. Considering the generalization performance of the model, we discarded the group sum with "n\_estimator" greater than the number of samples, and the final selection parameters are as follows: criterion='friedman\_mse',loss='deviance',max\_depth=17,min\_samples\_leaf=3,min\_samples\_split=10, n\_estimators=211.

图片包含 图形用户界面

描述已自动生成

图2

Figure 2 Flow chart of model feature construction. Figure 2 Flow chart of model feature construction. Figure 2a shows the funnel chart from the 278 feature structure to the final 8 features. Figure 2b, when number of features is 8,the model achieved the highest accuracy. Figure 2c shows that the SMOTE method has the best performance, except that precision is fully ahead. Figure 2d is a parallel distribution diagram of the eight features corresponding to each peptide after data enhancement, and the values are uniformly scaled between 0-10. Red represents Bitter and blue represents Umami, and the boundaries are obvious. BCUT2D\_MWLOW, PEOE\_VSA14, SMR\_VSA1, MinEStateIndex, VSA\_EState5, VSA\_EState6, VSA\_EState7, MolLogP are abbreviated as BM, PV14, SV, ME, VS5, VS6, VS7, ML.Figure 2e shows the planar spatial distribution of the data before and after the PCA dimensionality reduction, and the data is evenly distributed. Figure 2f is a violin diagram of each two-classification model under 5-fold cross-validation. The blue part on the left represents ROC, and the red part on the right represents ACC. Figure 2g is the result of grid search of the GTB model on the three hyperparameters n\_estimators, max\_depth, and min\_samples\_splite. The evaluation criterion is accuracy.

## 2.4 Performance Evaluation

In order to evaluate the performance of each binary classifier model in a fair, objective, effective and quantitative manner, this study introduces a total of 7 widely used indicators, as follows:

. The above indicators were all calculated by scikit-learn 0.24.2[29]. The closer the AUC value is to 1, the better the comprehensive classification effect. If it is 0.5, it means that there is no difference from the random classifier; The Matthews correlation coefficient(MCC) is used in machine learning as a measure of the quality of binary and multiclass classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. The MCC is in essence a correlation coefficient value between -1 and +1. A coefficient of +1 represents a perfect prediction, 0 an average random prediction and -1 an inverse prediction. 【这段是官网摘录的，我觉得说得蛮清楚的】

## 2.5 Software implementation

TastePeptides-Meta has been developed and expand as a taste Peptides universe framework that integrates taste peptide query, taste peptide prediction, and python language-assisted modeling. The front-end of TastePeptidesDB and Umami\_YYDS were built using HTML and BootStrap4 framework, and then Nginx dynamic load balancing is adopted. Uwsgi was employed to responds the back-end modeling and Umami-SQL database query requests,which built by Django3.2. The web serve has been tested on the latest versions of Google Chrome and Apple Safari for a period of 3 months, and it has performed well. Auto\_Taste\_ML as an auxiliary modeling third-party package has been released on the python package management website at https://pypi.org/project/Auto-ML-C/.

# 3 Results

## 3.1 Model interpreter

The transparency of the model algorithm determines the interpretability of the model. Interpretable algorithms, such as tree models and regression models, help model debugging and rule mining; but their performance and adaptability are not strong. While unexplainable algorithms, often called ‘black-box’, such as the forest model, have good performance, but it is often difficult for researchers to understand the decision-making process, and it is more difficult for the debugging and maintenance of the model[47]. Therefore, it is necessary to use some interpretable methods to crack the black box model, and the SHAP(SHapley Additive explanation) algorithm based on game theory is one of the commonly used tools[48]. Based on SHAP, the decision path and feature importance of the GTB model are displayed. Eight features were obtained through feature-screening, namely BCUT2D\_MWLOW(BM), PEOE\_VSA14(PV14), SMR\_VSA1(SV1), MinEStateIndex (ME), VSA\_EState5(VS5), VSA\_EState6(VS6), VSA\_EState7(VS5), MolLogP(Table 1). According to the definition of each feature, attributes such as molecular weight, charge, van der Waals surface area, and water solubility are the most important physical and chemical properties in judging umami/bitterness. This is basically consistent with the important indicators (Hydrophilicity, Acidic Amino Acid, Low Molecular Weight) considered by Charoenkwan P and others in the iUmami-SCM model[15]. The 8 features were sorted according to Permutation Importance[49], as shown in Figure 3b. It can be found that the contributions of MloLogP, VSA\_ESate6 and BCUT2D\_MWLOW to SHAP are positively correlated【也是在图3b Note部分有所说明】. According to the definition of 8 indicators in Table 1, it can be divided into 3 categories, one is solubility, the other is charge and van der Waals radius, and the third is molecular weight.

Solubility is the most important indicator for judging whether peptides are umami and bitter peptides, including MolLogP(brief in LogP) and SMR\_VSA1 (represent polarizability). Highly water-soluble peptides often mean the possibility of becoming umami peptides. Through Partial Dependence Plot (PDP), it is clear that when LogP < - 0.83 (interval is [-1.18,-0.83]), there is more than 61.5% probability that peptides are umami peptides, in [-3.84,-2.51], [-2.11,-1.64 ] Stages have 92.6% and 80.8% success rates, respectively, to make correct judgments based on the single indicator LogP (Figure S5a); when SMR\_VSA1 is greater than 24.6 (interval is [24.6,29.39], there is a 76.3% probability of judging as Umami peptides, with 100% accuracy in [34.5,49.19], use polarization information to judge the umami or bitter taste of peptides(Figure S5b). The umami peptides obtained by Yu ZL by hydrolyzing silkworm pupa all include hydrophilic structures (TAY, AAPY, VPY, GFP) [50]. YanKong hydrolyzes shiitake mushrooms to obtain umami-flavored peptides Gly-Cys-Gly, Glu-Pro-Glu, Cys-Met, Val-Phe, and Gly-Glu, each of which contains extremely amino acids [51]. As shown in Figure 5a, only the two features of SMR\_VSA1 and LogP have a good real sense of distinguishing effect. Most of the bubbles have been displayed in relatively pure colors, and they are evenly distributed with the diagonal line as the dividing line., which means that two characteristics play a leading role in the judgment of umami/bitterness.

The second is the relationship between charged properties and van der Waals surface (VSA\_ESTAT5/ESTAT6/ESTAT7, MinEStateIndex, SMR\_VSA1). These indicators are all "non-intuitive indicators" obtained by comprehensively considering the nature of the charge and the volume of Van der Waals space through a complex matrix. As shown in Figure S5c, when VSA\_Estate6 <-1.84, there is more than 74.1% probability that the peptide is an umami peptide (interval is [-2.27,1.84]), and in [-5.21,-3.61] there is a 100% success rate based on a single indicator Make the right judgment. The charge properties include the amount of charge and the positive or negative of the charge. Ken Otagiri's experiment found that when the basic amino acid arginine is adjacent to the fatty amino acid proline such as Arg-Pro, Gly-Arg-Pro and Arg-Pro-Gly, Will produce a bitter taste [52].

The third is molecular weight. BUCT2D\_MWLOW is one of the four aspects of BCUT description. The BCUT descriptor is designed to encode atomic properties related to intermolecular interactions. They have been widely used for diversity analysis[53]. The BCUT value is based on an early descriptor developed by Burden, which is calculated based on the matrix representation of the molecular connection table[54]. From Figure S5d, it is clear that the regional judgment of this value is relatively fine and fluctuating, and it may be difficult to judge by a single index. However, the chemoinformatics significance of this value is consistent with the iUmami-SCM [15],which believes that molecular weight helps The umami taste is basically the same. <0.5 kDa and 0.5-3kDa are often used as screening conditions for umami-flavor peptides, supporting by Rhyu[55], Wang WL[56], Yu M[57], Kim, Y[58], Lioe HN[59], etc. When the molecular weight is large, the peptide tends to be tasteless or bitter, which is related to Norio Ishibashi's belief that the side chain backbone of a bitter peptide should have at least 3 carbons [60]

Table 1 Feature source and calculation module

|  |  |  |
| --- | --- | --- |
| RDKit Module (Rdkit.Chem.) | Feature Selected | Explanation |
| rdMolDescriptors.BCUT2D | BCUT2D\_MWLOW | Calculates lowest and highest eigenvalues of the original Burden matrix and the three variant introduced by Pearlamn and Smith[53] |
| MolSurf module | SMR\_VSA1 | polarizability |
| EState.EState.MinEStateIndex | MinEStateIndex | MOE-type descriptors using EState indices and surface area contributions (developed at RD, not described in the CCG paper) [61] |
| EState.EState\_VSA module | VSA\_EState5 |
| VSA\_EState6 |
| VSA\_EState7 |
| Crippen module | MolLogP | Indicators for describing ligands based on atomic contribution[62] |

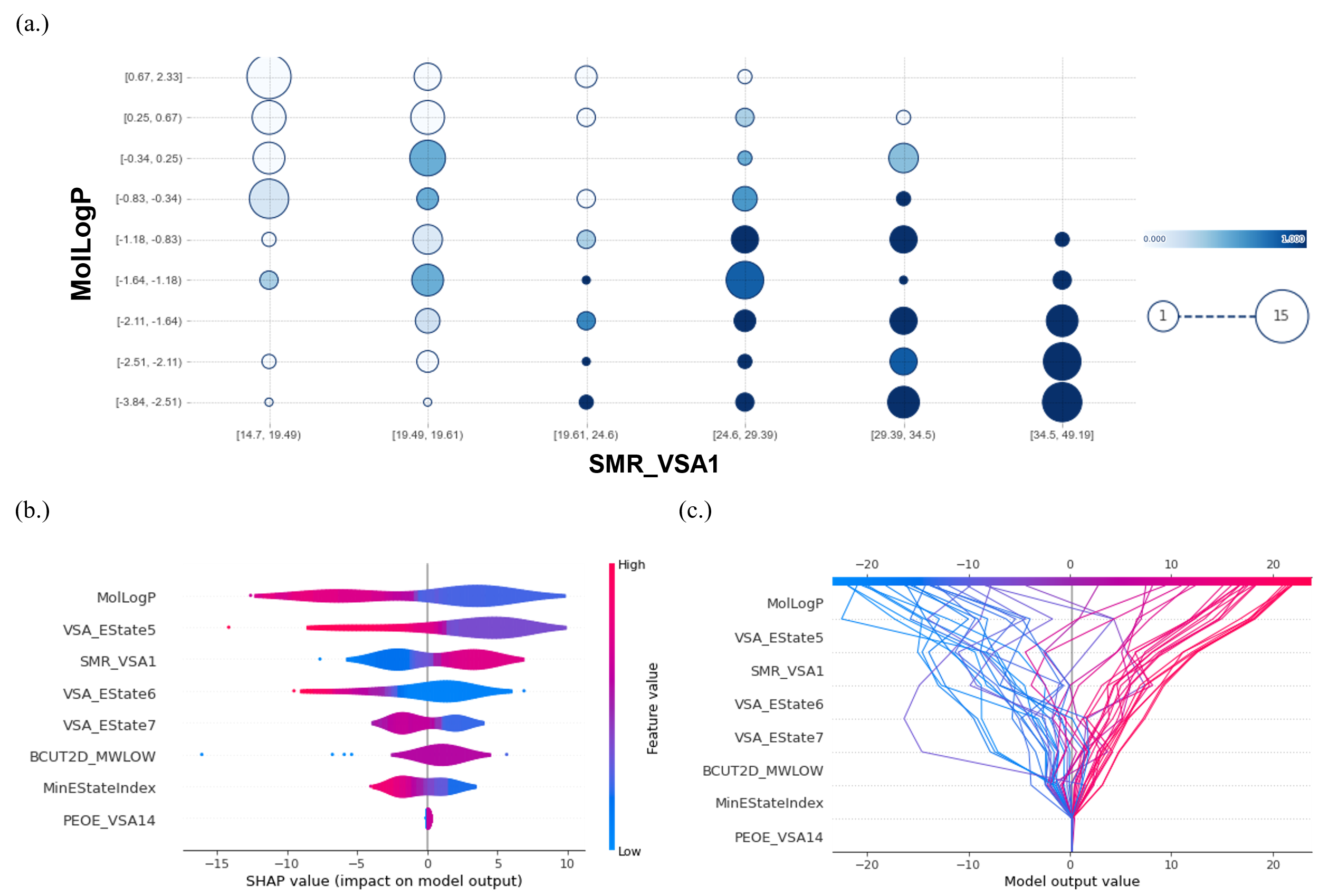


图3

Figure 3a shown the influence of MolLogP and SMR\_SVA on judgment. Figure 3b. According to the size of each characteristic SHAP, it is arranged in order from high to low on the y-axis. Most of the values of MolLogP, VSA\_ESate6 and BCUT2D\_MWLOW are on the right side of the X-axis, showing a blue dumbbell shape. Figure 3c, shown every peptides’ judgement path.

## 3.2 Comparison of Umami\_YYDS with Well-Known Taste Classifiers

The Umami\_YYDS model showed a good performed in the calibration set, which achieve 89.6% accuracy and the 98% AUC (Figure 4a&b). Umami\_YYDS shows a good taste discrimination effect for ATPD(all taste peptides dataset). Confusion matrix(Figure S4) shows that Umami\_YYDS maintains the highest accuracy rate (under the premise of 73% with iUmami-SCM[15]). The number of umami and bitter judgment examples is 46:63, which is the closest to the 198:215 ratio of umami and bitter peptide data in ATPD. This phenomenon shows that Umami\_YYDS has equally learned the umami and bitter attribute characteristics of taste peptides, while other models may overemphasize the judgment of bitter, which is reflected in more misjudgments for umami and few misjudgments for bitter. Umami\_YYDS is very similar to iUmami-SCM[15] in Accuracy(Figure 4c) (Umami\_YYDS = 0.735, iUmami-SCM[15] = 0.738) and MCC (Umami\_YYDS = 0.474, iUmami-SCM[15] = 0.485).

precision位于中等水平，考虑到是由于其他模型的过于保守的判断所致（对于bitter判断的过于倾斜）。但是在Recall方面大幅领先，表现为较高的准确率的时候较少的错判（Recall：0.768），所以F1值也是最高的，作为recall和precission的调和平均数，该结果表明Umami\_YYDS具有相对而言最理想、无偏的判断。【这里我怎么翻译都怪怪的,可以帮我写一下吗！！！】

In order to better reflect the generalization performance of Umami\_YYDS, we bring generalization test set(GTS) into the Umami\_YYDS model and compare it with other taste peptide judgment models that have been published, using confusion matrix, accuracy, precision, recall, F1 and MCC as the criterion for judgment. In terms of accuracy and F1, Umami\_YYDS is leading from hexapeptide, the first effect is stable, showing an upward trend. In terms of precision, the effect of Umami\_YYDS’s value is gradually improved by lens of peptides. Due to its "unbiased" judgment property, it’s easy to understand the reason why precision is not as good as others. In terms of Recall, although the model showed a slight downward trend and met the Q model[63] at 10 peptides, it still leads the way. In terms of MCC, Umami\_YYDS has not been widened by the best model BERT\_bitter[17], and it gradually overtook the mid-to-long peptide range. In general, the judgment of Umami\_YYDS at 4 peptides and above is reliable and extremely competitive.

图形用户界面

中度可信度描述已自动生成

**图4**

**Figure 4a** The confusion matrix of Umami\_YYDS.

## 3.3 鉴定新的鲜味肽——待完成

下周做，争取两次做完

这四种肽均具有鲜味，但鲜味程度存在显着差异。GFP、VPY和AAPY在水溶液中具有强烈的甜味，这可能是由于肽本身或合成肽过程中其他一些甜味氨基酸的甜味造成的。在肽合成过程中产生的疏水性氨基酸残基可能会导致大多数合成肽样品的[涩味](https://www.sciencedirect.com/topics/food-science/astringency" \o "从 ScienceDirect 的 AI 生成的主题页面中了解有关 Astringency 的更多信息)[64]

[YanKong](https://www.sciencedirect.com/science/article/pii/S0963996918309451?via%3Dihub#!) 文章认为 ：（[Arai、Yamashita 和 Noguchi，1973](https://www.sciencedirect.com/science/article/pii/S0963996918309451?via%3Dihub" \l "bb0005)) 观察到一些在 N 末端位置带有 L-Glu 的二肽，虽然它们会产生酸味，但在 pH 6.0 的含有 NaCl 的水溶液中也可以提供肉汤味。所以酸味肽被认为是鲜味肽的一部分[51]

暂存一篇文献，等到鲜味肽实验做完后用于进行权重解释的时候，可以用[48]

## 3.4 TastePeptides-Meta

TastePeptides-Meta currently includes three parts, namely, an recording taste peptide database TastePeptidesDB, a machine learning package Auto\_Taste\_ML for processing taste peptide data, and a web serve Umami\_YYDS for the prediction web of umami and bitter peptides.

*TastesPeptidesDB database*

The TastePeptidesDB is a database for information storage and display of taste peptides. At this stage, it contains 483 taste peptide information. It is the largest taste peptide database in published studies. The entry of each peptide includes the name (FASTA format), the taste (Taste), verified(Vitro\_verit), the simplified molecular-input line-entry system (Canonical SMILES), the literature, the paper Author (Contributor), update time and other information. The above information is obtained from nearly 100 previous studies. The query page has 4 basic functions containing precise search, taste screening, submission of new discoveries, and cross-page jump(Figure 5a). The information required to submit the newly discovered part is shown in Figure 5b.

The taste peptides in TastePeptidesDB are sorted according to their taste attributes, namely, umami, bitter, sweet, sour, kokumi, astringent and salty (Figure 5c). The taste of umami and the bitter account for most of the reported studies (79.4% in all taste peptides). This huge quantity advantage indicates that the structure or certain characteristics of peptides are easy to respond to umami taste receptors T1R1-T1R3[65], bitter taste receptor GABA [66] or T2Rs[67]. The sweet taste receptor T1R2-T1R3[68] is not easy to be activated by peptides. According to the taste of each peptide (Figure 5b), it is found that the bitter peptide and umami peptide with a single taste are still the most, followed by sweet/umami peptides, and bitter/umami peptides, indicating that there are some peptides, which can activate the receptors of umami, bitter and sweet taste. Whether these peptides have some key conformations that can activate the above two or more receptors at the same time is worthy of further consideration. As shown in Figure 5e, dipeptides and tripeptides occupy almost half of the capacity. With increasing peptide length, the number of peptides has gradually decreased. This is consistent with the study of xx that macromolecules may not have taste attributes (xxxxxxxxxxxxx)xx.

图形用户界面

描述已自动生成

**图5**

Figure 5a Taste peptide library search page; Figure 5b New taste peptide submission page; Figure 5c Taste peptide single taste distribution; Figure 5d Taste peptide multi-taste peptide distribution; Figure 5e Taste peptide length distribution

*Auto\_Taste\_ML 滋味数据建模数据包*

In terms of taste data analysis and processing, feature construction, model selection, data visualization and other aspects, there is a need for a standard workflow. Setting a set of framework systems can better help novices get started, reduce the workload of their own researchers, and recommend one API(Application Programming Interface) for communication in the suite. Auto\_Taste\_ML, which written in Python and complying with BSD protocol, is a set of third-party scientific and numerical toolkits. It serves TastePeptides-Meta and is committed to expose the entire process of TastePeptidesDB data processing and Umami\_YYDS model building, including feature construction, model selection and visualization. The model has been published in Pypi(The Python Package Index), which is the software repository of the Python programming language. The corresponding function can be realized within 1 min. The speed measurement file and detailed instructions are in the GitHub documentation README.md and README.pdf, the address is [https://github.com/SynchronyML/Auto\_Taste\_ML](https://github.com/SynchronyML/Auto_ML_C).

*Umami\_YYDS Web Server*

In order to build a direct connection between academia and industry and identify more taste peptides in a convenient way, a user-friendly web server has been developed as introduced by TastePeptides-Meta. This research published the modeling results on the Umami\_YYDS server at <http://tastepeptides-meta.com:7777/cal> (the website may be changed to [http://tastepeptides-meta.com/cal](http://tastepeptides-meta.com:7777/cal)). The webpage has been tested on the latest versions of Google Chrome and Apple Safari for a period of 3 months, and performed well. Step-by-step instructions about web server can be found in the support information.

# 4 Conclusions

In this study, the taste peptides published in recent years were sorted and published as TastepeptidesDB in the form of a web query system. Based on the above data, a umami/bitterness predictor based on chemoinformatics and amino acid sequence is proposed and published as a web service Umami\_YYDS. Through the interpretability analysis of the model, it is found that water solubility, polarization rate and van der Waals radius are the main factors affecting the taste characteristics of short peptides. Finally, the modeling process is encapsulated as an automatic machine learning package Auto\_Taste\_ML, which constitutes the TastePeptides-Meta universe.

Although the above-mentioned models have achieved excellent performance in terms of accuracy and robustness, there is still a lot of room for improvement. First, the method is based on the modeling results of a single model. At this stage, the fusion model method shows better in recognition performance in many fields. There are multiple classification judgments between basic taste such as umami, and there is even little research on its synergy, which can be studied through future database expansion. Finally, in the feature construction, we can add ligand interaction information based on molecular docking to achieve consensus judgement.

Reference

[1] Maehashi K, Huang L. Bitter peptides and bitter taste receptors[J]. Cellular and molecular life sciences, 2009, 66(10):1661-1671.

[2] Kim M J, Son H J, Kim Y, et al. Umami–bitter interactions: The suppression of bitterness by umami peptides via human bitter taste receptor[J]. Biochemical and Biophysical Research Communications, 2015, 456(2):586-590.

[3] Zhang C, Alashi A M, Singh N, et al. Beef Protein-Derived Peptides as Bitter Taste Receptor T2R4 Blockers[J]. Journal of Agricultural and Food Chemistry, 2018, 66(19):4902-4912.

[4] Zhang J, Zhao M, Su G, et al. Identification and taste characteristics of novel umami and umami-enhancing peptides separated from peanut protein isolate hydrolysate by consecutive chromatography and UPLC–ESI–QTOF–MS/MS[J]. Food Chemistry, 2019, 278:674-682.

[5] Sebald K, Dunkel A, Schäfer J, et al. Sensoproteomics: A New Approach for the Identification of Taste-Active Peptides in Fermented Foods[J]. Journal of Agricultural and Food Chemistry, 2018, 66(42):11092-11104.

[6] Yu Z L, Jiang H R, Guo R C, et al. Taste, umami-enhance effect and amino acid sequence of peptides separated from silkworm pupa hydrolysate[J]. Food Research International, 2018, 108:144-150.

[7] Lo Y C, Rensi S E, Torng W, et al. Machine learning in chemoinformatics and drug discovery[J]. Drug Discovery Today, 2018, 23(8):1538-1546.

[8] Mahmoodi-Reihani M, Abbasitabar F, Zare-Shahabadi V. In Silico Rational Design and Virtual Screening of Bioactive Peptides Based on QSAR Modeling[J]. ACS Omega, 2020, 5(11):5951-5958.

[9] Oussama C, Abdellah E, Youssef E, et al. In silico Prediction of Novel SARS-CoV 3CL(pro) Inhibitors: a Combination of 3D-QSAR, Molecular Docking, ADMET Prediction, and Molecular Dynamics Simulation[J]. Biointerface Research in Applied Chemistry, 2022, 12(4):5100-5115.

[10] Neves B J, Braga R C, Melo C C, et al. QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery[J]. Frontiers in Pharmacology, 2018, 9.

[11] Zheng S, Chang W, Xu W, et al. e-Sweet: A Machine-Learning Based Platform for the Prediction of Sweetener and Its Relative Sweetness[J]. Frontiers in Chemistry, 2019, 7(35).

[12] Charoenkwan P, Shoombuatong W, Lee H C, et al. SCMCRYS: Predicting Protein Crystallization Using an Ensemble Scoring Card Method with Estimating Propensity Scores of P-Collocated Amino Acid Pairs[J]. Plos One, 2013, 8(9).

[13] Zhong M, Chong Y, Nie X L, et al. Prediction of Sweetness by Multilinear Regression Analysis and Support Vector Machine[J]. Journal of Food Science, 2013, 78(9):S1445-S1450.

[14] Charoenkwan P, Yana J, Schaduangrat N, et al. iBitter-SCM: Identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides[J]. Genomics, 2020, 112(4):2813-2822.

[15] Charoenkwan P, Yana J, Nantasenamat C, et al. iUmami-SCM: A Novel Sequence-Based Predictor for Prediction and Analysis of Umami Peptides Using a Scoring Card Method with Propensity Scores of Dipeptides[J]. Journal of Chemical Information and Modeling, 2020, 60(12):6666-6678.

[16] Fritz F, Preissner R, Banerjee P. VirtualTaste: a web server for the prediction of organoleptic properties of chemical compounds[J]. Nucleic Acids Research, 2021, 49(W1):W679-W684.

[17] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J], 2018.

[18] Bai G L, Wu T T, Zhao L B, et al. CBDPS 1.0: A Python GUI Application for Machine Learning Models to Predict Bitter-Tasting Children's Oral Medicines[J]. Chemical & Pharmaceutical Bulletin, 2021, 69(10):989-994.

[19] Charoenkwan P, Nantasenamat C, Hasan M M, et al. BERT4Bitter: a bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides[J]. Bioinformatics, 2021, 37(17):2556-2562.

[20] Iwaniak A, Minkiewicz P, Darewicz M, et al. BIOPEP database of sensory peptides and amino acids[J]. Food Research International, 2016, 85:155-161.

[21] Zhang D, Tian Y, Tian Y, et al. A data-driven integrative platform for computational prediction of toxin biotransformation with a case study[J]. Journal of Hazardous Materials, 2021, 408:124810.

[22] Zheng S, Jiang M, Zhao C, et al. e-Bitter: Bitterant Prediction by the Consensus Voting From the Machine-Learning Methods[J]. Frontiers in Chemistry, 2018, 6(82).

[23] Liu B Y, Zhu K X, Guo X N, et al. Effect of deamidation-induced modification on umami and bitter taste of wheat gluten hydrolysates[J]. Journal of the Science of Food and Agriculture, 2017, 97(10):3181-3188.

[24] Landrum G. RDKit: Open-source cheminformatics[J], 2006.

[25] Marcou G, Horvath D, Solov'ev V, et al. Interpretability of SAR/QSAR Models of any Complexity by Atomic Contributions[J]. Molecular Informatics, 2012, 31(9):639-642.

[26] Frecer V. QSAR analysis of antimicrobial and haemolytic effects of cyclic cationic antimicrobial peptides derived from protegrin-1[J]. Bioorganic & Medicinal Chemistry, 2006, 14(17):6065-6074.

[27] Adamczak A, Ozarowski M, Karpinski T M. Antibacterial Activity of Some Flavonoids and Organic Acids Widely Distributed in Plants[J]. Journal of Clinical Medicine, 2020, 9(1).

[28] Bhonsle J B, Venugopal D, Huddler D P, et al. Application of 3D-QSAR for identification of descriptors defining bioactivity of antimicrobial peptides[J]. Journal of Medicinal Chemistry, 2007, 50(26):6545-6553.

[29] Buitinck L, Louppe G, Blondel M, et al. API design for machine learning software: experiences from the scikit-learn project[J]. Eprint Arxiv, 2013.

[30] Virtanen P, Gommers R, Oliphant T E, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python (vol 33, pg 219, 2020)[J]. Nature Methods, 2020, 17(3):352-352.

[31] Arjun P, Manoj K G. Improved Hybrid Bag-Boost Ensemble With K-Means-SMOTE–ENN Technique for Handling Noisy Class Imbalanced Data[J]. The Computer Journal, 2021.

[32] Han H, Wang W Y, Mao B H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning[J]. Lecture Notes in Computer Science, 2005.

[33] Zhang X, Wang S. Efficient Steganographic Embedding by Exploiting Modification Direction[J]. IEEE Communications Letters, 2006, 10(11):0-783.

[34] Mckinney W. Data Structures for Statistical Computing in Python[J]. proc.python sci.conf, 2010.

[35] Harris C R, Millman K J, Van Der Walt S J, et al. Array programming with NumPy[J]. Nature, 2020, 585(7825):357-362.

[36] Hunter J D. Matplotlib: A 2D Graphics Environment[J]. Computing in Science & Engineering, 2007, 9(3):90-95.

[37] Waskom M. seaborn: statistical data visualization[J]. The Journal of Open Source Software, 2021, 6(60):3021.

[38] Soltani S, Haghaei H, Shayanfar A, et al. QSBR Study of Bitter Taste of Peptides: Application of GA-PLS in Combination with MLR, SVM, and ANN Approaches[J]. BioMed Research International, 2013, 2013:501310.

[39] Platt J C. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods[J]. Advances in Large Margin Classifiers, 2000.

[40] Hastie T, Tibshirani R J, Friedman J H. The Elements of Statistical Learning: Springer[J]. Elements, 2009, 1.

[41] Ji Z, Hui Z, Rosset S, et al. Multi-class AdaBoost[C]. Statistics & Its Interface Volume, 2009.

[42] Gunopulos D, Hofmann T, Malerba D, et al. Machine Learning and Knowledge Discovery in Databases[J]. Lecture Notes in Computer Science, 2011.

[43] Trevor H, Robert T, Jerome F. The Elements of Statistical Learning[J]. Mathematical Intelligencer, 2005, 27(2):83-85.

[44] Vigneaux E, Courcoux P, Symoneaux R, et al. Random forests: A machine learning methodology to highlight the volatile organic compounds involved in olfactory perception[J]. Food Quality & Preference, 2018, 68:135-145.

[45] Manning C D, Raghavan P, Schütze H. Introduction to information retrieval[M]. Introduction to information retrieval, 2010.

[46] Sanderson, Mark. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. ISBN-13 978-0-521-86571-5, xxi + 482 pages[J]. Natural Language Engineering, 2010, 16(01):100.

[47] Parsa A B, Movahedi A, Taghipour H, et al. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis[J]. Accident Analysis & Prevention, 2020, 136:105405.

[48] Lundberg S M, Nair B, Vavilala M S, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery[J]. Nature Biomedical Engineering, 2018, 2(10):749-760.

[49] Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests[J]. Statistics and Computing, 2017, 27(3):659-678.

[50] Yu Z, Jiang H, Guo R, et al. Taste, umami-enhance effect and amino acid sequence of peptides separated from silkworm pupa hydrolysate[J]. Food Research International, 2018, 108:144-150.

[51] Kong Y, Zhang L-L, Zhao J, et al. Isolation and identification of the umami peptides from shiitake mushroom by consecutive chromatography and LC-Q-TOF-MS[J]. Food Research International, 2019, 121:463-470.

[52] Otagiri K, Nosho Y, Shinoda I, et al. Studies on a Model of Bitter Peptides Including Arginine, Proline and Phenylalanine Residues. I. Bitter Taste of Di- and Tripeptides, and Bitterness Increase of the Model Peptides by Extension of the Peptide Chain[J]. Agricultural and Biological Chemistry, 1985, 49(4):1019-1026.

[53] Beno B R, Mason J S. The design of combinatorial libraries using properties and 3D pharmacophore fingerprints[J]. Drug Discovery Today, 2001, 6(5):251-258.

[54] Pearlman R S, Smith K M. Metric validation and the receptor-relevant subspace concept[J]. Journal of Chemical Information and Computer Sciences, 1999, 39(1):28-35.

[55] Rhyu M-R, Kim E-Y. Umami taste characteristics of water extract of Doenjang, a Korean soybean paste: Low-molecular acidic peptides may be a possible clue to the taste[J]. Food Chemistry, 2011, 127(3):1210-1215.

[56] Wang W, Ning M, Fan Y, et al. Comparison of physicochemical and umami characterization of aqueous and ethanolic Takifugu obscurus muscle extracts[J]. Food and Chemical Toxicology, 2021, 154:112317.

[57] Yu M, He S, Tang M, et al. Antioxidant activity and sensory characteristics of Maillard reaction products derived from different peptide fractions of soybean meal hydrolysate[J]. Food Chemistry, 2018, 243:249-257.

[58] Kim Y, Kim E Y, Son H J, et al. Identification of a key umami-active fraction in modernized Korean soy sauce and the impact thereof on bitter-masking[J]. Food Chemistry, 2017, 233:256-262.

[59] Lioe H N, Selamat J, Yasuda M. Soy Sauce and Its Umami Taste: A Link from the Past to Current Situation[J]. Journal of Food Science, 2010, 75(3):R71-R76.

[60] Ishibashi N, Ono I, Kato K, et al. Role of the Hydrophobic Amino Acid Residue in the Bitterness of Peptides[J]. Agricultural and Biological Chemistry, 1988, 52(1):91-94.

[61] Hall L H, Mohney B, Kier L B. The Electrotopological State: An Atom Index for QSAR[J]. Quantitative Structure-Activity Relationships, 1991, 10(1):43-51.

[62] Wildman S A, Crippen G M. Prediction of Physicochemical Parameters by Atomic Contributions[J]. Journal of Chemical Information and Computer Sciences, 1999, 39(5):868-873.

[63] Ney K H. Bitterness of peptides: amino acid composition and chain length [M]. ACS Publications. 1979.

[64] Behrens M, Meyerhof W, Hellfritsch C, et al. Sweet and Umami Taste: Natural Products, Their Chemosensory Targets, and Beyond[J]. Angewandte Chemie-International Edition, 2011, 50(10):2220-2242.

[65] Kusuhara Y, Yoshida R, Ohkuri T, et al. Taste responses in mice lacking taste receptor subunit T1R1[J]. Journal of Physiology-London, 2013, 591(7):1967-1985.

[66] Duan D, Zhang H, Yue X, et al. Sensory Glia Detect Repulsive Odorants and Drive Olfactory Adaptation[J]. Neuron, 2020.

[67] Lee R J, Xiong G, Kofonow J M, et al. T2R38 taste receptor polymorphisms underlie susceptibility to upper respiratory infection[J]. Journal of Clinical Investigation, 2012, 122(11):4145-4159.

[68] Cheron J-B, Golebiowski J, Antonczak S, et al. The anatomy of mammalian sweet taste receptors[J]. Proteins-Structure Function and Bioinformatics, 2017, 85(2):332-341.