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 prediction, database, machine learning, web serve

# Abstract

Taste as one of the most important basic attributes of food, profoundly affects many processes in food industry, such as production, circulation, and trade. As one of the basic components of food, peptide have a variety of tastes including umami and bitterness, which have great influence on the taste of food. Therefore, it is necessary to clarify the law【要不要换为regulations 】 of the taste presented by the taste peptide. In this study, we used the Django framework to build the first taste peptide database (TastePeptidesDB, published at http://tastepeptides-meta.com:7777/database/son/1), based on the reported taste peptide information. Peptides with 2-3 amino acids were selected to generate QSAR descriptors (Quantitative Structure–Activity Relationship) whose number is 278【这里能不能改为非定语从句，本质意思说就是生成 278种QSAR，用 a amount of ？？？】. After statistical and modeling screening, 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 through a series of indicators including accuracy, recall, precision, F1 and Matthews correlation coefficient (MCC). Furthermore, taste peptides with 4-10 amino acids were tested by the Umami\_YYDS to illustrate the robustness of the model, which proved that Umami\_YYDS had a leading performance.【借助Umami\_YYDS模型，本研究鉴定出了三条新增滋味肽，ECH，NQS，ATQ；和一条苦味肽RVF】 Nowadays, Umami\_YYDS has been published on a prediction web site <http://tastepeptides-meta.com:7777/cal>. The above binary-model construction process was independently packaged and released as the Auto\_Taste\_ML machine learning package, which was published on https://pypi.org/project/Auto\_Taste\_ML/, aiming to help researchers in the field of taste peptide research. In summary, we established the taste peptide system TastePeptides-Meta, containing 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

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

With the improvement of computer performance and chemoinformatics, the quantitative structure-activity model (QSAR) has shown great potential in the prediction of material properties[7] successively, especially 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 regularity, previous studies have always been restricted by some factors, such as insufficient data size and simplistic models (linear regression[11], Scoring Card Method[12], support vector machine and ridge regression [13]), etc. As a result, the accuracy and generalization performance of those models are not ideal. In addition, those models can only achieve a single taste judgment, such as iBitter-SCM[14] and iUmami-SCM[15]. On the other hand, some studies put more efforts on performance and give up the exploration of the models, for example, using "black box" algorithms based on random forest [16], neural network[17], XGBoost [18], etc. to predict taste like BERT4Bitter [19], etc., which is unfavorable for the exploration of the taste regularity of peptides. Therefore, it is necessary to build a QSAR model with excellent performance and model interpretability to deal with these problems.

So far, most of the model research is still in the stage of method releasing, which only contains the introduction and description of the model construction method. None of these studies have finished the encapsulation of the code, methods, or functions of the model to a certain extent for release. Fortunately, there are still few studies published in the area of database construction (such as BIOPEP[20], AroCageDB[7], Toxindb[21]), web prediction services (for example VirtualTaste[16]) and software package (e-Bitter[22], CBDPS 1.0 [18]), which have played an exemplary role. As for related model names, algorithms, and judgments, please refer to Table S1 of Supplementary Materials.

Due to the lack of umami peptides information and the huge cost on manpower, material resources and time, there is an urgent need to establish an effective and professional flavor peptide information summary platform. In this study, we compiled the information of the reported taste peptide and released it in the form of a database TastePeptidesDB; use the data of umami and bitter peptide to construct a umami-bitterness judgment model, and encapsulated the modeling process into Auto\_Taste\_ml as Auto-Machine Helper. Up to now, there is no similar platform published like the TastePeptides-Meta in this study, which builds a systematic taste peptides universe, consisting of TastePeptidesDB, umami/bitter taste prediction model Umami\_YYDS and automatic taste model building package Auto\_Taste\_ML (Figure 1).Thereinto, TastePeptidesDB is the largest taste peptide database with the most information now, 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

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

# 2.Materials and Methods

## 2.1 Benchmark Data Sets

A total of 203 reported umami/bitter peptides were collected to support the model construction. Among them, there were 99 dipeptides (31 umami and 68 bitter) and 104 tripeptides (53 umami and 61 bitter). Considering the inhibitory effect of umami substances on bitterness, we labeled umami peptides as positive and bitter peptides as negative[2,23]. 【我们把数据发布在wwww,】

we 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 showed them on TastepeptidesDB with the address http://tastepeptides-meta.com:7777/database/son/1.

## 2.2 Feature structure

Feature selection was divided into 4 steps (Figure 2a):

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

**Step 2**: Use the variance checking algorithm (scikit-learn 0.24.2 [29]) to discard the features with 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 perspective, and the significance of 51 features with pvalue <= 0.0001 that does not show the same distribution and the data mean are not equal either, showing in 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 parameter(indicator) in turn, and the features with the largest 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, which concluded that these 8 features were retained as the final features.

Taking the imbalance of the data into account, we choose the imblearn0.8.1[29] package to oversample the umami peptide data, and use the KMeans-SMOTE[31], SMOTE[32], SVM-SMOTE[33] algorithms for data enhancement, which is different from the initial process【这里我是从以上几个方法里面选择一个合适的进行数据增强，不是说三个方法都使用】. In addition, the balanced data is compared with Random Forest Model (n\_estimators = 1000, max\_depth = 9). The SMOTE algorithm shows the best effect on the data, with excellent performance on the four indicators related with accuracy and recall, although the precision performance is slightly low (Figure 2c). The 8 features’ values of the enhanced data are scaled up to 0-10 for the visualization as shown in Figure 2d. People can easily distinguish the 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 in the method of PCA (Principal Component Analysis) to visualize the data as shown in Figure 2e. It can be found that there is a clear distinction between these data, which proves that the SMOTE data enhancement has the generalization performance.

For the purpose of independent testing and verification, the training set and 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, we used 4-10 peptides to constitute a generalization test set, referred to as GTS. All the umami and bitter peptides were constructed as a dataset called ATPD (all taste peptides dataset).

This research is done in Python3.8.10, with the numerical calculation and transmission conducted in Pandas 1.3.3[34] and Numpy 1.2.0[35], and the drawing part based on 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 dig out the internal laws of the data, this study selects 19 popular and widely recognized binary classification algorithms for the model construction[38], which contain LogisticRegression(LR)[31], RidgeClassifier, Perceptron, Stochastic Gradient Descent(SGD), Linear Discriminant Analysis(LDA), 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]and Gaussian Naive Bayes (GNB)[46]. The hyperparameter search range of each model is shown in Table S2. The accuracy (ACC) and the AUC (Area under roc curve) were used to evaluated by 5-fold cross-validation, respectively, which were shown in Figure2f. The model ensemble has higher median value in both ACC and AUC, together with more convergent box distribution, showing stronger robust effect. Among them, the Bagging, GradientBoosting and RandomForest algorithm have the best effects. Combined with the ROC trend of each model in Figure S3, the GradientBoosting algorithm with a higher 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 the hyperparameter range of Table 2 . A total of 551, 840 combinations were developed, and the accuracy was used as the grid search evaluation index. Each combination was evaluated by 5-fold cross-validation. Figure 2g is a schematic diagram of the results of the first hyperparameter search. The "n\_estimator" is the main factor that affects the results, followed by the "max\_depth" and "min\_samples\_split", while, the "min\_samples\_leaf" is not shown due to little influence. Considering the generalization performance of the model, we discarded the sum of group【为了模型的泛化性能考虑，我们抛弃了n\_estimators 参数大于滋味肽数量（483 peptides）的组合】 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, and n\_estimators=211.

图片包含 图形用户界面

描述已自动生成Figure 2 Flow chart of the model feature construction. Figure 2a shows the funnel chart from the 278 features condensed to the final 8 features. Figure 2b, when the number of features is 8, the model achieved the highest accuracy. Figure 2c shows that the SMOTE method has the best performance, except that its 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 to 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, and 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, and the evaluation criterion is accuracy.

## 2.4 Performance Evaluation

In order to evaluate the performance of each binary classifier model fairly, objectively, effectively and quantitatively, this study introduces 7 widely used indicators as follow:

The above indicators were all calculated based on scikit-learn 0.24.2[29]. The closer the AUC value is to 1, the better the comprehensive classification effect is. If the value is 0.5, which 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 the binary and multiclass classifications. It takes true & false, and positives & negatives into account, and is generally regarded as a balanced measure which can be used even if the classes are of quite different sizes. The MCC is essentially a correlation coefficient value between ‘-1’ and ‘+1‘. A coefficient of ‘+1’ represents a perfect prediction, and ‘0’ represents an average random prediction, and ‘-1’ represents an inverse prediction.

## 2.5 Sensory Evaluation

According to xxx's research, 15 panelists (6 males and 9 females, aged 22-29 years) with more than 6 months of experience in sensory evaluation of umami peptides were recruited for sensory evaluation. The experiment was carried out in an air-conditioned sensory panel room at 23 ± 2 °C and 60% humidity. In order to analyze the taste characteristics of the synthetic peptides, 5 ml of 0.05, 0.1, 0.15, 0.2, 0.4, 0.6, 0.8 mg/ml liquid was respectively put into a sensory plastic cup with a three-digit random code. Team members were asked to rotate around the sample in their mouth and spit it out. The taste attributes of bitter, sour, sweet, salty, and umami are described in the sample. In order to avoid taste fatigue, before the next analysis, team members were asked to relex for 5 minutes and rinse their mouths with ultrapure water at least twice

The ultrapure water used comes from the NW10VF water purifier system (Hong Kong, China). All peptides(ATQ, LPG, ECH, RVF, RGG, NQS ) were analytically pure and purchased from Geer Group Chemical Reagent Co., Ltd. (Shanghai, China).

## 2.6 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 by HTML and BootStrap4 framework, and then Nginx dynamic load balancing is adopted in the study. Uwsgi was employed to respond the back-end modeling and Umami-SQL database query requests, 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 showed good performance. As an auxiliary modeling third-party package, Auto\_Taste\_ML has been released on the python package management website at https://pypi.org/project/Auto-Taste-ML/.

# 3 Results

## 3.1 Model interpreter

The transparency of the model algorithm determines the interpretability of the model. Interpretable algorithms, such as the tree models and regression models, are conducive to the model debugging and rule mining, but their performance and adaptability are not always strong. On the contrary, unexplainable algorithms, often called ‘black-box’, such as the forest model, always show good performance, but it is often difficult for researchers to understand the decision-making process, and it is more difficult to debug and maintain 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-making 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 and PEOE\_VSA14 (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, which is basically consistent with the important indicators (Hydrophilicity, Acidic Amino Acid, Low Molecular Weight) concluded 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 MolLogP, VSA\_ESate6 and BCUT2D\_MWLOW to SHAP are positively correlated. According to the definition of the 8 indicators in Table 1, it can be divided into 3 categories, solubility, charge & van der Waals radius, and molecular weight.

Solubility is the most important indicator for judging whether peptides are umami or bitter peptides, compared with MolLogP (brief in LogP) and SMR\_VSA1 (represent polarizability). High water solubility often means a higher possibility to be umami peptides. According to 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 the peptides are umami peptides, while, in the Stages of [-3.84,-2.51] and [-2.11,-1.64 ] , the success rates of correct judgments are 92.6% and 80.8%, respectively, based on the single indicator LogP (Figure S5a). When the SMR\_VSA1 is greater than 24.6 (interval is [24.6,29.39], the probability of successfully determining Umami peptides is 76.3%, and the accuracy in [34.5,49.19] is 100%, using the polarization information to identify the umami or bitter taste of peptides(Figure S5b). All of the umami peptides obtained by Yu ZL hydrolyzing silkworm pupa include hydrophilic structures (TAY, AAPY, VPY, GFP) [50]. Each of the umami-flavored peptides (Gly-Cys-Gly, Glu-Pro-Glu, Cys-Met, Val-Phe, and Gly-Glu) obtained by YanKong extracted from shiitake mushrooms, contains extremely amino acids [51]. As shown in Figure 3a, only the two features of SMR\_VSA1 and LogP have a good discriminating effect. Most of the bubbles have been displayed in relatively pure colors, and they are evenly distributed along the diagonal line, which means that two characteristics play a dominant role in the judgment of umami/bitterness.

The second important indicator is the relationship between charged properties and van der Waals surface (VSA\_ESTAT5/ESTAT6/ESTAT7, MinEStateIndex, SMR\_VSA1, PEOE\_VSA14). These indicators are all "non-intuitive indicators" obtained by comprehensively considering the nature of the charge and the volume of Van der Waals space by 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 to make the right judgment. The charge properties contain the amount of charge and 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, a bitter taste will be produced[52].

The third important indicator 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 in 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 fluctuating, which makes it difficult to judge by a single index. However, the chemoinformatics significance of this value is consistent with the result of iUmami-SCM [15],which suggests 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. 【但是该值的化学信息学意义与iUmami-SCM:[14]模型认为分子量有助于鲜味基本一致。<0.5 kDa和 0.5-3kDa在很多时候被用作鲜味肽筛选条件，得到Rhyu[49]，Wang WL[50]，Yu M[51]，Kim, Y[52], Lioe HN[53]等实验结果支持】With large molecular weight, the peptide tends to be tasteless or bitter, which is related to Norio Ishibashi's law/principle 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 |
| Chem.MolSurf module | PEOE\_VSA14 | Exposes functionality for MOE-like approximate molecular surface area descriptors[62] . |
| Crippen module | MolLogP | Indicators for describing ligands based on atomic contribution[63] |

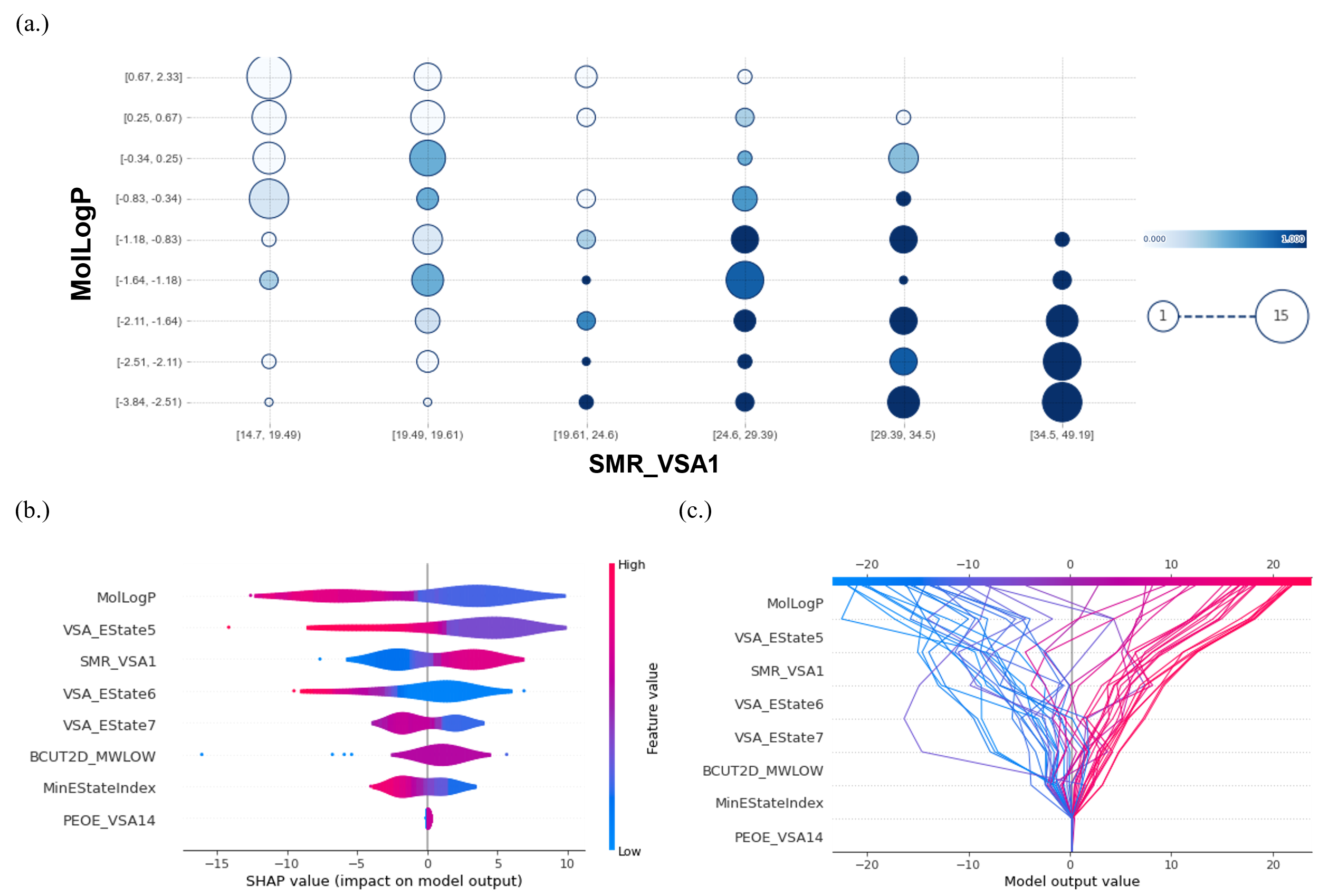


Figure 3a shows the influence of MolLogP and SMR\_SVA on judgment. Figure 3b. According to the size of each characteristic of SHAP, indexes are arranged in order from high to low on the y-axis. Most of the values of MolLogP, VSA\_ESate6 and BCUT2D\_MWLOW are on the right side of the X-axis, showing a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

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

The Umami\_YYDS model showed a good performance in the calibration set, which achieve 89.6% accuracy and the 98% AUC (Figure 4a&b). In addition, Umami\_YYDS also has a good taste recognition effect on 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]). 【其混淆矩阵展示如图4(A)，可以看到该模型判断对了几乎全部的滋味肽，各滋味仅有1-2个错误判断。当带入全部的滋味肽数据进行判断比较，Umami\_YYDS体现出了良好的滋味判别效果，从补充材料图4的各混淆矩阵可以看出，Umami\_YYDS在保持最高准确率(与iUmami-SCM都是73%）】The number of umami and bitterness judgments identified by this model is 46:63, which is the closest to 198:215 that is the ratio of umami and bitter peptide data in ATPD. This evidence 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 has very similar accuracy and MCC as iUmami-SCM[15] (Figure 4c) (Accuracy: Umami\_YYDS = 0.735, iUmami-SCM[15] = 0.738) (MCC: Umami\_YYDS = 0.474, iUmami-SCM[15] = 0.485). While, the precision of Umami\_YYDS is at the medium level, which is due to the overly conservative judgments of other models (biased towards bitter). However, it has a significant lead in the Recall section, which is manifested by less misjudgments for the relatively higher accuracy identification. Therefore, as the harmonic mean of recall and precission, F1 also shows the highest value, which indicates that Umami\_YYDS has relatively ideal and unbiased judgments.

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 published taste peptide judgment models , using confusion matrix, accuracy, precision, recall, F1 and MCC as the criterion for judgment. In terms of the 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 the length 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[64] at 10 peptides, it was still in the leading position. In terms of MCC, Umami\_YYDS has not been widened beaten by the best model BERT\_bitter[17], and it gradually overtook the mid-to-long peptide range. In general, the judgment of Umami\_YYDS for ≥ 4 peptides 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. Figure 4b ROC; Figure 4c Performance comparison among five models; Figure 4d Model generalization performance comparison

## 3.3 Identify new umami peptides

感官实验可以明确肽的真实滋味，Umami\_YYDS 从未报道的2-3肽潜在化学组合中随机检索筛选了NQS、ATQ等6条食品来源的肽进行化学合成，并感官评价以评价模型的性能（来源见表S3）。NQS等肽的滋味属性与验证结果见附表S4和图5 a.，ATQ、ECH、RVF、NQS实际滋味感知与预测结果高度一致，其中ATQ、ECH、NQS表现出较强的鲜味，识别阈值分别是0.164 mg/ml, 0.184 mg/ml, 0.148 mg/ml；RVF表现出强烈的苦味，识别阈值是0.150 mg/ml。此外，ATQ、ECH、NQS还感受到了甜味，识别阈值分别为0.134 mg/ml，0.181 mg/ml，0.137 mg/ml，甜味作为与鲜味同时存在的滋味，会增强人们对于鲜味的感知[51], 但是暂不清楚这种滋味甜味是不是和化学合成时的工艺有关[65]。而RGG和LPG预测为苦味，实际中也呈现出一定的苦味，但是鲜味感知才是主导滋味。分析RGG和LPG的特征属性值（图5b），发现除SV1 和VS7以外的属性与Umami属性的均值差不多，而Umami 肽的SV1 均值为30.647，LPG和RGG都是19.40，这与Bitter肽的18.567十分接近。同样，VS7属性方面，Umami肽为-0.366，而LPG和RGG为1.843和0.905，十分接近Bitter肽的1.213。所以认为是这两个参数导致了模型的误判，为日后模型升级提供可能。

图表

描述已自动生成

Figure 5a Synthetic peptide taste radar chart; Figure 5b Comparison of the feature values of LPG and RGG peptides with the mean values of umami/bitter peptides

Table 2 Synthetic peptide perception threshold table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Really** | **Prediction** | **Umami** | **Sweet** | **Bitter** |
| ATQ | Umami, Sweet | Umami | 0.164 | 0.134 |  |
| LPG | Umami | Bitter | 0.171 |  |  |
| ECH | Umami, Sweet | Umami | 0.184 | 0.181 |  |
| RGG | Sweet | Bitter |  | 0.226 |  |
| RVF | Bitter | Bitter |  |  | 0.150 |
| NQS | Umami, Sweet | Umami | 0.148 | 0.137 |  |

## 3.4 TastePeptides-Meta

TastePeptides-Meta currently contains three parts, that are 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 the information storage and display of taste peptides. At this stage, it contains 483 taste peptide information. It is the largest taste peptide database that have been published. 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 etc. 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 link (Figure 6a). The information required to submit for the newly discovered part is shown in Figure 6b, the method for users to submit flavor peptide information is in the supplementary material Step-by-step guide to update flavor peptide information..

The taste peptides in TastePeptidesDB are sorted according to their taste attributes, that are umami, bitter, sweet, sour, kokumi, astringent and salty (Figure 6c). The taste of umami and 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 susceptible to umami taste receptors T1R1-T1R3[66], bitter taste receptor GABA [67] or T2Rs[68]. The sweet taste receptor T1R2-T1R3[69] is not easy to be activated by peptides. According to the taste of each peptide (Figure 6d), it is found that the bitter peptide and umami peptide with a single taste still take the most amount, followed by sweet/umami peptides, and bitter/umami peptides, indicating that there are some peptides, which can activate all 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 6e, dipeptides and tripeptides occupy almost half of the capacity. With the increasing peptide length, the number of peptides has gradually decreased.

图形用户界面

描述已自动生成

Figure 6a Taste peptide library search page; Figure 6b New taste peptide submission page; Figure 6c Taste peptide with single taste distribution; Figure 6d Taste peptide with multi-taste peptide distribution; Figure 6e Taste peptide length distribution

*Auto\_Taste\_ML 滋味数据建模数据包*

In terms of the taste data processing & analysis, feature construction, model selection, data visualization and other aspects, a standard workflow is required. Setting a set of framework systems can better help novices get started and reduce the workload of their own researchers, and API (Application Programming Interface) for communication in the suite is also recommended. 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 dedicated to reveal the entire 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, and 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 study 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 showed good performance. 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 data above, an 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 the 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. Firstly, 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 by future database expansion. Finally, in the feature construction, we can add the information of ligand interaction 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] Labute P. A widely applicable set of descriptors[J]. J Mol Graph Model, 2000, 18(4-5):464-77.

[63] 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.

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

[65] 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.

[66] 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.

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

[68] 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.

[69] 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.