题目（暂定）：

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

**TastePeptides-Meta：一种新的滋味肽研究系统，包括基于化学计量学和梯度提升分类树的味肽判断模型Umami\_YYDS、滋味肽肽数据库TastePeptidesDB和开源自动机器学习包Auto\_Taste\_ML**

# Abstract

Taste is one of the important basic attributes of food, which profoundly affects many processes such as production, processing, circulation, and trade. Peptides, as one of the basic components of food, have taste diversity including umami and bitterness. It is very important to clarify the law of taste peptides presenting the sense of taste. This research first uses the published taste peptide information as data. Then use the Django framework to build the first taste peptide database, TastePeptidesDB, published at http:/ /tastepeptides-meta.com:7777/database/son/1. The database is

built specifically for recording taste peptide information, including amino acid sequences, taste attributes, and SMILES. Secondly, in this study, umami or bitter peptides with lengths 2 and 3 were selected to generate QSAR descriptors. Feature selection was performed from the perspective of statistics and modeling, and 8 features including BCUT2D\_MWLOW, PEOE\_VSA14, SMR\_VSA1, MinEStateIndex (The Electrotopological State), VSA\_EState5, VSA\_EState6, VSA\_EState7, MolLogP, etc. were obtained as feature descriptors. Then go through three steps of data enhancement, comparison of 19 algorithm and model optimization, finally select criterion='friedman\_mse',loss='deviance',max\_depth=17,min\_samples\_leaf=3,min\_samples\_split=10,n\_estimators=211 as parameters to establish a Gradient Boosting Decision Tree model Umami\_YYDS. The model achieved 89.6% accuracy and 98% AUC in the validation set. In order to better reflect the superiority of the model’s performance, compare Umami\_YYDS model with similar models iBitter-SCM, Q and other models, and analyse the results throught indicators including accuracy, recall variable, precision, F1, and Matthews corrcoef. Umami\_YYDS leads by a large margin in recall variable and F1, almost ranks first in Matthews corrcoef and accuracy, and

only ranks third in precision. In order to better illustrate the mobility of the model, the study additionally verified the model with a taste peptide of length 4-10. The results are still the same as the last round of comparison. Umami\_YYDS takes into account the mobility and robustness of the model with good accuracy. In order to facilitate research, Umami\_YYDS is published as a web prediction service at http://tastepeptides-meta.com:7777/cal. At the same time, the above 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 to build faster The binary model in professional research. In short, the research first established the taste peptide research system TastePeptides-Meta, including the largest taste peptide database TastePeptidesDB, the excellent umami/bitter taste prediction model Umami\_YYDS, and the first open source machine learning package Auto\_Taste\_ML in the field of taste peptide research.

滋味是食品重要的基本属性之一，深刻影响生产、加工、流通、贸易等多个环节。肽作为食品的基本组分之一，具有鲜味、苦味在内的味觉多样性，厘清滋味肽呈味的规律至关重要。本研究首先以已发表的滋味肽信息为数据，利用Django框架搭建首个专为记录滋味肽信息，包括氨基酸序列、滋味属性、SMILES式等资料在内的滋味肽数据库TastePeptidesDB，发布在http://tastepeptides-meta.com:7777/database/son/1。其次，本研究选择其中长度为2和3的鲜味或苦味肽生成QSAR描述符，分别从统计学和模型学角度进行特征筛选，得到BCUT2D\_MWLOW, PEOE\_VSA14, SMR\_VSA1, MinEStateIndex (The Electrotopological State), VSA\_EState5, VSA\_EState6, VSA\_EState7, MolLogP等8个特征作为特征描述符。然后经历数据增强、19种算法比较与模型优化，最终选择以criterion='friedman\_mse',loss='deviance',max\_depth=17,min\_samples\_leaf=3,min\_samples\_split=10,n\_estimators=211为参数建立梯度提升树判断模型Umami\_YYDS。该模型在验证集实现了89.6%的准确率与98%的AUC。为了更好地体现模型性能的优越性，将Umami\_YYDS与iBitter-SCM、Q等同类模型进行比较，分别以准确率、召回率、精准率、F1、马斯克系数等指标考量结果，Umami\_YYDS在召回率和F1评价中大幅领先，马斯克系数和准确率几乎是第一，只是在精准率中排名第三。为了更好地说明模型的迁移性，该研究以长度4-10的滋味肽额外验证模型，结果依旧如上一轮比较一样，Umami\_YYDS在准确率良好的情况下兼顾模型的迁移性与鲁棒性。为了方便业界研究需要，Umami\_YYDS被发布为网页预测服务，在<http://tastepeptides-meta.com:7777/cal>。同时，以上模型构建环节经过独立包装发布为Auto\_Taste\_ML 机器学习包，发布在<https://pypi.org/project/Auto-ML-C/>，旨在帮助滋味肽研究领域研究人员更快地搭建专有研究领域下的二分类模型。总之，该研究首先构建起了滋味肽研究体系TastePeptides-Meta，包括体量最大的滋味肽数据库TastePeptidesDB，性能优异的鲜味/苦味预测模型Umami\_YYDS以及第一个滋味肽研究领域的开源机器学习包Auto\_Taste\_ML。

图形用户界面, 应用程序

描述已自动生成

# 1.Introduction

Food taste is an indispensable basic 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. **The enhancement of the good taste of food by umami is closely related to the pleasant mood, while bitterness represents a perception that is not conducive to eating**[1]. Studies have reported that bitterness and umami **have a relationship that cover each up【相互掩盖】**, hence the two flavors are often compared to each other in research. Peptides, as the main degradation products of proteins, not only play a role in the nutritional intake, but also cannot be substituted in the process of taste formation. In recent years, there have been **increasingly number of relevant【层出不穷】** researches on taste peptides, and they have been reported to be found in beef[2], peanuts[3], fermented products[4] 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[5]. The time and economic cost of this work is relatively high, not to mention the toxicological evaluation of the new substance. Therefore, every piece of flavour peptide research data is precious, and there is an urgent need for an effective and professional flavour peptide information summary platform。

食品滋味是食品不可或缺的基本属性，深刻影响着食品加工、生产、贸易、营养等多个环节。鲜味和苦味作为五大基本滋味之一，鲜味对于食品美好味道的强化与愉悦心情产生密切相关，苦味则代表一种不利于食用的感知[1]。有研究报道苦味和鲜味存在相互掩盖的关系，因此两种滋味在研究中常常互为对照。肽，作为蛋白质的主要降解产物，不仅仅起到营养摄入环节发挥作用，在风味形成过程中也不可替代。近年来滋味肽的研究层出不穷，被报道在牛肉[2]、花生[3]、发酵食品[4]等食品中发现。然而传统意义上的滋味肽鉴别，是一个包括滋味物质预处理、分离、纯化、合成表征、感官评价等多环节在内的复杂工作[5]。该工作的时间与经济成本较高，更不用提新物质的毒理学评价。因此，每一份滋味肽的研究资料都弥足珍贵，迫切需要一个有效的、专业的滋味肽信息汇总平台。

With the improvement of Computer performance and the development of chemoinformatics, the quantitative structure-activity model (QSAR) has shown great potential in the prediction of material properties [6], successively in the activity judgment of biological peptides [7], ADMET [8] And molecular multi-dimensional description [9] showed excellent performance. However, in terms of the analysis of the taste regluations, previous studies have been affected by factors such as insufficient amount of flavor peptide data and simple models (linear regression [10], Scoring Card Method [11], support vector machine and ridge regression [12]), etc. As a result, the accuracy of judgment and the robustness of the model are not ideal, and in many cases the model can only achieve a single taste judgment, such as iBitter-SCM[13] and iUmami-SCM[14]. 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 [15], neural network [16], XGBoost [17], etc. to predict taste, such as BERT4Bitter [18], etc. This type of model is unfavorable for the exploration of the taste regluation of peptides. Therefore, it is necessary to adopt a QSAR model with excellent performance and model interpretability.

随着计算机性能的提升与化学信息学的发展，定量构效模型（QSAR）在物质性质预测方面表现出了巨大的潜力[6]，相继在生物肽活性判断[7]，ADMET[8]和分子多维描述等[9]领域展现出了卓越的性能。但是在呈味规律解析方面，以往研究一方面受到滋味肽数据量不足、模型简单(线性回归[10]，Scoring Card Method[11]，支持向量机与岭回归[12])等因素的影响，导致判断的准确率与模型的鲁棒性并不理想，并且在很多情况下模型只能实现单一滋味的判断，如iBitter-SCM[13]和iUmami-SCM[14]等。另一方面，部分研究追求性能而舍弃规律探索，采用基于随机森林[15]、神经网络[16]、XGBoost[17]等“黑箱”算法进行滋味预测，如BERT4Bitter[18]等，这类模型对于小分子呈味规律的探索是不利的。因此，采用一种性能优异且具备模型解释性的QSAR模型是必要的。

So far, most of the model research is still in the method release stage, only the description and introduction 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. This action is obviously contrary to the open source spirit of science and the fairness of the results. Fortunately, there are still a few studies on database construction (such as BIOPEP[19], AroCageDB[6], Toxindb[20]), web prediction services (such as VirtualTaste[15]), software (e-Bitter[21], CBDPS 1.0 [17]) . These three aspects have played an exemplary role. For related model names, algorithms and judgments, please refer to Table 1 of Supplementary Materials.

到目前为止，大多数模型研究仍提留在方法发布阶段，仅仅是模型构建方法的说明与介绍。这些工作未能将搭建模型的代码、方法或功能进行某种程度的包装以发布，这种做法显然有悖于科学的开源精神，与成果检验的公平性。所幸，仍然有少量研究在数据库构建（如BIOPEP[19]，AroCageDB[6]，T[oxindb](http://www.rxnfinder.org/toxindb/)[20]），网页预测服务（如VirtualTaste[15]），软件支持（e-Bitter[21]，CBDPS 1.0[17]）等环节起到了良好的表率作用，相关模型名称、算法及判断滋味见补充材料表1。

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 0 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.

据作者所知，目前尚无已发表研究如本研究TastePeptides-Meta一般，构建一个集滋味肽数据库TastePeptidesDB、鲜味/苦味滋味判断Umami\_YYDS、自动滋味预测模型构建包Auto\_Taste\_ML这般于一体的系统性滋味肽研究宇宙（详见图1）。其中，TastePeptidesDB是如今体量最大，收录信息最多的滋味肽数据库，Auto\_Taste\_ML是第一个开源的滋味领域机器学习包。

图形用户界面, 应用程序

描述已自动生成

**图1**

# 2.Materials and Methods

## 2.1 Benchmark Data Sets

Umami peptide data with literature support is the basis of model training. This study uses "Tastes", "Sour", "Sweet", "Bitter", "Salty", "Umami", "Kokumi", "Astringent", and "Peptides" as keywords to search in Web of Science. There are 483 flavor peptides, of which 474 have been synthesized and verified in vitro. Considering that Umami and Bitter are mutually inhibiting, the taste contradiction group (such as VE, Phe-Glu, reported in the literature, is umami[22] and bitter[23] in different literature, and the taste characteristics are confusing) is discarded , and there are 437 remaining. Among them, dipeptides and tripeptides have a total of 203, tetrapeptides and above and above 234. Considering the inhibitory effect of umami substances on bitterness, we labeled umami peptides as positive and bitter peptides as negative[24,25].For relevant data, please visit the GitHub website: or check the latest data on the official website: http://tastepeptides-meta.com:7777/database/son/1**等待最后确认后上传.滋味肽来源详见网页【这里不确定是直接丢网址比较好，还是说放在附加材料里】**

具有文献支撑的鲜味肽数据是模型训练的基础。本研究以”Tastes”, ”Sour”, ”Sweet”, ”Bitter”, “Salty”,”Umami”, “Kokumi”,” Astringent”,“Peptides” 为关键词在Web of Science进行检索，共检索到滋味肽483条，除没有经过验证的剩下474条。考虑到Umami和Bitter是相互抑制的关系，所以丢弃滋味矛盾组[比如XX肽，在文献A中报道是Umami而在文献B中报道为Bitter】，剩下437条。其中2肽与3肽共有 203条【2肽99条（31条鲜味，68条苦），3肽104条（53条鲜味，51条苦味）】，4肽及以上234条。考虑到鲜味物质对于苦味的抑制作用，我们将鲜味肽标注为阳性，苦味肽标注为阴性[24,25]。

相关数据可以访问GitHub网址xXXXX：，或者在官网查看最新数据：http://tastepeptides-meta.com:7777/database/son/1

## 2.2 Feature structure

Taking into account the diversity of peptide flavors and the insufficient number of studies, which have only been reported in recent years, this study selects dipeptides and tripeptides with sufficient research data and long-term demonstration as the basic data set. Among them, 99 dipeptides, of which 31 are umami and 68 are bitter; 104 peptide tripeptides, of which 53 are umami and 51 are bitter. Feature selection can be divided into 4 steps (as shown in the funnel diagram in Figure 1(A)):

**Step 1**: Calculate 208 molecular descriptors from 203 short peptides【**这个短肽式这么拼的嘛】** through the chemometrics special toolkit RDKit 2020.9.1[26]. This type of descriptors well describe the water solubility, electrostatic properties, atomic properties, etc. Chemical information included【病句，等我完善指标后再找天行看看】. Based on the published research, this study specially added 69 descriptors describing the planar properties of peptides, cyclic properties, aromatic properties, first and last amino acid properties and electrostatic properties (Such as the presence or absence of C-terminal hydrophobic amino acids[13]), for a total of 278 features. The descriptor data distribution of short peptides is shown in Figure 1 in the Supplementary Materials.

**Step 2**: Use the variance check algorithm of scikit-learn 0.24.2 [27] to discard the features with a variance of 0, leaving 207 features.

**Step 3**: The Kolmogorov-Smirnov test and t test of scipy [28] 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 supplementary materials figure 2.

**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 2(b), when the number of features is 8, the model It 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[27] package to Over Sampling the umami peptide data, and use the KMeans-SMOTE[29], SMOTE[30], SVM-SMOTE[31] 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 1c). The 8 feature values of the enhanced data are scaled to 0-10 for visualization as shown in Figure 2(d). 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 2(e). 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, according to the ratio of 4:1, it is divided into training set and calibration set to ensure that the two sets have the same number of Umami and Bitter peptides. The data of tetrapeptides to decapeptides are used as the test set. Considering the scarcity of undecapeptides and later peptides, this data is not representative, so all the taste peptide data sets of this study will not include the data of undecapeptides and above.

This research is completed in Python3.8.10 , the numerical calculation and transmission adopt Pandas 1.3.3[32] and Numpy 1.2.0[33], and the drawing adopts Matplotlib 3.4.2[34], Seaborn 0.11.2[35] and plotly. express 0.4.1.

考虑到多肽呈味的多样性、研究数量的不足、近年来才有所报道，本研究选择研究资料充足且经过长时间论证的二肽与三肽作为基本数据集。其中99条二肽，其中31条鲜味，68条苦味；104肽三肽，其中53条鲜味，51条苦味。特征筛选的可分为4步(如图2(a)中漏斗图所示）：

**Step 1**(Characterization of Umami Peptides):将203条滋味短肽通过化学计量学专用工具包RDKit 2020.9.1[26]计算得到208个分子描述符，该类描述符很好的描述了包括水溶性，静电性，原子特性等在内的化学信息。本研究根据已经发表的研究，加上描述肽平面特性、环特性、芳香特性和首尾氨基酸特性（如C端疏水氨基酸存在与否[13]）与静电特性等69描述符，共计278特征，短肽的描述符数据分布详见补充材料图1.

**Step 2**(Statistical screening1):通过scikit-learn 0.24.2[27]的方差检验算法丢弃方差为0的特征，还剩下207个特征。***【这里是不是将全部指标的value整理并附加在补充材料比较好！——11-5 新增疑问】***

**Step 3**(Statistical screening2):通过scipy[28]的Kolmogorov-Smirnov test和t检验从统计学角度进行特征筛选，保留51个在p-value=0.0001显著性下Umami和Bitter不服从同一分布且数据均值不相等的特征，共见补充材料图2。

**Step 4**(Model-based feature screening):采用基于随机森林的递增式特征筛选，将51个特征依次作为特征，保留每一轮迭代中效果最好的特征，如图2(b)所示，特征数量为8的时候，模型基本达到增幅收敛区，得以确认最终的保留8个特征作为最终特征。【这里的八个特征放到模型的解释性分析里面进行讲解】

考虑到模型数据的不平衡性，我们选择imblearn0.8.1[27]包对鲜味肽数据进行over sampling，分别采用KMeansSMOTE[29], SMOTE[30], SVMSMOTE[31]算法进行数据增强，与原始不平衡数据通过随机森林进(n\_estimators=1000, max\_depth=9)行结果对比。SMOTE算法效果最好，在准确率与召回率等4个指标中表现优异，虽然precision表现稍微差一点（图2c）。增强后的数据其8个特征值放缩到0-10的区间内做可视化如图2d所示，人们可以完全可以通过肉眼直观地区分出umami与bitter peptides。为了观察SMOTE数据增强前后变化，本研究通过PCA降维地方式将8维度数据降低为2维度进行数据可视化如图2e，可以发现数据间存在鲜明的区分，证明SMOTE数据增强是具有泛化性能的。

出于独立验证测试的目的，按照4：1的比例分层划分为训练集和校正集以确保两集合拥有相同数量的Umami和Bitter肽。将4肽到10肽的数据作为测试集，考虑到11肽及以后开始肽数量稀少，该数据代表性不明显，所以本研究的全部滋味肽数据集将不包括11肽及以上数据。

本研究采用Python3.8.10语言完成，其数值计算与传递采用Pandas 1.3.3[32] 与 Numpy 1.2.0[33]，绘图采用Matplotlib 3.4.2[34],Seaborn 0.11.2[35]与plotly express 0.4.1。

## 2.3 Model Selection and Optimization

The nature of the data determines the modeling algorithm, in order to select the best algorithm to better dig out the internal laws of the data[36]. This study selects 19 popular and widely recognized binary classification algorithms for model construction, including: LogisticRegression(LR)[29], RidgeClassifier, Perceptron, Stochastic Gradient Descent(SGD), Linear Discriminant Analysis, LinearSVC (LD), Support Vector Classification (SVC)[37], Nu-Support Vector Classification (NuSVC) [37], DecisionTree/ExtraTree(DT)[38], AdaBoost(ABT)[39], Bagging(BT)[40], GradientBoosting(GTBT)[41], RandomForest(RF)[42], XGBoost(XGBT)[42], KNeighbors(KN), NearestNeighbors, Naive Bayes classifier for multivariate Bernoulli models (BNB)[43], Gaussian Naive Bayes (GNB)[44], etc. For the search parameters of these models, see Table 1 of Supplementary Materials. Through 5-fold cross-validation, the accuracy(ACC) of the above two classifiers and the AUC value(Area Under Curve) can be evaluated respectively. From Figure 2(A), it can be found that the ensemble model has a higher median value of both AUC and ACC, and the upper and lower 25% and 75% quantiles are more clustered, showing a more robust effect. Among them, Bagging, GradientBoosting and RandomForest have the best effects. Combined with the ROC image trend of each model in Figure 2 of the supplementary material, 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 2(B) 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.

数据的性质决定了建模的算法，为了选择最佳的算法以更好地挖掘出数据内部的规律[36]。本研究选择19种流行且广受认可的二分类算法进行模型构建，包括：LogisticRegression[29], RidgeClassifier，Perceptron，Stochastic Gradient Descent，LinearDiscriminantAnalysis，LinearSVC(Support Vector Machines)，SVC，NuSVC，DecisionTree/ExtraTree，AdaBoost，Bagging，GradientBoosting，RandomForest，XGBoost，KNeighbors，NearestNeighbors，BernoulliNB，GaussianNB等，这些模型搜索的参数范围见补充材料表1.通过5折交叉验证可以分别评价以上二分类器的准确率与ROC的AUC值。通过图2(A)图可以发现，集成模型，无论是AUC还是ACC值的中位值都较高，且上下25%和75%分位数更加聚集，呈现出更加稳健的效果，其中Bagging，GradientBoosting和RandomForest的效果最为出色，结合补充材料图2中每一个模型的ROC图像走势，最终选择模型上限较高(ROC=0.934)的GradientBoosting算法作为建模算法。

为了尽可能全面的探索GradientBoosting算法的建模可能性，本研究采用不同范围搜索超参数：” n\_estimators”,” max\_depth”,” min\_samples\_split”,” min\_samples\_leaf”，超参数范围附表2。共展开551840种组合，以准确度为网格搜索评价指标，每种组合通过5折交叉验证进行评价。图2(B)是超参数搜索结果示意图，”n\_estimator”是影响结果的主要因素，其次是” max\_depth”与” min\_samples\_split”，而”min\_samples\_leaf”由于影响不大，暂不显示。考虑到模型的泛化性能，我们抛弃了”n\_estimator”大于样本数的组和，最终选择参数如下：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

## 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:

If you further consider the category attributes of the classification, you can calculate the average attributes of each category separately, such as precision-Recall and other indicators, Receiver operating characteristic (ROC) curves were used to assess the prediction performance of the proposed model using threshold -independent parameters. The above indicators are all calculated by scikit-learn 0.24.2[27]. It is worth mentioning that 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 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. The statistic is also known as the phi coefficient.

如果进一步考虑到分类的类别属性，那么就可以分别计算各类别属性求平均得到average属性，比如precision-Recall等指标，Receiver operating characteristic (ROC) curves were used to assess the prediction performance of the proposed model using threshold-independent parameters。以上指标皆是由scikit-learn 0.24.2 [27]计算得到。值得一提的是，AUC值越接近于1表示综合分类效果越好，如果是0.5则表示和随即分类器没有区别；The Matthews correlation coefficient 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. The statistic is also known as the phi coefficient.

## 2.5 Software implementation

TastePeptides-Meta has been developed as a TastePeptides-Meta universe framework that integrates taste peptide query, taste peptide prediction, and python language-assisted modeling, and will continue to expand. The front-ends of TastePeptidesDB and Umami\_YYDS are built using HTML language and BootStrap4 framework, using Nginx to dynamically load balance the corresponding resources, Uwsgi dynamically connecting Django 3.2 and Python 3.6 to build the software back-end, back-end processing modeling and Umami-SQL Database query request. The webpage has been tested on the latest versions of Google Chrome and Apple Safari for a period of 2 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/.

TastePeptides-Meta被开发为一个集成滋味肽查询、滋味肽预测、python语言辅助建模的滋味肽元宇宙框架，并将不断拓展。其中TastePeptidesDB，Umami\_YYDS的前端均采用HTML语言与BootStrap4框架搭建，利用Nginx动态负载均衡相应资源，Uwsgi动态连接Django3.2与Python3.6搭建的软件后端，后端处理建模与Umami-SQL数据库查询请求，该网页已在最新版本的 Google Chrome 和 Apple Safari 上进行了为期2个月的测试，表现良好。Auto\_Taste\_ML作为辅助建模第三方包已经在python包管理网站发布，网址为<https://pypi.org/project/Auto-ML-C/>。

# 3 Results

## 3.1 模型的可解释性分析

The modeling algorithm determines the transparency of the model. Through interpretability analysis based on game theory(SHAP,SHapley Additive explanation), explanatory analysis of the black box model of gradient boosting tree helps to understand the process of model judgment, thereby providing clues for model upgrading and parameter optimization. 8 model features are obtained through model screening, namely BCUT2D\_MWLOW, PEOE\_VSA14, SMR\_VSA1, MinEStateIndex (The Electrotopological State), VSA\_EState5, VSA\_EState6, VSA\_EState7, MolLogP), the source and explanation are shown in Table 1. According to the definition of each index, 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. The 8 judgment indicators are sorted according to Permutation Importance[45], as shown in Figure 5A. It can be found that the contributions of MloLogP, VSA\_ESate6 and BCUT2D\_MWLOW to SHAP are positively correlated.

建模的算法决定了模型的透明度，通过基于博弈论的可解释性分（SHAP，SHapley Additive exPlanation），对于梯度提升树这种黑箱模型进行解释性分析有助于理解模型判断的过程，从而为模型提升、参数优化提供线索。通过模型筛选得到8个模型特征,分别是BCUT2D\_MWLOW, PEOE\_VSA14, SMR\_VSA1, MinEStateIndex(The Electrotopological State), VSA\_EState5, VSA\_EState6, VSA\_EState7, MolLogP），其来源与解释见表1.根据各指标的定义，分子量、电荷、范德华表面积、水溶性等属性是在鲜味/苦味判断中至关重要的物理化学性质。这与Charoenkwan P等在iUmami-SCM模型中认为的重要指标（Hydrophilicity，Acidic Amino Acid， Low Molecular Weight）基本一致。

将8个判断指标按照Permutation Importance（特征重要度[45]）排序结果见图5A，可以发现MloLogP，VSA\_ESate6和BCUT2D\_MWLOW对于SHAP的贡献成正相关。

Solubility is the most important indicator for judging whether peptides are umami and bitter peptides, including LogP and SMR\_VSA1 (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 Supplementary Material 5B); 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. The umami peptides obtained by Yu ZL by hydrolyzing silkworm pupa all include hydrophilic structures (TAY, AAPY, VPY, GFP) [48]. 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 [46].

As shown in Figure 5B, 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. 水溶性是判断肽是否是鲜味肽与苦味肽的最重要的指标，包括LogP与SMR\_VSA1（polarizability），高水溶性的肽往往意味着成为鲜味肽的可能，通过Partial Dependence Plot（PDP）可以明确，当LogP <-0.83（区间是[-1.18,-0.83]），有超过61.5%的概率肽是鲜味肽，在[-3.84,-2.51],[-2.11,-1.64]阶段分别有92.6%和80.8%的成功率仅仅凭LogP这单一指标进行正确判断（图补充材料5B）；当SMR\_VSA1大于24.6的时候（区间为[24.6,29.39]，有76.3%的概率判断为鲜味肽，在[34.5,49.19]有100%的正确率通过极化率信息来判断肽鲜味或者苦味信息。[Yu](https://www.sciencedirect.com/science/article/pii/S0963996918301406?via%3Dihub" \l "!) ZL通过水解蚕蛹得到的鲜味肽都包括亲水结构（TAY，AAPY，VPY，GFP）[46]。[YanKong](https://www.sciencedirect.com/science/article/pii/S0963996918309451?via%3Dihub" \l "!)水解香菇得到鲜味肽Gly-Cys-Gly, Glu-Pro-Glu, Cys-Met, Val-Phe, and Gly-Glu，每一条都包括极氨基酸[47]。如图5B所示，仅仅依据SMR\_VSA1和 LogP两个特征即具有良好的实觉区分效果，大多数气泡都已比较纯粹的颜色进行展示，并且以对角线为分割线分布均匀。

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, but they are more different in feature selection than indicators such as the number of positive and negative charges constructed from an empirical point of view. As shown in Figure 5B, 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 [49].

二是带电性质与范德华表面之间的关系（VSA\_ESTAT5/ESTAT6/ESTAT7, MinEStateIndex, SMR\_VSA1)。这些指标都是通过复杂的矩阵综合考虑电荷性质和范德华空间体积运算得的“非直观指标”，但是在特征筛选中比从经验角度构造的单一的正负电荷数等指标更具有差异性。如图5B所示，当VSA\_Estate6 <-1.84有超过74.1%的概率肽是鲜味肽(区间是[-2.27,1.84])，在[-5.21,-3.61]有100%成功率仅仅凭单一指标进行正确判断。带电性质包括电荷的多少与电性的正负，[Ken Otagiri](javascript:;)实验发现当碱性氨基酸精氨酸与脂肪氨基酸脯氨酸如 Arg-Pro、Gly-Arg-Pro 和 Arg-Pro-Gly 相邻时，会产生苦味[48]。

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. The BCUT value is based on an early descriptor developed by Burden [待补充引用], which is calculated based on the matrix representation of the molecular connection table. From the supplementary material Figure 5D, 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: [5] model 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[50], Wang WL[51], Yu M[52], Kim, Y[53], Lioe HN[54], 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 [55]

第三是分子量。BUCT2D\_MWLOW 是 BCUT描述四个方面的一种，BCUT 描述符旨在编码与分子间相互作用相关的原子特性。它们已广泛用于多样性分析。BCUT 值基于 Burden [补充引文]开发的早期描述符，该描述符是根据分子连接表的矩阵表示计算。通过补充材料图5D明确，该值的区域判定较为精细且存在波动变化，可能难以仅靠单一指标进行判断，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](https://www.sciencedirect.com/science/article/pii/S0308814617316151?via%3Dihub" \l "!) M[51]，[Kim, Y](https://www.webofscience.com/wos/alldb/general-summary?queryJson=%5B%7B%22rowBoolean%22:null,%22rowField%22:%22AU%22,%22rowText%22:%22Kim,%20Yiseul%22%7D%5D&eventMode=oneClickSearch)[52], [Lioe HN](https://www.webofscience.com/wos/alldb/general-summary?queryJson=%5B%7B%22rowBoolean%22:null,%22rowField%22:%22AU%22,%22rowText%22:%22Lioe,%20Hanifah%20Nuryani%22%7D%5D&eventMode=oneClickSearch)[53]等实验结果支持。当分子量较大时，肽倾向于无味或者苦味，这与[Norio Ishibashi](javascript:;)认为苦味肽侧链骨架应至少有3个碳相关[54]

表1 特征来源与计算方式

|  |  |  |  |
| --- | --- | --- | --- |
| RDKit Module (Rdkit.Chem.) | Feature Selected | Explanation | Lib |
| rdMolDescriptors.BCUT2D | BCUT2D\_MWLOW | Calculates lowest and highest eigenvalues of the original Burden matrix and the three variant introduced by Pearlamn and Smith | Pearlman and K.M. Smith: Novel Software Tools for Chemical Diversity, Perspectives in Drug Discovery and Design, 9/10/11: 339-353, 1998[这个我稍后添加] |
| 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) | [55] |
| EState.EState\_VSA module | VSA\_EState5 |
| VSA\_EState6 |
| VSA\_EState7 |
| Crippen module | LogP | Indicators for describing ligands based on atomic contribution | [56] |

电脑萤幕画面

描述已自动生成

图3【图b和图a的顺序发生过变动】

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

The Umami\_YYDS model performed well in the calibration set. The accuracy rate of 89.6% and the AUC of 98% were shown in the 2-3 peptide set (Figure 4B). The confusion matrix is shown in Figure 4(A). The model judges almost all taste peptides correctly, with only 1-2 wrong judgments for each taste. When all the taste peptide data sets are used for judgment and comparison, Umami\_YYDS shows a good taste discrimination effect. From the confusion matrix in Figure 4 of the supplementary material, it can be seen that Umami\_YYDS maintains the highest accuracy rate ( Under the premise of 73% with iUmami-SCM), 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 the test machine. It can be seen that the model has learned the most balanced The umami and bitter attribute characteristics of taste peptides, and other models may overemphasize the judgment of bitter, which is reflected in more umami misjudgments and very few ibitter missed judgments. It can be seen from Figure 3 that Umami\_YYDS is basically the same as iUmami-SCM in Accuracy (Umami\_YYDS: 0.735, iUmami-SCm: 0.738) and Matthews corrcoef (Umami\_YYDS: 0.474, iUmami-SCM: 0.485), and precision is at a medium level. This is due to the overly conservative judgments of other models (too inclined to bitter judgments). However, it has a large lead in Recall, and it has fewer misjudgments (Recall: 0.768) when it shows a higher accuracy rate. Therefore, the F1 value is also the highest. As the harmonic average of recall and precission, the result shows that Umami\_YYDS has a relatively high accuracy. The most ideal and unbiased judgment.

Umami\_YYDS模型在校正集表现良好，在2-3肽集种体现出了高达89.6%的准确率与98%的AUC(图4B)，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%）的前提下对于umami和bitter判断实例的数量为46：63，与测试机中鲜味与苦味肽数据比例198：215最为接近，可见该模型最均衡地学习到了滋味肽的umami和bitter属性特征，而别的模型可能过于强调了bitter的判断，体现为较多的umami错判与极少的bitter漏判。从图3可知，Umami\_YYDS在Accuracy(Umami\_YYDS:0.735,iUmami-SCm:0.738)和Matthews corrcoef(Umami\_YYDS:0.474,iUmami-SCm:0.485)方面与iUmami-SCM基本持平，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。

In order to better reflect the generalization performance of Umami\_YYDS, we bring the test set data 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 Matthews corrcoef 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 is gradually improved. Due to its own "unbiased" judgment attribute, this indicator is not as good as others. The model is easy to understand. In terms of Recall, although the model showed a slight downward trend and met the Q model at 10 peptides, indicating that there is still a lot of room for improvement, it still leads the way. In terms of Matthews corrcoef, the model has not been widened by the best model BERT\_bitter, 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.

为了更好的体现Umami\_YYDS的泛化性能，我们将测试集数据带入Umami\_YYDS模型与已经发表的其他滋味肽判断模型进行比较，以混淆矩阵，准确率，精准率，召回率，F1和马修斯指数为判断标准。在准确率和F1方面，Umami\_YYDS从hexapeptide开始领先，先效果稳定，呈现出上升趋势；在precision方面，Umami\_YYDS效果逐级提升，碍于其自身“无偏“的判断属性，该指标不如别的模型也就易于理解；在Recall方面，虽然该模型呈现出细微的下降趋势并在10肽处与Q模型相遇，表明仍有较大提升空间，但依旧全程领先；在Matthews corrcoef方面，该模型并没有被最佳模型BERT\_bitter拉开较大差距，并在中长肽区间逐步反超。总的来说，Umami\_YYDS在4肽及以上的判断是可靠的，是极具竞争力的。

图形用户界面

中度可信度描述已自动生成

**图4**

## 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)[57]

[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[47]

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[58]

## 3.4 TastePeptides-Meta

TastePeptides-Meta currently includes three parts, namely TastePeptidesDB, which records the taste peptide data, the machine learning package Auto\_Taste\_ML for processing taste peptide data, and the Umami and bitter peptide identification web model Umami\_YYDS.

*TastesPeptidesDB database*

The TastePeptidesDB taste peptide database 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. he entry of each peptide includes the name of the peptide (FASTA format), the taste of the peptide (Taste), whether it has been verified by synthesis (Vitro\_verit), the simplified molecular-input line-entry system (Canonical SMILES), the source of the literature (Lib), and the paper Author (Contributor), and update time (update\_at) and other necessary display information, the above information is obtained from reading and collating from nearly 100 previous research documents, which can be consulted in Table 2 of Supplementary Materials. The TastePeptidesDB query page, as shown in Figure 3(A), contains 4 basic functions: precise search, taste screening, submission of new discoveries, and cross-page jump. The information required to submit the newly discovered part is shown in Figure 3(B), and the relevant operating instructions are in the supplementary materials.

As shown in Figure 3 (C), the taste peptides included in TastePeptidesDB are sorted according to taste attributes, followed by umami, bitter, sweet, sour, kokumi, astrigent and salty. Among them, the taste of umami and the taste of bitter account for most of the reported studies (79.4%) This huge quantity advantage indicates that the structure or certain characteristics of peptides are easy to respond to umami taste receptors T1R1-T1R3[59], bitter taste receptor GABA[60], and belong to GPCRs (G-protein coupled receptors). The sweet taste receptor T1R2-T1R3[61] is not easy to be activated. This difference is worthy of further study. Sort according to the taste of each peptide (Figure 3(B)), it is found that the bitter peptide and umami peptide with a single taste are still the most, followed by Sweet and Umami peptides, and bitter and umami peptides, indicating that it is easy to activate umami sensory receptors The presence of some of the taste peptides is also easy to activate the GABA bitter receptor and the T1R2-T1R3 sweet receptor. 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. The data included in TastesPeptidesDB is related to the length of the peptide from dipeptide to **16 peptides（不确定该怎么说）**. As shown in Figure 3 (E), dipeptides and tripeptides occupy almost half of the capacity, and then follow the peptide chain length. The increase in the number of peptides has gradually decreased, which 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3**

TastePeptidesDB滋味肽数据库以滋味肽为对象进行信息存储、展示的库，现阶段包含483个条味肽信息，是现在已发表的研究中体量最大的滋味肽数据库。每一条肽的词条包括肽的名字(FASTA 格式)，肽的滋味(Taste)，是否经过合成验证(Vitro\_verit)，Simplified molecular-input line-entry system(Canonical SMILES)[62],文献来源(Lib)，论文作者(Contributor), 以及更新时间(update\_at)等必要的展示信息，以上这些信息都是从近100篇前人研究的文献中阅读并整理得到的，在补充材料表2可以查阅。TastePeptidesDB查询页面如图3(A)所示包含4种基本功能，精准查找，滋味筛选，提交新发现和跨页跳转。其中提交新发现部分需要的信息如图3(B)所示，相关操作说明在补充材料中。

如图3（C),TastePeptidesDB收录的滋味肽按照味觉属性进行排序，依次是Umami，Bitter，Sweet，Sour，Kokumi，Astrigent和Salty，其中Umami和Bitter占据已报道研究中的大多数(79.4%)，这种巨大的数量优势说明肽的结构或者某些特征容易与鲜味受体T1R1-T1R3[59],苦味受体GABA[60]产生响应，而与同属于GPCRs(G-protein coupled receptors)的甜味受体T1R2-T1R3[61]却不易产生激活，这种差异现象值得进一步研究。按照每一条肽的滋味进行排序(图3(B))，发现单一滋味的bitter肽和Umami肽依旧是最多的，其次是Sweet和Umami肽，与 Bitter 和 Umami肽，表明易于激活Umami感知受体的滋味肽存在一部分也易于激活GABA苦味受体和T1R2-T1R3甜味受体，这些肽是不是具有一些可以同时激活以上两种及以上受体的关键构象，也是值得进一步思考的。TastesPeptidesDB中收录的数据按照肽段的长度从2肽到16肽皆有所涉及，主要展示如图3(E)所示，二肽与三肽占据了几乎一半的容量，之后随着肽链长度的增长，收录肽的数据逐渐减少，这与xx的研究认为大分子可能没有味觉属性相一致（不知道这里插入什么文献比较好）。

*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 interface for communication in the suite. Auto\_Taste\_ML is a set of third-party scientific and numerical toolkits written in Python and complying with the BSD protocol. It serves the TastePeptides-Meta research architecture and is committed to opening up the entire process of TastePeptidesDB data processing to Umami\_YYDS model building, including feature selection and visualization, and a two-class model. Build and visualize. The model has been published in Pypi, the software repository of the Python programming language. If used properly, the corresponding function can be realized within 1min. 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).

在滋味数据分析与处理，特征构造，模型选择，数据可视化方面等环节有需要标准的工作流程，设定一组框架体系，可以更好地帮助新手入门，减轻自身研究人员工作量，并建议一套方面交流的API接口。Auto\_Taste\_ML是一组用Python书写并遵守BSD协议的第三方科学与数值工具包，服务于TastePeptides-Meta研究架构，致力于打通TastePeptidesDB数据处理到Umami\_YYDS构建模型全流程，包括特征筛选及可视化，二分类模型构建及可视化。该模型已经在Python编程语言的软件存储库Pypi进行发布。运用得当的话，可以在1min之内实现对应的功能，测速文件及详细使用说明在GitHub的说明文件README.md与README. pdf中，地址为https://github.com/SynchronyML/Auto\_Taste\_ML

*Umami\_YYDS Web Server*

As introduced by TastePeptides-Meta, in order to help open up the last mile from academia to industry and help the industry to identify umami peptides in a more convenient way, a user-friendly web server must be developed. This research publishes the modeling results on the Umami\_YYDS server at http://cuilab.xyz:7777/cal (the website may be changed to http://cuilab.xyz/cal). Step-by-step instructions on using the web server can be found in the support information.

正如TastePeptides-Meta所介绍的那样，为了助力打通学术界向工业界的最后一公里，帮助业界以一种更加便捷的方式进行鲜味肽鉴别，必须开发一种用户友好的网络服务器。本研究将建模结果发布在Umami\_YYDS服务器上，网址：<http://tastepeptides-meta.com:7777/cal> （网址可能变更为http://tastepeptides-meta.com/cal）。可以在[支持信息中](https://pubs.acs.org/doi/suppl/10.1021/acs.jcim.0c00707/suppl_file/ci0c00707_si_001.pdf)找到有关使用 Web 服务器的分步说明。

# 4 Conclusions

本研究将近年来发表的滋味肽经过整理，以网页查询系统的形式发布为TastepeptidesDB ，网址为 。基于以上数据提出了一种基于化学信息学和氨基酸序列的鲜味/苦味预测器【 这一段 不会写，怎么写都感觉像 摘要 或者前言】

尽管该模型在准确性和鲁棒性等方面取得了优异的性能，但仍然存在很大的改进空间。首先，该方法是基于单一模型的建模结果，融合模型在很多领域表现出更好的表现；其次，滋味肽数据之间除了二分类，对于甜味、酸味、鲜味等基本味之间的协同作用研究甚少，数据库可以扩展记录并加以研究；最后，特征构造环节，可以加入基于分子对接等的受配体互作信息。

Reference

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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).

[11] 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).

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

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

[20] 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.

[21] 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).

[22] Kong Y, Yang X, Ding Q, et al. Comparison of non-volatile umami components in chicken soup and chicken enzymatic hydrolysate[J]. Food Research International, 2017, 102:559-566.

[23] Aspevik T, Thoresen L, Steinsholm S, et al. Sensory and Chemical Properties of Protein Hydrolysates Based on Mackerel (Scomber scombrus) and Salmon (Salmo salar) Side Stream Materials[J]. Journal of Aquatic Food Product Technology, 2021, 30(2):176-187.

[24] 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.

[25] 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.

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

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

[28] 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.

[29] 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.

[30] 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.

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

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

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

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

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

[36] 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.

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

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

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

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

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

[42] 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.

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

[44] 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.

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

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

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

[61] 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.

[62] O'boyle N M, Banck M, James C A, et al. Open Babel: An open chemical toolbox[J]. Journal of Cheminformatics, 2011, 3.