食品中甜味分子发掘模型构建

任海斌[1]，冯宝龙[2]，范 蓓[3]，贺斌彬[1]，李知陆[1]，
王清华[1]，高 飞[2]，王玉堂[1,3※]

（1. 东北农业大学乳品科学教育部重点实验室，哈尔滨 150030；2. 东北农业大学现代教育技术中心，哈尔滨 150030；
3. 中国农业科学院农产品加工研究所，北京 100193）

摘 要：食品工业一直在积极地发现新的甜味分子，传统发掘方法费时费力，效率较低。该研究基于分子的甜味和分子结构相关的假设，利用文献、专利及数据库中的数据，建立甜味、非甜味分子数据集和甜度分子数据集，采用随机森林和支持向量机算法建立定性构效关系模型定性预测甜味分子；采用主成分回归、*k* 最邻近回归、随机森林回归和偏最小二乘回归四种算法建立定量构效关系模型定量预测甜味分子的甜度。研究发现，随机森林算法模型的分类效果最好，接受者操作特性曲线下的面积为 0.987，准确度为 0.966；随机森林回归模型的甜度预测效果最好，决定系数为 0.82，误差均方根为 0.60。联用这两个模型在食品成分数据库中，发现 542 个具有甜味剂潜力的食品分子。

关键词：机器学习；甜味剂；预测；定性构效关系；定量构效关系
doi：10.11975/j.issn.1002-6819.2021.19.035
中图分类号：TS202.3 **文献标志码**：A **文章编号**：1002-6819(2021)-19-0303-06

任海斌，冯宝龙，范蓓，等. 食品中甜味分子发掘模型构建[J]. 农业工程学报，2021，37(19)：303-308.
doi：10.11975/j.issn.1002-6819.2021.19.035 http://www.tcsae.org
Ren Haibin, Feng Baolong, Fan Bei, et al. Establishment of the mining model for sweet molecules in food[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2021, 37(19): 303-308. (in Chinese with English abstract) doi：10.11975/j.issn.1002-6819.2021.19.035 http://www.tcsae.org

0 引 言

人类在与食物漫长的演化中，形成了甜味偏好[1]。这种进化而来的偏好，编码在人类基因中深深的影响着今天人类对食物的选择。甜味成为食物中基本味觉之一，可以让人产生愉悦的感觉，绝大多数人都不会拒绝甜味[2]。因此，糖和甜味剂等呈甜化合物在食品工业中得到了广泛的应用[3-5]。人们日常食用的蔗糖由于具有较高的热量，会引起肥胖、代谢紊乱和一系列疾病，如心血管疾病、高血脂、高血糖等[6-9]。研究表明高血糖是引发癌症的原因之一，长期患有糖尿病或高血糖是导致胰腺癌的一个危险因素[10]。目前已经开发了各种天然及人工合成的甜味剂，在满足对甜味味感需求的同时，减少能量的摄入，减轻患病风险[11]。但也有研究表明，长期、大量食用合成的非营养型甜味剂会有引发癌症等副作用[12]，因此食品行业一直热衷于发现更多新型、安全的甜味剂[13]。传统发现甜味剂的方法，除偶然发现外，主要采用结构改变的方法寻找新型的甜味剂，浪费了大量的时间和精力[14]，最近几年，基于数据发现新型甜味剂的研究越来越多[15]。

随着化合物的味觉信息及分子描述符越来越丰富，基于味觉信息和分子描述符，利用构效关系（Structure-activity relationship）[16]建立数学模型对分子进行定性和定量预测，从而快速发掘甜味分子并预测其甜度成为一种重要的方法[17]。2002 年，Alexander 等[18]公布了第一个甜味库 Sweet-DB，并提出发掘具有甜味的碳水合物的方法。2010 年，Ahmed 等[19]在前者的基础上，建立了可公开访问的 SuperSweet 数据库，并提出了基于构效关系和分子模拟方法的甜味发掘方法。2011 年，Yang 等[20]建立了预测糖和甜味化合物甜度的方法，但并没有公布数据库。这些研究时间久远，没有囊括一些新的天然或人工合成的化合物，没有使用大数据和机器学习的新技术。2016 年，Rojas 等[3]进一步深入研究了甜味和分子结构之间的关系。在此基础上，Cheron 等[21]提出了利用神经网络预测天然化合物甜味的方法。目前，最新的甜味分子发掘成果是 2019 年 Zheng 等[13]建立的预测甜味和甜味相关文字的机器学习平台 e-Sweet。这些最新的研究往往关注于预测一个分子是否具有甜味的定性研究，而忽略了要成为甜味剂的主要原因，应该包括预测甜度的定量研究问题。只有同时进行甜味的定性和定量研究，才能预测一个分子是否具有成为甜味剂的潜在价值，才能让研究贴近实际。另外，这些研究的数据库，无法直接获取，且只能利用这些研究内建的模型和算法进行甜味预测，无法形成数据累积，无法充分利用新的数学方法发掘新型的甜味分子，进而生产既能满足人类对甜味的需求，又具有较高安全性的甜味剂。

本研究以甜味研究文献、专利及公开数据库为数据

收稿日期：2021-05-01 修订日期：2021-08-16
基金项目：国家重点研发计划项目（2019YFF0217601-02）；中国农业科学院农产品加工研究所知识创新计划（125161015000150013）
作者简介：任海斌，研究方向为食品营养与安全。Email：renhb@neau.edu.cn
※通信作者：王玉堂，博士，副研究员，研究方向为食品营养与安全。
Email：wangyt@neau.edu.cn

源，采用人工交叉验证的方法搜集清洗数据，在 Mysql[22]中建立最大的人工修正甜味、非甜味数据集和甜味分子甜度公开数据集。利用最新的机器学习算法，首先建立定性构效关系模型，鉴别出给定分子是否呈现甜味，进一步建立定量构效关系模型，对分子的甜度做出预测，最后利用模型发掘 FooDB 数据库中潜在的甜味分子。本研究对于快速挖掘新型的潜在甜味剂，促进食品添加剂的发展具有实际意义，并对甜味数据的累积，预测方法的逐渐进步，提供了数据和方法基础。

1 材料与方法

1.1 甜味和非甜味分子数据集的建立及数据质量控制

数据来源于已有的数据库，包括 SweetDB[18]、SuperSweet[19]、PubChem[23]等数据库以及文献[20]。非甜分子是从 FlavorDB[24]以及文献[25-26]中根据氢键原子数、手性中心、分子量、油水分配系数、水溶性、疏水性和辛醇-水分配系数等性质人工筛选出的。分别获取甜味和非甜分子的名称、PubChem 化合物登录标识符（CID 或 SID）及分子结构（SMILES）用于后续研究。经人工查询已去除分子结构重复的以及分子结构过于复杂无法转化为描述符的分子，并筛选出甜度已知的甜味分子以及等数量的非甜味分子进行研究。

1.2 描述符的生成和选择

利用 MOE 软件（Molecular Operating Environment，MOE 2015.10）生成 206 个 2D 分子描述符表征分子结构[27]。采用本实验室自有软件四步法筛选分子描述符：首先用近零方差筛选和去除共线性的方法对描述符进行初步筛选；将初步筛选后得到的描述符采用相关性检验的方法对描述符做进一步筛选，即对决定分子的甜味与甜度的描述符行为进行分析，计算描述符之间、描述符与分类结果或甜度之间的相关系数，若两个描述符之间的相关系数大于 0.95，则删除对分类或对甜度贡献率小的描述符；采用主成分分析的方法对描述符进行最后筛选，删除对分类或对甜度贡献率小于 0.5 的描述符。描述符的筛选可以优化构效关系模型，提高模型的精度和预测准确度[28-29]。

1.3 甜味分子识别模型的建立和评价

采用 R 语言的 e1071（版本 1.7-4）支持向量机算法包和 RandomForest（版本 4.6-14）随机森林算法包建立甜味分子识别模型，将 80%的数据用作训练集，20%的数据用作测试集，判断给定分子是否呈现甜味。算法的实现均采用 R 软件并自行编写代码。

采用模型分类的准确度和受试者特征曲线面积来评价模型的预测效果，准确度用公式计算：

$$\mathrm{acc}(f;D)=\frac{1}{m}\sum_{i=1}^{m} II(f(x_i)=y_i) \tag{1}$$

式中 D 表示样本集$\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$，$y_i$ 代表分子的分类结果，x_i 代表自变量，即每个分子描述符，y_i 是 x_i 的真实标记，$f(x_i)$表示模型预测结果，m 代表样本数。

对于甜味分子识别模型，还可以采用受试者操作特征 ROC（Receiver Operating Characteristic）来评估模型预测质量，ROC 曲线下的面积 AUC（Area Under ROC Curve）越大，则模型预测效果越好。ROC 曲线的横轴“1-Specificity”代表“误诊率”，即“假正例率”（False Positive Rate，FPR），纵轴 Sensitivity 代表“灵敏度”，即“真正例率”（True Positive Rate，TPR），二者的定义分别是：

$$\mathrm{FPR}=\frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}} \tag{2}$$

$$\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{3}$$

式中 TP、FP、TN、FN 分别表示真正例（True Positive）：预测正确的甜味分子，假正例（False Positive）：预测错误的甜味分子，真反例（True Negative）：预测正确的非甜分子，假反例（False Negative）：预测错误的非甜分子对应的样例数，TP + FP + TN + FN = 样例总数。

1.4 甜度预测模型的建立和评价

采用 R 语言的 caret 包（版本 6.0-86）建立主成分回归（Principal Component Regression, PCR）、k 最邻近法回归（kNNR, k-Nearest Neighbor Regression）、偏最小二乘回归（PLSR, Partial Least Square Regression）、随机森林回归（RFR, Random Forest Regression）四种甜度预测模型，将 80%的数据用作训练集，20%的数据用作测试集，预测给定甜味分子的甜度。算法的实现均采用 R 软件并自行编写代码。

参数优化采用网格搜索和 10 折交叉法。10 折交叉验证法是将训练集随机划分成 10 个互补的子样本，每次选取其中 1 个子样本用作测试集，其余 9 个作训练集构建模型，重复此步骤 10 次，直到每个子样本都被用作测试集，再对每次测试集的表现结果进行综合分析[30]。通过该方法可以得出使模型预测效果达到最佳时的参数值。对于甜度预测模型，用决定系数（R^2）和均方根误差（RMSE）来评估模型的预测能力，R^2 越接近 1，RMSE 越接近 0，模型拟合效果越好。

决定系数 R^2 和均方根误差 RMSE 用公式表示为

$$R=\frac{\sum_{i=1}^{n}(y_{\exp_i}-\overline{y_{\exp_i}})(y_{\mathrm{pred}_i}-\overline{y_{\mathrm{pred}_i}})}{n\sigma_{y_{\exp}}\sigma_{y_{\mathrm{pred}}}} \tag{4}$$

$$\mathrm{RMSE}=\sqrt{\sum_{i=1}^{n}\frac{(y_{\exp}-y_{\mathrm{pred}})^2}{n}} \tag{5}$$

式中 n 是测试集中分子的个数，$y_{\exp}$ 是测试集中分子甜度的对数值 lg(S)（S 表示分子的甜度），y_{pred} 是由模型预测的分子甜度的对数值，$\sigma_{y_{\exp}}$ 和 $\sigma_{y_{\mathrm{pred}}}$ 分别是 $y_{\exp}$ 和 y_{pred} 的标准差。

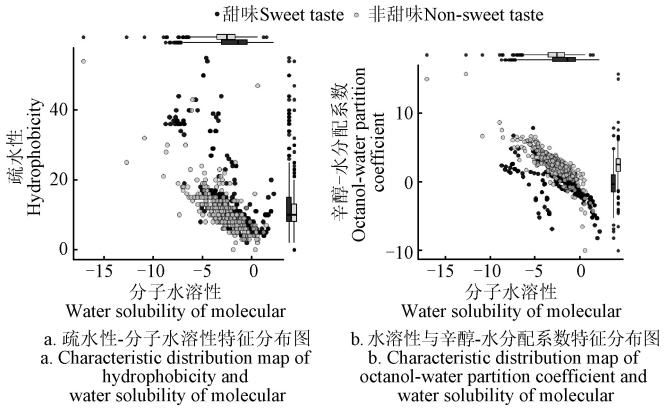
1.5 甜味分子的发掘

使用已建立的甜味分子识别模型预测 FooDB 数据库中可能具有甜味的分子，该数据库中共包含分子 28 772 个，删除掉被 MOE 识别为重复结构的分子和因结构复杂不能转化为描述符的分子，剩余分子 24 735 个。将所有分子结构转化为分子描述符后输入模型预测潜在的甜味物质，如果具有甜味，则使用甜度预测模型预测其甜度。所有代码存储在 https://gitee.com/wang_lab/EMMSM。

2　结果与分析

2.1　甜味和非甜味分子数据集

数据集包含 356 个甜味分子和 356 个非甜味分子，建立甜度预测模型所用的数据集来源于 SuperSweet 网站[19]和相关文献，共包含 356 个甜度（本文中甜度值均为以 10 为底对数处理后结果）范围在−0.744 7 到 7.350 0 之间的甜味化合物，定义蔗糖溶液在 20 ℃时的甜度为 0，其他分子的甜度为相同条件下与之相比得到的相对甜度。本研究建立的数据集是从几个数据库中严格筛选出的符合研究条件的分子，其中主要包括有机物和盐类。其中甜味分子数据集包括糖类化合物、甜味剂和其他具有甜味的化合物。本研究也分析了甜味分子和非甜味分子的氢键原子数、手性中心、分子量、油水分配系数、疏水性和辛醇-水分配系数等其他描述符性质。甜味与非甜味分子水溶性接近，疏水性和辛醇-水分配系数不同，化学空间分布如图 1 所示。图中横轴代表分子的水溶性，横轴上方的箱线图代表两类分子的水溶性分布。其中图 1a 纵轴代表疏水性，纵轴右侧的箱线图代表两类分子的疏水性分布；图 1b 纵轴代表辛醇-水分配系数，纵轴右侧的箱线图代表两类分子的辛醇-水分配系数分布。可以看出两种分子的疏水性和辛醇水分配系数差异显著，这是由于这两种特征与分子的甜度密切相关，甜度依赖于疏水基，亲水基会降低甜度，疏水基会增加甜度[31]。疏水性和辛醇-水分配系数是甜味分子的重要特征，为了提高甜味识别模型的准确度和精确度，以及甜度预测模型的决定系数，本研究筛选了疏水性较为相近的分子，使甜味、非甜味数据集较为接近，从而使训练得到的模型在 FooDB 上得到更好的应用。



a. 疏水性-分子水溶性特征分布图
a. Characteristic distribution map of hydrophobicity and water solubility of molecular

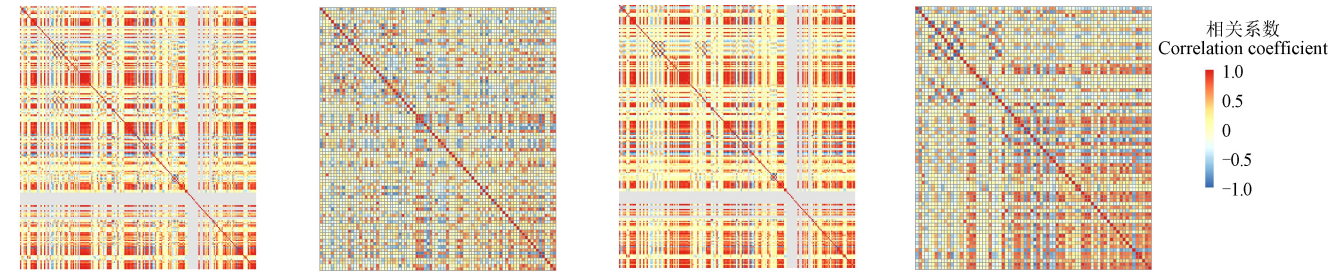
b. 水溶性与辛醇-水分配系数特征分布图
b. Characteristic distribution map of octanol-water partition coefficient and water solubility of molecular

图 1　化学空间分布图

Fig.1　Chemical spatial distribution maps

2.2　描述符的生成和筛选

甜味感觉是由分子同受体结合位点作用产生的，但结合位点往往很多，在以往的研究中，多集中于分子二维空间的研究并能取得较好的性能，而在立体异构等三维空间的研究中性能较差，这可能是由于分子三维结构的复杂性导致。因此，本研究采用 2D 描述符建模。我们在对描述符数据进行了近零方差筛选和去除共线性方法处理后，再根据相关性检验和 PCA 分析对描述符进行筛选后，甜味分子识别模型用描述符 110 个，甜度预测模型用描述符 88 个。图 2 为变量相关图，显示相关矩阵中每两个描述符之间线性关系的强度和方向，其中红色表示正相关系数，蓝色表示负相关系数，颜色越深表示相关系数的绝对值越大。通过描述符相关图分析可知，在未进行描述符筛选之前，可以明显地观察到来自所有描述之间的多重共线性非常高，经筛选后描述符之间的相关性都相对较低，从而使描述符冗余性显著降低，利于建立良好的甜味识别模型和甜度预测模型。



a. 甜味分子识别—筛选前热图
a. Molecular recognition of sweet taste-molecular heat map before the descriptor screening

b. 甜味分子识别—筛选后热图
b. Molecular recognition of sweet taste-molecular heat map after the descriptor screening

c. 甜度预测—筛选前热图
c. Prediction of sweetness-molecular heat map before the descriptor screening

d. 甜度预测—筛选后热图
d. Prediction of sweetness-molecular heat map after the descriptor screening

图 2　变量相关图

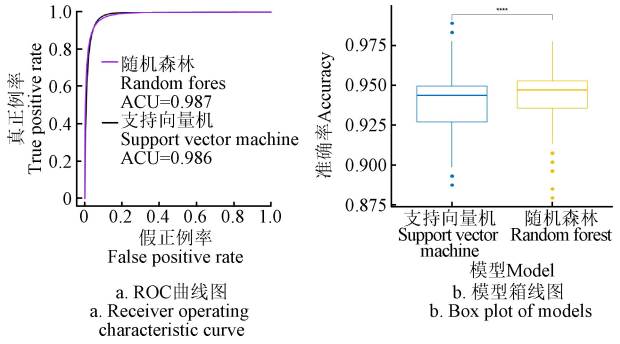
Fig.2　Variable correlation plots

2.3　甜味分子识别模型

通过无放回分层随机抽样得到包含 276 个甜味分子和 276 个非甜味分子的训练集，其余 178 个分子作为测试集，采用 RF（Random Forest）和 SVM（Support Vector Machines）两种算法建立甜味分子识别模型，对测试集样本进行分类。

在 SVM 中，选择径向基函数（radial）作为内核函数，为了优化支持向量机模型中的惩罚参数和核参数，采用了网格搜索和 10 折交叉验证的方法，这里 cost 的范围是 [10^{-6}：10^{-1}]，gamma 的范围是[10^{-10}：10^{10}]，选择交叉验证精度最好的参数 cost 为 10，gamma 为 0.01。在 10 折交叉验证中，训练集被分成 10 个相同大小的子集，使用其余 9 个子集上的训练器依次测试每一个子集，因此，整个训练集的每个实例都被预测一次，因此经过交叉验证的数据能够准确预测。RF 是一个未修剪分类和回归树的集合，并为 Bootstrap 抽样增加了额外的随机性层。RF 的主要参数是 mtry 值和 ntree 值，分别表示节点中用于二叉树的变量个数以及决策树的个数。经过网格搜索和 10 折交叉验证，确定最佳参数 mtry 值为 2，ntree 为 81。

两个模型的分类效果如图 3 所示。图 3a 中横坐标代表模型误诊率，纵坐标代表灵敏度，ROC 曲线下的面积越大表明模型分类效果越好，RF 和 SVM 二者 ACU 值分别为 0.987 和 0.986，且通过模型准确度的箱线图（图 3b）分析，两模型存在显著性差异（P<0.01），对比可以发现 RF 模型的分类效果优于 SVM 模型。Zheng 等[13]构建了甜味分子预测模型，分类准确率为 0.91。肖凌俊等[15]于 2021 年构建了甜味识别模型，分类准确率为 0.934。与以上研究相比，本研究所包含样本数据公开可用，甜味识别模型更加优秀，准确度达到了 0.966，对甜味分子有较好的预测效果。



a. ROC曲线图 a. Receiver operating characteristic curve　b. 模型箱线图 b. Box plot of models

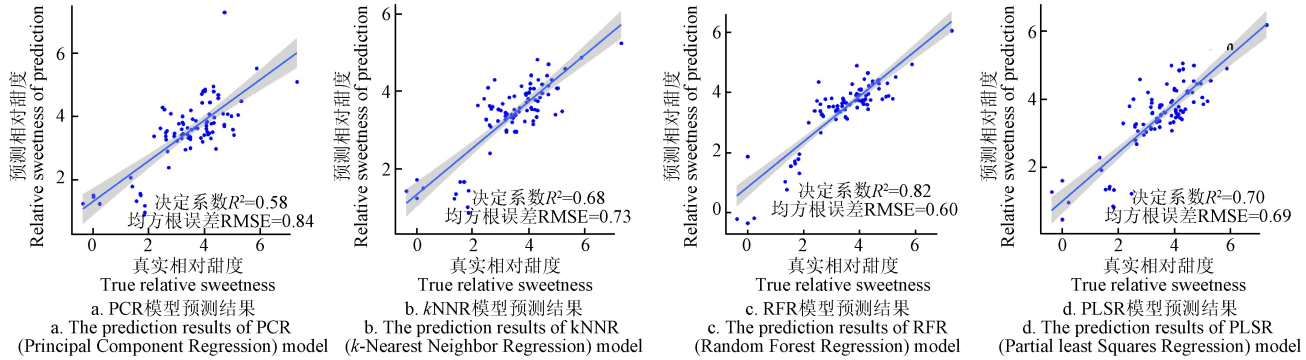
图 3　甜味分子识别模型结果

Fig.3　The results of sweet taste molecular recognition models

2.4　甜度预测模型

有学者对甜度进行了预测，舒俊生等人通过构效关系的方法对卷烟甜度进行预测，在 30 种化合物的训练集以及 10 种化合物的测试集上 R^2 达到了 0.95，模型具有较强预测能力[32]；孟骏等人通过逐步回归建立豆浆甜度预测模型，分析了 30 个大豆品种加工成豆浆的甜度值，预测模型 R^2 达到了 0.747，模型验证结果显示平均相对误差为 4.61%，因此该模型能够准确地预测豆浆甜度[26]。

本研究采用无放回抽样的方法随机将甜味化合物分为包含 267 个分子的训练集和包含 89 个分子的测试集，并对化合物甜度值进行对数处理，建立 PCR、kNNR、RFR、PLSR 四种甜度预测模型，模型通过 10 折交叉验证法选择各自最优参数后，结果如图 4 所示。图中直线代表回归拟合曲线，数据点代表测试集样本中分子的真实甜度，阴影部分代表置信区间，置信水平为 95%。图 4a 为主成分回归模型预测结果，结果显示 R^2=0.58，RMSE=0.84。图 4b 为 kNNR 模型预测结果，当 k=5 时，模型最稳定，预测效果最佳，结果显示 R^2=0.68，RMSE=0.73，甜度预测效果略优于 PCR 模型。图 4c 为 RF 回归模型预测结果，当 mtry 值为 2，ntree 为 81 时模型预测效果最好，结果显示 R^2=0.82，RMSE=0.60，甜度预测效果较为理想。图 4d 为 PLSR 模型预测结果，真实值和预测值的拟合回归线结果显示 R^2=0.70，RMSE=0.69。基于 RF 的回归模型均优于其他算法建立的模型（R^2=0.82 和 RMSE=0.60），甜度预测效果最好。



a. PCR模型预测结果 a. The prediction results of PCR (Principal Component Regression) model　b. kNNR模型预测结果 b. The prediction results of kNNR (k-Nearest Neighbor Regression) model　c. RFR模型预测结果 c. The prediction results of RFR (Random Forest Regression) model　d. PLSR模型预测结果 d. The prediction results of PLSR (Partial least Squares Regression) model

图 4　甜度预测模型预测结果

Fig.4　The prediction results of sweetness prediction models

2.5　发掘潜在的甜味物质

联用前述甜味分子定性识别模型和甜味分子甜度预测模型，预测食品中潜在的甜味成分。FooDB 是目前最大的食品成分数据库。为了发掘新的甜味分子，本研究对 FooDB 数据库进行人工和机器交叉验证，对验证后的 24 735 个分子进行了甜味与甜度预测。首先使用 RF 甜味分子识别模型预测 FooDB 分子，接着用 RF 甜度预测模型对发现理论上的甜味分子的甜度进一步预测，最终筛选出潜在甜味剂分子 542 个。根据预测概率和在食品领域的应用范围，2，3-二羟基-2-异戊酸，乙酸甲酯，肌醇，维生素 B15，6-O-α-鼠李糖-D-葡萄糖等尚未有文献报道的物质被发掘出来。所有数据存储在 https://gitee.com/wang_lab/EMMSM。通过甜味分子定性识别模型和甜味分子甜度预测模型新发掘的甜味化合物可以进一步试验测定。

表 1　部分分子结构式及甜度预测结果

Table 1　Structure formulas and sweetness prediction results of partial molecules

名称 Name	甜度值 Sweetness value
肌醇 Inositol	1.638
6-O-α-鼠李糖-D-葡萄糖 6-O-alpha-L-rhamnopyranosyl-D-glucose	3.447
2，3-二羟基-2-异戊酸 2,3-Dihydroxy-2-methylbutanoic acid	4.311
乙酸甲酯 Methyl acetate	4.885
维生素 B15 Vitamin B15	4.925

3　结　论

本研究建立了食品中甜味分子发掘模型，主要得到以下结论：

1）本研究建立了一个人工修正的、持续更新、可公开访问的非甜味、甜味物质及甜度数据集。

2）本研究建立的甜味分子识别模型，准确度达到 0.966，ROC 曲线下的面积为 0.987，具有良好的甜味分子识别能力；建立的甜度预测模型，决定系数达 0.82，均方根误差为 0.60，具有优良的甜味分子甜度预测能力。

3）本研究联用定性的甜味分子识别模型和定量的甜度预测模型，在食品成分数据库中发掘出潜在的甜味剂分子 542 个。

本研究所有数据和代码开源，其他研究人员既可以利用本研究的代码，继续发掘其他甜味剂，也可以设计新的算法，获得更为准确的预测结果。可以广泛应用于甜味分子发掘，具有较高的实际应用价值。

［参　考　文　献］

[1] Jayaram C, Mark A, Hoon N. The receptors and cells for mammalian taste[J]. Nature, 2006, 444(7117): 288-294.

[2] Burke N, Saikaly S K, Motaparthi K, et al. Malignancy-associated sweet syndrome presenting with simultaneous histopathologic and morphologic Variants[J]. JAAD Case Reports, 2021(6). DOI: 10.1016/i.jdcr.2021.06.007

[3] Rojas C, Tripaldi P, Duchowicz P R. A new qspr study on relative sweetness[J]. International Journal of Quantitative Structure-Property Relationships, 2016,1(1):78-93.

[4] Rojas C, Todeschini R, Ballabio D, et al. A qstr-based expert system to predict sweetness of molecules[J]. Front Chem, 2017,5:53.

[5] Altunayar U C, Unsalan O. Structural and anharmonic vibrational spectroscopic analysis of artificial sweetener alitame: A dat study for molecular basis of sweet taste[J]. Journal of Molecular Structure, 2021,1246:131157.

[6] Lustig R H, Schmidt L A, Brindis C D. Public health: The toxic truth about sugar[J]. Nature, 2012,482(7383):27.

[7] Goel A, Gajula K, Gupta R, et al. In-silico prediction of sweetness using structure-activity relationship models[J]. Food Chemistry, 2018,253(1):127-131.

[8] Ojha P K, Roy K. Development of a robust and validated 2d-qsar model for sweetness potency of diverse functional organic molecules[J]. Food and Chemical Toxicology, 2018,112:551-562.

[9] Bellisle F. Intense sweeteners, appetite for the sweet taste, and relationship to weight management[J]. Current Obesity Reports, 2015,4(1):106-110.

[10] Dooley J, Lagou V, Goveia J, et al. Heterogeneous effects of calorie content and nutritional components underlie dietary influence on pancreatic cancer susceptibility[J]. Cell Reports, 2020,32(2):107880.

[11] Cheron J B, Casciuc I, Golebiowski J, et al. Sweetness prediction of natural compounds[J]. Food Chemistry, 2017,221:1421.

[12] Mishra A, Ahmed K, Froghi S, et al. Systematic review of the relationship between artificial sweetener consumption and cancer in humans: Analysis of 599, 741 participants[J]. International Journal of Clinical Practice, 2015, 69(12): 1418-1426.

[13] 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. DOI: 10.3389/fchem.2019.00035.

[14] Ben S Y, Niv M Y. Structure-based screening for discovery of sweet compounds[J]. Food Chemistry, 2020,315:126286.

[15] 肖凌俊，陈爱斌，周国雄，等. 基于深度学习的甜味剂分类模型[J]. 农业工程学报，2021，37(11)：285-291.
Xiao Lingjun, Chen Aibin, Zhou Guoxiong, et al. Sweetener classification model based on deep learning[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2021, 37(11): 285-291. (in Chinese with English abstract)

[16] Lin K, Zhang L, Han X, et al. Quantitative structure-Activity relationship modeling coupled with molecular docking analysis in screening of angiotensin i-converting enzyme inhibitory peptides from qula casein hydrolysates obtained by two-enzyme combination hydrolysis[J]. J Agric Food Chem, 2018,66(12):3221-3228.

[17] Rojas C, Ballabio D, Consonni V, et al. Quantitative structure-activity relationships to predict sweet and non-sweet tastes[J]. Theoretical Chemistry Accounts, 2016, 135(3): 1-13.

[18] Alexander L, Peter B, Andreas B, et al. Sweet-db: An attempt to create annotated data collections for carbohydrates[J]. Nucleic Acids Research, 2002, 30(1): 405-408.

[19] Jessica A, Saskia P, Mathias D, et al. Supersweet—a resource on natural and artificial sweetening agents[J]. Nucleic Acids Research, 2010,39:377-382.

[20] Yang X, Chong Y, Yan A, et al. In-silico prediction of sweetness of sugars and sweeteners[J]. Food Chemistry, 2011, 128(3): 653-658.

[21] Cheron J B, Casciuc I, Golebiowski J, et al. Sweetness prediction of natural compounds[J]. Food Chemistry, 2017, 221: 1421.

[22] Jose B, Abraham S. Performance analysis of nosql and relational databases with mongodb and Mysql[J]. Materials Today: Proceedings, 2020, 24(7): 2036-2043.

[23] Štekláč M, Zajaček D, Bučinský L. 3Clpro and plpro affinity, a docking study to fight covid19 based on 900 compounds from pubchem and literature. Are there new drugs to be found?[J]. Journal of Molecular Structure, 2021, 1245: 130968.

[24] Neelansh G, Apuroop S, Rudraksh T, et al. Flavordb: A database of flavor molecules[J]. Nucleic Acids Research, 2017, 46. DOI: 10.1093/nar/gkx957

[25] Tuwani R, Wadhwa S, Bagler G. BitterSweet: Building machine learning models for predicting the bitter and sweet taste of small molecules[J]. Sci Rep, 2019, 9(1): 7155.

[26] 孟骏，汪芳，孙璐，等. 基于大豆原料蛋白质和氨基酸组成的豆浆甜度预测模型研究[J]. 食品工业科技，2019，40(10)：18-23.
Meng Jun, Wang Fang, Sun Lu, et al. Predictive model of soymilk sweetness based on protein and amino acid compositions of soybean materials[J]. Science and Technology of Food Industry, 2019, 40(10): 18-23. (in Chinese with English abstract)

[27] Wang Y, Russo D P, Liu C, et al. Predictive modeling of angiotensin i-converting enzyme inhibitory peptides using various machine learning approaches[J]. Journal of

Agricultural and Food Chemistry, 2020,68(43):12132-12140.
[28] Martínez M J, Razuc M, Ponzoni I. Modesus: a machine learning tool for selection of molecular descriptors in qsar studies applied to molecular informatics[J]. BioMed Research International, 2019, 2019: 1-12.
[29] Zhou Q, Yin J, Liang W, et al. Various machine learning approaches coupled with molecule simulation in the screening of natural compounds with xanthine oxidase inhibitory activity[J]. Food & function, 2021, 12(4): 1580-1589.
[30] Wong T T. Parametric methods for comparing the performance of two classification algorithms evaluated by k-fold cross validation on multiple data sets[J]. Pattern Recognition the Journal of the Pattern Recognition Society, 2016, 65: 97-107
[31] Deutsch E W, Hansch C. Dependence of relative sweetness on hydrophobic bonding[J]. Nature, 1966, 211(5044): 75.
[32] 舒俊生，徐志强，朱青林，等. 卷烟烟气中甜味化合物甜度的理论预测[J]. 食品工业科技，2013，34(19)：111-114.
Shu Junsheng, Xu Zhiqiang, Zhu Qinglin, et al. Theoritical predictions for sweetness of some sweet compounds in cigarette smoke[J]. Science and Technology of Food Industry, 2013, 34(19): 111-114. (in Chinese with English abstract)

Establishment of the mining model for sweet molecules in food

Ren Haibin[1], Feng Baolong[2], Fan Bei[3], He Binbin[1], Li Zhilu[1],
Wang Qinghua[1], Gao Fei[2], Wang Yutang[1,3※]

(1. *Key Laboratory of Dairy Science*, *Ministry of Education*, *Northeast Agricultural University*, *Harbin* 150030, *China*;
2. *Center for Education Technology*, *Northeast Agricultural University*, *Harbin* 150030, *China*;
3. *Institute of Food Science and Technology, Chinese Academy of Agricultural Sciences*, *Beijing* 100193, *China*)

Abstract: Sweet taste is one of the most important tastes in food flavor and quality. Sweet molecules that can be used to produce new sweeteners have also been actively explored in food processing. However, the traditional methods cannot meet the rapid development of the economy and market demand, due mainly to time-consuming, laborious, and inefficient methods. Therefore, an effective and reliable strategy is essential to produce the sweet stuff. Currently, machine learning and structure-activity relationship can be utilized to realize accurate predictions of sweet molecules in the food industry. In this study, a new database of sweeteners and non-sweeteners together with the scores of sweetness was established using molecular sweetness and structure-activity correlation between molecular structures. MOE software was selected to compute molecular descriptors, to fully characterize the properties of molecules. These descriptors were then filtered through neighborhood variance screening, collinearity removal, and principal component contribution rate screening. Specifically, the feature descriptors were screened by removing the descriptors with high correlation. 80% of the dataset was then divided into training sets for model construction, and 20% were divided into test sets for model validation. Random forest and support vector machines were utilized to establish a qualitative structure-activity relationship for the prediction and identification of potential sweet molecules. Evaluation indexes were taken as the area under the receiver characteristic curve (AUC) and accuracy rate. The higher the AUC and accuracy rate represented the better classification. As such, the optimal model was obtained. Subsequently, the principal component, K-nearest neighbor, random forest, and partial least squares regression were used to establish the quantitative structure-activity relationship for better prediction of sweet molecules. The determination coefficient R^2 and Root Mean Square Error (RMSE) were used as evaluation indexes of the quantitative structure-activity model. The higher R^2 and lower RMSE showed the better model. The optimal model was obtained to compare the performance. The food composition database (FooDB) was applied to predict the possible sweet food ingredients and the sweetness. Correspondingly, the publicly accessible dataset was established ranging from artificially revised and continuously updated on sweetener, non-sweetener substances, and sweetness values. A new model was established to identify sweet molecules using the random forest. The accuracy of the model was 0.966 on the test set, and the area under the ROC curve was 0.987, indicating excellent predictive ability. The prediction model of sweetness was also established using the random forest. Specifically, the R^2 was 0.82 and RMSE was 0.60. A manually modified data set was established to combine qualitative and quantitative sweetener prediction. 542 potential sweetener molecules, including lycopene were discovered in the food composition database. All data and code were then stored at the website of https://gitee.com/wang_lab/EMMSM for a better extension. Consequently, the new model indicated universal applicability and high practical application in searching for new sweet molecules.

Keywords: machine learning; sweetener; prediction; qualitative structure-activity relationship; quantitative structure- activity relationship