DOI：10. 13471/j. cnki. acta. snus. 2023E040

基于嗅觉受体激活关系模拟的气味感知预测*

左敏[1]，胡静珺[1]，颜文婧[1]，王瑞东[1]，张青川[1]，范大维[2]

1. 北京工商大学农产品质量安全追溯技术及应用国家工程研究中心，北京 100048
2. 北京市房山区教师进修学校，北京 102401

摘　要：气味分子与嗅觉受体相互作用是引起气味感知的重要环节，对于揭示气味感知机制具有重要意义。然而，获得气味分子与人类嗅觉受体激活关系的实验性结果耗时耗力，且目前可用的激活关系数据数量不足以支持智能气味感知研究。因此，本研究构建了嗅觉受体蛋白质关系网络，并提取特征来训练气味分子-嗅觉受体激活关系预测模型。在气味感知预测中综合考虑气味分子特征和嗅觉受体蛋白激活模拟关系，实现了对人类气味感知的高精度回归预测。实验结果表明，融合气味分子-嗅觉受体激活关系的人类气味感知预测相关度指标为0.94，明显优于现有的气味感知预测模型。此外，研究还在预测基础上总结了气味分子-嗅觉受体激活-气味感知模式。本研究为气味感知预测引入了可观测的嗅觉受体激活机制特征，为深入探索和理解气味感知机制提供了新思路。

关键词：分子特征提取；蛋白质特征提取；嗅觉受体激活预测；气味感知预测；图卷积；机器学习

中图分类号：Q-31　　**文献标志码：**A　　**文章编号：**2097－0137（2024）01－0086－10

Prediction of olfactory perception based on simulation of olfactory receptor activation relationships

ZUO Min[1], *HU Jingjun*[1], *YAN Wengjing*[1], *WANG Ruidong*[1], *ZHANG Qingchuan*[1], *FAN Dawei*[2]

1. National Engineering Research Centre for Agri-Product Quality Traceability, Beijing Technology and Business University, Beijing 100048, China
2. Beijing Fangshan District Teachers Training School, Beijing 102401, China

Abstract: The interaction between odor molecules and olfactory receptors is a crucial step in olfactory perception and holds significant importance in unraveling the mechanism of olfactory perception. However, obtaining experimental results on the activation relationship between odor molecules and human olfactory receptors is time-consuming and labor-intensive, and the available data on activation relationships is currently insufficient to support intelligent olfactory perception research. Therefore, this study constructed a network of olfactory receptor protein relationships and extracted features to train a model for predicting the activation relationship between odor molecules and olfactory receptors. By integrating the features of odor molecules and the simulated activation relationship of olfactory receptor proteins in olfactory perception prediction, high-precision regression prediction of human olfactory perception was achieved. Experimental results showed that the correlation coefficient of human olfactory perception prediction fused with odor molecule-olfactory receptor activation relationship reached 0.94, signifi-

* **收稿日期：**2023－08－01　　**录用日期：**2023－08－22　　**网络首发日期：**2023－10－23
基金项目：国家重点研发计划项目（2021YFD2100605）；北京市属高校教师队伍建设支持计划高水平科研创新团队项目（BPHR20220104）
作者简介：左敏（1973年生），男；**研究方向：**食品大数据、深度学习；E-mail：zuomin@btbu.edu.cn
通信作者：颜文婧（1985年生），女；**研究方向：**生物信息智能处理；E-mail：yanwenjing@btbu.edu.com

cantly outperforming existing olfactory perception prediction models. Additionally, the study summarized the odor molecule-olfactory receptor activation-olfactory perception pattern, enriching our understanding of the mechanism of smell perception. This study introduced observable features of olfactory receptor activation mechanisms into olfactory perception prediction, providing new insights for further exploration and understanding of the mechanism of olfactory perception.

Key words: molecular feature extraction; protein feature extraction; olfactory receptor activation prediction; olfactory perception prediction; graph convolution; machine learning

人类生理嗅觉系统十分复杂，气味分子和嗅觉受体（ORs，olfactory receptors）在气味感知表现中起着关键性作用。气味分子与嗅觉受体结合并激活嗅觉受体，将气味信号传递给大脑（Li et al.，2018），最终，人类对气味信号的感知被转化为相应的描述性词语（Lapid et al.，2011；Debnath et al.，2020；Francia et al.，2021）。受文化、语言和经验的影响，对于同一个气味分子人们可能会使用不同的感知词进行描述（Majid et al.，2018）。因此，对气味分子的气味感知进行预测是一项极具挑战性的任务。为解决这个问题，近年来智能信息研究领域尝试使用机器学习（ML，machine learning）方法构建气味感知预测模型（Keller et al.，2017），并获得了较好的效果。

目前大多数的气味感知预测模型都是从分子结构出发预测气味感知，该方式强烈依赖于分子表征（Pattanaik et al.，2020）。通常采用的方法是利用计算机表示方法对分子特征进行描述，进而构建机器学习模型。Shang et al.（2017）基于气味分子参数（MPs，molecular parameters），采用支持向量机（SVM，support vector machine）对 1 026 个分子的 10 种气味感知实现了正确率为 97.08% 的预测。Li et al.（2018）同样基于 MPs，并采用随机森林算法（RF，random forest）对 DREAM（dialogue on reverse engineering assessment and methods）数据集进行气味感知回归预测，气味强度预测的皮尔逊相关性指标达到了近似 0.6。Kasyap et al.（2022）采用图神经网络（GNNs，graph neural networks）提取分子结构特征并在 DREAM 数据集上进行气味感知多分类预测，模型的 AUC 指标为 0.89。

然而，从生理学机制上看，仅仅考虑分子物化特性无法对气味感知的形成进行解释，相似的分子结构可能产生不同的感知，而不同的分子结构也可能会产生相同的感知。研究者已经对人类嗅觉生理学机制进行揭秘，发现激活的嗅觉受体是气味感知产生的关键（Buck，2008）。目前只有少数研究基于气味分子-嗅觉受体激活关系进行气味感知预测。Kowalewski et al.（2020）发现，在气味感知预测任务上，结合嗅觉受体激活特征对气味分子进行感知预测更具优势，可取得更好的效果。

本研究首先创新性地构建了嗅觉受体蛋白质关系网络，通过引入人类嗅觉受体蛋白之间的复杂关系来学习气味分子和嗅觉受体之间的复杂非线性高维关系。其次，采用图卷积网络，在分子拓扑结构和蛋白质网络结构上提取气味分子和嗅觉受体蛋白质关系网络上的关键特征，在大规模气味感知数据集 DREAM 上实现对气味感知的精准预测。最后，基于预测的嗅觉受体激活信息，并结合模型正确决策的解释性分析，对气味分子-嗅觉受体活动-气味感知之间的模式进行分析，为人类嗅觉研究提供新的视角。

1　研究方法

1.1　研究框架

本研究首先基于人类嗅觉受体蛋白质关系网络构建嗅觉受体激活预测模型，通过图卷积方法分别提取气味分子和嗅觉受体蛋白的特征。其次，基于嗅觉受体激活预测模型的模拟结果，融合分子摩根指纹，基于 DREAM 数据集实现对气味感知的回归预测。

工作流程如图 1 所示。

1.2　蛋白质特征构建

1.2.1　嗅觉受体蛋白质关系网络构建　本研究收集了 43 个经过生物实验验证的确定可以被特定配体激活的人类嗅觉受体（Vassar et al.，1993；Matarazzo et al.，2005；Jacquier et al.，2006；Neuhaus et al.，2006；Braun et al.，2007；Fujita et al.，2007；Keller et al.，2007；Menashe et al.，2007；Schmiedeberg et al.，2007；Cook et al.，2009；Saito et al.，2009；Jaeger et al.，2013；Topin et al.，2014；Shirasu

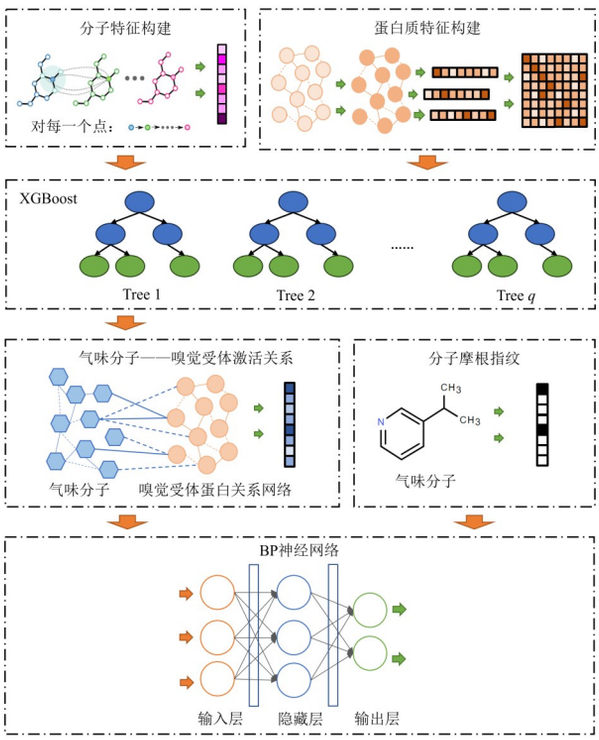

图 1 气味感知预测工作流程图

Fig. 1 Olfactory perception prediction workflow diagram

et al.,2014），采用 One-Hot 编码表示嗅觉受体蛋白氨基酸序列。嗅觉受体蛋白质三级结构信息有两个不同的来源。经过实验验证的结构来自于 Uniport 蛋白质数据库，未知的嗅觉受体蛋白质三级结构则采用 AlphaFold 蛋白质 3D 结构预测模型进行预测。

嗅觉受体蛋白质关系网络以嗅觉受体蛋白质作为节点，其氨基酸序列作为节点特征，嗅觉受体蛋白质三级结构的相似关系作为边。根据已获得的嗅觉受体蛋白质三级结构，本研究采用 TM-score（template modeling score）方法计算蛋白质之间的相似度。TM-score 是一种用于评估蛋白质结构拓扑相似性的指标，通过比较两个蛋白质全局结构的相似性来评估它们的匹配程度，其取值范围介于 0 到 1 之间。TM-score 低于 0.17 被认为对应于随机选择的不相关蛋白质（Zhang et al.,2004），而大于 0.5 则表示具有相似的折叠状态（Xu et al.,2010）。TM-score 的计算公式为

$$\text{TM-score} = \max\left[\frac{1}{L_{\text{target}}}\sum_{t}^{L_{\text{commom}}}\frac{1}{1+\left(\dfrac{d_t}{d_0\left(L_{\text{target}}\right)}\right)^2}\right], \quad (1)$$

其中 L_{target} 是目标蛋白质的氨基酸序列长度，L_{commom} 是在模板结构和目标结构中均存在的残基数量，d_t 是模板和目标结构中第 t 对残基之间的距离，$d_0(L_{\text{target}})$ 是用来归一化距离的距离尺度。获得嗅觉受体蛋白质三级结构相似度后，可以构建嗅觉受体蛋白质关系网络。

1.2.2 蛋白质图卷积特征 图卷积（graph convolution）是一种适用于处理具有节点间关联关系的图数据的卷积操作方法。在本研究中，嗅觉受体蛋白质关系网络表示为 $\boldsymbol{G}_{\text{p}}=(\boldsymbol{V}_{\text{p}},\boldsymbol{E}_{\text{p}})$，其中节点集合 $\boldsymbol{V}_{\text{p}}$ 表示嗅觉受体蛋白氨基酸序列集合，边集合 $\boldsymbol{E}_{\text{p}}$ 表示嗅觉受体蛋白质三级结构之间的相似度集合。每个节点的特征向量定义为 $\boldsymbol{v}_{\text{p}}$，$\boldsymbol{v}_{\text{p}}\in\boldsymbol{V}_{\text{p}}$，边的特征向量定义为 $\boldsymbol{e}_{\text{p}}$，$\boldsymbol{e}_{\text{p}}\in\boldsymbol{E}_{\text{p}}$。

嗅觉受体蛋白质关系网络是通过对嗅觉受体蛋白氨基酸序列和嗅觉受体蛋白质三级结构相似度进行编码得到的。嗅觉受体蛋白氨基酸序列被编码为一个具有 20 种氨基酸和 331 个序列位置的特征向量，其维度为[20,331]，嗅觉受体蛋白质三级结构间相似度被编码为一个维度为 1 的特征向量。

蛋白质图卷积特征的构建方法如下：

设 $\boldsymbol{H}_{\text{p}}^{i}$ 是第 i 层节点的特征表示矩阵，邻接矩阵 $\boldsymbol{A}_{\text{p}}$ 表示节点间的连接关系，度矩阵 $\boldsymbol{D}_{\text{p}}$ 为 $\boldsymbol{A}_{\text{p}}$ 的对角矩阵。图卷积计算公式

$$\tilde{\boldsymbol{A}}_{\text{p}}=\boldsymbol{A}_{\text{p}}+\boldsymbol{I}_{\text{p}}, \quad (2)$$

$$\tilde{\boldsymbol{H}}_{\text{p}}^{i}=\tilde{\boldsymbol{D}}_{\text{p}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}_{\text{p}}\tilde{\boldsymbol{D}}_{\text{p}}^{-\frac{1}{2}}\boldsymbol{H}_{\text{p}}^{i}, \quad (3)$$

其中式（2）表示将自环添加到邻接矩阵中，$\tilde{\boldsymbol{D}}_{\text{p}}$ 表示 $\tilde{\boldsymbol{A}}_{\text{p}}$ 的度矩阵，$\boldsymbol{I}_{\text{p}}$ 是单位矩阵。

对矩阵 $\tilde{\boldsymbol{H}}_{\text{p}}^{i}$ 进行线性变换，并应用非线性激活函数 σ_{p}，得到一个新的特征向量输出为

$$\boldsymbol{H}_{\text{p}}^{i+1}=\sigma_{\text{p}}(\tilde{\boldsymbol{H}}_{\text{p}}^{i}\boldsymbol{W}_{\text{p}}^{i}), \quad (4)$$

其中 $\boldsymbol{W}_{\text{p}}^{i}$ 是第 i 层到第 i+1 层的权重矩阵。

1.3 分子特征构建

1.3.1 分子摩根指纹 在本研究中，任意分子图表示为 $\boldsymbol{G}_{\text{m}}=(\boldsymbol{V}_{\text{m}},\boldsymbol{E}_{\text{m}})$，其中节点集合 $\boldsymbol{V}_{\text{m}}$ 表示原子集合，边集合 $\boldsymbol{E}_{\text{m}}$ 表示化学键集合。每个原子的特征向量定义为 $\boldsymbol{v}_{\text{m}}$，$\boldsymbol{v}_{\text{m}}\in\boldsymbol{V}_{\text{m}}$，化学键的特征向量定义为 $\boldsymbol{e}_{\text{m}}$，$\boldsymbol{e}_{\text{m}}\in\boldsymbol{E}_{\text{m}}$。

摩根指纹（Morgan fingerprints）方法是一种用于描述分子结构的化学指纹方法。它基于分子的拓扑结构，对于节点 $\boldsymbol{v}$ 通过递归遍历分子的邻居节点 $\boldsymbol{u}\in\boldsymbol{R}_{\boldsymbol{v}}$，$\boldsymbol{R}_{\boldsymbol{v}}$ 是与节点 $\boldsymbol{v}$ 相连的节点集合，并将邻

居节点的特征向量进行累积求和。然后，将累积特征向量$\boldsymbol{F}_u$与连接边的信息$\boldsymbol{G}_{u,v}$进行异或操作，并通过哈希函数进行映射，最终得到摩根指纹。摩根指纹计算公式

$$\boldsymbol{F}_v = \mathrm{Hash}(\boldsymbol{F}_u \oplus \boldsymbol{G}_{u,v}), \tag{5}$$

1.3.2　分子图卷积指纹　分子图卷积指纹基于分子拓扑结构进行分子特征提取，分子和化学键的特征基于原子符号、相邻原子、相邻氢原子、隐含价、芳香性以及化学键类型等进行编码。具体如表1所示。

表1　分子特征向量构成

Table 1　Molecular feature vector composition

分子结构	特征名	编码描述
原子	原子符号	44维向量
	相邻原子	6维向量
	相邻氢原子	5维向量
	隐含价	6维向量
	芳香性	1维向量
键	化学键类型	5维向量

对分子图进行图卷积操作

$$\boldsymbol{H}_{\mathrm{m}}^{j} = \sigma_{\mathrm{m}}\left(\sum_{u \in R_v} \frac{1}{c_{u,v}} \boldsymbol{W}_{\mathrm{m}}^{j} \boldsymbol{H}_{u}^{j-1} + \boldsymbol{b}_{\mathrm{m}}^{j}\right), \tag{6}$$

其中$\boldsymbol{H}_{\mathrm{m}}^{j}$是经过$j$次图卷积操作后节点$\boldsymbol{v}$的特征向量，$\boldsymbol{H}_{u}^{j-1}$是$j-1$层节点$\boldsymbol{v}$的邻居节点$\boldsymbol{u} \in \boldsymbol{R}_v$的特征向量，$\boldsymbol{W}_{\mathrm{m}}^{j}$和$\boldsymbol{b}_{\mathrm{m}}^{j}$是第$j$层的权重矩阵和偏置项，$c_{u,v}$是归一化常数，$\sigma_{\mathrm{m}}$是激活函数。

1.4　预测模型

1.4.1　SVM　支持向量机(SVM，support vector machine)是一种常用的监督学习算法，其基本原理是寻找一个最优的超平面，将样本空间分成两个不同类别，并最大化样本与超平面之间的间隔。对每一个样本数据，SVM决策函数

$$g(\boldsymbol{x})_{\mathrm{SVM}} = \mathrm{sign}\left(\boldsymbol{W}_{\mathrm{SVM}}^{\mathrm{T}} \boldsymbol{x} + \boldsymbol{b}_{\mathrm{SVM}}\right), \tag{7}$$

其中$\boldsymbol{x}$是输入样本特征向量，$\boldsymbol{W}_{\mathrm{SVM}}$是决策函数的权重矩阵，$\boldsymbol{b}_{\mathrm{SVM}}$是偏置项，sign是符号函数。

1.4.2　ELM　极限学习机(ELM，extreme learning machine)通过随机初始化输入层和输出层之间的权重，然后利用解析解的方式直接计算隐藏层的权重。这使得ELM能够快速地训练神经网络，并在很短的时间内生成准确的预测结果。对每一个样本数据，ELM决策函数

$$g(\boldsymbol{x})_{\mathrm{ELM}} = \mathrm{sign}\left(\boldsymbol{H}_{\mathrm{ELM}}(\boldsymbol{x})\boldsymbol{W}_{\mathrm{ELM}} + \boldsymbol{b}_{\mathrm{ELM}}\right), \tag{8}$$

其中$\boldsymbol{x}$是输入样本特征向量，$\boldsymbol{H}_{\mathrm{ELM}}(\boldsymbol{x})$是基于输入特征计算得到的隐藏层输出矩阵，$\boldsymbol{W}_{\mathrm{ELM}}$是输出层到隐藏层的权重矩阵，$\boldsymbol{b}_{\mathrm{ELM}}$是偏置项。

1.4.3　XGBoost　XGBoost是一种基于梯度提升树的集成学习算法。它通过迭代训练多个弱分类器(通常是决策树)，并将它们组合成一个强大的模型。对全部N个样本数据，XGBoost的目标函数

$$\mathrm{Obj}(\boldsymbol{\Phi}) = \sum_{n=1}^{N}\left[\mathrm{Loss}_{\mathrm{XGB}}\left(\boldsymbol{y}_n, \hat{\boldsymbol{y}}_n\right) + \Omega(\boldsymbol{\Phi})\right] + \gamma Q, \tag{9}$$

其中$\mathrm{Loss}_{\mathrm{XGB}}\left(\boldsymbol{y}_n, \hat{\boldsymbol{y}}_n\right)$是第$n$个样本的损失函数，$\boldsymbol{y}_n$是样本$n$的标签，$\hat{\boldsymbol{y}}_n$是样本$n$的预测值，$\Omega(\boldsymbol{\Phi})$表示模型中的每个子模型的正则化项，$Q$是决策树的个数，$\gamma$是正则化系数。

1.4.4　BP神经网络　BP(back propagation)人工神经网络模型基于反向传播算法，通过不断调整网络中连接权重和偏置，使网络能够学习输入与输出之间的高维非线性映射关系。

BP神经网络的标准前向传播公式为

$$\boldsymbol{Y}_{\mathrm{BP}}^{k} = \sigma_{\mathrm{BP}}(\boldsymbol{W}_{\mathrm{BP}}^{k} \cdot \boldsymbol{Y}_{\mathrm{BP}}^{k-1} + \boldsymbol{B}_{\mathrm{BP}}^{k}), \tag{10}$$

其中$\boldsymbol{Y}_{\mathrm{BP}}^{k}$是第$k$层的神经元输出矩阵，$\boldsymbol{Y}_{\mathrm{BP}}^{k-1}$是第$(k-1)$层的神经元输出矩阵，$\boldsymbol{W}_{\mathrm{BP}}^{k}$是第$k$层的权重矩阵，$\boldsymbol{B}_{\mathrm{BP}}^{k}$是第$k$层偏置项，$\sigma_{\mathrm{BP}}$是激活函数。

通过反向传播算法，BP神经网络根据误差信号从输出层反向传播到隐藏层，利用梯度下降法不断调整连接权重和偏置，以最小化损失函数，使得网络输出$\hat{\boldsymbol{Y}}_{\mathrm{BP}}$与真实标签$\boldsymbol{Y}_{\mathrm{BP}}$之间的差距尽可能小。训练过程通过不断迭代更新参数来提高模型的预测性能。损失函数

$$\mathrm{Loss}_{\mathrm{BP}} = \frac{1}{2P}(\boldsymbol{Y}_{\mathrm{BP}} - \hat{\boldsymbol{Y}}_{\mathrm{BP}})^2, \tag{11}$$

其中P表示训练样本的个数。

1.5　SVD-PCA

基于奇异值分解(SVD，singular value decomposition)的主成分分析(PCA，principal component analysis)是一种常用的降维技术。SVD-PCA的优点是可以处理高维数据，并且对异常值具有较好的鲁棒性。

给定一个数据矩阵$\boldsymbol{X}_{\mathrm{SP}}$，首先对$\boldsymbol{X}_{\mathrm{SP}}$进行标准化处理获得矩阵$\boldsymbol{X}'_{\mathrm{SP}}$，使得每个特征均值为0，方差为1。然后，对标准化后的数据矩阵进行SVD分解

$$\boldsymbol{X}'_{\mathrm{SP}} = \boldsymbol{COS}^{\mathrm{T}}, \tag{12}$$

其中$\boldsymbol{C}$和$\boldsymbol{O}$是由SVD计算得到的矩阵，$\boldsymbol{S}$是由SVD

得到的正交矩阵。

PCA 的结果是通过选择奇异值及其对应的左奇异向量来进行降维。主成分矩阵可以通过以下公式计算得到

$$\boldsymbol{Z} = \boldsymbol{X}'_{\mathrm{SP}}\boldsymbol{S}\ , \tag{13}$$

其中 $\boldsymbol{Z}$ 是降维后的数据矩阵。

2 实验过程

2.1 实验数据集

2.1.1 数据库 1：气味分子–嗅觉受体激活关系数据库 本文基于现有发表文献建立气味分子–嗅觉受体激活关系数据库，所有数据都来自于截至在 2023 年 7 月之前 Web of Science 数据库中收录的文献。数据库共收集了 43 个人类嗅觉受体，以及它们对选定的 170 个化合物的 254 条激活关系和 61 条非激活关系数据。

2.1.2 数据库 2：气味分子–气味感知关系数据库 DREAM 数据集使用包括强度、愉悦度和熟悉度在内的 23 个感知定义气味感知。数据集包括 49 名健康参与者（没有专业气味感知训练）对 476 种气味分子产生的 21 种气味感知数据，评分范围为 0~100。本研究选用标记为“高浓度”的数据共 405 条。

2.2 嗅觉受体激活预测模型训练

嗅觉受体激活预测 XGBoost 模型参数设置如表 2 所示。

表 2 XGBoost 模型参数调节范围[1)]

Table 2 Parameter adjustment range of XGBoost model

训练参数	参数值
学习率	[**0.3** 0.4 0.5]
随机种子	[5 10 **20**]
迭代次数	[100 500 **1 000**]

1）加粗数据为最终选定参数。

模型评价指标选取准确率（accuracy）、F_1-score、受试者工作特征（ROC，receiver operating characteristic curve）的曲线下面积（AUC，area under the curve）。

2.3 气味感知预测模型训练

气味感知预测模型训练采用 5 折交叉验证，即将数据划分为大致相等的 5 个子数据集，依次采用不同数据集作为训练集和测试集。取 5 次训练平均精度的平均值即得到模型精度，这样得到的模型精度更具有泛化性。

气味感知预测 BP 模型参数设置如表 3 所示。

表 3 BP 模型参数调节范围[1)]

Table 3 Parameter adjustment range of BP model

训练参数	参数值
隐藏层神经元个数	[25 **40** 55 100 200]
学习率	[0.001 0.005 **0.009** 1]
迭代次数	[50 **100** 150 200 400]
批量大小	[**1** 2 3]

1）加粗数据为最终选定参数。

模型评价指标选取 R^2-score、皮尔逊相关性、均方根误差（RMSE，root mean square erro）。

3 实验结果与分析

3.1 嗅觉受体蛋白质关系网络

本研究使用嗅觉受体蛋白质关系网络中 100% 的相似度、前 70% 的相似度、前 50% 相似度网络关系，获取相关网络性质指标，并使用基于模块度的社区发现算法分析网络的模块性（Blondel et al.，2008）。分析如表 4 所示。本研究基于相似度排名前 50% 的数据绘制出嗅觉受体蛋白质关系网络图（图 2）。使用相似度排名前 50% 的网络呈现出明显的 3 个子模块，且不存在孤立节点。属于同一模块的嗅觉受体具有相似的蛋白质结构，比如，图 2 中嗅觉受体 OR2J3 与 OR2J2 同属于一个社区模块，同时，研究也证实它们是人类嗅觉受体中最为相似的嗅觉受体对之一（Crasto et al.，2002）。

3.2 基于不同特征提取方式的嗅觉受体激活预测结果比较

分子的表征方式在化学领域中尚未形成统一的标准，不同的表征方法各具优势和局限性。本文对气味分子和嗅觉受体蛋白分别采用了两种不同的特征提取方法，并进行对比实验。结果如表 5 所示。结果表明，当分别使用图卷积进行分子特征和嗅觉受体蛋白氨基酸序列特征提取时，采用 XGBoost 算法实现了最佳的嗅觉受体激活预测效果，准确率为 77%，F_1-score 为 0.78，AUC 值为 0.77。4 种特征提取方式 AUC 比较结果如图 3 所示。

3.3 基于不同分类器的嗅觉受体激活预测模型比较

基于图卷积特征提取，本文采用 XGBoost、

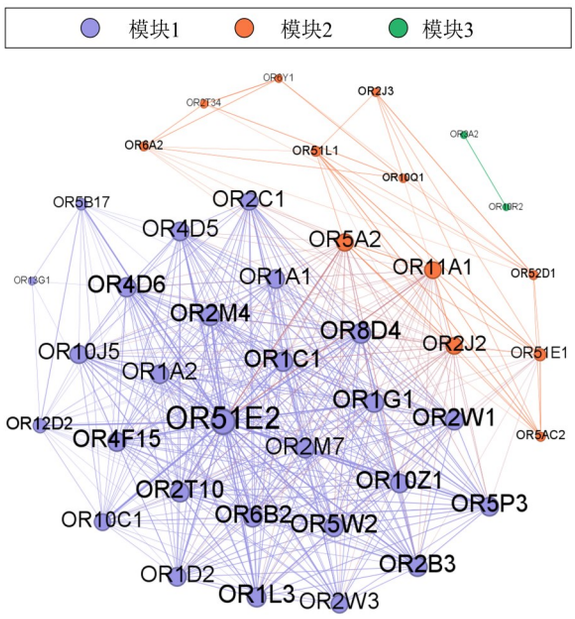

图2　嗅觉受体蛋白质关系网络(前 50%)

Fig. 2　Olfactory receptor protein relationship network (Top 50%)

SVM 以及 ELM 3 种机器学习方法进行嗅觉受体激活预测，并进行对比实验，结果如表 6 所示。实验结果表明，XGBoost 算法在气味分子-嗅觉受体激活关系数据库上表现结果最优，准确率为 77%，

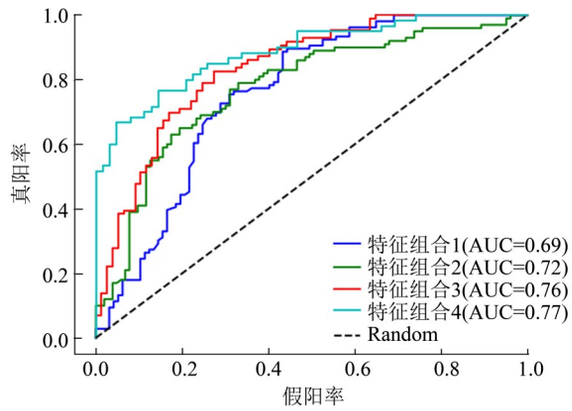

图3　不同特征提取方式组合的ROC曲线及AUC值

Fig. 3　ROC curves and AUC values of different feature extraction methods

F_1-score 为 0.78，AUC 为 0.77。3 种分类器的嗅觉受体激活预测模型 AUC 比较结果如图 4 所示。

3.4　气味感知预测结果比较

本研究在嗅觉受体激活预测模型的基础上，对 DREAM 数据集中的化合物与 43 个嗅觉受体的激活关系进行预测，将获得的新气味分子-嗅觉受体激活关系作为分子特征应用于气味感知预测模型。在数据集和回归预测模型相同的情况下，引入气味分子-嗅觉受体激活关系进行气味感知预测

表4　嗅觉受体蛋白质关系网络概览

Table 4　Network overview of olfactory receptor protein relationship

网络参数	相似度/%		
	前 50	前 75	100
平均度数	20. 98	30. 28	40. 51
平均加权度	17. 19	22. 02	25. 47
社区数量	3	3	3
模块度	0. 11	0. 08	0. 05

表5　不同分子特征提取方式组合在数据库 1 上的准确率、F_1-score 和 AUC

Table 5　Accuracy, F_1-score, and AUC of different feature extraction methods for database 1

特征组合（嗅觉受体特征&分子特征）	训练集			验证集		
	准确率/%	F_1-score	AUC	准确率/%	F_1-score	AUC
特征组合 1（One-hot 编码&摩根指纹）	100	1	1	68	0. 72	0. 68
特征组合 2（One-hot 编码&图卷积分子指纹）	100	1	1	72	0. 73	0. 72
特征线组合 3（图卷积&摩根指纹）	99	0. 99	0. 99	76	0. 77	0. 76
特征组合 4（图卷积&图卷积分子指纹）	100	1	1	77	0. 78	0. 77

表 6 不同分类器的嗅觉受体激活预测模型在数据库 1 的准确率、F_1-score 和 AUC

Table 6 Accuracy, F_1-score and AUC of olfactory receptor activation prediction models for different classifiers on database 1

特征组合	训练集			验证集		
	准确率/%	F_1-score	AUC	准确率/%	F_1-score	AUC
XGBoost	100	1	1	77	0. 78	0. 77
ELM	58	0. 67	0. 57	51	0. 57	0. 55
SVM	70	0. 72	0. 7	48	0. 48	0. 49

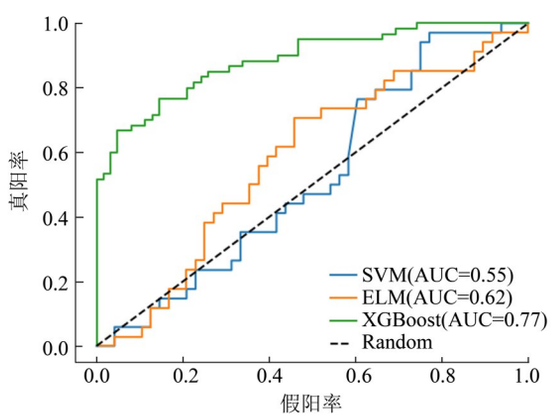

图 4 不同分类器的嗅觉受体激活预测模型的 ROC 曲线和 AUC 值

Fig. 4 ROC curves and AUC values of olfactory receptor activation prediction models for different classifiers

结果明显优于仅基于分子结构进行气味感知预测。实验结果说明在进行气味感知预测时，考虑嗅觉受体的活动情况是必要的。实验结果如表 7 所示。

在 3.3 节中，对于嗅觉受体激活预测任务，图卷积特征提取方法明显优于摩根指纹特征提取。然而，在本节的气味感知预测任务中，摩根指纹方法表现更优。这是由于图卷积方法和摩根指纹方法对分子特征表达方式不同造成的。图卷积方法基于图结构进行特征提取，考虑了原子之间的连接关系，在捕捉分子的全局信息上具有优势。而摩根指纹根据分子的物理化学性质进行有效编码，更擅长总结分子的理化特征（Cereto-Massagué et al.，2015；Duvenaud et al.，2015；Kipf et al.，2016）。

表 7 不同特征提取方式在 DREAM 数据集上的 R^2-score、r 和 RMSE

Table 7 R^2-score, r and RMSE on the DREAM dataset with different feature extraction methods

特征提取方式	训练集			验证集		
	R^2-score	r	RMSE	R^2-score	r	RMSE
摩根指纹	0. 91	0. 96	0. 05	0. 87	0. 93	0. 06
分子图卷积指纹	0. 79	0. 89	0. 07	0. 79	0. 89	0. 07
摩根指纹&激活关系	0. 92	0. 96	0. 04	0. 87	0. 94	0. 05
分子图卷积指纹&激活关系	0. 82	0. 91	0. 07	0. 81	0. 90	0. 07

3. 5 气味分子-嗅觉受体激活-气味感知模式

本研究通过嗅觉受体蛋白质关系网络，整合了 DREAM 数据集和气味分子-嗅觉受体激活关系信息。采用基于奇异值分解的主成分分析方法对嗅觉受体在特定气味感知中的贡献进行分析。嗅觉受体对 21 种气味感知的贡献度归一化后的结果如图 5 所示。大部分嗅觉受体会对特定气味感知产生较高的贡献度（Audouze et al.，2014）。

此外，本研究采用密度聚类算法（Campello et al.，2020），对来自 DREAM 数据集的 405 个气味分子的 43 个嗅觉受体激活特征进行聚类，将分子分为 4 个类别，并绘制了气味分子-嗅觉受体激活-气味感知模式图。如图 6 所示，产生激活关系少于 20 条的嗅觉受体并没有被绘制，DREAM 数据集中气味感知评分低于 5 分的气味感知描述词没有被绘制。

研究结果表明，经由气味分子-嗅觉受体激活关系对分子进行分类在气味感知上出现了明显的模式上的不同。例如，“腐烂（decayed）”只与第 1 类分子激活的 3 个嗅觉受体相连；“花（flower）”只与第 4 类分子激活的 4 个嗅觉受体相连等，本研究部分结果与已得到的生物实验结果验证一致（Chaput et al.，2012；El Mountassir et al.，2016；Keller et al.，2016）。本研究同时尝试了使用 SMILES 分子表达式和摩根指纹对分子进行聚类，所获得的结果难以提取出明显的气味分子-嗅觉受体激活-气味感知模式。

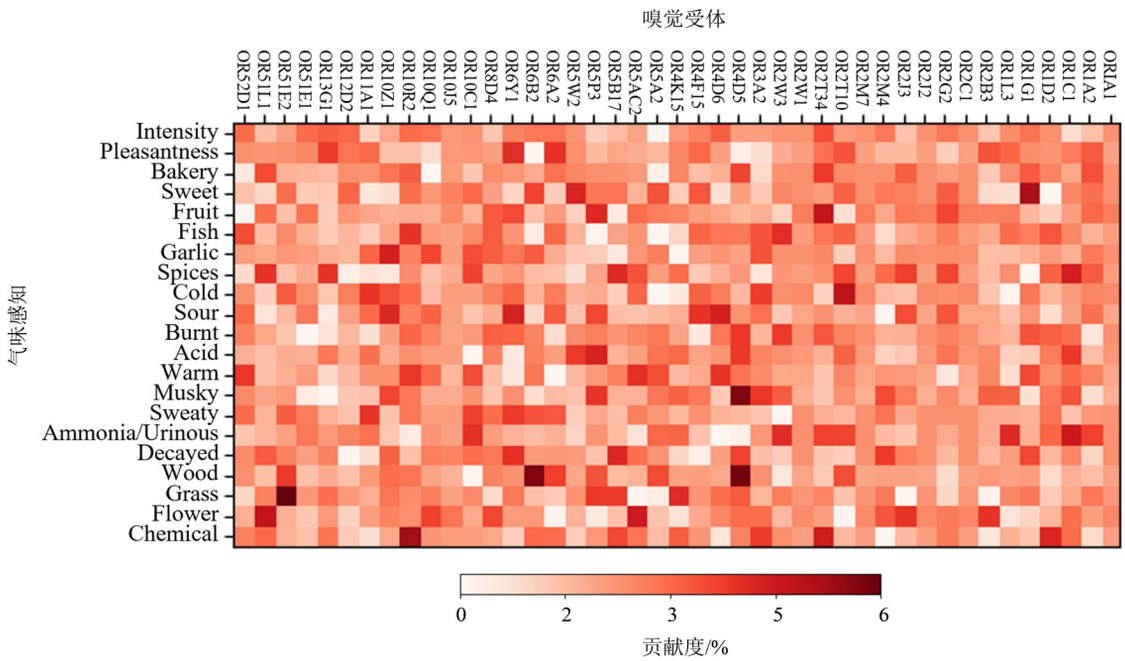

图 5　嗅觉受体对气味感知贡献度

Fig. 5　Olfactory receptor contribution to olfactory perception

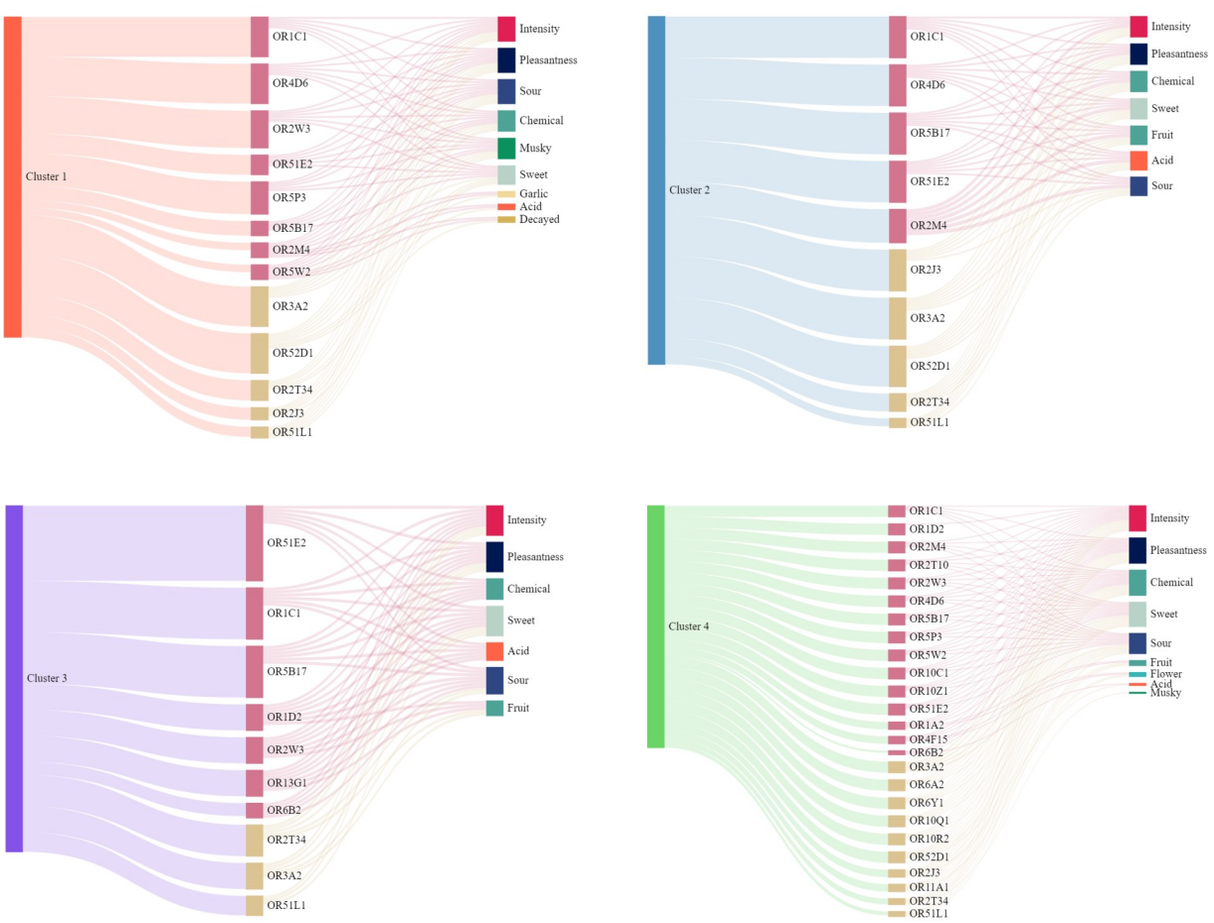

图 6　气味分子–嗅觉受体激活–气味感知模式

Fig. 6　Odor molecule-olfactory receptor activation-olfactory pattern

4 结 语

本研究旨在提出一种基于数据驱动方法的气味感知预测和分析的新解决方案。首先，构建了嗅觉受体蛋白质关系网络，采用图卷积方法以获得更全面有效的嗅觉受体蛋白特征。在嗅觉受体激活关系数据的基础上，构建了嗅觉受体激活预测模型。其次，面向DREAM数据集并引入其嗅觉受体激活数据，以提供必要的生理信息补充，实现对气味分子感知的精准预测。最后，对模型形成的正确决策机制进行解释分析，并总结了气味分子-嗅觉受体激活-气味感知模式。研究结果表明，综合考虑气味分子特征和气味分子-嗅觉受体激活关系构建预测模型，能够获得更好的预测结果，并获得对人类气味感知模式的有效总结。

尽管研究结果仍需要进一步验证，但本研究为进一步探索和理解气味感知机制提供了有价值的参考和启示。未来的工作将面向更多的气味感知数据集进一步优化模型，基于数据驱动技术进一步学习气味分子与嗅觉受体激活的对接模型，为气味感知的预测提供更多有用的信息，进一步推进人类嗅觉机理研究。

参考文献：

AUDOUZE K，TROMELIN A，le BON A M，et al，2014. Identification of odorant-receptor interactions by global mapping of the human odorome[J]. PLoS One，9(4)：e93037.

BLONDEL V D，GUILLAUME J L，LAMBIOTTE R，et al，2008. Fast unfolding of communities in large networks[J]. J Stat Mech，2008(10)：P10008.

BRAUN T，VOLAND P，KUNZ L，et al，2007. Enterochromaffin cells of the human gut：Sensors for spices and odorants[J]. Gastroenterology，132(5)：1890-1901.

BUCK L B，2008. Olfactory receptors and odor coding in mammals[J]. Nutr Rev，62：S184-S188.

CAMPELLO R J G B，KRÖGER P，SANDER J，et al，2020. Density-based clustering[J]. Wiley Interdiscip Rev Data Min Knowl Discov，10(2)：e1343.

CERETO-MASSAGUÉ A，OJEDA M J，VALLS C，et al，2015. Molecular fingerprint similarity search in virtual screening[J]. Methods，71：58-63.

CHAPUT M A，EL MOUNTASSIR F，ATANASOVA B，et al，2012. Interactions of odorants with olfactory receptors and receptor neurons match the perceptual dynamics observed for woody and fruity odorant mixtures[J]. Eur J Neurosci，35(4)：584-597.

COOK B L，STEUERWALD D，KAISER L，et al，2009. Large-scale production and study of a synthetic G protein-coupled receptor：Human olfactory receptor 17-4[J]. Proc Natl Acad Sci USA，106(29)：11925-11930.

CRASTO C，MARENCO L，MILLER P，et al，2002. Olfactory Receptor Database：A metadata-driven automated population from sources of gene and protein sequences[J]. Nucleic Acids Res，30(1)：354-360.

DEBNATH T，NAKAMOTO T，2020. Predicting human odor perception represented by continuous values from mass spectra of essential oils resembling chemical mixtures[J]. PLoS One，15(6)：e0234688.

DUVENAUD D K，MACLAURIN D，AGUILERA-IPARRAGUIRRE J，et al，2015. Convolutional networks on graphs for learning molecular fingerprints[J/OL]. arXiv：1509.09292v2.

EL MOUNTASSIR F，BELLOIR C，BRIAND L，et al，2016. Encoding odorant mixtures by human olfactory receptors[J]. Flavour Fragr J，31(5)：400-407.

FRANCIA S，LODOVICHI C，2021. The role of the odorant receptors in the formation of the sensory map[J]. BMC Biol，19(1)：174.

FUJITA Y，TAKAHASHI T，SUZUKI A，et al，2007. Deorphanization of Dresden G protein-coupled receptor for an odorant receptor[J]. J Recept Signal Transduct，27(4)：323-334.

JACQUIER V，PICK H，VOGEL H，2006. Characterization of an extended receptive ligand repertoire of the human olfactory receptor OR17-40 comprising structurally related compounds[J]. J Neurochem，97(2)：537-544.

JAEGER S，McRAE J，BAVA C，et al，2013. A Mendelian trait for olfactory sensitivity affects odor experience and food selection[J]. Curr Biol，23(16)：1601-1605.

KASYAP V L V S K B，BHAGAVAN V S，JAGADEESH M S，2022. Graph neural networks based model for aroma prediction using molecular structures[C]//IEEE 3rd GCAT，Bangalore，India：1-6.

KELLER A，GERKIN R C，GUAN Y，et al，2017. Predicting human olfactory perception from chemical features of odor molecules[J]. Science，355(6327)：820-826.

KELLER A，VOSSHALL L B，2016. Olfactory perception of

chemically diverse molecules[J]. BMC Neurosci, 17(1): 1-17.

KELLER A, ZHUANG H, CHI Q, et al, 2007. Genetic variation in a human odorant receptor alters odour perception [J]. Nature, 449(7161): 468-472.

KIPF T N, WELLING M, 2016. Semi-supervised classification with graph convolutional networks [EB/OL]. arXiv: 1609.02907.

KOWALEWSKI J, RAY A, 2020. Predicting human olfactory perception from activities of odorant receptors [J]. iScience, 23(8): 101361.

LAPID H, SHUSHAN S, PLOTKIN A, et al, 2011. Neural activity at the human olfactory epithelium reflects olfactory perception[J].Nat Neurosci, 14(11): 1455-1461.

LI H, PANWAR B, OMENN G S, et al, 2018. Accurate prediction of personalized olfactory perception from large-scale chemoinformatic features[J]. Gigascience, 7(2): gix127.

MAJID A, KRUSPE N, 2018. Hunter-gatherer olfaction is special[J]. Curr Biol, 28(3): 409-413.

MATARAZZO V, CLOT-FAYBESSE O, MARCET B, et al, 2005. Functional characterization of two human olfactory receptors expressed in the baculovirus Sf9 insect cell system[J]. Chem Senses, 30(3): 195-207.

MENASHE I, ABAFFY T, HASIN Y, et al, 2007. Genetic elucidation of human hyperosmia to isovaleric acid [J]. PLoS Biol, 5(11): e284.

NEUHAUS E M, MASHUKOVA A, ZHANG W, et al, 2006. A specific heat shock protein enhances the expression of mammalian olfactory receptor proteins[J]. Chem Senses, 31(5): 445-452.

PATTANAIK L, COLEY C W, 2020. Molecular representation: Going long on fingerprints[J]. Chem, 6(6): 1204-1207.

SAITO H, CHI Q, ZHUANG H, et al, 2009. Odor coding by a Mammalian receptor repertoire[J]. Sci Signal, 2(60): ra9.

SCHMIEDEBERG K, SHIROKOVA E, WEBER H P, et al, 2007. Structural determinants of odorant recognition by the human olfactory receptors OR1A1 and OR1A2[J]. J Struct Biol, 159(3): 400-412.

SHANG L, LIU C, TOMIURA Y, et al, 2017. Machine-learning-based olfactometer: Prediction of odor perception from physicochemical features of odorant molecules [J]. Anal Chem, 89:11999 - 12005.

SHIRASU M, YOSHIKAWA K, TAKAI Y, et al, 2014. Olfactory receptor and neural pathway responsible for highly selective sensing of musk odors[J]. Neuron, 81(1): 165-178.

TOPIN J, de MARCH C A, CHARLIER L, et al, 2014. Discrimination between olfactory receptor agonists and non-agonists[J]. Chem - A Eur J, 20(33): 10227-10230.

VASSAR R, NGAI J, AXEL R, 1993. Spatial segregation of odorant receptor expression in the mammalian olfactory epithelium[J]. Cell, 74(2): 309-318.

XU J, ZHANG Y, 2010. How significant is a protein structure similarity with TM-score = 0.5? [J]. Bioinformatics, 26(7): 889-895.

ZHANG Y, SKOLNICK J, 2004. Scoring function for automated assessment of protein structure template quality [J]. Proteins Struct Funct Bioinform, 57(4): 702-710.

（责任编辑　张　冰）