

基于系统发育排序的多尺度卷积特征融合模型 用于代谢谱预测

冀庆荣¹, 石凯^{1,2}

(1. 桂林理工大学 计算机科学与工程学院;

2. 桂林理工大学 广西嵌入式技术与智能系统重点实验室, 广西 桂林 541004)

摘要: 代谢组是微生物组调控宿主健康的重要中介, 其浓度变化不仅会影响宿主生理功能, 还关系微生物群落稳态。代谢组学实验的高成本与低覆盖度导致代谢谱数据稀缺, 而微生物组测序技术可高效、经济地获取群落组成数据, 使基于微生物组的数据推断代谢谱成为研究热点。针对现有代谢谱预测方法普遍忽视微生物间的系统发育关系, 从而限制了模型预测能力这一问题, 提出一种基于系统发育排序的多尺度卷积特征融合模型 PCMMP。首先, 通过系统发育树后序遍历、卷积核引导的特征重排策略来提取局部、全局微生物的多尺度关联信息; 其次, 构建一维卷积神经网络(1D-CNN)与多层感知机(MLP)的混合架构实现代谢谱预测。5个公开数据集与7种先进方法比较发现, PCMMP在多个关键指标上表现优越, 特别是在IBD独立验证集上所有代谢物的Spearman均值(ALL_SCC)由0.249提升至0.291, 良好预测的代谢物数量(WP_Nums)从3 669增至4 263, 相较于mNODE分别提升了16.9%、16.2%。

关键词: 微生物组; 代谢谱预测; 系统发育信息; 卷积神经网络

DOI: 10.11907/rjdk.251092

扫描二维码阅读全文:



中图分类号: TP301

文献标识码: A

文章编号: 1672-7800(2026)005-0050-09

Phylogeny-Aware Multi-Scale Convolutional Feature Fusion Model for Metabolomic Prediction

Ji Qingrong¹, Shi Kai^{1,2}

(1. College of Computer Science and Engineering;

2. Guangxi Key Laboratory of Embedded Technology and Intelligent System, Guilin University of Technology, Guilin 541004, China)

Abstract: Metabolome is an important mediator in the regulation of host health by the microbiome, and its concentration changes not only affect host physiological functions, but also relate to the stability of microbial communities. The high cost and low coverage of metabolomics experiments can lead to a scarcity of metabolic profile data, while microbiome sequencing technology can efficiently and economically obtain community composition data, making the inference of metabolic profiles based on microbiome data a research hotspot. A multi-scale convolutional feature fusion model PCMMP based on phylogenetic sorting is proposed to address the problem that existing metabolic spectrum prediction methods generally ignore the phylogenetic relationships between microorganisms, which limits the predictive ability of the model. Firstly, the multi-scale correlation information of local and global microorganisms is extracted through a feature rearrangement strategy guided by the post traversal of phylogenetic trees and convolutional kernels; Secondly, a hybrid architecture of one-dimensional convolutional neural network (1D-CNN) and multi-layer perceptron (MLP) is constructed to achieve spectral prediction. Comparing 5 publicly available datasets with 7 advanced methods, PCMMP showed superior performance in multiple key indicators, especially in the IBD independent validation set where the Spearman mean (ALL_SCC) of all metabolites increased from 0.249 to 0.291, and the number of well predicted metabolites (WP_Cums) increased from 3 669 to 4 263, which were 16.9% and 16.2% higher than mNNE, respectively.

Key Words: microbiome; metabolomic prediction; phylogenetic information; convolutional neural network

收稿日期: 2025-03-18

基金项目: 国家自然科学基金项目(62162019); 中央引导地方科技发展专项资金项目(ZY22096025); 广西研究生教育创新项目(YC-SW2024357)

作者简介: 冀庆荣(2000-), 男, 桂林理工大学计算机科学与工程学院硕士研究生, 研究方向为机器学习、生物信息学; 石凯(1980-), 男, 博士, 桂林理工大学计算机科学与工程学院副教授、硕士生导师, 研究方向为机器学习、计算系统生物学。本文通讯作者: 石凯。

0 引言

人类微生物组是指定植于人体内外不同生境(肠道、口腔、皮肤等)的微生物群落及其基因组成,在维持宿主生理稳态、调控免疫系统及介导营养代谢等方面发挥着重要作用^[1]。随着高通量测序技术的快速发展,微生物组研究已成为揭示宿主健康与疾病发生机制的重要方向。大量研究表明,微生物群落结构的失调与多种复杂疾病密切相关,包括炎症性肠病(Inflammatory Bowel Disease, IBD)、2型糖尿病(Type 2 Diabetes, T2D)和结直肠癌(Colorectal Cancer, CRC)等^[2-4]。

从病理生理学角度而言,疾病发生与进展并非由微生物群落组成直接决定,而是其代谢活动所产生的代谢物与宿主之间的相互作用起到了关键作用^[5]。例如,短链脂肪酸(Short-Chain Fatty Acids, SCFAs)是肠道微生物发酵膳食纤维的代谢产物,能调节宿主免疫稳态,维持肠道屏障功能,还具有抗炎作用^[6]。此外,吲哚衍生物(Indole Derivatives, ID)作为色氨酸代谢的产物,在慢性炎症性疾病中可通过调节芳香烃受体(Aryl hydrocarbon Receptor, AhR)信号通路,影响炎症反应和肠道稳态^[7]。实例表明,相较于单纯的微生物组成数据,代谢谱(代谢物种类及丰度数据)能直接反映微生物—宿主互作的生物学功能,在疾病预测等领域应用潜力巨大^[8]。

代谢组学作为专门研究生物体内代谢物及其动态变化的学科,提供了多种方法来测量、分析代谢谱,推动了个性化医疗精准化的进程。例如,基于血浆代谢组学的机器学习模型诊断胃癌的敏感性>90%,可有效将患者分为不同的风险组进行精准干预^[9];餐后代谢谱中的烯丙基苯酚硫酸盐和辛酸显著提升了糖尿病前期风险分层效能(AUC=0.80)^[10]。药物代谢组学通过分析给药前的代谢谱,可预测个体对特定药物的疗效与毒性风险,为剂量调整提供依据,表明了精准解析、量化代谢谱是实施个性化医疗的核心前提^[11]。

由于代谢组学实验成本高昂、代谢物覆盖度有限,且代谢谱数据仍然较为稀缺,导致其应用受限。微生物组高通量测序技术以其高效性、经济性,已广泛应用于微生物群落组成分析,这一技术代差使得从微生物组数据中预测代谢谱成为可能^[12-15]。为此,学者们基于现有的高质量微生物组—代谢组配对数据,提出一系列代谢谱预测方法,不仅有助于揭示微生物—代谢物—宿主之间的潜在关联,还为疾病早期预测和个体化健康监测提供了数据支撑。

1 相关工作

研究者们基于现有的高质量微生物组—代谢组配对数据,提出一系列数据驱动方法,利用微生物组数据推测

新样本的代谢谱,主要分为传统机器学习方法和深度学习方法。传统机器学习方法通常针对每个代谢物单独训练回归模型,以构建微生物与代谢物之间的映射关系。Mallik等^[16]提出计算框架 MelonnPan,结合弹性网络回归与特定环境的宏基因组数据训练多个回归模型,并通过迭代优化微生物特征选择来提升预测精度。在此基础上,Xie等^[17]提出 ENVIM 模型,引入随机森林权重分配策略,通过权重排名筛选最佳分类群以提升代谢谱的预测精度。Muller等^[18]采用随机森林回归与随机效应模型,研究不同个体和环境下微生物—代谢物关联的稳定性,以分析关键分类群对代谢物可预测性的贡献。Tang等^[19]基于 O2PLS 算法提出 MMNP 模型,通过迭代筛选预测性较强的代谢特征,以提升模型捕捉微生物—代谢物相互作用的能力。然而,传统方法往往难以充分挖掘微生物与代谢物之间的非线性、高维复杂关系。

近年来,深度学习方法正逐步应用于代谢谱预测任务。Le等^[20]提出的稀疏编码器 Sparse NED,利用 IBD 患者微生物—代谢物配对数据以学习潜在特征表示,精准预测 IBD 相关代谢物的丰度及其疾病状态。Reiman等^[21]开发了基于多层感知机(Multilayer Perceptron, MLP)的 Mi-MeNet,在模拟群落代谢谱时,通过功能模块聚类分析微生物—代谢物间的潜在关联。Shtossel等^[22]结合奇异值分解(Singular Value Decomposition, SVD)与低秩近似技术,解析微生物群落组成对代谢物浓度的影响,在疾病预测任务中表现突出。Wang等^[23]开发的 mNODE(基于神经常微分方程 NODE)模型,结合饮食信息来提升代谢谱的预测能力,探索了微生物组—饮食—代谢组之间的复杂相互作用。然而,上述方法普遍忽略了微生物间的系统发育关系,可能会限制模型的生物学可解释性,并影响代谢谱预测性能。

目前,系统发育树作为描述微生物亲缘关系的重要工具,已被广泛应用于微生物—宿主关联研究,用来增强疾病预测的准确性,但在微生物组驱动的代谢谱预测任务中,其应用仍然有限^[24]。研究表明,进化关系相近的微生物通常具有相似的代谢功能,由于水平基因转移和趋同进化等机制,不同进化支系的微生物也可能具备相同的代谢能力,存在微生物功能冗余现象^[25]。

为此,本文提出一种基于系统发育排序的多尺度卷积特征融合模型系统(Phylogeny-Aware Convolutional Network for Microbiome-based Metabolomic Prediction, PC-MMP)(代码网站为 <https://github.com/GUT-ShiLab/PC-MMP>),将微生物的系统发育信息整合至代谢谱预测任务中。首先,通过卷积神经网络(Convolutional Neural Network, CNN)学习微生物的局部特征和全局特征信息;其次,将学习到的特征表示输入 MLP 进行代谢谱预测。主要贡献为:①将系统发育信息融入代谢谱预测任务,通过顺序遍历对微生物数据进行排序,以保留近支(局部)关联信

息;②引入全局特征建模策略,通过调整卷积核来引导特征顺序,以捕捉系统发育远支(全局)微生物之间的特征关联,从而提升模型预测性能;③创新性地代谢谱预测任务中引入CNN,结合多尺度特征融合策略高效整合远支与近支信息,以显著提升模型泛化能力和预测精度。

2 代谢谱预测方法设计

2.1 模型整体框架

PCMMP是一个基于系统发育排序的多尺度卷积特征融合模型,主要包括:基于系统发育排序的多尺度特征构

建、基于卷积网络的特征学习与代谢丰度预测,旨在从微生物组数据中预测代谢谱。

首先,PCMMP基于系统发育树对微生物特征进行后序遍历排序,确保系统发育近支微生物在特征空间中相邻排列,保留其在代谢功能上的相似性;其次,为了增强远支微生物之间的特征关联建模,PCMMP在保持系统发育结构的基础上调整特征排列方式,使CNN能学习局部(近支)和全局(远支)特征,从而提升模型对多样化代谢模式的建模能力;最后,将远支与近支特征融合后输入CNN进行多尺度特征提取,通过多层感知机预测代谢丰度。PCMMP的工作流程如图1所示。

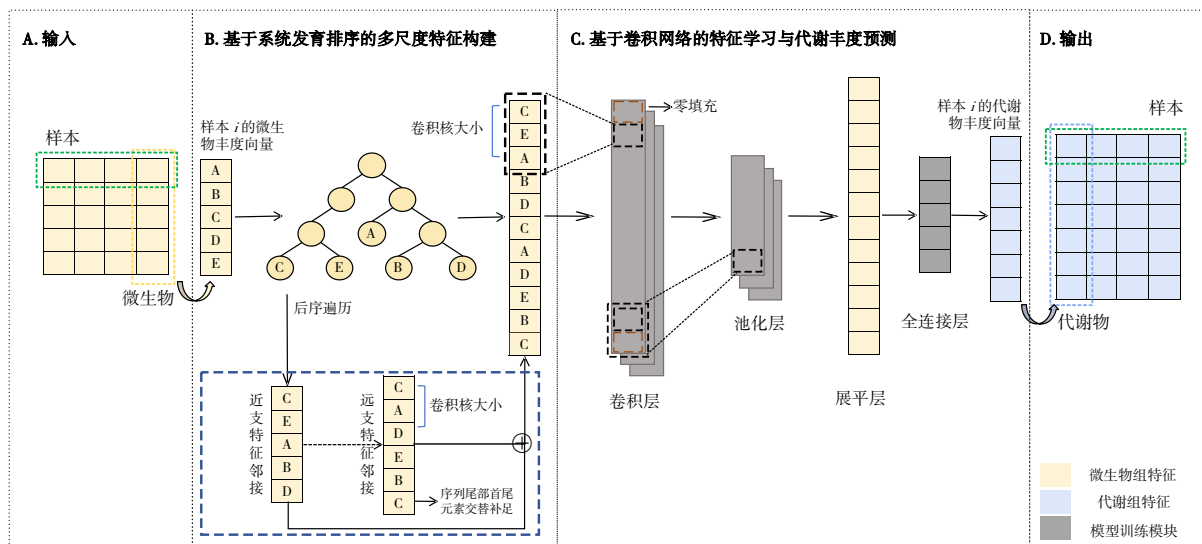


Fig. 1 PCMMP workflow

图1 PCMMP工作流程

2.2 基于系统发育排序的多尺度特征构建

系统发育信息在微生物组研究中具有重要的生物学价值,可解析微生物的进化关系及其功能作用。在系统发育树中,叶节点代表微生物分类单元,内部节点对应共同祖先,分支长度和层级则反映进化距离,距离越近的微生物分类单元往往具有相同的功能特征,并在与宿主相互作用的过程中发挥类似的作用^[26]。研究表明,在疾病预测任务中,利用系统发育信息进行特征重排或加权,可有效提升模型的生物学合理性和预测性能^[24,27]。

为此,本文在代谢谱预测任务中引入系统发育排序,以捕捉微生物之间的功能相似性,进而提升模型预测代谢谱的精度。然而,代谢谱预测涉及大量代谢物,受微生物群功能冗余的影响,只关注局部进化关系可能会导致特征表达同质化,忽略远支微生物在代谢调控中的潜在作用^[25]。因此,本文基于卷积核大小的间隔特征选择策略,使CNN在学习局部特征的同时能更有效捕捉远支微生物的全局信息。系统发育特征优化的具体步骤如图1所示。

(1)系统发育排序。基于NCBI Taxonomy数据库获取属层级微生物的NCBI ID,利用ETE3构建系统发育树进行后序遍历。如图1中B部分所示,微生物A-E经后序遍历

可生成特征序列“CEABD”,使进化相邻节点(如C-E属级分支)在特征矩阵中连续排列,强化了局部代谢关联。

(2)间隔特征选择。设定卷积核尺寸 k ,采用卷积核引导的等间隔特征选择策略,从后序遍历序列中提取全局关联特征。以 $k=3$ 为例(图1中B部分),由于卷积核尺寸与特征序列长度 $L=5$ 不匹配,采用尾部交替填充首尾元素的策略使特征序列长度能被卷积核整除,从而得到新的序列“CEABDC”。研究表明,该填充策略能最大限度保持系统发育间隔关系,避免冗余特征和无关噪声的干扰。随后,从“CEABDC”序列中等距选取第1、3、5位特征(C-A-D),形成远支特征组“CAD”,剩余特征按相同规则重排为“EBC”,最终生成全局特征序列“CADEBC”。

(3)特征融合。首先,将后序序列与全局特征序列拼接为“CEABDCADEBC”,通过局部卷积核(如“CEA”)捕获近支保守代谢模式;其次,利用跨分支窗口(如“CAD”)学习远支微生物的协同代谢机制;最后,将融合后的特征矩阵输入CNN进行端到端训练,通过层级特征抽象实现代谢谱的精准预测。

2.3 基于卷积网络的特征学习与代谢丰度预测

基于微生物组数据的代谢谱预测是一项典型的多元

回归任务,但弹性网络回归、随机森林及多层感知机等传统方法,既难以有效挖掘微生物间的拓扑关系,也难以捕捉代谢物之间的共享信息。由于微生物数据特征高维且样本量有限,过于复杂的模型(图神经网络等)易发生过拟合,从而限制模型泛化能力^[28]。为此,本文基于融合系统发育信息的微生物组数据,采用一维卷积神经网络(1D Convolutional Neural Networks, 1D-CNN)进行特征学习,以提升模型预测代谢谱的性能。

1D-CNN是一种专用于处理序列数据设计的深度学习模型,已广泛应用于自然语言处理、时间序列分析和脑电波信号处理等领域^[29]。在代谢谱预测任务中,基于系统发育排序的微生物数据可视为一种具有生物学结构的序列数据,利用1D-CNN进行特征提取可有效建模微生物间拓扑关系,挖掘其在代谢调控中的作用。本文设样本数据为 $\mathbf{x} \in \mathbb{R}^L$, L 为微生物特征数,卷积核权重为 $\mathbf{w} \in \mathbb{R}^k$ (大小为 k),偏置项为 b , j 的取值区间为 $[0, L - k]$,激活函数为 ReLU,卷积计算公式为:

$$z_j = \max\left(\sum_{m=0}^{k-1} \mathbf{w}_m \mathbf{x}_{j+m} + b, 0\right) \quad (1)$$

随后,将提取到的特征经过最大池化层(MaxPooling)进行降维,以减少冗余信息并提升计算效率。最终,将卷积层输出的特征展平后输入MLP以预测代谢丰度。本文采用均方误差(Mean Squared Error, MSE)作为损失函数,以衡量预测代谢谱与真实代谢谱之间的误差。

$$L_{\text{MSE}} = \frac{1}{N} \|\mathbf{Y} - \mathbf{Y}'\|_F^2 \quad (2)$$

式中: N 为样本数; $\mathbf{Y}, \mathbf{Y}' \in \mathbb{R}^{N \times D_y}$, 分别代表真实代谢物丰度矩阵和预测代谢物丰度矩阵; D_y 为代谢物特征数量; $\|\cdot\|_F$ 表示 Frobenius 范数。

3 实验结果与分析

3.1 数据集与预处理

为了评估模型性能并进行基准比较,使用了5个公开的人类肠道宏基因组—代谢组配对数据队列,涵盖炎症性肠病(IBD)、结直肠癌(CRC)、胃癌(Gastric Cancer, GC)和终末期肾病(End-Stage Renal Disease, ESRD)等4种疾病类型^[30-33]。本文将上述数据集经过文献[34]的方法进行统一预处理,确保初始输入数据一致。具体为,原始宏基因组测序数据经过标准质控、宿主去除及分类学分配,可生成微生物属层级相对丰度的数据。数据集的详细信息如表1所示(数据获取链接为 <https://github.com/borenstein-lab/microbiome-metabolome-curated-data>)。

尽管原始宏基因组数据均经过了统一的预处理,但由于原文中代谢组的数据处理方法不同,各数据集的代谢物数量存在差异,如表2所示。其中,IBD(PRISM队列和NLIBD队列)采用非靶向液相色谱—质谱联用技术(Liquid Chromatography-Mass Spectrometry, LC-MS)与内部化

Table 1 Datasets information

表1 数据集信息

数据集	疾病	样本总数	微生物特征数量	代谢物特征数量
IBD (PRISM 队列)	Inflammatory Bowel Disease	155	11 720	8 848
IBD (NLIBD 队列)	Inflammatory Bowel Disease	65	11 720	8 848
GC	Gastric Cancer	96	10 527	524
ESRD	End-Stage Renal Disease	287	11 845	276
CRC	Colorectal Cancer	347	11 942	450

合物库进行质荷比(m/z)匹配,覆盖的代谢物种类最多;ESRD数据集采用非靶向气相色谱—质谱联用技术(Gas Chromatography-Mass Spectrometry, GC-MS),通过严格的质谱匹配(相似性评分 ≥ 700)及质量控制($RSD < 0.05$)进行筛选,仅保留276种;GC、CRC数据集均采用靶向毛细管电泳—飞行时间质谱(Capillary Electrophoresis-Time of Flight Mass Spectrometry, CE-TOF-MS),基于KEGG数据库进行匹配,检测范围较窄。

Table 2 Metabolomic analysis methods across different datasets

表2 不同数据集的代谢组学分析方法

数据集	代谢物数量	质谱方法	靶向/非靶向	代谢物筛选标准
IBD	8 848	LC-MS (多模式)	非靶向	m/z 匹配+内部化合物库
GC	524	CE-TOF-MS	靶向	KEGG ID 匹配
ESRD	276	GC-MS	非靶向	质谱相似性评分 ≥ 700 且 $RSD < 0.05$
CRC	450	CE-TOF-MS	靶向	KEGG ID 匹配

同时,为了消除个体间的测量差异,降低特征维度的量纲效应,对微生物相对丰度的数据和代谢谱数据进行中心对数比(Centered Log-Ratio, CLR)变换,如式(3)所示。此外,严格按照原始文献中的数据预处理策略,分别进行归一化、特征筛选和特征变换等预处理操作,确保实验结果的公平性。

$$\text{CLR}(\mathbf{x}_i) = \log \mathbf{x}_{ij} - \frac{1}{D} \sum_{j=1}^D \log \mathbf{x}_{ij} \quad (3)$$

式中: $\mathbf{x}_{ij}$ 为样本 i 在特征 j 上的值;减数表示 D 维特征的几何平均值。

3.2 评估指标

本文将代谢谱预测建模为回归任务,常见的回归评估指标包括均方根误差(Root Mean Square Error, RMSE)、平均绝对误差(Mean Absolute Error, MAE)。然而,这些传统指标会受到不同方法的预处理策略所影响,导致跨模型的可比性不强。例如,Melonnpan对代谢物进行arcsin平方根变换以抑制高丰度离群值;ENVIM、MMINP采用Box-Cox变换优化数据分布;本文方法、MiMeNet、mNODE等比较模型,对微生物、代谢数据统一应用中心化对数比(Centered Log Ratio Transformation, CLR)变换,以解决高维稀疏性问

题并增强跨组学关联建模。

由于上述非线性预处理会扭曲原始量纲,跨模型比较绝对误差不够准确。为此,采用平均反正切绝对百分比误差(Mean Arctangent Absolute Percentage Error, MAAPE)替换绝对误差。通过反正切函数将相对误差映射至无单位区间,其尺度不变性可抵消不同预处理(CLR、Box-Cox等)的量化偏差,可在保留各方法技术性选择的前提下,为异构流程提供无偏评估基准。具体计算公式如式(4)所示。

$$\text{MAAPE} = \frac{1}{N} \sum_{i=1}^N \tan^{-1} \left(\left| \frac{y_i - y'_i}{y_i} \right| \right) * \frac{2}{\pi} \quad (4)$$

为了评估预测值与真实值的相关性,采用Spearman相关系数(Spearman Correlation Coefficient, SCC)作为衡量指标,SCC的取值范围为 $[-1, 1]$,数值越接近1表示正相关性越强,接近-1表示负相关性越强。具体计算公式如式(5)所示。

$$\text{SCC} = \frac{\text{cov}(R(Y), R(Y'))}{\sigma_{R(Y)} \sigma_{R(Y')}} \quad (5)$$

式中: $R(Y)$ 、 $R(Y')$ 分别代表真实值和预测值的等级转换; $\text{cov}()$ 为等级变量的协方差函数; σ 代表等级变量的标准差。

为了公平比较不同方法的预测能力,考虑到某些方法可能涉及代谢物筛选,本文参考文献[16]将 $\text{SCC} > 0.3$ 的代谢物定义为预测良好的代谢物,并以其数量作为另一项评估指标,具体计算公式为:

$$\text{WP_Nums} = |\{m | \text{SCC}_m > 0.3\}| \quad (6)$$

为了进一步评估代谢谱预测在疾病分类任务中的应用价值,还额外采用了分类任务中的常见评估指标,包括受试者工作特征曲线下面积(Area Under the Curve, AUC)、分类准确性(Accuracy)、精确率(Precision)、召回率(Recall)和F1分数(F1-score)。

$$\text{AUC} = \frac{1}{N_{\text{pos}} * N_{\text{neg}}} \sum_{i=1}^{N_{\text{pos}}} \sum_{j=1}^{N_{\text{neg}}} \mathbf{I}(s_i > s_j) \quad (7)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (9)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (10)$$

$$\text{F1} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11)$$

式中: N_{pos} 、 N_{neg} 分别为患病和健康样本的数量; s_i 、 s_j 分别表示模型对患病和健康样本的预测得分(即患病概率); $\mathbf{I}(\cdot)$ 为一个指示函数,当 $s_i > s_j$ 时值为1,否则为0; TP 、 TN 、 FP 、 FN 分别表示真阳性(True Positive)、真阴性(True Negative)、假阳性(False Positive)和假阴性(False Negative)。

3.3 实验环境与模型参数

本文实验环境为Windows 11(64位)系统,配置16 GB内存,采用Intel Core(TM) i7-13620H(2.40 GHz)处理器,

并搭载NVIDIA RTX4060(8 GB)显卡。模型基于PyTorch框架实现,编程语言为Python 3.9,开发环境为PyCharm 2024.2。训练过程中,batch size为32,最大迭代1 000轮,学习率为 $1e-3$,同时施加L2正则化(权重衰减 $1e-4$),通过连续20轮实验来验证损失无改善的早停策略,以确保训练过程稳定高效。

在模型架构设计方面,采用1D-CNN与MLP的混合结构(详细参数见表3)进行实验。其中,1D-CNN通过多层卷积与池化操作提取微生物特征的局部模式与全局关联信息;MLP通过全连接层与正则化机制实现特征的高维映射与回归预测。为了验证所选卷积层深度、通道数及MLP维度的合理性,选择代谢物规模极值数据集IBD(8 848种)、ESRD(276种)开展消融实验,如图2所示。

Table 3 Core parameter settings of 1D-CNN and MLP architecture
表3 1D-CNN与MLP架构的核心参数设置

层类型	模型参数	取值/配置
1D-CNN	层数	3
	通道数	32 → 64 → 128
	卷积核大小/步幅	5 / 1
	填充策略	Same Padding
	激活函数	ReLU
	池化类型	MaxPool1d
	池化核大小/步幅	2 / 2
MLP	全连接层数	3(2隐藏层+1输出层)
	隐藏层尺寸	256 → 128
	隐藏层激活函数	ReLU
	Dropout率	0.4
	输出层激活函数	Linear

由图2可见,采用3层卷积网络(32-64-128通道)及256-128维MLP时,所提模型在IBD数据集上的性能最优(ALL_SCC=0.373 9),相较于次优的3层架构(16-32-64通道,ALL_SCC=0.370 8)提升0.8%;在ESRD数据集上,所提模型的ALL_SCC为0.247 8,相较于次优结果(16-32-64通道,ALL_SCC=0.241 5)提升2.6%。综上,不同的卷积结构或MLP维度均会导致性能发生变化,证明了当前网络结构配置在模型复杂度与特征提取能力之间的平衡性较好。

3.4 实验评估

为了验证PCMMP框架的有效性,在GC、ESRD、CRC和IBD(PRISM队列)数据集中进行五折交叉验证和跨队列验证实验,并将所提方法与7种先进预测方法的性能进行比较,以探讨PCMMP预测得到的代谢谱在疾病分类任务中的应用价值,如表4所示。上述7种代谢谱预测模型包括:MelonnPan、ENVIM、MMINP等3个传统机器学习算法;BiomeNED、MiMeNet、LOCATE、mNODE等4个深度学习模型^[16, 17, 19, 20-23]。评估指标包括MAAPE、前50个良好预测代谢物的Spearman均值(TOP50_SCC)、所有预测代谢物的Spearman均值(ALL_SCC)及良好的预测代谢物数量(WP_Nums)。

由表4可知,PCMMP在MAAPE、TOP50_SCC和WP_Nums等关键指标上表现突出。在IBD数据集中,PC-

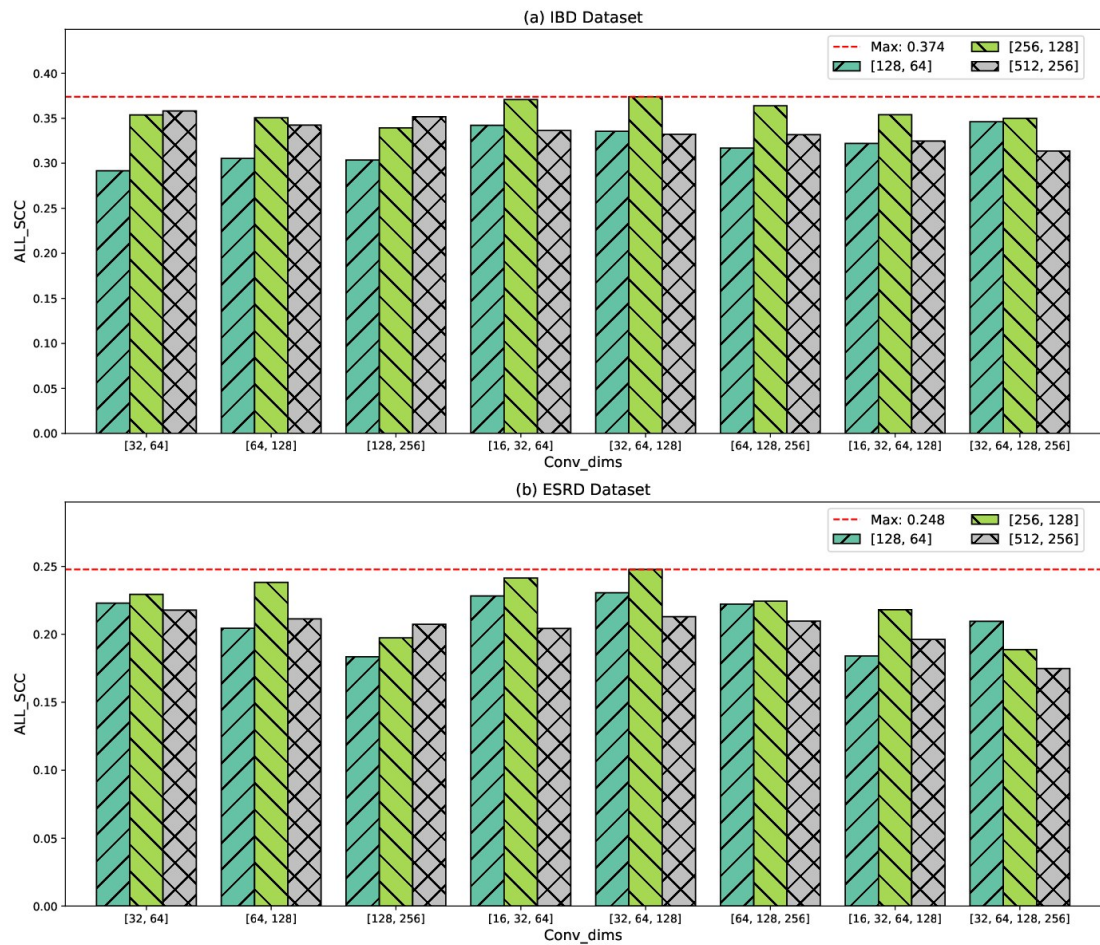


Fig. 2 Impact of different convolutional network structures and MLP parameter combinations on predictive performance

图 2 不同卷积网络结构与 MLP 参数的组合对预测性能的影响

Table 4 Multi-metric comparison and overall Friedman test ranking of eight methods across four datasets

表 4 8 种方法在 4 个数据集上的多指标比较及综合 Friedman 检验排名

数据	指标	MelonnPan	ENVIM	MMINP	MiMeNet	BiomeNED	mNODE	LOCATE	PCMMP
IBD	MAAPE	0.4959 (8)	0.3993 (3)	0.4403 (5)	0.4184 (4)	0.4768 (7)	0.3921 (2)	0.4481 (6)	0.3879 (1)
	TOP50_SCC	0.7847 (2)	0.7477 (5)	0.7739 (3)	0.7019 (8)	0.7710 (4)	0.7466 (6)	0.7386 (7)	0.7936 (1)
	ALL_SCC	0.5611 (1)	0.5015 (2)	0.4659 (3)	0.2996 (7)	0.2522 (8)	0.3229 (5)	0.3038 (6)	0.3739 (4)
	WP_Nums	651 (8)	827 (7)	954 (6)	4550 (3)	3027 (5)	4827 (2)	4338 (4)	5711 (1)
	Aver_Rank	4.75	4.25	4.25	5.50	6.00	3.75	5.75	1.75
GC	MAAPE	0.5463 (8)	0.3741 (4)	0.4496 (5)	0.3397 (3)	0.4908 (7)	0.2165 (1)	0.4686 (6)	0.2535 (2)
	TOP50_SCC	0.4473 (3)	0.5373 (1)	0.4284 (6)	0.4396 (5)	0.4177 (7)	0.3991 (8)	0.4453 (4)	0.4685 (2)
	ALL_SCC	0.4473 (2)	0.4698 (1)	0.4284 (3)	0.1488 (8)	0.2421 (5)	0.2197 (6)	0.1621 (7)	0.2513 (4)
	WP_Nums	17 (8)	55 (5)	48 (7)	67 (4)	54 (6)	185 (2)	111 (3)	269 (1)
	Aver_Rank	5.25	2.75	5.25	5.00	6.25	4.25	5.00	2.25
ESRD	MAAPE	0.7915 (8)	0.4889 (7)	0.4621 (4)	0.3483 (3)	0.4666 (5)	0.3326 (2)	0.4845 (6)	0.3069 (1)
	TOP50_SCC	0.4759 (4)	0.5716 (1)	0.4693 (5)	0.4052 (8)	0.4972 (2)	0.4532 (7)	0.4608 (6)	0.4879 (3)
	ALL_SCC	0.4759 (1)	0.3673 (3)	0.4349 (2)	0.2441 (7)	0.2821 (4)	0.2793 (5)	0.1330 (8)	0.2478 (6)
	WP_Nums	34 (8)	145 (1)	54 (6)	122 (4)	119 (5)	139 (2)	40 (7)	124 (3)
	Aver_Rank	5.25	3.00	4.25	5.50	4.00	4.00	6.75	3.25
CRC	MAAPE	0.5108 (8)	0.3564 (4)	0.4484 (5)	0.3480 (3)	0.4498 (6)	0.2493 (1)	0.4756 (7)	0.2571 (2)
	TOP50_SCC	0.4825 (5)	0.4943 (4)	0.4959 (3)	0.4532 (6)	0.3671 (8)	0.5065 (2)	0.3991 (7)	0.5077 (1)
	ALL_SCC	0.4593 (2)	0.4283 (3)	0.4683 (1)	0.2109 (7)	0.2290 (6)	0.2604 (4)	0.1311 (8)	0.2399 (5)
	WP_Nums	50 (7)	60 (6)	62 (5)	79 (3)	45 (8)	239 (1)	66 (4)	115 (2)
	Aver_Rank	5.50	4.25	3.50	4.75	7.00	2.00	6.50	2.50
ALLDATA_RANK		5.875	2.625	4.500	5.625	6.875	2.375	6.625	1.500

MMP的MAAPE误差率相较于mNODE、ENVIM降低0.4%、1.1%;PCMMP的TOP50_SCC相较于MelonnPan、MMINP提升0.008 9、0.019 7;PCMMP的WP_Nums领先第二名(mNODE)884个代谢物,证明其在代谢谱预测任务中的竞争优势明显。在GC、ESRD和CRC数据集上,PCMMP的表现稳定保持在前三,在ALL_SCC指标上,包括PCMMP在内的深度学习方法整体表现一般,原因为传统机器学习方法通常采用微生物特征筛选策略对每个代谢物单独建模,使其在特定代谢物上的SCC相关性更高。然而,MelonnPan、MMINP过滤了预测效果较差的代谢物,使ALL_SCC主要反映效果更好的可预测子集,能更灵活地对单一代谢物进行预测,但与代谢谱预测的全局性需求存在冲突,会导致计算复杂度剧增,忽视了代谢物之间的协同效应。PC-MMP的全局学习机制通过整合跨代谢物的微生物关联模式,实现了全谱代谢物的无偏预测。

尽管这种全局学习可能在部分低相关性代谢物的预测精度上有所牺牲,但在关键代谢物精度(TOP50_SCC)和全谱覆盖性(WP_Nums)指标上优势明显,能与传统方法一样有效预测高关联代谢物,更符合复杂代谢系统的真实建模需求。在IBD、GC数据集中,PCMMP在ALL_SCC上优于其他深度学习模型,验证了其综合性能优势明显。

为了进一步评估各方法在不同数据集上的综合性能,采用Friedman检验进行统计分析。结果显示,Friedman统计值为20.473,对应的 p 值为0.004 6($N=4, K=8$),表明不同方法在4个数据集上的性能存在显著差异($p<0.01$)。综合排名显示,PCMMP、mNODE和ENVIM表现最佳,而LOCATE和BiomeNED结果相对较差。总体而言,PCMMP的MAAPE、TOP50_SCC、WP_Nums指标表现优越,尽管在ALL_SCC上略逊于传统方法,但整体预测能力稳健,并在Friedman统计分析中排名首位,验证了其在代谢谱预测任务中的有效性与泛化性。

为了评估各模型在跨队列任务中的泛化能力,将IBD

的PRISM队列作为训练集,并在IBD的NLIBD独立队列上进行测试,如图3所示。其中,横轴表示不同方法,左纵轴(柱状图)表示平均Spearman相关性(ALL_SCC),右纵轴(折线图)反映良好预测的代谢物数量(WP_Nums)。

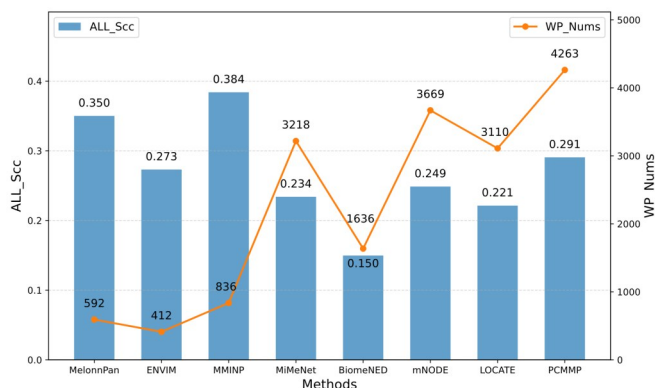


Fig. 3 8 methods for predicting metabolic profiles based on ALL_SCC and WP-Nums indicators in cross queue tasks

图3 8种方法在跨队列任务中基于ALL_SCC和WP_Nums指标的代谢谱预测性能

由图3可见,PCMMP的良好预测代谢物数量为4 263个,领先第二名594个。在ALL_SCC指标上,PCMMP的结果为0.291,超越了ENVIM和所有深度学习方法,这种高精度、高覆盖性的平衡使其在大规模代谢物预测任务中的优势明显,更符合全局代谢谱预测的需求。此外,MelonnPan、ENVIM在跨队列中的ALL_SCC,相较于五折交叉性能下降显著(>0.2),表明这些模型可能存在过拟合现象。

为了进一步验证该推测,对微生物组和代谢组数据分别进行PCA分析和置换检验,如图4所示。由此可见,微生物组训练集与测试集的中心距离为31.38($p=0.001$),代谢组为79.38($p=0.001$),表明两组数据分布存在显著差异,证明了传统方法独立建模单一代谢物更容易过拟合,而PCMMP通过学习全局特征具有更好的泛化性和鲁棒性。

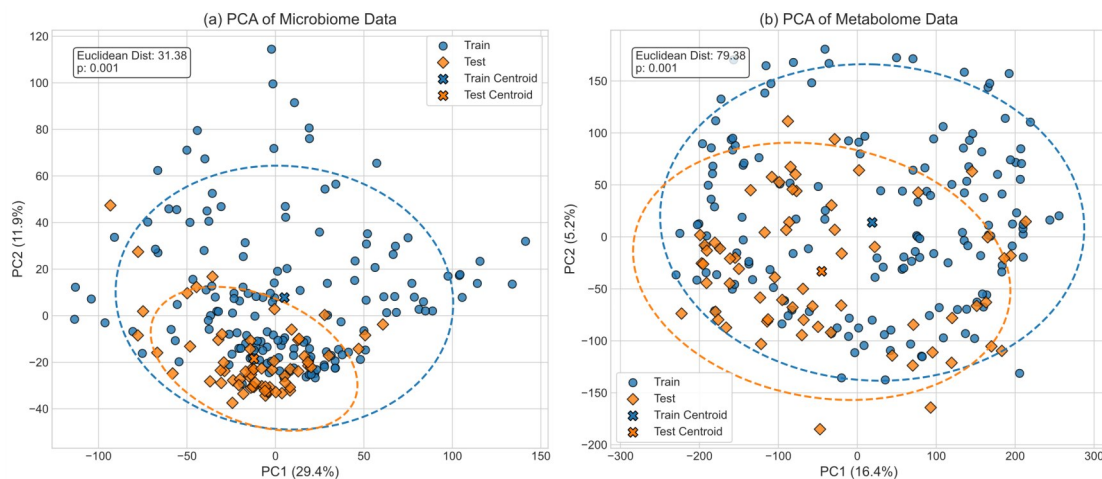


Fig. 4 Differences in PCA distributions between two data types in IBD training and testing sets

图4 两种数据类型在IBD训练集与测试集中的PCA分布差异

为了验证 PCMMP 预测代谢谱的应用价值,对 PCMMP 预测的 NLIBD 队列代谢组 (PCMMP_Metabolome)、基于 MLP 结合远近支融合数据预测的代谢组 (MLP_Metabolome)、真实 NLIBD 队列中的微生物组 (True_Microbiome) 和代谢组 (True_Metabolome) 数据进行分类性能比较,实验采用 Lasso 逻辑回归分类器进行五折交叉验证,如表 5 所示。由此可知,基于 True_Metabolome 训练的分类器在各项指标上表现最佳,AUC 达到 0.902 1;基于 PCMMP_Metabolome 训练的分类器,AUC 略低于基于 True_Microbiome 训练的分类器(0.870 4 vs 0.887 8),但其他指标更高;基于 MLP_Metabolome 训练的分类器相较于 True_Microbiome,AUC 下降约 6%(0.828 4 vs 0.887 8)。实验表明,CNN 模型具备更强的表示学习能力,基于 PCMMP 预测的代谢谱在疾病分类任务中具有较高的应用价值。

3.5 消融实验

为了验证 PCMMP 模型各结构的有效性,在数据层面

Table 5 Evaluation of the application value of four different data in disease classification tasks

表 5 4 种不同数据在疾病分类任务中的应用价值评估					
数据	Accuracy	AUC	Precision	Recall	F1-score
True_Microbiome	0.769 2	0.887 8	0.830 6	0.836 5	0.825 0
True_Metabolome	0.907 7	0.902 1	0.960 0	0.915 1	0.935 0
MLP_Metabolome	0.784 6	0.828 4	0.847 1	0.819 1	0.830 0
PCMMP_Metabolome	0.800 0	0.870 4	0.851 7	0.850 1	0.842 1

和模型层面分别设计消融实验。在数据层面,采用不同的数据处理策略,包括未进行系统发育排序的原始数据(O)、基于近支微生物特征的后序遍历数据(P)、基于远支微生物特征重排的数据(F)及远近支融合数据(C)。在模型层面,比较了 MLP 与 CNN+MLP 模型架构的性能。具体实验结果如图 5 所示。其中,纵轴列出了不同的数据处理策略与模型架构组合,下方横轴表示平均 Spearman 相关系数 (ALL_SCC),上方横轴表示良好预测的代谢物数量 (WP_Nums)。

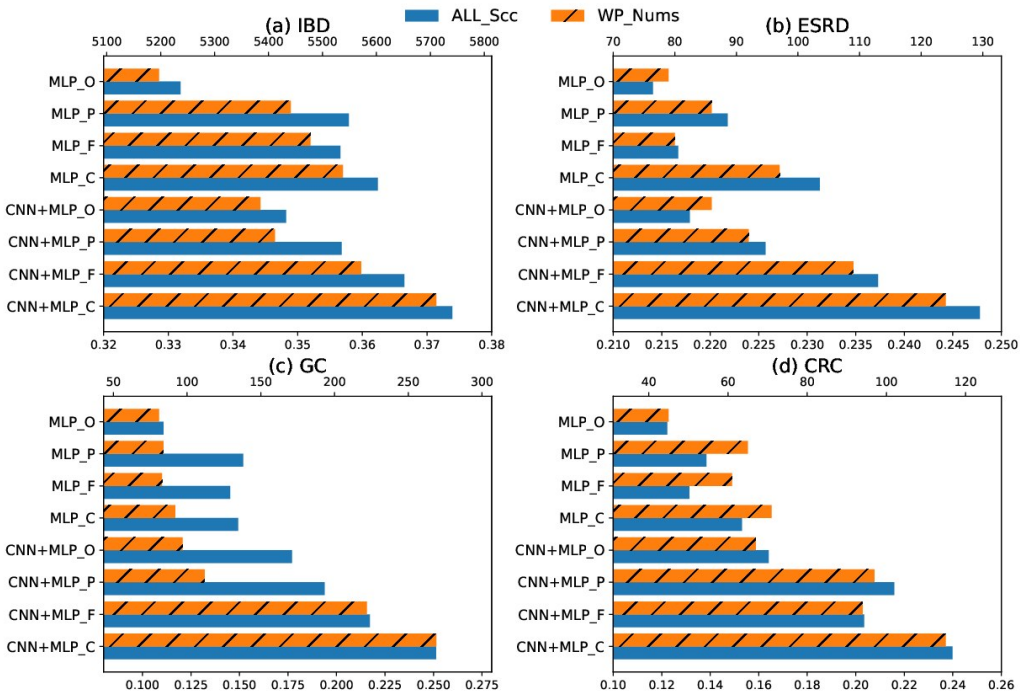


Fig. 5 Bar chart of metabolic profile prediction performance across different data types and model architectures

图 5 不同数据类型与模型架构组合下的代谢谱预测性能柱状图

由图 5 可知,PCMMP(CNN+MLP_C)在 4 个数据集上预测性能最佳。在 IBD、ESRD 和 GC 数据集上,CNN+MLP 架构的远支数据(F)预测效果优于近支数据(P),尤其在 GC 数据集上,远支数据(F)显著提升了模型预测性能,原因可能为该数据集中微生物代谢功能分布较为分散,远支数据能捕捉到因水平基因转移和趋同进化引起的跨进化支系功能冗余。在 CRC 数据集中,CNN+MLP 架构的近支数据(P)表现更佳,表明该框架能适应不同数据集的微生物代谢关联模式。对于 MLP 架构,远支和近支数据的性能差异较小,且整体表现低于相同数据下的 CNN+MLP,表明 MLP 在建模微生物系统发育信息方面存在一定的局限性。

值得注意的是,无论是 MLP 还是 CNN+MLP,一旦引入系统发育排序(P,F,C),其预测精度与覆盖度均高于未排序的原始数据(O),证实了系统发育排序在挖掘微生物一代代谢物关联中的重要性。实验结果充分验证了,PCMMP 各结构在代谢谱预测任务中均具有积极作用。

4 结语

本文提出一种融合系统发育信息的代谢谱预测模型 PCMMP,旨在解决现有方法在代谢谱预测中忽略微生物系统发育关系所带来的性能局限性。首先,基于系统发育

树的排序策略有效捕捉了近支与远支微生物间的潜在代谢关联;其次,结合 1D-CNN 对处理后的微生物组数据进行特征学习,并利用 MLP 预测代谢丰度。

五折交叉验证、跨队列验证和疾病分类等实验结果表明,PCMMP 在代谢谱预测中的优势明显、潜力巨大。消融实验证明了 PCMMP 组合架构的有效性,表明间隔特征选择策略可无偏地捕捉由水平基因转移引起的代谢冗余信息,并将其与局部代谢关联信息互补来提升预测精度。

然而,模型主要依赖微生物丰度数据,未整合功能基因组信息,而代谢过程中酶促反应会直接依赖特定功能基因的表达,仅靠丰度数据难以全面解析代谢功能的分子机制。未来,将整合基因表达或宏转录组数据,以进一步提升模型对代谢调控的解析能力,并在更大规模、多样化的队列中评估模型在不同环境和人群中的适应性。

参考文献:

- [1] AGGARWAL N, KITANO S, PUAH G R Y, et al. Microbiome and human health: current understanding, engineering, and enabling technologies[J]. *Chemical Reviews*, 2023, 123(1): 31–72.
- [2] TURPIN W, LEE S H, CROITORU K. Gut microbiome signature in pre-disease phase of inflammatory bowel disease: prediction to pathogenesis to prevention[J]. *Gastroenterology*, 2025, 168(5): 902–913.
- [3] METWALY A, REITMEIER S, HALLER D. Microbiome risk profiles as biomarkers for inflammatory and metabolic disorders[J]. *Nature Reviews Gastroenterology & Hepatology*, 2022, 19(6): 383–397.
- [4] BAI X, LIU B, FAN D, et al. Modulating the gut microbiota: a novel perspective in colorectal cancer treatment[J]. *Cancer Letters*, 2025, 612: 217459.
- [5] KRAUTKRAMER K A, FAN J, BACKHED F. Gut microbial metabolites as multi-kingdom intermediates[J]. *Nature Reviews Microbiology*, 2021, 19(2): 77–94.
- [6] SUN M, WU W, LIU Z, et al. Microbiota metabolite short chain fatty acids, GPCR, and inflammatory bowel diseases[J]. *Journal of Gastroenterology*, 2017, 52(1): 1–8.
- [7] GASALY N, DE V P, HERMOSO M A. Impact of bacterial metabolites on gut barrier function and host immunity: a focus on bacterial metabolism and its relevance for intestinal inflammation[J]. *Frontiers in Immunology*, 2021, 12: 658354.
- [8] KHAJEH T, REIMAN D, MORLEY R, et al. Integrating microbiome and metabolome data for host disease prediction via deep neural networks[C]// *IEEE EMBS International Conference on Biomedical and Health Informatics*, 2021: 1–4.
- [9] CHEN Y, WANG B, ZHAO Y, et al. Metabolomic machine learning predictor for diagnosis and prognosis of gastric cancer[J]. *Nature Communications*, 2024, 15(1): 1657.
- [10] GOODRICH J A, WANG H, WALKER D I, et al. Postprandial metabolite profiles and risk of prediabetes in young people: a longitudinal multi-cohort study[J]. *Diabetes Care*, 2024, 47(1): 151–159.
- [11] EVERETT J R. NMR-based pharmacometabonomics: a new paradigm for personalised or precision medicine[J]. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 2017, 102–103: 1–14.
- [12] YANG Q, ZHANG A H, MIAO J H, et al. Metabolomics biotechnology, applications, and future trends: a systematic review[J]. *RSC Advances*, 2019, 9(64): 37245–37257.
- [13] ZHENG C, ZHANG S, RAGG S, et al. Identification and quantification of metabolites in (1)H NMR spectra by bayesian model selection[J]. *Bioinformatics*, 2011, 27(12): 1637–1644.
- [14] BRUMFIELD K D, HUQ A, COLWELL R R, et al. Microbial resolution of whole genome shotgun and 16s amplicon metagenomic sequencing using publicly available neon data [J]. *PLoS One*, 2020, 15 (2) : e0228899.
- [15] AYLING M, CLARK M D, LEGGETT R M. New approaches for metagenome assembly with short reads[J]. *Briefings in Bioinformatics*, 2020, 21(2): 584–594.
- [16] MALLICK H, FRANZOSA E A, MCLVER L J, et al. Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences[J]. *Nature Communications*, 2019, 10(1): 3136.
- [17] XIE J, CHO H, LIN B M, et al. Improved metabolite prediction using microbiome data-based elastic net models[J]. *Frontiers in Cellular and Infection Microbiology*, 2021, 11: 734416.
- [18] MULLER E, ALGAVI Y M, BORENSTEIN E. A meta-analysis study of the robustness and universality of gut microbiome-metabolome associations[J]. *Microbiome*, 2021, 9(1): 203.
- [19] TANG W, ZHENG H, XU S, et al. MMNP: a computational framework of microbe-metabolite interactions-based metabolic profiles predictor based on the O₂-pls algorithm [J]. *Gut Microbes*, 2023, 15 (1) : 2223349.
- [20] LE V, QUINN T P, TRAN T, et al. Deep in the bowel: highly interpretable neural encoder-decoder networks predict gut metabolites from gut microbiome[J]. *BMC Genomics*, 2020, 21(4): 256.
- [21] REIMAN D, LAYDEN B T, DAI Y. Mimenet: exploring microbiome-metabolome relationships using neural networks[J]. *PLoS Computational Biology*, 2021, 17(5): 1009021.
- [22] SHTOSSEL O, KOREN O, SHAI I, et al. Gut microbiome-metabolome interactions predict host condition[J]. *Microbiome*, 2024, 12(1): 24.
- [23] WANG T, WANG X W, LEE-SARWAR K A, et al. Predicting metabolomic profiles from microbial composition through neural ordinary differential equations [J]. *Nature Machine Intelligence*, 2023, 5 (3) : 284–293.
- [24] XIAO J, CHEN L, YU Y, et al. A phylogeny-regularized sparse regression model for predictive modeling of microbial community data[J]. *Frontiers in Microbiology*, 2018, 9: 3112.
- [25] LOUCA S, POLZ M F, MAZEL F, et al. Function and functional redundancy in microbial systems [J]. *Nature Ecology & Evolution*, 2018, 2 (6): 936–943.
- [26] WASHBURNE A D, MORTON J T, SANDERS J, et al. Methods for phylogenetic analysis of microbiome data [J]. *Nature Microbiology*, 2018, 3(6): 652–661.
- [27] SHI K, LIU Q, JI Q, et al. Microhdf: predicting host phenotypes with metagenomic data using a deep forest-based framework[J]. *Briefings in Bioinformatics*, 2024, 25(6): 530.
- [28] QUINN T P, ERB I. Examining microbe-metabolite correlations by linear methods[J]. *Nature Methods*, 2021, 18(1): 37–39.
- [29] KIRANYAZ S, AVCI O, ABDELJABER O, et al. 1D convolutional neural networks and applications: a survey[J]. *Mechanical Systems and Signal Processing*, 2021, 151: 107398.
- [30] FRANZOSA E A, SIROTA-MADI A, AVILA-PACHECO J, et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease[J]. *Nature Microbiology*, 2019, 4(2): 293–305.
- [31] YACHIDA S, MIZUTANI S, SHIROMA H, et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer [J]. *Nature Medicine*, 2019, 25 (6) : 968–976.
- [32] ERAWIJANTARI P P, MIZUTANI S, SHIROMA H, et al. Influence of gastrectomy for gastric cancer treatment on faecal microbiome and metabolome profiles[J]. *Gut*, 2020, 69(8): 1404–1415.
- [33] WANG X, YANG S, LI S, et al. Aberrant gut microbiota alters host metabolome and impacts renal failure in humans and rodents [J]. *Gut*, 2020, 69(12): 2131–2142.
- [34] MULLER E, ALGAVI Y M, BORENSTEIN E. The gut microbiome-metabolome dataset collection: a curated resource for integrative meta-analysis[J]. *NPJ Biofilms and Microbiomes*, 2022, 8(1): 79.

(责任编辑:刘嘉文)