

污水处理系统中微生物群落作用机制的统计学方法研究进展

郑 嘉¹, 何 锋², 程华宇³, 李惊鑫⁴, 田永兰^{1*}

¹华北电力大学数理学院, 北京

²华北电力大学环境科学与工程学院, 北京

³华北电力大学核科学与工程学院, 北京

⁴华北电力大学水利与水电工程学院, 北京

收稿日期: 2026年6月8日; 录用日期: 2026年6月30日; 发布日期: 2026年7月13日

摘 要

污水生物处理系统的运行性能与微生物群落结构、功能潜力及代谢响应密切相关。随着高通量测序和代谢组学数据的积累, 统计学方法已成为解析污水处理微生物机制的重要工具。本文围绕微生物群落结构与功能研究, 系统梳理 α 多样性、 β 多样性、差异检验、降维排序、互作网络和群落稳定性等结构统计方法, 并进一步总结功能预测、代谢组统计及功能-物种关联分析的应用思路。结合填料强化多级生物接触氧化工艺和C/N比调控需求, 本文指出现有方法在组成型数据处理、时间滞后刻画、因果解释和多组学整合方面仍存在局限。未来应加强动态统计模型、机器学习、多组学联合建模和工程预警应用, 以提升污水处理系统的精准调控能力。

关键词

污水处理, 微生物群落, 统计学方法, 功能预测, 多组学整合

Research Progress on Statistical Methods for Mechanisms of Microbial Communities in Wastewater Treatment Systems

Jia Zheng¹, Feng He², Huayu Cheng³, Congxin Li⁴, Yonglan Tian^{1*}

¹College of Mathematics and Physics, North China Electric Power University, Beijing

²College of Environmental Science and Engineering, North China Electric Power University, Beijing

³School of Nuclear Science and Engineering, North China Electric Power University, Beijing

⁴College of Water Conservancy and Hydropower Engineering, North China Electric Power University, Beijing

*通讯作者。

文章引用: 郑嘉, 何锋, 程华宇, 李惊鑫, 田永兰. 污水处理系统中微生物群落作用机制的统计学方法研究进展[J]. 统计学与应用, 2026, 15(7): 52-64. DOI: 10.12677/sa.2026.157149

Received: June 8, 2026; accepted: June 30, 2026; published: July 13, 2026

Abstract

The operation performance of wastewater biological treatment system is closely related to microbial community structure, functional potential and metabolic responses. With the accumulation of high-throughput sequencing and metabolomics data, statistical methods have become an important tool for analyzing the microbial mechanism of wastewater treatment. Focusing on the research of microbial community structure and function, this paper systematically reviews the structural statistical methods such as α diversity, β diversity, difference test, dimension reduction sorting, interaction network and community stability, and further summarizes the application ideas of functional prediction, metabolome statistics and functional-species association analysis. Combined with the packing-enhanced multi-stage biological contact oxidation process and C/N ratio control requirements, this paper points out that the existing methods still have limitations in compositional data processing, time lag characterization, causal interpretation and multi-omics integration. In the future, dynamic statistical models, machine learning, multi-omics joint modeling and engineering early warning applications should be strengthened to improve the precise control ability of wastewater treatment systems.

Keywords

Wastewater Treatment, Microbial Communities, Statistical Methods, Functional Prediction, Multi-Omics Integration

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

城镇污水排放量大, 水质组分复杂, 污染物形态与浓度随季节、生活方式和工业排放变化而波动。物理法和化学法能够在预处理、应急处理或特定污染物去除中发挥作用, 但从长期运行成本、二次污染控制和工程可持续性看, 生物处理仍是城镇污水处理的主流路径[1][2]。生物处理的核心不只是反应器或填料本身, 而是依附于活性污泥、生物膜和填料表面的微生物群落; 不同微生物通过有机物氧化、硝化、反硝化、释磷吸磷、胞外聚合物形成和絮体维持等过程, 共同影响有机物、氨氮、总氮和总磷的去除效果[2]-[4]。因此, 污水处理系统的性能和稳定性, 本质上取决于微生物群落结构、功能潜力、代谢活性及其对环境因子的响应[5][6]。

污水生物处理研究已经从单纯关注污染物去除率, 逐渐转向研究“微生物群落如何参与处理过程”的机制问题。填料强化的多级生物接触氧化、固定生物膜-活性污泥(IFAS)等泥膜复合工艺能够为微生物提供更复杂的生活环境, 使悬浮污泥与附着生物膜共同发挥作用, 也使缺氧、好氧和局部微氧环境中的功能菌群分化更明显[7]-[9]。与此同时, C/N比、溶解氧、pH、水力停留时间、污泥龄和温度等环境因子会改变功能菌群的竞争关系和代谢方向, 尤其是C/N比变化不仅影响反硝化碳源供给, 也可能引起群落组成、互作网络、功能预测和代谢物谱的滞后响应[10]-[12]。仅凭常规水质指标难以解释这些过程, 因此需要引入统计学方法对微生物高维数据进行系统解析[13]-[16]。

随着 16S rRNA 扩增子测序、宏基因组测序和代谢组学技术的发展,污水处理微生物研究的数据类型快速扩展[15]-[17]。相较于传统培养方法,高通量测序可以在较高通量下识别优势类群、稀有类群和潜在功能类群;代谢组学则进一步从小分子代谢物层面反映微生物功能是否被激活以及激活程度[2] [17]-[19]。然而,这类数据通常具有高维、稀疏、零值多、组成性强、批次效应明显和动态非平稳等特征[5] [14]。如果只采用简单均值比较或经验判断,容易忽略群落整体结构、类群互作和环境驱动机制。统计学方法可将复杂的操作分类单元(OTU)/扩增子序列变异体(ASV)表、物种丰度表、功能预测表和代谢物矩阵转化为可解释的多样性指标、距离矩阵、排序图、显著性检验、网络拓扑和多组学关联结果[13]-[16] [20]。

因此,统计学不仅用于描述微生物群落“有哪些成员”,还用于回答“群落之间是否不同”、“差异由哪些类群或功能贡献”、“环境因子如何驱动群落变化”、“群落互作网络是否稳定”、“结构变化是否能够传导到功能与代谢层面”等问题[13] [20]-[23]。本文基于应用数学和环境工程交叉视角,结合污水处理微生物研究需求,从微生物群落结构解析和微生物群落功能研究两个方面梳理常用统计方法及其应用进展,并进一步讨论现有方法的局限性与未来发展方向。

2. 微生物群落结构研究的统计学方法

微生物群落结构研究主要关注样本内多样性、样本间差异、群落组成变化、关键类群识别和类群互作模式。对污水处理系统而言,群落结构并不是孤立的生态学描述,而是连接工艺运行与处理效果的重要中间变量[1] [4]-[6]。例如,同一污水厂不同处理单元之间、不同季节之间、碳源添加前后或停止添加碳源后,微生物群落可能在优势菌门、功能属、共现网络模块和稳定性指标上表现出差异[3] [6] [10]-[12]。结构统计方法的基本思路是先将测序数据转化为丰度矩阵,再通过标准化、距离计算、假设检验、降维排序和网络建模等步骤,揭示群落变化规律[13]-[16]。常用工具包括 QIIME、mothur、vegan 等,它们为多样性计算、群落距离、排序分析和置换检验提供了较为成熟的实现路径[15] [16]。

2.1. α 多样性统计

α 多样性用于衡量单一样本或同一处理组内部的物种丰富度和均匀度,描述一个样本内部的微生物群落有多丰富、多均衡[15] [16]。常用指标包括 Observed species 或 Observed ASVs、Chao1、ACE、Shannon、Simpson 和 Pielou 均匀度等。Observed species/ASVs 直接统计观测到的物种或序列变体数量,反映表观丰富度;Chao1 和 ACE 更关注低丰度类群,常用于估计真实丰富度;Shannon 指数同时考虑丰富度和均匀度,数值越高通常表示群落越丰富且分布越均匀;Simpson 指数更强调优势类群的集中程度;Pielou 均匀度则用于判断群落中不同类群的相对分布是否均衡[15]。

在污水处理研究中, α 多样性可用于比较不同处理单元、不同运行阶段和不同环境扰动下的群落复杂度[3] [5] [6]。例如,在填料强化的多级生物接触氧化工艺中,缺氧区和好氧区由于溶解氧、底物类型和氧化还原环境不同,可能形成不同的生物膜群落[7]-[9]。若某一处理区的 Chao1 和 Shannon 指数升高,说明其可检测类群和生态位复杂度可能增加;若 Simpson 指数显示优势菌高度集中,则说明系统可能依赖少数核心类群完成主要功能[5] [15]。对于实际工程而言,适度较高的多样性通常有助于提高系统对冲击负荷的缓冲能力,但多样性升高并不必然等同于处理效率提升,还需要结合功能菌丰度、代谢通路和出水指标共同解释[1]-[4]。

α 多样性统计需要注意测序深度、数据稀疏性和组成型数据问题。不同样本的测序量差异会直接影响 Observed species 和 Chao1 等指标,因此常需要进行抽平、标准化或采用更稳健的模型方法[14]-[16]。同时,微生物测序得到的是相对丰度而非绝对数量,某一类群相对丰度升高可能来自其自身增长,也可

能来自其他类群下降[14]。因此,在解释 α 多样性变化时,应避免把指数变化简单等同于微生物总量变化,必要时可结合 qPCR、MLSS、污泥活性或功能基因绝对定量进行验证[5] [14]。

2.2. β 多样性

β 多样性用于衡量不同样本或不同处理组之间的群落差异,用于描述两个样本的微生物组成是否相似、差异有多大[15]。与 α 多样性关注样本内部不同, β 多样性强调样本之间的距离或相异度,是后续排序分析、聚类分析和组间差异检验的基础[3] [5]。常用距离包括 Bray-Curtis 距离、Jaccard 距离、UniFrac 距离、Euclidean 距离等。其中,Bray-Curtis 距离基于丰度信息,适合反映优势群落和丰度梯度变化;Jaccard 距离基于存在/缺失信息,适合关注群落组成是否发生替换;weighted UniFrac 和 unweighted UniFrac 进一步引入系统发育信息,前者同时考虑系统发育距离和丰度,后者更关注谱系组成的有无[15] [16]。

在污水处理微生物研究中, β 多样性常用于比较不同反应区、不同采样时期、不同 C/N 比或不同工艺条件下群落结构是否发生整体迁移[3] [4] [6]。例如,当碳源添加后反硝化菌相关类群逐渐富集,Bray-Curtis 距离矩阵可能显示处理组之间距离增加;当停止添加碳源后群落向初始状态回归,时间序列上的 β 多样性轨迹可能表现为逐步靠近基线样本[10]-[12]。Jaccard 距离则有助于识别稀有类群替换,例如低丰度功能菌是否在环境变化后出现或消失[6]。

β 多样性结果通常需要与主坐标分析(PCoA)、非度量多维尺度分析(NMDS)或聚类树结合展示,也可以进一步进入置换多元方差分析(PERMANOVA)或相似性分析(ANOSIM)进行显著性检验[13]。需要注意的是, β 多样性受数据预处理、距离选择和样本离散度影响较大。若组间差异显著,需判断差异来自组间中心位置变化,还是来自组内离散程度不同[13]。对于污水处理厂这种受季节、进水波动和运行调控共同影响的系统,单一距离指标往往不足以说明机制,较合理的做法是同时报告距离方法、排序图、组间检验结果以及主要贡献类群[5] [6] [13]。

2.3. 差异检验

差异检验用于判断不同处理组、时间点或反应区之间的微生物群落是否存在统计学差异。根据研究对象不同,差异检验可分为单变量检验和多变量检验:单变量检验常用于比较某个菌门、菌属、功能基因或代谢物在不同组之间的丰度差异,常用方法包括 t 检验、Wilcoxon 秩和检验、Kruskal-Wallis 检验、方差分析和多重检验校正等;多变量检验则用于判断整个群落矩阵是否存在组间差异,常用方法包括 ANOSIM、PERMANOVA 和多重响应排列程序(MRPP)等[13]。

ANOSIM 基于距离矩阵比较组内距离和组间距离的秩差异,R 值越接近 1,说明组间差异越明显;PERMANOVA 通过置换检验评估分组变量对群落距离矩阵的解释度,常以 R^2 和 P 值表示结果[13]。与传统多元方差分析(MANOVA)相比,PERMANOVA 不要求多元正态分布,适合微生物群落这类高维、非正态、稀疏数据,因此在污水处理微生物研究中应用广泛[5] [13]。例如,研究碳源添加前后群落差异时,可用 PERMANOVA 检验“运行阶段”是否显著解释群落结构变化;研究缺氧区、好氧区和沉淀区差异时,可用 ANOSIM 或 PERMANOVA 判断不同处理单元是否形成稳定分化[7]-[11]。

差异检验的关键不在于获得显著 P 值,而在于建立合理的统计设计。污水处理样本常具有时间连续性和空间相关性,若把同一反应器不同日期样本简单当作完全独立样本,可能导致伪重复[6] [13]。较稳妥的做法是结合采样设计,明确重复层级,并在可行时使用置换限制、分层置换、混合效应模型或按反应器/时期汇总后再检验[13]。此外,在筛选差异菌属和差异功能时,应进行 FDR 校正,并同时报告效应量、丰度变化方向和生态功能含义,避免只根据 P 值解释工程机制[13] [14]。

2.4. 降维排序

降维排序是将高维微生物群落数据压缩到二维或三维空间,以便直观展示样本之间的相似性、分组趋势和环境梯度[24]。无约束排序不预设环境因子,主要用于探索群落内部结构;约束排序则把环境变量纳入模型,用于解释群落变化与环境因子之间的关系[5]。常用无约束排序包括主成分分析(PCA)、PCoA和NMDS。PCA适合线性关系较强且经过适当变换的连续变量;PCoA可基于Bray-Curtis、Jaccard等任意距离矩阵,是微生物群落分析中常见的展示方法之一;NMDS基于秩次关系,适合非线性生态距离,但需要关注stress值, stress值过高说明低维图难以准确表示原始距离[24]。

约束排序包括冗余分析(RDA)、典范对应分析(CCA)和基于距离的冗余分析(db-RDA)等。RDA可看作多元线性回归与PCA的结合,适合物种数据与环境因子之间近似线性响应的场景;CCA适合物种沿环境梯度呈单峰响应的情形;db-RDA则可以基于非欧式距离,更适合Bray-Curtis等生态距离[24]。污水处理研究中常把温度、pH、溶解氧(DO)、化学需氧量(COD)、氨氮、总氮、总磷、C/N比、HRT、SRT等作为解释变量,分析它们对群落结构的贡献[3][4][10]-[12]。若RDA结果显示C/N比向量与反硝化菌或相关功能模块方向一致,则可说明碳源供给可能是驱动该类群富集的重要因素;若DO与硝化菌类群方向一致,则可提示好氧硝化环境对群落结构具有筛选作用[7][10]-[12]。

降维排序具有直观优势,但也容易被过度解释。排序图中样本点距离越近通常代表群落越相似,但轴解释率、变量共线性、数据转换方式和离群样本都会影响结果[24]。使用RDA或CCA前,应先检查环境变量共线性,必要时删除高度相关变量或采用逐步筛选;使用PCA前,应考虑Hellinger转换、中心化对数比转换等,以降低高丰度类群和组成型约束的影响[13]。对于动态污水处理过程,排序图最好结合时间箭头、阶段分组和环境因子变化趋势,才能更清楚地揭示群落从扰动到恢复的过程[6][14]。

2.5. 互作网络

互作网络用于描述微生物类群之间潜在的共现、互斥和模块化关系[21][22]。网络中的节点通常代表OTU、ASV、菌属或功能基因,边表示节点之间达到统计阈值的相关关系。正相关边常被解释为共现关系,可能反映共同环境偏好、代谢互补或协同作用;负相关边常被解释为互斥关系,可能与竞争、生态位差异或环境选择有关[21][22]。需要强调的是,共现网络主要揭示统计关联,不等同于真实生物互作,因此需要结合生态知识和实验验证进行解释[5][21][22]。

网络构建方法主要包括Pearson相关、Spearman相关、稀疏相关性分析(SparCC)、稀疏逆协方差估计(SPIEC-EASI)等[21][22]。Pearson相关适合线性关系,Spearman相关适合单调关系,对非正态数据更稳健。由于微生物测序数据具有组成型特征,常规相关系数可能产生伪相关,SparCC等方法通过考虑组成型约束来推断更稳健的相关网络[14][21]。网络构建通常包括低丰度/低出现率类群过滤、相关系数计算、显著性检验、多重校正、阈值筛选和网络可视化等步骤[21][22]。过滤阈值不宜过低,否则网络会被大量稀有噪声类群干扰;也不宜过高,否则可能遗漏具有生态意义的低丰度关键类群[22][25]。

网络拓扑信息统计是互作网络分析的核心。常用指标包括节点数、边数、平均度、聚类系数、平均路径长度、网络密度、模块化程度、正负边比例、介数中心性和接近中心等[21][22]。平均度和网络密度较高通常说明类群关联更紧密;模块化程度较高说明群落可分成多个相对独立的功能模块;负相关比例增加可能提示竞争或环境分化增强[21][22][25]。在污水处理系统中,若碳源添加后反硝化相关模块内部连接增强,可能说明碳源供给促进了相关功能类群的协同;若停止添加碳源后网络连接减弱或模块解体,则可能提示原有功能互作难以维持[10]-[12]。

关键种识别是互作网络的另一个重要方向。常用判据包括高度节点、模块中心节点、连接不同模块

的 connector 以及高介数中心性的节点[22] [25]。关键种不一定是丰度最高的类群，它们可能通过连接多个功能模块、维持网络结构或调节资源流动而影响系统稳定性[22] [23] [25]。在污水处理工程中，识别关键种有助于解释为何某些低丰度类群变化会导致处理性能波动，也可为后续菌群强化、碳源调控和运行预警提供依据[5] [25]。

2.6. 群落稳定性

群落稳定性是指微生物群落在外界扰动下维持结构和功能的能力，以及扰动结束后恢复到原有状态或形成新稳定状态的能力[23]。稳定性通常包括抵抗力、恢复力、时间稳定性和鲁棒性等维度。抵抗力强调扰动发生时群落变化幅度是否较小；恢复力强调扰动解除后群落能否快速回到扰动前状态；时间稳定性关注长期运行中群落波动是否较低；鲁棒性则常用于网络层面，表示网络在节点或边被移除后保持连通和功能的能力[23] [25]。

常用稳定性评价方法包括平均变异度(AVD)、时间序列稳定性和网络鲁棒性等。AVD 可理解为平均变异程度，通常基于不同时点与基准群落之间的差异计算，数值越低代表群落相对变化越小、结构越稳定[6] [23]。时间序列稳定性可用变异系数、Bray-Curtis 距离随时间的变化、相邻时间点群落漂移速率或回归斜率表示[6] [23]。网络鲁棒性通常通过随机或定向移除节点，观察网络连通性、最大连通子图比例或自然连通度下降程度，从而判断网络是否依赖少数关键节点[22] [23] [25]。

在污水处理系统中，稳定性分析具有重要工程意义。污染物去除率短期稳定并不意味着微生物群落也稳定，因为群落可能已经发生结构替换，只是功能冗余暂时维持了出水指标[1] [5] [23]。相反，群落结构出现一定波动也不必然导致处理性能下降，若功能冗余较高，系统仍可维持氮磷去除能力[2] [3] [23]。因此，稳定性研究应同时关注结构稳定、功能稳定和性能稳定[5] [23]。对于 C/N 比调控问题，稳定性统计可用于回答碳源添加后群落是否形成稳定反硝化模块、停止添加碳源后功能菌群是否持续存在，以及群落恢复是否存在时间滞后[10]-[12]。

3. 微生物群落功能研究的统计学方法

微生物群落结构分析可以回答“有哪些微生物”和“群落如何变化”，而功能研究进一步回答“这些微生物可能做什么”和“功能是否真正发生”[2] [17]-[19] [26] [27]。污水处理过程中的关键功能包括有机物降解、硝化、反硝化、厌氧释磷、好氧吸磷、发酵、水解、胞外聚合物形成和抗逆响应等[2] [3]。由于 16S rRNA 测序主要提供分类信息，不能直接测量功能基因或代谢产物，因此常需要借助功能预测、宏基因组、代谢组和多组学关联方法来推断或验证功能变化[17]-[20] [26] [27]。

功能统计与结构统计之间存在递进关系。首先，基于物种组成可预测潜在功能；其次，基于代谢组数据可观察代谢产物和通路层面的实际变化；最后，通过功能 - 物种关联分析，可把优势类群、关键功能和环境因子连接起来，形成“运行条件 - 群落结构 - 功能潜力 - 代谢产物 - 处理性能”的解释链条[17]-[20] [26] [27]。对于填料强化的多级生物接触氧化工艺，这一链条尤其重要，因为不同处理区的微环境会造成不同类群和功能的空间分化[7]-[9]。

3.1. 功能预测统计

功能预测统计是指利用微生物分类组成或系统发育信息，推断样本中可能存在的功能基因、代谢通路或生态功能类群[26] [27]。该方法常用于 16S rRNA 测序研究，因为 16S 数据成本低、样本量大，但不能直接提供基因功能信息[16]。通过功能预测，可以在缺乏宏基因组数据的情况下初步判断不同样本在碳代谢、氮代谢、磷代谢、能量代谢和抗逆功能上的潜在差异[19] [26] [27]。

PICRUSt2 是较常用的功能预测思路之一,其基本逻辑是基于系统发育关系,将 ASV 或 OTU 序列放置到参考系统发育树上,再根据近缘参考基因组推断基因家族丰度,并进一步映射到 KEGG Orthology、EC 编号或 MetaCyc 通路等功能层级[19]。实际分析中, PICRUSt2 输出结果通常还需要进行标准化、差异分析和可视化, ggpicrust2 等工具可用于对 PICRUSt2 预测功能表进行组间比较、通路展示和富集分析。在污水处理研究中,这类方法可用于比较不同 C/N 比、不同反应区或不同运行阶段样本的氮代谢、碳代谢和能量代谢潜力[10]-[12]。

FAPROTAX 主要面向细菌和古菌生态功能注释,其根据已知培养菌和文献资料,将分类单元映射到硝化、反硝化、发酵、甲烷氧化、硫氧化、光能自养等生态功能类群[26]。相比 PICRUSt2, FAPROTAX 更强调生态功能解释,适合快速判断污水处理系统中氮循环、碳循环和硫循环相关类群的潜在功能变化[3] [26]。例如,当缺氧区反硝化相关功能类群相对丰度升高,同时总氮去除率提高时,可为碳源促进反硝化提供微生物学证据[10]-[12]。

FUNGuild 则主要用于真菌群落功能注释,可将真菌类群划分为腐生、病原、共生等营养型或生态功能型[27]。虽然污水处理研究通常以细菌为主,但真菌在复杂有机物降解、生物膜形成和生态互作中也可能发挥作用,因此在真菌测序数据中可作为辅助分析工具[2] [27]。

功能预测统计的优势是成本较低、适合大样本比较、能够从结构数据延伸到功能解释[26] [27]。但其局限也很明显:第一,预测结果依赖参考数据库,未培养类群或缺乏参考基因组的类群容易被低估;第二,功能预测反映的是潜在功能,不代表基因一定表达或代谢过程一定发生;第三,分类注释错误会传导到功能预测结果[26] [27]。因此,功能预测结果最好与宏基因组、宏转录组、功能基因 qPCR、酶活性或代谢组数据相互验证,避免把“潜在功能”直接等同于“实际功能” [17] [18]。

3.2. 代谢组统计

代谢组学通过检测样本中的小分子代谢物,反映微生物群落和环境条件共同作用下的代谢状态[17] [18]。与 16S 功能预测相比,代谢组更接近功能结果层面,可用于回答“哪些代谢产物发生变化”、“变化通路是否与污染物去除有关”、“代谢响应是否滞后于群落结构变化”等问题[2] [17] [18]。代谢组学可分为靶向代谢组和非靶向代谢组:靶向代谢组针对预设代谢物进行定量,适合验证特定通路;非靶向代谢组覆盖范围更广,适合探索未知差异代谢物和潜在通路[17] [18]。

代谢组统计通常包括数据预处理、质量控制、归一化、缺失值处理、数据变换、单变量分析、多变量分析和通路富集分析[17] [18]。常用的多变量方法包括 PCA、偏最小二乘判别分析(PLS-DA)和正交偏最小二乘判别分析(OPLS-DA)。PCA 主要用于观察样本整体分布、离群样本和批次效应;PLS-DA 属于有监督判别模型,可用于区分不同处理组并筛选贡献较大的代谢物;OPLS-DA 进一步分离组间相关变异和组内无关变异,提高判别解释性[17] [18]。筛选差异代谢物时,常将变量重要性投影(VIP)值、Fold Change、P 值和阳性发现错误率(FDR)校正结合使用,并通过火山图展示显著上调和下调代谢物[17] [18]。

KEGG 富集分析是代谢组统计中的重要步骤。研究者可将差异代谢物映射到 KEGG 通路,进一步识别氨基酸代谢、碳水化合物代谢、脂质代谢、能量代谢和氮代谢等显著变化通路[17]-[19]。在污水处理研究中,如果碳源添加后短链脂肪酸、氨基酸或核苷酸相关代谢物发生变化,并且这些代谢物与反硝化菌富集或总氮去除率提高相一致,则可以从代谢层面支持碳源调控微生物功能的解释[10]-[12];若停止添加碳源后代物谱仍维持一段时间,则可能提示系统存在代谢记忆或功能滞后效应[6] [17] [18]。

代谢组统计也存在一些局限。PLS-DA 等有监督模型容易过拟合,因此必须进行交叉验证、置换检验和外部验证[17] [18]。火山图和 VIP 筛选应结合生物学意义,不能只以阈值机械筛选;非靶向代谢组存在注释不确定性,同一质荷比峰可能对应多个候选化合物,因此,代谢组结果应与微生物结构、功能预测、

环境因子和水质指标联合解释，并在关键代谢物上进行靶向验证[17][18]。

3.3. 功能 - 物种关联

功能 - 物种关联分析用于连接“谁在变化”和“功能如何变化”。在污水处理微生物研究中，研究者不仅需要知道某些菌属是否富集，还需要判断这些菌属是否与氮去除、碳代谢、磷转化或关键代谢物相关[2][3][20]。常用方法包括 Spearman 相关、Pearson 相关、Mantel 检验、Procrustes 分析、RDA/CCA、加权基因共表达网络分析(WGCNA)和偏最小二乘路径模型等[20]。

Spearman 相关是常用的基础方法，适合分析优势菌属与功能基因、代谢物或环境指标之间的单调关系[20]。例如，可计算反硝化菌属与硝酸盐、总氮去除率、反硝化功能预测丰度和相关代谢物之间的相关系数，构建“菌属 - 功能 - 环境因子”热图[3][10]-[12]。Pearson 相关适合近似正态且线性关系明显的数数据。由于微生物和代谢组数据常不满足正态分布，实际研究中更常采用 Spearman 相关，并进行 FDR 校正[14][20]。

WGCNA 可将大量菌属、功能基因或代谢物划分为若干共表达或共变化模块，再分析模块特征值与环境因子或处理性能之间的关系[20]。与逐一相关相比，WGCNA 更适合高维数据，可以识别与 C/N 比、DO、COD 去除率或总氮去除率密切相关的功能模块[10]-[12][20]。若某一模块同时包含反硝化菌、氮代谢功能和特定代谢物，并且与总氮去除率显著相关，则该模块可被解释为潜在的脱氮功能模块[3][10][11][20]。RDA 和 CCA 也可用于功能 - 物种关联，尤其适合将环境变量、群落组成和功能预测结果放在同一解释框架中，分析环境因子对结构和功能的共同驱动[24][28]。

需要注意的是，功能 - 物种关联多为相关性分析，并不能直接证明因果关系：某一菌属与某一代谢物显著相关，可能是该菌属产生该代谢物，也可能是二者共同受 C/N 比或 DO 驱动[12]。因此，后续研究应在统计关联的基础上引入时间滞后分析、结构方程模型、干预实验或因果推断方法，进一步区分直接作用、间接作用和共同驱动[28][29]。

4. 现有方法的局限性与未来挑战

4.1. 现有方法的优势与局限性

总结基于 16S 高通量测序和代谢组学数据的常用微生物群落结构与功能统计分析方法，如表 1 所示。现有统计学方法已经显著提升了污水处理微生物研究的解释能力。 α 多样性和 β 多样性可以描述群落复杂度和组间差异；差异检验能够判断不同工艺条件下群落是否显著变化；降维排序可以把环境梯度与群落结构联系起来；互作网络能够揭示潜在协同、竞争和关键类群；稳定性指标有助于评价群落对扰动的抵抗和恢复能力；功能预测、代谢组统计和多组学关联则进一步推动研究从“结构描述”走向“机制解释”。这些方法为理解填料强化多级生物接触氧化工艺、C/N 比调控和微生物群落响应提供了重要工具。

但是，现有方法仍存在明显局限。第一，微生物测序数据具有组成型特征，相对丰度不能直接代表绝对丰度，常规相关和差异检验可能产生伪相关[14]。第二，微生物数据零值多、稀疏性强，不同过滤阈值和标准化方法会影响结果[14]。第三，污水处理厂样本具有时间连续性和工艺相关性，若忽略时间自相关和重复测量结构，可能造成统计推断偏差[6][13]。第四，功能预测依赖参考数据库，不能替代宏基因组和代谢组实测数据[17]-[19][26][27]。第五，许多研究采样周期较长，难以捕捉微生物对 C/N 比、DO 或进水水质变化的快速响应和滞后效应[6][10]-[12]。第六，网络分析和相关分析多停留在关联层面，仍难证明微生物类群与处理性能之间的因果关系[21][22]。

Table 1. Common statistical analysis methods for microbial communities
表 1. 常用微生物群落统计分析方法

微生物分析角度	分析指标	常用统计分析方法
结构分析	α 多样性	Chao1、Shannon、Simpson、ACE 等
	β 多样性	Bray-Curtis、Jaccard、UniFrac 距离等
	差异检验	t 检验、ANOSIM、PERMANOVA 等
	降维排序	PCA、PCoA、NMDS、RDA/CCA 等
	互作网络	Pearson、Spearman 相关、SparCC 等
	群落稳定性	AVD、时间序列稳定性、鲁棒性等
功能分析	功能预测	PICRUSt2、FAPROTAX、FUNGuild 等
	代谢路径	PLS-DA、OPLS-DA、KEGG 富集分析等
	关联分析	Mantel 检验、RDA/CCA、WGCNA 等

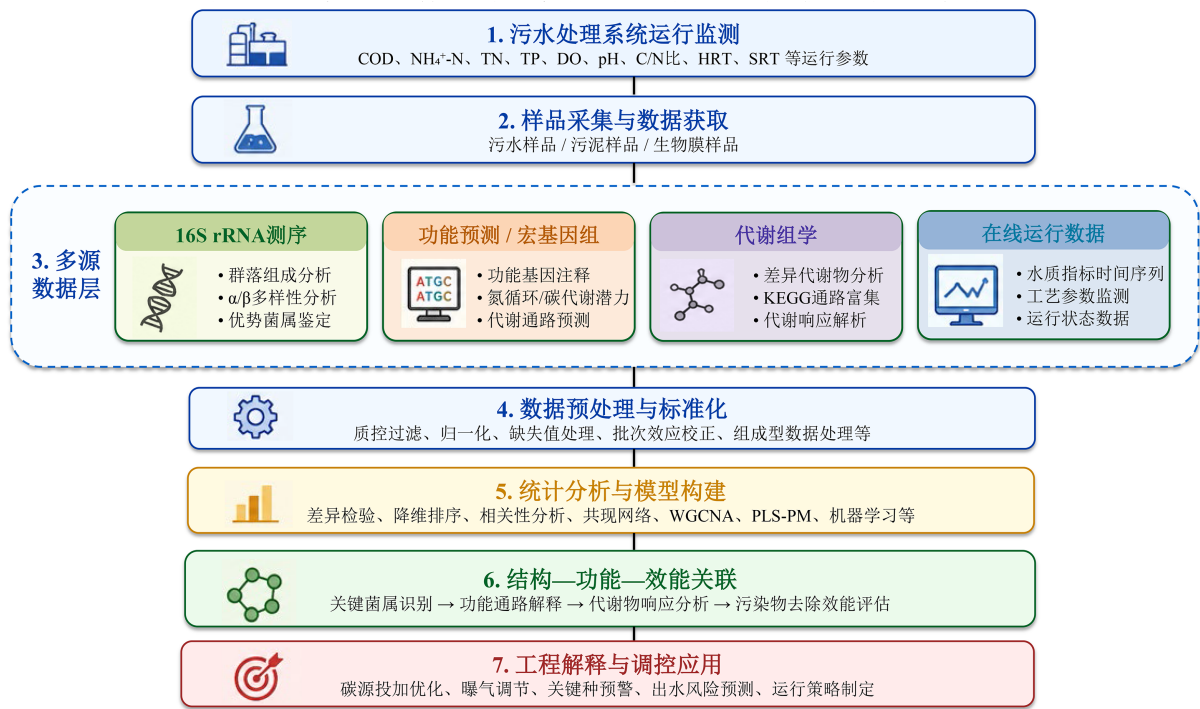


Figure 1. Example flowchart of multi-omics integrated analysis of microbial communities in wastewater treatment
图 1. 污水处理微生物群落多组学整合分析示例流程图

为进一步说明多种统计方法之间的衔接关系，可构建污水处理微生物群落多组学整合分析流程，如图 1 所示。该流程以污水处理系统运行数据为基础，将 COD、NH₄⁺-N、DO、pH、C/N 比等工程参数与 16S rRNA 测序、功能预测、宏基因组或代谢组学数据进行联合分析。首先对不同来源数据进行质控、归一化、缺失值处理和批次效应校正；随后分别开展多样性分析、差异检验、降维排序、共现网络、功能预测和代谢通路富集分析；最后通过相关性分析、WGCNA、PLS-PM 或机器学习模型建立“环境因子 - 微生物群落结构 - 功能潜力 - 代谢响应 - 处理效能”的关联路径。与单一组学分析相比，多组学整合能够更清晰地揭示微生物群落从结构变化到功能表达再到污染物去除效能变化的传导过程。

4.2. 未来研究展望

4.2.1. 动态统计模型

未来污水处理微生物研究需要从静态比较转向动态建模。时间序列网络、动态 RDA、状态空间模型、广义加性模型和滞后相关分析可用于刻画微生物群落对环境因子的非线性响应、记忆效应和滞后效应[6] [13]。例如, C/N 比升高后反硝化菌丰度可能并非立即变化, 而是在数天后通过网络模块重组和代谢物变化体现出来[10]-[12]。若只采用单时间点比较, 容易低估这种过程。动态模型还可以结合在线监测数据, 如 COD、氨氮、硝酸盐、DO 和 pH, 实现对群落变化与水质波动之间关系的连续追踪[3] [6]。例如, 在低 C/N 比污水处理系统中, 若连续监测进水 COD、 $\text{NH}_4^+\text{-N}$ 、C/N 比以及典型反硝化菌、硝化菌和有机物降解菌的相对丰度, 可利用时间滞后相关分析判断微生物群落是否存在延迟响应。假设 C/N 比在第 t 天升高, 而反硝化菌丰度并未在当天立即增加, 而是在 $t+1$ 或 $t+2$ 天达到峰值, 则说明群落结构调整可能存在 1~2 天的滞后效应。进一步地, 可建立广义 Lotka-Volterra 模型, 将不同菌属丰度变化表示为其自身增长、菌群间相互作用以及 C/N 比、DO 等环境因子的共同作用。该模型不仅能够描述某一功能菌对碳源波动的响应速度, 还可用于识别哪些菌属在碳源升高后发挥促进作用, 哪些菌属在资源竞争中受到抑制。由此, 动态模型能够把静态的丰度比较转化为连续的响应过程分析, 使微生物群落对进水波动的滞后响应机制更加具体。

4.2.2. 机器学习融合

机器学习可用于从高维微生物和环境数据中识别关键预测变量。随机森林、极端梯度提升(XGBoost)、支持向量机和深度学习模型能够处理非线性关系和复杂交互, 可用于预测出水水质、识别关键功能菌、判断运行异常或评估碳源投加效果。例如, 以菌属丰度、功能预测、代谢物和环境因子为输入, 以总氮去除率或出水 COD 为输出, 可建立处理性能预测模型[3] [30] [31]。模型解释方面, 可结合变量重要性、置换重要性或 SHAP 值识别最重要的类群和环境因子。

但是, 机器学习模型容易受到样本量不足、批次效应和过拟合影响, 因此需要严格划分训练集和测试集, 进行交叉验证和外部验证[30] [31], 更重要的是, 机器学习识别的是预测贡献, 不等同于因果作用, 未来可引入因果推断框架, 如图因果图、有向无环图和结构因果模型, 区分“真正驱动处理性能变化的因子”和“只与结果共同变化的指标”[29]。

4.2.3. 多组学整合统计

多组学整合是未来研究的重要方向。单一 16S 数据能够提供分类组成, 但不能充分解释功能表达和代谢结果; 代谢组能够反映功能结果, 但难以单独定位产生代谢物的微生物来源[17] [18]。因此, 将微生物组、功能预测、宏基因组、代谢组和环境组联合建模, 可以更完整地揭示污水处理机制, 常用思路包括相关网络、WGCNA、共惯性分析、Procrustes 分析、偏最小二乘回归和偏最小二乘路径模型等。

偏最小二乘路径模型适合构建链式机制假设。例如, 可设置“运行条件/C/N 比”影响“微生物群落结构”, 进而影响“功能潜力/代谢物”, 最终影响“污染物去除性能”的路径, 该模型既能估计直接效应, 也能估计间接效应, 有助于把分散的统计结果整合为较清晰的机制框架[10]-[12]。对于填料强化多级生物接触氧化工艺, 多组学整合还可用于比较悬浮污泥与附着生物膜在功能分工上的差异[7]-[9]。

4.2.4. 统计模型与机理模型融合

未来污水处理微生物研究还应进一步加强统计学方法与传统机理模型的融合。活性污泥模型 (Activated Sludge Models, ASM) 等机理模型能够描述有机物降解、硝化、反硝化和污泥增长等过程, 是污水处理工程设计和运行模拟的重要工具。然而, ASM 模型中的许多动力学参数通常依赖经验取值或批量

实验校准, 对实际污水处理系统中微生物群落结构变化、关键功能菌波动和组学层面的功能响应考虑仍然有限。因此, 将统计学方法与 ASM 等机理模型结合, 有助于提高模型参数识别和工程解释能力。

一方面, 机器学习可用于辅助预测 ASM 模型中的关键动力学参数。例如, 可将进水 COD、 NH_4^+-N 、DO、pH、C/N 比、污泥龄以及关键菌属丰度作为输入变量, 以最大比生长速率、半饱和常数、衰减系数或反硝化速率等参数为输出, 通过随机森林、XGBoost 或神经网络建立参数预测模型, 从而减少完全依赖经验校准带来的不确定性。另一方面, 共现网络 and 关键种识别可用于简化机理模型中的微生物群落描述。通过网络中心性、模块化和功能关联分析, 可筛选与硝化、反硝化、有机物降解密切相关的关键菌属或功能模块, 并将其作为模型中功能菌群划分和参数设定的依据。

此外, 多组学数据还可以为 ASM 模型扩展提供新的信息来源。例如, 功能预测和宏基因组数据可辅助判断氮循环、碳代谢和磷代谢潜力, 代谢组数据可反映碳源利用和关键中间代谢物变化, 在线水质数据则可提供连续的工程运行状态。若能将这些数据与 ASM 模型中的状态变量和过程速率相结合, 就可以构建兼具机理解释和数据驱动预测能力的混合模型。该思路有助于推动微生物生态学与污水处理工程模型的深度融合, 为低 C/N 比污水处理系统的碳源投加、曝气控制、污泥龄调节和运行预警提供更加精准的决策支持。

4.2.5. 工程应用

统计学方法最终应服务于工程运行。未来可将高通量测序、快速 qPCR、在线水质传感器和机器学习预警模型结合, 建立污水处理系统的微生物运行画像。当关键功能菌下降、网络鲁棒性降低或代谢物异常积累时, 系统可提前发出风险提示, 为碳源投加、曝气强度调节、回流比控制和污泥龄调整提供依据。

在 C/N 比精准调控方面, 统计模型可用于识别不同处理阶段的最优碳源需求, 避免外加碳源不足导致反硝化受限, 也避免碳源过量造成运行成本增加和出水 COD 风险。对于关键种强化, 可根据网络中心性、功能关联和稳定性贡献筛选潜在核心类群, 并通过填料优化、运行参数调节或接种策略促进其维持。由此, 统计学方法不只是后期数据分析工具, 也可以成为污水处理系统实时诊断、过程优化和风险预警的重要组成部分。

5. 总结

污水处理微生物群落研究正在从传统的“水质指标评价”逐步转向“结构-功能-环境-性能”综合解析。 α 多样性、 β 多样性、差异检验和降维排序可以揭示群落组成与空间时间差异; 互作网络和稳定性统计能够进一步刻画类群关系、关键种和扰动响应; 功能预测、代谢组统计和功能-物种关联则将结构变化与功能机制连接起来。对于填料强化多级生物接触氧化工艺、C/N 比调控和碳源添加/停止添加等工程问题, 这些方法能够帮助研究者从微生物层面解释处理性能变化。

未来研究应更加重视动态统计模型、机器学习与因果推断、多组学整合以及工程应用转化。特别是在实际污水处理厂中, 应提高采样频率, 结合在线运行数据和多组学数据, 分析微生物群落对环境因子变化的即时响应、滞后效应和恢复过程。只有将统计学方法与工程机制、生态理论和实验验证相结合, 才能更准确地揭示污水处理微生物群落的作用机制, 并为污水处理系统的稳定运行和精准调控提供可靠依据。

基金项目

国家自然科学基金资助项目(No. 52400035)、中央高校基本科研业务费专项资金(No. 2025MS078)。

参考文献

- [1] Cydzik-Kwiatkowska, A. and Zielińska, M. (2016) Bacterial Communities in Full-Scale Wastewater Treatment Systems.

- World Journal of Microbiology and Biotechnology*, **32**, Article No. 66. <https://doi.org/10.1007/s11274-016-2012-9>
- [2] Ferrera, I. and Sánchez, O. (2016) Insights into Microbial Diversity in Wastewater Treatment Systems: How Far Have We Come? *Biotechnology Advances*, **34**, 790-802. <https://doi.org/10.1016/j.biotechadv.2016.04.003>
 - [3] Yang, Y., Wang, L., Xiang, F., Zhao, L. and Qiao, Z. (2020) Activated Sludge Microbial Community and Treatment Performance of Wastewater Treatment Plants in Industrial and Municipal Zones. *International Journal of Environmental Research and Public Health*, **17**, Article 436. <https://doi.org/10.3390/ijerph17020436>
 - [4] Zhang, L., Shen, Z., Fang, W. and Gao, G. (2019) Composition of Bacterial Communities in Municipal Wastewater Treatment Plant. *Science of the Total Environment*, **689**, 1181-1191. <https://doi.org/10.1016/j.scitotenv.2019.06.432>
 - [5] Xia, Y., Wen, X., Zhang, B. and Yang, Y. (2018) Diversity and Assembly Patterns of Activated Sludge Microbial Communities: A Review. *Biotechnology Advances*, **36**, 1038-1047. <https://doi.org/10.1016/j.biotechadv.2018.03.005>
 - [6] Ju, F. and Zhang, T. (2015) Bacterial Assembly and Temporal Dynamics in Activated Sludge of a Full-Scale Municipal Wastewater Treatment Plant. *The ISME Journal*, **9**, 683-695. <https://doi.org/10.1038/ismej.2014.162>
 - [7] Singh, N.K., Kazmi, A.A. and Starkl, M. (2016) Treatment Performance and Microbial Diversity under Dissolved Oxygen Stress Conditions: Insights from a Single Stage IFAS Reactor Treating Municipal Wastewater. *JTICE*, **65**, 197-203. <https://doi.org/10.1016/j.jtice.2016.05.002>
 - [8] Al-Amshawee, S., Yunus, M.Y.B.M., Vo, D.N. and Tran, N.H. (2020) Biocarriers for Biofilm Immobilization in Wastewater Treatments: A Review. *Environmental Chemistry Letters*, **18**, 1925-1945. <https://doi.org/10.1007/s10311-020-01049-y>
 - [9] Waqas, S., Bilal, M.R., Man, Z., Wibisono, Y., Jaafar, J., Indra Mahlia, T.M., *et al.* (2020) Recent Progress in Integrated Fixed-Film Activated Sludge Process for Wastewater Treatment: A Review. *Journal of Environmental Management*, **268**, Article ID: 110718. <https://doi.org/10.1016/j.jenvman.2020.110718>
 - [10] Chen, Y., Li, B., Ye, L. and Peng, Y. (2015) The Combined Effects of COD/N Ratio and Nitrate Recycling Ratio on Nitrogen and Phosphorus Removal in Anaerobic/Anoxic/Aerobic (A²/O)-Biological Aerated Filter (BAF) Systems. *Biochemical Engineering Journal*, **93**, 235-242. <https://doi.org/10.1016/j.bej.2014.10.005>
 - [11] Fu, Z., Yang, F., Zhou, F. and Xue, Y. (2009) Control of COD/N Ratio for Nutrient Removal in a Modified Membrane Bioreactor (MBR) Treating High Strength Wastewater. *Bioresource Technology*, **100**, 136-141. <https://doi.org/10.1016/j.biortech.2008.06.006>
 - [12] Mannina, G., Ekama, G.A., Capodici, M., Cosenza, A., Di Trapani, D. and Ødegaard, H. (2017) Moving Bed Membrane Bioreactors for Carbon and Nutrient Removal: The Effect of C/N Variation. *Biochemical Engineering Journal*, **125**, 31-40. <https://doi.org/10.1016/j.bej.2017.05.005>
 - [13] Anderson, M.J. (2001) A New Method for Non-Parametric Multivariate Analysis of Variance. *Austral Ecology*, **26**, 32-46. <https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x>
 - [14] Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V. and Egozcue, J.J. (2017) Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, **8**, Article 2224. <https://doi.org/10.3389/fmicb.2017.02224>
 - [15] Schloss, P.D. (2020) Reintroducing Mothur: 10 Years Later. *Applied and Environmental Microbiology*, **86**, e02343-19. <https://doi.org/10.1128/aem.02343-19>
 - [16] Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., *et al.* (2010) QIIME Allows Analysis of High-Throughput Community Sequencing Data. *Nature Methods*, **7**, 335-336. <https://doi.org/10.1038/nmeth.f.303>
 - [17] Pang, Z., Chong, J., Zhou, G., de Lima Morais, D.A., Chang, L., Barrette, M., *et al.* (2021) Metaboanalyst 5.0: Narrowing the Gap between Raw Spectra and Functional Insights. *Nucleic Acids Research*, **49**, W388-W396. <https://doi.org/10.1093/nar/gkab382>
 - [18] Worley, B. and Powers, R. (2013) Multivariate Analysis in Metabolomics. *Current Metabolomics*, **1**, 92-107. <https://doi.org/10.2174/2213235x11301010092>
 - [19] Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. and Morishima, K. (2017) KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs. *Nucleic Acids Research*, **45**, D353-D361. <https://doi.org/10.1093/nar/gkw1092>
 - [20] Langfelder, P. and Horvath, S. (2008) WGCNA: An R Package for Weighted Correlation Network Analysis. *BMC Bioinformatics*, **9**, Article No. 559. <https://doi.org/10.1186/1471-2105-9-559>
 - [21] Friedman, J. and Alm, E.J. (2012) Inferring Correlation Networks from Genomic Survey Data. *PLOS Computational Biology*, **8**, e1002687. <https://doi.org/10.1371/journal.pcbi.1002687>
 - [22] Faust, K., Sathirapongsasuti, J.F., Izard, J., Segata, N., Gevers, D., Raes, J., *et al.* (2012) Microbial Co-Occurrence Relationships in the Human Microbiome. *PLOS Computational Biology*, **8**, e1002606. <https://doi.org/10.1371/journal.pcbi.1002606>
 - [23] Shade, A., Peter, H., Allison, S.D., Baho, D.L., Berga, M., Bürgmann, H., *et al.* (2012) Fundamentals of Microbial Community Resistance and Resilience. *Frontiers in Microbiology*, **3**, Article 417. <https://doi.org/10.3389/fmicb.2012.00417>

-
- [24] Oksanen, J., Simpson, G., Blanchet, F., Kindt, R., Legendre, P., Minchin, P., *et al.* (2026) Vegan: An R Package for Community Ecologists. R Package Version 2.8-0. <https://vegandevs.github.io/vegan/>
- [25] Xun, W., Liu, Y., Li, W., Ren, Y., Xiong, W., Xu, Z., *et al.* (2021) Specialized Metabolic Functions of Keystone Taxa Sustain Soil Microbiome Stability. *Microbiome*, **9**, Article No. 35. <https://doi.org/10.1186/s40168-020-00985-9>
- [26] Louca, S., Parfrey, L.W. and Doebeli, M. (2016) Decoupling Function and Taxonomy in the Global Ocean Microbiome. *Science*, **353**, 1272-1277. <https://doi.org/10.1126/science.aaf4507>
- [27] Nguyen, N.H., Song, Z., Bates, S.T., Branco, S., Tedersoo, L., Menke, J., *et al.* (2016) Funguild: An Open Annotation Tool for Parsing Fungal Community Datasets by Ecological Guild. *Fungal Ecology*, **20**, 241-248. <https://doi.org/10.1016/j.funeco.2015.06.006>
- [28] Tenenhaus, M., Vinzi, V.E., Chatelin, Y. and Lauro, C. (2005) PLS Path Modeling. *Computational Statistics & Data Analysis*, **48**, 159-205. <https://doi.org/10.1016/j.csda.2004.03.005>
- [29] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. <https://doi.org/10.1145/2939672.2939785>
- [30] Pearl, J. (2009) Causal Inference in Statistics: An Overview. *Statistics Surveys*, **3**, 96-146. <https://doi.org/10.1214/09-ss057>
- [31] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. <https://doi.org/10.1023/a:1010933404324>