



MicroDIG: 基于自监督学习的大规模微生物组纵向数据有向互作估计

献给钱敏平教授 86 寿辰

喻姚洋^{1,2†}, 章文浩^{1†}, 孙丰珠³, 王颖^{1,2,4*}

1. 厦门大学自动化系, 厦门 361101;

2. 厦门大学国家健康医疗大数据研究院, 厦门 361101;

3. Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA;

4. 厦门市大数据智能分析与决策重点实验室, 厦门 361101

† 同等贡献

E-mail: yuyaoyang@stu.xmu.edu.cn, zscotty@stu.xmu.edu.cn, fsun@usc.edu, wangying@xmu.edu.cn

收稿日期: 2024-09-25; 接受日期: 2025-03-07; 网络出版日期: 2025-04-10; * 通信作者

国家自然科学基金 (批准号: 62173282 和 62472363) 资助项目

摘要 微生物群落内的物种通过复杂的作用关系互相影响、相互依存, 维持动态平衡. 探索微生物群落的物种作用, 对理解群落功能和机理十分重要. 本文基于微生物群落的纵向时序物种丰度数据, 开发一种自监督学习框架 MicroDIG (microbiome directed interaction on GLV), 融合数学机理和智能学习估计并校验大规模微生物群落内物种间的有向作用关系. MicroDIG 利用 GLV (generalized Lotka-Volterra) 方程进行数据模拟与增强, 通过自编码器学习物种间的有向作用强度, 并设计有效性检验策略对有向作用强度进行统计校验, 以获得具有高置信度的物种间有向作用关系. 本文通过多角度多层次的仿真数据及人类肠道微生物群落的真实数据对 MicroDIG 的性能进行系统、综合评估, 实验表明, MicroDIG 的推断精度优于五类基准方法, 并且其能够有效降低估计结果的假阳性率, 同时能够准确识别物种之间的作用方向、去除作用关系传递造成的假阳性, 并适用于各种群落规模. 本文所提出的自监督学习框架, 不需要具备标准答案的训练数据, 模型具有清晰的可解释性和灵活的可拓展性, 可拓展到大规模微生物群落真实场景中, 实现更加精细的作用关系探索.

关键词 微生物间有向作用 自监督学习 自编码器 微生物组 纵向数据

MSC (2020) 主题分类 62F30, 62P10, 92B10

1 引言

微生物群落广泛存在于人体与自然界中, 其物种组成、结构改变、动态特性与其表型和功能息息相关 (参见文献 [19]). 微生物群落影响着人类健康的维持和疾病的发生、发展, 参与人体消化、代谢和

英文引用格式: Yu Y Y, Zhang W H, Sun F Z, et al. MicroDIG: Estimating directed interactions in large-scale microbiome longitudinal data based on self-supervised learning (in Chinese). Sci Sin Math, 2025, 55: 1277–1300, doi: 10.1360/SSM-2024-0293

免疫调节等多种生理功能, 与肥胖、心血管疾病、结肠癌、类风湿性关节炎和炎症性肠病等息息相关 (参见文献 [11])。同时, 微生物群落通过参与生态系统中的养分循环、污染物降解、土壤健康维护等进程, 影响、塑造地球的生态功能 (参见文献 [6])。群落内部的微生物并非孤立存在, 而是相互影响、相互作用, 构成动态的作用关系网络 (参见文献 [8])。解析群落中物种的有向作用网络, 对于探索群落特征、分析群落表型与功能、研究群落与外界环境或宿主的关系十分重要。

然而, 估计微生物群落中物种的作用关系十分具有挑战性, 受到生命科学、信息和数学领域的研究者的广泛关注。物种作用关系^[21] 包括共生 (symbiosis)、捕食 (predation) 和寄生 (parasitism) 等。传统方法通过湿实验培养少量类型微生物, 观察其生长行为推断微生物之间的作用关系 (参见文献 [30]), 但该方法受到可培养微生物种类稀少和时间成本等限制。随着高通量测序技术的发展, 整个微生物群落可以通过 16S 核糖体 RNA (rRNA) 测序^[12]、宏基因组测序^[25] 等技术获得序列信息, 对应的序列信息通过参考数据库配准、序列聚类等手段处理可以获得生物类群 OTU (operational taxonomic unit) 的丰度信息, 从而为研究群落的组成、物种间的作用关系提供重要的数据支持。

基于微生物群落的丰度矩阵, 经典的相关性统计方法 Pearson^[24] 和 Spearman^[28] 能够估计物种间的相关性, 但由于丰度数据本身存在固有的组分性 (compositionality) 和稀疏性 (sparsity)^[8, 15] (即所有物种丰度加和为定值, 且矩阵存在大量零值), 因此 Pearson 和 Spearman 的分析结果中会存在大量的假阳性 (参见文献 [18])。为此, 生物统计学家针对微生物丰度数据开发了专门的相关性度量方法, 例如, SparCC^[9] 通过稀疏相关矩阵迭代来估计微生物丰度间的显著关联, CCLasso^[7] 利用稀疏协方差估计和 LASSO (least absolute shrinkage and selection operator) 回归来捕捉微生物丰度数据中的条件独立结构。这类相关性方法旨在从不包含时间信息的独立样本所组成的静态丰度数据中推断出物种间丰度的相关性, 从而判断物种间的作用关系。但由于缺乏时间信息, 该相关性忽视物种间作用的有向性与时滞性, 仅能识别出物种间的共存竞争关系或共现互斥模式, 推断结果有偏差且是无向的 (参见文献 [1])。

同一微生物群落在时间维数上的多次采样所产生的纵向 (longitudinal) 丰度数据为研究物种间的有向作用关系提供重要的信息。针对时序数据, 孙丰珠团队^[33] 提出的 eLSA (extended local similarity analysis) 可处理多重样本时间序列, 基于数据特性发现局部的、存在时间滞后的物种共现关系, 但该工具只有在物种间的丰度存在明显时间滞后关系时才能进行物种间的有向作用关系分析。在时序数据的基础上, 为了进一步引入微生物群落物种作用的生态系统机理描述, 可以使用 Lotka-Volterra 方程和 GLV (generalized Lotka-Volterra)^[29] 方程来建模宏观和微观的生态系统中物种丰度的动态变化, 进而探索微生物群落的动态特性, 挖掘群落物种间的作用关系。GLV 方程被生物学家们广泛运用于微生物群落的建模过程中^[32], 具有成熟的理论和生物湿实验基础^[13, 31], 大量的研究表明, GLV 方程能在各类不同的微生物环境中准确地建模预测对应微生物群落的时序动态, 如湖泊^[5]、小鼠肠道^[22] 和奶酪制作环境^[23] 等。GLV 方程定义了群落物种间的有向作用, 可以用于预测群落中微生物丰度随时间的变化, 并用于微生物群落物种的丰度数据模拟, 例如, 工具 miaSim^[10] 能基于给定的物种间有向作用关系生成微生物群落时序丰度数据; 另外, 针对已知的群落时序丰度数据, 可通过对 GLV 方程的参数估计来辨识物种间的有向作用关系, 如基于物种绝对丰度数据的工具 MDSINE^[2]、MetaMIS^[27] 和 web-glv^[17] 以及基于物种相对丰度数据的工具 CLV^[14]。然而, 上述工具通常基于最小二乘算法进行有向作用关系的参数辨识, 本质为多元方程的拟合求解, 故针对微生物群落中物种数量庞大 (一般情况下, 物种数量超过 40) 的场景, 无法进行有效的有向作用关系估计, 只适用于小规模微生物群落。

长期以来, 微生物群落纵向丰度数据的作用关系分析始终受到数据量少以及物种间作用关系标签缺失两方面的挑战。自监督学习作为一种无需大量人工标注的学习范式, 展现出深远的跨学科应用潜

力, 已经在蛋白质结构预测、药物发现和基因功能注释以及时间序列分析等领域得到了应用, 其无需人工标注的特点和对不同任务的适应性, 为解决微生物时序丰度数据不足以及作用关系标签缺失的问题提供了全新的思路. 受此启发, 本文基于微生物群落的时序丰度纵向数据, 在物种丰度与有向作用关系满足 GLV 方程的前提假设下, 尝试将 GLV 方程的数学机理与神经网络相融合, 建立自监督的微生物作用关系学习框架 MicroDIG (microbiome directed interaction on GLV), 估计物种的有向作用关系. MicroDIG 以 GLV 方程驱动的数据模拟策略自监督式地进行数据增强, 通过自编码器学习物种间的有向作用关系强度, 并设计估计结果的有效性校验, 降低结果的假阳性率, 从而获得大规模微生物群落中物种间的有向作用关系. 本文在不同角度和不同层次的仿真数据以及人类肠道微生物组时序数据上的实验表明, MicroDIG 的推断精度优于其他五类基准方法, 并且其有效地降低了估计结果中的假阳性, 同时能够准确识别两两物种之间作用关系的有向性、去除作用关系传递造成的假阳性, 并能够适用于各种群落规模. 在此基础上, MicroDIG 还具有清晰的可解释性和灵活的可拓展性, 其整体框架基于 GLV 方程定义的微生物动态时变特性设计, 模型的估计输出能够清晰地从微生物作用关系角度进行解释; 同时框架内置的自编码器也可根据使用场景和性能需求灵活地替换成诸如 GCN (graph convolutional networks), Transformer 或 LSTM (long short-term memory) 等网络模型, 对 MicroDIG 进行扩展与升级. 通过知识和数据共驱动, MicroDIG 将小规模无标签数据下的方程参数估计问题转化为大数据驱动下通过数据增强的自监督学习, 不需要具备标准答案的训练数据, 模型具有清晰的可解释性和灵活的可拓展性, 可拓展到大规模复杂微生物群落真实场景中更加精细的作用关系探索. 为方便复现和进一步研究, MicroDIG 工具的代码已公开, 访问地址为 <https://github.com/Ying-Lab/MicroDIG>.

2 MicroDIG 方法描述

2.1 微生物群落的时序纵向丰度数据描述

针对一个微生物群落, 其组成可以用操作分类单元 (operational taxonomic unit, OTU) 作为单元进行描述, OTU 是一种在微生物学和生态研究中常用的群落组成的代表性单位. 群落内部随时间变化的 OTU 丰度数据可以表示为 (2.1) 所示的矩阵 $\mathbf{X}$, 作为 MicroDIG 的标准输入数据:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_N \end{bmatrix} = \begin{bmatrix} x_1(t_1) & x_1(t_2) & \cdots & x_1(t_T) \\ x_2(t_1) & x_2(t_2) & \cdots & x_2(t_T) \\ \vdots & \vdots & \ddots & \vdots \\ x_N(t_1) & x_N(t_2) & \cdots & x_N(t_T) \end{bmatrix}, \quad (2.1)$$

其中, $\mathbf{X}$ 是一个 $N \times T$ 的矩阵, N 为 OTU 个数, T 代表时间点个数, $x_i(t_\tau)$ ($i = 1, \dots, N; \tau = 1, \dots, T$) 表示 OTU_{*i*} 在时间点 t_τ 的丰度, $\mathbf{X}_i \in \mathbb{R}_0^{+1 \times T}$ ($i = 1, \dots, N$) 表示 OTU_{*i*} 在 T 个时间点的丰度向量, $[t_1, t_2, \dots, t_T]$ 为非均匀采样点.

微生物群落是一个动态平衡的生态系统, 在某个扰动下物种的丰度会波动然后逐步趋近于一个稳态. 然而, 由于稳态后时间序列包含的动态信息不足, 在系统辨识中通常需要将不同初始情况下的时序数据拼接起来以获得独立的时变序列 (参见文献 [3]), 因此 OTU_{*i*} 的丰度向量 $\mathbf{X}_i$ 如 (2.2) 所示, 由 K 个初始丰度不同的 OTU_{*i*} 丰度时序变化向量拼接而成:

$$\mathbf{X}_i = [\mathbf{X}_i^{(1)} \ \mathbf{X}_i^{(2)} \ \cdots \ \mathbf{X}_i^{(K)}], \quad (2.2)$$

其中 $\mathbf{X}_i^{(k)}$ ($k = 1, \dots, K$) 表示 $\mathbf{X}_i$ 中第 k 个丰度初值对应的时序丰度向量, 则 $\mathbf{X}^{(k)}$ 如 (2.3) 所示, 代表群落内所有 OTU 在第 k 个丰度初值下共同构成的丰度矩阵:

$$\mathbf{X}^{(k)} = \begin{bmatrix} \mathbf{X}_1^{(k)} \\ \mathbf{X}_2^{(k)} \\ \vdots \\ \mathbf{X}_N^{(k)} \end{bmatrix}. \quad (2.3)$$

本文采用 GLV 方程^[27] 描述微生物群落中 OTU 丰度与物种有向作用关系的动态时变特性, OTU_{*i*} 的丰度 $x_i(t)$ 随时间 t 的变化可表示为

$$\frac{d}{dt}x_i(t) = u_i x_i(t) + x_i(t) \sum_{j=1}^N M_{ij} x_j(t), \quad (2.4)$$

其中, u_i 表示 OTU_{*i*} 自身的生长速率, M_{ij} 表示 OTU_{*j*} 对 OTU_{*i*} 的有向作用强度, 将 (2.4) 除以 x_i 并离散, 可以得到 GLV 方程的离散形式, 如

$$\frac{\Delta \ln x_i(t_\tau)}{\Delta t_\tau} = u_i + \sum_{j=1}^N M_{ij} x_j(t_\tau), \quad (2.5)$$

最终群落所有 N 个 OTU 间的有向作用关系表示为矩阵 $\mathbf{M}$, 所有 OTU 的生长速率用向量 $\mathbf{u}$ 表示:

$$\mathbf{M} = \begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1N} \\ M_{21} & M_{22} & \cdots & M_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ M_{N1} & M_{N2} & \cdots & M_{NN} \end{bmatrix} \in \mathbb{R}^{N \times N}, \quad (2.6)$$

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix} \in \mathbb{R}^{+N \times 1}. \quad (2.7)$$

2.2 MicroDIG 整体框架

本文提出基于自监督的微生物作用关系学习框架 MicroDIG, 基于群落的 OTU 丰度时序矩阵 $\mathbf{X}$ 估计出 OTU 间的有向作用关系 $\mathbf{M}$. MicroDIG 由自监督数据增强、有向作用关系估计和有效性校验三个模块串行构成, 如图 1 所示.

- 自监督数据增强模块, 采用自监督数据模拟, 从群落的原始丰度矩阵中采样物种单元 OTU_{*j*} 的丰度数据作为模拟数据的已知时序丰度数据, 并基于采样设定的物种间有向作用强度 $\widetilde{M}_{ij}$ 利用 GLV 方程生成虚拟 $\widetilde{\text{OTU}}_i$ 的时序丰度数据, 从而实现无需实际群落中的有向作用强度作为参考标注的训练数据生成.

- 有向作用关系估计模块, 基于深度网络强大的非线性模拟能力, 从增强训练数据中学习获得物种间的有向作用强度的估计值 $\hat{M}_{ij}$.

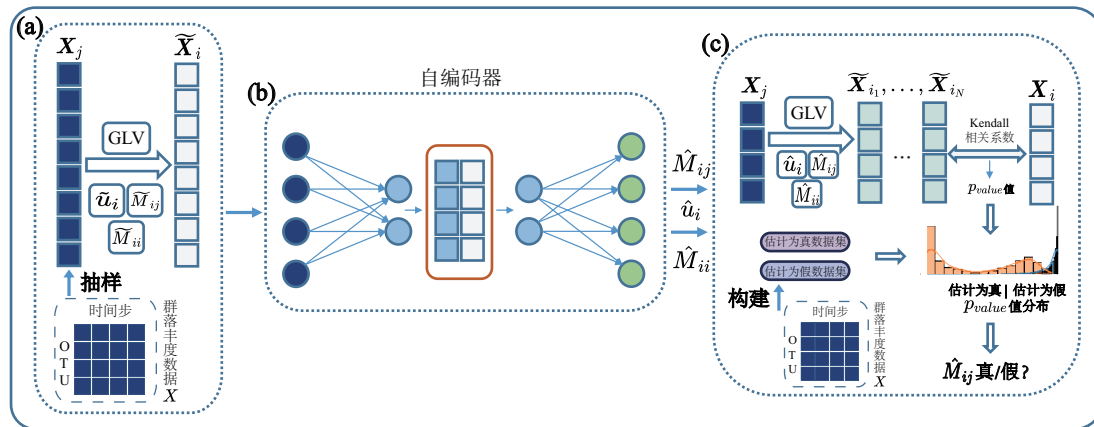


图 1 (网络版彩图) MicroDIG 框架图. (a) 自监督数据增强模块, 从原始丰度时序矩阵 X 中随机采样抽取物种单元 OTU_j , 基于其丰度数据 X_j 和采样设定的物种间有向作用强度 $[\tilde{M}_{ij}, \tilde{u}_i, \tilde{M}_{ii}]$ 生成虚拟 OTU_i 的丰度数据 $\tilde{X}_i$, $[X_j, \tilde{X}_i]$ 为一对具有 $[\tilde{M}_{ij}, \tilde{u}_i, \tilde{M}_{ii}]$ 作用关系标签的丰度时序配对数据. (b) 有向作用关系估计模块, 将配对数据 $[X_j, \tilde{X}_i]$ 输入自编码器, 其对应的作用标签 $[\tilde{M}_{ij}, \tilde{u}_i, \tilde{M}_{ii}]$ 作为参考输出, 进行神经网络的训练. (c) 有效性校验模块, 构建估计值为真/假时生成丰度向量与真实丰度向量的相关性显著 p 值分布, 从而基于极大似然估计校验某个估计值 $\hat{M}_{ij}$ 的真/假

• 有效性校验模块, 针对学习获得的有向作用强度估计值 $\hat{M}_{ij}$, 通过仿真生成估计值为真/假时的两类 p 值分布, 通过极大似然估计来判断 OTU_j 和 OTU_i 有向作用强度估计值的真/假, 降低结果的假阳性.

2.3 自监督数据增强模块

在现实应用场景中, 通过测序手段采集获得的微生物群落数据, 无法获得物种间有向作用强度的标准答案, 无法进行有监督的训练. 因此, 本文提出自监督数据增强策略, 在微生物时序丰度与物种间有向作用强度满足 GLV 方程的前提假设下, 生成大量基于采样获得物种间有向作用强度的配对 OTU 时序丰度模拟数据, 供训练有向作用关系估计网络使用.

本质上, 群落 N 个 OTU 间的有向作用关系和生长速率的估计是寻找最优 M 和 u 使得 (2.8) 目标函数取最小值. 本文通过最小二乘算法可以对 M 和 u 进行粗略估计 (使用 Python.scipy 包中的算法 optimize.leastsq 进行估算), 从而获得 M 和 u 的估计值 M_{init} 和 u_{init} , 如 (2.8) 所示:

$$M_{\text{init}}, u_{\text{init}} = \arg \min_{M, u} \|(M u)Y - F\|_2^2, \quad (2.8)$$

其中,

$$Y = \begin{bmatrix} x_1(t_1) & x_1(t_2) & \cdots & x_1(t_T) \\ x_2(t_1) & x_2(t_2) & \cdots & x_2(t_T) \\ \vdots & \vdots & \ddots & \vdots \\ x_N(t_1) & x_N(t_2) & \cdots & x_N(t_T) \\ 1 & 1 & \cdots & 1 \end{bmatrix}, \quad (2.9)$$

$$F = \begin{bmatrix} \frac{\Delta \ln x_1(t_1)}{\Delta t_1} & \frac{\Delta \ln x_1(t_2)}{\Delta t_2} & \dots & \frac{\Delta \ln x_1(t_T)}{\Delta t_T} \\ \frac{\Delta \ln x_2(t_1)}{\Delta t_1} & \frac{\Delta \ln x_2(t_2)}{\Delta t_2} & \dots & \frac{\Delta \ln x_2(t_T)}{\Delta t_T} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\Delta \ln x_N(t_1)}{\Delta t_1} & \frac{\Delta \ln x_N(t_2)}{\Delta t_2} & \dots & \frac{\Delta \ln x_N(t_T)}{\Delta t_T} \end{bmatrix}. \quad (2.10)$$

随后, 基于 $\mathbf{M}_{\text{init}}$ 和 $\mathbf{u}_{\text{init}}$, 自监督数据增强模块通过单次模拟生成一组具备物种间有向作用强度标准答案的配对物种单元 ($\text{OTU}_j, \widetilde{\text{OTU}}_i$) 的丰度数据 $[\mathbf{X}_j, \widetilde{\mathbf{X}}_i]$, 其中 OTU_j 为群落随机抽取得到的真实 OTU, 其对应的丰度数据为 $\mathbf{X}_j$; $\widetilde{\text{OTU}}_i$ 为模拟生成的虚拟 OTU, 其对应的丰度数据为 $\widetilde{\mathbf{X}}_i$. 本文均用 $\sim$ 表示仿真生成的虚拟对象. 模拟步骤如下.

(1) 从丰度矩阵 $\mathbf{X}$ 中随机抽取物种单元 OTU_j 的丰度向量 $\mathbf{X}_j$.

(2) 参考 Schluter 等^[26] 针对有向作用关系 M_{ij} 和生长速率 u_i 的研究, 本文设置 OTU 的生长速率服从正态分布 $N(\mu_1, \sigma_1^2)$, 其中 $\mu_1 = \text{mean}_i(\mathbf{u}_{\text{init}}(i))$, $\sigma_1^2 = \text{var}_i(\mathbf{u}_{\text{init}}(i))$; OTU 间的有向作用关系服从均匀分布 $U(a_1, b_1)$, 其中 $a_1 = \min_{i,j}(\mathbf{M}_{\text{init}}(i, j))$, $b_1 = \max_{i,j}(\mathbf{M}_{\text{init}}(i, j))$; OTU 对自身的作用强度服从均匀分布 $U(a_2, b_2)$, 其中 $a_2 = \min_{i,j}(\text{diag}(\mathbf{M}_{\text{init}}(i, j)))$, $b_2 = \max_{i,j}(\text{diag}(\mathbf{M}_{\text{init}}(i, j)))$.

(3) 通过如下采样方式获得有向作用强度 $[\widetilde{M}_{ij}, \widetilde{u}_i, \widetilde{M}_{ii}, \widetilde{M}_{ji}, \widetilde{u}_j, \widetilde{M}_{jj}]$: OTU_j 对 $\widetilde{\text{OTU}}_i$ 的有向作用 $\widetilde{M}_{ij} \sim U(a_1, b_1)$, $\widetilde{\text{OTU}}_i$ 对 OTU_j 的有向作用 $\widetilde{M}_{ji} = 0$, $\widetilde{\text{OTU}}_i$ 和 OTU_j 的自我生长速率 $\widetilde{u}_i \sim N(\mu_1, \sigma_1^2)$ 和 $\widetilde{u}_j \sim N(\mu_1, \sigma_1^2)$, $\widetilde{\text{OTU}}_i$ 和 OTU_j 对自身的作用强度 $\widetilde{M}_{ii} \sim U(a_2, b_2)$ 和 $\widetilde{M}_{jj} \sim U(a_2, b_2)$.

(4) 基于 $\mathbf{X}_j$ 和 $[\widetilde{M}_{ij}, \widetilde{u}_i, \widetilde{M}_{ii}, \widetilde{M}_{ji}, \widetilde{u}_j, \widetilde{M}_{jj}]$, 通过 GLV 方程来生成虚拟 $\widetilde{\text{OTU}}_i$ 的时序丰度 $\widetilde{\mathbf{X}}_i$. $\mathbf{X}_j$ 如 (2.2) 所示, 由 K 个时序丰度向量拼接而成, 针对每一个 $\mathbf{X}_j^{(k)}$, 可通过求解微分方程组 (2.11) 生成虚拟 $\widetilde{\text{OTU}}_i$ 的时序丰度序列 $\widetilde{\mathbf{X}}_i^{(k)}$:

$$\begin{cases} \frac{d}{dt} \widetilde{x}_i(t) = \widetilde{u}_i \widetilde{x}_i(t) + \widetilde{M}_{ij} \widetilde{x}_i(t) \widetilde{x}_j(t) + \widetilde{M}_{ii} \widetilde{x}_i(t) \widetilde{x}_i(t), \\ \frac{d}{dt} \widetilde{x}_j(t) = \widetilde{u}_j \widetilde{x}_j(t) + \widetilde{M}_{jj} \widetilde{x}_j(t) \widetilde{x}_j(t), \\ \widetilde{x}_i(t_1) = \frac{1}{T} \sum_{\tau=1}^T x_j(t_\tau), \\ \widetilde{x}_j(t_\tau) = x_j(t_\tau), \quad \tau = 1, \dots, T. \end{cases} \quad (2.11)$$

(5) 将 K 个 $\widetilde{\mathbf{X}}_i^{(k)}$, $k = 1, \dots, K$ 拼接得到 $\widetilde{\mathbf{X}}_i = [\widetilde{\mathbf{X}}_i^{(1)} \ \widetilde{\mathbf{X}}_i^{(2)} \ \dots \ \widetilde{\mathbf{X}}_i^{(K)}]^T$, 从而生成了一对具备物种间有向作用强度标准值 $[\widetilde{M}_{ij}, \widetilde{u}_i, \widetilde{M}_{ii}]$ 的丰度数据 $[\mathbf{X}_j, \widetilde{\mathbf{X}}_i]$.

为简化后文的描述, 本文将 (2.11) 所描述的基于 GLV 微分方程、 $\mathbf{X}_j$ 和 $[\widetilde{M}_{ij}, \widetilde{u}_i, \widetilde{M}_{ii}]$ 求解 $\widetilde{\mathbf{X}}_i$ 的生成过程记为 $\widetilde{\mathbf{X}}_i = \text{GLV}(\mathbf{X}_j, \widetilde{M}_{ij}, \widetilde{u}_i, \widetilde{M}_{ii})$, 同时令 $\mathbf{h} = [\mathbf{X}_j, \widetilde{\mathbf{X}}_i]$, $\zeta = [\widetilde{M}_{ij}, \widetilde{u}_i, \widetilde{M}_{ii}]$. 考虑到现实微生物群落有向作用强度的复杂性, 即微生物群落中可能存在部分生长速率高度相似、有向作用关系却具有显著差异或生长速率具有显著差异、有向作用关系却高度相似的微生物, 本文在上述有向作用强度 $[\widetilde{M}_{ij}, \widetilde{u}_i, \widetilde{M}_{ii}, \widetilde{M}_{ji}, \widetilde{u}_j, \widetilde{M}_{jj}]$ 的采样过程中额外生成了一系列 OTU_j 对 $\widetilde{\text{OTU}}_i$ 的有向作用 $\widetilde{M}_{ij}$ 相同、 $\widetilde{\text{OTU}}_i$ 的自我生长速率 $\widetilde{u}_i$ 不同或者有向作用 $\widetilde{M}_{ij}$ 不同、自我生长速率 $\widetilde{u}_i$ 相同的作用关系情况进行数据模拟, 保证最终生成的丰度数据 $[\mathbf{X}_j, \widetilde{\mathbf{X}}_i]$ 涵盖现实场景中复杂的作用关系情况.

基于单次模拟过程 $\text{GLV}(\mathbf{X}_j, \widetilde{M}_{ij}, \widetilde{u}_i, \widetilde{M}_{ii})$ 可以构建出一组包含 OTU_j 和虚拟 $\widetilde{\text{OTU}}_i$ 时序丰度数据以及对应的作用关系强度的自监督配对数据 $(\mathbf{h}, \zeta)$. 重复随机抽取不同的 OTU_j 生成对应的虚拟 $\widetilde{\text{OTU}}_i$, 最终构造出带有标准有向作用强度信息的配对 OTU 时序丰度数据集, 即 $\mathcal{H} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_S]$

和 $\xi = [\zeta_1, \zeta_2, \dots, \zeta_S]$, S 为抽样次数, 即数据增强生成的配对数据个数.

MicroDIG 通过自监督数据增强模块, 将已知的先验知识、数学机理融入数据模拟过程中, 使得 MicroDIG 模型的训练无要求实际微生物群落具有有向作用关系的标准答案, 其具备自监督自训练的能力. 具体的数据模拟算法可参见附录 D.

2.4 基于自编码器的物种间有向作用关系估计

MicroDIG 通过自编码器架构, 将自监督数据生成模块生成的配对 OTU 时序丰度数据作为训练样本, 进行物种间有向作用强度的估计. 如图 2 所示, 基于某个微生物群落的丰度时序数据 $\mathbf{X}$, MicroDIG 数据增强得到训练数据 $\mathcal{H} = [h_1, h_2, \dots, h_S]$ 和 $\xi = [\zeta_1, \zeta_2, \dots, \zeta_S]$, 通过编码器将时序配对数据 h_k 编码映射到特定的低维隐空间 $\mathbf{Z}_k$ 中, 以提取 h_k 中最本质的时变特征信息:

$$\mathbf{Z}_k = \phi(W_e \cdot h_k + b_e), \quad (2.12)$$

同时为保证低维隐空间 $\mathbf{Z}_k$ 中特征提取的准确性, 通过解码器对配对数据 h_k 进行解码重构:

$$\hat{h}_k = \sigma(W_d \cdot \mathbf{Z}_k + b_d), \quad (2.13)$$

其中, W_e 和 W_d 是权重矩阵, b_e 和 b_d 是偏置向量, $\hat{h}_k$ 表示模型重构的配对数据, $\phi(\cdot)$ 和 $\sigma(\cdot)$ 是 ReLU 激活函数. 隐空间通过去除 h_k 中冗余的特征信息、降低数据本身存在的时滞特性来估计配对物种间的作用强度 $\hat{\zeta}_k$:

$$\hat{\zeta}_k = \phi_m(W_m \cdot \text{flatten}(\mathbf{Z}_k) + b_m), \quad (2.14)$$

其中, W_m 是权重矩阵, b_m 是偏置向量, $\phi_m(\cdot)$ 为 ReLU 激活函数.

自编码器的训练损失函数 L 由两部分组成, 即

$$L = m \cdot L_1 + n \cdot L_2, \quad (2.15)$$

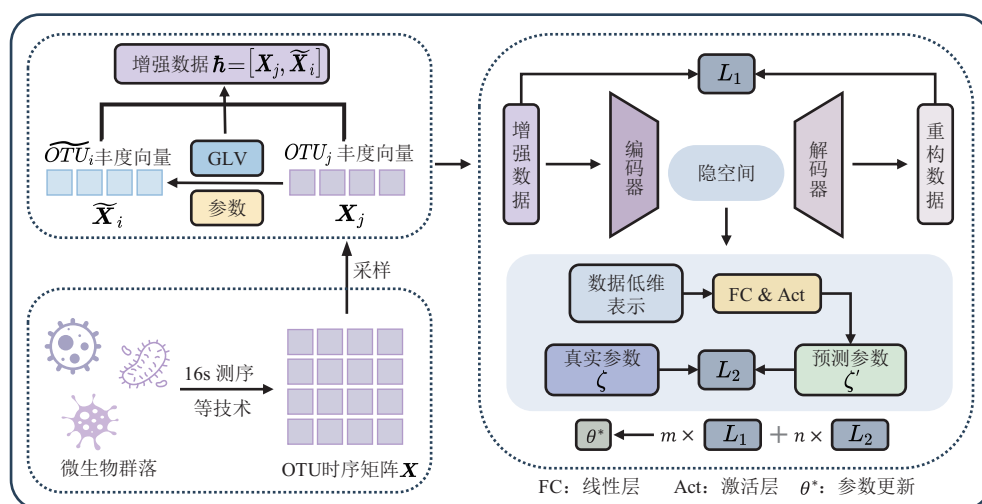


图 2 (网络版彩图) MicroDIG 有向作用估计网络结构图. 微生物群落通过测序技术获得的丰度时序数据 $\mathbf{X}$, 输入自监督数据增强模块, 可以获得包含作用关系标签的大量配对数据 $\mathcal{H}$ 作为训练数据, 通过数据编解码重构和作用关系估计来拟合训练神经网络

其中 m 和 n 为两个损失的权重; L_1 是自监督配对数据 $\mathcal{H}$ 的重构损失, 这里 $L_1 = \frac{1}{S} \sum_{k=1}^S (\hat{h}_k - \hat{h}_k)^2$; L_2 是 $\mathcal{H}$ 中 OTU_j 对 OTU_i 的作用强度 ξ 的估计损失, 这里 $L_2 = \frac{1}{S} \sum_{k=1}^S (\zeta_k - \hat{\zeta}_k)^2$.

MicroDIG 通过设置初始学习率为 0.001、重构损失权重 m 为 0.4、估计损失 n 为 0.6 对自编码器进行训练, 学习率随着训练迭代轮次的增加而自适应衰减, 最终得到使目标损失函数最小的参数 θ^* , 如

$$\theta^* = \arg \min_{\theta} (m \cdot L_1 + n \cdot L_2), \quad (2.16)$$

从而实现有向作用关系的估计. 神经网络的网络结构设置见附录 B.

2.5 有向作用强度估计值的有效性校验

微生物群落的物种作用关系网络通常是稀疏的, 许多 OTU 之间的作用关系微弱或不存在, 但 MicroDIG 的自编码器训练时仅针对有真正作用关系的微生物对, 因此针对任意一对 OTU_j 和 OTU_i 的丰度时序数据 $h = [\mathbf{X}_j, \mathbf{X}_i]$, 无论 OTU_j 对 OTU_i 是否存在真正的作用关系, 自编码器一定会输出一个作用关系的估计值 $\hat{M}_{ij}$, 因此需要建立严格的有效性检验, 去除估计结果的假阳性, 获得物种间真正有效的作用关系.

本文对微生物群落作用关系假设如 (2.11) 所示, 任一物种 OTU_i 仅受到 OTU_j 的作用, 基于以下思路对估计结果进行有效性校验.

当 OTU_j 对 OTU_i 存在作用关系 M_{ij} 且估计值 $\hat{M}_{ij}$ 与真实值 M_{ij} 越接近时, 由于 OTU_i 仅受一个物种作用, 基于 $\mathbf{X}_j$ 和 $\hat{M}_{ij}$ 所生成的丰度 $\tilde{\mathbf{X}}_i = \text{GLV}(\mathbf{X}_j, \hat{M}_{ij}, \hat{u}_i, \hat{M}_{ii})$ 与 OTU_i 的真实丰度 $\mathbf{X}_i$ 的相关性越显著, 此时判断估计值 $\hat{M}_{ij}$ 为真.

当 OTU_j 对 OTU_i 没有作用关系或估计值 $\hat{M}_{ij}$ 与真实值 M_{ij} 差别较大时, 基于 $\mathbf{X}_j$ 和 $\hat{M}_{ij}$ 所生成的丰度 $\tilde{\mathbf{X}}_i = \text{GLV}(\mathbf{X}_j, \hat{M}_{ij}, \hat{u}_i, \hat{M}_{ii})$ 与 OTU_i 的真实丰度 $\mathbf{X}_i$ 的相关性越差, 此时判断估计值 $\hat{M}_{ij}$ 为假.

因此, 校验模块通过相关性显著度的假设检验来判断作用关系估计值 $\hat{M}_{ij}$ 的真假. 如图 3(a) 和 3(c) 所示, 基于输入整个群落的丰度时序矩阵 $\mathbf{X}$, 构造估计值 $\hat{M}_{ij}$ 为真/假两个数据集, 从而通过相关性计算、检验来构建当前群落下真/假估计集合的 p 值分布. 如图 3(b), 实际应用中, 当输入任意一对 OTU_j 和 OTU_i 的丰度向量 $[\mathbf{X}_j, \mathbf{X}_i]$ 时, 自编码器估计获得 $\hat{M}_{ij}$, 通过 $\tilde{\mathbf{X}}_i = \text{GLV}(\mathbf{X}_j, \hat{M}_{ij}, \hat{u}_i, \hat{M}_{ii})$ 计算基于 $\mathbf{X}_j$ 和 $\hat{M}_{ij}$ 所生成的 $\widetilde{\text{OTU}}_i$ 的仿真丰度 $\tilde{\mathbf{X}}_i$, 随后计算获得 $\tilde{\mathbf{X}}_i$ 与 $\mathbf{X}_i$ 相关性的显著性, 从而判断 $\hat{M}_{ij}$ 的真假, 实现估计值的有效性校验. 具体步骤如下.

(1) 基于群落丰度时序矩阵 $\mathbf{X}$, 构建估计值为真/假时的 p 值分布 (图 3(c)).

- 针对群落随时间变化的 OTU 丰度矩阵 $\mathbf{X}$, 随机抽取出 OTU_p , 其对应的丰度向量记为 $\mathbf{X}_p$, 并基于自监督数据增强模块中提到的正态分布和均匀分布采样参数集合 $[\tilde{M}_{qp}, \tilde{u}_q, \tilde{M}_{qq}]$, 通过 $\tilde{\mathbf{X}}_q = \text{GLV}(\mathbf{X}_p, \tilde{M}_{qp}, \tilde{u}_q, \tilde{M}_{qq})$ 计算获得仿真生成 $\widetilde{\text{OTU}}_q$ 的丰度 $\tilde{\mathbf{X}}_q$. 因此 $(\text{OTU}_p, \widetilde{\text{OTU}}_q)$ 是具有确定作用关系 $\tilde{M}_{qp}$ 的一对 OTU, 且 $\widetilde{\text{OTU}}_q$ 只受 OTU_p 作用, 其丰度为 $[\mathbf{X}_p, \tilde{\mathbf{X}}_q]$.

- 将 $\tilde{\mathbf{X}}_q$ 保持不变, 在 OTU 丰度矩阵 $\mathbf{X}$ 中随机抽取 OTU_k ($k \neq p$), 其对应的丰度为 $\mathbf{X}_k$, 则 $(\text{OTU}_k, \widetilde{\text{OTU}}_q)$ 是无作用关系的一对 OTU, 其丰度为 $[\mathbf{X}_k, \tilde{\mathbf{X}}_q]$.

- 将 $[\mathbf{X}_p, \tilde{\mathbf{X}}_q]$ 和 $[\mathbf{X}_k, \tilde{\mathbf{X}}_q]$ 输入训练完毕的自编码器中, 分别输出估计值 $\hat{M}_{qp}$ 和 $\hat{M}_{qk}$. 由前文可知, $\hat{M}_{qp}$ 是针对确实存在作用关系的 $(\text{OTU}_p, \widetilde{\text{OTU}}_q)$ 所估计得到的, 因此估计值为真; 而 $(\text{OTU}_k, \widetilde{\text{OTU}}_q)$ 并不存在作用关系, 因此 $\hat{M}_{qk}$ 为假.

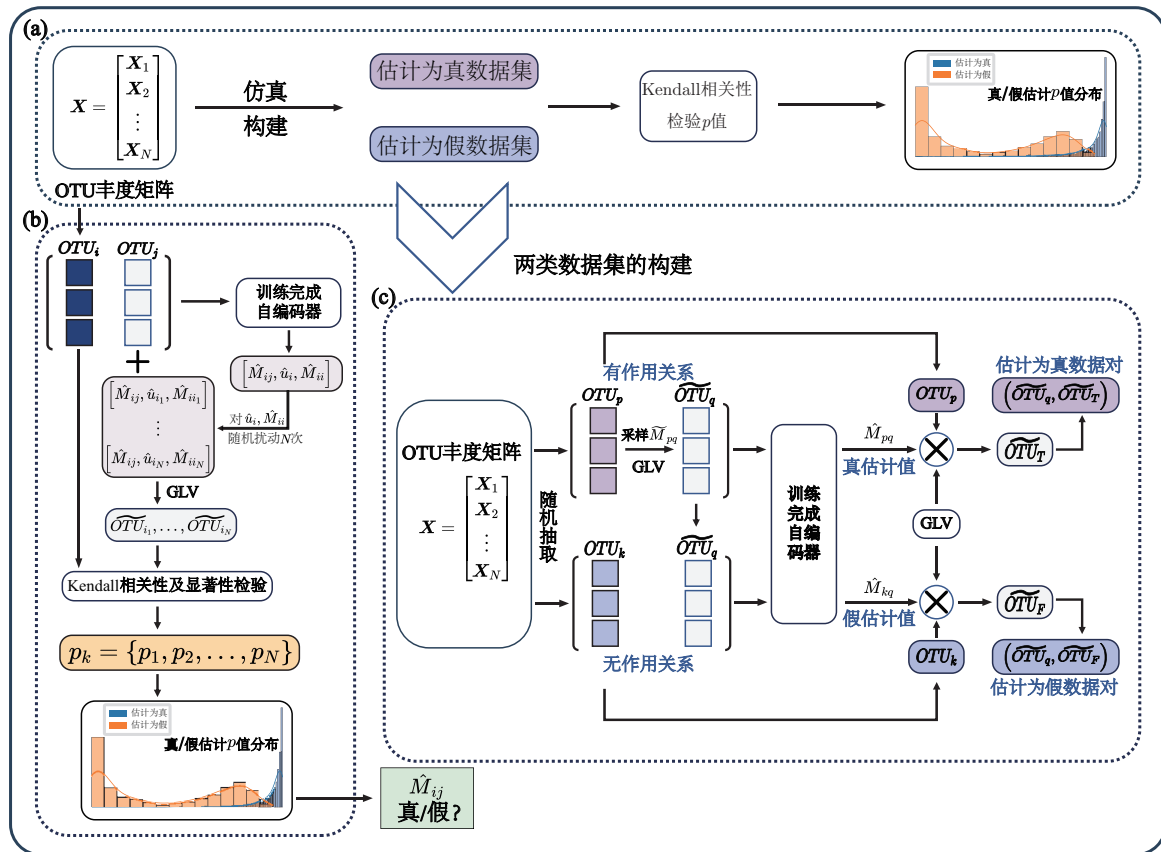


图 3 (网络版彩图) MicroDIG 校验模块流程图. (a) MicroDIG 构建估计为真和估计为假两类数据集, 计算数据集内每一组数据之间的 Kendall 相关性及其检验 p 值, 获得当前群落下真/假估计集合的 p 值分布. (b) 针对当前群落下任意的一组配对 OTU 丰度数据, 通过自编码器得到估计的 $[\hat{M}_{ij}, \hat{u}_i, \hat{M}_{ii}]$, 对 $\hat{u}_i$ 和 $\hat{M}_{ii}$ 施加 N 次扰动, 并根据扰动后的 GLV 参数生成对应的不同虚拟 OTU 的丰度 $[\hat{X}_{i1}, \hat{X}_{i2}, \dots, \hat{X}_{iN}]$, Kendall 相关性计算及 p 值检验得到对应的 p 值, 基于极大似然估计判断有向作用关系的估计值 $\hat{M}_{ij}$ 为真/假. (c) 两类数据集的具体构造流程, 通过原始群落随机抽取、自编码器估计、估计值 GLV 仿真等流程构造数据集

• 基于 $\text{GLV}(\mathbf{X}_p, \hat{M}_{qp}, \hat{u}_q, \hat{M}_{qq})$ 计算生成 $\widetilde{\text{OTU}}_T$ 的丰度 $\widetilde{\mathbf{X}}_T$, 这里 T 表示 True, 即基于真估计值仿真获得的丰度值, 则 $[\widetilde{\mathbf{X}}_q, \widetilde{\mathbf{X}}_T]$ 应具有显著相关性; 基于 $\text{GLV}(\mathbf{X}_k, \hat{M}_{qk}, \hat{u}_q, \hat{M}_{qq})$ 计算生成 $\widetilde{\text{OTU}}_F$ 的丰度 $\widetilde{\mathbf{X}}_F$, 这里 F 表示 False, 即基于假估计值仿真获得的丰度值, 则 $[\widetilde{\mathbf{X}}_q, \widetilde{\mathbf{X}}_F]$ 相关性弱.

• 使用 Kendall 相关性系数^[16] 及其检验 p 值来描述 $[\widetilde{\mathbf{X}}_q, \widetilde{\mathbf{X}}_T]$ 时序丰度之间以及 $[\widetilde{\mathbf{X}}_q, \widetilde{\mathbf{X}}_F]$ 时序丰度之间的相关性, 其计算方式见附录 A.

• 重复执行以上过程, 构建估计值为真/假时的 p 值分布 d_{pT} 和 d_{pF} .

(2) 针对任意一对 OTU_j 和 OTU_i 的有向作用估计值 $\hat{M}_{ij}$ 的检验 (图 3(b)).

• 将 OTU_j 和 OTU_i 的丰度时序向量 $[\mathbf{X}_j, \mathbf{X}_i]$ 输入训练完毕的自编码器, 输出 $[\hat{M}_{ij}, \hat{u}_i, \hat{M}_{ii}]$. 对 $\hat{u}_i$ 和 $\hat{M}_{ii}$ 施加 10% 以内的随机扰动, 生成 $[\hat{M}_{ij}, \hat{u}_{ik}, \hat{M}_{ii,k}], k = 1, \dots, N$. 基于 $\text{GLV}(\mathbf{X}_j, \hat{M}_{ij}, \hat{u}_{ik}, \hat{M}_{ii,k})$ 计算获得 $[\widetilde{\mathbf{X}}_{i1}, \widetilde{\mathbf{X}}_{i2}, \dots, \widetilde{\mathbf{X}}_{iN}]$.

• 计算 $[\mathbf{X}_i, \widetilde{\mathbf{X}}_{ik}]$ 的 Kendall 相关性系数 γ_k 及对应的置信度检验值 $p_k, k = 1, \dots, N$. 基于极大似然判断出 p 值序列 $\{p_1, p_2, \dots, p_N\}$ 来自估计值为真时的 p 值分布 d_{pT} 还是估计值为假时的分布 d_{pF} , 从而判断 OTU_j 和 OTU_i 有向作用关系的估计值 $\hat{M}_{ij}$ 为真/假.

3 实验结果

3.1 实验数据集的构造

为全面综合评估 MicroDIG 对微生物群落物种间有向作用关系估计的性能, 本文分别在微生物群落丰度模拟数据集和人类肠道宏基因组真实测序数据集上进行测试. 模拟数据集包括四类物种有向作用关系模式, 并涵盖不同的群落规模. 如表 1 所示, 群落规模从 40 个 OTU 至 100 个 OTU 共 7 种规模; 每种群落规模下, 建立从简单到复杂的四种微生物群落物种作用模式, 依次为“一对一”“级联”“一对多”和“混合模式”, 产生 28 个微生物群落丰度模拟数据集. 此外, 在 100 个 OTU 的群落规模下, 我们额外构造了共 9 个从 50 个到 250 个不同采样时间点个数 (采样时间点个数变化步长为 25 个采样时间点) 的模拟数据集, 检验时序长度对 MicroDIG 性能的影响.

为避免自监督数据增强所采用的仿真策略影响评估的客观性, 本文采用第三方微生物群落时序丰度模拟工具 miaSim^[10] 构造模拟微生物群落丰度数据集, 具体构造流程如图 4(a) 所示. 对于四种不同的微生物群落物种作用模式, 通过设置该模式下的作用关系矩阵 $\mathbf{M}$ 和生长速率向量 $\mathbf{u}$, 来进行丰度数据的模拟生成 (具体模式及其参数设置详见附录 C). 其中, miaSim 可基于 GLV 方程, 根据一组物种丰度初值、生长速率向量 $\mathbf{u}$ 和作用关系矩阵 $\mathbf{M}$ 生成一个微生物群落丰度时序矩阵, 时序长度可根据需求调整. 对于同一个微生物群落丰度数据集, 本文通过设置不同的物种丰度初值, 配合相同的 $\mathbf{u}$ 和 $\mathbf{M}$, 使用 miaSim 生成多个丰度时序矩阵. 通过对不同丰度时序矩阵的时间点进行采样得到 (2.3) 所定义的 $\mathbf{X}^{(k)}$, 并将采样结果拼接作为一个最终的微生物群落丰度数据集矩阵输入 $\mathbf{X}$, 从而模拟真实群落环境下多次扰动后群落的动态变化.

本文所分析的人体肠道群落真实丰度数据来自健康女性受试者在 6 个月内对肠道 (粪便) 进行取样测序得到的共计 131 个时间点的肠道微生物组时序丰度数据 (参见文献 [4]), 用于验证 MicroDIG 在真实的人类肠道微生物群落数据中的有向作用关系识别效果.

3.2 实验设计

实验流程如图 4(b) 所示, 输入模拟微生物群落丰度数据集的丰度矩阵 $\mathbf{X}$, MicroDIG 首先通过自监督数据增强模块生成 6,000 组配对数据 $\mathcal{H}$, 并将 $\mathcal{H}$ 按照 8 : 2 的比例划分训练集和验证集. MicroDIG 的有向作用关系估计网络基于训练集数据进行迭代训练, 并通过有效性检验模块构建出两类具有显著差异性的分布, 随后可在验证集上初步评估 MicroDIG 的各类性能, 最终将模拟微生物群落丰度数据集本身作为测试集输入 MicroDIG, 得到 $\mathbf{X}$ 中经过校验后的高置信物种作用强度.

表 1 模拟微生物群落丰度数据集设计

群落规模 (单位: OTU 个数)	物种作用模式
40	一对一模式: $\mathbf{X}_j \rightarrow \mathbf{X}_i$
50	
60	级联模式: $\mathbf{X}_j \rightarrow \mathbf{X}_{i_1} \rightarrow \mathbf{X}_{i_2}$
70	
80	一对多模式: $\mathbf{X}_j \rightarrow \mathbf{X}_{i_1}, \mathbf{X}_j \rightarrow \mathbf{X}_{i_2}, \mathbf{X}_j \rightarrow \mathbf{X}_{i_3}$
90	
100	混合模式: 一对一模式、级联模式、一对多模式

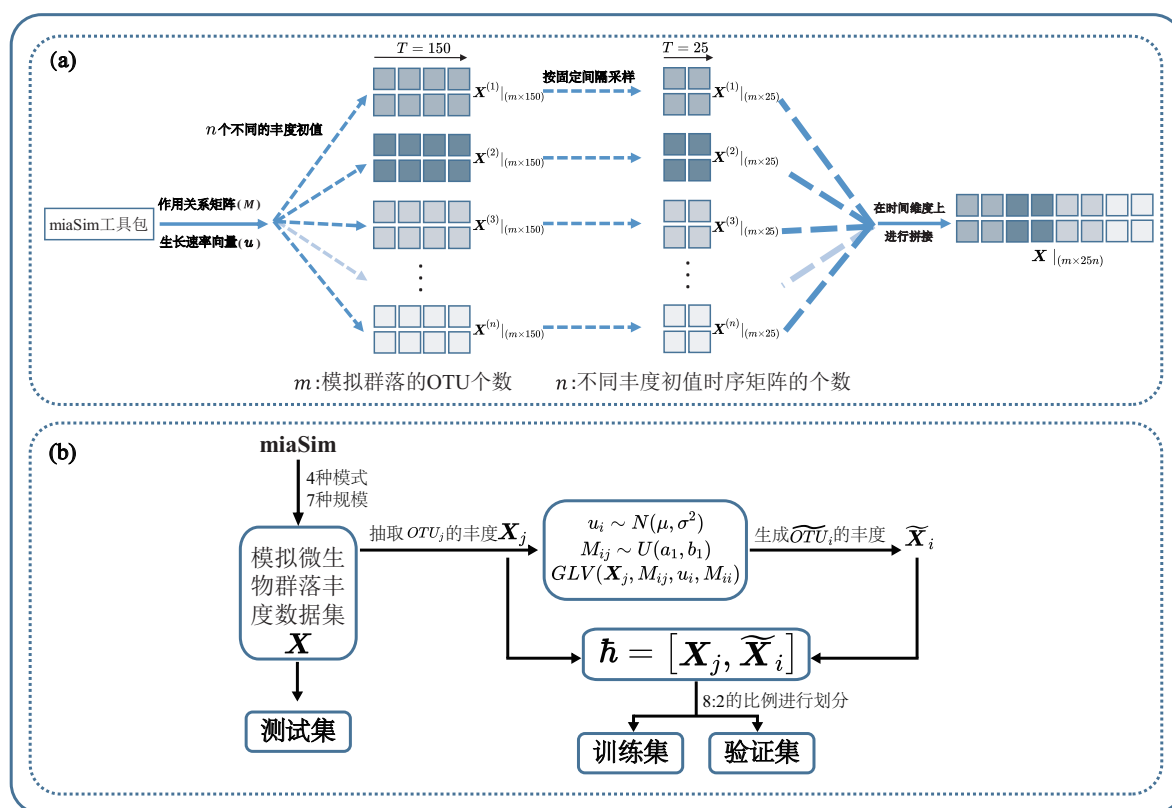


图 4 (网络版彩图) 实验基础流程图。(a) 给定一个群落物种作用关系矩阵 M 、生长速率向量 u 和 n 个不同的丰度初值，miaSim 工具包对应生成 n 个丰度时序矩阵，每个矩阵包含 150 个时间点，对其按照固定时间点采样至 25 个时间点，并在时间维数上进行拼接，得到最终的模拟微生物群落丰度数据。(b) 本文中的实验流程为：一个模拟微生物群落丰度数据集 X 输入 MicroDIG 框架，通过自监督数据增强获得大量包含作用关系标签的配对数据 $h = [X_j, \tilde{X}_i]$ ，按照 8 : 2 的比例划分为训练集和验证集进行作用关系估计网络的拟合训练，并在训练好的网络模型的基础上进行作用关系估计值有效性的校验，最终对模拟微生物群落丰度数据集本身（测试集）进行测试，得到其对应的物种作用关系。

为全面评估 MicroDIG 的性能，本文引入传统的相关性度量 Pearson^[24]、微生物领域的相关性算法 SparCC^[9] 和 CCLasso^[7] 以及基于 GLV 方程的参数辨识算法 GLV-based models^[14] 和 CLV^[13] 等五类基准方法，如表 2 所示。

MicroDIG 框架所对应的模型和分析实验均运行在 1 块 12 GB 的 NVIDIA GeForce GTX 1080Ti GPU 以及 32 核的 Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz CPU 的硬件设备上，具体的模型训练时间、计算资源消耗和硬件依赖见附录 E。

表 2 五类微生物间作用关系估计的基准方法

基准模型	描述
Pearson ^[24]	Pearson 相关性系数
Sparcc ^[9]	分析微生物群落中不同成员之间的相关性模式的统计工具
CCLasso ^[7]	估计高维数据中的稀疏相关性结构的统计方法
GLV-based models ^[14]	基于 GLV 方程参数辨识的作用关系估计方法
CLV ^[14]	将 GLV 方程参数辨识由绝对丰度扩展到相对丰度的方法

3.3 实验一: MicroDIG 能够准确估计大规模群落物种间有向作用关系

本文在 28 个模拟微生物群落丰度数据集上验证 MicroDIG 准确估计 OTU 间作用关系的能力. 使用 R^2 衡量数据中有向作用关系估计值 $\hat{M}_{ij}$ 与作用关系标签 M_{ij} 之间的拟合程度, 其中 N 为群落中 OTU 的个数, $\text{mean}(M_{ij}) = \frac{1}{N^2} \sum_{i,j=1}^N M_{ij}$ 表示所有作用关系标签 M_{ij} 的均值 (当 R^2 为负值时, 代表模型的预测误差低于均值预测):

$$R^2 = 1 - \frac{\sum_{i,j=1}^N (M_{ij} - \hat{M}_{ij})^2}{\sum_{i,j=1}^N (M_{ij} - \text{mean}(M_{ij}))^2}. \quad (3.1)$$

由于验证集和训练集属于同源数据, 本文首先在验证集上整体评估 MicroDIG 估计物种间作用关系的能力. 本文将同一作用模式下不同物种规模的模拟微生物群落丰度数据集对应的验证集上的结果进行整合, 分别对不同作用模式下 MicroDIG 取得的 R^2 结果绘制箱线图, 如图 5(a) 所示, MicroDIG 估计结果 $\hat{M}_{ij}$ 与标签 M_{ij} 的 R^2 均在 0.99 左右, 极高的拟合度证明了 MicroDIG 在各类复杂物种作用关系的微生物群落中能够准确地估计物种间的作用关系强度. 以 100 个 OTU 的物种规模为例, 图 5(b) 展示了在大规模物种数据集上的 MicroDIG 的估计性能, MicroDIG 估计的作用关系 $\hat{M}_{ij}$ 与作用关系标签 M_{ij} 所绘制的散点都分布于斜渐近线附近, 并且极少出现离散点, 证明了 MicroDIG 在大规模物种的微生物群落上的分析能力.

在独立的模拟微生物群落丰度数据集, 即测试集上, 本文与五类基准方法进行比较, 对 MicroDIG 的作用关系强度估计性能进行了更全面综合的评估, 除 R^2 外, 本文还采用 PCC (Pearson product-moment correlation coefficient) 进行评估:

$$\text{PCC} = \frac{\sum_{i,j=1}^N (M_{ij} - \text{mean}(M_{ij}))(\hat{M}_{ij} - \text{mean}(\hat{M}_{ij}))}{\sqrt{\sum_{i,j=1}^N (M_{ij} - \text{mean}(M_{ij}))^2 \sum_{i,j=1}^N (\hat{M}_{ij} - \text{mean}(\hat{M}_{ij}))^2}}, \quad (3.2)$$

其中 N 为群落中 OTU 的个数, $\text{mean}(M_{ij}) = \frac{1}{N^2} \sum_{i,j=1}^N M_{ij}$ 表示所有作用关系标签 M_{ij} 的均值, $\text{mean}(\hat{M}_{ij}) = \frac{1}{N^2} \sum_{i,j=1}^N \hat{M}_{ij}$ 表示所有作用关系估计值 $\hat{M}_{ij}$ 的均值.

如图 6 所示, 在四种物种作用模式的模拟微生物群落丰度数据集中, 当群落规模较小 (OTU 个数为 40 或 50) 时, Pearson^[24], CCLasso^[7], GLV 和 CLV^[14] 在特定的物种作用模式数据下取得了与 MicroDIG 相差无几的性能, 对应的 R^2 和 PCC 指标也处于较高值; 但当群落规模逐步扩大到 OTU 个数为 100 时, 五类基准方法的性能出现明显下降, 而 MicroDIG 仍能保持最优的估计效果. 同时, MicroDIG 不同物种作用模式下的大规模群落作用关系估计散点图也密集分布于对角线两侧, 充分展示了其优越的估计性能.

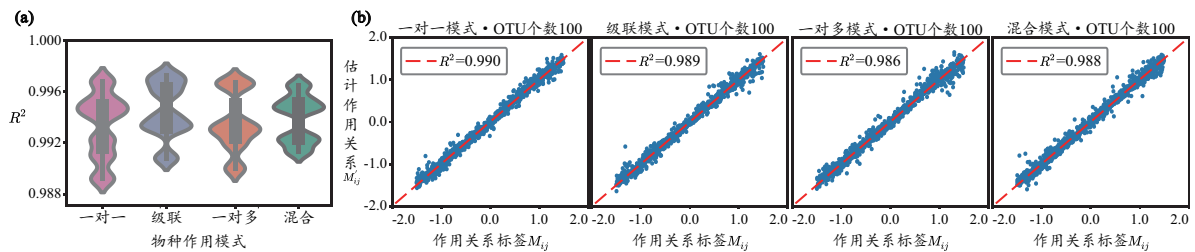


图 5 (网络版彩图) 网络训练效果图. (a) 28 个模拟微生物群落丰度数据集生成的验证集按照物种作用模式划分成四类, 每一类验证集中的数据由 MicroDIG 估计得到的估计结果 $\hat{M}_{ij}$ 与标签 M_{ij} 之间的 R^2 结果箱线图; (b) 以 100 个 OTU 的物种规模为例, 四类不同物种作用模式的估计结果 $\hat{M}_{ij}$ 与标签 M_{ij} 的散点图

本文在测试集上进一步分析 MicroDIG 与基准方法在四类不同物种作用模式下的平均 R^2 值随物种群落规模增大的变化趋势, 如图 7 所示. MicroDIG 在所有不同群落规模的测试集中都取得了最好的估计效果, 其性能不受群落规模大小的影响, 估计结果的 R^2 折线图曲线较为平稳, 而 Pearson^[24] 和

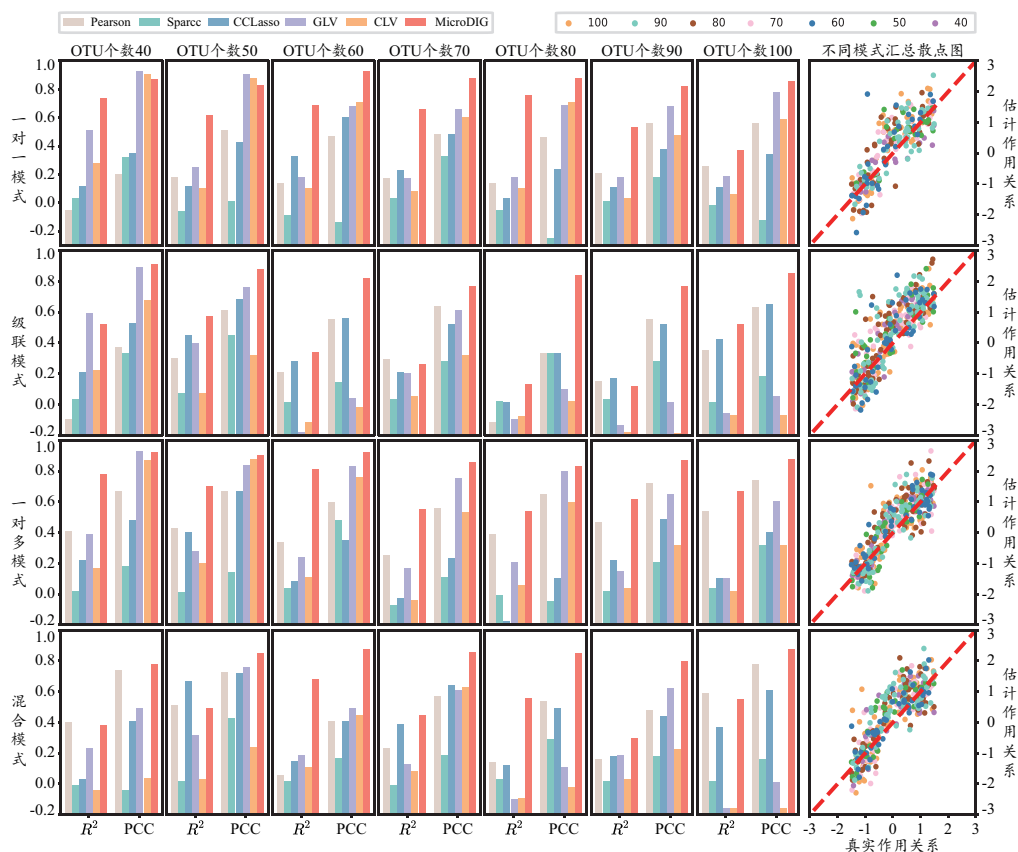


图 6 MicroDIG 与五类基准方法对比图. 柱状图: 在 28 个模拟微生物群落丰度数据集中, MicroDIG 与五类基准方法在 R^2 和 PCC 两类指标上的性能对比. 散点图: MicroDIG 在四类不同物种作用模式下的估计结果 $\hat{M}_{ij}$ 与标签 M_{ij} 的散点图, 散点的不同颜色代表不同的群落规模

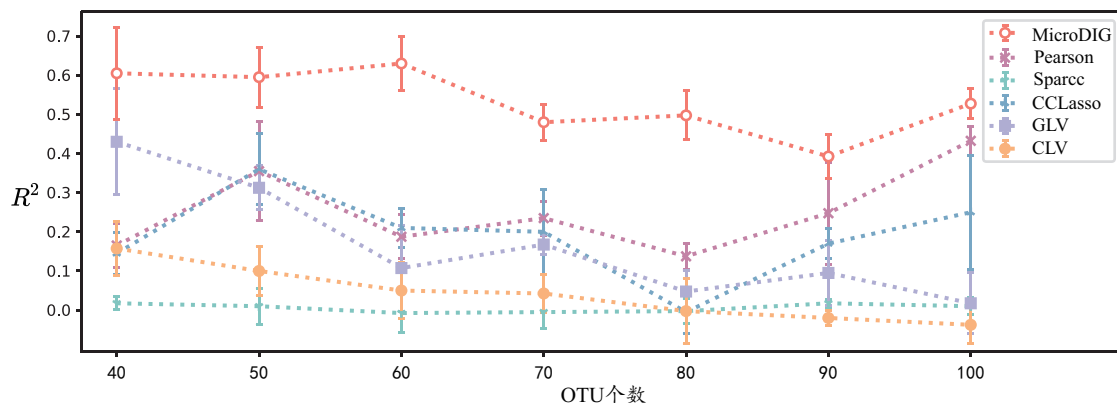


图 7 MicroDIG 与五类基准方法的平均性能分析折线图. MicroDIG 与五类基准方法在四类不同物种作用模式下的平均性能 (平均 R^2 值) 随物种群落规模增大的变化趋势折线图. 折线的不同颜色和散点图案表示不同的方法, MicroDIG 在所有群落规模下都取得了最佳的估计效果

CCLasso^[7] 在不同规模的测试集中的估计性能波动较大, SparCC^[9] 即使在较小规模群落上也无法进行有效的作用关系估计, GLV 和 CLV^[14] 效果随着群落规模的增大而显著下降.

在验证集和测试集的实验表明, MicroDIG 对群落间作用关系的估计优于另外五类基准方法, 且不受群落规模大小的影响, 在进行大规模物种群落分析时, 具有更明显的优势.

3.4 实验二: MicroDIG 能够处理不同时序长度的微生物群落数据

群落的丰度时序数据包含了 OTU 变化的丰富信息, MicroDIG 通过时间维数上的数据信息波动挖掘 OTU 间的有向作用关系. 因此, 本实验评估不同采样时间点个数对 MicroDIG 的性能影响, 在使用 miaSim^[10] 构建模拟微生物群落丰度数据集时, 通过控制不同物种丰度初值生成的微生物群落丰度矩阵的数量, 得到一批群落 OTU 个数为 100、物种作用模式为混合模式、采样时间点个数在 50 到 250 之间波动变化的模拟微生物群落丰度数据集进行实验. 如图 8 所示, 本文分析在不同采样时间点个数下 MicroDIG 估计的作用关系 $\hat{M}_{ij}$ 与作用关系标签 M_{ij} 之间的 R^2 值, 结果表明, 当输入数据仅有 50 个采样时间点时, 其包含的信息过少, MicroDIG 无法作出有效的作用强度估计; 当采样时间点提升到 75 个并逐步扩展到 250 个时, MicroDIG 在时间点足够的情况下得到了准确的估计结果, 并且其效果随采样时间点的增多而得到提升. 可见, MicroDIG 的物种作用关系估计效果随着时序数据采样点的增多而提升.

3.5 实验三: MicroDIG 能够对作用关系强度的估计值进行有效性校验

本实验在 28 个模拟微生物群落丰度数据集的验证集和测试集上对 MicroDIG 的有效性检验模块进行性能评估, 测试 MicroDIG 是否能准确校验作用关系存在与否, 从而获得物种间高置信度的有向作用关系.

本文通过随机置换 (permutation) 从验证集中构造不具有作用关系的配对数据组成负样本集, 与对应的验证集正样本一同输入校验模块, 评估 MicroDIG 准确校验作用关系存在与否的性能. 如图 9(a) 所示, 群落规模为 100 个 OTU 的四类不同物种作用模式的验证集得到的混淆矩阵反映出较高的准确率、召回率和精确率; 图 9(b) 则显示所有 28 个验证集数据中正负样本集校验性能的总体箱线分布图,

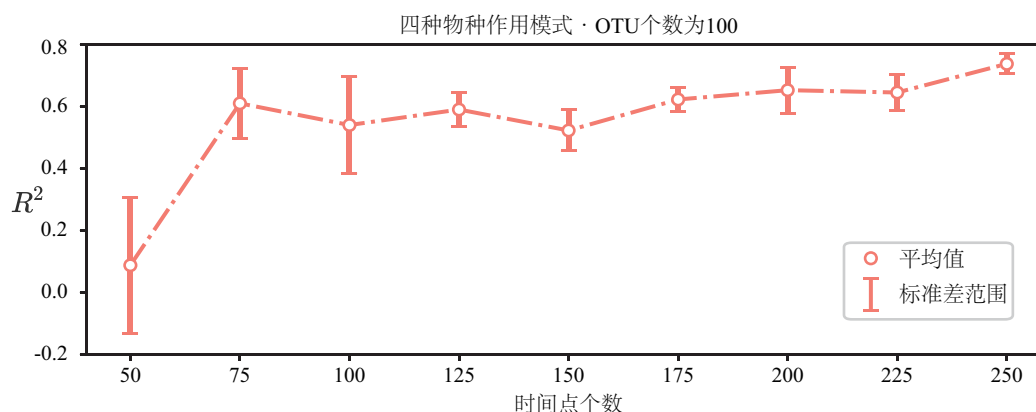


图 8 (网络版彩图) MicroDIG 在不同数据采样点下的性能变化折线柱状图. 当输入丰度时序数据 X 的时间点个数不同 (从 50 个时间点逐步变化到 250 个时间点) 时, MicroDIG 在群落规模为 100 个 OTU、物种作用模式为混合模式的情况下, 估计结果 $\hat{M}_{ij}$ 与标签 M_{ij} 间 R^2 的变化

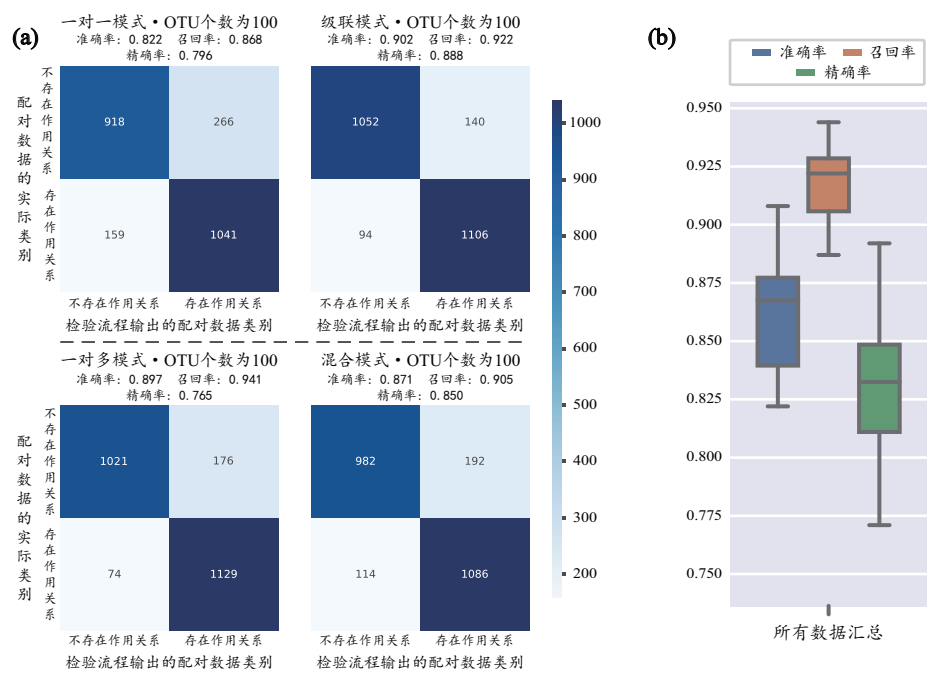


图 9 (网络版彩图) 校验流程有效性验证. (a) 群落规模为 100 个 OTU 下, 四类不同物种作用模式的验证集数据与不具有作用关系的配对数据组成的负样本数据经由有效性校验后得到的混淆矩阵. (b) 28 个模拟微生物群落丰度数据验证集混淆矩阵的总体准确率、召回率和精确率箱线图

准确率、召回率和精确率均分布在 0.85 左右, 说明 MicroDIG 校验流程能够有效地去除假阳性的物种作用关系, 从而识别出真正存在的种间关系.

验证集正负样本数量均衡且与训练集同源, 但在真实的微生物群落中的种间作用关系通常较为稀疏, 因此真阴性种间关系 (即两个物种间不存在作用关系) 的识别也十分重要. 本实验在验证集的性能评估基础上, 使用测试集进一步验证 MicroDIG 在不同群落 OTU 规模的四类不同物种作用模式下的作用关系真阴性的判别情况, 如图 10 所示. 在所有情况下, MicroDIG 对于真阴性的识别率均高于

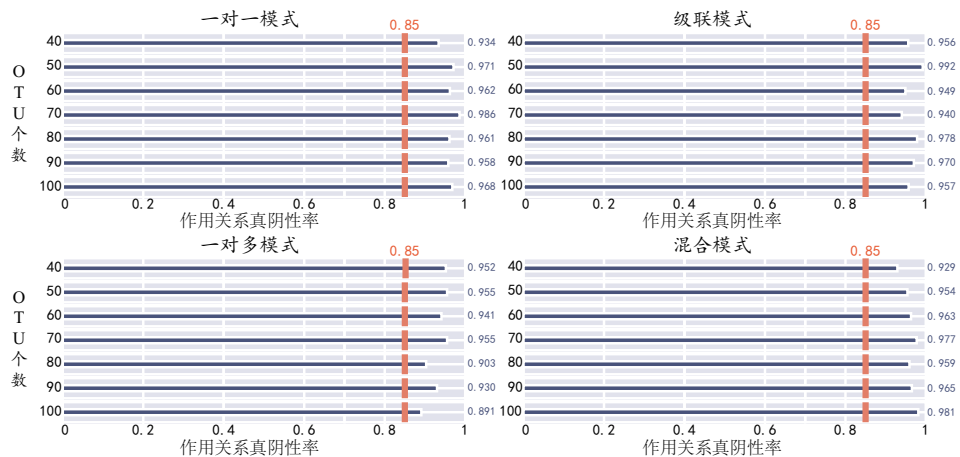


图 10 (网络版彩图) MicroDIG 真阴性率分析效果图. 28 个模拟微生物群落丰度数据集中, MicroDIG 对真阴性种间关系 (即两个物种间不存在作用关系) 的识别效果, 竖状分界线为识别率 85%, 蓝色横向矩形为具体的真阴性的识别率. MicroDIG 对于真阴性的识别效果均超过橙色分界线

85%, 即 MicroDIG 能够有效地剔除有向作用估计网络初步估计结果的假阳性, 并保证估计结果的真阴性.

对于测试集中具有真实物种间作用关系的 OTU, MicroDIG 也能进行有效的作用关系估计, 结果如图 11 所示, 当某一对配对数据中 OTU_j 对 OTU_i 具有较大的作用强度时, 校验模块能较为准确地保留对应的作用关系强度估计结果; 当 OTU_j 对 OTU_i 具有很微弱的作用强度时, 校验模块倾向删除对应的估计结果, 即认为二者之间不存在作用关系.

3.6 实验四: MicroDIG 能准确识别物种关联的方向

在自然界的微生物群落中, 两类不同微生物之间的作用关系通常是有向的, 例如, 针对两类特定的微生物 OTU_j 和 OTU_i , 它们之间的作用关系是由 OTU_j 指向 OTU_i 的, 而 OTU_i 指向 OTU_j 的作用关系不存在, 即从微生物丰度的变化方面来说, OTU_j 的丰度变化导致了 OTU_i 的丰度变化. 为了探究 MicroDIG 在上述情况中对于两类不同微生物 OTU_j 与 OTU_i 之间作用关系 (OTU_j 对 OTU_i 存在作用强度 M_{ij} , OTU_i 对 OTU_j 不具备反向的作用关系, 如图 12(a) 所示) 的估计效果, 本实验批量构造了四类不同物种作用模式下, OTU_j 对 OTU_i 存在作用强度 M_{ij} 而 OTU_i 对 OTU_j 不具备反向的作用关系的微生物配对数据, 并使用 MicroDIG 对其作用关系进行分析.

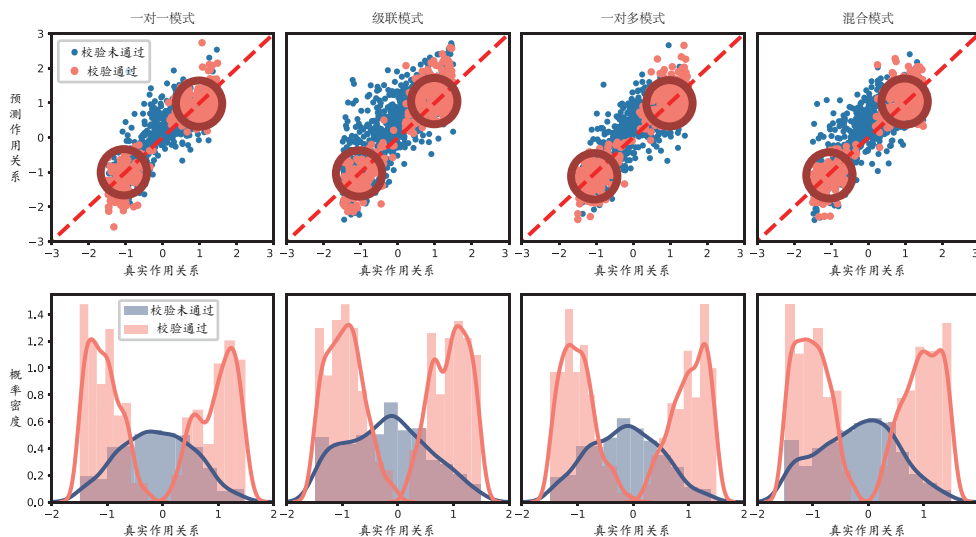


图 11 (网络版彩图) MicroDIG 校验流程效果图. 橙红色表示 MicroDIG 有效性校验通过的作用关系估计值 $\hat{M}_{ij}$, 蓝色表示未通过校验的 $\hat{M}_{ij}$. 散点图: 校验通过作用关系与校验未通过作用关系散点分布情况. 概率密度图: 校验通过作用关系与校验未通过作用关系概率密度的差异情况

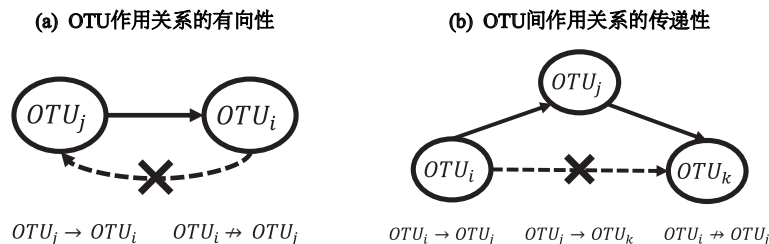


图 12 (网络版彩图) OTU 间的作用关系. (a) OTU 作用关系的有向性; (b) OTU 间作用关系的传递性

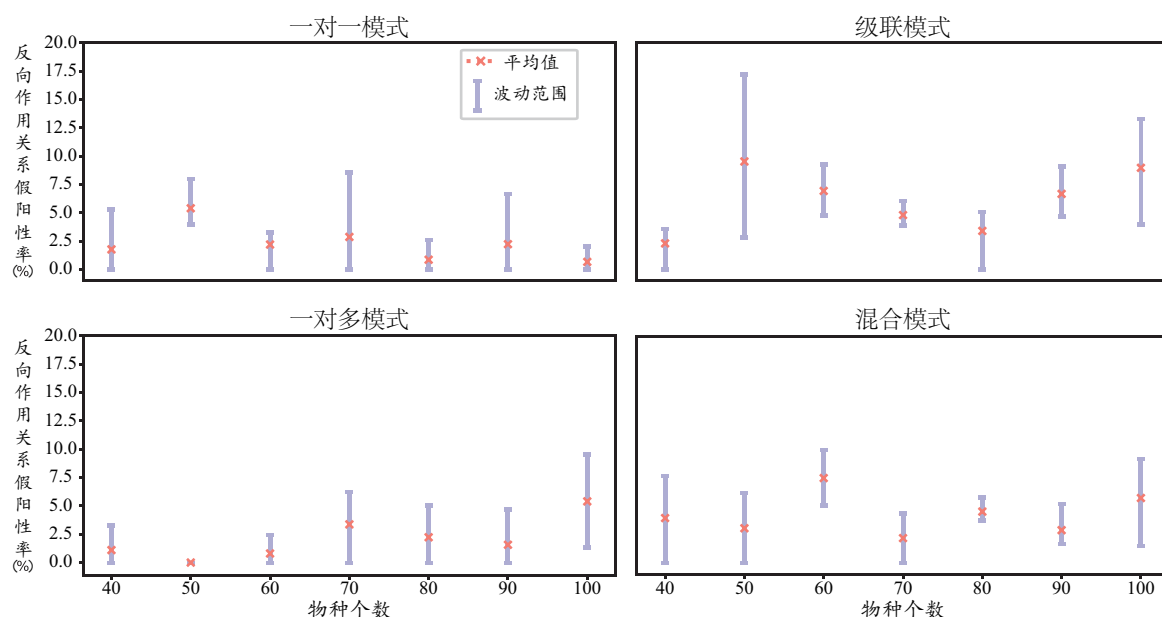


图 13 (网络版彩图) MicroDIG 反向作用关系校验分析图. 四类不同物种作用模式所涵盖的 28 个模拟微生物群落丰度数据集中, MicroDIG 针对图 12(a) 所示的 OTU_i 对 OTU_j 的反向作用关系的识别假阳性率的波动范围与平均值情况

实验结果如图 13 所示, 四类作用模式下, MicroDIG 针对 OTU_i 对 OTU_j 是否具备反向作用关系分析校验的假阳性率仅在 2%–10% 左右, 不存在的反向作用关系无法逃脱校验模块的检测从而被有效地剔除, 说明 MicroDIG 能准确地识别出两两 OTU 之间作用关系的方向.

3.7 实验五: MicroDIG 能够有效降低传递作用带来的假阳性

微生物群落种间作用关系分析中, 除了作用关系的有向性以外, 还需要考虑由于链路传递导致的假阳性. 具体而言, 若已知三类不同的微生物 OTU_i , OTU_j 和 OTU_k , 当 OTU_i 对 OTU_j 有作用关系, OTU_j 对 OTU_k 有作用关系, 而 OTU_i 对 OTU_k 没有作用关系 (如图 12(b) 所示) 时, 仅从丰度变化上观察, OTU_i 的丰度变化同样引起了 OTU_k 的丰度变化, 但本质上, 两者之间并不存在对应的作用关系. 因此, 由于这种作用关系的传递性, 大量相关性分析方法会估计出 OTU_i 对 OTU_k 具有作用关系的假阳性结果, 即 OTU_i 对 OTU_k 存在虚假的作用关系. 为了评估 MicroDIG 对于此类虚假作用关系的分析效果, 本实验在级联模式的物种作用关系下, 构造了大量 OTU_i 对 OTU_j 有作用关系, OTU_j 对 OTU_k 有作用关系, 而 OTU_i 对 OTU_k 没有作用关系的微生物数据, 并使用 MicroDIG 进行了分析. 实验结果如图 14(b) 所示, 在七类不同群落规模的汇总数据中, 仅有 85% 左右的虚假作用关系都能被 MicroDIG 检测辨识, 同时图 14(a) 中不同群落规模的级联模式模拟微生物群落丰度数据集所对应的扇形区域中, 链路导致的假阳性作用关系大部分能被 MicroDIG 的校验流程有效拒绝, 说明 MicroDIG 能够有效降低传递作用带来的假阳性.

3.8 实验六: MicroDIG 估计值能够预测 OTU 丰度在扰动下的动态变化

本文从 28 个模拟微生物群落丰度数据集的验证集数据中随机选取了四组包含有向作用关系的配对数据 h_i ($i = 1, 2, 3, 4$), 分别为 $OTU_j \rightarrow OTU_i$, $OTU_k \rightarrow OTU_l$, $OTU_q \rightarrow OTU_p$, $OTU_r \rightarrow OTU_t$.

在 OTU_j , OTU_k , OTU_q 和 OTU_r 的丰度基础上, 通过 MicroDIG 分析估计得到对应的作用关系参数 $\zeta_1 = [\hat{M}_{ij}, \hat{u}_i, \hat{M}_{ii}], \dots, \zeta_4 = [\hat{M}_{tr}, \hat{u}_t, \hat{M}_{tt}]$, 并结合 GLV 微分方程模拟得到四组作用关系估计值所预测的 OTU_i , OTU_l , OTU_p 和 OTU_t 的丰度变化, 如图 15 所示, 预测的丰度时序点均在真实丰度时序曲线附近波动, 说明 MicroDIG 估计的有向作用强度能够准确模拟出被影响物种的丰度动态特性.

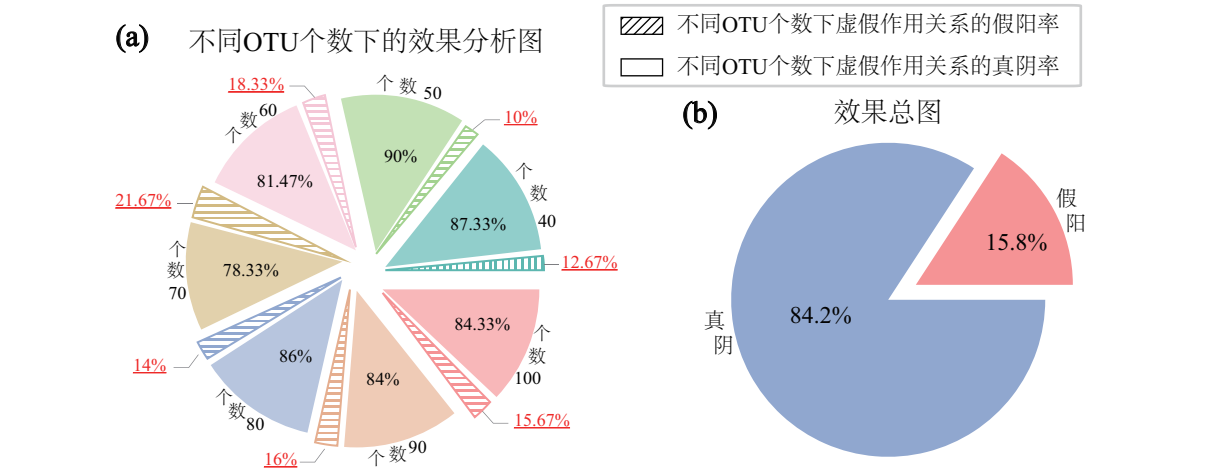


图 14 (网络版彩图) MicroDIG 虚假作用关系识别分析图. 级联模式模拟微生物群落丰度数据集下, MicroDIG 针对图 12(b) 所示的 OTU_i 对 OTU_k 的虚假作用关系的识别假阳性率与真阴性率分析饼图. (a) 不同群落规模独立分析, 每一个扇形区域代表一种不同的群落规模; (b) 所有数据的汇总分析

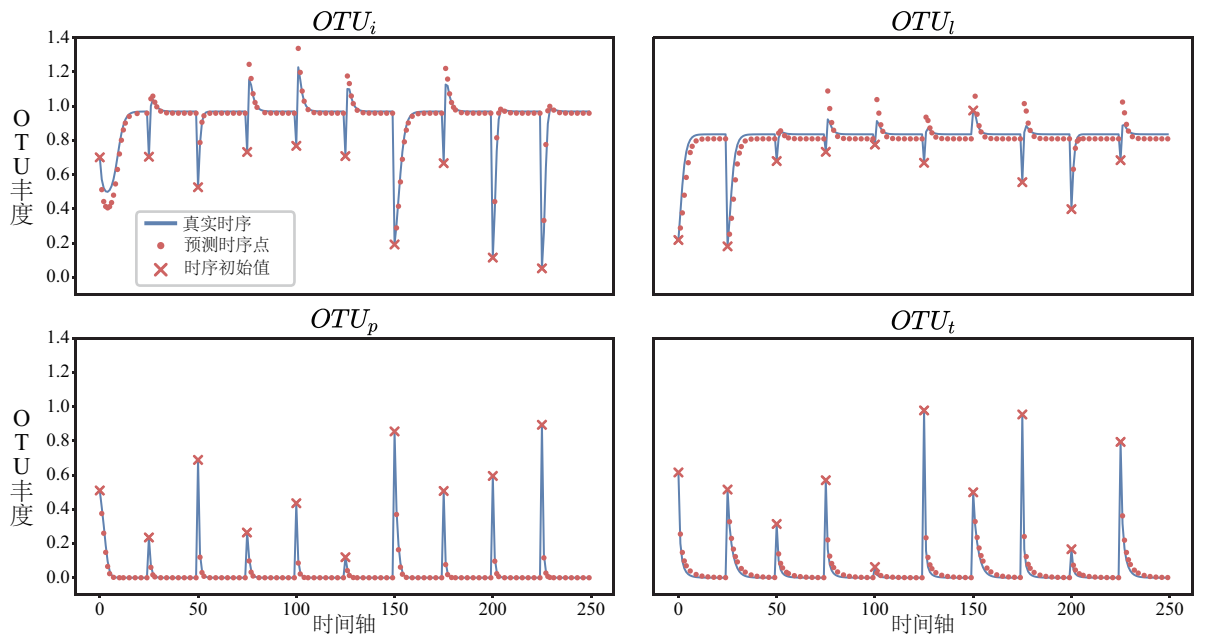


图 15 (网络版彩图) OTU 时序波动预测图. 蓝色实线为四组包含有向作用关系的配对数据 $OTU_j \rightarrow OTU_i, OTU_k \rightarrow OTU_l, OTU_q \rightarrow OTU_p, OTU_r \rightarrow OTU_t$ 中 OTU_i, OTU_l, OTU_p 和 OTU_t 的真实丰度变化; 红色散点为 OTU_i, OTU_l, OTU_p 和 OTU_t 根据作用关系估计值预测的丰度变化

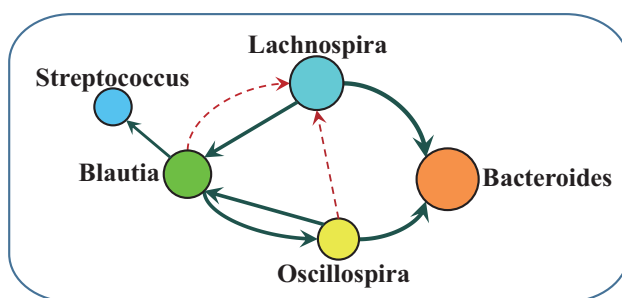


图 16 (网络版彩图) 人类肠道内五大优势物种的作用关系图. 红色虚线表示积极作用关系 (positive), 绿色实线表示消极作用关系 (negative); 线段的粗细代表作用关系的强弱; 箭头指明作用关系的方向; 圆圈的面积代表物种在群落中的丰度占比

3.9 实验七: MicroDIG 识别人体肠道微生物组中优势物种的作用关系

本文进一步针对健康女性肠道微生物组 131 个时间点的丰度数据探究在真实微生物组中优势物种间的有向作用关系估计. MicroDIG 得到人类肠道中五个优势物种 *Streptococcus*, *Blautia*, *Lachnospira*, *Oscillospira* 和 *Bacteroides* 的有向作用网络, 如图 16 所示. *Lachnospira* 对 *Blautia* 和 *Bacteroides* 具有消极影响, 同时受到来自 *Blautia* 和 *Oscillospira* 的积极影响; *Blautia* 和 *Oscillospira* 互相抑制对方的生长等. 该结果与现有的研究结果^[20]中对该数据的作用关系分析结果高度一致, 进一步验证了 MicroDIG 在实际微生物群落中的有向作用关系估计分析能力.

4 结论与讨论

本文开发了一种自监督学习框架 MicroDIG, 融合了数学机理和智能学习估计并校验大规模微生物群落内物种间的有向作用关系. MicroDIG 设计自监督数据增强策略构建大量自带有向作用关系标签的训练数据, 通过自编码器准确拟合微生物群落物种间的有向作用关系, 并提供了一系列严格的校验流程来确保分析结果的可靠性. 本文在多角度多层次的仿真数据上, 进行了大量的实验分析, 首先是通过作用关系估计的准确性评估实验, 验证了 MicroDIG 估计物种间作用关系的能力优于五类基准方法, 并且其性能不受群落规模的限制, 适用于大规模群落的作用关系推断; 其次是基于校验流程的有效性评估实验, 论证了 MicroDIG 对真阴性种间关系以及高置信物种作用关系的准确识别; 最后是涉及两类特殊情况的物种间作用关系分析实验, 证明了 MicroDIG 具备识别两两物种之间作用关系的有向性以及去除作用关系传递导致的假阳性的能力. 在模拟数据实验基础之上, 本文进一步将 MicroDIG 应用于真实数据, 分析了健康女性肠道微生物组的纵向丰度数据, 结果表明, MicroDIG 在实际应用场景中同样能够高效且准确地揭示微生物群落中的作用关系, 这一实验证实了 MicroDIG 的实用性和准确性.

MicroDIG 框架虽然展现了出色的性能, 但目前其设计思路仍局限于两个物种间有向作用关系的建模分析, 无法对涉及更复杂的多物种共同作用的种间关系进行分析. 而真实环境中的微生物群落通常更加复杂化和多元化, 不同于简化的模拟数据, 来源于真实微生物群落的时序丰度数据时间点稀少且不均匀, 其丰度变化可能受到多种外部因素的扰动影响, 如测序仪器的偏差、宿主的健康状态、天气温度的变化等. 标准的 GLV 方程在进行微生物群落的建模时, 会忽视这些外部因素的影响, 从而导致估计结果的误差, 这对 MicroDIG 框架提出了更深层次的应用挑战. 在后续的研究中, 可以充分

考虑 MicroDIG 的可拓展性与可解释性, 在知识导向的前提下, 为其嵌入更有效、更全面的深度学习网络架构, 提升有向作用关系识别的能力, 例如, 使用图网络结构, 引入群落拓扑结构信息; 或使用 Transformer 或 LSTM 等网络模型, 对时序信息进行更准确的提炼等, 从而实现对多样化复杂作用关系的识别. MicroDIG 内置的标准 GLV 方程也可以引入特定的误差项将不同的外部因素融入微生物群落的建模过程之中, 以提高 MicroDIG 的预测精度. 针对真实微生物群落数据时间点不均匀且稀少的问题, 可以考虑通过时间点差值、时间点拷贝等手段来进行优化, 以提高 MicroDIG 的通用性. 在其他领域的动力学研究中, 也可以考虑采用 MicroDIG 提出的动力学方程自监督模拟 - 神经网络训练预测的范式进行相应的分析, 例如在蛋白质组学中, 可以通过 MicroDIG 框架和 Langevin 动力学方程自监督模拟蛋白质从线性氨基酸序列到其最终三维结构的折叠过程, 结合网络训练来预测蛋白质折叠的路径和蛋白质的稳定状态等; 在代谢组学中, 可以结合 Michaelis-Menten 方程和 MicroDIG 的自监督范式来模拟细胞内的代谢催化反应, 进而研究细胞的代谢调控等. 总而言之, MicroDIG 不仅具有独特的数据模拟与有效性校验功能, 能够适应大规模微生物群落, 同时还能够灵活地替换嵌入不同的功能模块, 可拓展到大规模复杂微生物群落真实场景中实现更加精细的互作关系探索. MicroDIG 的自监督框架还具备结合不同动力学方程以扩展到其他领域的潜力, 能为不同领域的研究提供一个新颖的分析思路.

致谢 感谢来自厦门大学信息与网络中心的方少荣和吴添富在高性能计算方面的帮助.

参考文献

- Berry D, Widder S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Front Microbiol*, 2014, 5: 219
- Bucci V, Tzen B, Li N, et al. MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. *Genome Biol*, 2016, 17: 1–17
- Cao H-T, Gibson T E, Bashan A, et al. Inferring human microbial dynamics from temporal metagenomics data: Pitfalls and lessons. *Bio Essays*, 2017, 39: 1600188
- Caporaso J G, Lauber C L, Costello E K, et al. Moving pictures of the human microbiome. *Genome Biol*, 2011, 12: 1–8
- Dam P, Fonseca L L, Konstantinidis K T, et al. Dynamic models of the complex microbial metapopulation of lake mendota. *NPJ Syst Biol Appl*, 2016, 2: 1–7
- Falkowski P G, Fenchel T, Delong E F. The microbial engines that drive Earth's biogeochemical cycles. *Science*, 2008, 320: 1034–1039
- Fang H, Huang C, Zhao H, et al. CCLasso: Correlation inference for compositional data through Lasso. *Bioinformatics*, 2015, 31: 3172–3180
- Faust K, Raes J. Microbial interactions: From networks to models. *Nat Rev Microbiol*, 2012, 10: 538–550
- Friedman J, Alm E J. Inferring correlation networks from genomic survey data. *PLoS Comput Biol*, 2012, 8: e1002687
- Gao Y, Şimşek Y, Gheysen E, et al. miaSim: An R/Bioconductor package to easily simulate microbial community dynamics. *Methods Ecol Evol*, 2023, 14: 1967–1980
- Gilbert J A, Quinn R A, Debelius J, et al. Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature*, 2016, 535: 94–103
- Hamady M, Knight R. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res*, 2009, 19: 1141–1152
- Hu J, Amor D R, Barbier M, et al. Emergent phases of ecological diversity and dynamics mapped in microcosms. *Science*, 2022, 378: 85–89
- Joseph T A, Shenhav L, Xavier J B, et al. Compositional Lotka-Volterra describes microbial dynamics in the simplex. *PLoS Comput Biol*, 2020, 16: e1007917
- Jovel J, Patterson J, Wang W, et al. Characterization of the gut microbiome using 16S or shotgun metagenomics. *Front Microbiol*, 2016, 7: 459

- 16 Kendall M G. A new measure of rank correlation. *Biometrika*, 1938, 30: 81–93
- 17 Kuntal B K, Gadgil C, Mande S S. Web-gLV: A web based platform for Lotka-Volterra based modeling and simulation of microbial populations. *Front Microbiol*, 2019, 10: 288
- 18 Kurtz Z D, Müller C L, Miraldi E R, et al. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biology*, 2015, 11: e1004226
- 19 Ley R E, Peterson D A, Gordon J I. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell*, 2006, 124: 837–848
- 20 Li C, Chng K R, Kwah J S, et al. An expectation-maximization algorithm enables accurate ecological modeling using longitudinal microbiome sequencing data. *Microbiome*, 2019, 7: 118
- 21 Lidicker W Z Jr. A clarification of interactions in ecological systems. *BioScience*, 1979, 29: 475–477
- 22 Marino S, Baxter N T, Huffnagle G B, et al. Mathematical modeling of primary succession of murine intestinal microbiota. *Proc Natl Acad Sci USA*, 2014, 111: 439–444
- 23 Mounier J, Monnet C, Vallaes T, et al. Microbial interactions within a cheese microbial community. *Appl Environ Microbiol*, 2008, 74: 172–181
- 24 Pearson K VII. Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philos Trans R Soc Lond Ser A Math Phys Eng Sci*, 1896, 187: 253–318
- 25 Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 2010, 464: 59–65
- 26 Schluter J, Hussey G, Valeriano J, et al. The MTIST platform: A microbiome time series inference standardized test. *Research Square*, 2024, 8: 4343683
- 27 Shaw G T W, Pao Y Y, Wang D. MetaMIS: A metagenomic microbial interaction simulator based on microbial community profiles. *BMC Bioinform*, 2016, 17: 488
- 28 Spearman C. The proof and measurement of association between two things. *Am J Psychol*, 1904, 15: 72–101
- 29 Stein R R, Bucci V, Toussaint N C, et al. Ecological modeling from time-series inference: Insight into dynamics and stability of intestinal microbiota. *PLoS Comput Biology*, 2013, 9: e1003388
- 30 Stolyar S, Van Dien S, Hillesland K L, et al. Metabolic modeling of a mutualistic microbial community. *Mol Syst Biol*, 2007, 3: 92
- 31 Sulaiman J E, Thompson J, Cheung P L K, et al. *Phocaeicola vulgatus* shapes the long-term growth dynamics and evolutionary adaptations of *Clostridioides difficile*. *Cell Host Microbe*, 2025, 33: 42–58
- 32 Tilman D. Resource Competition and Community Structure. Princeton: Princeton Univ Press, 1896
- 33 Xia L C, Steele J A, Cram J A, et al. Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. *BMC Syst Biol*, 2011, 5: S15

附录 A Kendall 相关性系数计算

给定两个向量 $A = [a_1 \ a_2 \ \cdots \ a_n]$ 和 $B = [b_1 \ b_2 \ \cdots \ b_n]$, n 为对应的元素总个数. A 与 B 之间的 Kendall 相关性系数计算公式为

$$\begin{cases} \gamma = \frac{2 \times (C - D)}{n(n-1)}, \\ C = \sum_{0 \leq i_1 < i_2 \leq n} I((a_{i_1} - a_{i_2})(b_{i_1} - b_{i_2}) > 0), \\ D = \sum_{0 \leq i_1 < i_2 \leq n} I((a_{i_1} - a_{i_2})(b_{i_1} - b_{i_2}) < 0). \end{cases}$$

其中, $I(\cdot)$ 为指示函数, 当括号内的条件为真时为 1, 否则为 0.

附录 B 自编码器结构设置

自编码器层类型与结构设置如表 B1.

表 B1 自编码器层类型与结构

网络层类型	维数	激活函数	备注
输入层	$\mathbf{X}$ 中的时间点个数 T	—	配对数据输入
编码层 1	64	ReLU	—
隐藏层	32	ReLU	编码的潜在表示
展平层	64	—	将数据展平, 维数扩展为隐藏层的两倍
估计值输出层 1	32	ReLU	—
估计值输出层 2	3	—	估计的作用关系
解码层 1	64	ReLU	—
解码层 2	$\mathbf{X}$ 中的时间点个数 T	—	重构的配对数据输入

附录 C 微生物群落丰度模拟数据集的模式及其参数设置

本文采用第三方微生物群落时序丰度模拟工具 miaSim^[10] 来构造模拟微生物群落绝对丰度数据集, 以验证 MicroDIG 的性能. 本文的模拟微生物群落数据集共涉及四种不同物种作用模式, 不同模式的物种间作用关系使用作用关系矩阵 $\mathbf{M}$ 进行定义, 每一个 $\mathbf{M}$ 会对应一个生长速率向量 $\mathbf{u}$, 作为后续性能验证的标准, 也是 miaSim 构造时序数据的参数基础.

(1) 参考 miaSim^[10] 模拟时的默认初值范围, 从标准正态分布 $N(0, 1^2)$ 中采样获得不同的物种初值.

(2) 不同物种作用模式下的具体作用关系矩阵 $\mathbf{M}$ 形式如下, 以四个 OTU 组成的群落为例 (当群落 OTU 个数大于 4 时, 以此类推).

一对一模式 除对角线以外, 一个物种只会对另一个物种具有作用关系, 其余位置为 0,

$$\mathbf{M} = \begin{bmatrix} M_{11} & 0 & 0 & 0 \\ M_{21} & M_{22} & 0 & 0 \\ 0 & 0 & M_{33} & 0 \\ 0 & 0 & M_{43} & M_{44} \end{bmatrix}.$$

级联模式 除对角线以外, 作用关系仅存在于对角线下侧, 其余位置为 0, 从而模拟物种 1 作用于物种 2 进而作用于物种 3 的级联作用情况,

$$\mathbf{M} = \begin{bmatrix} M_{11} & 0 & 0 & 0 \\ M_{21} & M_{22} & 0 & 0 \\ 0 & M_{32} & M_{33} & 0 \\ 0 & 0 & M_{43} & M_{44} \end{bmatrix}.$$

一对多模式 除对角线以外, 作用关系仅存在于同一列, 其余位置为 0, 从而模拟一个物种对多个物种产生作用的情况,

$$\mathbf{M} = \begin{bmatrix} M_{11} & 0 & 0 & 0 \\ M_{21} & M_{22} & 0 & 0 \\ M_{31} & 0 & M_{33} & 0 \\ M_{41} & 0 & 0 & M_{44} \end{bmatrix}.$$

混合模式 通过随机选取上述三种不同的作用关系矩阵单元结构构造最终的 $\mathbf{M}$, 使得 $\mathbf{M}$ 中涵盖三种不同的物种作用关系. 在此作用模式下, 通过对角线拼接一对一模式、级联模式及一对多模式的作用关系矩阵, 获得该模式下的最终作用关系矩阵 $\mathbf{M}$, 从而使其包含更为复杂的物种间作用关系.

(3) 作用关系矩阵 $\mathbf{M}$ 中的非零元素服从均匀分布 $U(a, b)$; 生长速率向量 $\mathbf{u}$ 中各元素服从正态分布 $N(\mu, \sigma^2)$; a, b, μ 和 σ 的取值参照研究 MTIST^[26] 所提供的群落规模为 3 个 OTU、10 个 OTU 和 100 个 OTU 的经过验证的可以生成稳定的虚拟微生物群落的作用关系矩阵和生长速率向量进行设置.

(4) 针对每一组对应的作用关系矩阵 $\mathbf{M}$ 、生长速率向量 $\mathbf{u}$ 和 OTU 初值, 通过 miaSim 来构造对应的模拟微生物群落丰度数据.

附录 D 自监督数据生成算法

自监督数据增强模块的具体操作细节如下列伪代码 (算法 D1) 所示.

算法 D1 MicroDIG 自监督数据增强算法

输入: 微生物 OTU 时序丰度数据 $\mathbf{X}$, 配对数据的生成个数 S .

输出: 增强配对数据 $\mathcal{H} = [h_1, h_2, \dots, h_S] \in \mathbb{R}^{+S}$ 、对应的 GLV 参数 $\xi = [\zeta_1, \zeta_2, \dots, \zeta_S] \in \mathbb{R}^S$.

```

1: function Augmentation( $\mathbf{X}, S$ )
2:    $\mathcal{H} \leftarrow \emptyset, \xi \leftarrow \emptyset$ 
3:   while  $S > 0$  do
4:      $\mathbf{X}_j \leftarrow \text{RandomChoseOne}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N)$ 
5:      $\tilde{u}_i \sim N(\mu_1, \sigma_1^2), \tilde{u}_j \sim N(\mu_1, \sigma_1^2), \tilde{M}_{ii} \sim U(a_2, b_2), \tilde{M}_{jj} \sim U(a_2, b_2), \tilde{M}_{ij} \sim U(a_1, b_1)$ 
6:      $\text{Set} \leftarrow \emptyset$ 
7:     for  $\mathbf{X}_j^{(k)}$  in  $\mathbf{X}_j$  do
8:        $\tilde{x}_i(t_1) = \text{mean}(\mathbf{X}_j^{(k)}), \tilde{x}_j(t_1) = x_j(t_1)$ 
9:       while True do
10:         $\frac{d}{dt_\tau} \tilde{x}_i(t_\tau) = \tilde{u}_i \tilde{x}_i(t_\tau) + \tilde{M}_{ij} \tilde{x}_i(t_\tau) \tilde{x}_j(t_\tau) + \tilde{M}_{ii} \tilde{x}_i(t_\tau) \tilde{x}_i(t_\tau)$ 
11:         $\frac{d}{dt_\tau} \tilde{x}_j(t_\tau) = \tilde{u}_j \tilde{x}_j(t_\tau) + \tilde{M}_{jj} \tilde{x}_j(t_\tau) \tilde{x}_j(t_\tau)$ 
12:        if  $t_\tau$  in  $[t_1 \ t_2 \ \dots \ t_T]$  do
13:           $\tilde{x}_j(t_\tau) = x_j(t_\tau)$ 
14:          if  $t_\tau = t_T$  do
15:            break
16:        end while
17:         $\text{Set} \leftarrow \text{append}(\tilde{\mathbf{X}}_i^{(k)})$ 
18:     end for
19:      $\tilde{\mathbf{X}}_i \leftarrow \text{concat}([\tilde{\mathbf{X}}_i^{(1)}, \tilde{\mathbf{X}}_i^{(2)}, \dots, \tilde{\mathbf{X}}_i^{(K)}])$  while  $\tilde{\mathbf{X}}_i^{(k)}$  ( $k = 1, \dots, K$ ) in  $\text{Set}$ 
20:      $\mathcal{H} \leftarrow \text{append}(h = [\mathbf{X}_j, \tilde{\mathbf{X}}_i])$ 
21:      $\xi \leftarrow \text{append}(\zeta = [\tilde{M}_{ij}, \tilde{u}_i, \tilde{M}_{ii}])$ 
22:      $S = S - 1$ 
23:   end while
24:   return  $\mathcal{H}, \xi$ 
25: end function

```

表 E1 MicroDIG 在不同规模数据集下的运行时间

数据集规模 (OTU 数量)	自监督增强数据生成 (单位: 分钟)	自编码器训练 (单位: 分钟)
40	0.043	4.169
50	0.060	5.346
60	0.075	6.283
70	0.090	7.483
80	0.121	8.152
90	0.164	8.930
100	0.215	10.052

附录 E 计算资源与计算复杂度

MicroDIG 目前运行于 1 块 12 GB 的 NVIDIA GeForce GTX 1080Ti GPU 以及 32 核的 Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10 GHz CPU 的硬件设备上, 其对 GPU 硬件资源无特殊要求. MicroDIG 在不同规模数据集下的运行时间如表 E1 所示, 其算法复杂度为 $O(n)$.

MicroDIG: Estimating directed interactions in large-scale microbiome longitudinal data based on self-supervised learning

Yaoyang Yu, Wenhao Zhang, Fengzhu Sun & Ying Wang

Abstract Species within microbial communities influence and depend on each other through complex interactions, thereby maintaining a dynamic balance. Investigating these interactions is crucial for understanding the functions and underlying mechanisms of microbial communities. In this paper, we introduce a self-supervised learning framework, the MicroDIG (microbiome directed interaction on GLV), which utilizes longitudinal abundance data to explore species interactions within large-scale microbial communities. The MicroDIG combines mathematical modeling with self-supervised learning to estimate and validate directed interactions between species. It leverages the GLV (generalized Lotka-Volterra) equations for data simulation and augmentation, and employs an autoencoder to infer the strengths of directed interactions between species. Additionally, it implements a statistical testing strategy to validate these interaction strengths, identifying high-confidence directed interactions. Through comprehensive evaluation on both simulated data and real human gut microbiome data, the MicroDIG has been shown to outperform five baseline methods. Specifically, the MicroDIG effectively reduces false positives in its estimations, accurately identifies the directionality of species interactions, and rejects interactions that arise from transient relationships. The MicroDIG is applicable to communities of various sizes, particularly large-scale ones that traditional methods struggle to handle. The self-supervised learning framework proposed in the MicroDIG does not require training data with ground truth, offers clear interpretability, and is highly flexible and scalable. It can be extended to explore more intricate interactions in real-world, large-scale, and complex microbial communities.

Keywords directed interactions between microbes, self-supervised learning, autoencoder, microbiome, longitudinal data

MSC(2020) 62F30, 62P10, 92B10

doi: 10.1360/SSM-2024-0293