微生物学通报
Microbiology China
http://journals.im.ac.cn/wswxtbcn

Sep. 20, 2024, 51(9): 3673−3689
DOI: 10.13344/j.microbiol.china.231078

前瞻与评论

机器学习在微生物生态领域的应用文献计量分析

林珊珊，李哲，杨柳，鲁伦慧*

中国科学院重庆绿色智能技术研究院，重庆 400714

林珊珊, 李哲, 杨柳, 鲁伦慧. 机器学习在微生物生态领域的应用文献计量分析[J]. 微生物学通报, 2024, 51(9): 3673-3689.
LIN Shanshan, LI Zhe, YANG Liu, LU Lunhui. Bibliometric analysis of machine learning in microbial ecology[J]. Microbiology China, 2024, 51(9): 3673-3689.

摘　要：【背景】随着新型台站和测序方法的迅猛发展，微生物生态学数据已经呈现出爆炸式的增长，为理解微生物在全球生态系统中的功能提供了坚实的基础。然而，这些庞大的数据集的处理和分析对传统方法来说是一个挑战，机器学习因其在处理大数据方面的优势而成为解决这一问题的关键技术。【目的】基于文献计量学，本文全面探索机器学习在微生物生态学研究中的应用，包括其发展趋势、现状及热点，为未来机器学习和微生物生态研究的结合指明方向。【方法】获取并分析Web of Science (WOS)核心合集数据库中1991–2023年间发表的相关文献，运用可视化软件CiteSpace探究发文量演变特征、国际合作情况及学科交叉现状，利用 Carrot2 对文本数据进行挖掘，构建可视化知识图谱。【结果】机器学习在微生物生态学研究中应用的数量以2018年为界经历了稳定增长和暴发增长两个时期，相关应用正逐渐成为各国研究的热点，研究成果持续增长，相关学科之间的交叉融合越来越紧密，特别是微生物生态学与化学、物理、环境、计算机等学科之间的合作，为科学研究的进展提供了新的视角。机器学习在微生物生态学中的应用广泛。在早期，研究主要集中在序列识别和物种分类上。自2018年以来，随着深度学习和计算机视觉技术的发展，研究焦点转向复杂系统的预测。这两个时期关键词的对比展示了机器学习技术在微生物生态学中的应用从基础的数据处理和分析逐渐转向更加复杂、高级的预测模型。【结论】基于文献计量分析结果，结合机器学习在微生物生态中应用的数据缺乏、模型选择困难和可解释性差等挑战，后续研究应更加重视国际合作和数据共享，加强学科交流，推动可解释机器学习的发展。

关键词：机器学习；文献计量；CiteSpace；微生物生态；预测

资助项目：国家自然科学基金(U2340222)；水利部重大科技项目(SKS-2022081)；中国长江三峡集团有限公司科研项目(202403005)；中国科学院西部之光“西部青年学者”计划
This work was supported by the National Natural Science Foundation of China (U2340222), the Key Project of the Ministry of Water Resources (SKS-2022081), the China Three Gorges Corporation Research Projects (202403005) and the “Light of West” Program from the Chinese Academy of Sciences.
*Corresponding author. E-mail: lulunhui@cigit.ac.cn
Received: 2023-12-26; Accepted: 2024-02-04; Published online: 2024-03-21

Bibliometric analysis of machine learning in microbial ecology

LIN Shanshan, LI Zhe, YANG Liu, LU Lunhui*

Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, China

Abstract: [Background] With the rapid development of new types of stations and sequencing methods, the data in microbial ecology has experienced explosive growth, providing rich resources for revealing the role of microorganisms in global ecosystems. However, the processing and analysis of these large datasets pose challenges to conventional methods. Machine learning, with unique advantages in handling big data, has become a key technology to address these challenges. **[Objective]** This study comprehensively explored the developmental trends, current status, and hotspots in the application of machine learning in microbial ecology through bibliometric analysis, aiming to guide the future integration of machine learning with the research of microbial ecology. **[Methods]** The relevant articles published between 1991 and 2023 in the Web of Science (WOS) Core Collection were collected and analyzed. CiteSpace was used to visualize the evolution of the number of publications, international collaboration, and interdisciplinary status. Carrot2 was employed to mine textual data and build knowledge maps. **[Results]** The application of machine learning in the research of microbial ecology has undergone two distinct phases: stable growth followed by explosive growth, with 2018 marking a turning point. This research field has gained increasing attention, which led to continuous growth in the research output. The integration of machine learning with other disciplines, especially chemistry, physics, environmental sciences, and computer science, has become increasingly tight-knit, providing new perspectives for the advancement of scientific research. The application of machine learning in microbial ecology is extensive. The early studies primarily focused on sequence recognition and species identification. However, as deep learning and computer vision technologies have kept advancing since 2018, the research focus shifted towards predicting complex systems. The comparison of keywords between the two phases highlighted the evolution of machine learning in microbial ecology from basic data processing and analysis to complex and advanced prediction models. **[Conclusion]** According to the results of bibliometric analysis and considering challenges like data scarcity, difficulties in model selection, and poor interpretability in applying machine learning in microbial ecology, we suggest that international collaboration, data sharing, and interdisciplinary exchange should be emphasized in the future to promote the development of interpretable machine learning.

Keywords: machine learning; bibliometrics; CiteSpace; microbial ecology; prediction

微生物在古老的地球生命系统进化过程中发挥关键作用,他们具有强大的代谢能力和丰富的多样性,几乎是所有生物地球化学循环过程的驱动者[1-2]。微生物生态学是一门研究微生物彼

此之间或与其生活环境之间相互作用的学科。它聚焦于研究微生物在生态系统中扮演的角色和相互作用关系。这包括对宿主健康、生物地球化学循环以及微生物与人类活动导致的环境污染或气候变化等方面的相互作用的研究[3]。微生物生态学在帮助我们理解自然生态进化[4]、应用工业微生物[5]和维持自然生态系统的健康[6]方面起着重要作用。

微生物生态学正经历一场变革，其影响遍及微生物学、生态学和生态系统科学[7]。新型台站，如美国国家生态观测站网络(national ecological observatory network, NEON)[8]和全球湖泊生态观测站网络(global lake ecological observatory network, GLEON)[9]的建立为生态学提供了研究平台，它们使用数千个传感器收集不同的环境参数。另外，由于二代测序的发展，使得大量分子及多组学数据快速积累，为深入理解微生物的巨大多样性、丰富的未培养微生物群和新的微生物功能提供了有力支撑。在“大科学”联网数据驱动下[10-11]，生态学日益成为一门数据密集型学科[12]，这些大数据为我们提供了前所未有的研究机遇。然而，这些数据的海量和复杂特性，使得传统统计方法显得力不从心。计算效率低、无法处理高维数据，以及对动态数据分析的不足，都严重制约了我们对这些数据的深入理解[7]。当前的微生物生态研究主要集中在用生物指标对微生物群落及其功能进行描述，大量分子生态学数据的积累亟须运用新的理论方法来预测或指示更多的环境系统功能。在这样的背景下，机器学习技术应运而生，为我们提供了新的解决方案。

机器学习(machine learning)最早由 Arthur Samuel 于 1959 年提出，是人工智能的一个特殊分支(子领域)，旨在从大规模异构数据中寻找特征。机器学习需要获取大量数据，并使用算法对其进行训练，这一过程中最基本的是使用算法解析数据，自动分析数据中的模式，然后利用这些模式对真实世界的事件进行预测和决策[13]。机器学习在微生物生态学领域的应用是人工智能和生态学两个领域的交叉研究，生态学和人工智能都追求对复杂系统的预测性理解，这些系统通常具有非线性、多尺度和多维相互作用的特点[12]。为了更好地理解微生物群落的组成和功能，研究者们采用了多种机器学习方法，包括非监督学习中的分层聚类、监督学习中的随机森林、深度神经网络等，这些方法在微生物分类学[14]、结构与功能分析[15-18]、群落相互作用研究以及生态系统建模[19-20]等方面取得了显著进展。然而，由于生态系统具有弹性、突变性、非线性动力学、系统现象等不可预测的多维复杂特征[15,19]，对生态系统全面建模仍具有挑战性，微生物生态学与人工智能的结合急需新的更强大的人工智能架构。

基于此，本文利用文献计量学方法对使用机器学习技术进行微生物生态学研究的文章进行检索与概述，探寻机器学习在微生物生态研究领域中应用的热点方向及中心变化，展示机器学习在微生物生态学中成功应用的实例，为其潜力提供证据，提出目前在微生物生态学中使用机器学习的挑战和未来方向。

1 材料与方法

1.1 数据来源

机器学习在微生物生态领域应用广泛，涉及数据预处理、模式识别、预测等研究。为了更全面地了解该领域的发展情况，我们使用 Web of Science (WOS)核心合集作为数据来源。为了确保检索数据的全面性和准确性，选择 Science Citation Index Expanded (SCIE)作为索引，首先，我们希望确保当前广泛使用的机器学习方法都

包含在主题搜索结果中。因此，在标题、摘要和/或关键词列表中提及深度学习、神经网络、随机森林和支持向量机等方法的出版物都将被包括在内。其次，研究的领域是微生物生态，而微生物生态的范围非常广泛，为了提高检索的精确度，将其分为研究对象和研究范围两个层次。研究对象是微生物，研究范围根据生态学的研究尺度确定，包括种群、群落、生态系统、景观和全球。但仅使用上述关键词做限定，范围较模糊，机器学习有可能应用于与微生物生态学无关的上下文，因此我们将应用机器学习进行微生物生态研究的主要内容也作为主题词纳入检索范围，例如预测、分类和相互关系等。最终，我们确定的检索策略为 TS=[(“machine learning” OR “deep learning” OR “artificial intelligence” OR “neural networks” OR “supervised learning” OR “unsupervised learning” OR “reinforcement learning” OR “semi-supervised learning” OR “random forest” OR “SVM”) AND (“microbi*” OR “Microorganism*” OR “bacteria” OR “fungi” OR “virus”) AND (“communit*” OR “population” OR “ecology” OR “ecological” OR “ecosystem” OR “landscape” OR “global” OR “forecast*” OR “predict*” OR “classif*” OR “interaction”)]。为了获得尽可能全面的结果，未限制检索开始的时间，检索截止时间为 2023 年 8 月 5 日。选择的文章类型是论文、综述论文和在线发表论文，最终获得了 9 025 篇有效论文，构成了本次研究的数据集。

1.2 分析方法

本文利用文献计量学工具 CiteSpace (v5.6.R3, https://citespace.podia.com/)和 Carrot2 (v3.16.0, https://github.com/carrot2/carrot2)绘制知识图谱[21]，用可视化技术量化机器学习在微生物生态研究中应用的前沿。本文利用 CiteSpace 时间切片与国家间合作网络分析功能，对 WOS 检索结果中的文献发表时间和来源国家进行描述性统计分析，研究其发展趋势。利用 CiteSpace 软件双图叠加的方法研究文献集合的施引与被引关系，双图指的是施引期刊集群图和被引期刊集群图[22]，这些期刊集群利用 Blondel 算法形成，描绘了 10 000 种科学期刊之间的相互联系[23]。在期刊集群图上叠加期刊数据分析图，获得期刊引用关系的可视化图谱。通过研究施引文献与被引文献之间的联系，探索不同学科的文献分布、被引轨迹和主题变化等信息[24]，结合突现和聚类功能展现知识流转与学科交叉。借助 Carrot2 聚类分析功能构建可视化知识图谱，探索机器学习在微生物生态领域应用的研究热点和前沿方法，并进行概括性的总结分析。

2 结果与分析

2.1 机器学习在微生物生态领域应用的发展趋势分析

2.1.1 发文量演变特征

通过分析某一领域的年度发文数量，可以在一定程度上揭示该领域的发展趋势和受到的关注程度。图 1 为 1991 年以来微生物生态领域中使用机器学习进行研究的发文量和发文量相较前一年的增幅变化。整体来看，在微生物生态领域中，与机器学习方法有关的文章数量呈逐年上升趋势。根据发文量的变化趋势，可将该领域研究分为 2 个时期：1991–2017 年为稳定增长期，这一时期中应用机器学习方法研究微生物生态的学者较少，发文量较少且整体发文量增长较为平稳；2018–2023 年为暴发增长期，2018 年，对人工智能和机器学习的炒作逐渐冷却，更多关注公平、可解释性或因果关系等具体问题[25]，同时深度学习框架 PyTorch 在处理复杂模型时的优异表现引起巨大的轰动[26]，让研究者们看到了机器学习在微生物生态领域应用的潜力，相关论文激增。

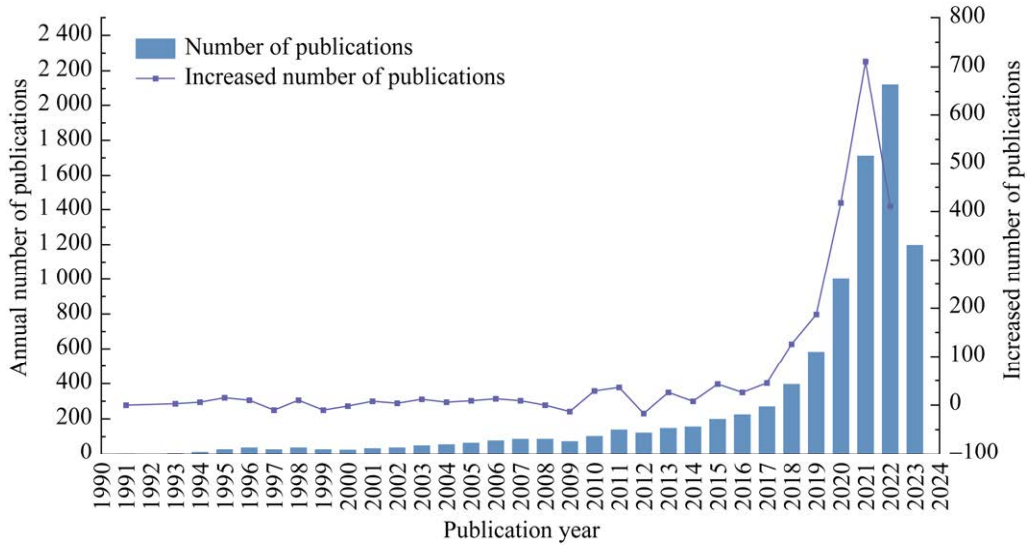

图 1　机器学习在微生物生态学中应用相关的发文数量的变化和趋势

Figure 1　The changes and trends in the number of publications on the application of machine learning in microbial ecology.

2.1.2　国家间合作情况

在 CiteSpace 软件中，节点类型选择“Country”，阈值设定为 Top 30%，使用“minimum spanning tree”和“pruning the merged network”进行裁剪以更好地突出主体之间的关系，其他参数使用默认设置。

在国际合作网络图谱中(图 2)，节点的中介中心性反映了该节点在整个合作网络中的重要程度和影响力[27]。从整体来看，连线较为稀疏，说明在该领域中国际合作较为匮乏。荷兰表现出最积极的国际合作，其中介中心性为 1.03，丹麦和加拿大紧随其后，中介中心性分别为 0.87 和 0.64，这表明荷兰、丹麦、加拿大等国家与其他国家之间有着较密切的合作关系。美国的研究体量最大，同时其中介中心性也有 0.4，说明美国在这一领域的研究有很大的影响力。中国的发文量虽然位居前二，但其中介中心性仅 0.05，说明中国与国外机构的合作较为缺乏，发文量虽多但影响力不高。

通过同心圆的颜色变化及色带的粗细来看时间线变化，中心的颜色越深表示相关研究开始的时间越早。机器学习在微生物生态领域的研究始于美国、英国等西方发达国家，中国、印度、韩国等国家次之，伊朗、土耳其、马来西亚等国家近 10 年也开始进行相关研究，说明机器学习在微生物生态这一领域的应用正逐步受到更多国家的关注。

从地区来看，在微生物生态领域，东南亚、西亚和欧洲地区的国家合作较为密切。究其原因，首先是地域上的优势，使得这些国际合作更容易进行，其次这些国家的研究体量较小，更需要借助国际合作的力量来完成相应的研究。

2.2　机器学习在微生物生态领域应用的学科交叉分析

2.2.1　双图叠加分析

双图叠加分析通过期刊集群之间的引文链接指示了不同学科之间知识的流转(图 3)。数据集中的文章施引主要分为三大区域：兽医/动物/科学(Veterinary/Animal/Science)、分子/生物学/免疫学(Molecular/Biology/Immunology)和医药/医学/临床(Medicine/Medical/Clinical)，被引主要分为环境/毒理学/营养学(Environmental/Toxicology/Nutrition)和分子/生物学/遗传学(Molecular/Biology/Genetics)两大区域。

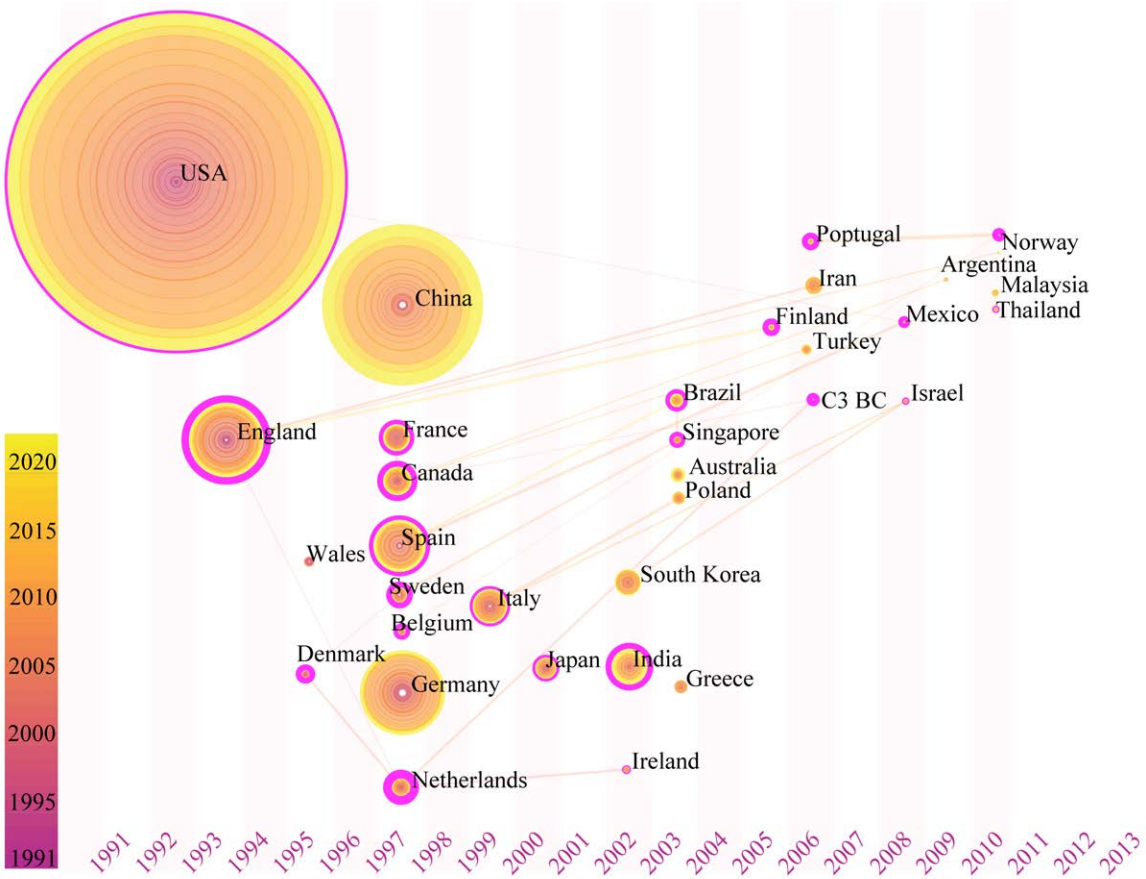

图 2 机器学习在微生物生态领域应用研究的国际合作网络 年轮中心所在的位置代表该国家在领域中首次发表文章的年份，年轮的整体大小反映该国发文量的多少，不同的年轮颜色和连线代表不同的发文时间，一个年轮的厚度与相应时间分区内的发文量成正比，年轮之间的连线代表不同国家的相互合作关系，最外层有粉色标记的代表中心性较强

Figure 2 The international collaborative network in the application research of machine learning in microbial ecology. The position of the center of the tree ring represents the year in which a country first published articles in the field. The overall size of the tree rings reflects the volume of publications from each country. Different colors and connecting lines of the tree rings represent different publication periods. The thickness of a tree ring is proportional to the volume of publications within the corresponding time zone. Connections between tree rings represent collaborative relationships between different countries. The outermost layer, marked in pink, represents countries with higher centrality.

1997–2017 年的数据分析表明，黄色引文链接被分成了 2 个主要的流，表明动物学和兽医学相关的出版物主要引用了两个领域的文章，分别是环境/毒理学/营养学领域和分子/生物学/遗传学领域。分子/生物学/免疫学的引文链接和医药/医学/临床的引文链接也指向了分子/生物学/遗传学领域期刊组。分子/生物学/遗传学领域的文章受到了其他学科大量引用的影响，说明该领域中使用机器学习方法的研究有着许多重要的成果，同时这些成果为动物医学、临床医学等领域的研究做出了贡献。2018 年以后的集体引用行为与之前不同，分子/生物学/免疫学领域的

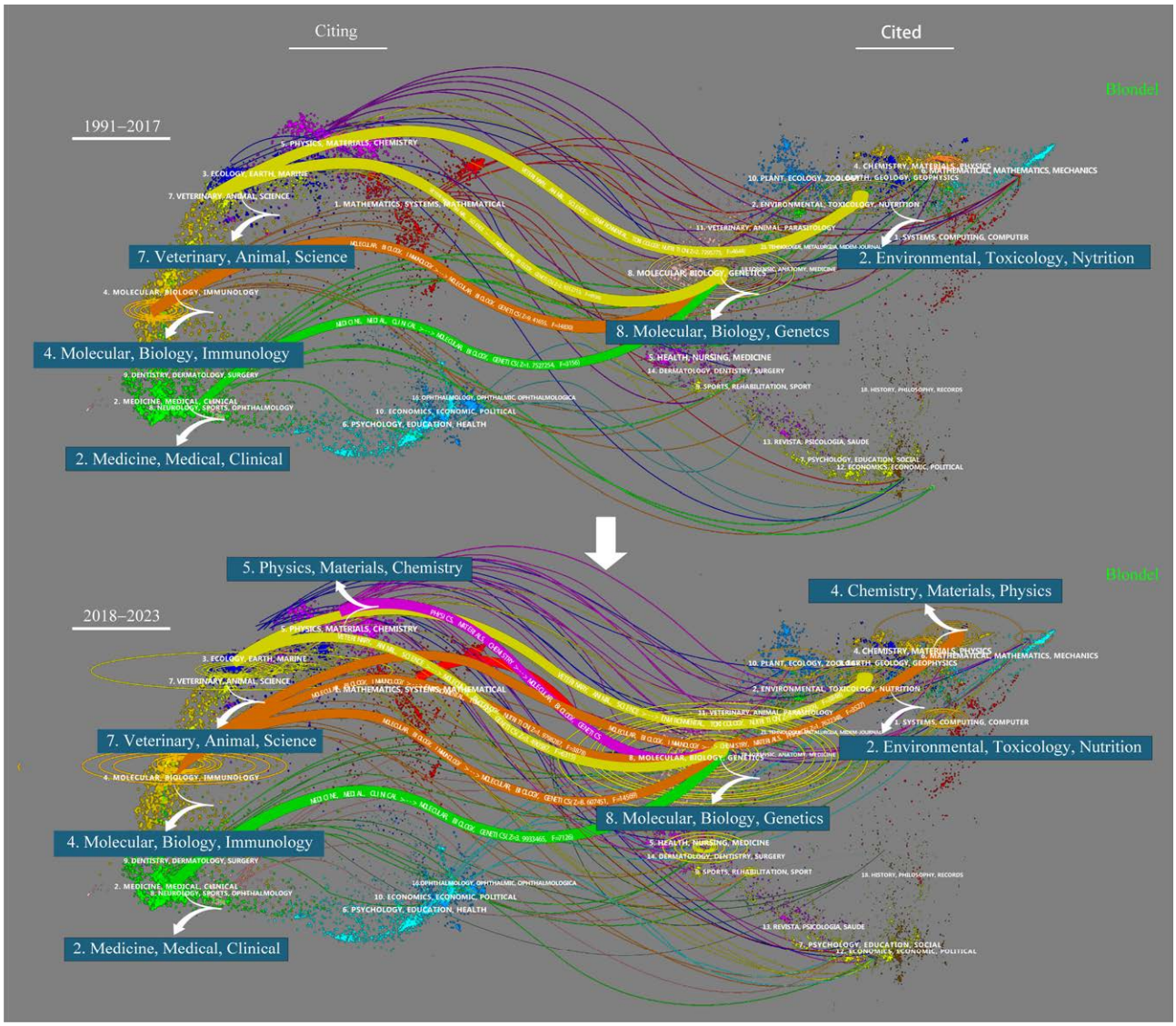

图 3　双图叠加知识流转图　知识流左侧为施引期刊，右侧为被引期刊，彩色曲线表示应用机器学习方法研究微生物生态的相关论文施引和被引路径. *z*-score 代表连线合并的显著性，*z* 值越大表示合并程度越高

Figure 3　A dual-map overlay of the knowledge trajectory. The left side illustrates citing journals, while the right side shows cited journals. Colored curves represent citation paths of papers related to the application of machine learning methods in the study of microbial ecology. The *z*-score indicates the significance of the merging of connecting lines, with a larger *z*-value indicating a higher degree of merging.

文章开始增加化学/材料/物理领域和环境/毒理学/营养学领域文章的引用数量。另外，新的学科领域“物理/材料/化学”，也加入了主要引用和被引的行列中。这一变化得益于新技术的发展。2018 年 12 月，英国皇家学会举办了一场以“单细胞生态学”为主题的跨学科会议[28]，该会议利用物理和分子领域相关的最新方法，在单细胞尺度研究生物现象，揭示了同一物种个体(或个体群)与其他个体、环境，以及不同种个体之间的相互作用。这一会议集结了操纵细胞的物理学家、研究微生物群落性质的微生物学家及开发新单细胞方法的基因组学家，为各个领域带来了许多新的见解与灵感。单细胞技术中的纳米二次离子质谱技术(nanoscale secondary ion mass spectrometry,

NanoSIMS)、单细胞拉曼光谱技术和 $^{15}N_2$ 稳定同位素标记法的结合为表征固氮微生物固氮活性提供了直接手段[29]，挖掘了群落中的高活性固氮微生物，并研究了其空间分布及与其他生物的共生关系，成为了固氮微生物研究的前沿工具。深度学习方法在分析单细胞技术产生的大量高维、稀疏和复杂的数据上具有巨大的潜力[30]。

2.2.2 共被引文献突现性检测

突现性指的是某一时期内某个关键文献在引文合集中出现次数快速增加的现象，而该领域的特定研究热点可以通过引文中具有突现性的文章来进行特征化[24]。通过对数据集(1991–2023 年)中的文章进行共被引突现性分析(表1)可以发现，获得了突现性的文章大多是关于技术方法的研究。在列表顶部，Pedregosa 等的文章“Scikit-learn: machine learning in python”，在 2016–2019 年之间具有一个突现期，其突现强度为 50.74。该文章介绍了机器学习库 scikit-learn，该库封装了多种机器学习模型，用简洁的代码就可以调用，让机器学习更易于被非专业人员使用，并广泛应用于神经影像数据分析[41-42]、蛋白质突变预测[43]和流行病预测[44]等方面。2018 年以来突现强度最高的文章由 Bolyen 等于 2019 年发表，题目为“Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2”，突现强度为 49.40，这篇文章中描述的 QIIME 2 是一个完全重新设计和重写的系统，它提供多种接口，包括命令行接口、Python 接口和图形用户界面，可以处理大规模的微生物组数据，并提供一系列的工具和技术，以便更好地理解微生物组的结构和功能，该平台在 COVID-19 的检测中发挥了重要的作用[45-47]。

2.3 机器学习在微生物生态领域应用的研究热点探索

本文利用 Carrot2 (https://search.carrot2.org)分析研究主题的变化和趋势，使用 Lingo 聚类算法：“cluster count base”设定为 10，“cluster merging threshold”设定为 0.5，其他参数默认设置。分别对 1991–2017 年 2 027 篇文献以及 2018–2023 年 6 998 篇文献的主题进行研究，得到机器学习在微生物生态研究中应用的主题图谱(图 4)。

表 1 爆发强度前 10 的引文

Table 1 Top 10 citations by burst intensity

Title	References	Year	Strength	Begin	End
Scikit-learn: machine learning in python	[31]	2011	50.74	2016	2019
Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2	[32]	2019	49.40	2020	2023
Deep residual learning for image recognition	[33]	2016	47.54	2021	2023
DADA2: high-resolution sample inference from illumina amplicon data	[34]	2016	45.65	2021	2023
Automated detection of COVID-19 cases using deep neural networks with x-ray images	[35]	2020	45.43	2021	2023
COVID-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks	[36]	2020	40.93	2021	2023
Very deep convolutional networks for large-scale image recognition	[37]	2015	40.74	2021	2023
LIBSVM: a library for support vector machines	[38]	2011	40.23	2011	2019
QIIME allows analysis of high-throughput community sequencing data	[39]	2010	32.50	2015	2018
XGBoost: a scalable tree boosting system	[40]	2016	32.20	2021	2023

1991–2017 年的文献更注重序列分析(“Sequencing Studies” 223 篇，“Sequence Analysis” 185 篇)，以及序列分析的应用如蛋白质研究(“Predicting Protein” 397 篇)和物种鉴定(“Species were Used” 272 篇，“Microbial Samples” 148 篇，“Bacteria Species” 114 篇)等分子生物学方面的研究。这一时期高通量测序技术的发展极大地降低了测序成本，使得个人研究成为可能[48]，也促进了各种大型项目如人类基因组计划[49]、地球微生物组计划[50]等的推进。这些研究和大型项目为科学界提供了大量且宝贵的数据资源，但是越来越多的数据对研究者使用的分析方法和拥有的计算资源提出了要求。在这一方面，机器学习技术成为了一种强大的工具，能够处理和分析大规模的序列数据，以应对这一挑战[51]。主要的应用包括利用特定序列识别蛋白质并研究其结构和功能，利用序列识别微生物物种进行物种鉴定与分类。例如，Wang 等[52]使用 SWISS-PROT 及 GenBank 数据库中的蛋白质序列，利用神经网络(neural network)和隐马尔可夫模型(hidden Markov models)鉴定了泡沫病毒(foamy viruses)的糖蛋白信号肽、切割位点、融合肽、跨膜结构域和独特的内质网检索信号。Kalate 等[53]开发了一种使用错误-反向传播(error back propagation)算法训练的多层前人工神经网络 (multilayered feed-forward artificial neural network)架构，用于判断给定的核苷酸序列是否为分枝杆菌启动子序列，准确性约为 97%；并且进一步与卡尺随机化(caliper randomization)方法结合使用，以确定启动子序列中与结构或功能相关的重要区域。Dollhopf 等[54]比较了主成分分析(principal component analysis, PCA)和自组织图神经网络(self-organizing map neural network, SOM)这 2 种机器学习方法在分析 16S rRNA 基因数据，了解不同群落微生物的相似性，破译群落动态变化并描述特定的生态系统类型上的效果，结果表明 2 种分析都可用于 16S rRNA 基因数据，然而，相较于 PCA 排序，SOM 擅长解析更复杂和可变数据中的模式。

2018–2023 年的发文量是前一时期的 3 倍多，然而其聚类数量不增反减，说明近 5 年利用机器学习方法研究微生物生态的热点问题较为集中。应用大量集中在病毒研究上(“Virus Patients” 1 417 篇，“Method for COVID-19” 1 308 篇)。这一时期正值 COVID-19 大流行，如何快速识别病毒感染者、准确预测病毒的扩散与传播等研究备受关注[55]，在短短数月内就有多个重要的 COVID-19 疫情预测模型出现[56-58]。这些模型大多都是利用公开的病例数据进行模型训练。例如，Ardabili 等[56]收集了意大利、德国、伊朗、美国和中国 30 d 内的总病例数，基于微生物生长 Logistic 模型，应用了多层感知器(multi-layered perceptron, MLP)和基于自适应网络的模糊推理系统(adaptive network-based fuzzy inference system, ANFIS)来估计疫情的发展趋势。Alanazi 等[59]收集各国家/地区的 COVID-19 数据集，包括死亡率、康复率和感染率，并将总人口分为易感、感染和康复 3 个子集，利用易感-感染-康复(susceptible-infected-recovered, SIR)派生的常微分(ordinary differential equations, ODE)模型，预测采取不同防治手段下疫情的进展。除了病毒的研究，机器学习方法也大量地应用于土壤生态学的研究(“Forest Soils” 440 篇)，包括土壤微生物群落结构的驱动机制[60-62]、土壤有机碳评估[63-65]及土壤微生物生态系统网络[66-68]。其中，随机森林(random forest, RF)是最常用的机器学习方法，它通过使用输入特征中包含的信息(例如微生物分类群的丰度)来构建由决策树组

成的“森林”，以根据其分配的真实值连续拆分样本。随机森林是一种现成的、计算上易于处理的、性能最佳的分类器，对异常值、固有的噪声和非线性数据(如宏基因组)具有鲁棒性[69]，在微生物生态学中多用于要素筛选。Lammel 等[60]利用 RF 根据环境要素对 OTU 的丰度进行建模，评估彼此高度相关的环境要素变量的相对重要性，从而确定对 OTU 的丰度(响应变量)影响更强的环境要素变量。Chen 等[70]利用 RF 模型评估了非生物和生物特性在驱动受冻土影响的土壤中氨氧化和反硝化微生物功能基因丰度、土壤潜在氨氧化和反硝化速率中的相对重要性，结果表明，氨氧化和反硝化速率分别由古细菌 *amoA* 基因和 *nosZ* 基因等功能基因的丰度决定。

结合突现性检测的结果(表 1)，2018 年以后应用机器学习研究微生物生态的方向变化，其表象下隐藏的是机器学习方法的快速发展及与微生物生态学的多方位结合(“Learning Image” 1 021 篇，“Model Development Approach” 725 篇，“Prediction Models Trained” 722 篇)。深度学习网络中网络的深度对模型的性能至关重要，但是当网络深度增加时，网络准确度会出现饱和，甚至下降，这一现象称为网络退化(degradation)，这一现象的存在导致了深度卷积神经网络(deep convolutional neural network, DCNN)的训练存在困难，进一步导致了计算机图像识别发展的缓慢[71]。2015 年微软研究院的 He 等在 ImageNet 大规模视觉识别挑战赛中凭借深度残差神经网络(deep residual network, ResNet)架构获得冠军[33]，研究人员针对深度学习退化现象发明了“短路连接(shortcut connection)”，极大地消除了深度过大的神经网络训练困难问题，并且改变了图像分类的研究前景。

机器学习技术的发展很快地反应在微生物生态学研究方向的变化上。早期机器学习应用于微生物分类工作脱离不了分子生物学的方法，它通过分析 DNA 或 RNA 序列中的特征来区别不同的微生物。然而随着深度学习、计算机视觉等技术的发展，现在的微生物分类可以通过算法直接分析显微镜图像实现。Kosov 等[72]提出了一种环境微生物分类引擎，结合条件随机场(conditional random field, CRF)和 DCNN 自动分析显微图像，通过预训练的 DCNN 提取显微图像的像素级特征，结合全局特征利用 CRF 模型对微生物进行准确分类和定位，平均准确率达到 91.4%。Tahir 等[73]利用光学传感器获取不同来源的 40 800 张真菌孢子图像，开发了一种基于 ResNet 架构的 CNN 方法以 94.8%的准确率精确检测和分类真菌。2018 年多伦多大学 Chen 等[74]将深度学习与微分方程结合，提出了神经常微分方程(neural ordinary differential equation, NODE)，可以视作 ResNet 架构的连续极限，在近似常数级的内存成本上实现高效的图像分类和生成任务。由于其可控的内存开销、高效且精确的预测及连续等优点，该方法很快被用到微生物生态学领域。哈佛医学院的 Michel-Mata 等[75]利用 NODE 方法建立了组合神经常微分方程(compositional neural ordinary differential equation, cNODE)，可以在不了解控制微生物动力学的各种物理、生化和生态过程的情况下通过微生物的存在与否预测微生物物种在群落中的相对丰度。Wang 等[76]在 cNODE 的基础上建立了从微生物组成预测微生物代谢产物相对浓度常微分方程模型(metabolomic profile predictor using neural ordinary differential equation, mNODE)，该模型可以节省大量测量代谢组学的实验成本，利用具有相对良好效益的 16S rRNA 基因测序的方法预测代谢组学特征。Wang 等[18]基于 cNODE 方法创建了数据驱动的关键物种

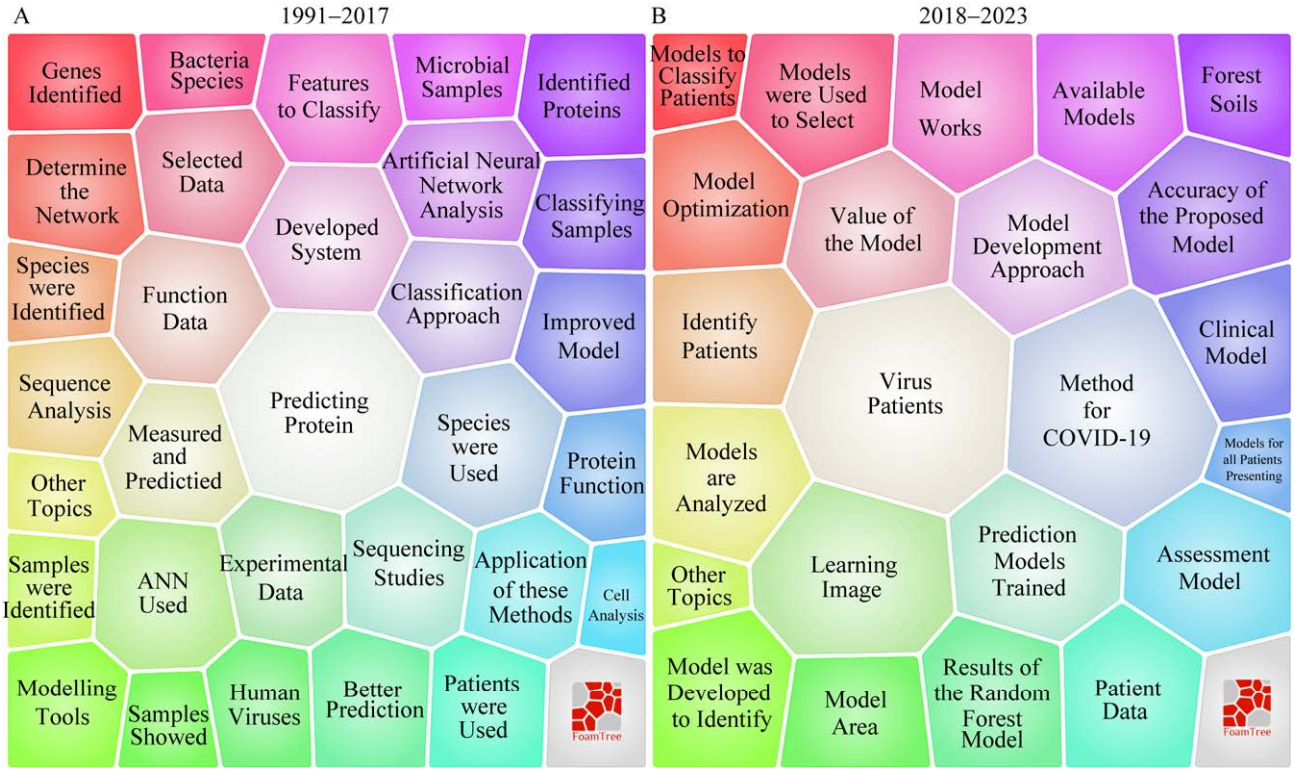

图 4　1991–2023 年机器学习在微生物生态研究中应用的主题词气泡图　A：1991–2017 年 2 027 篇文章的 31 个聚类. B：2018–2023 年 6 998 篇文章的 23 个聚类. 气泡面积的大小代表聚类的大小

Figure 4　Bubble diagram of the subject terms of machine learning research in microbial ecology, 1991−2023. A: 31 clusters for 2 027 articles from 1991−2017. B: 23 clusters for 6 998 articles from 2018−2023. The size of the bubble area represents the size of the clusters.

识别框架(data-driven keystone species identification, DKI)，该框架可以量化并识别微生物群落中物种的关键性。

总体来看，2 个时期都共同关注模型的开发方法和优化，2018 年以前研究热点为识别(identified)、分类(classification)等分子生物学方面的研究，2018 年以后随着机器学习方法的进步，该领域的研究热点转向复杂系统的预测，如病毒传播预测、微生物代谢产物预测、微生物物种群落组成预测等。

3　结论与展望

3.1　机器学习在微生物生态领域中应用的结论

借助 CiteSpace 和 Carrot2，本文使用 WOS 核心合集数据库作为数据源选取 1991–2023 年的 9 025 篇文献，基于文献计量学的研究探讨机器学习如何在微生物生态研究中应用，包括其现状、热点和未来发展方向，为微生物生态研究提供新的视角。主要结论如下：

(1) 基于时间脉络分析发现，自 1991 年以来，在微生物生态研究中应用机器学习方法发表的文章数量呈现逐年上升的明显趋势。由于深度学习方法的突破性发展，其应用量于 2018 年开始暴发增长。从合作网络分析来看，机器学习在微生物生态学的应用最早起源于欧美发达国家，并在欧洲国家间形成了密切的合作网络。近几年来，越来越多的发展中国家也选择了机器学习方法作为研究手段，形成了自己特色的研究合作网络。同时，中国的发文量虽位居第二，但中介中

心性较低，缺乏一定的国际合作和国际影响力。

(2) 研究使用双图叠加分析，通过研究施引文献与被引文献之间的联系，揭示了机器学习在微生物生态中应用的关键研究领域和发展路径。在不同的领域之间，引文链接的变化展示了交叉学科合作日益增强的趋势，尤其是生命科学领域与化学、物理、环境等学科之间的交叉合作，为科学研究提供了新的视角。突现检测的结果展示了计算机学科的发展，Scikit-learn 架构、QIIME 2 等计算方法的出现使机器学习的方法更容易被其他学科领域的研究人员接受，促进了微生物生态学的发展。

(3) 从 Carrot2 聚类获得的信息来看，2018 年前后的研究热点由于计算技术的进步有了明显的变化。当前的研究热点侧重于基于对复杂系统的预测，前沿研究主要围绕病毒传播预测、患者诊断、微生物生态系统建模开展。主要的研究思路是通过机器学习方法减少研究成本，更好地理解微生物生态。能完成这样的任务主要归功于深度学习的突破性发展，学习的深度能达到数千层，显著提高了其获得特征的能力。将深度学习应用于微生物生态学可以在缺乏先验知识的情况下准确地预测系统变化，利用已获取的信息预测难获取的信息，极大降低了研究门槛，也为微生物生态学理论的发展提供了依据。

3.2 机器学习在微生物生态领域中应用的展望

尽管机器学习作为微生物生态学研究中强大的分析工具，但它同时受到各种障碍的挑战，这些挑战阻碍了其广泛的应用[77]。常见的障碍与数据缺乏、模型评估和选择及可解释性有关，这些方面也是机器学习应用中关键的步骤(图 5)。

第一个挑战是缺少训练一个可靠的机器学习模型所需的大量、高质量和正确标记的数据[78]。监督式深度学习面临的最大挑战在于需要大型训练数据集才能获得高精度模型。训练数据集通常包含数千到数百万个样本，具体取决于任务、要检测的项目数量和所需的性能，但总体来说训练数据集越大，分类准确性就越高[79]。微生物生态学研究中产生的数据量庞大且多样化，包括环境样品的基因测序数据、代谢组数据等。然而，许多数据的质量并不高，由于采用不同的实验方法，不同项目之间的数据往往不具有可比性，获取这些数据涉及昂贵或耗时的过程[78]，所以需要开发高效的算法和工具处理这些数据，包括数据获取、清洗、对齐、去噪和归一化等步骤。Lo 等[80]从负二项分布中对微生物组图谱进行建模和采样，以扩大他们的训练数据集，提高其 CNN 模型的宿主表型分类性能。Sayyari 等[81]通过引入基于树的关联数据增强(tree-based associative data augmentation, TADA)方法，从推断的系统发育树中生成新的 OTU 样本，解决了样本数量少和代表性不足的普遍局限性。

第二个挑战是微生物生态研究者难以为给定任务选择和调整适当的机器学习模型。在众多模型中进行选择并寻找一组合适的超参数对生态学家是一项艰巨的跨学科任务，但如今持续开发的 Python 库、R 库和更易上手的高级框架促进了这项任务的完成。例如，Scikit-learn[41]、FastAI (https://docs.fast.ai/)、LightningAI (https://lightning.ai/)和 Keras (https://keras.io/)。另外，微生物组数据集还可以利用现有的框架生成，并用于调整和开发机器学习模型，常用的框架如宏基因组解读关键评估(critical assessment of metagenome interpretation, CAMI)计划提供的模拟宏基因组和微生物群落框架[82]。

第三个挑战是机器学习模型通常难以解释。机器学习凭经验在输入变量和响应变量之间建

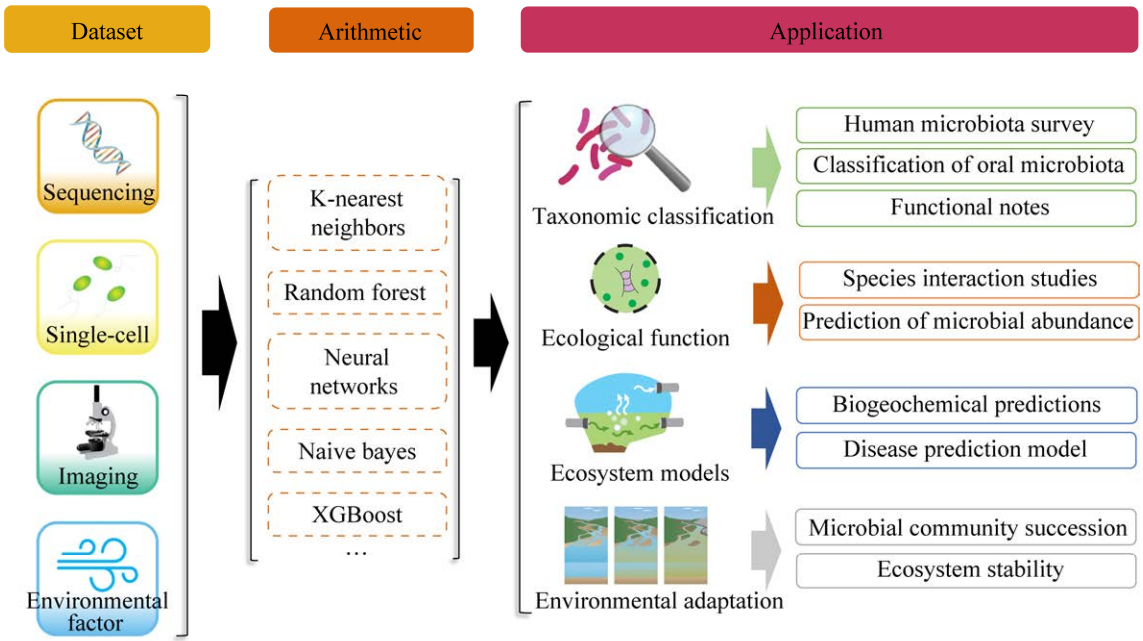

图 5 机器学习在微生物生态领域的应用与面对的挑战概述 图标由马里兰大学环境科学中心集成与应用网络提供(ian.umces.edu/symbols/)

Figure 5 Overview of machine learning applications and challenges in microbial ecology. Symbols courtesy of the Integration and Application Network, University of Maryland Center for Environmental Science (ian.umces.edu/symbols/).

立联系，但对这种关系背后的底层逻辑无任何机制理解，其内部结构被认为是无法解释的黑匣子[83]。人们对研究可解释机器学习的兴趣日益浓厚，但可解释性的概念定义不明确[84]，目前的研究通常让机器学习结合机理模型和非参数模型来提高可解释性，例如使用机理模型作为高斯过程的平均函数[85]或使用机理模型作为非参数模型的正则化先验[86]。一些特殊框架也可以提高机器学习的可解释性，如神经编码器-解码器模型 SparseNED[87]，该模型通过稀疏且可解释的潜在空间(latent space)捕获了与炎症性肠病相关的微生物-代谢物关系。Guidotti 等[88]对解释黑匣子的常用方法进行了全面综述。

综上所述，机器学习在微生物生态学中应用广泛，研究内容多元化，至今已积累了大量研究成果，为后续微生物生态的研究奠定了基础，但同时也存在一些挑战。根据热点趋势，未来机器学习的应用需重点关注三点。

(1) 加强数据可用性。相关研究者要对自己的数据结构有充分的了解，确定数据集是否包含不平衡的类，确定是否需要对其进行插补或特征工程[89]。同时加强国际合作，建立开放的公共注释数据库，以提高深度神经网络的训练效果。

(2) 关注新兴技术。机器学习方法在快速地更新迭代，但微生物生态领域对这一变化的响应有很多的滞后性，很多先进的方法没有第一时间应用到微生物生态学的研究。多关注最新的技术进展、加强学科融合，才能在微生物生态领域获得更多的突破。

(3) 结合先验知识提高模型的可解释性。目前机器学习在微生物生态学中很受欢迎，但几乎仅用于纯粹的预测目的，这些方法的全部潜力尚未完全发掘。机器学习模型在探索和假设生成方面通常有优越表现，而在正式验证大多数假设时

仍然需要更稳健的统计方法。可解释的机器学习模型将传统的模型与机器学习模型结合起来，可以各取所长。后续的研究要探索机器学习模型与先验模型的结合，提高模型的解释性，为微生物生态学理论的发展作出贡献。

REFERENCES

[1] HAZARD C, GOSLING P, van der GAST CJ, MITCHELL DT, DOOHAN FM, BENDING GD. The role of local environment and geographical distance in determining community composition of arbuscular mycorrhizal fungi at the landscape scale[J]. The ISME Journal, 2013, 7(3): 498-508.

[2] RAES J. Crowdsourcing earth's microbes[J]. Nature, 2017, 551: 446-447.

[3] SONG LY. Toward understanding microbial ecology to restore a degraded ecosystem[J]. International Journal of Environmental Research and Public Health, 2023, 20(5): 4647.

[4] GOLDMAN AD, KAÇAR B. Very early evolution from the perspective of microbial ecology[J]. Environmental Microbiology, 2023, 25(1): 5-10.

[5] VOLMER J, SCHMID A, BÜHLER B. Guiding bioprocess design by microbial ecology[J]. Current Opinion in Microbiology, 2015, 25: 25-32.

[6] LEMKE M, DeSALLE R. The role of microbial ecology in restoration ecology in the age of genomics: a summary of the microbial ecology special issue[J]. Microbial Ecology, 2023, 85(3): 1136-1141.

[7] GALLOWAY-PEÑA J, HANSON B. Tools for analysis of the microbiome[J]. Digestive Diseases and Sciences, 2020, 65(3): 674-685.

[8] KELLER M, SCHIMEL DS, HARGROVE WW, HOFFMAN FM. A continental strategy for the National Ecological Observatory Network[J]. Frontiers in Ecology and the Environment, 2008, 6(5): 282-284.

[9] HANSON PC, WEATHERS KC, DUGAN HA, GRIES C. The global lake ecological observatory network[J]. Ecological Informatics: Data Management and Knowledge Discovery, 2018: 415-433.

[10] COLLINS SL, BETTENCOURT LM, HAGBERG A, BROWN RF, MOORE DI, BONITO G, DELIN KA, JACKSON SP, JOHNSON DW, BURLEIGH SC, WOODROW RR, McAULEY JM. New opportunities in ecological sensing using wireless sensor networks[J]. Frontiers in Ecology and the Environment, 2006, 4(8): 402-407.

[11] PORTER JH, NAGY E, KRATZ TK, HANSON P, COLLINS SL, ARZBERGER P. New eyes on the world: advanced sensors for ecology[J]. BioScience, 2009, 59(5): 385-397.

[12] KELLING S, HOCHACHKA WM, FINK D, RIEDEWALD M, CARUANA R, BALLARD G, HOOKER G. Data-intensive science: a new paradigm for biodiversity studies[J]. BioScience, 2009, 59(7): 613-620.

[13] JORDAN MI, MITCHELL TM. Machine learning: trends, perspectives, and prospects[J]. Science, 2015, 349(6245): 255-260.

[14] YU T, SU SX, HU J, ZHANG J, XIANYU YL. A new strategy for microbial taxonomic identification through micro-biosynthetic gold nanoparticles and machine learning[J]. Advanced Materials, 2022, 34(11): e2109365.

[15] DUTTA A, GOLDMAN T, KEATING J, BURKE E, WILLIAMSON N, DIRMEIER R, BOWMAN JS. Machine learning predicts biogeochemistry from microbial community structure in a complex model system[J]. Microbiology Spectrum, 2022, 10(1): e0190921.

[16] KE MJ, XU NH, ZHANG ZY, QIU DY, KANG J, LU T, WANG TZ, PEIJNENBURG WJGM, SUN LW, HU BL, QIAN HF. Development of a machine-learning model to identify the impacts of pesticides characteristics on soil microbial communities from high-throughput sequencing data[J]. Environmental Microbiology, 2022, 24(11): 5561-5573.

[17] OYETUNDE T, LIU D, MARTIN HG, TANG YJ. Machine learning framework for assessment of microbial factory performance[J]. PLoS One, 2019, 14(1): e0210558.

[18] WANG XW, SUN Z, JIA HJ, MICHEL-MATA S, ANGULO MT, DAI L, HE XS, WEISS ST, LIU YY. Identifying keystone species in microbial communities using deep learning[J]. Nature Ecology & Evolution, 2024, 8: 22-31.

[19] MEI R, KIM J, WILSON FP, BOCHER BTW, LIU WT. Coupling growth kinetics modeling with machine learning reveals microbial immigration impacts and identifies key environmental parameters in a biological wastewater treatment process[J]. Microbiome, 2019, 7(1): 65.

[20] PACHECO AR, SEGRÈ D. An evolutionary algorithm for designing microbial communities via environmental modification[J]. Journal of the Royal Society Interface, 2021, 18(179): 20210348.

[21] CHEN CM. CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature[J]. Journal of the American Society for Information Science and Technology, 2006, 57(3): 359-377.

[22] CHEN CM, DUBIN R, KIM MC. Emerging trends and new developments in regenerative medicine: a scientometric update (2000—2014)[J]. Expert Opinion on Biological Therapy, 2014, 14(9): 1295-1317.

[23] CHEN CM, LEYDESDORFF L. Patterns of connections and movements in dual-map overlays: a new method of publication portfolio analysis[J]. Journal of the

Association for Information Science and Technology, 2014, 65(2): 334-351.

[24] 李杰, 陈超美. CiteSpace: 科技文本挖掘及可视化[M]. 北京: 首都经济贸易大学出版社, 2016.
LI J, CHEN CM. CiteSpace: Text Mining and Visualization in Scientific Literature[M]. Beijing: Capital University of Economics & Business Press, 2016 (in Chinese).

[25] PEARL J, MACKENZIE D. The Book of Why: the New Science of Cause and Effect[M]. New York: Basic Books, 2018.

[26] NOVAC OC, CHIRODEA MC, NOVAC CM, BIZON N, OPROESCU M, STAN OP, GORDAN CE. Analysis of the application efficiency of TensorFlow and PyTorch in convolutional neural network[J]. Sensors, 2022, 22(22): 8872.

[27] 叶靓俏, 尹彩春, 赵文武. 基于文献计量分析的联合国可持续发展目标研究[J]. 生态学报, 2023, 43(24): 10480-10489.
YE LQ, YI CC, ZHAO WW. Bibliometric analysis of research on UN Sustainable development goals based on Web of Science [J]. Acta Ecologica Sinica, 2023, 43(24): 10480-10489 (in Chinese).

[28] RICHARDS TA, MASSANA R, PAGLIARA S, HALL N. Single cell ecology[J]. Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 2019, 374(1786): 20190076.

[29] 辛雨菡, 崔丽. 单细胞稳定同位素标记技术在固氮微生物中的应用研究[J]. 微生物学报, 2020, 60(9): 1772-1783.
XIN YH, CUI L. Application of single-cell stable isotope probing approach to investigate N2-fixing microorganisms[J]. Acta Microbiologica Sinica, 2020, 60(9): 1772-1783 (in Chinese).

[30] MA Q, XU D. Deep learning shapes single-cell data analysis[J]. Nature Reviews Molecular Cell Biology, 2022, 23: 303-304.

[31] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, MICHEL V, THIRION B, GRISEL O, BLONDEL M, PRETTENHOFER P, WEISS R, DUBOURG V, VANDERPLAS J, PASSOS A, COURNAPEAU D, BRUCHER M, PERROT M, DUCHESNAY É. Scikit-learn: machine learning in python[J]. Journal of Machine Learning Research, 2011, 12(85): 2825-2830.

[32] BOLYEN E, RIDEOUT JR, DILLON MR, BOKULICH NA, ABNET CC, AL-GHALITH GA, ALEXANDER H, ALM EJ, ARUMUGAM M, ASNICAR F. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2[J]. Nature biotechnology, 2019, 37(8): 852-857.

[33] HE K, ZHANG X, REN S, SUN J. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 770-778.

[34] CALLAHAN BJ, MCMURDIE PJ, ROSEN MJ, HAN AW, JOHNSON AJA, HOLMES SP. DADA2: high-resolution sample inference from illumina amplicon data[J]. Nature Methods, 2016, 13(7): 581-583.

[35] OZTURK T, TALO M, YILDIRIM EA, BALOGLU UB, YILDIRIM O, RAJENDRA ACHARYA U. Automated detection of COVID-19 cases using deep neural networks with X-ray images[J]. Computers in Biology and Medicine, 2020, 121: 103792.

[36] APOSTOLOPOULOS ID, MPESIANA TA. COVID-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks[J]. Physical and Engineering Sciences in Medicine, 2020, 43(2): 635-640.

[37] SIMONYAN K. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv: 1409.1556, 2014

[38] CHANG CC, LIN CJ. LIBSVM: a library for support vector machines[J]. ACM transactions on intelligent systems and technology (TIST), 2011, 2(3): 1-27.

[39] CAPORASO JG, KUCZYNSKI J, STOMBAUGH J, BITTINGER K, BUSHMAN FD, COSTELLO EK, FIERER N, PEÑA AG, GOODRICH JK, GORDON JI, HUTTLEY GA, KELLEY ST, KNIGHTS D, KOENIG JE, LEY RE, LOZUPONE CA, MCDONALD D, MUEGGE BD, PIRRUNG M, REEDER J, et al. QIIME allows analysis of high-throughput community sequencing data[J]. Nature Methods, 2010, 7(5): 335-336.

[40] Chen T, Guestrin C. XGBoost: a scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery, 2016: 785-794.

[41] ABRAHAM A, PEDREGOSA F, EICKENBERG M, GERVAIS P, MUELLER A, KOSSAIFI J, GRAMFORT A, THIRION B, VAROQUAUX G. Machine learning for neuroimaging with scikit-learn[J]. Frontiers in Neuroinformatics, 2014, 8: 14.

[42] CAPANEMA CGS, de OLIVEIRA GS, SILVA FA, SILVA TRMB, LOUREIRO AAF. Combining recurrent and Graph Neural Networks to predict the next place's category[J]. Ad Hoc Networks, 2023, 138: 103016.

[43] DENG L, ZHU F, HE Y, MENG FW. Prediction of post-translational modification cross-talk and mutation within proteins via imbalanced learning[J]. Expert Systems with Applications, 2023, 211: 118593.

[44] ZHANG J, ZHOU PF, ZHENG YJ, WU HY. Predicting influenza with pandemic-awareness via dynamic virtual graph significance networks[J]. Computers in Biology and Medicine, 2023, 158: 106807.

[45] DILLON MR, BOLYEN E, ADAMOV A, BELK A, BORSOM E, BURCHAM Z, DEBELIUS JW, DEEL H, EMMONS A, ESTAKI M, HERMAN C, KEEFE CR, MORTON JT, OLIVEIRA RRM, SANCHEZ A, SIMARD A, VÁZQUEZ-BAEZA Y, ZIEMSKI M, MIWA HE, KERERE TA, et al. Experiences and lessons learned from two virtual, hands-on microbiome

bioinformatics workshops[J]. PLoS Computational Biology, 2021, 17(6): e1009056.
[46] KNYAZEV S, CHHUGANI K, SARWAL V, AYYALA R, SINGH H, KARTHIKEYAN S, DESHPANDE D, BAYKAL PI, COMAROVA Z, LU A, POROZOV Y, VASYLYEVA TI, WERTHEIM JO, TIERNEY BT, CHIU CY, SUN R, WU AP, ABEDALTHAGAFI MS, PAK VM, NAGARAJ SH, et al. Unlocking capacities of genomics for the COVID-19 response and future pandemics[J]. Nature Methods, 2022, 19: 374-380.
[47] ELDRED LE, THORN RG, SMITH DR. Simple matching using QIIME 2 and RDP reveals misidentified sequences and an underrepresentation of fungi in reference datasets[J]. Frontiers in Genetics, 2021, 12: 768473.
[48] REUTER JA, SPACEK DV, SNYDER MP. High-throughput sequencing technologies[J]. Molecular Cell, 2015, 58(4): 586-597.
[49] SCHLOSS JA, GIBBS RA, MAKHIJANI VB, MARZIALI A. Cultivating DNA sequencing technology after the human genome project[J]. Annual Review of Genomics and Human Genetics, 2020, 21: 117-138.
[50] THOMPSON LR, SANDERS JG, McDONALD D, AMIR A, LADAU J, LOCEY KJ, PRILL RJ, TRIPATHI A, GIBBONS SM, ACKERMANN G, NAVAS-MOLINA JA, JANSSEN S, KOPYLOVA E, VÁZQUEZ-BAEZA Y, GONZÁLEZ A, MORTON JT, MIRARAB S, XU ZZ, JIANG LJ, HAROON MF, et al. A communal catalogue reveals Earth's multiscale microbial diversity[J]. Nature, 2017, 551(7681): 457-463.
[51] GOODSWEN SJ, BARRATT JLN, KENNEDY PJ, KAUFER A, CALARCO L, ELLIS JT. Machine learning and applications in microbiology[J]. FEMS Microbiology Reviews, 2021, 45(5): fuab015.
[52] WANG G, MULLIGAN MJ. Comparative sequence analysis and predictions for the envelope glycoproteins of foamy viruses[J]. The Journal of General Virology, 1999, 80(Pt 1): 245-254.
[53] KALATE RN, TAMBE SS, KULKARNI BD. Artificial neural networks for prediction of mycobacterial promoter sequences[J]. Computational Biology and Chemistry, 2003, 27(6): 555-564.
[54] DOLLHOPF SL, HASHSHAM SA, TIEDJE JM. Interpreting 16S rDNA T-RFLP data: application of self-organizing maps and principal component analysis to describe community dynamics and convergence[J]. Microbial Ecology, 2001, 42(4): 495-505.
[55] HABIB AR, LO NC. Predicting COVID-19 outcomes[J]. BMJ, 2022, 376: o354.
[56] ARDABILI S, MOSAVI A, GHAMISI P, FERDINAND F, VARKONYI-KOCZY A, REUTER U, RABCZUK T, ATKINSON P. COVID-19 outbreak prediction with machine learning[J]. Algorithms, 2020, 13(10): 249.
[57] NASUTION H, KHAIRANI N, AHYANINGSIH F, ALAMSYAH F. Mathematical modeling of the spread of corona virus disease 19 (COVID-19) with vaccines[C]// The 8th Annual Interational Seminal on Trends in Science and Science Education (AISTSSE) 2021, AIP Conference Proceedings. Medan, Indonesia. AIP Publishing, 2022, 2659(1): 110009.
[58] HE S, TANG SY, RONG LB. A discrete stochastic model of the COVID-19 outbreak: forecast and control[J]. Mathematical Biosciences and Engineering: MBE, 2020, 17(4): 2792-2804.
[59] ALANAZI SA, KAMRUZZAMAN MM, ALRUWAILI M, ALSHAMMARI N, ALQAHTANI SA, KARIME A. Measuring and preventing COVID-19 using the SIR model and machine learning in smart health care[J]. Journal of Healthcare Engineering, 2020, 2020: 8857346.
[60] LAMMEL DR, BARTH G, OVASKAINEN O, CRUZ LM, ZANATTA JA, RYO M, de SOUZA EM, PEDROSA FO. Direct and indirect effects of a pH gradient bring insights into the mechanisms driving prokaryotic community structures[J]. Microbiome, 2018, 6(1): 106.
[61] FEESER KL, van HORN DJ, BUELOW HN, COLMAN DR, McHUGH TA, OKIE JG, SCHWARTZ E, TAKACS-VESBACH CD. Local and regional scale heterogeneity drive bacterial community diversity and composition in a polar desert[J]. Frontiers in Microbiology, 2018, 9: 1928.
[62] LUO GW, RENSING C, CHEN H, LIU MQ, WANG M, GUO SW, LING N, SHEN QR. Deciphering the associations between soil microbial diversity and ecosystem multifunctionality driven by long-term fertilization management[J]. Functional Ecology, 2018, 32(4): 1103-1116.
[63] THANGAVEL R, KANCHIKERIMATH M, SUDHARSANAM A, AYYANADAR A, KARUNANITHI R, DESHMUKH NA, VANAO NS. Evaluating organic carbon fractions, temperature sensitivity and artificial neural network modeling of CO_2 efflux in soils: impact of land use change in subtropical India (*Meghalaya*)[J]. Ecological Indicators, 2018, 93: 129-141.
[64] ZHANG H, WU PB, FAN MM, ZHENG SY, WU JT, YANG XH, ZHANG M, YIN AJ, GAO C. Dynamics and driving factors of the organic carbon fractions in agricultural land reclaimed from coastal wetlands in Eastern China[J]. Ecological Indicators, 2018, 89: 639-647.
[65] JIANG YJ, ZHOU H, CHEN LJ, YUAN Y, FANG H, LUAN L, CHEN Y, WANG XY, LIU MQ, LI HX, PENG XH, SUN B. Nematodes and microorganisms interactively stimulate soil organic carbon turnover in the macroaggregates[J]. Frontiers in Microbiology, 2018, 9: 2803.
[66] DiMUCCI D, KON M, SEGRÈ D. Machine learning reveals missing edges and putative interaction

mechanisms in microbial ecosystem networks[J]. mSystems, 2018, 3(5): e00181-18.
[67] DAMASO N, MENDEL J, MENDOZA M, von WETTBERG EJ, NARASIMHAN G, MILLS D. Bioinformatics approach to assess the biogeographical patterns of soil communities: the utility for soil provenance[J]. Journal of Forensic Sciences, 2018, 63(4): 1033-1042.
[68] RAMIREZ KS, KNIGHT CG, de HOLLANDER M, BREARLEY FQ, CONSTANTINIDES B, COTTON A, CREER S, CROWTHER TW, DAVISON J, DELGADO-BAQUERIZO M, DORREPAAL E, ELLIOTT DR, FOX G, GRIFFITHS RI, HALE C, HARTMAN K, HOULDEN A, JONES DL, KRAB EJ, MAESTRE FT, et al. Detecting macroecological patterns in bacterial communities across independent studies of global soils[J]. Nature Microbiology, 2018, 3: 189-196.
[69] KNIGHTS D, COSTELLO EK, KNIGHT R. Supervised classification of human microbiota[J]. FEMS Microbiology Reviews, 2011, 35(2): 343-359.
[70] CHEN YL, KOU D, LI F, DING JZ, YANG GB, FANG K, YANG YH. Linkage of plant and abiotic properties to the abundance and activity of N-cycling microbial communities in Tibetan permafrost-affected regions[J]. Plant and Soil, 2019, 434(1): 453-466.
[71] PEI YT, HUANG YP, ZOU Q, ZHANG XY, WANG S. Effects of image degradation and degradation removal to CNN-based image classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(4): 1239-1253.
[72] KOSOV S, SHIRAHAMA K, LI C, GRZEGORZEK M. Environmental microorganism classification using conditional random fields and deep convolutional neural networks[J]. Pattern Recognition, 2018, 77: 248-261.
[73] TAHIR MW, ZAIDI NA, RAO AA, BLANK R, VELLEKOOP MJ, LANG W. A fungus spores dataset and a convolutional neural network based approach for fungus detection[J]. IEEE Transactions on NanoBioscience, 2018, 17(3): 281-290.
[74] CHEN RTQ, RUBANOVA Y, BETTENCOURT J, DUVENAUD D. Neural ordinary differential equations[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada. ACM, 2018: 6572-6583.
[75] MICHEL-MATA S, WANG XW, LIU YY, ANGULO MT. Predicting microbiome compositions from species assemblages through deep learning[J]. iMeta, 2022, 1(1): e3.
[76] WANG T, WANG XW, LEE-SARWAR KA, LITONJUA AA, WEISS ST, SUN YZ, MASLOV S, LIU YY. Predicting metabolomic profiles from microbial composition through neural ordinary differential equations[J]. Nature Machine Intelligence, 2023, 5: 284-293.
[77] CHING T, HIMMELSTEIN DS, BEAULIEU-JONES BK, KALININ AA, DO BT, WAY GP, FERRERO E, AGAPOW PM, ZIETZ M, HOFFMAN MM, XIE W, ROSEN GL, LENGERICH BJ, ISRAELI J, LANCHANTIN J, WOLOSZYNEK S, CARPENTER AE, SHRIKUMAR A, XU JB, COFER EM, et al. Opportunities and obstacles for deep learning in biology and medicine[J]. Journal of the Royal Society, Interface, 2018, 15(141): 20170387.
[78] ADADI A. A survey on data-efficient algorithms in big data era[J]. Journal of Big Data, 2021, 8(1): 24.
[79] CHRISTIN S, HERVET É, LECOMTE N. Applications for deep learning in ecology[J]. Methods in Ecology and Evolution, 2019, 10(10): 1632-1644.
[80] LO C, MARCULESCU R. MetaNN: accurate classification of host phenotypes from metagenomic data using neural networks[J]. BMC Bioinformatics, 2019, 20(Suppl 12): 314.
[81] SAYYARI E, KAWAS B, MIRARAB S. TADA: phylogenetic augmentation of microbiome samples enhances phenotype classification[J]. Bioinformatics, 2019, 35(14): i31-i40.
[82] FRITZ A, HOFMANN P, MAJDA S, DAHMS E, DRÖGE J, FIEDLER J, LESKER TR, BELMANN P, DeMAERE MZ, DARLING AE, SCZYRBA A, BREMGES A, McHARDY AC. CAMISIM: simulating metagenomes and microbial communities[J]. Microbiome, 2019, 7(1): 17.
[83] DOMINGOS P. A few useful things to know about machine learning[J]. Communications of the ACM, 2012, 55(10): 78-87.
[84] LUCAS TCD. A translucent box: interpretable machine learning in ecology[J]. Ecological Monographs, 2020, 90(4): e01422.
[85] RASMUSSEN CE. Gaussian Processes in Machine Learning[M]//Advanced Lectures on Machine Learning. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004: 63-71.
[86] LYDDON S, WALKER S, HOLMES C. Nonparametric learning from Bayesian models with randomized objective functions[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada. ACM, 2018: 2075-2085.
[87] LE V, QUINN TP, TRAN T, VENKATESH S. Deep in the bowel: highly interpretable neural encoder-decoder networks predict gut metabolites from gut microbiome[J]. BMC Genomics, 2020, 21(Suppl 4): 256.
[88] GUIDOTTI R, MONREALE A, RUGGIERI S, TURINI F, GIANNOTTI F, PEDRESCHI D. A survey of methods for explaining black box models[J]. ACM Computing Surveys, 2018, 51(5): 93.
[89] HERNÁNDEZ MEDINA R, KUTUZOVA S, NIELSEN KN, JOHANSEN J, HANSEN LH, NIELSEN M, RASMUSSEN S. Machine learning and deep learning applications in microbiome research[J]. ISME Communications, 2022, 2: 98.