微生物学报
Acta Microbiologica Sinica
2022, 62(12): 4646–4662
http://journals.im.ac.cn/actamicrocn
DOI: 10.13343/j.cnki.wsxb.20220389

环境微生物研究中机器学习算法及应用

陈鹤[1,2]，陶晔[1,2]，毛振镀[1,2]，邢鹏[1*]

1 中国科学院南京地理与湖泊研究所，湖泊与环境国家重点实验室，江苏 南京 210008
2 中国科学院大学，北京 100049

陈鹤, 陶晔, 毛振镀, 邢鹏. 环境微生物研究中机器学习算法及应用. 微生物学报, 2022, 62(12): 4646–4662.
Chen He, Tao Ye, Mao Zhendu, Xing Peng. A review of machine learning algorithms for environmental microbiology. *Acta Microbiologica Sinica*, 2022, 62(12): 4646–4662.

摘　要：微生物在环境中无处不在，它们不仅是生物地球化学循环和环境演化的关键参与者，也在环境监测、生态治理和保护中发挥着重要作用。随着高通量技术的发展，大量微生物数据产生，运用机器学习对环境微生物大数据进行建模和分析，在微生物标志物识别、污染物预测和环境质量预测等领域的科学研究和社会应用方面均具有重要意义。机器学习可分为监督学习和无监督学习 2 大类。在微生物组学研究当中，无监督学习通过聚类、降维等方法高效地学习输入数据的特征，进而对微生物数据进行整合和归类。监督学习运用有特征和标记的微生物数据集训练模型，在面对只有特征没有标记的数据时可以判断出标记，从而实现对新数据的分类、识别和预测。然而，复杂的机器学习算法通常以牺牲可解释性为代价来重点关注模型预测的准确性。机器学习模型通常可以看作预测特定结果的“黑匣子”，即对模型如何得出预测所知甚少。为了将机器学习更多地运用于微生物组学研究、提高我们提取有价值的微生物信息的能力，深入了解机器学习算法、提高模型的可解释性尤为重要。本文主要介绍在环境微生物领域常用的机器学习算法和基于微生物组数据的机器学习模型的构建步骤，包括特征选择、算法选择、模型构建和评估等，并对各种机器学习模型在环境微生物领域的应用进行综述，深入探究微生物组与周围环境之间的关联，探讨提高模型可解释性的方法，并为未来环境监测、环境健康预测提供科学参考。

关键词：机器学习；微生物组；环境微生物；16S rRNA 基因；宏基因组

基金项目：国家自然科学基金(91751111，31670505，31722008)；江苏省自然科学基金(BK20220015)；中国科学院青年创新促进会(2014273)

Supported by the National Natural Science Foundation of China (91751111, 31670505, 31722008), by the Natural Science Foundation of Jiangsu (BK20220015) and by the Youth Innovation Promotion Association of CAS (2014273)

***Corresponding author.** Tel: +86-25-86882112; Fax: +86-25-57714759; E-mail: pxing@niglas.ac.cn

Received: 24 May 2022; **Revised:** 18 July 2022; **Published online:** 22 July 2022

A review of machine learning algorithms for environmental microbiology

CHEN He[1,2], TAO Ye[1,2], MAO Zhendu[1,2], XING Peng[1*]

1 State Key Laboratory of Lake Science and Environment, Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Nanjing 210008, Jiangsu, China

2 University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: Ubiquitous microorganisms, key players in biogeochemical cycles and environmental evolution, are involved in environmental monitoring as well as ecological governance and protection. The booming high-throughput technologies have generated massive microbial data and expanded the scope of microbiome research. Constructing machine learning models to analyze complex microbial data is of great importance to microbial marker identification, pollutant prediction, and environmental quality prediction. Machine learning algorithms can be classified into two categories: supervised learning and unsupervised learning. In microbiome research, unsupervised learning grasps the characteristics of input data through clustering and dimensionality reductions, enabling the integration and classification of microbial data. Supervised learning uses microbial datasets with features and labels to train and build models that can be used to classify, identify, and predict new data without labels. However, sophisticated machine learning algorithms often focus on the accuracy of model predictions at the expense of interpretability. Machine learning models can often be regarded as a "black box" that predicts a specific outcome. Little is known about how the prediction is obtained by the model. Improving model interpretability is critical for the accurate application of machine learning and the extraction of valuable biological information in microbiome research. This review introduced the machine learning algorithms commonly used in environmental microbiology and the construction steps (including feature selection, algorithm selection, model construction and evaluation) of machine learning models based on microbiome data. Furthermore, we summarized several application scenarios of machine learning models in environmental microbiology for in-depth exploration of the relationship between the microbiome and the surrounding environment, attempting to improve the interpretability of the model and provide a reference for future environmental monitoring and environmental health prediction.

Keywords: machine learning; microbiome; environmental microorganisms; 16S rRNA gene; metagenome

微生物广泛存在于地球上所有环境之中甚至是在极端环境中(如极地、高盐湖泊和热泉等)[1]，其具有物种多样性和功能多样性，在驱动生物地球化学循环和物质能量代谢等方面发挥着重要作用[2]。环境微生物学(environmental microbiology)在于探究自然环境中的微生物群落、结构、功能多样性以及微生物与不同环境之间的关联。微生物组学(microbiome)是研究给定环境中的所有微生物及其遗传信息和功能的集合[3]。对微生物组的研究还包括不同微生物之间的相互作用[4]、微生物与其他生物之间的相互作用[5]以及微生物与环境之间的相互作用[6]。大量研究利用培养技术探索微生物的生长条件以及它们与环境的相互作用，然而微生物在培养基与真实环境中的生长状况并不相同，并且存在很多无法被培养的微生物，因此

仅仅使用培养技术无法全面了解特定环境中的微生物多样性。

近年来，扩增子测序、宏基因组测序、宏转录组测序等高通量测序技术广泛用于揭示微生物种群结构、基因功能活性、微生物间的相互协作关系以及微生物与环境因子之间的关系[7]。在扩增子测序中，具有高度保守性的特定基因、基因片段或序列[如 16S/18S rRNA 基因和真菌的核糖体 DNA 内转录间隔区(ITS)序列]会被扩增，以确定样本中的物种组成和群落多样性等[7]。宏基因组测序无需分离培养，可直接对样本中全部微生物的完整基因组进行测序，不仅可以揭示微生物多样性、种群结构和进化关系，还可以进行基因功能层面的深入研究。宏转录组侧重于研究活跃微生物组基因的表达及其对环境的响应。随着高通量测序技术的进步和测序成本的不断下降，大量的基因数据随之产生。一项微生物组研究可能包含数百 Gb 或更多的原始测序数据。据统计，在 2025 年为基因组数据所提供的存储空间(~2−40 exabytes per year)将会超过天文数据所需的存储量(~1 exabyte per year)[8]。这些大量且复杂的微生物数据往往蕴涵着很多有价值的信息。例如某些天然的微生物群落与环境因子密切相关，因此它们可以用来预测环境现象。然而传统的统计学方法已经无法满足大样本量的微生物组学研究，而适用于复杂数据分析的机器学习逐渐成为首选方法[9]。近年来，机器学习逐渐成了微生物组学研究中的常见词。Yatsunenko 等的研究证明了机器学习能够对物种水平的 OTU 进行有效且准确的分类，并且可以找出能够区分组间差异的关键成分(OTU 或物种)[10]。本综述重点介绍在环境微生物研究中常用的机器学习算法及其应用。目前机器学习方法主要应用于分类问题和预测问题，因此我们着重专注于这 2 个方面。

1 环境微生物领域的机器学习概述

1.1 机器学习定义、分类和模型构建

机器学习是一种结合了统计学、概率论和线性代数等多种数学方法可以实现从经验中自动学习的计算机算法，其在解决聚类、分类及回归等问题中表现出独有的优势。目前，机器学习已被广泛应用于许多科学领域，例如生物信息学[11−13]、生物化学[14−15]、医学[16]、水产养殖[17−18]和气候学[19]等。在微生物组学领域，机器学习运用了合适的算法使得计算机从复杂且大量的基因数据中归纳逻辑和总结规律，从而建立经验模型，根据模型去识别、分类或预测新数据[20]。

根据训练数据是否有标记(label)，可以将机器学习分为 2 大类——监督学习和无监督学习。有标记的为监督学习，无标记的为无监督学习(图 1 和图 2A)。无监督学习算法可以对没有标记的数据集进行学习，根据聚类得到数据间的关系，并将其可视化。除了聚类，无监督学习算法还可以利用降维去降低数据的疏散程度和复杂程度[13]。这类算法常用于对高纬度的宏基因组数据进行初步的探索性分析，并为后续分析提出假设。监督学习运用有已知特征和标记的微生物组数据集作为训练集，通过训练让计算机找到特征与标记间的联系，从而构建规则(即模型)，在面对只有特征没有标记的数据(即样本)时，可以判断出标记。

构建运用于微生物组研究的机器学习模型主要包含 3 个步骤。以扩增子测序为例，第一步是特征选择(feature selection)，从特征集中选择出重要的特征子集。即从 16S rRNA 测序产生的数据集中创建特征集(带有物种注释的 OTU/ASV 表)，再通过相应的特征选择方法(具

体的特征选择方法将会在下文中介绍)选择出重要的特征子集作为模型的训练数据集(training datasets)。第二步是根据所要解决的问题选择出合适的机器学习算法，并利用训练数据集来训练模型。例如选择朴素贝叶斯算法，利用细菌16S rRNA数据集训练模型，最终构建的朴素贝叶斯分类器可对扩增子数据进行物种分类。最后一步是利用测试集通过交叉验证等方法来检验模型参数、评估模型预测的好坏。图1展示了机器学习应用于环境微生物领域的常规思路。

1.2 微生物组研究中常用的特征选择方法

特征选择是从众多特征中选择可用子集的过程。微生物组学的数据特征往往超过样本的数量，这会导致模型出现过拟合，提供过度乐观的模型评估，可能影响交叉验证[21]。因此，通过特征选择来优化模型性能，以提高模型对于新数据的泛化程度。过滤法是特征选择方法中最简单的一种方法，它通常是在建模之前执行的预处理步骤，不依赖于模型。然而过滤法对于解释潜在的特征异质性、微生物群落数据的多种共线性和复杂的协方差结构时具有挑战性[2]。包裹法不独立于模型，而是使用模型本身来选择特征子集，因此所选出的特征子集可以使模型发挥最佳性能。然而特征选择的过程需要重复训练模型，因此对于宏基因组的高纬数据结构，包裹法进行特征选择的计算成本往往比过滤法高得多[2]。嵌入法与包裹法类似，但它既不需要较高的计算成本，又具有过滤法的高效性。因此，嵌入法在优化基于微生物组学研究的机器学习模型的特征选择时更加适

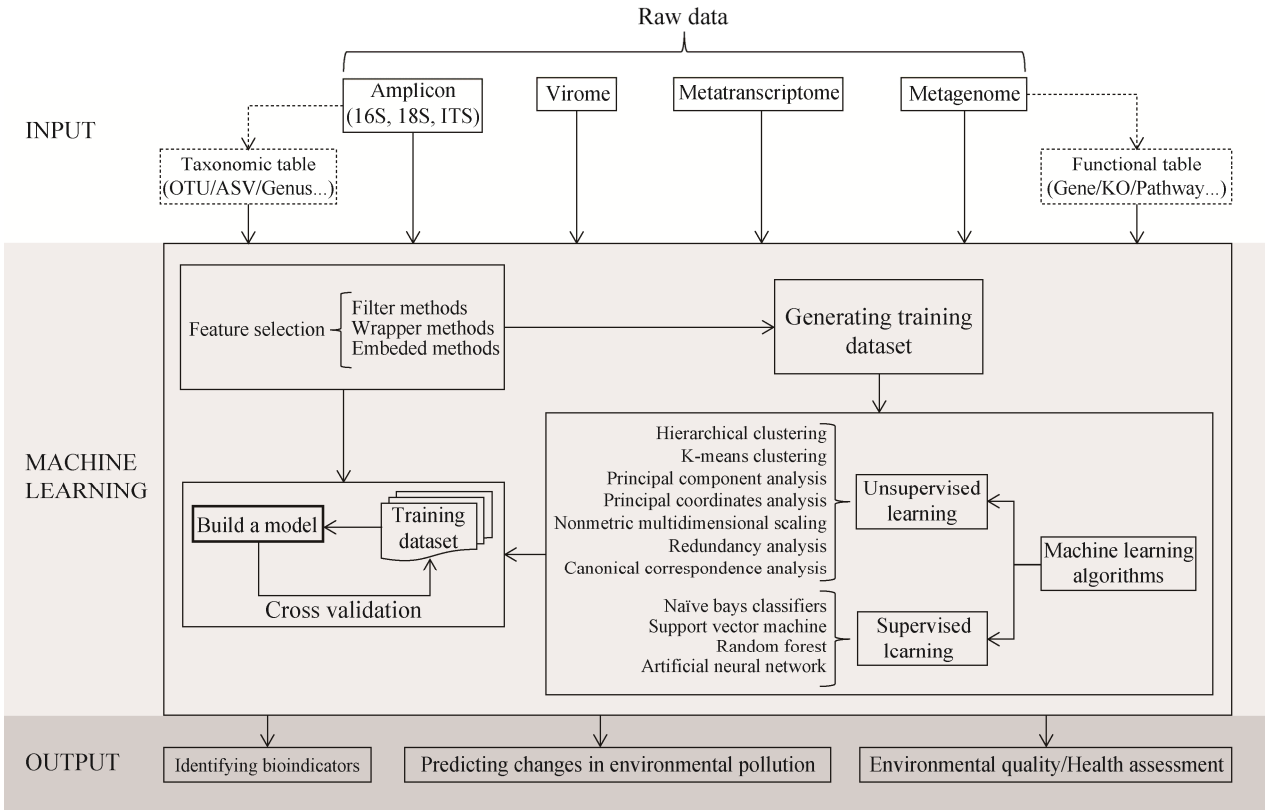


图1 机器学习应用于环境微生物学的思路框架

Figure 1 A framework of machine learning based on environmental microbiology.

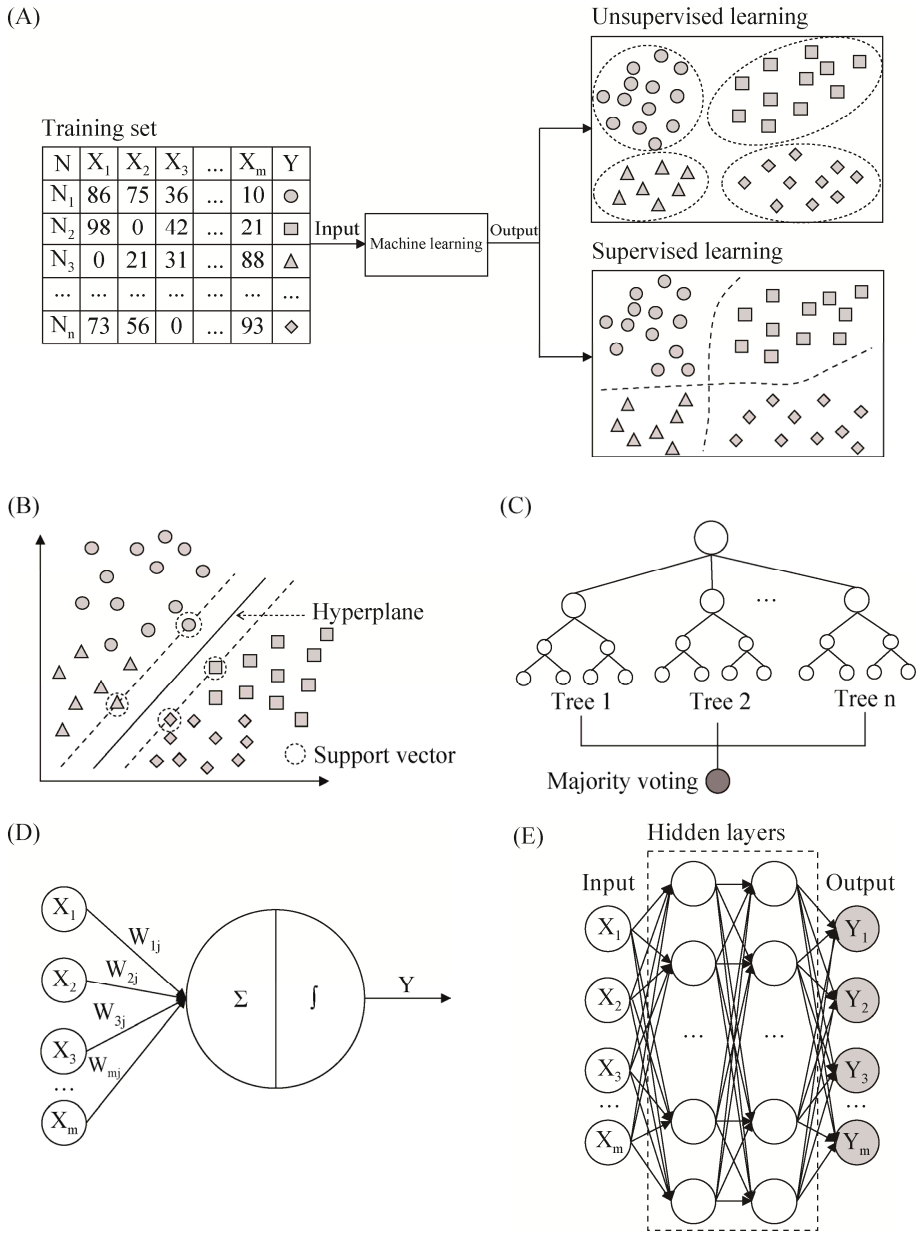


图 2　监督学习和无监督学习原理示意图

Figure 2　Schematic diagrams of supervised and unsupervised forms of learning (A) schematic representation of training set, supervised learning, and unsupervised learning. A training set containing samples (N), features (X_1–X_m) and multiple class labels (Y) could be used to build a model to either predict which cluster Y belongs to (unsupervised) or to find a best fit decision boundary between X and Y (supervised). B: a schematic representation of SVM. The hyperplane maximizes the distance between the data points on both sides of the hyperplane. C: a schematic representation of RF. RF consists of multiple decision tree. D: a schematic representation of M-P model. The model is similar to a neuron (j), which can simultaneously accept multiple input signals (X_i) with different weights (w_{ij}) before summing all the received signals ($\sum$) and processing them through an activation function ($\int$) to produces the output (Y). E: a schematic representation of an artificial neural network. In neural network, the input data (X) is processed through multiple hidden layers to predict the output (Y).

用。例如 metagenomic prediction analysis based on machine learning (MetAML)在随机森林和支持向量机模型中结合嵌入式特征选择方法(如least absolute shrinkage and selection operato (Lasso)和 elastic net (ENet))进行微生物的宏基因组分类[22]。近年来，具有生物学动机的新特征选择技术，例如 taxonomy-aware hierarchical feature engineering (HFE)[21]开始受到关注。尽管该技术在目前的微生物组学研究中还没有得到广泛的应用和报道，但在处理非常高维的数据集时，它可以解决嵌入法难以使用完整搜索空间的问题[2]。

1.3 环境微生物研究中常用的机器学习算法

1.3.1 无监督学习算法

无监督学习主要用于 2 个方面：基于成对相似性度量的数据点分组(聚类)或从高纬度数据中提取代表性特征(降维)。在微生物组学的研究中，无监督学习模型的输入数据可以是连续的数据集，也可以是群落组成之间相似性的距离矩阵。聚集在一起的样本代表了它们之间的微生物群落组成相似。常用的聚类方法有层次聚类和 K-means 聚类等。在微生物组学的研究中，Bray-Curtis 相异度和 UniFrac 相异度是广泛使用的指标。Bray-Curtis 相异度量化了 2 个样本之间计数的组成差异[23]。Schmitt 等基于 Bray-Curtis 相异度对 OTU 进行聚类分析，以根据采样位置分析海洋微生物群落的特征[24]。UniFrac 相异度是基于系统发育距离的β多样性指标，Unifrac 越大，β多样性也越高[25]。

降维是从高维特征集中获取数据点，并将它们投影到低维空间中，在保持结构和最小化信息损失的同时，获得最大的统计方差[26]。常用的降维方法有主成分分析(principal component analysis，PCA)、主坐标分析(principal coordinates analysis，PCoA)、非度量多维尺度分析(nonmetric multidimensional scaling，NMDS)、冗余分析(redundancy analysis，RDA)和典范对应分析(canonical correspondence analysis，CCA)等。在微生物多样性研究当中，PCA、PCoA 和 NMDS 常用于组间样本比较，反映微生物群落结构的相似性和差异性。RDA 和 CCA 可以反映环境因子对微生物群落结构的影响以及物种和环境因子之间的相关性。

1.3.2 监督学习算法

根据输出的数据是离散(例如“患病”或者“健康”等)或者连续的(例如物种年龄、存活率和污染物浓度等)，可将监督学习分为 2 类：解决分类问题(classification)和解决回归问题(regression)。解决分类问题的算法可预测样本的类别(即定性)；解决回归问题的算法可预测真实的数值(即定量)。回归与分类之间没有明确界定，连续值可以阈值化或离散化，从而将回归问题转化为分类问题。目前，许多监督学习算法(如支持向量机、随机森林、朴素贝叶斯分类器、人工神经网络和深度学习等)已经广泛用于微生物组研究。下面将会分别介绍这些算法。

(1) 支持向量机(SVM)

支持向量机(support vector machine，SVM)是一种常用的二分类算法。其原理是判断是否存在一条直线可以将样本分为 2 类(在二维特征空间中)，即线性可分问题(图 2B)。对于线性不可分问题，运用 SVM 的核函数(kernel function)[27−28]寻找一个最优的超平面可以分类不同的数据样本，从而将线性不可分转化为线性可分问题。从统计学理论来看，SVM 通过最大化边界导出的决策函数可以降低预期风险的理论上限，因此具有泛化能力[29]。另外，SVM 可基于小样本构建决策函数，通常作为一种基础算法在微生物组研究当中广泛使用，例如

SVM 模型分类粪便微生物组[30]、SVM 模型预测蛋白质序列[31–33]。在预测土壤健康的研究中[34]，研究人员发现 SVM 在分类土壤健康类别方面优于随机森林，而随机森林在基于回归分析的健康评级预测方面优于 SVM。

(2) 朴素贝叶斯分类器(NBC)

朴素贝叶斯分类器(naïve bayes classifiers，NBC)是在概率统计的基础上对样本数据集进行准确分类的算法。对标记基因进行物种分类是微生物组学研究的重要步骤，Werner 等[35]的研究表明仅在特定的目标序列上训练的 NBC 具有较准确的分类结果，因此基于 NBC 的 Quantitative Insights Into Microbial Ecology version 2 (QIIME2) 已被广泛应用于细菌 16S rRNA 和真菌 ITS 标记基因扩增子数据的分类[36]。

(3) 随机森林

随机森林(random forest，RF)是由多个决策树组成的森林。随机森林中的每棵决策树都可以看作是一个分类器或者回归器，不同决策树之间没有关联，在所有的决策树输出相应的结果之后，投票选择出最合适的结果作为随机森林模型的最终预测(图 2C)。因为随机森林的随机性，使其拥有不容易出现过拟合，并且对缺省值不敏感的优势。在微生物群落的研究当中，因为数据维度高且特征数量有限，运用其他监督学习算法容易出现过拟合，因此随机森林方法更受欢迎，且被广泛应用(表 3)。随机森林还提供了一种衡量单个变量预测能力的参数，称为变量重要性(VIMP)，由此可以选择出关键特征已进行进一步调查。

(4) 人工神经网络

人工神经网络(artificial neural network，ANN)的基本思想是仿生学，它是一种模拟生物神经元运作机制的算法[37]。生物神经元的组成包括树突、细胞核、轴突、突触，外部的信号由树突进入神经元并传递给细胞核，经细胞核加工后作为输出传递给轴突。根据这一原理，科学家们提出 M-P 模型(图 2D)作为人工神经网络的一个神经元。它是由一个乘以权重的数值输入向量(树突)、一个激活函数(细胞核)和一个产生输出的函数(轴突)组成。模型的训练过程是一个不断学习权重的过程，实现以最小的错误率获得最佳的结果。为了能够解决复杂的问题、提高预测能力，神经网络模型当中会添加多个隐含层。通过增加或减少隐含层的节点数量，可以扩大或缩小数据维度(图 2E)。因此，无需对大数据进行预处理、特征选择或降维，就可以进行分类分析[38]。人工神经网络被广泛应用于蛋白质结构预测，功能预测[39–40]和蛋白质分类[41]。然而神经网络的每个节点的功能难以解释和验证，通常被称为“黑匣子”[2]。

1.3.3 生物信息学数据分析的深度学习算法

伴随高性能计算和大数据技术的发展，在传统机器学习(上文所总结的机器学习算法)的基础上又产生了可同时涵盖监督学习算法和无监督学习算法的深度学习(deep learning，DL)。深度学习的概念源于人工神经网络，深度学习的结构采用了与人工神经网络相似的分层结构，包括输入层、多个隐含层、输出层[42]。深度学习也分为监督学习和无监督学习，例如卷积神经网络(convolutional neural networks，CNNs)[43]是一种监督学习下的深度学习模型，而深度置信网络(deep belief nets，DBNs)[44]是一种无监督学习下的深度学习模型。深度学习在医学领域应用广泛，它能够很好地处理微生物基因组序列数据，自动确定关键特征(features)用于构建模型，从而对疾病进行准确的预测和诊断等[45]。例如基于微生物数据，卷积神经网络可以用于对个体的疾病状态进行分类[46]。更多深度学习在生物信息领域的应用如表 1 所示。

表 1　深度学习在生物信息学分析中的应用

Table 1　Applications of deep learning in bioinformatics

Application	Datasets	Data types	Algorithms/models	References
Sequence analysis	Sequence data (DNA sequence, RNA sequence, etc.)	1D Data	CNN, RNN	[47–51,32]
Protein and gene function prediction	Sequence data, structure traits, microarray gene expression	1D data, 2D data, structure data	DNN, CNN, RNN	[52–55]
Biomolecular interaction and correlation prediction	Microarray gene expression, gene-disease associations, Disease similarity network	1D data, 2D data, structure data, graph data	GCN, CNN	[56–61]

相较于传统的机器学习依赖于人为对输入特征进行选择，深度学习的特征提取在某些情况下是由模型自身完成的，即深度学习可以通过连续地提取数据抽象来改变原始的输入特征。另外，深度学习可以通过添加更多隐含层，拥有高复杂度的函数，输入大量的已做标记的样本数据集，来获得更高的预测性能。换句话说，随着数据量的增加，深度学习可能会超越其他机器学习算法(图 3)。然而，对特征重要性或模型预测的置信度的估算并不容易[62–63]，这两者在生物学研究中是必不可少的。另外，构建深度神经网络和对其进行训练是一件耗时且计算成本高的任务[37]。因此，在基于生物学数据的预测时(例如，利用转录组学数据预测表型[64])，谨慎的做法通常是先训练传统机器学习模型，再将其与深度学习模型进行比较。

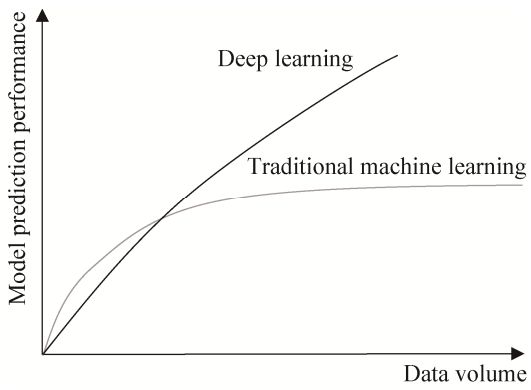


图 3　深度学习和传统机器学习模型预测性能比较[38]

Figure 3　Comparison of the prediction performance between deep learning and traditional machine learning[38]

1.4　机器学习模型评估方法

目前已有很多方法可用于评估机器学习模型的性能。对于回归模型，评估参数主要包括偏差(bias)、方差(variance)、平均绝对误差 (MAE)和均方误差(MSE)和 R^2 等。对于分类模型，精度(accurarcy)、准确率(precision，P)、召回率(recall，R)、受试者工作特征(receive operationg characteristic，ROC)和 ROC 曲线下面积(area under ROC curve，AUC)等是常用的评价指标。精度通常以百分数形式存在，越接近 100%认为模型进行分类的准确度越高。P 和 R 这 2 个评价指标一般同时使用，P 和 R 越接近 1 表明模型性能越佳。但是 P 和 R 是一对矛盾的度量，通常认为，P 高时，R 往往偏低，而 R 高时，P 往往偏低。ROC 描述模型的正确分类的正样本数占总正样本数的比例(true positive rate，TPR)与错误分类的正样本数占总负样本数的比例(false positive rate，FPR)之间的变化关系。理想状态下认为 TPR 为 100%，FPR 为 0%的模型分类效果最佳，但实际情况很难达到这个效果。AUC 可以解决 ROC 难以比较多个模型性能的问题，AUC 的取值在 0.5 到 1 之间，AUC 越大，模型的分类效果越好。Statnikov 等[65]基于微生物组数据对 18 种机器学习分类模型、5 种特征选择方法和

2 个精度指标进行了综合评估。结果表明，在大多数情况下支持向量机、随机森林、岭回归和贝叶斯逻辑回归是执行准确分类最有效的模型，且它们的分类精度在统计学上相似。

2 机器学习与传统统计模型在环境微生物研究中的比较

传统统计模型与机器学习在本质上都是统计数据、建立模型，对模型参数进行推断，以及评价模型拟合或预测效果。两者之间的区别在于传统统计模型更注重模型的拟合置信度，而机器学习更关注模型的预测效果。换句话说，前者的目标是描述和推断变量之间的关系，而后者旨在优化对外部数据集的预测性能。例如，监督学习使用训练集(有标记)来设计模型，并使用测试集(隐藏标记)来判断模型的性能。另一方面，传统统计模型主要关注值与结果之间的关联，并且大多数研究不需要对数据进行分区来衡量性能[2]。

另外，传统统计模型大多需要对数据进行假设，模型本身有明确的数学含义。对模型的优劣评价，需要对数据分布的假设进行检验来判断。然而，人们很难对于真实世界的数据分布进行假设，且复杂的数据也很难用有限的数学公式描述。机器学习建立在统计学的基础上，它对数据没有任何假定，摆脱了假设分布、明确数学模型、假设检验等过程，更关注于如何提高预测效果，并且通过预测的性能评估模型优劣[66]。随着测序技术发展，数据量的不断增多，机器学习更加适用于微生物组学研究。

相比于传统统计学，机器学习能够挖掘微生物组数据中的复杂关联，更有效地识别微生物群落结构的细微变化，并且能够识别出特定的细菌类群[2]。另外，机器学习还可以对细菌计数数据和环境参数之间的非线性关系建模，不需要假设复杂的转化过程和预处理[2]。然而，由于机器学习可以在没有明确的用户指令的情况下运行，具有高度可配置性，并且需要大量数据，因此可能会出现数据的过拟合。另外，有些机器学习算法对于预测结果具有可解释性(如随机森林)，而有些仅提供模糊的准确度统计数据，难以解释模型如何实现预测(如人工神经网络、深度学习)。因此认为没有最适用于某种机器学习算法的情景，应根据所要解决的问题、可用的数据去选择分析方法。如果研究的目标是基于微生物群落数据建立对结果的预测性理解，那么监督学习算法更加适用，因为这些算法是为提高预测准确性而定制的。然而，如果研究的目标是将特定微生物群与特定结果联系起来，则可以使用经典统计模型。虽然统计模型在微生物群落分析中有其优势，但是机器学习更适于分析从真实环境中(如土壤和水[67−68])采样得到的复杂数据、挖掘其中的复杂关联，并且在分析宏基因组数据方面优于传统的多变量统计。此外，为方便微生物组的研究，用户友好的机器学习操作途径也被不断地开发，例如可以通过网页、R[69]和 Python[70]等方式完成机器学习。相关的机器学习工具如表 2 所示。

3 机器学习在环境微生物领域的应用

3.1 机器学习识别环境指示生物

机器学习可以将环境基因组与环境质量相关联，同时识别出对生态环境质量有指示作用的生物，因此可被监管机构用于指导环境监测[79]。以鲑鱼养殖为例，大量研究运用监督学习识别出可以预测生物指数和环境质量的真核微生物和细菌群落。Cordier 等[80]使用细菌、纤毛虫和常见的真核生物的核糖体小亚基上的不同标记基因来训练模型，进而比较这些模型对

表 2　可用于微生物组预测的机器学习工具包

Table 2　Summary of machine learning tools used for microbiome research.

Tools	Application	URL	References
Mian	Training machine learning models to explore relevant relationships among microbiome data	https://miandata.org	[71]
SIAMCAT	Inferring associations between microbial communities and host phenotypes	R package: https://siamcat.embl. de/	[72]
MetAML	Quantitative assessment of microbiome-phenotype associations	Python: https://github.com/segatalab/metaml	[73]
mAML	A online machine learning system based on microbiome for disease classification	http://lab.malab.cn/soft/mAML/	[74]
Meta-Signer	Identifying feature classes in the model	R package: https://github.com/YDaiLab/Meta-Signer/tree/master/src	[75]
Scikit-learn	Machine learning in Python used for medium-scale supervised and unsupervised learning	Python: https://scikit-learn.org/stable/	[76]
Keras	Deep learning in Python	Python: https://keras.io	[77]
mlr	Machine learning in R	R package: https://mlr3.mlr-org.com	[78]

环境质量的预测性能。结果表明，这些模型均有较准确的预测能力，并且细菌相较于纤毛虫和真核生物能够更好地预测鲑鱼养殖场的环境质量。在这项研究之后，Frühe 等[81]将随机森林模型的预测性能与用于评价环境质量的IndVal方法进行比较。IndVal方法可以从环境DNA 中判断生态质量，且无需知道确切的物种信息，因此不需要对 OTU/ASV 进行物种注释。此次研究发现监督学习在预测环境质量方面优于 IndVal，表明了机器学习在生物监测中的效用。然而，这些研究仅用同一实验室生成的数据来训练和验证模型，因此模型不具有泛化性。为了使机器学习能够更好地实现环境监测，需要收集更多数据，提高训练数据量，从而提高模型的泛化能力。Alneberg 等[82]还运用机器学习来预测微生物群落的生态位。该研究对跨越波罗的海主要环境梯度的 123 个水样进行宏基因组分箱，由此产生了能涵盖该海域大部分原核生物的 1961 个宏基因组组装基因组(MAGs)，并运用多种机器学习算法(如岭回归算法、随机森林算法和梯度提升算法)，根据功能基因预测这些原核生物基因簇的生态位梯度。结果表明预测的生态位梯度和观察到的生态位梯度一致(spearman 等级相关系数为0.70−0.81)。该研究证明了基因组和生态位之间的紧密联系，且机器学习可以识别自然微生物群落的分布模式和预测生物体的生态位。Sun 等[83]的研究表明机器学习可以识别潜在的抗性基因宿主。该研究利用宏基因组数据构建随机森林模型，成功预测污水处理厂活性污泥中的抗性基因丰度，并且发现能够与一些抗性基因显著正相关的细菌和病原体(在属水平上)，例如 *Bacteroides*、*Clostridium* 和 *Streptococcus* 等，它们可以作为这些抗性基因的潜在宿主。

3.2　根据微生物群落识别环境污染

机器学习对环境污染物的预测是环境微生物领域的研究热点。通常通过直接测量某种污染物来判断该环境是否受到污染。然而，有些污染物在环境中可能是暂时存在，因此在取样时可能很难监测到它们。已有研究基于微生物数据构建随机森林模型来预测地下水样品是否被铀和硝酸盐污染[84]，该研究还表明随机森林

模型可以精准预测海洋是否受到石油污染(F1值高达 0.98)。值得注意的是，该模型还可以根据微生物群落识别出过去被石油污染的样品。过去被石油污染的样品是指在过去某个时间点检测到石油含量，但在取样时没有检测到石油。这些结果表明，即使人类活动产生的污染物已经耗尽，但污染的“痕迹”仍然存在，随机森林模型具有识别这些“痕迹”的能力。Janßen 等[85]运用人工神经网络和随机森林来预测波罗的海是否受到除草剂草甘膦的污染，并对比2种算法的预测性能。结果表明，2 种算法作为处理相同信息的不同方法，均具有较高的预测性能，且很难一分高下。可以根据提出的假设和数据的特征来选择使用哪种算法。根据经验，作者建议先使用随机森林算法，因为该算法是现成的且模型能够快速提供结果。当数据量扩大，数据集变得复杂，可以选择使用人工神经网络算法，该算法能够执行超参数优化。另外，该研究的另一个有趣的发现是使用随机森林邻近矩阵与使用 Bray-Curtis 相异矩阵相比，前者在 PCoA 分析中能够更清晰地将不同簇的样品分开。

3.3 根据微生物群落评价环境质量

环境中的微生物可以对其所在环境的变化做出响应，因此机器学习可以根据微生物群落的变化去预测环境变化，从而达到环境监测的目的[86]。Smith 等[84]的研究展示了由 16S rRNA 确定的微生物群落组成可用于预测环境因子和多种元素的地球化学特征，包括 pH 值、锰、镁、钾等。Hermans 等[87]的研究表明随机森林模型可以利用土壤细菌群落预测土壤理化指标(如 pH、养分浓度、体积密度等)和土壤质量。机器学习还可以利用微生物群落数据预测环境现象。例如，利用宏基因组数据构建随机森林模型，将土壤中存在的微生物群落与农作物的生产力联系起来[88]，从而根据微生物群落的变化来确定农作物的生产力。该模型对与农作物生产力的预测准确度可以达到 79%。还有研究利用随机森林和人工神经网络算法分别构建模型，试图将溶解有机碳(DOC)与微生物群落组成联系起来[89]，以根据微生物群落组成预测落叶的 DOC 浓度。通过对比随机森林模型和人工神经网络模型的预测性能，发现它们的预测准确度都很高，对于预测的 DOC 与观察到的 DOC 之间的 Pearson 相关系数分别为 0.636 和 0.676。另外，这项研究还将 2 个模型确定的重要物种与指示种分析中确定的指示种进行了比较，发现 2 种模型中识别的特征有一定程度的重叠，但只有大约 30%的特征与指示种分析的结果一致。这表明机器学习通常使用不同的特征进行分类，而不使用通过差异丰度或指示种分析得到特征，因此能够更敏感地从复杂的微生物群落中识别重要的特征子集。机器学习在环境微生物领域的应用见表 3。

4 总结与展望

微生物组大数据蕴含着大量信息，适用于分析高纬度复杂数据的机器学习技术开始被广泛应用。本综述旨在概述环境微生物领域常用的机器学习算法，运用不同的机器学习算法从微生物大数据中总结规律，创建模型，以实现对新数据的识别、分类和预测。大量的研究结果已经证明了机器学习在这个领域的贡献，尤其是对于物种分类、微生物标志物识别、污染物预测和环境质量预测等方面。

然而，机器学习在微生物组学数据的分析中仍有很大的发展空间。(1) 数据方面：环境微生物数据表示形式单一且存在大量冗余信息，无法为机器学习提供足够的数据信息，导致模型缺乏泛化能力。对于规模较小或者样本难以

表 3　机器学习在环境微生物领域的应用

Table 3　Studies using machine learning in microbial ecology

Research object	Prediction	Data types	Number of samples	Machine learning algorithms	Training and validation	References
Microbiome in wastewater treatment plants	Antibiotic resistance gene abundance	Shotgun metagenome	248	RF	60% samples as training set, 40% samples as validation set, five-fold cross validation	[83]
Soil microbiome	Crop productivity	Shotgun metagenome	12	RF	10 samples as training set, 2 samples as validation set	[88]
Soil microbiome	Dissolved organic carbon concentration	16S rRNA	302	ANN and RF	257 samples as training set, 51 samples as test set.	[89]
Soil microbiome	Soil quality	16S rRNA	3 030	RF	80% samples for training, 20% samples for validation	[87]
Brackish microbiome	Detecting glyphosate contamination	16S rRNA	32	ANN and RF	Leave-one-out cross validation	[85]
Marine microbiome	Marine benthic ecological quality	SSU RNA	144	RF	Four samples from the salmon farms as training set, others as test set	[90]
Marine microbiome	Bioindicator	SSU RNA	152	RF and SVM	Six samples from the salmon farms as training set, the seventh sample used to test	[80]
Different ecological categories (including human waste, animal waste, activated sludge from wastewater treatment plants, natural environment, etc.)	Contribution of antibiotic resistance gene from different sources	Shotgun metagenome	656	RF	Leave-one-out cross validation	[91]
Groundwater microbiome		16S rRNA	93 groundwater samples and 43 oil polluted samples	RF	Model performances were determined from a confusion matrix	[84]
Harmful algae	Marine water quality	Water quality (total inorganic nitrogen, phosphorus, chlorophyll a, dissolved oxygen, water temperature, etc.)	11 293	ANN and SVM	9 000 samples from 1988 to 2012 as training set, others as validation set, five-fold cross validation	[92]

收集的研究(采样点少)，使用结构复杂的机器学习算法容易出现过拟合现象，而使用结构简单的机器学习算法可能无法充分挖掘数据的特征，且对无用特征的过滤能力低。因此在使用机器学习算法前可以先进行特征选择，目前已有许多研究开始关注独立于机器学习的特征选择在微生物学领域的应用[21,93–94]。(2) 算法方面：目前常用的机器学习算法适用于许多领域，尤其在计算机视觉、自然语言处理和图像语音识别等“计算机应用技术”领域。对于微生物组学等交叉学科领域，机器学习仅停留于应用层面，对算法本身的研究不足，且缺乏特异性优化。因此，未来对适用于微生物领域的特异性机器学习算法研发值得关注。(3) “黑匣子”问题：尽管机器学习模型可以在复杂数据中生成高精度指标，但是决策系统通常是“黑匣子”，尤其对于深度学习，解释其预测的原理和逻辑具有挑战性。对于理解“黑匣子”的决策行为，可以尝试关注输入数据对模型输出结果的影响，通过每个样本的输入-输出关系来间接理解模型是如何实现预测的。(4) 模型方面：大多数运用微生物数据构建机器学习模型的研究倾向于应用特征重要性来识别重要的微生物类群。然而，由于模型的复杂性，从特征重要性推断生物学重要性可能存在一定局限[95]。因此还需要通过模拟实验或者收集更多的数据等方法来验证机器学习的结果和提高模型的准确性。

综上所述，机器学习在环境微生物领域已展示了良好的前景。然而能否正确使用机器学习技术至关重要。虽然机器学习功能强大，但并非适用于任何情况，不正确的使用将会带来不准确的分析结果和错误的结论。在使用机器学习前应对现有数据充分认识、明确分析目的和待解决的问题、深入理解各个机器学习算法的功能和差异，再选择合适的算法对数据进行分析。为了更好地应用各种机器学习算法，还需要对算法本身开展更深层次的研究，优化算法的特异性，从而提高模型的可解释性，并且收集更全面的环境微生物数据，以提高模型的泛化能力。

参考文献

[1] Rampelotto PH. Extremophiles and extreme environments. *Life: Basel, Switzerland*, 2013, 3(3): 482–485.

[2] Ghannam RB, Techtmann SM. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. *Computational and Structural Biotechnology Journal*, 2021, 19: 1092–1107.

[3] Qu KY, Guo F, Liu XR, Lin Y, Zou Q. Application of machine learning in microbiology. *Frontiers in Microbiology*, 2019, 10: 827.

[4] DiMucci D, Kon M, Segrè D. Machine learning reveals missing edges and putative interaction mechanisms in microbial ecosystem networks. *mSystems*, 2018, 3(5): e00181–e00118.

[5] Xie KB, Guo L, Bai Y, Liu WD, Yan JB, Bucher M. Microbiomics and plant health: an interdisciplinary and international workshop on the plant microbiome. *Molecular Plant*, 2019, 12(1): 1–3.

[6] Moitinho-Silva L, Steinert G, Nielsen S, Hardoim CCP, Wu YC, McCormack GP, López-Legentil S, Marchant R, Webster N, Thomas T, Hentschel U. Predicting the HMA-LMA status in marine sponges by machine learning. *Frontiers in Microbiology*, 2017, 8: 752.

[7] Di Bella JM, Bao YG, Gloor GB, Burton JP, Reid G. High throughput sequencing methods and analysis for microbiome research. *Journal of Microbiological Methods*, 2013, 95(3): 401–414.

[8] Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai CX, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE. Big data: astronomical or genomical? *PLoS Biology*, 2015, 13(7): e1002195.

[9] Zhou YH, Gallins P. A review and tutorial of machine learning methods for microbiome host trait prediction. *Frontiers in Genetics*, 2019, 10: 579.

[10] Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D,

Knight R, Gordon JI. Human gut microbiome viewed across age and geography. *Nature*, 2012, 486(7402): 222–227.

[11] Tarca AL, Carey VJ, Chen XW, Romero R, Drăghici S. Machine learning and its applications to biology. *PLoS Computational Biology*, 2007, 3(6): e116.

[12] Jasner Y, Belogolovski A, Ben-Itzhak M, Koren O, Louzoun Y. Microbiome preprocessing machine learning pipeline. *Frontiers in Immunology*, 2021, 12: 677870.

[13] Krause T, Wassan JT, Mc Kevitt P, Wang HY, Zheng HR, Hemmje M. Analyzing large microbiome datasets using machine learning and big data. *BioMedInformatics*, 2021, 1(3): 138–165.

[14] Richardson A, Signor BM, Lidbury BA, Badrick T. Clinical chemistry in higher dimensions: machine-learning and enhanced prediction from routine clinical chemistry data. *Clinical Biochemistry*, 2016, 49(16/17): 1213–1220.

[15] Wildenhain J, Spitzer M, Dolma S, Jarvik N, White R, Roy M, Griffiths E, Bellows DS, Wright GD, Tyers M. Prediction of synergism from chemical-genetic interactions by machine learning. *Cell Systems*, 2015, 1(6): 383–395.

[16] Cammarota G, Ianiro G, Ahern A, Carbone C, Temko A, Claesson MJ, Gasbarrini A, Tortora G. Gut microbiome, big data and machine learning to promote precision medicine for cancer. *Nature Reviews Gastroenterology & Hepatology*, 2020, 17(10): 635–648.

[17] Zhou C, Lin K, Xu DM, Chen L, Guo Q, Sun CH, Yang XT. Near infrared computer vision and neuro-fuzzy model-based feeding decision system for fish in aquaculture. *Computers and Electronics in Agriculture*, 2018, 146: 114–124.

[18] López-Cortés XA, Nachtigall FM, Olate VR, Araya M, Oyanedel S, Diaz V, Jakob E, Ríos-Momberg M, Santos LS. Fast detection of pathogens in salmon farming industry. *Aquaculture*, 2017, 470: 17–24.

[19] Fang K, Shen CP, Kifer D, Yang X. Prolongation of SMAP to spatiotemporally seamless coverage of continental US using a deep learning neural network. *Geophysical Research Letters*, 2017, 44(21): 11030–11039.

[20] Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D. Machine learning in agriculture: a review. *Sensors: Basel, Switzerland*, 2018, 18(8): E2674.

[21] Oudah M, Henschel A. Taxonomy-aware feature engineering for microbiome classification. *BMC Bioinformatics*, 2018, 19(1): 227.

[22] Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Computational Biology*, 2016, 12(7): e1004977.

[23] Bray JR, Curtis JT. An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, 1957, 27(4): 325–349.

[24] Schmitt S, Tsai P, Bell J, Fromont J, Ilan M, Lindquist N, Perez T, Rodrigo A, Schupp PJ, Vacelet J, Webster N, Hentschel U, Taylor MW. Assessing the complex sponge microbiota: core, variable and species-specific bacterial communities in marine sponges. *The ISME Journal*, 2012, 6(3): 564–576.

[25] Lozupone CA, Hamady M, Kelley ST, Knight R. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology*, 2007, 73(5): 1576–1585.

[26] Silva V, Tenenbaum J. Global versus local methods in nonlinear dimensionality reduction. *Advances in Neural Information Processing Systems*, 2002, 15: 721–728.

[27] Ben-Hur A, Weston J. A user's guide to support vector machines. *Data Mining Techniques for the Life Sciences*, 2009, 609: 223–239.

[28] Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G. Support vector machines and kernels for computational biology. *PLoS Computational Biology*, 2008, 4(10): e1000173.

[29] Vapnik VN. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 1999, 10(5): 988–999.

[30] Kung HC, Chen RM, Tsai JJP, Hu RM. Stratification of human gut microiome and building a SVM-based classifier. 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering. Taichung, Taiwan, China. Piscataway, NJ: IEEE, 14–17.

[31] Xu L, Liang GM, Liao CR, Chen GD, Chang CC. An efficient classifier for Alzheimer's disease genes identification. *Molecules: Basel, Switzerland*, 2018, 23(12): 3140.

[32] Xu L, Liang GM, Wang LJ, Liao CR. A novel hybrid sequence-based model for identifying anticancer peptides. *Genes*, 2018, 9(3): 158.

[33] Xu L, Liang GM, Shi SH, Liao CR. SeqSVM: a sequence-based support vector machine method for identifying antioxidant proteins. *International Journal of Molecular Sciences*, 2018, 19(6): 1773.

[34] Wilhelm RC, van Es HM, Buckley DH. Predicting

measures of soil health using the microbiome and supervised machine learning. *Soil Biology and Biochemistry*, 2022, 164: 108472.

[35] Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, Caporaso JG, Angenent LT, Knight R, Ley RE. Impact of training sets on classification of high-throughput bacterial 16S rRNA gene surveys. *The ISME Journal*, 2012, 6(1): 94–103.

[36] Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Gregory Caporaso J. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*, 2018, 6(1): 90.

[37] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, 521(7553): 436–444.

[38] Namkung J. Machine learning methods for microbiome studies. *Journal of Microbiology: Seoul, Korea*, 2020, 58(3): 206–216.

[39] Mathkour H, Ahmad M. An integrated approach for protein structure prediction using artificial neural network. 2010 Second International Conference on Computer Engineering and Applications. Bali, Indonesia. Piscataway, NJ: IEEE, 484–488.

[40] Hirst JD, Sternberg MJ. Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural networks. *Biochemistry*, 1992, 31(32): 7211–7218.

[41] Rossi ALD, De Oliveira Camargo-Brunetto MA. Protein classification using artificial neural networks with different protein encoding methods. Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007). Rio de Janeiro, Brazil. Piscataway, NJ: IEEE, 169–176.

[42] Min S, Lee B, Yoon S. Deep learning in bioinformatics. *Briefings in Bioinformatics*, 2016, 18(5): 851–869.

[43] Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 2019, 20(7): 389–403.

[44] Norouzi M, Ranjbar M, Mori G. Stacks of convolutional Restricted Boltzmann Machines for shift-invariant feature learning. 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA. Piscataway, NJ: IEEE, 2735–2742.

[45] Li Y, Huang C, Ding LZ, Li ZX, Pan YJ, Gao X. Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. *Methods*, 2019, 166: 4–21.

[46] Lo C, Marculescu R. MetaNN: accurate classification of host phenotypes from metagenomic data using neural networks. *BMC Bioinformatics*, 2019, 20(Suppl 12): 314.

[47] Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biology*, 2017, 18(1): 67.

[48] Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 2015, 33(8): 831–838.

[49] Umarov R, Kuwahara H, Li Y, Gao X, Solovyev V. Promoter analysis and prediction in the human genome using sequence-based deep learning models. *Bioinformatics*, 2019, 35(16): 2730–2737.

[50] Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 2015, 12(10): 931–934.

[51] Li Y, Han RM, Bi CW, Li M, Wang S, Gao X. DeepSimulator: a deep simulator for Nanopore sequencing. *Bioinformatics*, 2018, 34(17): 2899–2908.

[52] Kulmanov M, Khan MA, Hoehndorf R, Wren J. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics: Oxford, England*, 2018, 34(4): 660–668.

[53] Zou ZZ, Tian SY, Gao X, Li Y. mlDEEPre: multi-functional enzyme function prediction with hierarchical multi-label deep learning. *Frontiers in Genetics*, 2019, 9: 714.

[54] Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 2017, 33(21): 3387–3395.

[55] Li Y, Wang S, Umarov R, Xie BQ, Fan M, Li LH, Gao X. DEEPre: sequence-based enzyme EC number prediction by deep learning. *Bioinformatics*, 2017, 34(5): 760–769.

[56] Ma JZ, Yu MK, Fong S, Ono K, Sage E, Demchak B, Sharan R, Ideker T. Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods*, 2018, 15(4): 290–298.

[57] Zeng HY, Edwards MD, Liu G, Gifford DK. Convolutional neural network architectures for predicting DNA-protein binding. *Bioinformatics*, 2016, 32(12): i121–i127.

[58] Zong NS, Kim H, Ngo V, Harismendy O. Deep mining heterogeneous networks of biomedical linked data to predict novel drug-target associations. *Bioinformatics*, 2017, 33(15): 2337–2344.

[59] Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 2018, 34(13): i457–i466.

[60] Kordopati V, Salhi A, Razali R, Radovanovic A, Tifratene F, Uludag M, Li Y, Bokhari A, AlSaieedi A, Bin Raies A, Van Neste C, Essack M, Bajic VB. DES-mutation: system for exploring links of mutations and diseases. *Scientific Reports*, 2018, 8: 13359.

[61] Wan FP, Hong LX, Xiao A, Jiang T, Zeng JY. NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions. *Bioinformatics*, 2018, 35(1): 104–111.

[62] 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, Lavender CA, Turaga SC, Alexandari AM, Lu ZY, Harris DJ, DeCaprio D, Qi YJ, Kundaje A, Peng YF, Wiley LK, Segler MHS, Boca SM, Swamidass SJ, Huang A, Gitter A, Greene CS. Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society, Interface*, 2018, 15(141): 20170387.

[63] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. 2015: arXiv: 1506.02142[stat.ML].

[64] Smith AM, Walsh JR, Long J, Davis CB, Henstock P, Hodge MR, Maciejewski M, Mu XJ, Ra S, Zhao SR, Ziemek D, Fisher CK. Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC Bioinformatics*, 2020, 21(1): 119.

[65] Statnikov A, Henaff M, Narendra V, Konganti K, Li ZG, Yang LY, Pei ZH, Blaser MJ, Aliferis CF, Alekseyenko AV. A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome*, 2013, 1(1): 11.

[66] 吴喜之. 复杂数据统计方法: 基于 R 的应用. 北京: 中国人民大学出版社, 2012.

[67] Belk A, Xu ZZ, Carter DO, Lynne A, Bucheli S, Knight R, Metcalf JL. Microbiome data accurately predicts the postmortem interval using random forest regression models. *Genes*, 2018, 9(2): 104.

[68] Ghannam RB, Schaerer LG, Butler TM, Techtmann SM. Biogeographic patterns in members of globally distributed and dominant taxa found in port microbial communities. *mSphere*, 2020, 5(1): e00481–e00419.

[69] Team R. R: a language and environment for statistical computing. 2014

[70] Van RG, Drake FL. Python tutorial. New York: Iuniverse Inc, 2020.

[71] Jin BT, Xu F, Ng RT, Hogg JC. Mian: interactive web-based microbiome data table visualization and machine learning platform. *Bioinformatics*, 2021, 38(4): 1176–1178.

[72] Wirbel J, Zych K, Essex M, Karcher N, Kartal E, Salazar G, Bork P, Sunagawa S, Zeller G. Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. *Genome Biology*, 2021, 22(1): 93.

[73] Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLOS Computational Biology*, 2016, 12(7): e1004977.

[74] Yang FL, Zou Q. mAML: an automated machine learning pipeline with a microbiome repository for human disease classification. *Database*, 2020, 2020(10.1093): database.

[75] Reiman D, Metwally A, Sun J, Dai Y. Meta-signer: metagenomic signature identifier based on rank aggregation of features. *F1000Research, 2021,* 10(194): 194.

[76] 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, Duchesnay E. Scikit-learn: machine learning in Python. *The Journal of Machine Learning Research*, 2011, 12: 2825–2830.

[77] Gulli A, Pal S. Deep learning with Keras. Birmingham: Packt Publishing Ltd, 2017.

[78] Lang M, Binder M, Richter J, Schratz P, Pfisterer F, Coors S, Au Q, Casalicchio G, Kotthoff L, Bischl B. mlr3: a modern object-oriented machine learning framework in R. *Journal of Open Source Software*, 2019, 4(44): 1903.

[79] Cordier T, Lanzén A, Apothéloz-Perret-Gentil L, Stoeck T, Pawlowski J. Embracing environmental genomics and machine learning for routine biomonitoring. *Trends in Microbiology*, 2019, 27(5): 387–397.

[80] Cordier T, Forster D, Dufresne Y, Martins CIM, Stoeck T, Pawlowski J. Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Molecular Ecology Resources*, 2018, 18(6): 1381–1391.

[81] Frühe L, Cordier T, Dully V, Breiner HW, Lentendu G, Pawlowski J, Martins C, Wilding TA, Stoeck T, et al. Supervised machine learning is superior to indicator

value inference in monitoring the environmental impacts of salmon aquaculture using eDNA metabarcodes. Molecular Ecology, 2021, 30(13): 2988–3006.

[82] Alneberg J, Bennke C, Beier S, Bunse C, Quince C, Ininbergs K, Riemann L, Ekman M, Jürgens K, Labrenz M, Pinhassi J, Andersson AF. Ecosystem-wide metagenomic binning enables prediction of ecological niches from genomes. *Communications Biology*, 2020, 3: 119.

[83] Sun YP, Clarke B, Clarke J, Li X. Predicting antibiotic resistance gene abundance in activated sludge using shotgun metagenomics and machine learning. *Water Research*, 2021, 202: 117384.

[84] Smith MB, Rocha AM, Smillie CS, Olesen SW, Paradis C, Wu LY, Campbell JH, Fortney JL, Mehlhorn TL, Lowe KA, Earles JE, Phillips J, Techtmann SM, Joyner DC, Elias DA, Bailey KL, Hurt RA Jr, Preheim SP, Sanders MC, Yang J, Mueller MA, Brooks S, Watson DB, Zhang P, He ZL, Dubinsky EA, Adams PD, Arkin AP, Fields MW, Zhou JZ, Alm EJ, Hazen TC. Natural bacterial communities serve as quantitative geochemical biosensors. *mBio*, 2015, 6(3): e00326–15.

[85] Janßen R, Zabel J, Von Lukas U, Labrenz M. An artificial neural network and Random Forest identify glyphosate-impacted brackish communities based on 16S rRNA amplicon MiSeq read counts. *Marine Pollution Bulletin*, 2019, 149: 110530.

[86] Techtmann SM, Hazen TC. Metagenomic applications in environmental monitoring and bioremediation. *Journal of Industrial Microbiology and Biotechnology*, 2016, 43(10): 1345–1354.

[87] Hermans SM, Buckley HL, Case BS, Curran-Cournane F, Taylor M, Lear G. Using soil bacterial communities to predict physico-chemical variables and soil quality. *Microbiome*, 2020, 8(1): 79.

[88] Chang HX, Haudenshield JS, Bowen CR, Hartman GL. Metagenome-wide association study and machine learning prediction of bulk soil microbiome and crop productivity. *Frontiers in Microbiology*, 2017, 8: 519.

[89] Thompson J, Johansen R, Dunbar J, Munsky B. Machine learning to predict microbial community functions: an analysis of dissolved organic carbon from litter decomposition. *PLoS One*, 2019, 14(7): e0215502.

[90] Cordier T. Bacterial communities' taxonomic and functional turnovers both accurately predict marine benthic ecological quality status. *Environmental DNA*, 2020, 2(2): 175–183.

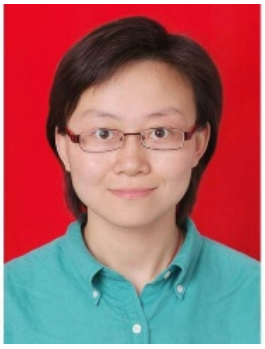
[91] Li LG, Yin XL, Zhang T. Tracking antibiotic resistance gene pollution from different sources using machine-learning classification. *Microbiome*, 2018, 6(1): 93.

[92] Deng TA, Chau KW, Duan HF. Machine learning based marine water quality prediction for coastal hydro-environment management. *Journal of Environmental Management*, 2021, 284: 112051.

[93] Dhindsa A, Bhatia S, Agrawal S, Sohi BS. An improvised machine learning model based on mutual information feature selection approach for microbes classification. *Entropy: Basel, Switzerland*, 2021, 23(2): 257.

[94] Jiang L, Haiminen N, Carrieri AP, Huang S, Vázquez-Baeza Y, Parida L, Kim HC, Swafford AD, Knight R, Natarajan L. Utilizing stability criteria in choosing feature selection methods yields reproducible results in microbiome data. *Biometrics*, 2021. DOI: https://doi.org/10.1111/biom.13481.

[95] Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: a survey on methods and metrics. *Electronics*, 2019, 8(8): 832.



邢鹏，中国科学院南京地理与湖泊研究所，研究员。研究方向为水域微生物生态学，重点开展湖泊微生物驱动的元素循环过程和机制研究。先后主持国家基金委 NSFC-云南联合基金项目、优秀青年科学基金项目、科技部基础资源调查专项课题、国家自然科学基金面上项目、基金委重大研究计划培育项目等。在国际期刊上发表论文 60 余篇，他引 1 500 余次。2018 年获得中国科学院青年创新促进会优秀会员，入选江苏省“333 工程”中青年科学技术带头人。