ACCSCIENCE PUBLISHING

Microbes & Immunity

PERSPECTIVE ARTICLE

Using artificial intelligence to study plant–microbiome interactions for sustainable agriculture

Gaurav Sablok*

St. Joseph-Krankenhaus, Alexianer, Berlin, Germany

Abstract

Plant microbiome research is a rapidly developing and productive field, with substantial attention directed toward the identification of the major determinants of plant growth, nutrient acquisition, stress tolerance, disease resistance, and associated agricultural productivity. Advances in high-throughput sequencing and multi-omics technologies have generated large and complex datasets describing plant-associated microbial communities. However, the high dimensionality, sparsity, compositionality, and heterogeneity of microbiome datasets create substantial analytical challenges for conventional statistical approaches. Machine learning and artificial intelligence (AI) provide new opportunities to identify microbial biomarkers, predict plant phenotypes, characterize plant–microbe interactions, and design beneficial microbial communities. Applications include plant disease prediction, crop productivity, abiotic-stress tolerance, nutrient cycling, biomarker discovery, microbiome-assisted breeding, and synthetic microbial community (SynCom) design. Recent studies increasingly combine AI with multi-omics, synthetic biology, and predictive modeling, indicating a transition from descriptive microbiome analysis toward microbiome engineering. Nevertheless, model generalizability remains constrained by small sample sizes, compositional data, technical variation, batch effects, environmental heterogeneity, and insufficient external validation. The combination of microbiome science and AI has considerable potential to support sustainable, climate-resilient crop production, but successful translation will require rigorous biological validation and field-scale testing.

Keywords: Plant microbiome; Rhizosphere; Machine learning; Artificial intelligence; Microbiome engineering; Multi-omics; Synthetic microbial communities; Crop improvement

***Corresponding author:**
Gaurav Sablok
(gsablok@proton.me)

Citation: Sablok G.Using artificial intelligence to study plant–microbiome interactions for sustainable agriculture. *Microbes & Immunity*.
doi: 10.36922/MI026380111

Received: September 17, 2026

Revised: October 8, 2026

Accepted: October 10, 2026

Published online: October 19, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Plants and Microbes

Plants form associations with microbial communities inhabiting the rhizosphere, phyllosphere, endosphere, and surrounding soil (Table 1). These microbial communities influence plant nutrition, development, immune responses, pathogen resistance, and tolerance to environmental stresses. The rhizosphere is particularly important because root exudates influence microbial recruitment and activity, while microorganisms reciprocally modify nutrient availability, hormone signaling, and plant stress responses

(Table 2). Consequently, the plant and its associated microbiota can be considered an interconnected biological system in which microbial functions contribute substantially to plant performance.[1,2]

Recent advances in amplicon sequencing, shotgun metagenomics, metatranscriptomics, metabolomics, and other omics technologies have transformed plant microbiome research, allowing the characterization of the microbial composition and functional potential at unprecedented resolution. However, the resulting datasets often contain thousands of microbial and molecular features measured across relatively few biological samples. Such datasets are high-dimensional, sparse, and compositional, and are strongly influenced by environmental and technical factors. These characteristics make plant microbiome datasets particularly challenging for conventional statistical analyses and create a strong rationale for machine-learning approaches.[3]

Machine learning (ML) in particular can identify nonlinear relationships between microbiome features and plant phenotypes, classify disease states, select informative microbial biomarkers, and predict ecosystem or plant responses. Recent studies have increased the emphasis on integrating microbiome information with plant genotype, environmental conditions, and multi-omics measurements and highlighted the growing role of ML in decoding plant–microbiome dynamics for sustainable agriculture.[4] More recently, Ma *et al.*[5] reviewed artificial intelligence (AI) applications specifically for the rhizosphere microbiome and proposed an integrated framework combining microbiome-enabled genomic selection, synthetic microbial communities, federated learning, large language models, and digital twins. In this perspective, recent advances in ML and AI for plant microbiome research were summarized, major applications were examined, methodological limitations were discussed, and emerging opportunities for microbiome-enabled agriculture were identified.

2. Plant microbiome data and the need for machine learning

Microbiome datasets are heterogeneous, encompassing amplicon sequencing data that characterize bacterial and fungal community composition. Shotgun sequencing provides broader information on microbial genomes and functional genes, genome-wide metatranscriptomic approaches, and metabolomics, which together provide complementary information for characterizing microbial communities and the metabolites produced by associated microorganisms. Accordingly, integrating these complementary data modalities may help mitigate biases associated with compositional features and sparse taxonomic representations and may provide a more robust analytical framework than conventional classification and regression approaches alone. Busato *et al.*[3] emphasized that low sample size, soil heterogeneity, and technical variation can adversely affect ML performance in plant microbiome studies. They also identified compositionality and sparsity as important sources of difficulty for predictive modeling. These issues remain highly relevant to current AI-based microbiome research.

Supervised learning and ensemble learning methods, including random forests, support vector machines, and gradient boosting, have been widely applied across biological domains and have demonstrated considerable utility for classification and prediction tasks. However, these approaches also have several limitations in microbial community research. Classification models can effectively distinguish phenotypic traits between healthy and susceptible plants, but their performance may be affected by high-dimensional microbial datasets, class imbalance, noise, and limited sample sizes. In addition,

Table 1. Classification of bacterial communities associated with plant microbiome

Plant microbiome location	Major bacterial classes/groups	Main findings/functions	Ref.
Rhizosphere	Alpha-, Gammaproteobacteria; Bacilli; Actinomycetia; Bacteroidia; Acidobacteriae	Nutrient cycling, root growth, hormone production, pathogen suppression	2–10
Endosphere	Alpha-/Gammaproteobacteria; Bacilli; Actinomycetia	Internal colonization, stress tolerance, nutrient acquisition and plant defense	2,9
Phyllosphere	Alphaproteobacteria, Gammaproteobacteria	Leaf colonization, nutrient use, pathogen competition and environmental adaptation	10
Soil-associated microbiome	Acidobacteriae, Actinomycetia, Proteobacteria, Bacteroidia, Gemmatimonadetes-related groups	Carbon decomposition, nutrient cycling and adaptation to soil conditions	2,9
Beneficial inoculants/ SynComs	Bacilli (*Bacillus*), Gammaproteobacteria (*Pseudomonas*), Alphaproteobacteria (rhizobia)	Nitrogen fixation, phosphate mobilization, biocontrol and stress resistance	9,11,12

ensemble methods can be computationally intensive and may provide limited biological interpretability, making it challenging to identify the specific microbial mechanisms underlying model predictions. Similarly, support vector machines can perform well in high-dimensional settings but require careful selection of kernel functions and hyperparameters, while gradient-boosting methods may be prone to overfitting when model complexity is not appropriately controlled.

Regularized regression models, such as Least Absolute Shrinkage and Selection Operator and elastic net, can reduce overfitting when the number of microbial features substantially exceeds the number of biological samples and can therefore facilitate the identification of microbial taxa associated with environmental and plant phenotypic variables (Table 3). Nevertheless, these models may be sensitive to correlated microbial features, compositionality of microbiome data, and the choice of regularization parameters. Moreover, associations identified by regression models do not necessarily imply causal relationships between microbial taxa and plant phenotypes.

Table 2. Multi-omics integration for artificial intelligence-based plant microbiome research

Data layer	Information obtained	Potential ML application
16S/ITS sequencing	Microbial community composition	Classification and biomarker discovery
Shotgun metagenomics	Microbial genes and pathways	Functional prediction
Metatranscriptomics	Active microbial processes	Prediction of functional activity
Metabolomics	Plant/microbial metabolites	Mechanistic prediction
Plant transcriptomics	Host responses	Plant–microbe interaction modeling
Plant genomics	Host genetic variation	Microbiome-assisted breeding
Soil chemistry	Nutrient/environmental status	Yield and microbiome prediction
Climate/environment	Temperature, rainfall, drought	Context-specific prediction
Plant phenotyping	Growth, yield and stress responses	Supervised learning targets

Table 3. Major applications of ML and artificial intelligence in plant microbiome research

Application	Typical ML task	Main data sources	Potential outcome	References
Disease prediction	Classification/risk prediction	16S/ITS, metagenomics, plant phenotype	Early disease detection and risk assessment	4,5
Crop yield and plant phenotype	Regression/prediction	Microbiome, soil, climate, plant phenotype	Yield and biomass prediction	5,6
Abiotic-stress tolerance	Classification/ regression	Microbiome, environmental and plant data	Prediction of drought/salinity/stress responses	2,5
Nutrient cycling and soil health	Regression/feature selection	Metagenomics, soil chemistry, environmental data	Functional prediction and soil-health assessment	2,5,6
Biomarker discovery	Feature selection/ classification	Amplicon and metagenomic data	Identification of diagnostic taxa or genes	3–5
Microbial interaction prediction	Network/graph modelling	Taxonomic, genomic and metabolomic data	Identification of microbial interactions	5,6
SynCom design	Optimization/ prediction	Microbial genomes, interaction networks and plant phenotypes	Rational design of microbial consortia	5,7
Microbiome-assisted breeding	Genomic prediction	Plant genotype + microbiome + phenotype	Selection of genotypes associated with beneficial microbiomes	5,6
Crop management	Predictive modelling	Microbiome, soil, weather and agronomic data	Microbiome-informed management decisions	5,7

Abbreviations: ITS: Internal transcribed spacer; ML: Machine learning.

The application of deep learning has remained comparatively limited in microbial community-based machine learning, primarily because microbial datasets often contain a very large number of features relative to the number of biological samples, illustrating the compositional heterogeneity of the samples. This imbalance makes deep-learning models particularly susceptible to overfitting and limits their generalizability. Deep-learning approaches also typically require substantial computational resources, careful model architecture and hyperparameter optimization, and sufficiently large, diverse, and well-annotated datasets. Consequently, their application to microbiome research may be constrained by the scarcity of large-scale, standardized datasets.

More recently, Ma *et al.*[5] proposed integrating large language models, knowledge graphs, retrieval-augmented generation, federated learning, digital twins, and autonomous AI agents into rhizosphere microbiome research. Although these emerging approaches may provide new opportunities for integrating heterogeneous biological data and supporting microbiome analysis, their practical application remains constrained by challenges related to data quality, model interpretability, computational requirements, privacy, standardization, and experimental validation. Therefore, further methodological development and validation using large, diverse, and biologically characterized datasets are required before these approaches can be broadly adopted in rhizosphere microbiome research.

One of the most promising applications of AI is the prediction of plant disease based on microbiome characteristics. Plant pathogens do not operate independently; disease development can be influenced by the composition and functional properties of surrounding microbial communities. Systematic analyses of soil microbiomes and associated plant microbial communities have highlighted the integration of computational approaches, host genetics, microbial interactions, and microbiome engineering as a promising strategy for sustainable agriculture.[6] ML models can integrate microbial abundance with soil properties, climatic variables, and plant genetic information. Such multimodal models may be more informative than models based exclusively on microbiome composition because plant performance is influenced simultaneously by host genotype and environmental conditions. Accordingly, the plant microbiome has become an important component of climate-resilient agriculture.[2] AI can potentially identify microbial combinations associated with stress tolerance and predict which microbial communities are most likely to benefit specific plant genotypes. The challenge is that microbiome–stress relationships are strongly context-dependent. A microbial taxon associated with drought tolerance in one soil or cultivar may not produce the same effect elsewhere. Models therefore require multi-environment datasets and external validation. A potential strategy would be to establish ensemble method of model evaluation under multi-cultural conditions and to evaluate the multi-genotype trials of the predicted model.

Machine learning can contribute by predicting microbial compatibility, identifying potentially complementary strains, and optimizing community composition. Integration of AI with synthetic biology may therefore facilitate rational design of plant-associated microbial consortia.[5,7] Dubey *et al.*[7] described plant-microbiome engineering as an emerging strategy for crop health and sustainability, while emphasizing the importance of microbial partners and targeted community design. The major challenge is ecological stability. A SynCom that performs well in a controlled laboratory experiment may behave differently when introduced into a complex agricultural soil containing thousands of competing microorganisms. Ma *et al.* proposed hologenome-based genomic selection as an "inside-out" approach that integrates host genetic information and microbiome characteristics to improve breeding decisions.[5] This framework represents an important shift from conventional host-centric breeding toward a more integrated plant–microbiome breeding framework.

Several developments are likely to shape plant microbiome AI research over the next decade. First, larger and better-standardized datasets will be essential. Data from multiple cultivars, locations, soil types, and growing seasons should be collected using harmonized protocols. Second, multimodal AI should integrate microbial, plant, soil, and environmental data rather than treating microbiome composition in isolation. Third, causal AI should complement predictive modeling. The objective should shift from identifying correlations toward determining which microbial changes actually cause improved plant performance. Fourth, explainable AI will be important for translating computational predictions into biological hypotheses. Fifth, federated learning could facilitate collaboration among institutions while allowing sensitive or proprietary datasets to remain locally stored. Finally, AI should increasingly be connected with synthetic microbial communities, robotics, sensors, and field experiments, creating an iterative cycle in which computational predictions are tested experimentally, and experimental outcomes improve subsequent predictions (Figure 1).

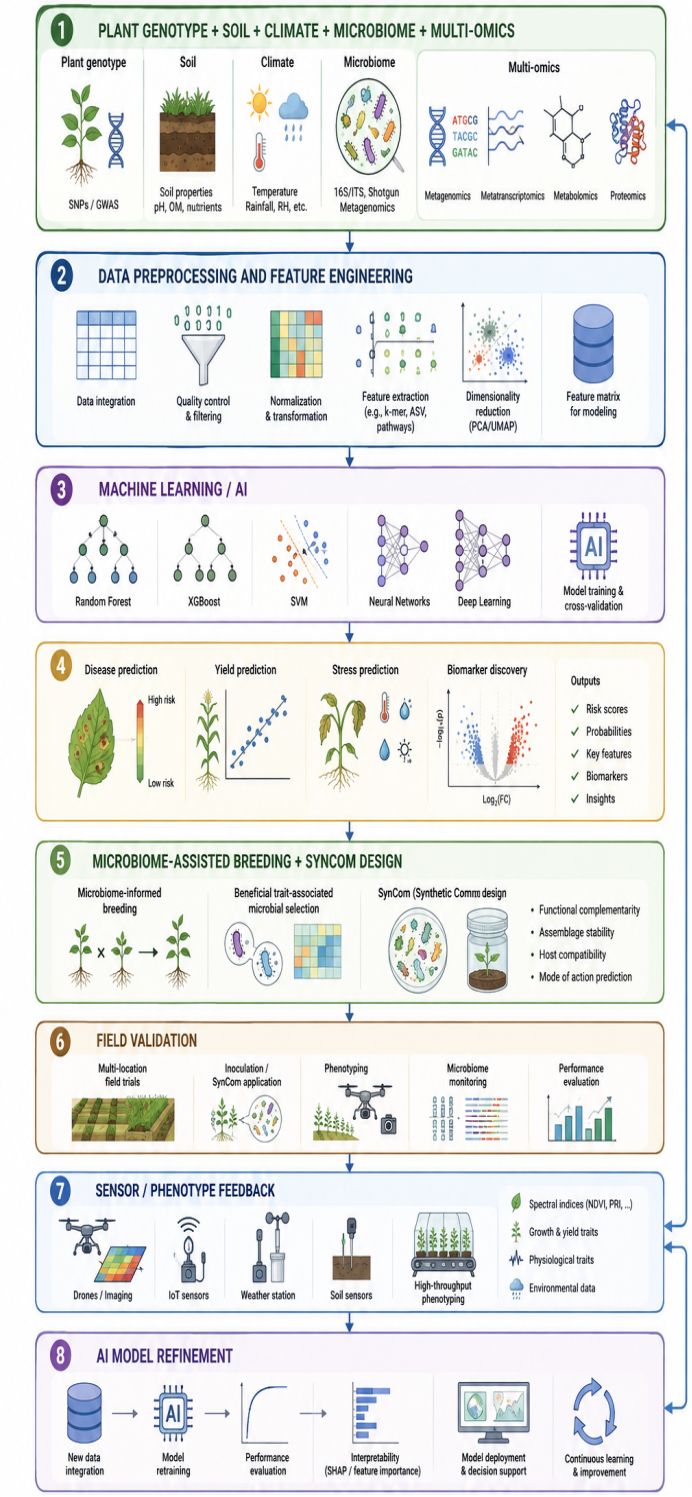


Figure 1. Role of multimodal machine learning and AI approaches in plant microbiome research

Abbreviations: AI: Artificial intelligence; ASV: Amplicon sequence variant; FC: Fold change; GWAS: Genome-wide association study; IoT: Internet of Things; ITS: Internal transcribed spacer; NDVI: Normalized difference vegetation index; PCA: Principal component analysis; PRI: Photochemical reflectance index; RH: Relative humidity; SHAP: SHapley Additive exPlanations; SNP: Single-nucleotide polymorphism; SVM: Support vector machine; UMAP: Uniform manifold approximation and projection. Created with BioRender.com.

Artificial intelligence can identify candidate microorganisms, predict interactions, and prioritize metabolic functions. Synthetic biology can then modify microbial strains or construct defined communities to test those predictions. Kamath *et al.* described this convergence of AI, omics, and synthetic biology as a route toward rational design of plant growth-promoting microorganisms and microbial consortia.[8] However, engineered microbial systems introduce additional considerations, including ecological stability, biosafety, horizontal gene transfer, environmental persistence, and regulatory approval. Therefore, future plant microbiome engineering should integrate computational prediction with controlled experimentation, ecological assessment, and appropriate regulatory frameworks. The most promising applications include disease prediction, plant phenotype prediction, stress tolerance, nutrient management, biomarker discovery, microbiome-assisted breeding, and SynCom design. Recent reviews emphasize that the ultimate objective is not merely to predict microbiome states but to use those predictions to engineer beneficial plant–microbe interactions.[5–8] Integrating AI with multi-omics, synthetic biology, and precision agriculture offers a promising route to predictive and engineered plant microbiomes. If these approaches can be validated across diverse field environments, microbiome-informed AI could become an important component of sustainable and climate-resilient crop production.[9–12]

3. Conclusion

The increasing application of ML and AI in microbiome research should be considered complementary to, rather than a replacement for, established statistical approaches. Traditional multivariate methods such as canonical correspondence analysis, redundancy analysis, and non-metric multidimensional scaling remain valuable for exploring microbial community structure, evaluating relationships between community composition and environmental variables, and generating readily interpretable ecological patterns.[13] Intermediate approaches, including sparse partial least squares discriminant analysis, compositional data analysis, and mixed-effects models, can address high dimensionality, feature selection, the compositional nature of microbiome data, and hierarchical or repeated-measures experimental designs while retaining comparatively strong interpretability.[14–16] In particular, compositional data analysis is important because sequencing-based microbiome measurements represent relative rather than absolute abundances, and inappropriate treatment of compositional data can lead to misleading associations.[15] ML/AI approaches are most likely to provide additional value when the objective extends beyond describing associations to prediction, classification, nonlinear pattern recognition, integration of heterogeneous data types, or modeling complex interactions that are difficult to capture using conventional statistical frameworks. Thus, AI should not necessarily be preferred solely because of its computational complexity; its added value should be demonstrated through improved predictive performance, generalizability, robustness, or the discovery of biologically meaningful patterns relative to appropriate simpler statistical baselines. ML represents a fundamental component of modern AI by providing algorithms that learn patterns from data and generate predictions or classifications without requiring all relationships to be explicitly specified in advance. The increasing availability of microbiome big data, including metagenomic, metatranscriptomic, metabolomic, proteomic, environmental, and host-phenotypic measurements, can therefore enhance ML by providing sufficiently rich and heterogeneous information for identifying complex, nonlinear, and potentially interacting biological signals. However, increasing data volume alone does not guarantee improved prediction, and appropriate preprocessing, dimensionality reduction, feature selection, independent validation, and careful control of overfitting remain essential.

Despite this potential, substantial practical barriers must be addressed before AI-driven microbiome engineering can be implemented routinely. Large-scale multi-omics integration requires substantial computational resources for data storage, preprocessing, model training, and validation, particularly when datasets contain very large numbers of microbial and molecular features.[17] In addition, effective AI development depends on sufficiently large, diverse, well-annotated, and independently validated datasets. Data sharing remains challenging because microbiome datasets are heterogeneous and are often accompanied by incomplete or inconsistently structured metadata, limiting cross-study integration and reproducibility. Standardized metadata frameworks, including the Minimum Information about any (x) Sequence standards, provide structured requirements for describing sequencing samples and their contextual information and can facilitate data discovery, comparison, and integration across studies.[18] The adoption of Minimum Information about any (x) Sequence together with FAIR data practices, persistent identifiers, harmonized ontologies, and reproducible computational workflows will therefore be important for developing interoperable datasets suitable for AI training and benchmarking. Ultimately, successful AI-driven microbiome engineering will require not only sophisticated algorithms but also coordinated data-sharing infrastructure, scalable

computational resources, standardized metadata, rigorous benchmarking against conventional statistical methods, and experimental validation of model-derived predictions.

Acknowledgments

The author acknowledges BioRender for creating Figure 1 and the Graphical Abstract, as well as for language corrections.

Funding

None.

Conflict of interest

The author declares no conflict of interest.

Author contributions

This is a single-authored article.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Not applicable.

Further disclosure

The paper has been uploaded to preprints.org (doi: 10.20944/preprints202609.1390.v1).

References

1. Berendsen RL, Pieterse CMJ, Bakker PAHM. The rhizosphere microbiome and plant health. *Trends Plant Sci*. 2012;17(8):478-486.

   doi: 10.1016/j.tplants.2012.04.001

2. Compant S, Cassan F, Kostić T, *et al.* Harnessing the plant microbiome for sustainable crop production. *Nat Rev Microbiol*. 2025;23(1):9-23.

   doi: 10.1038/s41579-024-01079-1

3. Busato S, Gordon M, Chaudhari M, *et al.* Compositionality, sparsity, spurious heterogeneity, and other data-driven challenges for machine learning algorithms within plant microbiome studies. *Curr Opin Plant Biol*. 2023;71:102326.

   doi: 10.1016/j.pbi.2022.102326

4. Yoosefzadeh Najafabadi M, Khalaf EM, Mysara M, El-Baz AM. Editorial: Harnessing machine learning to decode plant-microbiome dynamics for sustainable agriculture. *Front Microbiomes*. 2025;4:1602938.

   doi: 10.3389/frmbi.2025.1602938

5. Ma J, Qiao J, Cao Y, Cheng Z. Harnessing artificial intelligence to decode the rhizosphere microbiome. *aBIOTECH*. 2026;7(1):100005.

   doi: 10.1016/j.abiote.2025.100005

6. Berruto CA, Demirer GS. Engineering agricultural soil microbiomes and predicting plant phenotypes. *Trends Microbiol*. 2024;32(9):858-873.

   doi: 10.1016/j.tim.2024.02.003

7. Dubey A, Malla MA, Vimal SR, Kumar A, Prasad SM, Khan ML. Plant-microbiome engineering: synergistic microbial partners for crop health and sustainability. *Plant Growth Regul*. 2025;105(6):1825-1839.

   doi: 10.1007/s10725-025-01385-5

8. Kamath A, Patel D, Saraf M, Patel S. Rooted intelligence: Integrating AI, Omics, and Synthetic biology to engineer the plant microbiome. *3 Biotech*. 2026;16(1):10.

   doi: 10.1007/s13205-025-04614-w

9. Prasannath K, Peter Arulraj S, Shakthivel K, Rakulan T, Karunarathna SC. Revisiting the ecological significance, composition, and functions of plant microbiome: a review. *N Z J Crop Hortic Sci*. 2025;53(5):1329-1361.

   doi: 10.1080/01140671.2025.2458811

10. Zeng Q, Hu HW, Ge AH, *et al.* Plant–microbiome interactions and their impacts on plant adaptation to climate change. *J Integr Plant Biol*. 2025;67(3):826-844.

    doi: 10.1111/jipb.13863

11. Yadav J, Gehlot P, Soni P, Jain T. Plant microbiome engineering: from inoculation to genome editing. *Front Microbiol*. 2026;17:1781381.

    doi: 10.3389/fmicb.2026.1781381

12. Wei M, Lian T, Chen L, *et al.* The plant microbiome: From ecological foundations to precision microbial engineering for sustainable agriculture. *iMeta*. 2026;5(4).

    doi: 10.1002/imt2.70152

13. Legendre P, Gallagher ED. Ecologically meaningful transformations for ordination of species data. *Oecologia*.2001;129(2):271-280.

    doi: 10.1007/s004420100716

14. Lê Cao KA, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinform*. 2011;12(1):253.

    doi: 10.1186/1471-2105-12-253

15. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. *Front Microbiol*. 2017;8:2224.

doi: 10.3389/fmicb.2017.02224

16. Bolker BM, Brooks ME, Clark CJ, *et al.* Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol Evol*. 2009;24(3):127-135.

    doi: 10.1016/j.tree.2008.10.008

17. Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol*. 2017;13(11):e1005752.

    doi: 10.1371/journal.pcbi.1005752

18. Genomic Standards Consortium. *MIxS: Minimum Information about any (x) Sequence*. Genomic Standards Consortium; 2026. https://zenodo.org/records/21679212