AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems

| Ecology | Minireview

Advances and future directions in identifying specific taxa from microbial meta-omics data: from pipeline to deep learning

Jingkang Zhang,[1] Xingjie Wang,[1,2] Di Wang,[2] Zikui Zheng,[2] Hongmei Wang,[2] Liyuan Ma[2,3]

AUTHOR AFFILIATIONS See affiliation list on p. 14.

ABSTRACT Molecular profiling enabled by meta-omics technologies has significantly expanded our knowledge of microbial catalog across diverse environments. Increasing attention has now been focused on identifying ecologically significant taxa, particularly keystone that stabilize communities, rare taxa that underpin functional redundancy, and indicators that reflect environmental gradients. However, current pipeline methods remain limited in deciphering complex ecological relationships and modeling the evolution of community dynamics. As a transformative computational tool, deep learning (DL) offers novel strategies to address these challenges through autonomous feature extraction, nonlinear interaction modeling, and integration of multi-modal data sets. Nevertheless, there are still obstacles to the widespread adoption of DL for collaborative identification of specific microbial taxa, primarily including the intrinsic heterogeneity and imbalance of data sets, the difficulty of model generalization across diverse ecosystems, and the limited ecological interpretability of model outputs. This review summarizes existing research advances and proposes to build a unified DL framework for multi-modal data, exploring its implementation pathways, challenges, and potential coping strategies. The envisioned framework establishes a multi-task learning architecture for unified identification of keystone, rare, and indicator taxa, incorporating domain knowledge through ecological constraint layers and explainable AI modules, while providing flexible implementation pathways for heterogeneous data integration and model customization across microbial ecosystems. This framework has the potential to form a closed-loop verification in combination with synthetic microbial community experiments, reshape the paradigm of microbial community research, and promote the transition from empirical classification to mechanistic ecological cognition.

KEYWORDS specific taxa identification, meta-omics, deep learning, high-throughput sequencing, explainable AI

Microbial communities play crucial roles in global biogeochemical cycles and host–microbe symbiosis, being widely distributed across diverse environments ranging from the cryosphere to the deep lithosphere (1, 2). Although culturomics has made significant strides in specific ecosystems in recent years (for instance, over 50%-70% of microbes in the human gut are now culturable), the vast majority of microbes in broader complex environments remain unculturable (e.g., typically less than 1%–5% in soil and marine environments), persisting in an "invisible" state (3–5). These uncultivable microbes, along with silent functional genes and uncharacterized taxa, are collectively referred to as "microbial dark matter" (6, 7). Despite their undisputed ecological significance, the investigation of this vast, untapped repertoire of microbial diversity and metabolic potential has been hampered by methodological limitations (8–10).

Advances in high-throughput sequencing technologies, such as 16S rRNA gene amplicon sequencing, genome-resolved metagenomics, and functional genomics, have markedly enhanced our capacity to investigate microbial diversity and community

Editor Andrew Bartko, University of California, San Diego, La Jolla, California, USA

Address correspondence to Liyuan Ma, maly@cug.edu.cn.

The authors declare no conflict of interest.

See the funding table on p. 15.

Published 30 April 2026

composition in a variety of ecosystems (11–13). As microbial ecology continues to evolve toward a function-oriented and ecosystem holistic perspective, taxa that exert disproportionate ecological effects within communities have attracted growing attention (2, 14). In particular, keystone taxa that maintain community structure and stability (15), rare taxa that underpin functional redundancy (16), and indicator taxa that reflect environmental gradients (17). Crucially, these functional categories are not mutually exclusive in biological systems; their intersections often harbor taxa of unique functional significance (18). For instance, rare keystone, despite their scarcity, exert disproportionate influence on community stability, acting as "hidden drivers" of the ecosystem (16, 18). Furthermore, taxa sharing all three attributes simultaneously maintain network topology and reflect environmental shifts, serving as "critical ecosystem barometers" (19–21). Therefore, elucidating the ecological roles, interrelationships, and impacts of these specific taxa on community dynamics is becoming an important direction in microbial ecology research (22).

Although the emergence of high-throughput sequencing technology and the launch of large-scale studies such as the International Microbiome Projects have accelerated the rapid accumulation of massive microbiome data sets (23–25), the exponential growth of microbiome sequencing data has also introduced significant computational and analytical challenges (26). Traditional methods that rely on linear assumptions and manual feature selection struggle to accommodate the high dimensionality, sparsity, and zero-inflation characteristic of microbiome data sets (27, 28). Moreover, the highly nonlinear, hierarchical structures and complex interactions within microbial communities further limit the applicability of these conventional approaches in capturing community assembly patterns across habitats and disentangling ecological dependencies (29, 30). Data-driven methods such as machine learning (ML) and deep learning (DL) offer the ability to automatically extract features and model nonlinear relationships (22, 31). Especially, DL excels at learning hierarchical representations and detecting subtle ecological interaction patterns, making it a powerful analytical tool in microbial ecology (32, 33).

Based on the development of high-throughput sequencing technology and the attendant challenges in data analysis, this article discusses the current status of research on three types of taxa with specific ecological significance and key issues that remain unresolved and summarizes the applications and limitations of classical ML methods in microbial ecology research. Furthermore, we analyze the current applications and challenges of DL in identifying microbial taxa and highlight the transformative role of DL in advancing this field. The aim of this review is to provide insights into the construction of a unified DL framework that enables more accurate and comprehensive microbial taxa identification, facilitating a deeper understanding of microbial community structure and ecosystem functions.

IDENTIFYING SPECIFIC MICROBIAL TAXA FROM META-DATA: CURRENT STATUS AND CHALLENGES

High-throughput sequencing technologies have revolutionized microbial ecology by generating comprehensive and cultivation-independent data sets that capture both taxonomic composition and functional potential across diverse habitats (26, 29, 34). 16S rRNA gene amplicon sequencing rapidly obtains taxonomic information of microbial communities from diverse ecosystems (e.g., soils, marine environments, and the human gut) by targeting conserved regions interspersed with hypervariable segments, revealing numerous previously undetectable microbial taxa and substantially expanding our understanding of microbial diversity and ecosystem dynamics (24, 25, 35). In addition, functional genomics connects cultured isolates to their genetic and functional properties (36), and metagenomics reconstructs complete gene catalogs and metabolic pathways directly from environmental samples (24, 37). Collectively, these methods provide a rich meta-omics resource that allows researchers to move beyond simple taxa inventories toward identifying ecologically important microbial taxa.

In highly diverse microbial communities, ecologists classify certain taxa as ecologically significant due to their disproportionately large roles. Identifying specific microbial taxa that play critical ecological roles, particularly keystone, rare, and indicator taxa, is essential for a deep and comprehensive understanding of microbial community ecology. Keystone is classically defined as taxa that exert a disproportionately large influence on their environment relative to their abundance (19). Since Paine first introduced the concept of keystone taxa in 1969 (38), its precise ecological definition has been a longstanding subject of debate and has continually evolved over the years (39). Co-occurrence network analysis, a commonly employed method for identifying keystone taxa, investigates potential microbial interactions by examining abundance correlations within sequencing data, thereby constructing microbial association networks (30, 40). In these networks, putative keystone taxa often emerged as highly interconnected hub nodes or exhibited high centrality measures, suggesting extensive interactions with other community members (41). Such network-based scoring has been widely used to nominate candidate keystone taxa across various ecosystems, ranging from soil to human gut habitats (15, 22, 42). However, network correlations only provide suggestive insights and do not necessarily imply causation, while spurious associations (e.g., co-occurrence due to shared habitat preferences rather than direct interactions) may mislead interpretations (43). Moreover, co-occurrence networks are highly sensitive to the inherent noise and sparsity in microbial data sets (44), leading to false-positive correlations and overlooked interactions (45), particularly involving rare taxa (46).

High-throughput sequencing reveals that, in addition to a few dominant taxa, most communities contain organisms detected at very low frequencies (16). These rare taxa can number in the hundreds or thousands within a single sample, representing a vast reservoir of biodiversity collectively known as the rare biosphere (47, 48). Traditional studies have often overlooked the roles of rare taxa; however, in recent years, rare taxa are increasingly recognized as critical yet vulnerable components of ecosystems (49–51). Despite their low abundance, rare taxa significantly contribute to functional redundancy and serve as a potential reservoir for ecological resilience (16, 18). They persist in small numbers until environmental conditions become favorable for their growth, at which point some rare taxa rapidly proliferate, performing crucial biogeochemical transformations when needed, such as in response to specific substrates or environmental conditions (e.g., pollutants or changes in pH) (52, 53). Nevertheless, many researchers use arbitrary thresholds (e.g., taxa with relative abundances below 0.1% or 0.01% within samples) to define the rare biosphere (54, 55). These thresholds vary considerably and may lack universality across different habitats and ecosystems (56, 57). Such empirical categorization involves subjective assumptions and risks neglecting the dynamic characteristics of rare taxa in specific communities (58, 59). The inability to capture this dynamic nature has constrained advancements in identifying and studying rare taxa within the broader ecological context (60). Additionally, sequencing depth is equally critical, as low-abundance taxa may be missed or discarded as noise in shallow sequencing efforts (26, 61, 62).

Although indicator taxa may not drive community processes, their abundance patterns, such as presence, absence, or fluctuations, sensitively reflect changes in specific environmental conditions (14, 24, 63), offering valuable insights into habitat states (64). Due to their rapid responses to perturbations and relative ease of sampling (e.g., via water or soil samples or non-invasive body swabs), microbes represent attractive early-warning indicators (14, 65, 66). To identify indicator taxa, researchers typically adapt classical community ecology techniques to high-throughput sequencing data (67, 68). One prominent method is Indicator Value (IndVal) (69), which quantitatively evaluates the association of each taxon with specific groups of samples (such as habitat types or experimental conditions) based on its consistency and exclusivity within that group (58, 70). Another approach, Threshold Indicator Taxa Analysis (TITAN) (71), is specifically designed for continuous environmental gradients. TITAN identifies taxa exhibiting abrupt changes in abundance at certain thresholds of environmental parameters (e.g., salinity

or pollutant concentration), thereby pinpointing not only indicator taxa but also critical points at which community composition shifts occur. This approach has been effectively utilized in freshwater and soil research to detect early-warning signals (72, 73).

Despite significant progress achieved by the methods outlined above, the inherent high-dimensionality, sparsity, zero-inflation, and composition of sequencing data sets limit the effectiveness of traditional methods in accurately identifying specific microbial taxa (26, 74). Moreover, existing practices typically identify keystone, rare, and indicator taxa in isolation, without considering the association and overlaps of these categories, particularly regarding whether rare taxa universally possess keystone properties, and whether keystone taxa may also serve as indicator in specific environments. Due to these limitations, current studies have difficulty capturing the complex interactions of microbial ecological network and the inner contribution of these taxa to the specific ecosystems. These technical challenges and unresolved theoretical gaps in identification continue to impede in-depth ecological inference and hinder the development of effective ecosystem management strategies. To address these challenges, there is an urgent need for advanced analytical tools and systematic frameworks capable of modeling nonlinear relationships, effectively handling data sparsity, and improving the accuracy of identifying microbial taxa (75), thereby paving the way for the adoption of ML methodologies.

MACHINE LEARNING IN IDENTIFYING MICROBIAL TAXA: CONTRIBUTIONS AND LIMITATIONS

ML has been widely applied in analyzing complex data sets (some classic cases are summarized in Table 1) and uncover microbial coexistent patterns that traditional methods often fail to discern. By leveraging supervised and unsupervised learning techniques, ML facilitates more accurate and scalable approaches to microbial taxonomy, biomarker discovery, and functional inference (33, 75). These applications have demonstrated the robustness of ML in extracting valuable information from sequencing data and outperforming traditional methods in handling complex, high-dimensional data (76–79). As sequencing data accumulates and methodologies advance, researchers have begun to explore the use of ML methods to further identify taxa that have special effects on community ecological functions from massive microbiome data (80–82).

ML could serve as a powerful standalone tool that directly leverages sequencing-derived community profiles to identify and delineate ecologically critical taxa. A representative application used random forest (RF) models to investigate the stability of soil microbiomes under disturbance, identifying key metabolic functions and associated microorganisms that contribute to maintaining community stability (82). Among these, specific metabolic pathways, particularly nitrogen metabolism, were identified as "key functions" for maintaining stability, predominantly carried out by certain taxa such as *Nitrospira*. Without relying on network or simulated dynamics, this supervised ML model quantified the contribution of each feature to observed stability outcomes (via feature and permutation importance), thereby highlighting candidate keystone taxa together with the metabolic functions. Similarly, in agricultural microbiomes, ML models help identify key taxa that influence soil microbial community assembly and interactions (99). In addition to feature importance, researchers perform prediction degradation analysis (PDA), which involves removing taxa from community data or randomizing their values to observe the decline in model performance or changes in predicted output to evaluate the impact of taxa on community stability (100–102). This process is actually implemented by permutation importance in the RF model, which is similar to ecologists removing a taxa in an experiment to see if the ecosystem changes occur (103, 104). In summary, supervised ML models assess taxa by their contribution to predicting community-level outcomes, benefiting from using real functional or stability data as a benchmark, rather than relying solely on network structure (22). However, it requires appropriate measurable outcomes or proxies for "community impact" to train models (24).

TABLE 1 Examples of common tasks and machine learning methods used in microbiome research[a,b]

Task	Predictive goal	Method	Data source	Validation strategy	Reference
Specific taxon identification	Keystone taxa	DKI framework	Hybrid	Wet-lab (synthetic community)	Wang et al. (83)
		Linear discriminant analysis	Simulated	Additional simulations	Berry and Widder (40)
		Random forest	Observed	Computational evaluation (genome-resolved analysis)	Cheng et al. (81)
			Hybrid	Wet-lab (invasion experiments)	Wu et al. (84)
		LIMITS	Hybrid	Additional simulations	Fisher and Mehta (85)
	Rare taxa	Ulrb	Observed	Computational evaluation	Pascoal et al. (86)
	Indicator taxa	Random forest, neural networks	Observed	Computational evaluation (independent test set)	Thompson et al. (87)
		Random forest	Observed	Wet-lab (m-ddPCR validation)	Zheng et al. (88)
		Compare supervised ML models (SVMs, RF, linear regression)	Observed	Computational evaluation (genome-resolved analysis)	Liu et al. (89)
		Deep neural networks	Observed	Computational evaluation (comparison with IndVal)	Tsai et al. (90)
Phenotyping	Disease (inflammatory bowel disease)	Random forest, Lasso, elastic nets	Observed	Computational evaluation (cross-study validation)	Wirbel et al. (91)
	Colony morphology classification	Random forest	Observed	Wet-lab	Huang et al. (78)
Interaction analysis	Classification of interactions	Random forest	Hybrid	Wet-lab and additional simulations	DiMucci et al. (92)
Biomarker discovery	Antimicrobial candidates against *E. coli*	Deep neural networks	Observed	Wet-lab (*in vitro* and *in vivo*)	Stokes et al. (93)
Microbial classification	Disease (e.g., colonic screen-relevant neoplasias)	Compare supervised ML models (e.g., SVMs, RF, XGBoost)	Observed	Computational evaluation	Topcuoglu et al. (94)
Functional annotation	Antibiotic resistance genes	HMD-ARG	Observed	Wet-lab (heterologous expression)	Li et al. (95)
Dynamics prediction	Predict future complex community behavior	LSTM	Observed	Wet-lab	Baranwal et al. (96)
Metagenome assembly	Contaminant removal	Deepurify	Hybrid	Additional simulations	Zou et al. (97)
	Metagenomic binning	SemiBin	Observed	Computational evaluation	Pan et al. (98)

[a]Data source indicates whether studies used observed, simulated, or hybrid data. Observed: real-world sequencing data sets (e.g., 16S rRNA amplicon or metagenomics) obtained from natural environments or clinical samples. Simulated: synthetic data generated mathematically (e.g., using generalized Lotka-Volterra models). Hybrid: studies that utilize both real-world observed data sets and simulated data sets (either combined for augmentation or used in parallel for comprehensive benchmarking).
[b]Validation strategy including wet-lab, additional simulation, or computational evaluation. Wet-lab: findings verified through subsequent laboratory experiments (e.g., culturing, qPCR, or synthetic community construction). Additional simulations: validation performed by new data sets under varying parameters or theoretical conditions (e.g., interaction strengths or noise levels) to rigorously test the model's robustness and theoretical generalizability. Computational evaluation: validation performed solely using statistical metrics (e.g., AUC, precision-recall) or data splitting (e.g., cross-validation) without new experimental data.

Accordingly, researchers integrated machine learning with classical identification approaches and reduced spurious associations via reciprocal validation, thereby improving robustness. Berry and Widder (40) combined microbial co-occurrence network topology features with linear discriminant analysis (LDA) based on ML to classify specific taxa according to their network connectivity characteristics. Cheng et al. (81) identified highly connected "hub" taxa in plant rhizosphere microbial communities through co-occurrence network analysis and then verified the accuracy by training multiple ML models. Notably, the RF model highlighted the same taxa (e.g., *Nitrospiraceae*, *Xanthomonadaceae*, *Mycobacteriaceae*) marked as hubs by network analysis, confirming that these taxa were strong predictors of soil potential function, outperforming abiotic factors in explaining soil functional variation. Although the combination of network analysis and ML goes beyond simple hub identification, it also faces challenges in network inference accuracy and model interpretability (44, 105). Another integrative strategy combines classical ecological mechanistic models with ML through computational simulations (84, 85), and the embedding of ecological prior knowledge effectively improves the interpretability of the model. A landmark study by Fisher and Mehta (85) introduced the LIMITS (Learning Interactions from Microbial Time Series), a sparse regression method

with bootstrap aggregation. Unlike traditional methods, they inferred Lotka-Volterra species interaction models from human gut microbiome time-series data and successfully identified the keystone taxa that significantly impacted community structure (85). Their key roles in the model dynamics were examined by removal of these taxa from network; however, it requires relatively rich longitudinal data to fit the model. Wu et al. (84) built a microbial community invasion scenario data set based on generalized Lotka-Volterra (gLV) model, offering a novel analytical strategy for data-scarce scenarios by synthesizing labeled invasion scenarios from mechanistic knowledge. Then, they trained an RF classifier to learn a generalizable mapping and identify the most effective "key resistance taxa" by introducing invasive taxa, with predictions transferring to fecal-derived synthetic communities in wet-lab tests (84).

ML also provides novel solutions to address limitations associated with traditional methods for identifying rare microbial taxa. Pascoal et al. (86) introduced "ulrb," an unsupervised learning-based approach that applied k-medoids to the within-sample abundance vector and partitions taxa into rare, intermediate, and abundant categories without an artificial threshold. As the clustering operates along a single axis of abundance, standard data transformations preserve the relative distances used for classification of compositional and non-compositional inputs. Across distinct sequencing strategies, taxonomic units, and wide ranges of sample sizes, numbers of taxa, and sequencing depths, ulrb yielded consistent rarity assignments and showed better cross-method consistency than fixed-threshold rules. However, ulrb does not explicitly model zero inflation or remove compositional artifacts, as it classifies taxa from within-sample abundance distributions rather than via generative count models. Beyond the abundance classification, ML methods have also been utilized to uncover the functional or interactions of rare taxa (16, 40). Pust and Tümmler (106) revealed that rare taxa were the most critical variables differentiating healthy children versus children with cystic fibrosis (CF) by constructing a classification model. Additionally, anomaly detection algorithms have been employed to identify statistically rare yet highly active or strongly interactive taxa from microbial data sets (107). Ishiya and Aburatani (107) demonstrated that even extremely low-abundance taxa exhibit abundance fluctuations that could signify important ecological shifts or environmental influences. In summary, ML has enhanced the objectivity of rare taxa identification through adaptive abundance thresholds and clustering techniques and also provided valuable approaches to elucidate the ecological roles of these taxa within communities. Nonetheless, due to the inherently low counts, compositional constraints, and high noise associated with rare taxa, accurate identification and importance evaluation remain challenging, and current unsupervised approaches represent an initial step toward comprehensive understanding (61).

Leveraging supervised learning and permutation-based attribution, ML prioritized cross-validated predictive contribution rather than correlation alone, providing a new pathway for identifying indicator. Thompson et al. (87) used indicator species analysis (ISA) to identify 285 bacterial operational taxonomic unit (OTUs) from a total of 1,709 OTUs that significantly correlated with high dissolved organic carbon (DOC) concentrations. By comparing these indicator with those identified as important by RF and neural network models, the researchers found that a subset of 86 overlapping OTUs provided the most robust prediction of DOC concentrations. These overlaps showed that the indicator identified by supervised ML was both statistically significant and predictively informative across heterogeneous communities. Furthermore, supervised ML models have been widely utilized to directly pick biomarkers from abundance data that best differentiate environmental or phenotypic categories (108–111). Zheng et al. (88) employed RF to classify gut microbiome, identifying microbial biomarkers that effectively distinguished inflammatory bowel disease (IBD; including ulcerative colitis [UC] and Crohn's disease [CD]) patients from healthy controls. Cheng et al. applied an RF model to accurately distinguish the differences in rhizosphere microbiome communities between two rice varieties, identifying family-level indicator taxa as key indicator groups

associated with differential cadmium accumulation between cultivars (81). In another study involving anaerobic fermentation reactors, RF regression with Synthetic Minority Oversampling Technique (SMOTE) balancing and fivefold cross-validation accurately predicted product yields in a medium-chain fatty acid-producing system, successfully identified four genera as indicator microbial groups associated with high-yield conditions (89). Collectively, these cases demonstrate that ML models not only predict environmental or functional states but also leverage feature selection processes to uncover the local microbial indicator taxa.

Although ML has been preliminarily applied to identify specific microbial taxa and has yielded valuable insights, it still faces several challenges (44, 112). A fundamental constraint is the heavy reliance on manual feature engineering, which requires substantial domain-specific expertise (74, 79) and frequently introduces biases (30). Handling the high variability of microbial communities across different habitats poses additional challenges for manual feature engineering, as distinct data sets often necessitate entirely different feature sets, thereby restricting the scalability and generalizability of ML models across diverse ecosystems (25, 74, 76). Despite demonstrating superior performance over traditional methods in dealing with high-dimensional data, ML models remain vulnerable to issues such as data sparsity and zero-inflation, which can lead to overfitting or reduced generalization capabilities (113). Furthermore, ML models encounter significant difficulties in integrating multimodal data sets, resulting in insights that are often incomplete or fragmented, thus limiting their capacity to fully elucidate microbial community dynamics (114). Another critical challenge associated with applying ML in microbial research is model interpretability (94). Despite achieving high accuracy in classification and prediction tasks, these models typically function as "black boxes," obscuring the biological relevance of their outputs (77, 105). This opacity complicates the translation of computational findings into actionable insights, particularly when aiming to understand the mechanisms that drive microbial community structure and functionality (102, 115, 116). These challenges emphasize the pressing need for more effective computational approaches capable of modeling the complex, nonlinear, and hierarchical relationships inherent in microbial data.

THE RISE OF DEEP LEARNING: A NEW FRONTIER IN IDENTIFYING MICROBIAL TAXA

DL has emerged as a transformative approach to address the limitations of classical ML methods, such as reliance on manual feature engineering and insufficient scalability when handling complex data sets (117). One of the primary strengths of DL lies in its ability to autonomously extract features directly from data without requiring researchers to predefine them, enabling the modeling of intricate nonlinear and hierarchical relationships (118). DL models such as convolutional neural networks (CNNs) have been successfully applied to the classification of microbial taxa based on DNA sequences, achieving impressive accuracy that surpasses traditional methods (119–121). A notable example is the Data-driven Keystone Species Identification (DKI) framework (83), which trains a composition Neural ODE (cNODE2) on steady-state microbiome samples to learn the mapping from species assemblages to compositions and then conducts virtual species removals to compute community-specific structural and functional keystoneness from predicted post-removal changes. DKI recovered true keystoneness on gLV-simulated communities across network connectivities with high accuracy (Spearman $\rho \approx 0.96$–0.98) and, in an eight-species synthetic consortium, outperformed topological baselines in wet-lab tests (accuracy 0.85 vs 0.40 for degree and 0.25 for betweenness). Additionally, graph-based DL has gained attention by integrating network analysis with neural networks to identify keystone taxa within microbial communities (122). These applications underscore the critical role of DL in facilitating the rapid and accurate identification of keystone taxa.

The application of DL on multi-omics data analysis has further advanced microbiome research, including the prediction of metabolic pathways, identifying biomarkers, and

assessing microbial responses to environmental stressors (30, 32, 123). For example, deep neural network models were trained on tens of millions of promoter-expression pairs to learn an environment-aware sequence-to-expression mapping, enabling accurate cross-condition expression prediction and revealing condition-specific regulatory signals (124). Tsai et al. (90) introduced deep optimal feature learning to identify indicator across ecological habitats, learning low-dimensional representations from 16S rRNA amplicon profiles to prioritize taxa and reveal robust environment-taxon associations. The method was reported to maintain performance under stratified cross-validation, and the indicator sets were more reproducible than traditional feature selection. Attention mechanisms (125), a key architectural innovation in DL, help the model identify which input features matter most in complex data sets by assigning a weight to each candidate taxon, gene, or interaction, thereby facilitating deeper interpretations of microbial interactions and functional roles (126). Melnyk et al. (127) combined microbial network analysis with Transformer-based DL models and applied the interpretability technique of layer-wise relevance propagation (LRP) to elucidate model predictions. By examining the attention weights and relevance scores of the models, researchers were able to highlight which taxa within the community were the most influential in driving transitions from healthy to diseased states.

Identifying rare taxa presents a unique challenge, as signals from these taxa are extremely weak (reflected in low read counts or abundance values) and often obscured by noise and more abundant taxa (61, 128). Traditional methods and classical ML models frequently struggle with class imbalance and low signal-to-noise ratios (details of the cases and discussions are given above in Identifying Specific Microbial Taxa from Meta-Data: Current Status and Challenges and in Machine Learning in Identifying Microbial Taxa: Contributions and Limitations), resulting in ecologically significant rare taxa being overlooked (129, 130). DL approaches offer new strategies to address this issue, one of which is to use DL to learn from massive unlabeled data and denoise complex inputs (121). Autoencoders, for example, have been applied to microbiome data sets to compress high-dimensional community features into lower-dimensional representations (131), effectively filtering out noise while preserving key variation, thereby enhancing the detectability of signals from rare taxa (132). Another powerful approach involves the use of deep generative models for data augmentation to mitigate the paucity of rare taxa examples. Generative adversarial networks (GANs) and related generative models have been utilized to increase model robustness by alleviating class imbalance and expanding the training data sets for rare taxa (133, 134). Early successes have demonstrated that reliable detection of members of the "rare biosphere" is achievable without the need for arbitrary abundance thresholds (86).

and 3), resulting in ecologically significant rare taxa being overlooked (129, 130). DL approaches offer new strategies to address this issue, one of which is to use DL to learn from massive unlabeled data and denoise complex inputs (121). Autoencoders, for example, have been applied to microbiome data sets to compress high-dimensional community features into lower-dimensional representations (131), effectively filtering out noise while preserving key variation, thereby enhancing the detectability of signals from rare taxa (132). Another powerful approach involves the use of deep generative models for data augmentation to mitigate the paucity of rare taxa examples. Generative adversarial networks (GANs) and related generative models have been utilized to increase model robustness by alleviating class imbalance and expanding the training data sets for rare taxa (133, 134). Early successes have demonstrated that reliable detection of members of the "rare biosphere" is achievable without the need for arbitrary abundance thresholds (86).

Overall, DL methods have overcome many limitations of earlier approaches, enabling more accurate identification of specific microbial taxa. However, despite these significant advantages, there remain several key challenges to the broader application of DL in identifying specific microbial taxa. Issues such as overfitting, data bias, and class imbalance hinder the robustness and generalizability of DL models (118, 135–137).

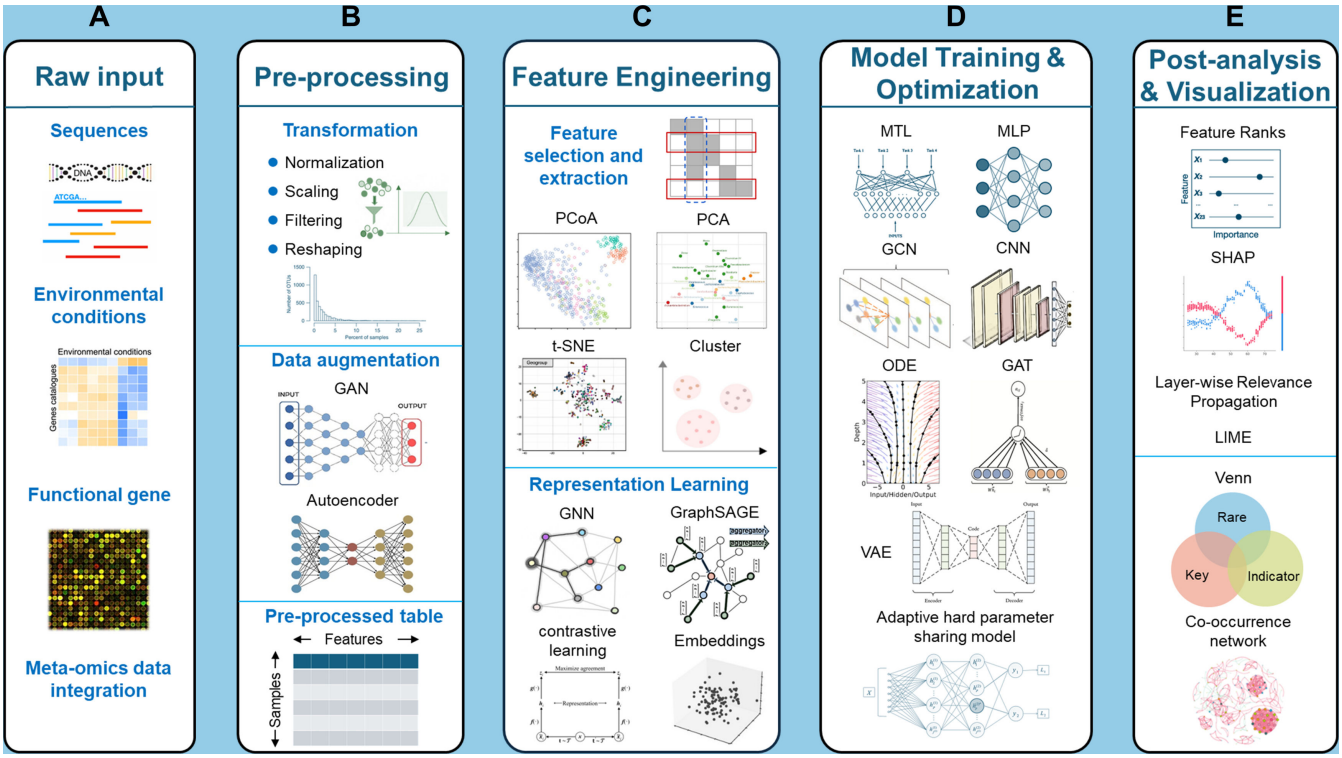

FIG 1 Overview of a unified deep learning framework for microbial taxa identification. (A) Raw input, define the range and resolution of the data. (B) Pre-processing, eliminate technical artifacts and improve comparability, preparing for data analysis. (C) Feature Engineering, convert data tables into structural feature information. (D) Model Training & Optimization, learn predictive mappings. (E) Post-analysis & Visualization, explain and validate. Abbreviations: MLP, multilayer perceptron; CNN, convolutional neural network; GCN, graph convolutional network; GAT, graph attention network; ODE, neural ordinary differential equation; VAE, variational autoencoder; SHAP, Shapley additive explanations; LRP, layer-wise relevance propagation; GAN, generative adversarial network; PCoA, principal coordinates analysis; PCA, principal component analysis; t-SNE, t-distributed stochastic neighbor embedding; CLR/ALR, centered/additive log-ratio; ZINB, zero-inflated negative binomial.

Additionally, the "black boxes" nature of DL models raises concerns about interpretability and reliability. Although tools such as Shapley additive explanations (SHAP) and local interpretable model-agnostic explanations (LIME) have been developed to elucidate DL predictions, ensuring their biological relevance remains an urgent task (138, 139). Furthermore, the computational cost and hardware requirements of DL models are typically high, raising concerns about scalability and deployment in real-world applications (140). Nevertheless, the rise of DL represents a paradigm shift in microbial taxa identification, providing unparalleled tools for analyzing complex data sets and modeling dynamic interactions within microbial ecosystems. As advances continue in data standardization, model interpretability, and computational infrastructure, DL holds the potential to drive transformative innovations in microbial research, spanning health, environmental sustainability, and biotechnology.

FUTURE DIRECTIONS OF MICROBIAL TAXA IDENTIFICATION BY DEEP LEARNING

Conceptual foundations

The application potential of DL in identifying specific microbial taxa has been initially demonstrated, showcasing its robust capabilities in analyzing complex microbial ecosystems (83, 86, 90). Despite these advances, these approaches are typically task-specific or tailored to individual ecological contexts, limiting their broader applicability. In addition, current DL applications are mainly focused on amplicon and shotgun metagenomic data sets, and comprehensive utilization of multi-omics data sets for

holistic microbial community studies remains to be further explored (33, 141). Against this backdrop, we propose a future direction for DL in identifying specific microbial taxa: a unified multitask framework based on multi-modal data integration that simultaneously identifies keystone, rare, and indicator taxa (Fig. 1). This unified framework aims to efficiently identify these three categories of taxa through standardized data processing and analytical pipelines, uncovering their interrelationships and ecological significance within microbial communities. By combining the ecological functions of keystone, the dynamic characteristics of rare taxa, and the environmental responses of indicator, this direction offers promising opportunities to explore the intricate complexities of microbial community structure and function.

Framework design

This section presents the modular blueprint of a unified framework for identifying ecologically important taxa in microbiomes. It focuses on the core design challenges and explains, in a layered manner, the construction logic and technical selection of each module. From data preprocessing and augmentation ("Data preprocessing and augmentation"), to compositional and hierarchy-aware representations ("Feature engineering and representation learning"), to multi-paradigm model architectures for prediction ("Multi-paradigm model architecture"), and finally to interpretability, validation, and communication with multi-omics integration ("Interpretability validation and optimization"), the aim is to provide a reusable analysis workflow. Additionally, to help policymakers and practitioners intuitively understand how deep-learning "black-box" models operate, we provide a plain-language explainer (see Table 2).

Data preprocessing and augmentation

Microbiome data preprocessing constitutes the foundational phase of framework construction (Fig. 1A and B). Because amplicon and metagenomic abundance tables are inherently sparse, subject to batch effects, and commonly zero-inflated, preprocessing should implement multi-stage correction and augmentation, including robust zero handling and a log-ratio transformation (centered log-ratio or additive log-ratio for targeted contrasts). For multi-omics analyses, preprocessing should also align sample identifiers and time points across data layers and construct layer-specific feature matrices with appropriate normalization, such as log-ratio features for compositional abundance layers and standard scaling procedures for expression and metabolite intensities. To address zero-inflation artifacts, the zero-inflated negative binomial (ZINB) model

TABLE 2 A plain-language guide to machine learning workflow in microbial ecology

Key aspect	Plain-language explanation
What data are used as input	Data on one or more categories of taxa and their abundances, optional genes and pathways, metabolites, and environmental or clinical variables
What does model training refer to	The model sees many labeled examples and adjusts internal weights so that similar patterns lead to similar outcomes
What does a prediction represent	For a new sample, the model computes a score or class that reflects patterns it has seen before
How do we know what matters	Metrics such as contribution scores (e.g., SHAP) and simple attention weights are aggregated to show which taxa, genes, or pathways were most influential for that sample
How do we validate results	By comparing potential (DNA), activity (RNA or proteins), and outputs (metabolites), and testing with independent test data, reporting uncertainty when confidence is low
What are the limitations of these models	These models reveal associations within the observed data. But they do not prove causation and they depend on data quality, sampling depth, and preprocessing choices

employs a mixture distribution framework to distinguish technical zeros (induced by insufficient sequencing depth) from true ecological zeros (reflecting actual taxa absence) (142, 143). The dual-component structure of this model has demonstrated superior performance in microbiome data analysis (144). For cross-data set integration scenarios, the ComBat algorithm based on empirical Bayesian estimation systematically eliminates technical biases introduced by heterogeneous sequencing platforms or experimental conditions through parameterization of batch-specific offsets (145), with its standardization process proving particularly advantageous for multicenter microbiome studies. To further mitigate underrepresentation of rare taxa, GANs synthesize ecologically plausible low-abundance samples via adversarial training between generator and discriminator modules (133, 146). In the human gut microbiome, GAN-based augmentation has produced realistic community profiles and improved downstream classification and cross-cohort generalization (147, 148). Taken together, this methodology has exhibited promising potential for data augmentation in gut microbiota rare taxa identification. The synergistic application of these preprocessing and augmentation steps produces harmonized, analysis-ready inputs for the representation, modeling, and interpretability stages that follow.

Feature engineering and representation learning

Extracting ecologically meaningful low-dimensional representations from high-dimensional OTU tables represents a pivotal challenge in framework design (Fig. 1C). Conventional approaches such as ANCOM-II (Analysis of Compositions of Microbiomes) employ log-ratio transformations coupled with multiple hypothesis testing to identify differentially abundant OTUs across experimental groups (149). However, their linear assumptions inadequately capture complex taxa interaction patterns (150, 151). Consequently, graph neural network-based embedding methods have emerged as a transformative solution. By constructing OTU co-occurrence networks through SparCC (Sparse Correlations for Compositional data) correlation analysis (45), the GraphSAGE (Graph Sample and Aggregate) model learns topological features of nodes within microbial interaction networks (152), where neighborhood aggregation mechanisms effectively encode synergistic or competitive relationships among taxa. Furthermore, contrastive learning techniques maximize representation consistency between OTU tables and environmental covariates (153), embedding niche theory into feature spaces to automatically associate microbial composition with host habitat characteristics. When multi-omics data are available, representation learning also benefits from layer-specific features and simple cross-layer links. Compositional abundance layers are expressed with log-ratio features, gene and pathway profiles are aggregated from metagenomes, and standardized expression or metabolite intensities are derived from metatranscriptomes, metaproteomes, and metabolomes. Two practical cross-omics features are the activity-to-abundance ratio, for example, RNA over DNA for a taxon or pathway which highlights low-abundance high-activity patterns relevant to rare taxa, and gradient-concordance vectors that summarize how taxa, pathway activity, and metabolites co-vary with an environmental or clinical gradient for indicator taxa. Joint or contrastive embeddings can then align the taxon space with pathway and metabolite spaces so that the learned manifold reflects ecological niches and functional constraints. The integration of these methodologies constructs discriminative and interpretable feature spaces that enable high-quality inputs for multitask modeling.

Multi-paradigm model architecture

The framework adopts a multi-paradigm fusion strategy at the architectural level to address ecosystem complexity (Fig. 1D). Multitask learning (MTL) extracts cross-task invariant features through shared encoders while employing task-specific modules for simultaneous prediction of keystone, rare, and indicator taxa (154). This hard parameter-sharing mechanism reduces overfitting risks while enhancing model discriminative capacity for ecological functional groups (155). With multi-omics input, the architecture

can use early fusion by concatenating standardized features or late fusion with per-omics encoders that feed a shared trunk. These distinct modules for keystone, rare, and indicator taxa allow attention or gating to reveal which data layer supports each decision. For dynamic community assembly processes, neural ordinary differential equations (Neural ODEs) integrate Lotka-Volterra dynamics into differentiable solvers (156), enabling temporal inference of taxa interaction networks from static snapshot data (157). In practice, trajectories can be constrained by short time series or pseudo-time, and the model can produce time-dependent interaction estimates together with dynamic importance profiles for taxa and pathways. In spatial-scale analyses, graph attention networks (GATs) utilize adaptive weight allocation mechanisms to capture heterogeneous interaction patterns within OTU co-occurrence networks (158). When multi-omics information is available, graphs can be extended to multi-relational forms in which edges encode co-occurrence, pathway linkage, or metabolite similarity, and attention weights highlight candidate interactions within and across layers. The coordinated design of these models equips the framework to resolve both steady-state ecosystem properties and simulate temporal community dynamics.

Interpretability validation and optimization

The framework systematically identifies taxa that span multiple categories (159). These overlapping taxa, which may simultaneously act as keystone, rare, and indicator taxa, represent a compelling area of study. To further elucidate the ecological significance of these taxa, the framework establishes a multimodal validation system that integrates explainable artificial intelligence (XAI) techniques (e.g., LRP [160], SHAP, or LIME). By integrating visualization tools such as Venn diagrams, co-occurrence networks, and heatmaps (Fig. 1E), this system visually translates computational patterns into ecological insights; specifically, the co-occurrence networks here are used to depict model-inferred non-linear dependencies (e.g., attention weights or interaction coefficients) rather than simple statistical correlations (83, 126, 127). When working with multi-omics data, attribution is first computed within each omics layer and then aggregated along the links of taxa, genes, pathways, and metabolites to form a cross-layer evidence profile for each taxa. SHAP quantifies contributions at the OTU level through cooperative game theory (116, 138), and the same procedure is applied to genes, pathways, and metabolites. Model importance can be linked to functional potential, realized activity, and pathway output in a multi-layer evidence graph. Taxon is prioritized when evidence is consistent across layers, such as keystone showing high model importance accompanied by pathway activity or metabolite changes that may affect stability. Rare taxa show low DNA abundance together with elevated RNA or protein activity or a distinct metabolite footprint. Indicator exhibits taxa, pathway activity, and metabolites that co-vary in consistent directions across cohorts along environmental or clinical gradients.

Synthetic community construction is used to validate model-predicted specific taxa and to assess community stability, establishing closed-loop feedback between computational predictions and laboratory observations (161). During model update and optimization, the adoption of dynamic curriculum learning strategies that progressively increase data complexity (e.g., incremental learning from dominant to rare taxa) (162) can improve the ability of the model to fit long-tailed distributions and training efficiency (163) without changing the model architecture. In addition, the temporal perspective can be incorporated to study microbial community assembly across different habitats and gradients (164), enabling systematic dissection of how the roles and interactions among keystone, rare, and indicator taxa evolve over time in response to environmental perturbations, disturbance regimes, or successional trajectories (59, 165). The systematic integration of these validation mechanisms contributes to a more comprehensive understanding of microbial ecosystem dynamics, thereby ensuring theoretical compatibility between framework outputs and ecological principles.

Core challenges and theoretical bottlenecks

The construction of current DL frameworks remains confronted with several cross-cutting challenges arising from fundamental tensions between biological system complexity, intrinsic data characteristics, and computational model capabilities. The primary challenge stems from the conflict between the heterogeneous, low signal-to-noise nature of microbiome data and the generalization capacity of models. Although existing preprocessing techniques (e.g., batch effect correction [166], data augmentation [167]) have partially mitigated data quality issues, the high dimensionality, sparsity, and multimodal characteristics of microbiome data continue to challenge model robustness. For instance, signals from low-abundance taxa are often faint and difficult to distinguish from technical noise. This signal overlap makes it challenging for models to discriminate between spurious statistical correlations and genuine biological associations (168). Furthermore, the lack of standardization across studies (e.g., inconsistencies in sampling protocols and sequencing depths) induces fundamental shifts in the underlying data distribution (feature space shifts) during pretrained model transfer (112, 169), which often causes models trained on one data set to fail when applied to another. Consequently, this necessitates the development of meta-learning architectures capable of projecting diverse data into universal representation spaces, where biological signals from diverse data sets of different studies are aligned and conserved.

A second pervasive challenge lies in balancing ecological interpretability with predictive accuracy across framework hierarchies. While DL models excel at capturing complex nonlinear relationships, their opaque decision-making mechanisms often create explanatory disjunctions between computational logic and ecological theory (170), particularly in causal inference of host-microbiome interactions (171). Current interpretability methods (e.g., feature importance ranking) provide OTU-level contribution analyses but fail to elucidate ecosystem-scale principles such as multi-taxa synergies or temporal succession dynamics (172, 173). A more fundamental tension arises from the inability of current discriminative objective-driven DL paradigms to holistically model emergent properties of microbiome systems (e.g., community stability, functional redundancy), demanding hybrid architecture-inference systems that integrate first principles of ecology.

Finally, the mismatch between computational resource constraints and the inherent complexity and multi-scale nature of biological systems limits the practical boundaries of frameworks. Microbiome studies frequently involve ultrahigh-dimensional OTU features ($\geq 10^4$ dimensions) (174) and multiple temporal scales (ranging from minute-level metabolic dynamics to decadal ecological succession) (175), posing dual challenges to model memory efficiency and long-term dependency modeling. Traditional DL approaches face a dilemma between the "curse of dimensionality" (i.e., too many microbial features with too few samples lead to a decrease in model performance) and "over-smoothing effects" (i.e., where distinct nodes in a graph neural network become too similar to distinguish) when addressing such problems—increasing network capacity enhances feature discrimination but exacerbates overfitting on scarce annotated data (118). Breakthroughs in physics-informed neural networks (PINNs) and sparse training strategies are urgently required to overcome computational bottlenecks (176, 177). Simultaneously, orders-of-magnitude disparities between synthetic biology techniques (for wet-lab validation), microbial cultivation cycles, and computational model iteration speeds hinder efficient coordination of "dry-wet closed loops" (161, 178, 179). To overcome these disparities, employing active learning strategies to iteratively select the most informative experimental validations can minimize the wet-lab workload. Anyway, overcoming these technical challenges is key to fully unlocking the potential of DL in advancing microbial ecology and broader biological research.

FUTURE PROSPECTS

Future developments should prioritize the integration of multi-modal data sources to capture the complete dynamics of microbial communities across genetic, transcriptional, and metabolic levels, thereby enriching ecological insights (180). The development of DL frameworks must also emphasize enhancing model interpretability by incorporating causal inference capabilities. The inclusion of dynamic feedback loops could provide a more accurate means to simulate temporal variations in microbial communities (157), capturing their responses to long-term evolutionary dynamics, including successional processes and environmental disturbances. Furthermore, to improve the accessibility and applicability of DL frameworks, it is essential to design more efficient and lightweight computational architectures. Simultaneously, advancing community-wide standardization of data formats and ensuring transparency in framework design (e.g., open source, documentation, and clearly defined application boundaries) will be critical for fostering consistency and reproducibility in future research (181, 182).

CONCLUSION

Deep learning is driving a transformative shift in microbial taxa identification from communities, offering novel perspectives through automated feature extraction and the modeling of complex nonlinear interactions. This review examines the progression of identifying specific microbial taxa, from traditional methodologies to the advancements brought by deep learning, emphasizing the paradigm shift enabled by these innovations. The categories of keystone, rare, and indicator taxa are dynamic and context-dependent, influenced by environmental gradients, community interactions, and functional redundancy. A unified deep learning framework was proposed by integrating the identification of keystone, rare, and indicator taxa into a cohesive computational workflow. Bridging the gap between microbial ecological theory and data-driven computational approaches using multi-task learning to model shared ecological patterns while preserving task-specific nuances. By harmonizing deep learning with ecological theory, it offers a powerful tool to unravel the complexities of microbial communities, with applications spanning environmental monitoring, biotechnology, and human health. Future work will focus on empirical validation across biomes and the development of user-friendly pipelines for broader adoption.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (42577239 and 42277193) and the Open Fund of Key Laboratory of Mine Ecological Effects and Systematic Restoration, Ministry of Natural Resources (MEER-2024-10).

This article was only translated and checked for English grammar using ChatGPT and DeepSeek, and the authors take full responsibility for the content of this paper.

AUTHOR AFFILIATIONS

[1]Institute of Natural Resources Survey, China University of Geosciences, Wuhan, China
[2]MOE Key Laboratory of Groundwater Quality and Health, School of Environmental Studies, China University of Geosciences, Wuhan, China
[3]Key Laboratory of Mine Ecological Effects and Systematic Restoration, Ministry of Natural Resources, Beijing, China

AUTHOR ORCIDs

Jingkang Zhang http://orcid.org/0009-0000-0900-2940
Hongmei Wang http://orcid.org/0000-0001-7621-7810
Liyuan Ma http://orcid.org/0000-0001-7351-836X

FUNDING

Funder	Grant(s)	Author(s)
National Natural Science Foundation of China	42577239, 42277193	Liyuan Ma
Open Fund of Key Laboratory of Mine Ecological Effects and Systematic Restoration, Ministry of Natural Resources	MEER-2024-10	Liyuan Ma

AUTHOR CONTRIBUTIONS

Jingkang Zhang, Investigation, Methodology, Resources, Visualization, Writing – original draft, Writing – review and editing | Xingjie Wang, Conceptualization, Formal analysis, Investigation, Methodology, Writing – review and editing | Di Wang, Data curation, Investigation, Resources, Visualization | Zikui Zheng, Formal analysis, Investigation, Validation, Writing – review and editing | Hongmei Wang, Conceptualization, Formal analysis, Methodology, Resources, Supervision | Liyuan Ma, Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Writing – review and editing

REFERENCES

1. Fuhrman JA. 2009. Microbial community structure and its functional implications. Nature 459:193–199. https://doi.org/10.1038/nature08058
2. Falkowski PG, Fenchel T, Delong EF. 2008. The microbial engines that drive Earth's biogeochemical cycles. Science 320:1034–1039. https://doi.org/10.1126/science.1153213
3. Steen AD, Crits-Christoph A, Carini P, DeAngelis KM, Fierer N, Lloyd KG, Cameron Thrash J. 2019. High proportions of bacteria and archaea across most biomes remain uncultured. ISME J 13:3126–3130. https://doi.org/10.1038/s41396-019-0484-y
4. Lagier J-C, Khelaifia S, Alou MT, Ndongo S, Dione N, Hugon P, Caputo A, Cadoret F, Traore SI, Seck EH, et al. 2016. RETRACTED ARTICLE: Culture of previously uncultured members of the human gut microbiota by culturomics. Nat Microbiol 1:16203. https://doi.org/10.1038/nmicrobiol.2016.203
5. Forster SC, Kumar N, Anonye BO, Almeida A, Viciani E, Stares MD, Dunn M, Mkandawire TT, Zhu A, Shao Y, Pike LJ, Louie T, Browne HP, Mitchell AL, Neville BA, Finn RD, Lawley TD. 2019. A human gut bacterial genome and culture collection for improved metagenomic analyses. Nat Biotechnol 37:186–192. https://doi.org/10.1038/s41587-018-0009-7
6. Lok C. 2015. Mining the microbial dark matter. Nature 522:270–273. https://doi.org/10.1038/522270a
7. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu WT, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499:431–437. https://doi.org/10.1038/nature12352
8. Amann RI, Ludwig W, Schleifer KH. 1995. Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. Microbiol Rev 59:143–169. https://doi.org/10.1128/mr.59.1.143-169.1995
9. Lloyd KG, Steen AD, Ladau J, Yin J, Crosby L. 2018. Phylogenetically novel uncultured microbial cells dominate earth microbiomes. mSystems 3:e00055-18. https://doi.org/10.1128/mSystems.00055-18
10. Janda JM, Abbott SL. 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol 45:2761–2764. https://doi.org/10.1128/JCM.01228-07
11. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37–43. https://doi.org/10.1038/nature02340
12. Wilmes P, Bond PL. 2006. Metaproteomics: studying functional gene expression in microbial ecosystems. Trends Microbiol 14:92–97. https://doi.org/10.1016/j.tim.2005.12.006
13. Ma L, Wang H, Wu J, Wang Y, Zhang D, Liu X. 2019. Metatranscriptomics reveals microbial adaptation and resistance to extreme environment coupling with bioleaching performance. Bioresour Technol 280:9–17. https://doi.org/10.1016/j.biortech.2019.01.117
14. Lloyd-Price J, Abu-Ali G, Huttenhower C. 2016. The healthy human microbiome. Genome Med 8:51. https://doi.org/10.1186/s13073-016-0307-y
15. Banerjee S, Walder F, Büchi L, Meyer M, Held AY, Gattinger A, Keller T, Charles R, van der Heijden MGA. 2019. Agricultural intensification reduces microbial network complexity and the abundance of keystone taxa in roots. ISME J 13:1722–1736. https://doi.org/10.1038/s41396-019-0383-2
16. Jousset A, Bienhold C, Chatzinotas A, Gallien L, Gobet A, Kurm V, Küsel K, Rillig MC, Rivett DW, Salles JF, van der Heijden MGA, Youssef NH, Zhang X, Wei Z, Hol WHG. 2017. Where less may be more: how the rare biosphere pulls ecosystems strings. ISME J 11:853–862. https://doi.org/10.1038/ismej.2016.174
17. Fierer N, Lauber CL, Ramirez KS, Zaneveld J, Bradford MA, Knight R. 2012. Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. ISME J 6:1007–1017. https://doi.org/10.1038/ismej.2011.159
18. Lynch MDJ, Neufeld JD. 2015. Ecology and exploration of the rare biosphere. Nat Rev Microbiol 13:217–229. https://doi.org/10.1038/nrmicro3400
19. Banerjee S, Schlaeppi K, van der Heijden MGA. 2018. Keystone taxa as drivers of microbiome structure and functioning. Nat Rev Microbiol 16:567–576. https://doi.org/10.1038/s41579-018-0024-1
20. Herren CM, McMahon KD. 2018. Keystone taxa predict compositional change in microbial communities. Environ Microbiol 20:2207–2217. https://doi.org/10.1111/1462-2920.14257
21. Wagg C, Schlaeppi K, Banerjee S, Kuramae EE, van der Heijden MGA. 2019. Fungal-bacterial diversity and microbiome complexity predict ecosystem functioning. Nat Commun 10:4841. https://doi.org/10.1038/s41467-019-12798-y
22. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E, Ajami NJ, Bonham KS, Brislawn CJ, et al. 2019. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature 569:655–662. https://doi.org/10.1038/s41586-019-1237-9
23. The Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207–214. https://doi.org/10.1038/nature11234
24. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, et al. 2015. Ocean plankton. Structure and function of the global ocean microbiome. Science 348:1261359. https://doi.org/10.1126/science.1261359
25. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, et al. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457–463. https://doi.org/10.1038/nature24621

26. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall L-I, McDonald D, Melnik AV, Morton JT, Navas J, Quinn RA, Sanders JG, Swafford AD, Thompson LR, Tripathi A, Xu ZZ, Zaneveld JR, Zhu Q, Caporaso JG, Dorrestein PC. 2018. Best practices for analysing microbiomes. Nat Rev Microbiol 16:410–422. https://doi.org/10.1038/s41579-018-0029-9
27. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, Brady A, Creasy HH, McCracken C, Giglio MG, McDonald D, Franzosa EA, Knight R, White O, Huttenhower C. 2017. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 550:61–66. https://doi.org/10.1038/nature23889
28. Bäckhed F, Fraser CM, Ringel Y, Sanders ME, Sartor RB, Sherman PM, Versalovic J, Young V, Finlay BB. 2012. Defining a healthy human gut microbiome: current concepts, future directions, and clinical applications. Cell Host Microbe 12:611–622. https://doi.org/10.1016/j.chom.2012.10.012
29. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. 2017. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol 35:833–844. https://doi.org/10.1038/nbt.3935
30. Faust K, Raes J. 2012. Microbial interactions: from networks to models. Nat Rev Microbiol 10:538–550. https://doi.org/10.1038/nrmicro2832
31. Kostic AD, Gevers D, Siljander H, Vatanen T, Hyötyläinen T, Hämäläinen A-M, Peet A, Tillmann V, Pöhö P, Mattila I, Lähdesmäki H, Franzosa EA, Vaarala O, de Goffau M, Harmsen H, Ilonen J, Virtanen SM, Clish CB, Orešič M, Huttenhower C, Knip M, Xavier RJ, DIABIMMUNE Study Group. 2015. The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. Cell Host Microbe 17:260–273. https://doi.org/10.1016/j.chom.2015.01.001
32. Lee JY, Sadler NC, Egbert RG, Anderton CR, Hofmockel KS, Jansson JK, Song HS. 2020. Deep learning predicts microbial interactions from self-organized spatiotemporal patterns. Comput Struct Biotechnol J 18:1259–1269. https://doi.org/10.1016/j.csbj.2020.05.023
33. Hernández Medina R, Kutuzova S, Nielsen KN, Johansen J, Hansen LH, Nielsen M, Rasmussen S. 2022. Machine learning and deep learning applications in microbiome research. ISME Commun 2:98. https://doi.org/10.1038/s43705-022-00182-9
34. Ma L, Huang X, Wang H, Yun Y, Cheng X, Liu D, Lu X, Qiu X. 2021. Microbial interactions drive distinct taxonomic and potential metabolic responses to habitats in karst cave ecosystem. Microbiol Spectr 9:e01152-21. https://doi.org/10.1128/Spectrum.01152-21
35. 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. 2012. Human gut microbiome viewed across age and geography. Nature 486:222–227. https://doi.org/10.1038/nature11053
36. Browne HP, Forster SC, Anonye BO, Kumar N, Neville BA, Stares MD, Goulding D, Lawley TD. 2016. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. Nature 533:543–546. https://doi.org/10.1038/nature17645
37. Zou Y, Xue W, Luo G, Deng Z, Qin P, Guo R, Sun H, Xia Y, Liang S, Dai Y, et al. 2019. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nat Biotechnol 37:179–185. https://doi.org/10.1038/s41587-018-0008-8
38. Paine RT. 1969. A note on trophic complexity and community stability. Am Nat 103:91–93. https://doi.org/10.1086/282586
39. Mills LS, Doak DF. 1993. The keystone-species concept in ecology and conservation. Bioscience 43:219–224. https://doi.org/10.2307/1312122
40. Berry D, Widder S. 2014. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. Front Microbiol 5:219. https://doi.org/10.3389/fmicb.2014.00219
41. Barberán A, Bates ST, Casamayor EO, Fierer N. 2012. Using network analysis to explore co-occurrence patterns in soil microbial communities. ISME J 6:343–351. https://doi.org/10.1038/ismej.2011.119
42. Mao J, Zheng Z, Ma L, Wang H, Wang X, Zhu F, Xue S, Srivastava P, Sapsford DJ. 2024. Polymetallic contamination drives indigenous microbial community assembly dominated by stochastic processes at Pb-Zn smelting sites. Sci Total Environ 947:174575. https://doi.org/10.1016/j.scitotenv.2024.174575
43. Vandeputte D, De Commer L, Tito RY, Kathagen G, Sabino J, Vermeire S, Faust K, Raes J. 2021. Temporal variability in quantitative human gut microbiome profiles and implications for clinical research. Nat Commun 12:6740. https://doi.org/10.1038/s41467-021-27098-7
44. Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, Xia LC, Xu ZZ, Ursell L, Alm EJ, Birmingham A, Cram JA, Fuhrman JA, Raes J, Sun F, Zhou J, Knight R. 2016. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. ISME J 10:1669–1681. https://doi.org/10.1038/ismej.2015.235
45. Friedman J, Alm EJ. 2012. Inferring correlation networks from genomic survey data. PLoS Comput Biol 8:e1002687. https://doi.org/10.1371/journal.pcbi.1002687
46. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. 2015. Sparse and compositionally robust inference of microbial ecological networks. PLoS Comput Biol 11:e1004226. https://doi.org/10.1371/journal.pcbi.1004226
47. Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci USA 103:12115–12120. https://doi.org/10.1073/pnas.0605127103
48. Salazar G, Cornejo-Castillo FM, Benítez-Barrios V, Fraile-Nuez E, Álvarez-Salgado XA, Duarte CM, Gasol JM, Acinas SG. 2016. Global diversity and biogeography of deep-sea pelagic prokaryotes. ISME J 10:596–608. https://doi.org/10.1038/ismej.2015.137
49. Mouillot D, Bellwood DR, Baraloto C, Chave J, Galzin R, Harmelin-Vivien M, Kulbicki M, Lavergne S, Lavorel S, Mouquet N, Paine CET, Renaud J, Thuiller W. 2013. Rare species support vulnerable functions in high-diversity ecosystems. PLoS Biol 11:e1001569. https://doi.org/10.1371/journal.pbio.1001569
50. Leitão RP, Zuanon J, Villéger S, Williams SE, Baraloto C, Fortunel C, Mendonça FP, Mouillot D. 2016. Rare species contribute disproportionately to the functional structure of species assemblages. Proc Biol Sci 283:20160084. https://doi.org/10.1098/rspb.2016.0084
51. Hillebrand H, Blasius B, Borer ET, Chase JM, Downing JA, Eriksson BK, Filstrup CT, Harpole WS, Hodapp D, Larsen S, Lewandowska AM, Seabloom EW, Van de Waal DB, Ryabov AB. 2018. Biodiversity change is uncoupled from species richness trends: consequences for conservation and monitoring. J Appl Ecol55:169–184. https://doi.org/10.1111/1365-2664.12959
52. Pedler BE, Aluwihare LI, Azam F. 2014. Single bacterial strain capable of significant contribution to carbon cycling in the surface ocean. Proc Natl Acad Sci USA 111:7202–7207. https://doi.org/10.1073/pnas.1401887111
53. Shade A, Jones SE, Caporaso JG, Handelsman J, Knight R, Fierer N, Gilbert JA. 2014. Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. mBio 5:e01371-14. https://doi.org/10.1128/mBio.01371-14
54. Dawson W, Hör J, Egert M, van Kleunen M, Pester M. 2017. A small number of low-abundance bacteria dominate plant species-specific responses during rhizosphere colonization. Front Microbiol 8:975. https://doi.org/10.3389/fmicb.2017.00975
55. Fuentes S, Barra B, Caporaso JG, Seeger M. 2016. From rare to dominant: a fine-tuned soil bacterial bloom during petroleum hydrocarbon bioremediation. Appl Environ Microbiol 82:888–896. https://doi.org/10.1128/AEM.02625-15
56. Jia X, Dini-Andreote F, Falcão Salles J. 2018. Community assembly processes of the microbial rare biosphere. Trends Microbiol 26:738–747. https://doi.org/10.1016/j.tim.2018.02.011
57. Delgado-Baquerizo M, Oliverio AM, Brewer TE, Benavent-González A, Eldridge DJ, Bardgett RD, Maestre FT, Singh BK, Fierer N. 2018. A global atlas of the dominant bacteria found in soil. Science 359:320–325. https://doi.org/10.1126/science.aap9516
58. Logares R, Audic S, Bass D, Bittner L, Boutte C, Christen R, Claverie J-M, Decelle J, Dolan JR, Dunthorn M, et al. 2014. Patterns of rare and abundant marine microbial eukaryotes. Curr Biol 24:813–821. https://doi.org/10.1016/j.cub.2014.02.050
59. Nemergut DR, Schmidt SK, Fukami T, O'Neill SP, Bilinski TM, Stanish LF, Knelman JE, Darcy JL, Lynch RC, Wickey P, Ferrenberg S. 2013. Patterns and processes of microbial community assembly. Microbiol Mol Biol Rev 77:342–356. https://doi.org/10.1128/MMBR.00051-12
60. Blaser MJ, Cardon ZG, Cho MK, Dangl JL, Donohue TJ, Green JL, Knight R, Maxon ME, Northen TR, Pollard KS, Brodie EL. 2016. Toward a predictive understanding of earth's microbiomes to address 21st century challenges. mBio 7:e00714-16. https://doi.org/10.1128/mBio.00714-16
61. Callahan BJ, McMurdie PJ, Holmes SP. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J 11:2639–2643. https://doi.org/10.1038/ismej.2017.119

62. Hillmann B, Al-Ghalith GA, Shields-Cutler RR, Zhu Q, Gohl DM, Beckman KB, Knight R, Knights D. 2018. Evaluating the information content of shallow shotgun metagenomics. mSystems 3:e00069-18. https://doi.org/10.1128/mSystems.00069-18
63. Siddig AAH, Ellison AM, Ochs A, Villar-Leeman C, Lau MK. 2016. How do ecologists select and use indicator species to monitor ecological change? Insights from 14 years of publication in *Ecological Indicators*. Ecol Indic 60:223–230. https://doi.org/10.1016/j.ecolind.2015.06.036
64. Fierer N, Strickland MS, Liptzin D, Bradford MA, Cleveland CC. 2009. Global patterns in belowground communities. Ecol Lett 12:1238–1249. https://doi.org/10.1111/j.1461-0248.2009.01360.x
65. Allison SD, Martiny JBH. 2008. Colloquium paper: resistance, resilience, and redundancy in microbial communities. Proc Natl Acad Sci USA 105:11512–11519. https://doi.org/10.1073/pnas.0801925105
66. Gilbert JA, Jansson JK, Knight R. 2014. The Earth Microbiome project: successes and aspirations. BMC Biol 12:69. https://doi.org/10.1186/s12915-014-0069-1
67. Legendre P, Legendre L. 2012. Numerical ecology. Vol. 24. Elsevier.
68. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. 2011. Metagenomic biomarker discovery and explanation. Genome Biol 12:R60. https://doi.org/10.1186/gb-2011-12-6-r60
69. Dufrêne M, Legendre P. 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecol Monogr 67:345–366. https://doi.org/10.1890/0012-9615(1997)067[0345:SAAIST]2.0.CO;2
70. Zhou J, Deng Y, Luo F, He Z, Yang Y. 2011. Phylogenetic molecular ecological network of soil microbial communities in response to elevated CO_2. mBio 2:e00122-11. https://doi.org/10.1128/mBio.00122-11
71. Baker ME, King RS. 2010. A new method for detecting and interpreting biodiversity and ecological community thresholds. Methods Ecol Evol 1:25–37. https://doi.org/10.1111/j.2041-210X.2009.00007.x
72. Taylor JM, DeVilbiss SE, Hicks M. 2023. Using taxa-based approaches to delineate stream macroinvertebrate assemblage responses to stressor gradients in modified alluvial agroecosystems. Ecol Indic 153:110377. https://doi.org/10.1016/j.ecolind.2023.110377
73. Sultana J, Tibby J, Recknagel F, Maxwell S, Goonan P. 2020. Comparison of two commonly used methods for identifying water quality thresholds in freshwater ecosystems using field and synthetic data. Sci Total Environ 724:137999. https://doi.org/10.1016/j.scitotenv.2020.137999
74. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. 2017. Microbiome datasets are compositional: and this is not optional. Front Microbiol 8:2224. https://doi.org/10.3389/fmicb.2017.02224
75. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, Lipson KS, Knight R, Caporaso JG, Segata N, Huttenhower C. 2018. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods 15:962–968. https://doi.org/10.1038/s41592-018-0176-y
76. Walsh C, Stallard-Olivera E, Fierer N. 2024. Nine (not so simple) steps: a practical guide to using machine learning in microbial ecology. mBio 15:e02050-23. https://doi.org/10.1128/mbio.02050-23
77. Weis CV, Jutzeler CR, Borgwardt K. 2020. Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review. Clin Microbiol Infect 26:1310–1317. https://doi.org/10.1016/j.cmi.2020.03.014
78. Huang Y, Sheth RU, Zhao S, Cohen LA, Dabaghi K, Moody T, Sun Y, Ricaurte D, Richardson M, Velez-Cortes F, Blazejewski T, Kaufman A, Ronda C, Wang HH. 2023. High-throughput microbial culturomics using automation and machine learning. Nat Biotechnol 41:1424–1433. https://doi.org/10.1038/s41587-023-01674-2
79. Knights D, Costello EK, Knight R. 2011. Supervised classification of human microbiota. FEMS Microbiol Rev 35:343–359. https://doi.org/10.1111/j.1574-6976.2010.00251.x
80. Levine NM, Alexander H, Bertrand EM, Coles VJ, Dutkiewicz S, Leles SG, Zakem EJ. 2025. Microbial ecology to ocean carbon cycling: from genomes to numerical models. Annu Rev Earth Planet Sci 53:595–624. https://doi.org/10.1146/annurev-earth-040523-020630
81. Cheng Z, Zheng Q, Shi J, He Y, Yang X, Huang X, Wu L, Xu J. 2023. Metagenomic and machine learning-aided identification of biomarkers driving distinctive Cd accumulation features in the root-associated microbiome of two rice cultivars. ISME Commun 3:14. https://doi.org/10.1038/s43705-023-00213-z
82. Xun W, Liu Y, Li W, Ren Y, Xiong W, Xu Z, Zhang N, Miao Y, Shen Q, Zhang R. 2021. Specialized metabolic functions of keystone taxa sustain soil microbiome stability. Microbiome 9:35. https://doi.org/10.1186/s40168-020-00985-9
83. Wang XW, Sun Z, Jia H, Michel-Mata S, Angulo MT, Dai L, He X, Weiss ST, Liu YY. 2024. Identifying keystone species in microbial communities using deep learning. Nat Ecol Evol 8:22–31. https://doi.org/10.1038/s41559-023-02250-2
84. Wu L, Wang XW, Tao Z, Wang T, Zuo W, Zeng Y, Liu YY, Dai L. 2024. Data-driven prediction of colonization outcomes for complex microbial communities. Nat Commun 15:2406. https://doi.org/10.1038/s41467-024-46766-y
85. Fisher CK, Mehta P. 2014. Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. PLoS One 9:e102451. https://doi.org/10.1371/journal.pone.0102451
86. Pascoal F, Branco P, Torgo L, Costa R, Magalhães C. 2025. Definition of the microbial rare biosphere through unsupervised machine learning. Commun Biol 8:544. https://doi.org/10.1038/s42003-025-07912-4
87. Thompson J, Johansen R, Dunbar J, Munsky B. 2019. Machine learning to predict microbial community functions: an analysis of dissolved organic carbon from litter decomposition. PLoS One 14:e0215502. https://doi.org/10.1371/journal.pone.0215502
88. Zheng J, Sun Q, Zhang M, Liu C, Su Q, Zhang L, Xu Z, Lu W, Ching J, Tang W, Cheung CP, Hamilton AL, Wilson O'Brien AL, Wei SC, Bernstein CN, Rubin DT, Chang EB, Morrison M, Kamm MA, Chan FKL, Zhang J, Ng SC. 2024. Noninvasive, microbiome-based diagnosis of inflammatory bowel disease. Nat Med 30:3555–3567. https://doi.org/10.1038/s41591-024-03280-4
89. Liu B, Sträuber H, Saraiva J, Harms H, Silva SG, Kasmanas JC, Kleinsteuber S, Nunes da Rocha U. 2022. Machine learning-assisted identification of bioindicators predicts medium-chain carboxylate production performance of an anaerobic mixed culture. Microbiome 10:48. https://doi.org/10.1186/s40168-021-01219-2
90. Tsai Y, Baldwin SA, Gopaluni B. 2021. Identifying indicator species in ecological habitats using Deep Optimal Feature Learning. PLoS One 16:e0256782. https://doi.org/10.1371/journal.pone.0256782
91. Wirbel J, Zych K, Essex M, Karcher N, Kartal E, Salazar G, Bork P, Sunagawa S, Zeller G. 2021. Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. Genome Biol 22:93. https://doi.org/10.1186/s13059-021-02306-1
92. DiMucci D, Kon M, Segrè D. 2018. Machine learning reveals missing edges and putative interaction mechanisms in microbial ecosystem networks. mSystems 3:e00181-18. https://doi.org/10.1128/mSystems.00181-18
93. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, MacNair CR, French S, Carfrae LA, Bloom-Ackermann Z, Tran VM, Chiappino-Pepe A, Badran AH, Andrews IW, Chory EJ, Church GM, Brown ED, Jaakkola TS, Barzilay R, Collins JJ. 2020. A deep learning approach to antibiotic discovery. Cell 180:688–702. https://doi.org/10.1016/j.cell.2020.01.021
94. Topçuoğlu BD, Lesniak NA, Ruffin MT IV, Wiens J, Schloss PD. 2020. A framework for effective application of machine learning to microbiome-based classification problems. mBio 11:e00434-20. https://doi.org/10.1128/mBio.00434-20
95. Li Y, Xu Z, Han W, Cao H, Umarov R, Yan A, Fan M, Chen H, Duarte CM, Li L, Ho PL, Gao X. 2021. HMD-ARG: hierarchical multi-task deep learning for annotating antibiotic resistance genes. Microbiome 9:40. https://doi.org/10.1186/s40168-021-01002-3
96. Baranwal M, Clark RL, Thompson J, Sun Z, Hero AO, Venturelli OS. 2022. Recurrent neural networks enable design of multifunctional synthetic human gut microbiome dynamics. eLife 11:e73870. https://doi.org/10.7554/eLife.73870
97. Zou B, Wang J, Ding Y, Zhang Z, Huang Y, Fang X, Cheung KC, See S, Zhang L. 2024. A multi-modal deep language model for contaminant removal from metagenome-assembled genomes. Nat Mach Intell 6:1245–1255. https://doi.org/10.1038/s42256-024-00908-5
98. Pan S, Zhu C, Zhao XM, Coelho LP. 2022. A deep siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments. Nat Commun 13:2326. https://doi.org/10.1038/s41467-022-29843-y
99. Mo Y, Bier R, Li X, Daniels M, Smith A, Yu L, Kan J. 2024. Agricultural practices influence soil microbiome assembly and interactions at

different depths identified by machine learning. Commun Biol 7:1349. https://doi.org/10.1038/s42003-024-07059-8
100. Rodriguez PA, Rothballer M, Chowdhury SP, Nussbaumer T, Gutjahr C, Falter-Braun P. 2019. Systems biology of plant-microbiome interactions. Mol Plant 12:804–821. https://doi.org/10.1016/j.molp.2019.05.006
101. Kursa MB, Rudnicki WR. 2010. Feature selection with the Boruta package. J Stat Softw 36:13. https://doi.org/10.18637/jss.v036.i11
102. Clark RL, Connors BM, Stevenson DM, Hromada SE, Hamilton JJ, Amador-Noguez D, Venturelli OS. 2021. Design of synthetic human gut microbiome assembly and butyrate production. Nat Commun 12:3254. https://doi.org/10.1038/s41467-021-22938-y
103. Pichler M, Hartig F. 2023. Machine learning and deep learning—A review for ecologists. Methods Ecol Evol 14:994–1016. https://doi.org/10.1111/2041-210X.14061
104. Altmann A, Toloşi L, Sander O, Lengauer T. 2010. Permutation importance: a corrected feature importance measure. Bioinformatics 26:1340–1347. https://doi.org/10.1093/bioinformatics/btq134
105. Goodswen SJ, Barratt JLN, Kennedy PJ, Kaufer A, Calarco L, Ellis JT. 2021. Machine learning and applications in microbiology. FEMS Microbiol Rev 45:fuab015. https://doi.org/10.1093/femsre/fuab015
106. Pust M-M, Tümmler B. 2022. Bacterial low-abundant taxa are key determinants of a healthy airway metagenome in the early years of human life. Comput Struct Biotechnol J 20:175–186. https://doi.org/10.1016/j.csbj.2021.12.008
107. Ishiya K, Aburatani S. 2019. Outlier detection for minor compositional variations in taxonomic abundance data. Appl Sci (Basel) 9:1355. https://doi.org/10.3390/app9071355
108. Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, Fleck JS, Voigt AY, Palleja A, Ponnudurai R, et al. 2019. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med 25:679–689. https://doi.org/10.1038/s41591-019-0406-6
109. Hermans SM, Buckley HL, Case BS, Curran-Cournane F, Taylor M, Lear G. 2020. Using soil bacterial communities to predict physico-chemical variables and soil quality. Microbiome 8:79. https://doi.org/10.1186/s40168-020-00858-1
110. Marcos-Zambrano LJ, Karaduzovic-Hadziabdic K, Loncar Turukalo T, Przymus P, Trajkovik V, Aasmets O, Berland M, Gruca A, Hasic J, Hron K, et al. 2021. Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. Front Microbiol 12:634511. https://doi.org/10.3389/fmicb.2021.634511
111. Vujkovic-Cvijin I, Sklar J, Jiang L, Natarajan L, Knight R, Belkaid Y. 2020. Host variables confound gut microbiota studies of human disease. Nature 587:448–454. https://doi.org/10.1038/s41586-020-2881-9
112. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. 2016. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Comput Biol 12:e1004977. https://doi.org/10.1371/journal.pcbi.1004977
113. Statnikov A, Henaff M, Narendra V, Konganti K, Li Z, Yang L, Pei Z, Blaser MJ, Aliferis CF, Alekseyenko AV. 2013. A comprehensive evaluation of multicategory classification methods for microbiomic data. Microbiome 1:11. https://doi.org/10.1186/2049-2618-1-11
114. Huang S, Chaudhary K, Garmire LX. 2017. More is better: recent progress in multi-omics data integration methods. Front Genet 8:84. https://doi.org/10.3389/fgene.2017.00084
115. Rudin C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1:206–215. https://doi.org/10.1038/s42256-019-0048-x
116. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI. 2020. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2:56–67. https://doi.org/10.1038/s42256-019-0138-9
117. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. 2019. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet 20:389–403. https://doi.org/10.1038/s41576-019-0122-6
118. Przymus P, Rykaczewski K, Martín-Segura A, Truu J, Carrillo De Santa Pau E, Kolev M, Naskinova I, Gruca A, Sampri A, Frohme M, Nechyporenko A. 2024. Deep learning in microbiome analysis: a comprehensive review of neural network models. Front Microbiol 15:1516667. https://doi.org/10.3389/fmicb.2024.1516667
119. Shen WX, Liang SR, Jiang YY, Chen YZ. 2023. Enhanced metagenomic deep learning for disease prediction and consistent signature recognition by restructured microbiome 2D representations. Patterns (N Y) 4:100658. https://doi.org/10.1016/j.patter.2022.100658
120. Vu D, Groenewald M, Verkley G. 2020. Convolutional neural networks improve fungal classification. Sci Rep 10:12628. https://doi.org/10.1038/s41598-020-69245-y
121. Roy G, Prifti E, Belda E, Zucker JD. 2024. Deep learning methods in metagenomics: a review. Microb Genom 10:001231. https://doi.org/10.1099/mgen.0.001231
122. Zhu Q, Jiang X, Zhu Q, Pan M, He T. 2019. Graph embedding deep learning guides microbial biomarkers' identification. Front Genet 10:1182. https://doi.org/10.3389/fgene.2019.01182
123. Zhou Z, Tran PQ, Breister AM, Liu Y, Kieft K, Cowley ES, Karaoz U, Anantharaman K. 2022. METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. Microbiome 10:33. https://doi.org/10.1186/s40168-021-01213-8
124. Vaishnav ED, de Boer CG, Molinet J, Yassour M, Fan L, Adiconis X, Thompson DA, Levin JZ, Cubillos FA, Regev A. 2022. The evolution, evolvability and engineering of gene regulatory DNA. Nature 603:455–463. https://doi.org/10.1038/s41586-022-04506-6
125. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. 2017. Attention is all you need
126. Zhao Z, Woloszynek S, Agbavor F, Mell JC, Sokhansanj BA, Rosen GL. 2021. Learning, visualizing and exploring 16S rRNA structure using an attention-based deep neural network. PLoS Comput Biol 17:e1009345. https://doi.org/10.1371/journal.pcbi.1009345
127. Melnyk K, Weimann K, Conrad TOF. 2023. Understanding microbiome dynamics via interpretable graph representation learning. Sci Rep 13:2058. https://doi.org/10.1038/s41598-023-29098-7
128. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. Nat Methods 11:1144–1146. https://doi.org/10.1038/nmeth.3103
129. Johnson JS, Spakowicz DJ, Hong B-Y, Petersen LM, Demkowicz P, Chen L, Leopold SR, Hanson BM, Agresta HO, Gerstein M, Sodergren E, Weinstock GM. 2019. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun 10:5029. https://doi.org/10.1038/s41467-019-13036-1
130. Branco P, Torgo L, Ribeiro RP. 2017. A survey of predictive modeling on imbalanced domains. ACM Comput Surv 49:1–50. https://doi.org/10.1145/2907070
131. Baig Y, Ma HR, Xu H, You L. 2023. Autoencoder neural networks enable low dimensional structure analyses of microbial growth dynamics. Nat Commun 14:7937. https://doi.org/10.1038/s41467-023-43455-0
132. Oh M, Zhang L. 2020. DeepMicro: deep representation learning for disease prediction based on microbiome data. Sci Rep 10:6026. https://doi.org/10.1038/s41598-020-63159-5
133. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. 2020. Generative adversarial networks. Commun ACM 63:139–144. https://doi.org/10.1145/3422622
134. Rong R, Jiang S, Xu L, Xiao G, Xie Y, Liu DJ, Li Q, Zhan X. 2021. MB-GAN: microbiome simulation via generative adversarial network. Gigascience 10:giab005. https://doi.org/10.1093/gigascience/giab005
135. Lever J, Krzywinski M, Altman N. 2016. Model selection and overfitting. Nat Methods 13:703–704. https://doi.org/10.1038/nmeth.3968
136. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. 2014. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15:1929–1958.
137. Dong Q, Gong S, Zhu X. 2019. Imbalanced deep learning by minority class incremental rectification. IEEE Trans Pattern Anal Mach Intell 41:1367–1381. https://doi.org/10.1109/TPAMI.2018.2832629
138. Lundberg SM, Lee SI. 2017. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. Vol. 30.Nips
139. Ribeiro MT, Singh S, Guestrin C. 2016. Why should I trust you? Abstr Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, California, USA.Association for Computing Machinery
140. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al. 2020. Language models are few-shot learners. Advances in Neural Information Processing Systems. Vol. 33, p 1877–1901.Neurips

141. Li P, Luo H, Ji B, Nielsen J. 2022. Machine learning for data integration in human gut microbiome. Microb Cell Fact 21:241. https://doi.org/10.1186/s12934-022-01973-4
142. Yu Z, Lu Y, Wang Y, Tang F, Wong K-C, Li X. 2022. ZINB-based graph embedding autoencoder for single-cell RNA-Seq interpretations. Proceedings of the AAAI Conference on Artificial Intelligence 36:4671–4679. https://doi.org/10.1609/aaai.v36i4.20392
143. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vázquez-Baeza Y, Birmingham A, Hyde ER, Knight R. 2017. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome 5:27. https://doi.org/10.1186/s40168-017-0237-y
144. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert JP. 2018. A general and flexible method for signal extraction from single-cell RNA-seq data. Nat Commun 9:284. https://doi.org/10.1038/s41467-017-02554-5
145. Johnson WE, Li C, Rabinovic A. 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8:118–127. https://doi.org/10.1093/biostatistics/kxj037
146. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. 2016. Improved techniques for training GANs. Advances in Neural Information Processing Systems. Nips.
147. Oh M, Zhang L. 2023. DeepGeni: deep generalized interpretable autoencoder elucidates gut microbiota for better cancer immunotherapy. Sci Rep 13:4599. https://doi.org/10.1038/s41598-023-31210-w
148. Sharma D, Lou W, Xu W. 2024. phylaGAN: data augmentation through conditional GANs and autoencoders for improving disease prediction accuracy using microbiome data. Bioinformatics 40:btae161. https://doi.org/10.1093/bioinformatics/btae161
149. Lin H, Peddada SD. 2020. Analysis of compositions of microbiomes with bias correction. Nat Commun 11:3514. https://doi.org/10.1038/s41467-020-17041-7
150. Lin H, Peddada SD. 2024. Multigroup analysis of compositions of microbiomes with covariate adjustments and repeated measures. Nat Methods 21:83–91. https://doi.org/10.1038/s41592-023-02092-7
151. Lin H, Peddada SD. 2020. Analysis of microbial compositions: a review of normalization and differential abundance analysis. NPJ Biofilms Microbiomes 6:60. https://doi.org/10.1038/s41522-020-00160-w
152. Hamilton WL, Ying R, Leskovec J. 2017. Inductive representation learning on large graphs. Advances in Neural Information Processing Systems. Nips.
153. Chen T, Kornblith S, Norouzi M, Hinton G. 2020. A simple framework for contrastive learning of visual representations, p 1597–1607. In Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria, PMLR 119
154. Ditzler G, Roveri M, Alippi C, Polikar R. 2015. Learning in nonstationary environments: a survey. IEEE Comput Intell Mag 10:12–25. https://doi.org/10.1109/MCI.2015.2471196
155. Sun X, Panda R, Feris R, Saenko K. 2020. Adashare: learning what to share for efficient deep multi-task learning. arXiv. https://doi.org/10.48550/arXiv.1911.12423
156. Arratia A, Ortiz C, Romaní M. 2022. On flows of neural ordinary differential equations that are solutions of Lotka-Volterra dynamical systems, p 191–198. In Artificial intelligence research and development. IOS Press.
157. Chen RTQ, Rubanova Y, Bettencourt J, Duvenaud D. 2018. Neural ordinary differential equations. Advances in Neural Information Processing Systems. Vol. 31.Nips
158. Velickovic P, Cucurull G, Casanova A, Romero A, Lio P, YJs B. 2017. Graph attention networks. arXiv. https://doi.org/10.48550/arXiv.1710.10903
159. Pfeil J, Siptroth J, Pospisil H, Frohme M, Hufert FT, Moskalenko O, Yateem M, Nechyporenko A. 2023. Classification of microbiome data from type 2 diabetes mellitus individuals with deep learning image recognition. Big Data Cogn Comput 7:51. https://doi.org/10.3390/bdcc7010051
160. Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One 10:e0130140. https://doi.org/10.1371/journal.pone.0130140
161. Lawson CE, Harcombe WR, Hatzenpichler R, Lindemann SR, Löffler FE, O'Malley MA, García Martín H, Pfleger BF, Raskin L, Venturelli OS, Weissbrodt DG, Noguera DR, McMahon KD. 2019. Common principles and best practices for engineering microbiomes. Nat Rev Microbiol 17:725–741. https://doi.org/10.1038/s41579-019-0255-9
162. Bengio Y, Louradour J, Collobert R, Weston J. 2009. Curriculum learning. ICML '09; Montreal, Quebec, Canada.Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1553374.1553380
163. Wang X, Chen Y, Zhu W. 2022. A survey on curriculum learning. IEEE Trans Pattern Anal Mach Intell 44:4555–4576. https://doi.org/10.1109/TPAMI.2021.3051099
164. Stegen JC, Lin X, Fredrickson JK, Chen X, Kennedy DW, Murray CJ, Rockhold ML, Konopka A. 2013. Quantifying community assembly processes and identifying features that impose them. ISME J 7:2069–2079. https://doi.org/10.1038/ismej.2013.93
165. De Cáceres M, Legendre P. 2009. Associations between species and groups of sites: indices and statistical inference. Ecology 90:3566–3574. https://doi.org/10.1890/08-1823.1
166. Danino R, Nachman I, Sharan R. 2024. Batch correction of single-cell sequencing data via an autoencoder architecture. Bioinform Adv 4:vbad186. https://doi.org/10.1093/bioadv/vbad186
167. Gligorijević V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, Chandler C, Taylor BC, Fisk IM, Vlamakis H, Xavier RJ, Knight R, Cho K, Bonneau R. 2021. Structure-based protein function prediction using graph convolutional networks. Nat Commun 12:3168. https://doi.org/10.1038/s41467-021-23303-9
168. Tipton L, Müller CL, Kurtz ZD, Huang L, Kleerup E, Morris A, Bonneau R, Ghedin E. 2018. Fungi stabilize connectivity in the lung and skin microbial ecosystems. Microbiome 6:12. https://doi.org/10.1186/s40168-017-0393-0
169. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, Collado MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C, Huttenhower C, Segata N. 2019. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell 176:649–662. https://doi.org/10.1016/j.cell.2019.01.001
170. Scholkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, Bengio Y. 2021. Toward causal representation learning. Proc IEEE 109:612–634. https://doi.org/10.1109/JPROC.2021.3058954
171. Lecca P. 2021. Machine learning for causal inference in biological networks: perspectives of this challenge. Front Bioinform 1:746712. https://doi.org/10.3389/fbinf.2021.746712
172. Widder S, Allen RJ, Pfeiffer T, Curtis TP, Wiuf C, Sloan WT, Cordero OX, Brown SP, Momeni B, Shou W, et al. 2016. Challenges in microbial ecology: building predictive understanding of community function and dynamics. ISME J 10:2557–2568. https://doi.org/10.1038/ismej.2016.45
173. Zhou J, Ning D. 2017. Stochastic community assembly: does it matter in microbial ecology? Microbiol Mol Biol Rev 81:e00002-17. https://doi.org/10.1128/MMBR.00002-17
174. Pasolli E, De Filippis F, Mauriello IE, Cumbo F, Walsh AM, Leech J, Cotter PD, Segata N, Ercolini D. 2020. Large-scale genome-wide analysis links lactic acid bacteria from food with the gut microbiome. Nat Commun 11:2610. https://doi.org/10.1038/s41467-020-16438-8
175. Maghini DG, Oduaran OH, Olubayo LAI, Cook JA, Smyth N, Mathema T, Belger CW, Agongo G, Boua PR, Choma SSR, et al. 2025. Expanding the human gut microbiome atlas of Africa. Nature 638:718–728. https://doi.org/10.1038/s41586-024-08485-8
176. Cuomo S, Di Cola VS, Giampaolo F, Rozza G, Raissi M, Piccialli F. 2022. Scientific machine learning through physics–informed neural networks: where we are and what's next. J Sci Comput 92:88. https://doi.org/10.1007/s10915-022-01939-z
177. Chen Z, Liu Y, Sun H. 2021. Physics-informed learning of governing equations from scarce data. Nat Commun 12:6136. https://doi.org/10.1038/s41467-021-26434-1
178. Carbonell P, Jervis AJ, Robinson CJ, Yan C, Dunstan M, Swainston N, Vinaixa M, Hollywood KA, Currin A, Rattray NJW, et al. 2018. An automated Design-Build-Test-Learn pipeline for enhanced microbial production of fine chemicals. Commun Biol 1:66. https://doi.org/10.1038/s42003-018-0076-9
179. Nielsen J, Keasling JD. 2016. Engineering cellular metabolism. Cell 164:1185–1197. https://doi.org/10.1016/j.cell.2016.02.004
180. Grazioli F, Siarheyeu R, Alqassem I, Henschel A, Pileggi G, Meiser A. 2022. Microbiome-based disease prediction with multimodal variational information bottlenecks. PLoS Comput Biol 18:e1010050. https://doi.org/10.1371/journal.pcbi.1010050
181. Glass EM, Dribinsky Y, Yilmaz P, Levin H, Van Pelt R, Wendel D, Wilke A, Eisen JA, Huse S, Shipanova A, Sogin M, Stajich J, Knight R, Meyer F, Schriml LM. 2014. MIxS-BE: a MIxS extension defining a minimum

information standard for sequence data from the built environment. ISME J 8:1–3. https://doi.org/10.1038/ismej.2013.176
182. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. Sci Data 3:160018. https://doi.org/10.1038/sdata.2016.18

AUTHOR BIOS

Jingkang Zhang is a third-year master student in the Institute of Natural Resources Survey and Research at the China University of Geosciences (Wuhan). He earned his B.S. in environmental engineering from Taiyuan University of Technology in 2023. His current research interests lie in the intersection of environmental microbiology and artificial intelligence. His work aims to leverage data-driven approaches to optimize biotechnological processes for environmental remediation.

Liyuan Ma, Ph.D., is a professor at the School of Environmental Studies, China University of Geosciences (Wuhan). She obtained her B.S. and Ph.D. degrees from Central South University and served as a visiting scholar (postdoctoral program) at the Geoenvironmental Research Centre, Cardiff University. Her research endeavors to bridge fundamental microbial ecology with practical applications in mine environmental remediation. Specifically, her team focuses on microbial-mediated geochemical cycling of iron, sulfur, and heavy metals in mining environments, along with the ecological restoration of mine sites. Her current research interests lie in the cross-kingdom interactions among bacteria, fungi, and viruses, as well as the application of machine learning approaches in microbiome studies. Dr. Ma is an editor for *mSystems*, a review editor for *Frontiers in Microbiology*, and an active member of the American Society for Microbiology (ASM).