

Harnessing machine learning for metagenomic data analysis: trends and applications

Shradha Sharma,^{1,2,3} Hari Priya Narahari,^{3,4} Karthik Raman^{2,3}

AUTHOR AFFILIATIONS See affiliation list on p. 8.

ABSTRACT Metagenomic sequencing has revolutionized our understanding of microbial ecosystems by enabling high-resolution profiling of microbes across diverse environments. However, the resulting data are high-dimensional, sparse, and noisy, posing challenges for downstream data analysis. Machine learning (ML) has provided an arsenal of tools to extract meaningful insights from such large and complex data sets. This review surveys the existing state of ML applications in metagenomic data analysis, from traditional supervised and unsupervised learning to time-series modeling, transfer learning, and newer directions such as causal ML and generative models. We highlight certain key challenges and delve into important issues like model interpretability, emphasizing the importance of explainable AI (XAI). We also compare ML with mechanistic models, commenting on their relative advantages, disadvantages, and prospects for synergy. Finally, we preview future directions, such as the incorporation of multi-omics data, synthetic data generation, and Agentic AI systems, highlighting the increasingly prominent role that AI and ML will play in the future of microbiome science.

KEYWORDS metagenomics, microbiome, machine learning, deep learning

With the advent of high-throughput sequencing and computational innovations, our knowledge of microbial diversity across ecosystems has expanded tremendously. These advances have revealed previously unrecognized species, metabolic pathways, and ecological interactions that shape microbial communities. The structure and composition of these communities are determined by their environment, and interactions among microbes create system-level properties that, in turn, influence their environment. Therefore, the knowledge of these communities is fundamental to the understanding of biogeochemical cycling, host health, and soil nutrient dynamics. Nevertheless, the complexity and vastness of biological data pose computational challenges in processing and interpreting the data correctly. One of the significant challenges with metagenomic data is its inherent compositional nature. Since the overall read count (i.e., the total number of sequencing reads assigned to taxa) is limited by sequencing depth, the abundance of a single taxon is affected by the presence and abundance of others, creating spurious correlations and biases (1). This compositional nature of the data needs to be addressed through proper normalization and data transformation techniques. In addition, microbiome data are large, high-dimensional, highly variable, and plagued by the “curse of dimensionality,” where the number of features (usually the abundances of different organisms) is much larger than the number of samples (2). This high dimensionality complicates data interpretation. In addition, microbiome data are often noisy and redundant, complicating downstream analysis and the discovery of meaningful biological patterns (2). To address these challenges, researchers are increasingly turning to ML techniques for more effective analysis and interpretation in microbiome studies.

Editor Ashkaan K. Fahimipour, Florida Atlantic University, Boca Raton, Florida, USA

Address correspondence to Karthik Raman, kraman@dsai.iitm.ac.in.

The authors declare no conflict of interest.

Published 7 October 2025

Copyright © 2025 Sharma et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

While previous reviews on ML in microbiome have discussed in detail the data transformation and normalization techniques (3–5), the scope of this review is to provide an overview of ML approaches and applications for microbiome analysis, outlining their core concepts, strengths, and limitations. Our goal is to help researchers choose methods that best fit their data and objectives. We then describe the next steps, *post hoc* interpretability techniques, and quantitative scoring metrics to increase confidence in model results. Finally, we emphasize that ML is just one of many tools for metagenomic data analysis and that methodological choices should always be guided by the data at hand and the specific research question.

CURRENT LANDSCAPE OF ML IN MICROBIOME RESEARCH

ML has become a key tool in microbiome research because it can handle complex, high-dimensional data and uncover patterns that traditional methods often miss. This is especially important given the noisy, sparse, and imbalanced nature of metagenomic data (2). In these data sets, features—that is, the relative abundances of different taxa are unevenly distributed, with many rare taxa and only a few dominant ones. This makes it difficult for models to generalize across samples. Feature engineering helps address this by transforming raw data to better capture the biologically relevant signals. Techniques broadly fall into two categories: feature extraction, which converts raw data into more structured and informative representations, and feature selection, which removes redundancy by retaining only the most relevant features. Several methods exist for each, but manually tuning the right combination of feature engineering and ML models is labor-intensive and error-prone. Automated approaches like BioAutoML (6) streamline this process by testing multiple pipelines to identify the most effective combinations, improving performance while reducing human effort.

Although robust feature engineering lays the foundation, the success of ML models also depends on how the learning process is structured, particularly in relation to data labeling. Currently, most of the ML models for microbiome analysis are based on supervised learning. This reflects the field's fundamental need to answer concrete questions like pathogen detection, functional annotation, and disease prediction, which require labeled data sets. However, high-quality labeled data remain scarce for many important scenarios, including studies of rare diseases, specialized populations, or longitudinal studies. This limitation has driven interest in semi-supervised approaches like meta model agnostic pseudo label learning (MMAPE), which employs a teacher–student framework to progressively improve predictions by leveraging both labeled and unlabeled data, even when dealing with out-of-distribution samples (7).

For cases where even semi-supervised learning struggles due to extreme data scarcity, transfer learning offers a powerful alternative by adapting models pre-trained on large data sets to smaller related ones. A prime example is EXPERT (8), which was first trained on the comprehensive MGnify (9) database before being fine-tuned for specific applications ranging from age-related microbiome changes to different stages of colorectal cancer.

As these diverse ML approaches continue to proliferate, the field faces growing needs for standardization and benchmarking to enable fair comparisons between methods. Critical assessment of massive data analysis (CAMDA) provides an open framework for metagenomic interpretation through community challenges (10), offering pre-curated data sets and defined metrics to objectively assess model performance. Similarly, critical assessment of metagenome interpretation (CAMI) benchmarks tools for assembly, taxonomic profiling, and binning using realistic data sets to ensure reproducibility (11).

With these developments in mind, it is essential to recognize the broad range of ML models currently used in microbiome research. Table 1 summarizes the recently developed ML tools and their applications in microbiome analysis. These tools span several ML methods, such as time-series analysis, supervised and unsupervised learning, and deep learning (DL) models, specifically created to address the specific challenges of microbiome data analysis.

TABLE 1 Cross-consistency matrix (CCM) of ML approaches and applications in metagenomic data analysis^a

Major application category	Subcategory	Time-series		Supervised learning			Unsupervised learning		
		ML	DL	ML	ML	DL	ML	ML	DL
Assembly and binning	Assembly			MetaVelvet-SL (12)		MetaVelvet-DL (13)	Reinforcement learning (14)		VAMB (15)
	Binning			BusyBee (16)		SemiBin (17)	MaxBin (18), MaxBin 2.0 (19)		COMEBin (20)
	Bin refinement and quality check			CheckM2 (21)			Agglomerative hierarchical clustering (22)		SemiBin (17)
Annotation	Taxonomy			R-SVM (23)		CHEER (24)	Latent Dirichlet allocation (25)		DBN (26)
	Phylogenetic tree					DeePhy (27)	Split-weight embedding (28)		DeLUCS (29)
	Functional profile			Random forest (30)		Meta-MFDL (31)	PICA (32)		DeepARG (33)
Microbiome analysis	Prediction	MDITRE (34)	phyLoSTM (35)	KernelBiome (36)		MDL4Microbiome (37)	VALENCIA (38)		DeepGum (39)
	Host-microbe interaction		Recurrent neural network (40)	Random forest (41)		Meta-Spec (42)	RCCA (43)		MEGMA (44)
	Microbe-microbe interaction	mbtransfer (45)	MicroGrowthPredictor (46)	SVM (47)		LSTM framework (48)	Deep latent space (49)		Transformer (50)
	Pattern recognition	ARGfore (51)	DeepVirFinder (52), EXPERT (8)	Recursive ensemble feature selection (53)		BPNNHMDA (54)	DMIM (55)		DeepBioGen (56)
	Feature engineering		DeepIDA-GRU (57)	BioAutoML (6)		MDeep (58)	Hierarchical feature engineering (59)		DeepGeni (60)
Preprocessing	Dimensionality reduction	EMBED (61)	TCAM (62)	Recursive feature elimination (63)		Autoencoder	Barnes-Hut stochastic neighbor embedding (65)		DeepMicro (66)
						Neural Network (64)			
Synthetic data generation			DeepMicroGen (67)			DeepBioGen (56)			Evo 2 (68)

^aThis matrix summarizes ML methods applied to various metagenomic analysis tasks. Columns represent ML paradigms, such as supervised, unsupervised, DL, and time-series approaches, while rows correspond to key application areas, including assembly, binning, taxonomic annotation, functional profiling, prediction of host-microbe interaction, and pattern recognition. Each filled cell indicates the existence of one or more tools that implement the corresponding ML method for the given task. Empty cells may be interpreted as (i) potential research gaps, (ii) infeasible method-task pairings due to data or modeling constraints, or (iii) areas outside the scope of this review.

INTERPRETABILITY

ML and DL algorithms have demonstrated remarkable success in analyzing complex, high-dimensional data sets. However, despite their predictive power, the internal decision-making processes of these models often remain opaque, which is why they are characterized as “black-box” models. This lack of transparency arises from the inherent complexity and non-linearity associated with the information they analyze, making it difficult for investigators to trace how inputs determine outputs. In biological scenarios, where interpretability is key, dependency on model predictions without understanding their underlying rationale can be misleading or even detrimental. To overcome this, the emerging field of explainable AI (XAI) offers frameworks designed to illuminate model reasoning.

XAI primarily aims to deliver (i) interpretability, yielding qualitative insight into predictions, usually presented visually or as text; (ii) explainability, facilitating humans’ ability to understand the internal workings of a model; and (iii) causality, determining the degree to which a model replicates the underlying causal relations among inputs and outputs (69). Among the most widely used XAI techniques are *post hoc* explanation methods like LIME (70) and SHAP (SHapley Additive exPlanations) (71), which diagnose model behavior after training.

LIME was among the first general-purpose model-agnostic methods created to explain ML models. Its core idea is to approximate the model locally using a simpler, interpretable surrogate model. More precisely, LIME slightly perturbs the input data and looks at how the predictions change. It then uses a simple model, for example, linear regression, to fit these locally produced data, effectively pointing out which features (e.g., microbial taxa) had the largest influence on a given prediction. However, this local fidelity can also be a weakness since it might not capture the model’s global decision-making rationale. LIME is valuable for questions where sample-specific predictions are needed. For example, why did a model classify a microbiome sample as “high risk,” and which characteristics contributed most to that result? Nonetheless, owing to its surrogate nature, LIME’s explanations may not always align with the original model’s reasoning (72).

Another *post hoc* method, SHAP, has been of special interest due to its strong theoretical foundations and reproducible outputs. It applies Shapley values—derived from cooperative game theory—to attribute credit or blame to features based on how they contribute to a model’s prediction (71). In this framework, each feature is treated as a “player” in a game, and the model prediction is the “payout” distributed among them based on their marginal contributions. For example, in the study by Novielli et al. (73), SHAP was used to analyze the drivers of soil respiration sensitivity (Q_{10}). The results revealed that glucose-induced soil respiration and the proportion of bacterial taxa positively associated with Q_{10} were among the most influential predictors. SHAP quantified the contribution of these features by measuring how the model’s output changed with and without each feature, providing insights aligned with established ecological understanding, such as the role of microbial metabolism in carbon cycling. This demonstrates how SHAP can help uncover key biological drivers in complex environmental systems. While SHAP as a technique is extremely robust, due to its game-theoretic foundations, its explanations still rely on the model’s behavior under perturbations—a vulnerability shared with LIME that adversarial attacks can exploit. Slack et al. (74) demonstrated how classifiers can be deliberately engineered to deceive *post hoc* explanation methods, with LIME proving particularly susceptible due to its local approximation approach compared to SHAP’s. These findings reveal fundamental vulnerabilities in current perturbation-based explanation methods, demanding new research directions in adversarially robust explainability that can maintain interpretability while withstanding manipulation attempts.

Given these limitations, it is essential to quantitatively evaluate ML models. In microbiome research, assessing models using appropriate performance metrics is a standard best practice. For readers interested in how to evaluate ML models in

microbiome applications, we provide a comprehensive overview of relevant metrics and their appropriate use cases in Note S1, along with a task-metric scoring matrix (Table S1) and a glossary of terms (Table S2). In Fig. 1, we present a general workflow for implementing ML for metagenomic data.

Most critically, these challenges underscore that XAI aims to explain the inner logic and choice-making of an ML model, shedding light on how a model came to its prediction. Explanations derived in this manner do not have to mirror the underlying biological processes. As a result, there are likely to be differences between outputs from models and biological ground truth. Hence, domain expert validation is crucial to establish the reliability and biological significance of the conclusions inferred from these models.

COMPARING AND COMBINING ML AND MECHANISTIC MODELS IN METAGENOMICS

Before the rise of ML, microbiome research relied heavily on mechanistic models such as Boolean models, ordinary differential equations (ODEs), and constraint-based models to describe microbial dynamics and interactions. Although this review has focused on ML for the analysis of metagenomic data, mechanistic models continue to play an important complementary role (75, 76). Both come with their strengths and limitations, and the decision to use one over the other should be based on the nature and quality of the data set, as well as the objectives of the study. While ML excels at identifying non-linear and complex patterns in large-scale, high-dimensional, and time-series data sets, it often fails to infer causality. Mechanistic models, on the other hand, are rooted in hypothesis generation and causal reasoning, aiming to capture the underlying processes driving microbial community dynamics (77, 78). For instance, while an ML model may distinguish between healthy and diseased microbiome states, a mechanistic model can elucidate how changes in nutrient concentrations lead to such shifts. A recent study by Kuppa Baskaran et al. (79) applied this approach by constructing metagenome-based metabolic models to predict metabolic exchanges and horizontal gene transfer events in the deep-sea hydrothermal vents' microbiome, offering insights into archaeal-bacterial interactions that underpin community structure. However, these models can be limited by their underlying assumptions and may not readily scale to complex data sets. Their development also demands domain-specific expertise and significant manual effort in curation and analysis. This has spurred interest in causal ML, which aims to extract causal relationships directly from observational data with minimal assumptions.

Causal ML offers a middle ground by bridging the correlation-causation gap. It extracts associations and potential causal relationships from data, adding to the interpretability of the results. DoWhy (80) is one framework that explicitly models causal assumptions and validates them, unlike traditional ML models. The advantage is that it reduces false positives due to spurious correlations and allows for designing mechanistic follow-ups of the results.

Another clever way to deploy ML is as a hybrid strategy with mechanistic or statistical models. A compelling example of this hybrid philosophy is mbtransfer, a method that blends ML with principles from control theory to study how interventions reshape microbial communities over time. Unlike conventional ML, which treats causality as an afterthought, mbtransfer embeds temporal reasoning into its framework: it uses transfer functions—a concept adapted from engineering to model delayed effects of perturbations (e.g., dietary changes or birth events)—and simulates counterfactual scenarios (e.g., “What if the intervention had not occurred?”). These simulations, combined with mirror statistics (a statistical method that controls false discoveries by comparing results across data splits), allow researchers to identify taxa most sensitive to interventions. While not a fully mechanistic model, mbtransfer bridges ML's scalability with causal ML's ambition, inferring when and how interventions matter (45). The resulting hypotheses can then be tested mechanistically (e.g., using metabolic models or generalized Lotka-Volterra equations), creating a feedback loop where ML-driven discoveries inform biological validation. This mirrors frameworks like DoWhy but directly addresses

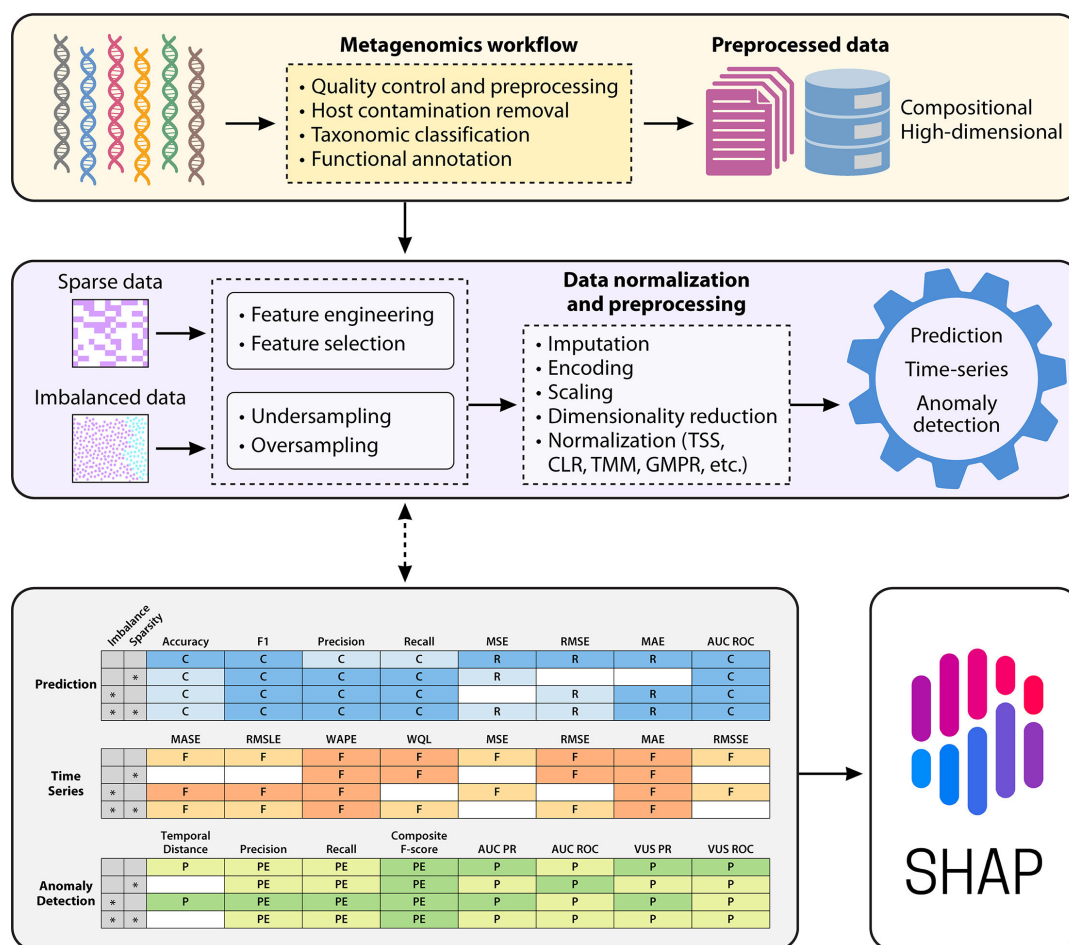


FIG 1 Overview of the ML workflow for metagenomic data analysis. The sequencing data are first processed through the metagenomics workflow to generate features that are inherently compositional and high-dimensional. Depending on data characteristics such as sparsity or class imbalance, appropriate preprocessing steps are applied to homogenize and prepare the data for ML. Based on the study objectives (prediction, time-series, or anomaly detection), an appropriate ML model is chosen and evaluated using the most suitable scoring metric. To obtain biological insights, interpretability techniques such as SHAP are employed.

microbiome-specific challenges, such as sparse time-series data and phylogenetic dependencies (80).

It is important to recognize that neither ML nor mechanistic models can compensate for the low-quality data or a flawed experimental design. An adequate sample size per experimental group is essential for both approaches. Both approaches demand sufficient sample sizes to yield reliable results. ML methods are particularly vulnerable to overfitting when applied to small data sets. On the other hand, mechanistic models require adequate observations to constrain parameters and validate assumptions. This brings us to statistical significance, often expressed through *P*-values, a concept developed to help researchers communicate confidence in their results. Unfortunately, *P*-values have been widely misused and even abused, intentionally or otherwise. Studies in the metagenomics space are often underpowered and increasingly vulnerable to *P*-hacking, where researchers test multiple hypotheses to find one that appears significant, despite the volume of data generated (81). The low sample size only worsens this, increasing the risk of false positives or negatives. Kers and Saccenti's study is one of the few studies that explicitly discuss how small sample sizes can skew alpha-diversity metrics and emphasize the need for statistical planning before the experiment begins. As a general guideline, a minimum of 25 samples per category is recommended for any analysis, whether the approach is mechanistic or ML-based (82).

Ultimately, it is not a question of choosing one approach over the other, but how we could strategically combine them to extract predictive power and mechanistic insights from microbiome data. Such synergies between ML and mechanistic modeling have been well explored in drug discovery and related domains (83), and there is much opportunity to carry over such learnings to metagenomic data analysis.

FUTURE PERSPECTIVES

The idea that genes are the codes that shape the trajectories of our lives is only half the truth. Genes do not function autonomously; instead, they are akin to a database that the biological system accesses and interprets contextually, rather than a program that executes independently. Metagenomics thus gives us a very static view of microbial DNA. To truly understand function, context, and dynamics, we must move beyond the DNA picture to include metatranscriptomics (RNA expression), metaproteomics (proteins in action), and metabolomics (the results of all activity). These layers reveal what microbes are actively doing in a given environment. To meet the complexity of such biological questions, computational tools must evolve accordingly. MMAPLE is one such tool, a multimodal, multi-omics framework that combines various data types and employs a meta-learning approach, promising results in extracting meaningful insights from complex microbiome data sets (7). Hence, new ML models tailored for multi-omics integration and analysis represent the next frontier in decoding microbial function, context, and dynamics.

However, integrating these diverse data sets is not straightforward. While multi-omics integration is conceptually promising, it is technically very challenging. Aligning heterogeneous data with different scales, missing values, noise characteristics, and batch effects poses a substantial barrier to seamless integration, demanding innovative normalization or imputation strategies. Furthermore, as multi-omics, single-cell, temporally and spatially resolved data sets grow, computational efficiency can become a limiting factor. Training large-scale models—especially with DL frameworks—requires significant compute, memory, time, and energy resources, which may be inaccessible to smaller research groups.

In parallel to traditional sequence-based metagenomics, recent efforts are beginning to explore the untapped potential of raw signal data produced by sequencing technologies. For instance, Urel et al. (84) developed a DL framework that infers microbial viability directly from raw nanopore electrical signals, bypassing the need for read-level taxonomic assignment. These signal-based ML models offer new possibilities for identifying functional states of microbes (e.g., live vs. dead) and present unique preprocessing challenges such as denoising, segmentation, and feature extraction from unstructured time-series data. As sensor-based metagenomics grows, developing robust signal processing and evaluation metrics tailored to such data will be essential.

Another emerging frontier in artificial intelligence (AI) is the development of agentic AI systems. These hybrid frameworks combine the structured logic and deterministic nature of software engineering with the adaptability of AI. Recent innovations such as AgentClinic (85), Agent Laboratory (86), and the AI co-scientist (87) exemplify this shift toward systems that not only analyze data but actively assist in experimental design, hypothesis generation, and interpretation. Such systems could play a vital role in microbiome research and help explore interventions or simulate ecological shifts *in silico*.

Alongside agentic AI, generative models and LLMs are now powering synthetic data for microbiome studies. Tools like Evo 2, for example, can infer and “fill in” low-coverage regions of MAGs from sparse reads (68), producing biologically plausible genomes. These reconstructed sequences can then be used to test how particular mutations or environmental shifts affect metabolic outputs, giving us an *in silico* window into evolutionary dynamics. Of course, each synthetic data set must be benchmarked against real measurements to guard against bias and ensure the models we build remain grounded in biology.

Despite the rapid rise of ML in metagenomics, many challenges remain to be addressed. First, there is a lack of standardized benchmarks for evaluating models across diverse microbiome data sets—sample sizes, sequencing platforms, annotation depths, and environmental contexts all vary wildly, making comparisons across studies difficult. Second, reproducibility suffers when studies omit detailed documentation or fail to share containerized workflows. We urgently need open-source tools, shareable pipelines, and strict adherence to findable, accessible, interoperable, reusable, reproducible (FAIRer) data principles (88). Finally, public data are heavily skewed toward European and North American cohorts—the South American MicroBiome Archive (saMBA) preprint reports that over 70% of sequenced human microbiomes come from these regions (89). Such geographic and population biases threaten the global generalizability of our ML models.

Overall, ML models and emerging advances in AI will have a telling impact on microbiome data analysis. The next frontier lies in hybrid approaches that marry ML with mechanistic models and statistical frameworks. As cartographers of the microbial world, we must summon every kind of map: from multi-omics integration that layers genomic, proteomic, and metabolic data, to algorithmic compasses that blend deep learning with ecological theory. Only by wielding this full spectrum of tools can we truly navigate the boundless complexity of these invisible ecosystems that quietly govern life on Earth.

ACKNOWLEDGMENTS

S.S. acknowledges the Half-Time Research Assistantship from the Ministry of Education, Government of India. H.P.N. acknowledges the pre-doctoral fellowship from the National Programme on Technology Enhanced Learning (NPTEL).

The authors apologize to colleagues whose work they could not cite due to space constraints.

AUTHOR AFFILIATIONS

¹Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology (IIT) Madras, Chennai, Tamil Nadu, India

²The Centre for Integrative Biology and Systems medicine (IBSE), Wadhwani School of Data Science and AI, IIT Madras, Chennai, Tamil Nadu, India

³Department of Data Science and AI, Wadhwani School of Data Science and AI, IIT Madras, Chennai, Tamil Nadu, India

⁴Department of Computer Science and Engineering, Shiv Nadar University, Chennai, Tamil Nadu, India

AUTHOR ORCIDs

Shradha Sharma  <http://orcid.org/0009-0003-7599-5460>

Hari Priya Narahari  <http://orcid.org/0009-0004-6227-416X>

Karthik Raman  <http://orcid.org/0000-0002-9311-7093>

AUTHOR CONTRIBUTIONS

Shradha Sharma, Data curation, Investigation, Writing – original draft | Hari Priya Narahari, Data curation, Investigation, Writing – original draft | Karthik Raman, Conceptualization, Supervision, Writing – review and editing

ADDITIONAL FILES

The following material is available [online](#).

Supplemental Material

Supplemental Material (mSystems01642-24-s0001.pdf). Scoring matrix for evaluating machine learning models.

REFERENCES

- 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>
- Armstrong G, Rahman G, Martino C, McDonald D, Gonzalez A, Mishne G, Knight R. 2022. Applications and comparison of dimensionality reduction methods for microbiome data. *Front Bioinform* 2:821861. <https://doi.org/10.3389/fbinf.2022.821861>
- 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>
- McKnight DT, Huerlimann R, Bower DS, Schwarzkopf L, Alford RA, Zenger KR. 2019. Methods for normalizing microbiome data: An ecological perspective. *Methods Ecol Evol* 10:389–400. <https://doi.org/10.1111/2041-210X.13115>
- Wang B, Sun F, Luan Y. 2024. Comparison of the effectiveness of different normalization methods for metagenomic cross-study phenotype prediction under heterogeneity. *Sci Rep* 14:7024. <https://doi.org/10.1038/s41598-024-57670-2>
- Bonidia RP, Santos APA, de Almeida BLS, Stadler PF, da Rocha UN, Sanchez DS, de Carvalho ACP. 2022. BioAutoML: automated feature engineering and metalearning to predict noncoding RNAs in bacteria. *Brief Bioinform* 23:bbac218. <https://doi.org/10.1093/bib/bbac218>
- Wu Y, Xie L, Liu Y, Xie L. 2024. Semi-supervised meta-learning elucidates understudied molecular interactions. *Commun Biol* 7:1104. <https://doi.org/10.1038/s42003-024-06797-z>
- Chong H, Zha Y, Yu Q, Cheng M, Xiong G, Wang N, Huang X, Huang S, Sun C, Wu S, Chen W-H, Coelho LP, Ning K. 2022. EXPERT: transfer learning-enabled context-aware microbial community classification. *Brief Bioinform* 23:bbac396. <https://doi.org/10.1093/bib/bbac396>
- Richardson L, Allen B, Baldi G, Beracochea M, Bileschi ML, Burdett T, Burgin J, Caballero-Pérez J, Cochrane G, Colwell LJ, Curtis T, Escobar-Zepeda A, Gurbich TA, Kale V, Korobeynikov A, Raj S, Rogers AB, Sakharova E, Sanchez S, Wilkinson DJ, Finn RD. 2023. MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res* 51:D753–D759. <https://doi.org/10.1093/nar/gkac1080>
- Contreras-Peruero H, Nuñez I, Vázquez-Rosas-Landa M, Santana-Quinteros D, Pashkov A, Carranza-Barragán ME, Perez-Estrada R, Guerrero-Flores S, Balanzario E, Muñoz Sánchez V, Nakamura M, Ramírez-Ramírez LL, Sélem-Mojica N. 2024. CAMDA 2023: finding patterns in urban microbiomes. *Front Genet* 15:1449461. <https://doi.org/10.3389/fgen.2024.1449461>
- Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, Gregor I, Majda S, Fiedler J, Dahms E, et al. 2017. Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nat Methods* 14:1063–1071. <https://doi.org/10.1038/nmeth.4458>
- Sato K, Sakakibara Y, Afiahayati. 2015. MetaVelvet-SL: an extension of the Velvet assembler to a *de novo* metagenomic assembler utilizing supervised learning. *DNA Res* 22:69–77. <https://doi.org/10.1093/dnares/dsu041>
- Liang K-C, Sakakibara Y. 2021. MetaVelvet-DL: a MetaVelvet deep learning extension for *de novo* metagenome assembly. *BMC Bioinform* 22:427. <https://doi.org/10.1186/s12859-020-03737-6>
- Bocicor MI, Czibula G, Czibula IG. 2011. A reinforcement learning approach for solving the fragment assembly problem. 13th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, p 191–198
- Nissen JN, Johansen J, Allesøe RL, Sønderby CK, Armenteros JJA, Grønbech CH, Jensen LJ, Nielsen HB, Petersen TN, Winther O, Rasmussen S. 2021. Improved metagenome binning and assembly using deep variational autoencoders. *Nat Biotechnol* 39:555–560. <https://doi.org/10.1038/s41587-020-00777-4>
- Laczny CC, Kiefer C, Galata V, Fehlmann T, Backes C, Keller A. 2017. BusyBee Web: metagenomic data analysis by bootstrapped supervised binning and annotation. *Nucleic Acids Res* 45:W171–W179. <https://doi.org/10.1093/nar/gkx348>
- 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>
- Wu YW, Tang YH, Tringe SG, Simmons BA, Singer SW. 2014. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* 2:26. <https://doi.org/10.1186/2049-2618-2-26>
- Wu YW, Simmons BA, Singer SW. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32:605–607. <https://doi.org/10.1093/bioinformatics/btv638>
- Wang Z, You R, Han H, Liu W, Sun F, Zhu S. 2024. Effective binning of metagenomic contigs using contrastive multi-view representation learning. *Nat Commun* 15:585. <https://doi.org/10.1038/s41467-023-4429-0-z>
- Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. 2023. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat Methods* 20:1203–1212. <https://doi.org/10.1038/s41592-023-01940-w>
- Plaza Onate F, Batto J-M, Juste C, Fadlallah J, Fougeroux C, Gouas D, Pons N, Kennedy S, Levenez F, Dore J, Ehrlich SD, Gorochov G, Larsen M. 2015. Quality control of microbiota metagenomics by k-mer analysis. *BMC Genomics* 16:183. <https://doi.org/10.1186/s12864-015-1406-7>
- Cui H, Zhang X. 2013. Alignment-free supervised classification of metagenomes by recursive SVM. *BMC Genomics* 14:1–12. <https://doi.org/10.1186/1471-2164-14-641>
- Shang J, Sun Y. 2021. CHEER: hierarchical taxonomic classification for viral metagenomic data via deep learning. *Methods* 189:95–103. <https://doi.org/10.1016/j.jmeth.2020.05.018>
- Pappalardo VY, Azarang L, Zaura E, Brandt BW, de Menezes RX. 2024. A new approach to describe the taxonomic structure of microbiome and its application to assess the relationship between microbial niches. *BMC Bioinform* 25:58. <https://doi.org/10.1186/s12859-023-05575-8>
- Fiannaca A, La Paglia L, La Rosa M, Lo Bosco G, Renda G, Rizzo R, Gaglio S, Urso A. 2018. Deep learning models for bacteria taxonomic classification of metagenomic data. *BMC Bioinform* 19:198. <https://doi.org/10.1186/s12859-018-2182-6>
- Mahapatra A, Mukherjee J. 2025. DeePhy: a deep learning model to reconstruct phylogenetic tree from unaligned nucleotide sequences. *bioRxiv*. <https://doi.org/10.1101/2025.03.09.642239>
- Kong Y, Tilely GP, Solis-Lemus C. 2023. Unsupervised learning of phylogenetic trees via split-weight embedding. *arXiv:2312.16074*. <https://doi.org/10.48550/arXiv.2312.16074>
- Millán Arias P, Alipour F, Hill KA, Kari L. 2022. Delucs: deep learning for unsupervised clustering of dna sequences. *PLoS One* 17:e0261531. <https://doi.org/10.1371/journal.pone.0261531>
- Maltecca C, Lu D, Schillebeeckx C, McNulty NP, Schwab C, Shull C, Tiezzi F. 2019. Predicting growth and carcass traits in swine using microbiome data and machine learning algorithms. *Sci Rep* 9:6574. <https://doi.org/10.1038/s41598-019-43031-x>
- Zhang SW, Jin XY, Zhang T. 2017. Gene prediction in metagenomic fragments with deep learning. *Biomed Res Int* 2017:4740354. <https://doi.org/10.1155/2017/4740354>
- Feldbauer R, Schulz F, Horn M, Rattei T. 2015. Prediction of microbial phenotypes based on comparative genomics. *BMC Bioinform* 16 Suppl 14:1–8. <https://doi.org/10.1186/1471-2105-16-S14-S1>
- Arango-Argoty G, Garner E, Pruden A, Heath LS, Vikesland P, Zhang L. 2018. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* 6:23. <https://doi.org/10.1186/s40168-018-0401-z>
- Maringanti VS, Bucci V, Gerber GK. 2022. MDITRE: scalable and interpretable machine learning for predicting host status from temporal microbiome dynamics. *mSystems* 7:e00132-22. <https://doi.org/10.1128/mSystems.00132-22>
- Sharma D, Xu W. 2021. phyLoSTM: a novel deep learning model on disease prediction from longitudinal microbiome data. *Bioinformatics* 37:3707–3714. <https://doi.org/10.1093/bioinformatics/btab482>
- Huang S, Ailer E, Kilbertus N, Pfister N. 2023. Supervised learning and model analysis with compositional data. *PLoS Comput Biol* 19:e1011240. <https://doi.org/10.1371/journal.pcbi.1011240>
- Lee SJ, Rho M. 2022. Multimodal deep learning applied to classify healthy and disease states of human microbiome. *Sci Rep* 12:824. <https://doi.org/10.1038/s41598-022-04773-3>
- France MT, Ma B, Gajer P, Brown S, Humphrys MS, Holm JB, Waetjen LE, Brotman RM, Ravel J. 2020. VALENCIA: a nearest centroid classification

- method for vaginal microbial communities based on composition. *Microbiome* 8:166. <https://doi.org/10.1186/s40168-020-00934-6>
39. Çiftcioğlu UGE, Nalbanoğlu OU. 2024. DeepGum: deep feature transfer for gut microbiome analysis using bottleneck models. *Biomed Signal Process Control* 91:105984. <https://doi.org/10.1016/j.bspc.2024.105984>
 40. Chen X, Liu L, Zhang W, Yang J, Wong KC. 2021. Human host status inference from temporal microbiome changes via recurrent neural networks. *Brief Bioinform* 22:bbab223. <https://doi.org/10.1093/bib/bbab223>
 41. Zhang M, Yang L, Ren J, Ahlgren NA, Fuhrman JA, Sun F. 2017. Prediction of virus-host infectious association by supervised learning methods. *BMC Bioinform* 18:60. <https://doi.org/10.1186/s12859-017-1473-7>
 42. Wu S, Li Z, Chen Y, Zhang M, Sun Y, Xing J, Zhao F, Huang S, Knight R, Su X. 2023. Host-variable-embedding augmented microbiome-based simultaneous detection of multiple diseases by deep learning. *Adv Intelligent Syst* 5:2300342. <https://doi.org/10.1002/aisy.202300342>
 43. Presley LL, Ye J, Li X, Leblanc J, Zhang Z, Ruegger PM, Allard J, McGovern D, Ippoliti A, Roth B, Cui X, Jeske DR, Elashoff D, Goodglick L, Braun J, Borneman J. 2012. Host-microbe relationships in inflammatory bowel disease detected by bacterial and metaproteomic analysis of the mucosal-luminal interface. *Inflamm Bowel Dis* 18:409–417. <https://doi.org/10.1002/ibd.21793>
 44. 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 (New York)* 4:100658. <https://doi.org/10.1016/j.patter.2022.100658>
 45. Sankaran K, Jeganathan P. 2024. mbrtransfer: Microbiome intervention analysis using transfer functions and mirror statistics. *PLoS Comput Biol* 20:e1012196. <https://doi.org/10.1371/journal.pcbi.1012196>
 46. Sun G, Zhou YH. 2024. Predicting microbiome growth dynamics under environmental perturbations. *Appl Microbiol* 4:948–958. <https://doi.org/10.3390/applmicrobiol4020064>
 47. Wen C, Zheng Z, Shao T, Liu L, Xie Z, Le Chatelier E, He Z, Zhong W, Fan Y, Zhang L, Li H, Wu C, Hu C, Xu Q, Zhou J, Cai S, Wang D, Huang Y, Breban M, Qin N, Ehrlich SD. 2017. Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome Biol* 18:142. <https://doi.org/10.1186/s13059-017-1271-6>
 48. 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>
 49. García-Jiménez B, Muñoz J, Cabello S, Medina J, Wilkinson MD. 2021. Predicting microbiomes through a deep latent space. *Bioinformatics* 37:1444–1451. <https://doi.org/10.1093/bioinformatics/btaa971>
 50. 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>
 51. Choi JM, Rumi MA, Vikesland PJ, Pruden A, Zhang L. 2025. ARGfore: a multivariate framework for forecasting antibiotic resistance gene abundances using time-series metagenomic datasets. *bioRxiv*:2025–01. <https://doi.org/10.1101/2025.01.13.632008>
 52. Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, Xie X, Poplin R, Sun F. 2020. Identifying viruses from metagenomic data using deep learning. *Quant Biol* 8:64–77. <https://doi.org/10.1007/s40484-019-0187-4>
 53. Peralta-Marzal LN, Rojas-Velazquez D, Rigters D, Prince N, Garssen J, Kraneveld AD, Perez-Pardo P, Lopez-Rincon A. 2024. A robust microbiome signature for autism spectrum disorder across different studies using machine learning. *Sci Rep* 14:814. <https://doi.org/10.1038/s41598-023-50601-7>
 54. Li H, Wang Y, Zhang Z, Tan Y, Chen Z, Wang X, Pei T, Wang L. 2020. Identifying microbe-disease association based on a novel back-propagation neural network model. *IEEE/ACM Trans Comput Biol Bioinform* 18:2502–2513. <https://doi.org/10.1109/TCBB.2020.2986459>
 55. Holmes I, Harris K, Quince C. 2012. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One* 7:e30126. <https://doi.org/10.1371/journal.pone.0030126>
 56. Oh M, Zhang L. 2022. Generalizing predictions to unseen sequencing profiles via deep generative models. *Sci Rep* 12:7151. <https://doi.org/10.1038/s41598-022-11363-w>
 57. Jain S, Safo SE. 2024. DeepIDA-GRU: a deep learning pipeline for integrative discriminant analysis of cross-sectional and longitudinal multiview data with applications to inflammatory bowel disease classification. *Brief Bioinform* 25:bbae339. <https://doi.org/10.1093/bib/bbae339>
 58. Wang Y, Bhattacharya T, Jiang Y, Qin X, Wang Y, Liu Y, Saykin AJ, Chen L. 2021. A novel deep learning method for predictive modeling of microbiome data. *Brief Bioinform* 22:bbaa073. <https://doi.org/10.1093/bib/bbaa073>
 59. Oudah M, Henschel A. 2018. Taxonomy-aware feature engineering for microbiome classification. *BMC Bioinformatics* 19:227. <https://doi.org/10.1186/s12859-018-2205-3>
 60. 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>
 61. Shahin M, Ji B, Dixit PD. 2023. EMBED: essential microbiome dynamics, a dimensionality reduction approach for longitudinal microbiome studies. *NPJ Syst Biol Appl* 9:26. <https://doi.org/10.1038/s41540-023-00285-6>
 62. Mor U, Cohen Y, Valdés-Mas R, Kviatkovsky D, Elinav E, Avron H. 2022. Dimensionality reduction of longitudinal 'omics data using modern tensor factorizations. *PLoS Comput Biol* 18:e1010212. <https://doi.org/10.1371/journal.pcbi.1010212>
 63. Luong HH, Phan NTL, Duong TT, Dang TM, Nguyen TD, Nguyen HT. 2021. Dimensionality reduction on metagenomic data with recursive feature elimination. *Complex, intelligent and software intensive systems: Proceedings of the 15th International Conference on Complex, Intelligent and Software Intensive Systems (CISIS-2021)*. p 68–79. https://doi.org/10.1007/978-3-030-79725-6_7
 64. 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>
 65. Laczny CC, Pinel N, Vlassis N, Wilmes P. 2014. Alignment-free visualization of metagenomic data by nonlinear dimension reduction. *Sci Rep* 4:4516. <https://doi.org/10.1038/srep04516>
 66. 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>
 67. Choi JM, Ji M, Watson LT, Zhang L. 2023. DeepMicroGen: a generative adversarial network-based method for longitudinal microbiome data imputation. *Bioinformatics* 39:btad286. <https://doi.org/10.1093/bioinformatics/btad286>
 68. Bixi G, Durrant MG, Ku J, Poli M, Brockman G, Chang D, Gonzalez GA, King SH, Li DB, Merchant AT. 2025. Genome modeling and design across all domains of life with evo 2. *bioRxiv*. <https://doi.org/10.1101/2025.02.18.638918>
 69. Molnar C, Casalicchio G, Bischl B. 2020. Interpretable machine learning—a brief history, state-of-the-art and challenges. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. p 417–431. https://doi.org/10.1007/978-3-030-65965-3_28
 70. Ribeiro MT, Singh S, Guestrin C. 2016. “Why should i trust you?”: explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. p 1135–1144
 71. Lundberg SM, Lee SI. 2017. A unified approach to interpreting model predictions. *Neural Inform Process Syst* 2017:4765–4774.
 72. Holzinger A, Saranti A, Molnar C, Biecek P, Samek W. 2020. Explainable AI methods—a brief overview. *International Workshop on Extending Explainable AI beyond Deep Models and Classifiers*. p 13–38
 73. Novielli P, Magarelli M, Romano D, Di Bitonto P, Stellacci AM, Monaco A, Amoroso N, Bellotti R, Tangaro S. 2025. Leveraging explainable AI to predict soil respiration sensitivity and its drivers for climate change mitigation. *Sci Rep* 15:12527. <https://doi.org/10.1038/s41598-025-96216-y>
 74. Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. 2020. Fooling lime and shap: adversarial attacks on post hoc explanation methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. p 180–186
 75. Kumar M, Ji B, Zengler K, Nielsen J. 2019. Modelling approaches for studying the microbiome. *Nat Microbiol* 4:1253–1267. <https://doi.org/10.1038/s41564-019-0491-9>
 76. Ravikrishnan A, Raman K. 2018. *Systems-level modelling of microbial communities: theory and practice*. CRC Press.
 77. 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:e0205023. <https://doi.org/10.1128/mbio.02050-23>
 78. Baker RE, Peña J-M, Jayamohan J, Jérusalem A. 2018. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biol Lett* 14:20170660. <https://doi.org/10.1098/rsbl.2017.0660>

79. Kuppa Baskaran DK, Umale S, Zhou Z, Raman K, Anantharaman K. 2023. Metagenome-based metabolic modelling predicts unique microbial interactions in deep-sea hydrothermal plume microbiomes. *ISME Commun* 3:42. <https://doi.org/10.1038/s43705-023-00242-8>
80. Sharma A, Kiciman E. 2020. DoWhy: an end-to-end library for causal inference. *arXiv:2011.04216*
81. Gelman A, Loken E. 2014. The statistical crisis in science. *Am Sci* 102:460. <https://doi.org/10.1511/2014.111.460>
82. Kers JG, Saccenti E. 2021. The power of microbiome studies: some considerations on which alpha and beta metrics to use and how to report results. *Front Microbiol* 12:796025. <https://doi.org/10.3389/fmicb.2021.796025>
83. Raman K, Kumar R, Musante CJ, Madhavan S. 2025. Integrating modelinformed drug development with AI: a synergistic approach to accelerating pharmaceutical innovation. *Clin Transl Sci* 18:e70124. <https://doi.org/10.1111/cts.70124>
84. Ürel H, Benassou S, Marti H, Reska T, Sauerborn E, Pinheiro Alves De Souza Y, Perlas A, Rayo E, Biggel M, Kesselheim S, Borel N, Martin EJ, Venegas CB, Schlöter M, Schröder K, Mittelstrass J, Prospero S, Ferguson JM, Urban L. 2024. Nanopore- and AI-empowered metagenomic viability inference. *bioRxiv*. <https://doi.org/10.1101/2024.06.10.598221>
85. Schmidgall S, Ziaei R, Harris C, Reis E, Jopling J, Moor M. 2024. AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. *arXiv:2405.07960*
86. Schmidgall S, Su Y, Wang Z, Sun X, Wu J, Yu X, Liu J, Liu Z, Barsoum E. 2025. Agent Laboratory: using LLM agents as research assistants. *arXiv:2501.04227*
87. Gottweis J, Weng W-H, Daryin A, Tu T, Palepu A, Sirkovic P, Myaskovsky A, Weissenberger F, Rong K, Tanno R, et al. 2025. Towards an AI co-scientist. *arXiv:2502.18864*
88. Tiwari DD, Hoffmann N, Didi K, Deshpande S, Ghosh S, Nguyen TVN, Raman K, Hermjakob H, Sheriff R. 2023. BioModelsML: building a fair and reproducible collection of machine learning models in life sciences and medicine for easy reuse. *bioRxiv*. <https://doi.org/10.1101/2023.05.22.540599>
89. Valderrama B, Calderon-Romero P, Bastiaanssen TFS, Lavelle A, Clarke G, Cryan JF. 2025. The South American MicroBiome Archive (saMBA): enriching the healthy microbiome concept by evaluating uniqueness and biodiversity of neglected populations. *bioRxiv*. <https://doi.org/10.1101/2025.04.03.647034>

AUTHOR BIOS

Shradha Sharma is a third-year PhD student at IIT Madras. Her research work focuses on understanding the marine microbiome using systems biology principles. She earned her master's degree in biomedical engineering from the Indian Institute of Engineering Science.



Hari Priya Narahari is working as a post-baccalaureate research fellow at the Computational Systems Biology Lab, WSAI, IIT Madras. She recently completed a bachelor of science (B.S.) in data science and applications from IIT Madras. Alongside, she pursued and earned a master of technology (M.Tech.) in artificial intelligence and data science from Shiv Nadar University, Chennai. Her research interests include AI for scientific discovery, quantum computing, and agentic AI.



Karthik Raman is a professor at the Department of Data Science and AI at the Wadhvani School of Data Science & AI (WSAI), IIT Madras. Karthik's research group works on developing scalable algorithms and computational tools to understand, predict, and manipulate complex biological networks. Broadly spanning computational aspects of synthetic and systems biology, key research areas in his group encompass microbiome systems biology, *in silico* metabolic engineering, biological network design, and biological big data analysis. Karthik also coordinates the Centre for Integrative Biology and Systems Medicine (IBSE) at IIT Madras. Karthik teaches courses on computational biology and systems biology at IIT Madras and has also authored a textbook on computational systems biology.

