CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	From microbiomes to predictive ecosystems: Challenges and opportunities in artificial intelligence-based approaches
Authors	Yu, Wanqian;Qu, Hua;Wang, Shaopu;Shi, Mang;Lu, Yingzhou;Xia, Shuting;Stanton, Catherine;Zhang, Chenhong;Wang, Yuhao;Zhu, Tongyu;Zhu, Dong;Zhou, Xin;Ross, R. Paul;Zhou, Zhemin;Huang, Yuan
Publication date	2026-05-28
Original Citation	Yu, W, Qu, H, Wang, S, Shi, M, Lu, Y, Xia, S, Stanton, C, Zhang, C, Wang, Y, Zhu, T, Zhu, D, Zhou, X, Ross, R P, Zhou, Z & Huang, Y 2026, 'From microbiomes to predictive ecosystems: Challenges and opportunities in artificial intelligence-based approaches', The Lancet Microbe, vol. 7, no. 10, 101428, pp. 1-14. https://doi.org/10.1016/j.lanmic.2026.101428
Type of publication	Review
Link to publisher's version	https://www.scopus.com/pages/publications/105040213874 - 10.1016/j.lanmic.2026.101428
Rights	© 2026, authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/ licenses/by-nc-nd/4.0/) - https://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2026-11-07 10:18:55
Item downloaded from	https://hdl.handle.net/10468/19526

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

From microbiomes to predictive ecosystems: challenges and opportunities in artificial intelligence-based approaches

Wanqian Yu, Hua Qu*, Shaopu Wang*, Mang Shi*, Yingzhou Lu*, Shuting Xia, Catherine Stanton, Chenhong Zhang, Yuhao Wang, Tongyu Zhu, Dong Zhu, Xin Zhou†, R Paul Ross†, Zhemin Zhou†, Yuan Huang*

Microbiome systems encompass diverse ecological niches and host associations, with their scale and complexity challenging traditional analytical frameworks. Advances in artificial intelligence are transforming microbiome research by enabling improved integrative analyses of microbial genomes, community structure, and functional potential. In this Review, we outline how these developments create opportunities for microbiome research to move beyond descriptive analyses towards predictive modelling and hypothesis generation, including emerging insights into microbial function, host–microbe interactions, and ecosystem-level organisation. As the field grapples with challenges in model interpretability, generalisability, and causal inference, emerging strategies, such as multiomics and multicohort integration, provide promising avenues to deepen biological understanding. Addressing ethical considerations, including data privacy, algorithmic bias, and equitable access, will be crucial for translating artificial intelligence-driven microbiome discoveries into robust and inclusive clinical applications.

Introduction: artificial intelligence and the microbiome revolution

Advances in sequencing and molecular biology technologies have substantially expanded the scope of microbiological research. Early culture-dependent approaches captured less than 1% of environmental microbes, whereas high-throughput sequencing enabled systematic characterisation of largely unculturable microbial communities.[1] Since the introduction of 16S rRNA sequencing in the mid-1990s and shotgun metagenomics subsequently,[2,3] microbiomes have been recognised as dynamic, multikingdom ecosystems governed by ecological principles such as competition, metabolic interdependence, and spatial organisation, similar in complexity to macroecological systems.[4]

In this context, the human microbiome became a central focus of research, leading to the launch of the human microbiome project in 2007.[5] Host-associated microbiomes, for instance, are profoundly influenced by diet, ageing, and environmental exposures, resulting in highly individualised microbial profiles.[6–8] This inherent variability, combined with the scale and heterogeneity of microbiome datasets, introduces substantial analytical challenges that exceed the capacity of conventional methods.[9]

However, traditional microbiome research has long been constrained by high dimensionality, insufficient data utilisation, and interpretive bottlenecks. The increasing complexity of microbiome data has driven the integration of artificial intelligence (AI) into microbiome research. AI has developed powerful approaches for extracting structure from complex, noisy biological data; the convergence of AI methods and microbiological research marks a fundamental shift in how microbial systems are studied, reaching a critical threshold at which integrated AI–microbiome frameworks enable deep insights into ecosystem functioning and therapeutic potential. Figure 1 summarises this convergence, tracing the parallel evolution of AI and microbiome research towards this point of integration.

Advances in AI have transformed biological data analysis. Traditional machine learning methods, such as support vector machines[10] and random forests,[11] enabled early pattern recognition and biomarker discovery in high-dimensional data,[12] whereas deep learning architectures enabled the modelling of more complex and structured biological systems. Recent breakthroughs, exemplified by deep neural networks,[13] graph-based models,[14] and transformer architectures (based on the evidence from a preprint),[15] have further accelerated the emergence of AI for science, a phrase that refers to the use of AI to accelerate scientific discovery through data-driven modelling, prediction, and hypothesis generation, thus reshaping discovery frameworks across genomics and biomedicine.

In AI, data-driven models are trained on large datasets to generate predictive insights. Building on traditional machine learning approaches, deep learning architectures—including convolutional and graph neural networks—capture spatial, relational, and hierarchical features of microbiome data,[16,17] whereas transformer-based and generative models enable flexible representation of ecosystem-scale complexity by modelling long context and functional relationships (figure 2).[18]

Together, the advances in AI and microbiome research support a shift from descriptive microbiome analyses to predictive modelling of community behaviour. We refer to this emerging capability as a predictive ecosystem, which is defined as a framework in which data-driven and mechanistic models are integrated to forecast ecosystem dynamics under perturbation. Such forecasting enables hypothesis testing, rational intervention design, and prospective evaluation of therapeutic strategies. This development, which extends beyond the introduction of new tools, shapes how microbial communities are analysed across scales, from molecular interactions to ecosystem-level dynamics, in coupled host-associated, environmental, or clinical systems.

Lancet Microbe 2026; 7: 101428

Published Online May 29, 2026 https://doi.org/10.1016/j.lanmic.2026.101428

*Contributed equally
†Contributed equally as senior authors

MOE Key Laboratory of Geriatric Diseases and Immunology, The Second Affiliated Hospital of Soochow University, Soochow University, Suzhou, China (W Yu BSc, S Xia PhD, Prof Z Zhou PhD); **Institute of Clinical Pharmacology, Xiyuan Hospital, China Academy of Chinese Medical Sciences, Beijing, China** (H Qu PhD); **Department of Pediatrics, Key Laboratory of Birth Defects and Related Diseases of Women and Children, West China Second University Hospital, Sichuan University, Chengdu, China** (Prof S Wang PhD); **National Key Laboratory of Intelligent Tracking and Forecasting for Infectious Diseases, School of Medicine, Sun Yat-sen University, Shenzhen, China** (Prof M Shi PhD); **Shenzhen Key Laboratory for Systems Medicine in Inflammatory Diseases, Sun Yat-sen University, Shenzhen, China** (Prof M Shi); **Guangdong Provincial Center for Disease Control and Prevention, Guangzhou, China** (Prof M Shi); **Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA** (Y Lu PhD, Prof X Zhou PhD); **Teagasc Food Research Centre, Moorepark, Fermoy, Ireland** (Prof C Stanton PhD); **APC Microbiome Ireland, University College Cork, Cork, Ireland** (Prof C Stanton, R Paul Ross); **State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China** (Prof C Zhang PhD); **Zhejiang Provincial Key Laboratory of Pancreatic Disease, The First**

Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China (Prof Y Wang PhD); **Department of Urology** (Prof T Zhu PhD, Prof D Zhu PhD) **and Department of Kidney Transplantation** (Prof T Zhu), **Zhongshan Hospital, Fudan University, Shanghai, China; Shanghai Key Laboratory of Organ Transplantation, Shanghai, China** (Prof T Zhu); **Shanghai Medical College, Fudan University, Shanghai, China** (Prof T Zhu); **Department of Periodontics, Shanghai Stomatological Hospital** (Prof X Zhou) **and Center on Phenomics and Precision Medicine, Intelligent Medicine Institute** (Prof X Zhou), **Fudan University, Shanghai China; PolyU-Stanford Joint Research Centre on Longitudinal Deep Omics, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region, China** (Prof X Zhou); **National Key Laboratory of Intelligent Tracking and Forecasting for Infectious Diseases, National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China** (Prof Z Zhou); **National Center of Technology Innovation for Biopharmaceuticals, Suzhou Biomedical Industry Innovation Center, Suzhou, China** (Prof Z Zhou); **Jiangsu Province Engineering Research Center of Precision Diagnostics and Therapeutics Development, Suzhou Medical College, Soochow University, Suzhou, China** (Prof Z Zhou); **Biomedical Basic Research Center (BBRC) of Jiangsu, Institute of Health Data Science, Soochow University, Suzhou 215127, China** (Prof Z Zhou); **State Key Laboratory of Cardiovascular Disease, Fuwai Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China** (Y Huang PhD)

Correspondence to: Dr Yuan Huang, State Key Laboratory of Cardiovascular Disease, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical

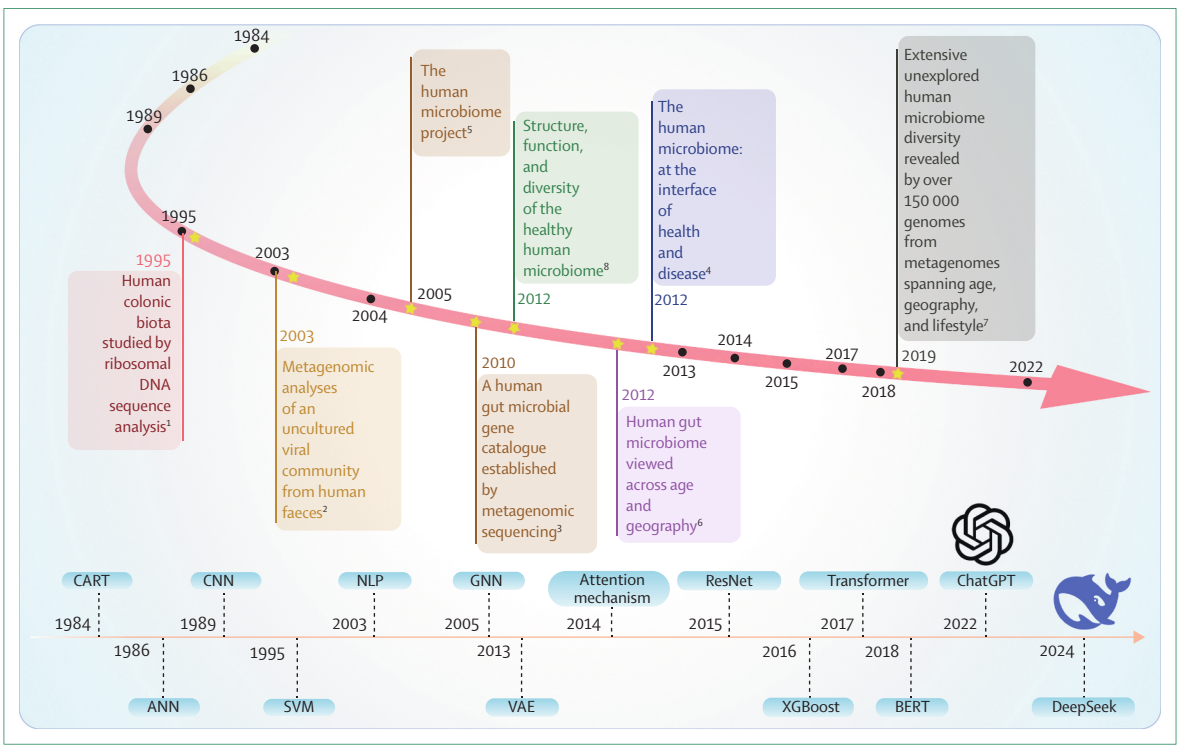


Figure 1: **The developmental timeline of artificial intelligence and landmark papers in microbiome research**
Figure presents a timeline illustrating the parallel development of artificial intelligence, marked by advances in algorithms, models, and microbiome research. The landmark studies that shaped the field are highlighted.[1–8] The circles denote the time points along the timeline; the stars indicate landmark studies. ANN=artificial neural network. BERT=bidirectional encoder representations from transformers. CART=classification and regression tree. CNN=convolutional neural network. GNN=graph neural network. NLP=natural language processing. SVM=support vector machine. VAE=variational autoencoder.

AI-based technologies in microbiome analysis

Current AI applications span multiple stages of the microbiome research pipeline, from data processing to system-level modelling. Together, these advances form an emerging infrastructure for predictive microbiome science, which has the potential to transform basic research and clinical applications. This predictive ecosystem can be viewed as a central integrative framework (figure 3), supported by modular analytical components that enable system-level inference and translational insights. The technologies surveyed in this Review (table 1) reflect increasing technical sophistication and a broader shift towards comprehensive ecosystem modelling and represent a subset of widely adopted and practically relevant tools.

Genome assembly and binning

High-quality genome reconstruction from metagenomic data represents the foundational layer of the predictive ecosystem, as it determines the resolution at which microbial structure and dynamics can be inferred. However, the process of genome reconstruction remains constrained by both technical limitations, such as repetitive genomic regions and short read length, and biological factors, such as the intrinsic spatial organisation of microbial chromosomes.[36,37] Thus, these issues are exacerbated in microorganisms with intricate chromosomal architectures, in which accurate genome reconstruction becomes particularly challenging.

AI-based approaches address these limitations by integrating diverse biological datasets and learning complex patterns that elude rule-based methods. By jointly modelling heterogeneous signals, AI approaches incorporate structural and spatial constraints, which are otherwise difficult to encode explicitly. For example, AutoHiC combines transformer-based models with high-throughput chromosome conformation capture proximity information to perform chromosome-level assemblies, illustrating a shift towards biologically informed reconstruction.[38] With improved assembly quality, emphasis has moved from algorithm development to evaluation and refinement under uncertainty. Tools such as CheckM2 apply machine learning to genome quality estimation, thus extending reliable assessment to novel microbial lineages.[19] Complementary approaches, such as misassembly detection tools (eg, metaMIC) further identify misassemblies and reduce contamination without relying on reference genomes or predefined markers, which is essential for studying previously unexplored host-associated or natural environments.[20,39,40]

Genome binning addresses the challenge of disentangling genome-level units from highly mixed and

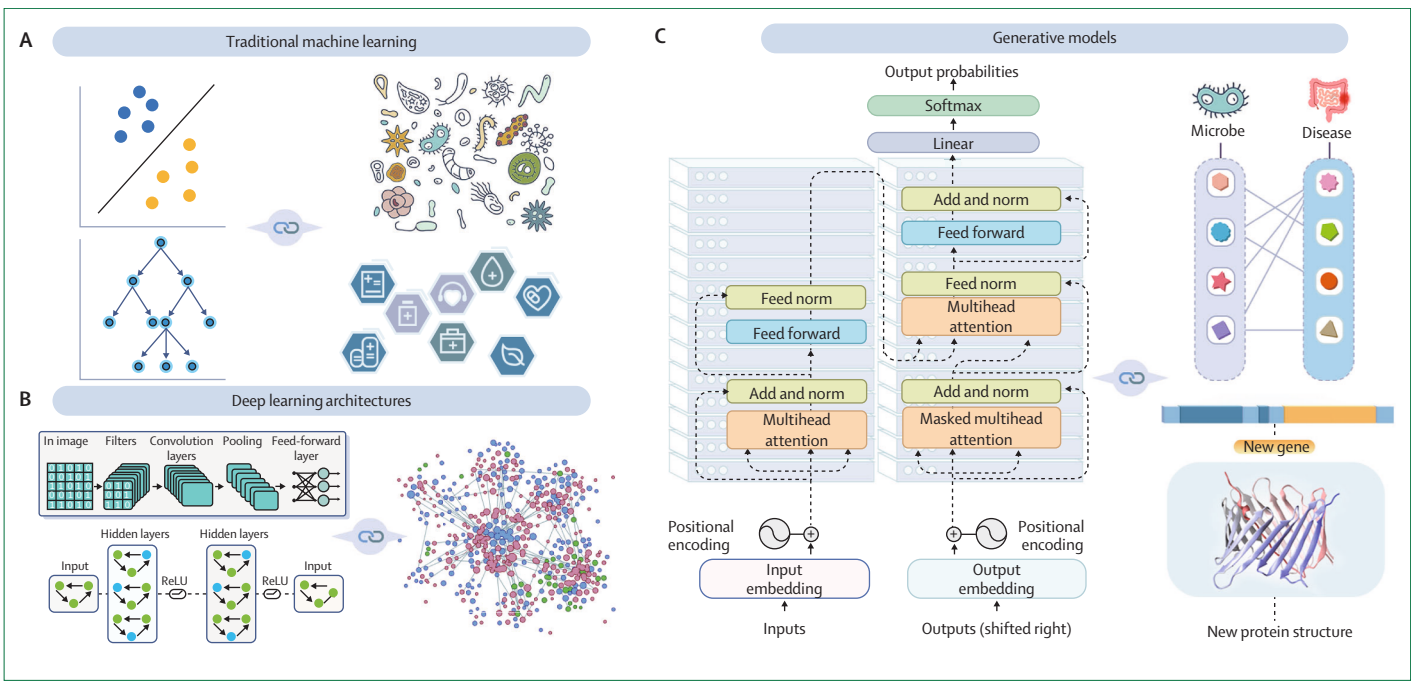


Figure 2: **Overview of representative artificial intelligence-based model classes and their commonly associated tasks in microbiome research**
(A) Traditional machine learning methods, such as decision trees and random forests, are commonly used for pattern recognition in microbiome data and for identifying potential microbial biomarkers. (B) Deep learning approaches, such as convolutional neural networks and graph-based neural networks, are well suited for modelling complex, structured, and interaction-rich nature of microbial communities. (C) Generative models, particularly large language models based on transformer architectures, enable flexible learning of microbiome representations and can be applied to diverse tasks, such as predicting microbe–disease associations, discovering novel genes, and predicting protein structure or function. Norm refers to layer normalisation in transformer blocks. Although transformer models belong to deep learning architectures, they are commonly employed as the backbone of generative language models. ReLU=rectified linear unit.

fragmented metagenomic assemblies. AI has the potential to improve this step further by learning intrinsic compositional patterns and reducing dependence on curated reference databases. For example, self-supervised and autoencoder models perform robustly in complex communities with uneven species abundances.[21,41] Collectively, AI-based advances refine genome-level representations and provide a robust foundation for the predictive modelling of microbial ecosystems.

Identification of mobile genetic elements and modifications

Predictive ecosystem modelling also requires the understanding of dynamic genomic features that mediate microbial adaptation to host-associated or environmental pressures, horizontal gene transfer, and regulatory plasticity. Mobile genetic element (MGE) and DNA methylation play central roles in the reorganisation of microbial genomes in response to environmental pressures, thereby shaping microbial adaptation and community dynamics. AI models in this domain show sophisticated pattern recognition capabilities and can decode complex biological signals embedded in genomic sequences.[42]

For MGE identification, AI-based methods have evolved from single-class identification tools to multiclass classification systems. Several tools now apply machine learning to identify MGEs within mixed assemblies; representative examples are discussed here, and additional commonly used tools are summarised in table 1. geNomad, which achieves robust multiclass classification of plasmids and viruses across diverse environments, is an example of the evolution of such classification systems.[28] Recent developments extend the applications of AI-based tools from classification to synthetic design. Notably, Evo, a large language framework, represents a promising tool that deciphers coevolutionary patterns in coding and non-coding regions, enabling the identification of functional elements such as novel CRISPR and transposon systems, with potential applications in the development of genome engineering tools.[29] Together, these advances signal a shift from descriptive identification to active biological engineering.

Similar to MGE identification, epigenetic modification detection has also progressed from basic identification to comprehensive profiling. Deep learning models operating directly on long-read sequencing data enable accurate detection of common modifications such as 6-methyladenine and 5-methylcytosine even under low-coverage conditions.[30] This real-time capability enables the investigation of epigenetic responses to environmental perturbations. Furthermore, organism-specific models show that tailored methods can improve the predictive precision in particular microbial contexts or for rare modifications, such as the identification of 4-methylcytosine sites in *Escherichia coli* or *Geobacter pickeringii*.[43,44] Emerging multiscale

College, Beijing 100037, China
develandhuangyuan@163.com

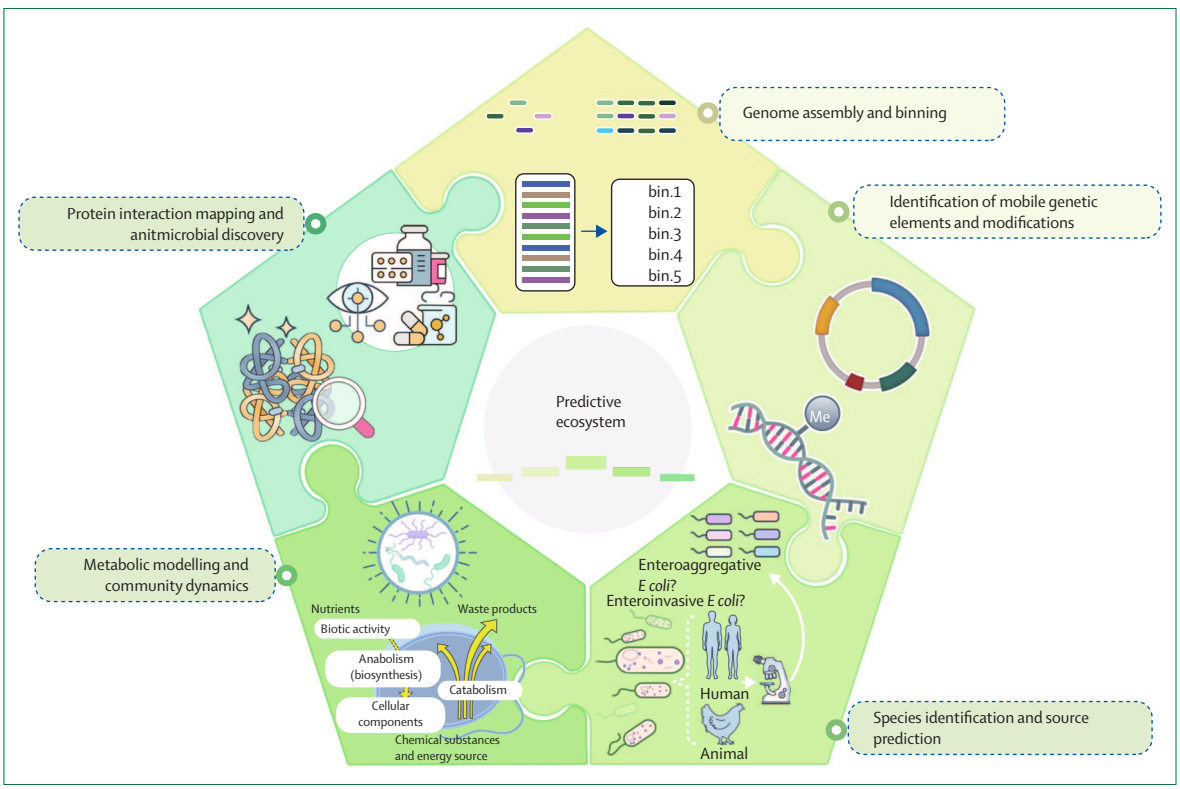


Figure 3: **Artificial intelligence in microbiome science for predictive ecosystem modelling**
The figure depicts five aspects representing a progressive workflow to form the predictive ecosystem. The workflow starts from the assembly and binning of raw sequence data, followed by the identification of mobile genetic elements and sequence modifications. The workflow then moves to species identification and source prediction, followed by characterisation of the dynamics and metabolism of microbial communities. The workflow culminates with the mapping of protein–protein interactions and discovery of antimicrobial peptides. Together, these steps provide a comprehensive understanding of the microbial system as a whole. *E coli*=*Escherichia coli*.

language models further integrate multiple modification types within a unified framework, thus supporting comprehensive profiling of microbial communities.[31]

Species identification and source prediction

Accurate taxonomic assignment and source prediction form crucial components of the predictive ecosystem framework that enable researchers to map ecosystem composition and trace origins across spatial and temporal scales. Traditional culture-based methods and marker gene sequencing often do not possess the speed and resolution required for high-throughput analysis of novel taxa. AI bridges this gap by enhancing both experimental and computational approaches.

Machine learning methods applied to spectroscopic techniques enable rapid, cultivation-free identification of microbial taxa.[45,46] In parallel, sequence-based identification methods extract discriminative features directly from raw reads and assembled contigs of uncultured or poorly characterised lineages, thus reducing overclassification and supporting real-time analysis of long-read sequencing data.[22,47]

Virus classification and discovery present unique challenges because of extreme diversity and incomplete representation in current databases.[48] AI-based methods have progressively expanded the scope of viral detection from metagenomic and metatranscriptomic data at multiple levels, including raw reads[49] and contigs,[50] thereby reducing dependence on predefined viral signatures. Advanced viral discovery tools enable the recovery of diverse types of viruses, spanning DNA and RNA viruses and highly divergent or previously uncharacterised lineages,[25,26] to ensure comprehensive characterisation. This advancement is exemplified by LucaProt, which uses deep learning frameworks to uncover tens of thousands of novel viruses, thus substantially enlarging the known virosphere and illuminating viral dark matter across ecosystems.[27]

Source tracking further extends taxonomic inference towards ecological interpretation with direct practical relevance. Environmental tracing tools can now be used to infer environmental origins, ecological niches, and host associations from genomic features, supporting applications such as contamination tracing, outbreak investigation, and microbial risk assessment.[23,51] These approaches have been applied to genome-based tracing of foodborne pathogens, such as *Listeria*[52] and *Salmonella*,[53] during epidemiological outbreaks. Similar methodologies have been extended to viral

	Applied area	Description	Algorithm(s)	Author, year
CheckM2	Genome assembly	A machine learning-based method for predicting MAG quality, offering improved accuracy, speed, and quick updates to the reference genome database	Gradient-boosted decision tree, neural network	Chklovski et al (2023)[19]
MetaMIC	Genome assembly	A machine learning-based tool that identifies and corrects misassemblies in metagenomic data, improving scaffolding and binning by localising and splitting misassembly breakpoints	Random forest	Lai et al (2022)[20]
SemiBin2	Genome binning	A self-supervised contrastive learning tool for binning metagenomic contigs from short-read and long-read data into high-quality MAGs; it processes a large number of genomes with low runtime and memory usage and is available as open source	Siamese neural network	Pan et al (2023)[21]
IDTAXA	Species identification	A machine learning-based taxonomic classifier that reduces overclassification, outperforms other classifiers (such as QIIME and RDP Classifier), and maintains low error rates across varying sequence lengths	LearnTaxa, IdTaxa (similar to decision tree)	Murali et al (2018)[22]
ONN4MST	Source prediction	An ontology-aware deep learning tool for large-scale microbial source tracking across 114 niches; it achieves near-optimal accuracy with high speed and interpretability and is available as open source	Ontology-aware neural network	Zha et al (2022)[23]
DeepHost	Source prediction	A deep learning tool for predicting phage hosts from genome sequences, achieving high genus-level accuracy and outperforming existing methods by 10–30%; it is available as open source	Spaced k-mer pair encoding, convolutional neural network	Ruohan et al (2022)[24]
VirSorter2	Virome identification	A DNA and RNA virus identification tool with high accuracy across diverse viral genomes, thus minimising errors; it is adaptable to new virus types through modular classifier design	Random forest	Guo et al (2021)[25]
VIBRANT (Virus Identification By iteRative ANnoTation)	Virome identification	A hybrid machine learning-based protein similarity method for automated recovery, annotation, and characterisation of viral genomes from metagenomic assemblies; it highlights viral functions and pathways	Neural network	Kieft et al (2020)[26]
LucaProt	Virome identification	A deep learning tool for detecting divergent RNA viruses by identifying RNA-dependent RNA polymerase sequences in metatranscriptomes	Transformer architecture	Hou et al (2024)[27]
geNomad	MGE detection	A classification and annotation framework to identify plasmids and viruses, achieving high precision and speed, and outperforming other tools (such as PlasFlow and PlasmidFinder) in detecting diverse MGEs	Deep neural network	Camargo et al (2024)[28]
Evo	MGE detection	A genomic model with seven billion parameters for prediction and generation tasks across DNA, RNA, and protein modalities	StripedHyena	Brixi et al (2026)[29]
DeepSignal	Epigenetic profiling	A tool to detect DNA methylation states from Nanopore sequencing reads	Neural network	Ni et al (2019)[30]
iDNA-ABF	Epigenetic profiling	A deep learning model for interpretable DNA methylation prediction; it outperforms existing methods and reveals functional insights	Transformer architecture	Jin et al (2022)[31]
CLEAN (Contrastive Learning enabled Enzyme ANnotation)	Metabolic modelling	A machine learning algorithm for accurate enzyme function annotation, improving enzyme commission number assignment, especially for understudied or promiscuous enzymes	Neural network	Yu et al (2023)[32]
EffectorP	Protein annotation	A machine learning-based tool for predicting fungal effectors from secretome data; it achieves high sensitivity and specificity using sequence-derived features	Naïve Bayes, logistic regression, multilayer perceptron, decision tree, random forest	Sperschneider et al (2016)[33]
AlphaFold3	Protein structural prediction	A model for predicting the structure of protein complexes with improved accuracy across protein–ligand, protein–nucleic acid, and antibody–antigen modalities	Pairformer, generative adversarial network	Abramson et al (2024)[34]
LSTM-PHV	Host–microbiome interaction	A sequence-based machine learning model for predicting protein–protein interactions between humans and viruses; it achieves high accuracy and outperforms existing methods	Long short-term memory, word2vec	Tsukiyama et al (2021)[35]

Tools were selected based on their methodological maturity, reported performance, and broad adoption or representativeness within each task category. MAG=metagenome-assembled genome. MGE=mobile genetic element.

Table 1: **Summary of representative artificial intelligence-powered tools used in microbiome research, categorised by analytical task and modelling approach**

host prediction, linking viral sequences to microbial or animal hosts and providing insights into host–virus interactions and zoonotic spillover potential.[24,54]

Metabolic modelling and community dynamics

Metabolic modelling represents the functional core of the predictive ecosystem framework, translating genomic potential into quantitative predictions of growth, interaction, and environmental response.[55] AI integration has improved metabolic inference by enabling high-resolution pathway reconstruction,[56] precise metabolite prediction,[57] and identification of keystone species that disproportionately influence community function.[58]

AI approaches address metabolic complexity across multiple biological scales rather than treating pathways as static annotations. At the genetic level, machine learning frameworks identify essential metabolic genes[59] and predict gene acquisition and loss patterns underlying microbial adaptation to environmental pressures.[60] Beyond essential gene identification, machine learning models can directly predict previously unannotated functional genes from sequence context.[61] For incomplete genomes, particularly relevant for understudied species, deep learning-based integration of Kyoto Encyclopedia of Genes and Genomes (KEGG) database annotations enhances metabolic inference reliability.[62] At the molecular level, learning-based

enzyme models move beyond homology-based annotation towards capturing underlying biochemical constraints, thus expanding the accessible reaction space and predicting previously uncharacterised functions.[32,63]

Metabolic flux analysis moves beyond static reconstruction and provides quantitative descriptions of reaction rates in metabolic networks. Metabolic flux analysis has been revolutionised by platforms that integrate AI with constraint-based modelling (ie, modelling metabolic behaviour under physicochemical and stoichiometric constraints). Modern frameworks can be used to infer metabolic fluxes by integrating data-driven learning with constraint-based metabolic network.[64] These frameworks can also be used to estimate kinetic parameters under uncertainty (eg, owing to insufficient or noisy experimental measurements) and reconstruct metabolic states from sparse or heterogeneous data.[65,66] These capabilities enable the predictive modelling of cellular behaviour instead of performing post-hoc interpretation and facilitate the identification of candidate metabolic targets that might be leveraged for rational microbiome modulation or therapeutic intervention.

Mechanistic insights have been translated into biotechnological applications using the practical applications of AI. These principles were first established at the level of single organisms. Data-driven metabolic models were used to streamline strain design and metabolic engineering by linking genotype–phenotype relationships to quantitative metabolic predictions; optimising metabolic pathways to enhance the production of target compounds such as fine chemicals or pharmaceuticals is an example.[67,68] The same framework was then extended to engineered and natural communities. Metabolism provides a shared mechanistic layer through which AI models infer resource-mediated interactions among species; this framework enables the prediction of cooperative and competitive behaviours and supports the rational design and modulation of functional microbial consortia, including the optimisation of community composition, metabolic outputs, and interspecies interaction mechanisms such as quorum sensing.[69–71] Together, this progression supports actionable prediction and design of microbial communities for real-world applications, including metabolic engineering, bioproduction, and ecosystem management.

Protein interaction mapping and antimicrobial discovery

Protein-level analysis represents the molecular foundation underlying cellular machinery and community interactions by linking genomic information to executable biological functions. Recent AI advances enable scalable functional annotation, accurate structure prediction, and comprehensive interaction mapping. AI technologies can thus accelerate protein research while opening new possibilities for decoding microbial behaviour and guiding therapeutic applications.

Machine learning methods address persistent challenges in protein function prediction, particularly in microbial and viral systems, in which remote homology and sparse annotation hinder the application of classic methods.[72] By exploiting discriminative sequence features, machine learning models have enabled reliable identification of effector proteins in poorly characterised fungal pathogens.[33]

Structural prediction has further shifted protein analysis from annotation to mechanistic understanding. Together, advances in structure prediction exploit large-scale sequence data to resolve structures of both poorly characterised proteins and higher-order assemblies, enabling mechanistic insights into macromolecular organisation and function, such as predicting protein complexes and biomolecular (eg, protein–protein or protein–nucleic acid) interactions.[34,73]

Besides predicting the structure or function of individual proteins, AI is also reshaping how we predict and interpret functional interactions between proteins, thus extending protein analysis from isolated interaction inference to network-level reconstruction. This shift from pairwise to network-based modelling enables inference of molecular coordination, regulation, and information flow across biological systems.[74]

These capabilities of AI tools enable the translation of molecular insight to practical discovery. This application is exemplified by the AI-driven identification of antimicrobial peptides from microbial community data, host-associated microbiome data, and cryptic genomic region data. Thus, protein-centric modelling is a key bridge linking molecular biology and translational applications such as antibiotic development.[75–77] Notably, several antimicrobial peptides discovered using AI tools have been validated in vitro, although clinical translation remains scarce.

AI-based technologies in host–microbe interactions: from mechanisms to clinical applications

In host-associated microbiomes, ecosystem-level prediction is intrinsically linked to microbial interactions with the host. These host–microbe relationships form highly complex interaction networks in which molecular exchanges shape both ecosystem stability and host physiology. Advances in AI now enable the predictive study of these dynamics instead of descriptive analysis and reveal mechanistic links between microbial ecology and clinical outcomes.[78]

Translating the information on microbial community structure into clinically relevant outcomes requires analytical frameworks capable of modelling high-dimensional, compositionally constrained data. AI-driven approaches capture the organisation of host–microbe interactions, spanning ecological networks of microbial coexistence and molecular interfaces of immune regulation. Using AI, insights from microbial interaction networks are translated into metabolite-level signals, mechanistic understanding, or actionable applications (figure 4).

Microbial ecological networks

Microbial communities exist as sophisticated ecological networks where species interactions influence functional output and ecosystem resilience. Early network inference approaches, largely based on correlation analyses, provided foundational insights into microbial association but struggled to capture the sparse, high-dimensional, and compositional nature of microbiome data.[79]

Subsequent methodological developments reframed microbial coexistence as a structured network inference problem rather than a simple association. Based on evidence from a preprint paper, such methodological approaches might improve the reliable reconstruction of microbial coexistence network architecture at large sample sizes, by learning dependency patterns directly from sparse and compositional community data.[80] These advances have supported scalable inference across large cohorts and have started to capture host-specific variability in microbial network organisation, for example, across different host populations.[81]

Recent advances have also broadened microbial coexistence networks beyond bacteria–bacteria associations to encompass bacteriophage–bacterial host relationships.[82] In parallel, metabolic dependencies are increasingly embedded into microbial community-level representations.[83] These advances support structured reconstructions of microbial communities that integrate ecological scope with functional constraint and provide a mechanistic bridge linking microbial organisation with host health.

Microbe–host immune system interaction

The molecular interface between microbes and host systems represents a domain of extraordinary complexity, in which transcriptional regulation, protein–protein interactions, metabolic exchanges, and immune signalling converge to influence health and disease outcomes. Recent breakthroughs have transformed researchers' ability to systematically explore these interactions, transcending the limitations of traditional experimental approaches that depend on controlled conditions and high-quality samples. Although high-throughput experimental platforms such as BASEHIT have unveiled extensive host–microbe protein interaction repertoires, the scope of such platforms remains constrained by experimental cost, sample availability, and controlled conditions.[84]

AI-driven computational approaches have expanded host–microbe interface analysis from individual molecular interactions to coordinated molecular systems. At the genomic and transcriptomic levels, learning-based models have the potential to uncover patterns of pathogen–host crosstalk that link microbial variation to host regulatory responses and subsequent metabolic coupling.[85,86] Studies on microbial bile acid metabolism that were published between 2020 and 2025 enabled the identification of previously unrecognised enzyme families and metabolite structures. These discoveries show broad population-level prevalence and clear relevance to host physiology,

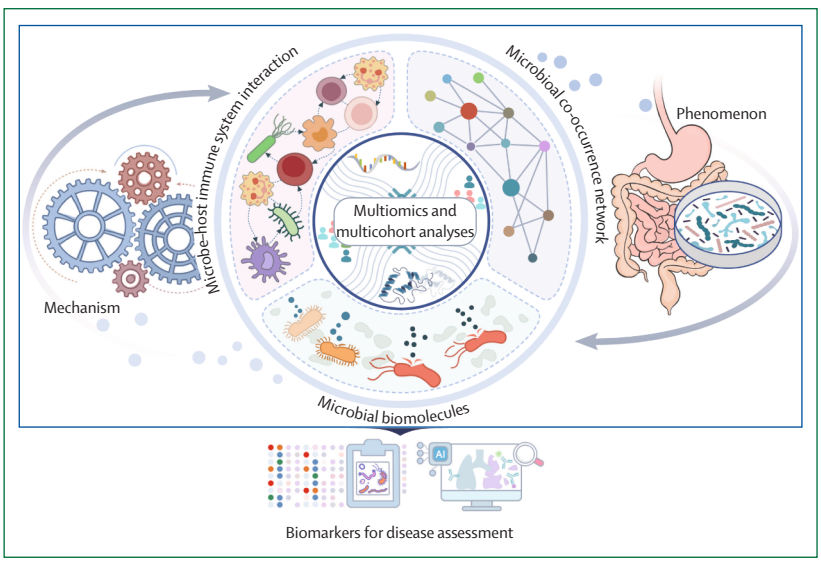


Figure 4: **AI-driven dissection of microbial ecology and host interactions**
AI contributes important insights across multiple layers of biology, from inferred microbial cooccurrence networks that capture community interactions to complex host–immune crosstalk, thereby facilitating biomarker identification and improving diagnostic applications. AI=artificial intelligence.

illustrating how computational approaches can support mechanistic inference with translational potential.[87] Such AI-enabled decoding of pathways provides a concrete route towards biomarker discovery and therapeutic target identification based on microbiome-derived metabolites. At the protein level, sequence-informed inference is increasingly being used to resolve direct interaction interfaces for viral and bacterial pathogens.[88,89] Protein-level interaction modelling enables host–microbe interactions to be interpreted as organised molecular programmes underlying immune modulation and pathogenesis.[35]

Predicting microbial biomarkers for disease assessment

Microbial ecosystems act as a tractable source of non-invasive signals linking community organisation to disease phenotypes. Typical workflows involve microbial profiling from samples such as stool or saliva, followed by AI-driven feature selection to identify taxa associated with disease presence or progression. Model performance is commonly reported using quantitative metrics (eg, area under the curve, accuracy, and precision–recall), reflecting the increasing clinical framing of microbiome-based biomarkers.[90]

Machine learning models have shown the utility of microbial signatures as diagnostic and prognostic biomarkers across diverse diseases. Microbiome-based classifiers often show high accuracy in distinguishing gastrointestinal disease states and subtypes, reflecting the close anatomical and functional coupling between gut microbial communities and host gastrointestinal tract. These findings, together with the observed disease-associated microbial compositional shifts, suggest a potential role of the microbiome in shaping inflammation,

metabolic imbalance, and disease progression in inflammatory bowel disease, metabolic disorders, and gastrointestinal cancers.[91–93] Similar modelling strategies have been extended to immune-mediated, metabolic, and neurological diseases. In these settings, disease-specific microbial signatures are increasingly used to generate biologically informed hypotheses, which link altered community structure to host pathophysiology, rather than serving solely as diagnostic markers.[94–96] Oncological applications further highlight the transformative potential of microbial biomarkers in early cancer detection. Associations between microbial dysbiosis and multiple cancer types support their use in early detection,[93,97] whereas integration of microbiome profiles with clinical variables can enhance the performance of models predicting disease risk and clinical outcomes, compared with the performance of models based on microbiome features alone.[98]

Besides supporting diagnosis, microbial biomarkers can capture dynamic aspects of diseases, including disease activity, progression, infection risk, and treatment response.[99–101] This expanded scope extends the application of the biomarkers from static classification to treatment stratification and outcome prediction. In this context, microbiome-informed models have shown particular value in forecasting therapeutic responses in cancer immunotherapy.[102] Microbial communities thus emerge as dynamic indicators of host physiological state and treatment sensitivity.

Multiomics and multicohort data integration

An increasingly important direction in host–microbe interaction research lies in integrating diverse data types to capture the full spectrum of biological complexity. Multiomics approaches transcend the limitations of single-omics studies by integrating genomic, transcriptomic, proteomic, and metabolomic data with host clinical features. Although these methods capture complementary biological processes, from microbial gene expression to host immune responses, integrating datasets remains challenging because of technical noise, batch effects, and dimensionality disparities.[103]

AI-driven integration provides a principled solution to these constraints by learning structured dependencies across heterogeneous omics layers. By jointly modelling microbial and host-derived features, these approaches transition biomarker discovery based on single-omics associations to that based on cross-omics patterns that reflect disease progression and recovery dynamics.[104,105] Such evidence supports the interpretation of the microbiome as a dynamic biomarker, whose relevance emerges from its functional coupling with host molecular states.

Multicohort analyses complement multiomics approaches by leveraging learning-based models to distinguish disease-associated signals that recur across populations, while accounting for other sources of variation. In gastrointestinal and oncological studies, such integration has revealed microbial patterns linked to clinical outcome, supporting biomarkers that generalise beyond single-cohort observations.[106,107] At the therapeutic level, AI-enabled cross-cohort analyses have linked recurrent microbiome configurations to treatment responses.[108] At a broader ecological scale, similar analyses have revealed stable microbiome states that recur across diseases and populations, thus bridging health and disease. This possibility has been illustrated by the data-driven identification of reproducible community structures and disease-associated microbial signatures across independent cohorts.[109,110]

The emergence of a predictive ecosystem

The convergence of AI technologies with multiscale data integration is shifting microbiome research towards a predictive ecosystem framework, in which biological systems are modelled as interconnected networks across molecular, metabolic, and ecological levels. By integrating multicohort and multiomics data, these approaches enable the identification of patterns that are not accessible to single-cohort or single-omics analyses.

Crucially, this framework extends beyond incremental gains in predictive performance to enable the integration of mechanistic understanding across scales, supporting both forward prediction and system-level interpretation. As these frameworks mature, they are expected to enable predictive ecosystem modelling capable of anticipating responses to perturbations and guiding microbiome-based therapeutic and biotechnological interventions.

Together, these advances mark a transition from descriptive characterisation to predictive and engineering-oriented microbiome science.

The unfinished shift: structural limitations to a predictive microbiome ecosystem

Despite the substantial breadth of AI methods and conceptual advances surveyed in this Review, current efforts remain far from effectively realising the predictive ecosystem framework proposed here. The gap between current AI applications and a true predictive ecosystem reflects structural limitations at multiple levels. These include how biological systems are represented in models, how data enable mechanistic inference, how analytical frameworks couple prediction with interpretation, and how predictive claims are rigorously validated across multiple levels and translated into real-world clinical contexts.

The most crucial barrier lies in the inherent fragility and heterogeneity of microbiome data. Unlike the relatively standardised datasets that enable AI success in other domains, such as computer vision or natural language processing,[111,112] microbiome profiles display extreme variability across individuals, temporal states, and environments. This heterogeneity arises from multiple sources including uneven sequencing depth, technical variability across experimental workflow, the reliance on relative-abundance data in the absence of absolute quantification, incomplete reference databases, and a persistent gap in taxonomic and functional annotation. Consequently, even

	Key considerations	Examples or frameworks	Relevance to microbiome AI
Data standards and sharing	Data standardisation, storage, and interoperability	NCBI or ENA submission standards; FAIR principles	Enables reproducibility and cross-cohort integration
Data privacy and governance	Personal data protection and reidentification risk	GDPR; HIPAA	Microbiome profiles can act as quasi-identifiers
Algorithmic considerations	Bias, robustness, and interpretability	Model validation guidelines	Prevents cohort bias and improves generalisability
Data processing and quality	Noise, batch effects, and composition	Quality control and contamination handling; normalisation and compositionality-aware preprocessing	Crucial for reliable subsequent modelling
Clinical translation	Validation and regulatory approval	AI or ML guidance by the US Food and Drug Administration; clinical trial frameworks	Ensures safe deployment in clinical settings
Equity and access	Population representation and data equity	Global health data initiatives	Avoids under-representation of population groups from low-income and middle-income countries

Representative examples include data governance and privacy regulations (eg, GDPR and HIPAA), data stewardship principles (eg, FAIR), and regulatory guidance for AI-based or ML-based medical technologies.[121–125] AI=artificial intelligence. ENA=European Nucleotide Archive. FAIR=Findable, Accessible, Interoperable, Reusable. GDPR=General Data Protection Regulation. HIPAA=Health insurance of Portability and Accountability Act. ML=machine learning. NCBI=National Center for Biotechnology Information.

Table 2: **Different levels of regulatory and technical considerations in microbiome AI research**

for a single disease, identified biomarkers can differ across cohorts or studies.

These data constraints fundamentally hinder the interpretability and predictive robustness of AI-derived results. For instance, *Akkermansia muciniphila*,[108] which has been frequently identified by machine learning models in human gut microbiome studies and has been found to be clinically relevant (eg, metabolic health and immune modulation),[113,114] displays context-dependent and sometimes contradictory effects across host conditions.[115] Similarly, *Ruminococcaceae* UCG-014, highlighted by machine learning algorithms as health-relevant, remains experimentally intractable under standard laboratory conditions.[116] These examples illustrate a fundamental disconnect: although AI methods excel at identifying statistical patterns, translating those patterns into biological insights requires not only domain knowledge but also the experimental tractability necessary for mechanistic validation, both of which remain incomplete in microbiome science.

A persistent focus on correlation rather than causation remains a fundamental barrier to the realisation of a predictive ecosystem. Despite advances integrating multiomics approaches, many studies remain descriptive and identify statistical associations without resolving underlying mechanisms. This limitation is particularly evident in clinical applications such as faecal microbiota transplantation, in which observed shifts in microbiome composition are frequently interpreted as therapeutic indicators. Although faecal microbiota transplantation has shown clinical efficacy in some contexts (eg, recurrent *Clostridioides difficile* infection and in some cases, ulcerative colitis),[117,118] the interpretation of these compositional changes as mechanistic drivers rather than mere correlates remains a major challenge. Such interpretations often overlook the complex network of host–microbiome interactions and intramicrobial ecological dynamics that ultimately shape ecosystem behaviour and therapeutic outcomes.

The widespread adoption of black box AI models exacerbates this interpretability crisis. Although deep learning architectures and ensemble methods often show impressive predictive performance, they provide only minimal insights into how predictions are generated or which biological features influenced the model decisions. Consequently, these models provide little support for mechanistic interpretation or hypothesis generation; they instead obscure the ecosystem structure and widen the gap between prediction and biological understanding.

Clinical translation faces additional concerns regarding data governance, model bias, and equitable access to emerging AI-driven microbiome technologies. A most pressing challenge is data privacy. Although microbiome data are derived from microbial rather than human genomes, their high individual specificity renders them as quasi-identifiers, raising concerns regarding consent, anonymisation, and crossinstitutional data sharing.[119]

The absence of robust frameworks for de-identifying microbiome data without eroding biological signal hinders crossinstitutional data integration, and in turn, restricts the scale and diversity required for reliable predictive modelling. These constraints are closely intertwined with the issues of algorithmic bias and generalisability. Many AI models are trained on datasets derived from a few high-income countries, leading to systematic underrepresentation of diverse geographical, ethnic, or socioeconomic backgrounds.[120] Therefore, models that perform well in one population might generalise poorly to others, undermining both clinical validity and equity. These challenges persist despite the existence of multiple regulatory and technical frameworks (table 2).

Equitable access further influences the real-world effects of predictive microbiome technologies. Most models are computationally intensive and reliant on high-resolution

omics data, which are rarely available in low-resource settings. Without deliberate efforts to develop resource-efficient algorithms and adaptable deployment strategies, microbiome AI risks reinforcing existing global health disparities. Ethical and infrastructural constraints are intrinsic to the development of predictive microbiome AI, shaping the validity, generalisability, and clinical credibility of predictive claims.

Rethinking the foundation: towards a translational predictive ecosystem

The limitations outlined previously indicate that further progress in microbiome AI cannot be achieved only by incremental improvements in task-specific prediction. A unifying perspective that links predictive modelling with biological interpretation and intervention logic at the ecosystem level should be established. Within such a perspective, the predictive ecosystem framework would reframe predictions as structured inferences over microbial and host processes rather than as the extrapolation of statistical associations.

Current limitations in microbiome AI arise less from the shortage of analytical tools than from the structural noise and heterogeneity inherent to microbiome data. Many existing methods focus on data assessment or filtering but leave the underlying biological and technical variability unresolved, tightly constraining predictive performance. Standardisation of sample collection, processing, and sequencing can reduce technical noise and improve cross-study comparability.[126] However, biological variability across hosts, environments, and temporal states persists beyond protocol harmonisation, motivating the use of AI to model latent ecological structure and dynamics, which otherwise remain inaccessible to conventional analytical approaches.

In this context, multicohort and multicondition study designs offer a practical path towards separating technical artifacts from biologically meaningful variation.[127] By accounting explicitly for population-level heterogeneity, these designs support more reliable inferences than conventional approaches, across diverse host backgrounds and ecological contexts, thus strengthening the empirical basis for predictive modelling.

Building on this empirical foundation, a predictive ecosystem calls for alignment between model structure, inferential logic, biological mechanism, and experimental validation.

From a modelling perspective, architectures aligned with ecological organisation offer a natural starting point. Generative models further offer a conceptual framework for exploring community dynamics under perturbation, enabling the simulation of alternative evolutionary and intervention scenarios. By contrast, graph-based models encode interspecies dependencies and ecological constraints, anchoring prediction in explicit community structure rather than isolated features, as exemplified by graph-based representations of microbial interaction and host–microbe association networks.[128,129]

However, structurally informed models alone are insufficient when inference remains correlation-driven. Translating predictive signals into biological insights requires inference frameworks that combine causal reasoning with interpretability, including causal machine learning approaches and model-level attribution methods (eg, Shapley Additive exPlanation [SHAP]-like analyses and attention-based or gradient-based attribution), to distinguish genuine effects from spurious associations and relate microbial changes to host outcomes.[130]

To establish causal claims, inferences need to be anchored to biological mechanisms. Integration of multiomics approaches acts as an essential bridge that links microbial composition to functional states and host responses. Establishing such links widens the interpretations to move beyond taxonomic association.[131] Coupling these data with mechanistic or constraint-based models—such as metabolic network or biophysical frameworks—offers a principled way to embed biological plausibility into inference and strengthen interpretability at the ecosystem level.[132]

Ultimately, predictive ecosystem intelligence requires experimental validation embedded in an iterative modelling cycle. Coupling AI-driven prediction with empirical feedback supports continuous refinement rather than static inference.

Clinical translation of the predictive ecosystem is constrained not only by technical performance but also by challenges related to data privacy, trust, and equitable deployment. Privacy-preserving distributed learning and trusted computing environments enable crossinstitutional modelling without direct data sharing; however, they do not obviate the need for robust governance and transparency. Moreover, the computational demands of large-scale ecosystem models risk exacerbating existing inequities in research capacity across institutions and regions, thus hindering the adoption of such models in resource-constrained settings and concentrating validation within well-funded institutions. Addressing these barriers will require not only algorithmic innovation but also deliberate design choices that prioritise efficiency, transparency, and broad accessibility.

Conclusions

The convergence of AI and microbiome research, from early computational approaches to current sophisticated platforms, represents more than technological advancement. Such a convergence marks the emergence of a predictive ecosystem as a unifying framework for understanding, modelling, and steering microbial communities as dynamic biological systems.

Despite the rapid expansion of AI-driven tools reviewed here, their transformative effects remain constrained by persistent challenges in data integration, interpretability, and clinical translation. These limitations are not

Search strategy and selection criteria

We adopted a scoping-informed, narrative approach and identified representative and influential studies at the intersection of artificial intelligence (AI) and microbiome research to inform this Review. We initially performed targeted searches of PubMed and Web of Science for studies (including original research articles, with a small number of review papers where relevant) published from the date of inception of the database to Dec 31, 2025. Only articles in English were included. The primary search strategy combined microbiome-related terms, including "microbiome", "microbiota", "gut microbiome", and "gut microbiota", with AI-related terms such as "machine learning", "deep learning", "artificial intelligence" and "neural networks". For specific sections, additional keywords were incorporated to refine the search according to the topic of interest. For example, terms such as "genome assembly" and "binning" were included when focusing on genome reconstruction, and other section-specific keywords were applied as appropriate. This initial search was complemented by iterative expansion, including screening of reference lists and identification of relevant studies not captured by predefined keywords. We prioritised studies published between 2020 and 2025, along with earlier landmark studies with clear relevance to microbiome data analysis and predictive or translational applications. Titles and abstracts were first reviewed to remove records that were clearly outside the scope of this Review, followed by full-text examination of potentially relevant articles.

During full-text screening, priority was given to studies in which AI-based methods were applied to microbiome data analysis or demonstrated clear relevance to predictive or translational applications. These included studies addressing microbiome-related questions across different biological and analytical levels, including microbial community composition, strain-level variation, functional potential, host–microbiome interactions, and clinical prediction. Studies were selected based on their relevance to the proposed conceptual framework and how well the studies exemplify important research methods, data integration strategies, and emerging translational directions rather than on formal methodological quality assessment.

peripheral technical issues but core requirements for responsible and generalisable microbiome intelligence; thus, a shift from fragmented, task-specific solutions to integrated, mechanistically grounded frameworks is required to address these limitations.

Achieving this transition will require deliberate integration of ecological theory, principled inference, and iterative experimental validation into cohesive systems capable of forward prediction, causal reasoning, and rational intervention. Microbiome research is entering a new phase, in which advances in genome reconstruction, taxonomic profiling, and metabolic modelling are increasingly integrated to support ecosystem-level intelligence. This shift from correlation to causation, from description to prediction, and from isolated tools to integrated systems opens new opportunities for studying and applying microbiome knowledge, with the predictive ecosystem serving as a unifying conceptual framework for both basic and translational research. In practice, realising this vision will require coordinated efforts to develop benchmarked, interoperable datasets; mechanistically interpretable models; and prospective validation pipelines (all of which can link predictions to experimental and clinical decision making), alongside robust governance frameworks that address data privacy, ethical oversight, and equitable access.

Contributors

WY participated in conceptualisation, methodology, investigation, visualisation, writing–original draft, and writing–review and editing. HQ and MS contributed to reviewing and editing of the manuscript, including critical evaluation and refinement of key arguments. SW, YL, SX, CS, YW, CZ, DZ, and TZ contributed to the review and editing of manuscript. RPR and XZ contributed to conceptualisation, methodology, supervision, and writing–review and editing. YH and ZZ contributed to conceptualisation, methodology, supervision, visualisation, and writing–review and editing. All authors reviewed and approved the final manuscript.

Declaration of interests

We declare no competing interests.

Acknowledgments

The work was funded by the Prevention and Control of Emerging and Major Infectious Diseases-National Science and Technology Major Project (number 2025ZD01900401); the Shenzhen Medical Research Fund (number B2403009); the National Natural Science Foundation of China (numbers 82300345, 32370099, and 82104678); the National High-Level Hospital Clinical Research Funding, China (number 2023-GSP-QN-23); the Provincial-level Talent Program for National Center of Technology Innovation for Biopharmaceuticals (number NCTIB2024JS0101); the Suzhou Top-Notch Talent Groups (number ZXD2022003); Beijing Municipal Natural Science Foundation (number JQ24060); Xiyuan Hospital of China Academy of Chinese Medical Sciences Enhancement Fund (number XYZY0304-04); Young Elite Scientists Sponsorship Program by the China Association for Science and Technology (number ZZ15-YQ-006); Fundamental Research Funds for the Central Universities, Peking Union Medical College (number 3332025033); Young Elite Scientists Sponsorship Program of the Beijing High Innovation Plan (number 20250948); and National Science and Technology Major Project on Prevention and Treatment of Cancer, Cardiovascular and Cerebrovascular Diseases, Respiratory Diseases and Metabolic Diseases (number 2026ZD0555105). We thank our colleagues and collaborators for the valuable discussions and constructive feedback that contributed to this work. During the preparation of this work, the authors used ChatGPT (OpenAI, GPT-4 class model; accessed in 2025) to edit the language and improve the clarity and readability of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication The tool did not contribute to the scientific interpretation or synthesis of the reviewed studies.

References

1 Wilson KH, Blitchington RB. Human colonic biota studied by ribosomal DNA sequence analysis. *Appl Environ Microbiol* 1996; **62:** 2273–78.

2 Breitbart M, Hewson I, Felts B, et al. Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* 2003; **185:** 6220–23.

3 Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 2010; **464:** 59–65.

4 Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. *Nat Rev Genet* 2012; **13:** 260–70.

5 Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. *Nature* 2007; **449:** 804–10.

6 Yatsunenko T, Rey FE, Manary MJ, et al. Human gut microbiome viewed across age and geography. *Nature* 2012; **486:** 222–27.

7 Pasolli E, Asnicar F, Manara S, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* 2019; **176:** 649–62.e20.

8 Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* 2012; **486:** 207–14.

9 Peters A, Nawrot TS, Baccarelli AA. Hallmarks of environmental insults. *Cell* 2021; **184:** 1455–68.

10 Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995; **20:** 273–97.

11 Breiman L. Random forests. *Mach Learn* 2001; **45:** 5–32.

12 Loomba R, Seguritan V, Li W, et al. Gut microbiome-based metagenomic signature for non-invasive detection of advanced fibrosis in human nonalcoholic fatty liver disease. *Cell Metab* 2017; **25:** 1054–62.e5.

13 Otter DW, Medina JR, Kalita JK. A survey of the usages of deep learning for natural language processing. *IEEE Trans Neural Netw Learn Syst* 2021; **32:** 604–24.

14 Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. The graph neural network model. *IEEE Trans Neural Netw* 2009; **20:** 61–80.

15 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *arXiv* 2023; published online Aug 2. https://doi.org/10.48550/arXiv.1706.03762 (preprint).

16 Ho CS, Jean N, Hogan CA, et al. Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. *Nat Commun* 2019; **10:** 4927.

17 Syama K, Jothi JAA, Khanna N. Automatic disease prediction from human gut metagenomic data using boosting GraphSAGE. *BMC Bioinform* 2023; **24:** 126.

18 Aminian-Dehkordi J, Parsa M, Dickson A, Mofrad MRK. SIMBA-GNN: mechanistic graph learning for microbiome prediction. *NPJ Syst Biol Appl* 2025; **12:** 8.

19 Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat Methods* 2023; **20:** 1203–12.

20 Lai S, Pan S, Sun C, et al. metaMIC: reference-free misassembly identification and correction of de novo metagenomic assemblies. *Genome Biol* 2022; **23:** 242.

21 Pan S, Zhao XM, Coelho LP. SemiBin2: self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing. *Bioinformatics* 2023; **39** (suppl 1): i21–29.

22 Murali A, Bhargava A, Wright ES. IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome* 2018; **6:** 140.

23 Zha Y, Chong H, Qiu H, et al. Ontology-aware deep learning enables ultrafast and interpretable source tracking among sub-million microbial community samples from hundreds of niches. *Genome Med* 2022; **14:** 43.

24 Ruohan W, Xianglilan Z, Jianping W, Shuai Cheng LI. DeepHost: phage host prediction with convolutional neural network. *Brief Bioinform* 2022; **23:** bbab385.

25 Guo J, Bolduc B, Zayed AA, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* 2021; **9:** 37.

26 Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* 2020; **8:** 90.

27 Hou X, He Y, Fang P, et al. Using artificial intelligence to document the hidden RNA virosphere. *Cell* 2024; **187:** 6929–42.e16.

28 Camargo AP, Roux S, Schulz F, et al. Identification of mobile genetic elements with geNomad. *Nat Biotechnol* 2024; **42:** 1303–12.

29 Brixi G, Durrant MG, Ku J, et al. Genome modelling and design across all domains of life with Evo 2. *Nature* 2026; published online March 4. https://doi.org/10.1038/s41586-026-10176-5.

30 Ni P, Huang N, Zhang Z, et al. DeepSignal: detecting DNA methylation state from nanopore sequencing reads using deep-learning. *Bioinformatics* 2019; **35:** 4586–95.

31 Jin J, Yu Y, Wang R, et al. iDNA-ABF: multi-scale deep biological language learning model for the interpretable prediction of DNA methylations. *Genome Biol* 2022; **23:** 219.

32 Yu T, Cui H, Li JC, Luo Y, Jiang G, Zhao H. Enzyme function prediction using contrastive learning. *Science* 2023; **379:** 1358–63.

33 Sperschneider J, Gardiner DM, Dodds PN, et al. EffectorP: predicting fungal effector proteins from secretomes using machine learning. *New Phytol* 2016; **210:** 743–61.

34 Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 2024; **630:** 493–500.

35 Tsukiyama S, Hasan MM, Fujii S, Kurata H. LSTM-PHV: prediction of human-virus protein-protein interactions by LSTM with word2vec. *Brief Bioinform* 2021; **22:** bbab228.

36 Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* 2020; **21:** 30.

37 Koren S, Phillippy AM. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Curr Opin Microbiol* 2015; **23:** 110–20.

38 Jiang Z, Peng Z, Wei Z, et al. A deep learning-based method enables the automatic and accurate assembly of chromosome-level genomes. *Nucleic Acids Res* 2024; **52:** e92.

39 Mineeva O, Rojas-Carulla M, Ley RE, Schölkopf B, Youngblut ND. DeepMAsED: evaluating the quality of metagenomic assemblies. *Bioinformatics* 2020; **36:** 3011–17.

40 Zou B, Wang J, Ding Y, et al. A multi-modal deep language model for contaminant removal from metagenome-assembled genomes. *Nat Mach Intell* 2024; **6:** 1245–55.

41 Nissen JN, Johansen J, Allesøe RL, et al. Improved metagenome binning and assembly using deep variational autoencoders. *Nat Biotechnol* 2021; **39:** 555–60.

42 Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet* 2019; **20:** 389–403.

43 Zulfiqar H, Sun Z-J, Huang QL, et al. Deep-4mCW2V: a sequence-based predictor to identify N4-methylcytosine sites in *Escherichia coli*. *Methods* 2022; **203:** 558–63.

44 Zulfiqar H, Huang QL, Lv H, Sun ZJ, Dao FY, Lin H. Deep-4mCGP: a deep learning approach to predict 4mC sites in *Geobacter pickeringii* by using correlation-based feature selection technique. *Int J Mol Sci* 2022; **23:** 1251.

45 Papagiannopoulou C, Parchen R, Rubbens P, Waegeman W. Fast pathogen identification using single-cell matrix-assisted laser desorption/ionization-aerosol time-of-flight mass spectrometry data and deep learning methods. *Anal Chem* 2020; **92:** 7523–31.

46 Lasch P, Stämmler M, Zhang M, Baranska M, Bosch A, Majzner K. FT-IR hyperspectral imaging and artificial neural network analysis for identification of pathogenic bacteria. *Anal Chem* 2018; **90:** 8896–904.

47 Bao Y, Wadden J, Erb-Downward JR, et al. SquiggleNet: real-time, direct classification of nanopore signals. *Genome Biol* 2021; **22:** 298.

48 Simmonds P, Adams MJ, Benkő M, et al. Virus taxonomy in the age of metagenomics. *Nat Rev Microbiol* 2017; **15:** 161–68.

49 Tampuu A, Bzhalava Z, Dillner J, Vicente R. ViraMiner: deep learning on raw DNA sequences for identifying viral genomes in human samples. *PLoS One* 2019; **14:** e0222271.

50 Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* 2017; **5:** 69.

51 Knights D, Kuczynski J, Charlson ES, et al. Bayesian community-wide culture-independent microbial source tracking. *Nat Methods* 2011; **8:** 761–63.

52 Gu W, Cui Z, Stroika S, et al. Predicting food sources of *Listeria monocytogenes* based on genomic profiling using random forest model. *Foodborne Pathog Dis* 2023; **20:** 579–86.

53 Zhang S, Li S, Gu W, et al. Zoonotic source attribution of *Salmonella enterica* serotype Typhimurium using genomic surveillance data, United States. *Emerg Infect Dis* 2019; **25:** 82–91.

54 Liu Y, Li J, Li P, et al. ARNLE model identifies prevalence potential of SARS-CoV-2 variants. *Nat Mach Intell* 2024; **7:** 18–28.

55 Ye D, Li X, Shen J, Xia X. Microbial metabolomics: from novel technologies to diversified applications. *TrAC Trends Anal Chem* 2022; **148:** 116540.

56 Dale JM, Popescu L, Karp PD. Machine learning methods for metabolic pathway prediction. *BMC Bioinform* 2010; **11:** 15.

57 Han S, Van Treuren W, Fischer CR, et al. A metabolomics pipeline for the mechanistic interrogation of the gut microbiome. *Nature* 2021; **595:** 415–20.

58 Xun W, Liu Y, Li W, et al. Specialized metabolic functions of keystone taxa sustain soil microbiome stability. *Microbiome* 2021; **9:** 35.

59 Nandi S, Subramanian A, Sarkar RR. An integrative machine learning strategy for improved prediction of essential genes in *Escherichia coli* metabolism using flux-coupled features. *Mol Biosyst* 2017; **13:** 1584–96.

60 Konno N, Iwasaki W. Machine learning enables prediction of metabolic system evolution in bacteria. *Sci Adv* 2023; **9:** eadc9130.

61 Kim GB, Kim JY, Lee JA, Norsigian CJ, Palsson BO, Lee SY. Functional annotation of enzyme-encoding genes using deep learning with transformer layers. *Nat Commun* 2023; **14:** 7370.

62 Geller-McGrath D, Konwar KM, Edgcomb VP, et al. Predicting metabolic modules in incomplete bacterial genomes with MetaPathPredict. *eLife* 2024; **13:** e85749.

63 Mellor J, Grigoras I, Carbonell P, Faulon JL. Semisupervised Gaussian process for automated enzyme search. *ACS Synth Biol* 2016; **5:** 518–28.

64 Wu SG, Wang Y, Jiang W, et al. Rapid prediction of bacterial heterotrophic fluxomics using machine learning and constraint programming. *PLoS Comput Biol* 2016; **12:** e1004838.

65 Choudhury S, Moret M, Salvy P, Weilandt D, Hatzimanikatis V, Miskovic L. Reconstructing kinetic models for dynamical studies of metabolism using generative adversarial networks. *Nat Mach Intell* 2022; **4:** 710–19.

66 Andreozzi S, Miskovic L, Hatzimanikatis V. iSCHRUNK–*In* silico approach to characterization and reduction of uncertainty in the kinetic models of genome-scale metabolic networks. *Metab Eng* 2016; **33:** 158–68.

67 Carbonell P, Jervis AJ, Robinson CJ, et al. An automated design-build-test-learn pipeline for enhanced microbial production of fine chemicals. *Commun Biol* 2018; **1:** 66.

68 Zhou Y, Li G, Dong J, Xing XH, Dai J, Zhang C. MiYA, an efficient machine-learning workflow in conjunction with the YeastFab assembly strategy for combinatorial optimization of heterologous metabolic pathways in *Saccharomyces cerevisiae*. *Metab Eng* 2018; **47:** 294–302.

69 Culley C, Vijayakumar S, Zampieri G, Angione C. A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth. *Proc Natl Acad Sci U S A* 2020; **117:** 18869–79.

70 Dama AC, Kim KS, Leyva DM, et al. BacterAI maps microbial metabolism without prior knowledge. *Nat Microbiol* 2023; **8:** 1018–25.

71 Wu S, Feng J, Liu C, et al. Machine learning aided construction of the quorum sensing communication network for human gut microbiota. *Nat Commun* 2022; **13:** 3079.

72 Flamholz ZN, Biller SJ, Kelly L. Large language models improve annotation of prokaryotic viral proteins. *Nat Microbiol* 2024; **9:** 537–49.

73 Zheng W, Wuyun Q, Li Y, Zhang C, Freddolino PL, Zhang Y. Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data. *Nat Methods* 2024; **21:** 279–89.

74 Sun T, Zhou B, Lai L, Pei J. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinform* 2017; **18:** 277.

75 Ma Y, Guo Z, Xia B, et al. Identification of antimicrobial peptides from the human gut microbiome using deep learning. *Nat Biotechnol* 2022; **40:** 921–31.

76 Maasch JRMA, Torres MDT, Melo MCR, de la Fuente-Nunez C. Molecular de-extinction of ancient antimicrobial peptides enabled by machine learning. *Cell Host Microbe* 2023; **31:** 1260–74.e6.

77 Santos-Júnior CD, Torres MDT, Duan Y, et al. Discovery of antimicrobial peptides in the global microbiome with machine learning. *Cell* 2024; **187:** 3761–78.e16.

78 Xia S, Jiang D, Zhou Q, et al. Unlocking the healthy human microbiome: redefining core microbial signatures. *Acta Pharm Sin B* 2025; **15:** 1189–92.

79 Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. *PLOS Comput Biol* 2012; **8:** e1002687.

80 Morton JT, Silverman J, Tikhonov G, Lähdesmäki H, Bonneau R. Scalable estimation of microbial co-occurrence networks with variational autoencoders. *bioRxiv* 2021; published online Nov 11. https://doi.org/10.1101/2021.11.09.467939 (preprint).

81 Vatanen T, Kostic AD, d'Hennezel E, et al. Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. *Cell* 2016; **165:** 842–53.

82 Tang T, Hou S, Fuhrman JA, Sun F. Phage-bacterial contig association prediction with a convolutional neural network. *Bioinformatics* 2022; **38** (suppl 1)**:** i45–52.

83 Silva-Andrade C, Rodriguez-Fernández M, Garrido D, Martin AJM. Using metabolic networks to predict cross-feeding and competition interactions between microorganisms. *Microbiol Spectr* 2024; **12:** e0228723.

84 Sonnert ND, Rosen CE, Ghazi AR, et al. A host-microbiota interactome reveals extensive transkingdom connectivity. *Nature* 2024; **628:** 171–79.

85 Priya S, Burns MB, Ward T, et al. Identification of shared and disease-specific host gene-microbiome associations across human diseases using multi-omic integration. *Nat Microbiol* 2022; **7:** 780–95.

86 Morton JT, Aksenov AA, Nothias LF, et al. Learning representations of microbe–metabolite interactions. *Nat Methods* 2019; **16:** 1306–14.

87 Ding Y, Luo X, Guo J, et al. Identification of gut microbial bile acid metabolic enzymes via an AI-assisted pipeline. *Cell* 2025; **188:** 6012–27.e20.

88 Kösesoy İ, Gök M, Öz C. A new sequence based encoding for prediction of host-pathogen protein interactions. *Comput Biol Chem* 2019; **78:** 170–77.

89 Ahmed I, Witbooi P, Christoffels A. Prediction of human-*Bacillus anthracis* protein-protein interactions using multi-layer neural network. *Bioinformatics* 2018; **34:** 4159–64.

90 Wirbel J, Zych K, Essex M, et al. Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. *Genome Biol* 2021; **22:** 93.

91 Das A, O'Herlihy E, Shanahan F, O'Toole PW, Jeffery IB. The fecal mycobiome in patients with irritable bowel syndrome. *Sci Rep* 2021; **11:** 124.

92 Zhu J, Yin J, Chen J, et al. Integrative analysis with microbial modelling and machine learning uncovers potential alleviators for ulcerative colitis. *Gut Microbes* 2024; **16:** 2336877.

93 Shah MS, DeSantis TZ, Weinmaier T, et al. Leveraging sequence-based faecal microbial community survey data to identify a composite biomarker for colorectal cancer. *Gut* 2018; **67:** 882–91.

94 Qian X, Liu Y-X, Ye X, et al. Gut microbiota in children with juvenile idiopathic arthritis: characteristics, biomarker identification, and usefulness in clinical prediction. *BMC Genomics* 2020; **21:** 286.

95 Gou W, Ling C, He Y, et al. Interpretable machine learning framework reveals robust gut microbiome features associated with type 2 diabetes. *Diabetes Care* 2021; **44:** 358–66.

96 Su Q, Wong OWH, Lu W, et al. Multikingdom and functional gut microbiota markers for autism spectrum disorder. *Nat Microbiol* 2024; **9:** 2344–55.

97 Zheng Y, Fang Z, Xue Y, et al. Specific gut microbiome signature predicts the early-stage lung cancer. *Gut Microbes* 2020; **11:** 1030–42.

98 Fonseca DC, Marques Gomes da Rocha I, Depieri Balmant B, et al. Evaluation of gut microbiota predictive potential associated with phenotypic characteristics to identify multifactorial diseases. *Gut Microbes* 2024; **16:** 2297815.

99 Tap J, Derrien M, Törnblom H, et al. Identification of an intestinal microbiota signature associated with severity of irritable bowel syndrome. *Gastroenterology* 2017; **152:** 111–23.e8.

100 Aasmets O, Lüll K, Lang JM, et al. Machine learning reveals time-varying microbial predictors with complex effects on glucose regulation. *mSystems* 2021; **6:** e01191-20.

101 Montassier E, Al-Ghalith GA, Ward T, et al. Pretreatment gut microbiome predicts chemotherapy-related bloodstream infection. *Genome Med* 2016; **8:** 49.

102 Stein-Thoeringer CK, Saini NY, Zamir E, et al. A non-antibiotic-disrupted gut microbiome is associated with clinical responses to CD19-CAR-T cell cancer immunotherapy. *Nat Med* 2023; **29:** 906–16.

103 Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: a review. *Biotechnol Adv* 2021; **49:** 107739.

104 Singh A, Shannon CP, Gautier B, et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* 2019; **35:** 3055–62.

105 Ren Z, Wang H, Cui G, et al. Alterations in the human oral and gut microbiomes and lipidomics in COVID-19. *Gut* 2021; **70:** 1253–65.

106 Dai Z, Coker OO, Nakatsu G, et al. Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. *Microbiome* 2018; **6:** 70.

107 Li M, Liu J, Zhu J, et al. Performance of gut microbiome as an independent diagnostic tool for 20 diseases: cross-cohort validation of machine-learning classifiers. *Gut Microbes* 2023; **15:** 2205386.

108 Lee KA, Thomas AM, Bolte LA, et al. Cross-cohort gut microbiome associations with immune checkpoint inhibitor response in advanced melanoma. *Nat Med* 2022; **28:** 535–44.

109 Wu G, Xu T, Zhao N, et al. A core microbiome signature as an indicator of health. *Cell* 2024; **187:** 6550–65.e11.

110 Wirbel J, Pyl PT, Kartal E, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat Med* 2019; **25:** 679–89.

111 Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: Proceedings of 2009 IEEE conference on computer vision and pattern recognition. IEEE, 2009: 248–55.

112 Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, eds. Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). Association for Computational Linguistics, 2019: 4171–86.

113 Plovier H, Everard A, Druart C, et al. A purified membrane protein from *Akkermansia muciniphila* or the pasteurized bacterium improves metabolism in obese and diabetic mice. *Nat Med* 2017; **23:** 107–13.

114 Routy B, Le Chatelier E, Derosa L, et al. Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science* 2018; **359:** 91–97.

115 Wang W, Li N, Xu H, et al. ILC3s regulate the gut microbiota via host intestinal galactosylation to limit pathogen infection in mice. *Nat Microbiol* 2025; **10:** 654–66.

116 Yang J, Moon HE, Park HW, et al. Brain tumor diagnostic model and dietary effect based on extracellular vesicle microbiome data in serum. *Exp Mol Med* 2020; **52:** 1602–13.

117 van Nood E, Vrieze A, Nieuwdorp M, et al. Duodenal infusion of donor feces for recurrent *Clostridium difficile*. *N Engl J Med* 2013; **368:** 407–15.

118 Paramsothy S, Kamm MA, Kaakoush NO, et al. Multidonor intensive faecal microbiota transplantation for active ulcerative colitis: a randomised placebo-controlled trial. *Lancet* 2017; **389:** 1218–28.

119 Valles-Colomer M, Blanco-Míguez A, Manghi P, et al. The person-to-person transmission landscape of the gut and oral microbiomes. *Nature* 2023; **614:** 125–35.

120 Condori-Catachura S, Ahannach S, Ticlla M, et al. Diversity in women and their vaginal microbiota. *Trends Microbiol* 2025; **33:** 1163–72.

121 Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016; **3:** 160018.

122 Kuziemsky CE, Chrimes D, Minshall S, Mannerow M, Lau F. AI quality standards in health care: rapid umbrella review. *J Med Internet Res* 2024; **26:** e54705.

123 Shiferaw KB, Roloff M, Balaur I, Welter D, Waltemath D, Zeleke AA. Guidelines and standard frameworks for artificial intelligence in medicine: a systematic review. *JAMIA Open* 2025; **8:** ooae155.

124 Kumar R, Sporn K, Waisberg E, et al. Navigating healthcare AI governance: the comprehensive algorithmic oversight and stewardship framework for risk and equity. *Health Care Anal* 2025; published online Aug 13. https://doi.org/10.1007/s10728-025-00537-y.

125 Liu M, Ning Y, Teixayavong S, et al. A scoping review and evidence gap analysis of clinical AI fairness. *NPJ Digit Med* 2025; **8:** 360.

126 Bindels LB, Watts JEM, Theis KR, et al. A blueprint for contemporary studies of microbiomes. *Microbiome* 2025; **13:** 95.

127 Lin Y, Lau HC-H, Liu C, et al. Multi-cohort analysis reveals colorectal cancer tumor location-associated fecal microbiota and their clinical impact. *Cell Host Microbe* 2025; **33:** 589–601.e3.

128 Zhang H, Lu G, Zhan M, Zhang B. Semi-supervised classification of graph convolutional networks with laplacian rank constraints. *Neural Process Lett* 2022; **54:** 2645–56.

129 Faust K, Raes J. Microbial interactions: from networks to models. *Nat Rev Microbiol* 2012; **10:** 538–50.

130 Hertel J, Heinken A, Fässler D, Thiele I. Causal inference on microbiome-metabolome relations in observational host-microbiome data via in silico in vivo association pattern analyses. *Cell Rep Methods* 2023; **3:** 100615.

131 Yan Z, Chen B, Yang Y, et al. Multi-omics analyses of airway host–microbe interactions in chronic obstructive pulmonary disease identify potential therapeutic interventions. *Nat Microbiol* 2022; **7:** 1361–75.

132 Dar D, Dar N, Cai L, Newman DK. Spatial transcriptomics of planktonic and sessile bacterial populations at single-cell resolution. *Science* 2021; **373:** eabi4882.

© 2026 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).