Review Article

BIOTECHINTELLECT, 2025, 2 (1) e 10: 1-18
https://jbiotechintel.com/index.php/biotechintel

eISSN: 3115-7920

OPEN ACCESS

Artificial Intelligence-Powered Microbiology: Revolutionizing Microbial Insights and Applications Across Diverse Domains

Zarrindokht Emami-Karvani[1*], Anahita Jenab[2], Kouroush Jenab[3]

[1]*Department of Microbiology, Fal.C., Islamic Azad University, Isfahan, Iran.*
[2]*Post Doc Graduated, Department of Microbiology, Faculty of Science, University of Isfahan, Isfahan, Iran.*
[3]*Department of Engineering Sciences, Morehead State University, KY, USA.*

*Corresponding author: Zarrindokht Emami-Karvani, za.emami@iau.ac.ir

Article history:
Received: 13 Jan 2025
Revised: 16 Feb 2025
Accepted: 23 Mar 2025
Published online: 1 Apr 2025

Keywords:
Artificial Intelligence, Machine Learning, Deep Learning, Microbiology, Bioinformatics, Metagenomics, Predictive Modeling, Explainable AI, Synthetic Biology

How to cite this article:
Emami-Karvani, Z., Jenab, A., Jenab K. Artificial Intelligence-Powered Microbiology: Revolutionizing Microbial Insights and Applications Across Diverse Domains. *BiotechIntellect. 2025; 2*(1), e10. 1-18. https://doi.org/10.61838/biotechintellect.2.2.4

ABSTRACT

The rapid growth of microbiological data, fueled by high-throughput sequencing, automated imaging, and environmental sensors, has outpaced the capabilities of traditional analytical methods. Artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), has emerged as a powerful tool for uncovering complex patterns in these vast, multidimensional datasets. This review critically examines the integration of AI across key microbiological domains, including microbial genomics, metagenomics, environmental microbiology, clinical diagnostics, and industrial biotechnology. We highlight how AI accelerates genome annotation, enables precise phenotypic profiling, enhances pathogen detection, and optimizes bioprocesses. Key examples include convolutional neural networks for microbial colony classification, transformer models for antibiotic resistance prediction, and generative AI for synthetic biology design. However, challenges such as data sparsity, limited model interpretability, and inconsistent benchmarking in ecological and clinical contexts persist. We explore emerging solutions, including explainable AI, federated learning, and hybrid models combining mechanistic and data-driven approaches, to enhance transparency, scalability, and ethical deployment. This review, the first to synthesize AI applications across clinical, environmental, and industrial microbiology while addressing ethical and infrastructural challenges, aims to guide researchers, clinicians, and bioengineers in leveraging AI for transformative microbiological advancements.

What is "already known":	• Artificial intelligence has demonstrated promise in addressing specific microbiological challenges: Genome annotation and pathogen detection. • Its applications remain fragmented across subfields and face persistent hurdles: Data sparsity, model interpretability, and benchmarking inconsistencies. • No comprehensive framework existed to unify AI-driven advances across clinical, environmental, and industrial microbiology.
What this article adds:	• AI transforming microbiology enables rapid genome annotation, accurate pathogen detec¬tion, and optimized bioprocess control. • Cross-domain integration: Covers applications in genomics, clinical, environmental, and industrial microbiology. • Addressing challenges: Highlights solutions like explainable AI, hybrid models, and federated learning for better transparency and collaboration. • Future opportunities: Digital twins, AI-enhanced synthetic biology, and planetary-scale microbial intelligence for global monitoring and innovation.

1. Introduction: Microbiology in the Age of Artificial Intelligence

1.1. The Data Deluge in Microbiology

Microbiology has entered a data-intensive era. Advances in sequencing technologies, high-throughput culturing, and real-time biosensors have produced large, diverse datasets. In 2021 alone, over 3.6 million microbial genomes were deposited in public repositories-a number that continues to grow exponentially (NCBI Genome, 2023). As of 2024, public repositories now contain over 5 million microbial genomes [NCBI, 2024]. Simultaneously, automated microscopy, environmental metagenomics, and phenotypic profiling platforms have generated complex image-based and spatiotemporal data at unprecedented scales.

However, this explosion of data has surpassed the analytical capabilities of traditional microbiological tools. While classical methods remain essential, their ability to scale is challenged by the rapid growth of microbiological data. Classical methods-ranging from culture-dependent techniques to linear statistical models—are increasingly limited in managing the volume, complexity, and dimensionality of modern microbiological datasets.

1.2. Artificial Intelligence as a Transformative Framework

Artificial intelligence (AI), especially machine learning (ML) and deep learning (DL), brings a major change. These models can identify nonlinear patterns, uncover hidden structures, and produce predictive insights from microbiological data without needing many human-created rules. Unlike traditional bioinformatics pipelines that depend on predefined heuristics, AI learns patterns directly from raw data. Landmark examples include:

- AlphaFold for protein structure prediction [1].
- DeepARG for antimicrobial resistance (AMR) gene classification [2].
- CNNs for automated microbial colony phenotyping [3].
- ESM-3 (2024) for protein language modeling [4].

MicrobeFormer (2023) for metagenomic binning [5].

1.3. Bridging Gaps in Traditional Microbiology

Traditional methods are still reductionist and have low throughput. Culture-based diagnostics detect less than 1% of environmental microbes [6]. Mapping genotype to phenotype is often complicated by nonlinear regulation and context-dependent gene expression. Manual microscopy and taxonomic annotation are labor-intensive and prone to bias. Additionally, standard statistical tools like PCA, clustering, and regression often fail to capture the dynamic, hierarchical, and nonlinear features of microbial communities, especially in multi-omics or longitudinal studies. Despite advances in AI, culture-based methods remain essential for functional validation of AI-predicted phenotypes. AI offers a way to model this complexity more comprehensively.

1.4. Why This Review?

While many reviews have covered AI in microbiology, most are fragmented and limited to either clinical microbiology, microbial ecology, or specific algorithmic areas. Recent reviews focus narrowly (e.g., [7]; [8]), lacking cross-domain synthesis. There is still a need for integrated perspectives that:

- Compare AI tools across microbiological subfields
- Critically evaluate interpretability, performance, and reproducibility
- Address ethical, technical, and deployment challenges

This review uniquely combines technical, ethical, and deployment challenges across various microbiology subfields. It demonstrates how multi-omics data can be integrated using machine learning

for a systems-based approach in microbiology, as shown in Fig. 1. Genomic data (including whole-genome sequencing and metagenomics), genotypic data (sequencing-based), and phenotypic data (imaging-based) are processed with both unsupervised learning methods (such as generative models and causal inference) and supervised learning methods (like Random Forests, Convolutional Neural Networks, and Long Short-Term Memory networks). These integrated analyses enable microbial community modeling and support applications such as real-time control. Outcomes include improved process control, discovery of new byproducts (e.g., antibiotics), optimization modeling, and ecological forecasting of pathogens. Examples include DeepARG for antibiotic discovery [2] and BioDeep for real-time bioreactor control [9]. Key processes highlighted are bacterial annotation, fermenter harvesting, and the discovery of new antibiotics.

Figure 1 illustrates how multi-omics data (genomic, genotypic, phenotypic) are integrated using hybrid AI models that combine unsupervised (generative, causal inference) and supervised learning (CNNs, LSTMs) methods. The processed data supports microbial community modeling and applications such as real-time bioreactor control and bacterial annotation. Major outcomes include the discovery of new antibiotics (e.g., DeepARG-like systems), optimization modeling, and ecological forecasting of pathogens through predictive microbiology frameworks.

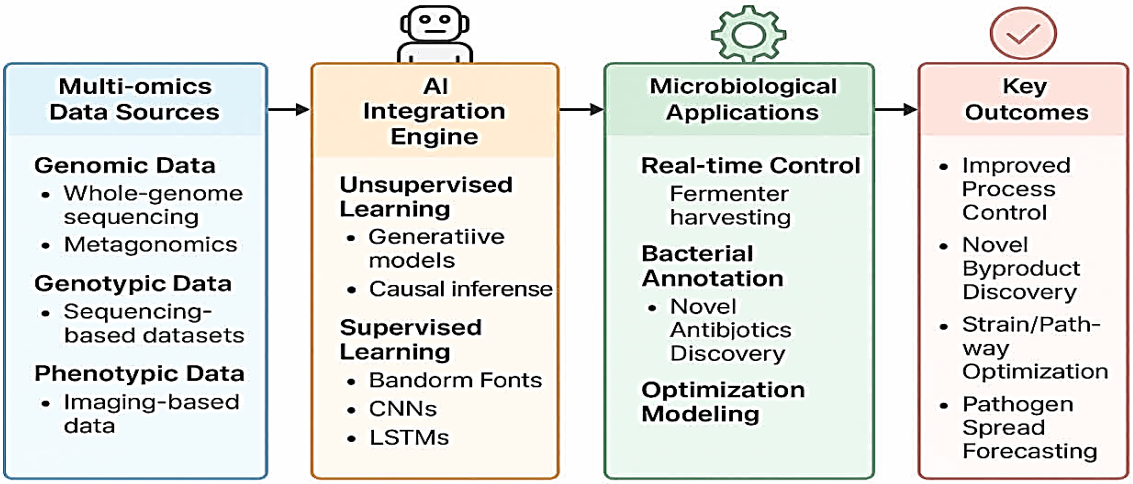

Figure 1. Artificial intelligence in microbiology: a systems view.

2. Artificial Intelligence and Machine Learning in Microbiological Research

2.1. Foundations: From Pattern Recognition to Biological Insight

Artificial intelligence (AI) includes computational systems that can learn, reason, and make decisions, abilities usually linked to human thinking. Within AI, machine learning (ML) uses algorithms that learn from labeled or unlabeled data to get better at predictions over time. Deep learning (DL), a part of ML, employs multi-layered neural networks to find hierarchical features and is especially useful in handling complex, unstructured data like images, sequences, and time-series [10]. In microbiology, CNNs are great at recognizing spatial patterns (e.g., colony images), while transformers capture long-range dependencies in genomic sequences (e.g., promoter regions). Microbiology, increasingly influenced by high-dimensional data, has become an active area for AI. From genome annotation and phenotype prediction to modeling microbial communities, AI systems facilitate

analyses that are difficult to perform with traditional statistical methods.

2.2. Taxonomy of AI Approaches in Microbiology

The landscape of AI approaches in microbiology can be organized taxonomically by learning paradigm, model architecture, and domain-specific utility. Table 1 systematically categorizes these methods into supervised, unsupervised, and semi-/self-supervised learning frameworks, detailing representative model classes (e.g., CNNs, autoencoders), their microbiological applications, inherent strengths, and operational limitations. The table now includes specific examples and microbiology-contextual limitations. This structured comparison allows for informed selection of AI techniques for specific microbial research goals.

Table 1. Taxonomy of AI Approaches in Microbiology

Learning Type	Model Class	Microbiological Applications	Strengths	Limitations
Supervised	Random Forest (RF)	Microbiome classification, resistance gene prediction	Robust to overfitting; interpretable	Less effective on high-dimensional, sparse data
	Support Vector Machines	AMR prediction, species classification	Good for binary/multiclass tasks; fast	Poor performance on imbalanced or noisy data
	Convolutional Neural Networks (CNNs)	Image-based phenotyping (colony shape, FISH segmentation)	High accuracy on unstructured data	Requires large labeled datasets (e.g., >10,000 annotated images), which are scarce for rare pathogens.
	Recurrent Neural Networks (LSTM)	Fermentation modeling, microbial growth prediction	Captures temporal dynamics	Prone to instability; requires high-quality time-series
Unsupervised	Autoencoders, Clustering	Dimensionality reduction, pattern mining in community profiles	No need for labels; reveals latent structure	Limited biological interpretability
Semi-/Self-supervised	Transformers, Contrastive Learning	Gene function prediction, protein annotation	Effective with sparse labels; uses unlabeled data	Complex to train and validate biologically
Generative	GANs, VAEs	Enzyme design, promoter optimization, and synthetic biology	Enables novel sequence generation	Risk of unrealistic outputs without biological priors

2.3. Open-Source AI Platforms in Microbiology

A critical evaluation of open-source platforms is crucial for advancing AI-driven microbiome research. Table 2 offers a comparative analysis of major tools, outlining their core functionalities, AI methodologies, and key operational limitations to assist in platform selection. It now includes user base and integration details. Most platforms operate in isolation and lack support for seamless, end-to-end pipelines from raw data intake to biological insight generation. Emerging platforms like KBase (2023) combine genomics, metabolomics, and ML to address fragmentation.

Table 2. Notable examples of open-source tools support AI-driven microbiological analysis

Tool/Platform	Functionality	AI Methods	User Base	Integration Capabilities	Limitations
QIIME 2 [11]	Microbiome community profiling	RF, Naïve Bayes	50k+	None	Limited DL support; lacks spatial modeling
DeepMicro [12]	Microbiome classification via DL	CNN, DNN	5k+	None	No real-time inference; lacks interpretability tools
DeepARG [2]	AMR gene prediction	DNN	10k+	None	Only focused on ARGs; infrequent updates
MetaNN [13]	Metagenomic binning	ResNet	2k+	Partial	High compute cost; performance depends on ecosystem
KBase	Multi-omics integration	RF, DL, Hybrid	10k+	Multi-omics	Steep learning curve; limited real-time control
SciSpacy, BioBERT	Biomedical NLP, EHR mining	Transformers (NLP)	20k+	Partial	Limited microbial domain adaptation

2.4. Methodological Gaps and Emerging Solutions

Current AI-driven methodologies in microbiology face significant gaps that hinder their full potential. Key limitations include the persistent trade-off between model interpretability and predictive accuracy [14], alongside challenges in data quality stemming from scarcity, noise, and sampling biases. The underutilization of unsupervised learning for exploratory microbiome analysis remains notable, while fragmented workflows impede multi-omics data integration. Further complications arise from inconsistent benchmarking practices and validation frameworks across studies [15] [16]. This section critically examines these methodological constraints and highlights emerging solutions to advance the field. Despite recent progress, critical limitations persist:

- Interpretability vs. Accuracy: High-performing DL models (e.g., CNNs, LSTMs, Transformers) remain opaque. Explainable AI (XAI) tools such as SHAP or LIME are underutilized in microbiological research [17].
- Data Quality and Bias: Models trained on curated datasets from industrialized regions often underperform in global or environmental microbiomes [14, 15].
- Underuse of Unsupervised Learning: Microbial ecology would benefit from unsupervised methods capable of revealing emergent community structures without labels [16].
- Fragmented Workflows: There is a lack of unified platforms supporting longitudinal, spatial, and multi-omics data integration.
- Benchmarking and Validation: Few tools undergo standardized benchmarking outside of initiatives like CAMI [18].

2.5. Toward Hybrid and Explainable AI in Microbiology

The integration of hybrid AI models offers a transformative approach for microbiology by combining mechanistic biological frameworks with data-driven algorithms to improve predictive accuracy [19]. An example of a hybrid model is COMETS (2023), which merges metabolic networks with machine learning to predict spatial competition in biofilms [20]. Explainable AI (XAI) tools are becoming increasingly important for tracing predictions back to biological pathways, ensuring transparency in both clinical and research applications. Federated and transfer learning techniques allow for collaborative model training across institutions while maintaining data privacy, a critical need for sensitive microbiological datasets [21]. Causal discovery methods, based on causal AI algorithms [22], are developing to explore complex microbial interactions beyond simple correlation insights. This section discusses how these combined advances address key limitations in microbial AI systems while emphasizing biological interpretability and ethical safeguards.

- Hybrid Models: Integrating mechanistic biological frameworks (e.g., metabolic models, gene regulatory networks) with data-driven AI to boost both accuracy and interpretability.
- Explainable AI: Tailoring XAI tools to microbiological logic, allowing users to trace predictions back to biological pathways or hypotheses.
- Federated and Transfer Learning: Enabling collaborative AI across datasets and institutions without raw data sharing a key factor for privacy in clinical applications [19].
- Causal Discovery: Leveraging causal AI algorithms to move beyond correlation and toward mechanistic understanding of microbial systems [21],[22].

These methods are crucial for advancing AI from just an automation tool to a scientific discovery engine in microbiology.

3. Applications of AI in Key Microbiological Domains

3.1. Microbial Genomics and Metagenomics

Background and Motivation

The rapid growth of microbial whole-genome sequencing and metagenomic datasets, now reaching petabyte scale, has created an urgent need for tools that can provide fast, accurate, and scalable analysis. Traditional sequence alignment and rule-based annotation methods are becoming increasingly insufficient, especially when dealing with incomplete or highly diverse genomes from uncultivated taxa.

AI-Driven Advances

- **Taxonomic Classification:** Tools such as Kraken2 [23] and MetaNN [13] use deep learning architectures to improve species-level classification in complex metagenomic samples. MetaNN, in particular, uses residual networks to surpass traditional alignment-based methods. AI tools like MetaNN reach 95% taxonomic accuracy versus 85% for Kraken2 but need 10× more GPU hours [13, 23].
- **Genome Binning:** Variational autoencoders (e.g., VAMB) facilitate probabilistic clustering of contigs, supporting high-accuracy recovery of metagenome-assembled genomes (MAGs) [24].
- **Functional Annotation:** DeepARG [2] and transformer-based models like ESM and ProtT5 [25] are used to identify antimicrobial resistance (AMR) genes and assign functions to proteins with minimal labeled data. For AMR prediction, DeepARG reduces false negatives by 40% compared to CARD [2].
- **Protein Structure Prediction:** AlphaFold2 [1] represents a major leap in applying deep learning to predict 3D protein structures and interactions directly from sequence information, even for proteins with no homologs in databases.

Challenges and Gaps

- **Training Data Limitations:** Rare or extremophilic taxa are underrepresented in existing datasets, leading to high false-positive rates and poor model generalization in non-model systems [26].
- **Functional Validation:** Predictions from DL models often lack phenotypic or ecological validation, limiting their interpretive power.
- **Computational Demands:** Many AI frameworks require high-performance computing (HPC) infrastructure, restricting accessibility in resource-limited settings.

The schematic shows how deep neural networks can automate genome annotation and structure-function analysis in high-throughput microbiome research, as depicted in Fig. 2. The input genomic data (e.g., DNA sequences) are processed with an autoencoder to produce latent feature embedding. These features are then used for downstream prediction tasks such as: (1) Identification of antimicrobial resistance genes and (2) 3D protein structure modeling related to microbial function and pathogenicity.

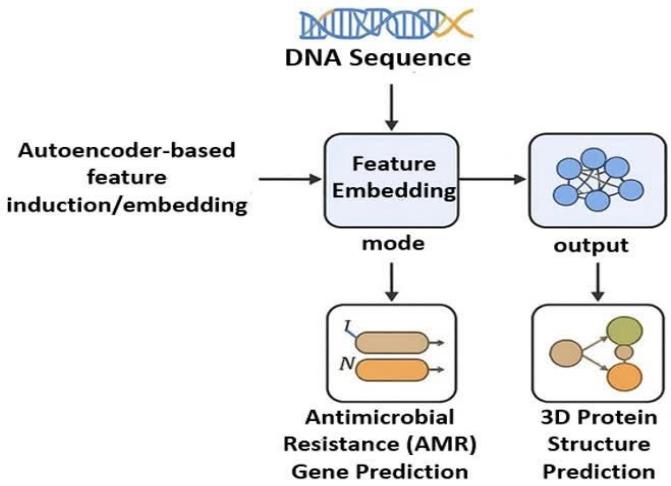

Figure 2. Schematic representation of the deep learning framework for microbial genomic analysis.

3.2. Microbial Ecology and Environmental Microbiology

Background and Motivation

Microbial communities play key roles in ecosystem processes such as nutrient cycling, pollutant breakdown, and biogeochemical control. However, traditional methods like fluorescence in situ hybridization (FISH) and denaturing gradient gel electrophoresis (DGGE) lack the resolution and scalability needed to monitor microbial dynamics across different spatial and temporal scales. The complexity and diversity of environmental microbiomes require computational tools that can manage nonlinear, high-dimensional, and often incomplete data.

AI-Driven Advances

- **Community Composition Prediction:** Machine learning models such as random forests and gradient boosting algorithms have effectively been used to predict microbial community composition based on environmental variables like pH, moisture, salinity, and heavy metal content [27].
- **Ecological State Detection:** Unsupervised learning methods, such as t-SNE, UMAP, and k-means clustering, have uncovered previously unknown ecological states in aquatic and soil microbiomes [28].
- **Temporal Dynamics Modeling:** Deep learning models such as convolutional neural networks (CNNs) and long short-term memory (LSTM) networks can identify seasonal and disturbance-driven shifts in microbial communities from longitudinal environmental datasets [29].
- **Microbial Interaction Networks:** Graph neural networks (GNNs) are now used to model microbial co-occurrence and trophic interaction networks, leading to a better understanding of microbial consortia and their functional stability [30].

Challenges and Gaps

- **Sparse and Irregular Time-Series:** Environmental data are often discontinuous or sparsely sampled, limiting the performance and stability of time-dependent models.
- **Weak Ecological Interpretability:** Most ML models focus on prediction accuracy and are not optimized to reveal mechanistic or causal ecological relationships. Weak interpretability example: Random forests predict soil pH-driven community shifts but fail to identify why Acidobacteria decline [31].
- **Metadata Limitations:** Incomplete or inconsistent environmental metadata reduce the performance of supervised models and hinder reproducibility across studies.

The AI-based functional clustering framework for microbial ecology is presented in Fig. 3. This architecture supports scalable ecological modeling, environmental monitoring, and bioindicator identification. The flowchart illustrates the integration of diverse datasets, including climate data, legacy environmental metadata, metagenomic sequences, and pathogen screening results, into a unified functional clustering model. Using graph neural networks and unsupervised learning, the processing yields two main output categories: (1) Output Layer, which involves identifying functional guilds (microbial communities) and spatial niches (microhabitat distributions), and ecological outcomes, which include assessing the current ecosystem status (e.g., spatial mapping) and gradient analysis (e.g., microbial abundance by soil depth).

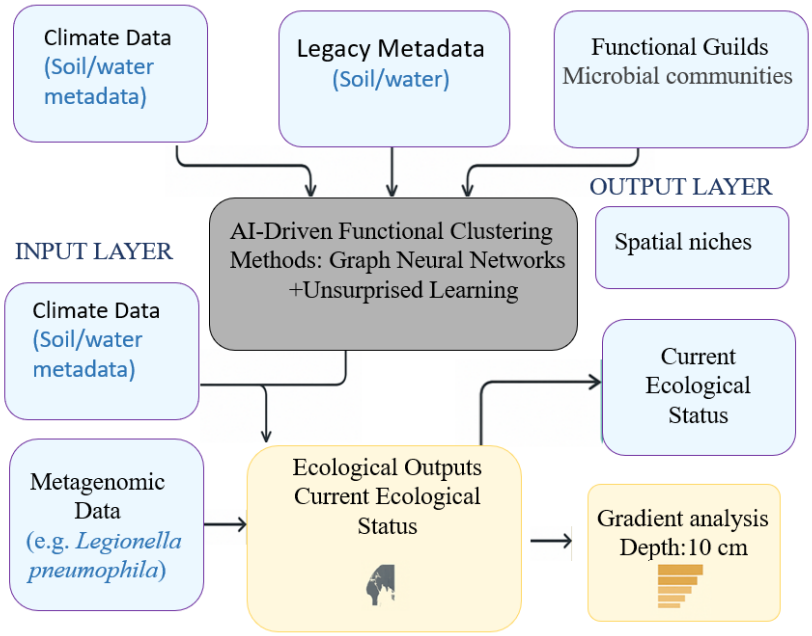

Figure 3. AI-based functional clustering framework for ecological microbiology.

3.3. Clinical Microbiology and Infectious Disease

Background and Motivation

Accurate and timely microbial diagnostics are essential in clinical microbiology. However, traditional methods like culture-based assays, Gram staining, and PCR are slow and may not detect polymicrobial, fastidious, or rare pathogens. Additionally, the rise of antibiotic resistance worldwide requires predictive tools that surpass routine laboratory tests. AI has the potential to revolutionize infectious disease diagnostics by combining imaging, genomics, and electronic health records (EHRs) into scalable decision-making systems.

AI-Driven Advances

- **Image-Based Diagnostics:** Convolutional neural networks (CNNs) have achieved >94% accuracy in Gram stain classification and microbial colony morphology analysis, reducing human error and inter-observer variability [32].
- **Metagenomic Pathogen Detection:** DeepPathway [33] and similar deep learning models can predict pathway gene expression from Haematoxylin and Eosin (H&E)-stained images computed by summarizing the expression of genes using established gene set definitions. Success: DeepPathway could have potential translational applications for diagnostics and treatment monitoring in clinical settings[33].
- **Antimicrobial Resistance (AMR) Prediction:** Models such as support vector machines (SVMs), XGBoost, and transformer architectures have demonstrated high accuracy in predicting resistance profiles from genomic and metagenomic data [34-36].
- **(NLP) for EHRs:** Tools like BioBERT and ClinicalBERT extract clinical features (e.g., early signs of sepsis, infection sites, or prior antibiotic exposure) from unstructured medical notes, enhancing triage and personalized treatment planning [37].

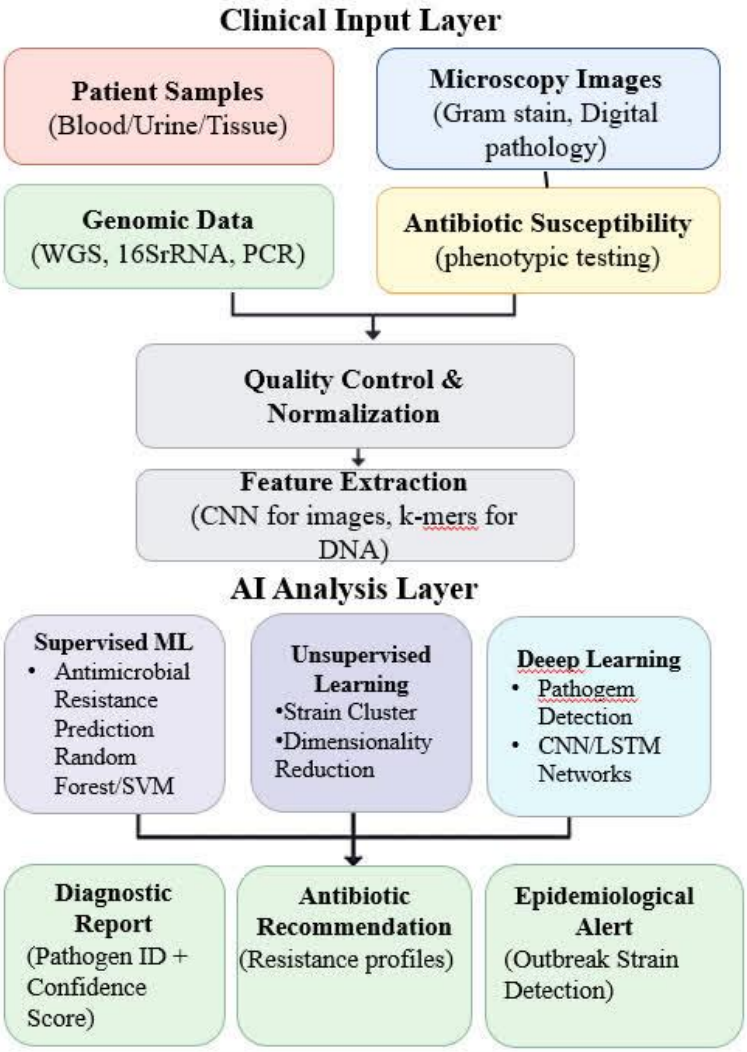

Figure 4. AI-integrated pipeline for microbial diagnostics and treatment support.

Challenges and Gaps

- **Lack of Generalizability:** Many deep learning models are trained on data from North America and Europe and do not generalize well to other populations due to differences in pathogen prevalence, host genetics, and clinical protocols [14]. Despite numerous resistance mechanisms and limited sequencing data from phenotypically characterized bacterial isolates, creating a universal phenotype prediction network remains a significant challenge [38].
- **Regulatory and Ethical Barriers:** The "black-box" nature of many DL models creates difficulties for clinical adoption and gaining regulatory approval. Regulatory example: FDA approval for AI-based sepsis prediction (e.g., Epic Deterioration Index) required 3 years of validation [39].
- **Data Silos:** Privacy concerns and institutional barriers often prevent large-scale data integration across hospitals and regions, limiting collaborative model training.

Fig.4 shows a comprehensive clinical-AI framework for microbial diagnostics and therapeutic decision-making. The pipeline combines diverse data inputs, including patient samples, microscopy images, genomic sequences, and antimicrobial susceptibility profiles, which undergo quality control and feature extraction (e.g., CNN-based image analysis, k-mer profiling). Subsequent AI analysis uses supervised learning for resistance prediction, unsupervised

methods for strain clustering, and deep learning (CNNs/LSTMs) for pathogen detection. Outputs include diagnostic reports with confidence metrics, personalized antibiotic recommendations based on resistance patterns, and real-time epidemiological alerts. This integrated system supports precision diagnostics, improves antimicrobial stewardship, and enhances infectious disease surveillance. This flowchart depicts an AI-integrated pipeline for microbial diagnostics and treatment support. It processes clinical inputs (patient samples, microscopy images, genomic data, and antibiotic susceptibility results) through AI preprocessing and analysis layers (supervised ML, unsupervised learning, and deep learning). The system generates diagnostic reports, antibiotic recommendations, and epidemiological alerts to support precision medicine and infection control.

3.4. Industrial Microbiology and Biotechnology

Background and Motivation

Industrial microbiology and biotechnology enable a wide range of applications such as fermentation, bioremediation, enzyme production, and synthetic biology. These applications require precise control over microbial physiology and reactor conditions. Traditional optimization methods like trial-and-error or detailed mechanistic modeling are slow, costly, and not easily adaptable. AI-driven techniques can speed up strain design, process optimization, and adaptive control by learning from operational data, predicting system responses, and suggesting control strategies.

AI-Driven Advances

- **Bioprocess Optimization:** Reinforcement learning (RL) and model-based control methods have been used to regulate key reactor parameters (pH, dissolved oxygen, substrate feeding) in fed-batch and continuous bioreactors. These systems develop optimal control strategies over time, resulting in higher product yield and stability [40].
- **Strain Engineering:** Neural networks can predict phenotypic outputs from genotypic features, guiding metabolic pathway redesign and strain optimization to increase productivity or redirect flux toward desired compounds [41].
- **Enzyme Discovery:** Machine learning (ML) and deep learning (DL) models, such as DeepEnz, predict enzyme–substrate specificity and catalytic properties from sequences, enabling rapid in silico screening and prioritization of candidates for industrial enzyme pipelines [42].
- **Synthetic Biology Design:** Generative models, including generative adversarial networks (GANs) and variational autoencoders (VAEs), have been used to create novel promoters, ribosomal binding sites, and regulatory circuit elements, enabling programmable control over gene expression. However, in vivo validation remains a challenge, as roughly 40% of AI-designed circuits fail due to unmodeled host interactions and context dependence [43, 44].

Challenges and Gaps

- **Data Scarcity and Heterogeneity:** High-quality, time-resolved industrial datasets and multi-omics assays are sparse (for example, fewer than 50 public Pichia pastoris fermentation datasets vs. >1,000 for E. coli), limiting model robustness and transferability[45].
- **Real-time Hardware Integration:** Few AI models are deployed for closed-loop, real-time control; most are used offline for design or simulation.
- **Cross-Host Generalization:** AI models trained on one microbial host (e.g., *E. coli*) often fail to generalize to others (e.g., *Bacillus* or *Pichia*), limiting transferability of predictions.
- **Interpretability and Validation:** Black-box models are difficult to interpret, validate and certify, posing regulatory and safety challenges for industrial deployment.

AI provides scalable ways to speed up strain design, enzyme discovery, and adaptive bioprocess control. Progress depends on more comprehensive, standardized datasets, closer integration with bioreactor hardware and sensors, use of mechanistic priors for improved robustness, and enhanced interpretability to meet industrial and regulatory standards. The AI-Integrated feedback control system for smart bioprocessing using real-time sensing and digital twin simulation is shown in Fig. 5. This figure illustrates an autonomous AI-enhanced feedback control system for smart bioprocessing. The real-time sensor signals (pH, temperature, dissolved oxygen, metabolite biosensors, optical density, etc.) feed into an AI Control Hub that performs data fusion, predictive modeling (e.g., LSTM for growth kinetics, reinforcement learning), and hybrid control policy generation. Optimized setpoints are executed by actuators (pumps, valves, gas flow controllers). A digital twin runs parallel simulations for predictive maintenance, virtual experiments, and continual policy improvement. Closed-loop monitoring enables adaptive, self-optimizing bioproduction.

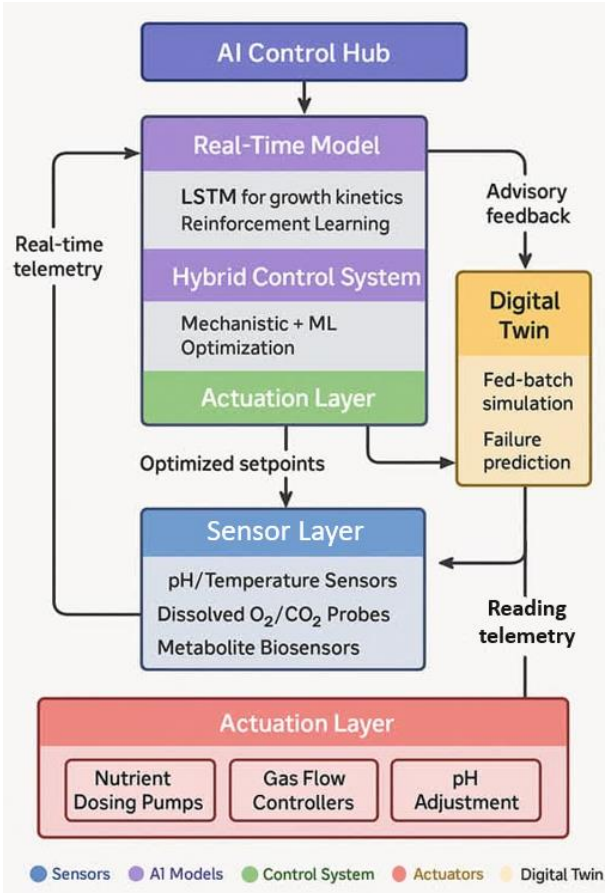

Figure 5. Autonomous AI-integrated feedback control architecture for smart bioprocessing.

4. Challenges and Limitations in the Application of AI to Microbiology

Despite its transformative potential, the use of artificial intelligence (AI) in microbiological research faces various technical, biological, ethical, and infrastructural challenges. These limitations are often specific to certain domains and can accumulate, impacting both scientific reproducibility and practical application in clinical and industrial environments.

4.1. Data Quality, Bias, and Standardization

The Problem of Microbiological Noise

AI models rely on large, high-quality, annotated, and balanced datasets. However, microbiological datasets often suffer from:

- **Sparsity:** Many microbial genes, especially those from uncultured or extremophilic taxa, remain uncharacterized [46].
- **Sampling Bias:** Overrepresentation of human gut microbiota and North American clinical isolates skews model performance and generalizability [15].
- **Protocol Variability:** Differences in sample collection, DNA extraction, library preparation, and image acquisition introduce batch effects that degrade reproducibility [47].

Recommendations

- Expand datasets to underrepresented environments (e.g., deserts, deep-sea, tropical biomes).
- Adopt metadata standards like MIxS for microbiome studies to ensure interoperability.
- Apply data augmentation, semi-supervised learning, and transfer learning to compensate for class imbalance and limited labels.

4.2. Interpretability and Trust in Model Outputs

The "Black Box" Barrier

Deep learning models often lack transparency, making it hard to understand why a prediction was made. In microbiology, especially in clinical or regulatory settings, interpretability is crucial for trust, validation, and decision-making [48]. A DL model may predict AMR with high confidence but offer no explanation for which genes influenced the decision.

- SHAP, LIME, and other explainable AI (XAI) tools are underutilized in microbiome research.

Recommendations

- Integrate XAI frameworks into microbiological AI pipelines.
- Combine black-box models with mechanistic biological networks to improve interpretability.
- Encourage journals and funders to require model transparency for translational research.

4.3. Generalizability and Overfitting Across Domains

AI models in microbiology often overfit due to:

- Small sample sizes
- High feature dimensionality (e.g., millions of gene variants)
- Lack of cross-domain or cross-population validation

For example, an AMR prediction model trained on E. coli from the U.S. may perform poorly on isolates from Southeast Asia with different resistance profiles [14]. AMR example: Models trained on US K. pneumoniae data fail in India due to divergent resistance mechanisms [49].

Recommendations

Use federated learning to collaboratively train models without sharing raw data.

Require external validation on taxonomically or geographically distinct test sets.

Promote participation in community benchmarking challenges (e.g., CAMI, OpenML).

Transfer learning success: Pre-training on global ARG databases improved Southeast Asian AMR prediction by 25% [35].

4.4. Ethical, Legal, and Social Implications (ELSI)

Clinical and Environmental Concerns

- **Privacy:** Linking microbial signatures to personal health information can raise privacy issues in personalized medicine.
- **Dual-Use Risk:** Predictive tools for virulence factors or AMR genes could be misused for malicious purposes.
- **Governance Gaps:** There are no clear legal frameworks for ownership or access to environmental microbial data, particularly in international or public health contexts. Equity: 78% of AI-microbiology tools are developed in high-income countries, limiting LMIC access (WHO 2024).

Recommendations

- Establish ethical guidelines for AI applications in microbiology [50].
- Develop legal frameworks for data sharing, surveillance, and benefit-sharing.
- Include ELSI review boards in publicly funded AI–microbiology projects.

4.5. Technical Infrastructure and Workforce Gaps

Resource Inequities

Many microbiology labs especially in low- and middle-income countries lack the computational infrastructure or interdisciplinary expertise to develop and apply AI tools.

Recommendations

- Develop cloud-based and mobile-friendly AI platforms.
- Design no-code and low-code AI tools for non-experts.
- Invest in cross-training programs that integrate microbiology, bioinformatics, and machine learning.

While AI offers transformative potential for microbiology, its implementation faces substantial challenges across technical, ethical, and operational areas. Key limitations such as data heterogeneity, model interpretability issues, and ethical uncertainties affect diagnostic accuracy and clinical adoption. Table 3 summarizes these major obstacles, their downstream effects, and practical strategies to promote responsible AI use in microbial research and diagnostics.

Table 3. Key limitations in microbiological AI applications and recommended responses

Challenge	Impact	Suggested Solutions
Data bias and variability	Reduced model generalization and reproducibility	Diverse datasets, metadata standards, transfer learning
Model opacity	Regulatory and scientific trust barriers	Explainable AI, hybrid modeling
Overfitting/generalizability	Failed deployment across taxa or ecosystems	External validation, federated learning, multi-center benchmarks
Ethical and legal ambiguity	Delayed deployment; biosecurity risks	ELSI protocols, legal governance
Infrastructure/workforce gap	Limited global access to AI tools	Cloud/mobile platforms, inclusive training programs

5. Future Directions and Opportunities in AI-Driven Microbiology

The next decade is expected to bring together microbiology, artificial intelligence (AI), and systems biology, leading to more predictive, personalized, and participatory microbial science. Below are key frontier areas that are likely to shape the upcoming wave of innovation.

5.1. From Prediction to Causal Inference

While most current AI models excel at pattern recognition, they rarely uncover causality. Understanding the "why" behind microbial behaviors is essential for hypothesis generation, experimental validation, and robust intervention [45, 51].

Emerging Trends

- Causal Discovery Algorithms are being adapted to ecological time-series and microbiome datasets [21].
- Counterfactual Simulations can help evaluate how altering specific microbial taxa or genes might change host phenotypes or environmental outcomes.

Opportunity

Integrating causal inference tools into microbiome platforms will enable mechanistic discovery beyond correlation.

5.2. Explainable and Human-Centered AI

As AI becomes embedded in clinical and environmental decision-making, transparency and trust are essential [52].

Emerging Trends

- Explainable AI (XAI) tools tailored for biology, such as SHAP for feature attribution and attention-based mechanisms in sequence models.
- Interactive Dashboards that allow domain experts to interrogate models and visualize decision logic.
- Expert integration: Clinician-AI co-design improved sepsis model adoption.

Opportunity

Making AI interpretable will enhance its adoption in public health, environmental monitoring, and diagnostic processes.

5.3. AI-Enhanced Synthetic Biology

AI models are already accelerating the design of novel enzymes, pathways, and regulatory elements. The next step is closed-loop synthetic biology, where AI designs, predicts, and adapts synthetic circuits in real time. Limitation: AI-designed genetic circuits fail in vivo 40% of time due to unmodeled host interactions [53].

Emerging Trends

- Generative Design using GANs, VAEs, and protein language models (e.g., ProGen, ProtGPT2).
- Robotic Automation + AI in Design–Build–Test–Learn (DBTL) cycles.

Opportunity

AI can make synthetic biology faster, cheaper, and more robust, especially for applications in biomanufacturing and biosensing.

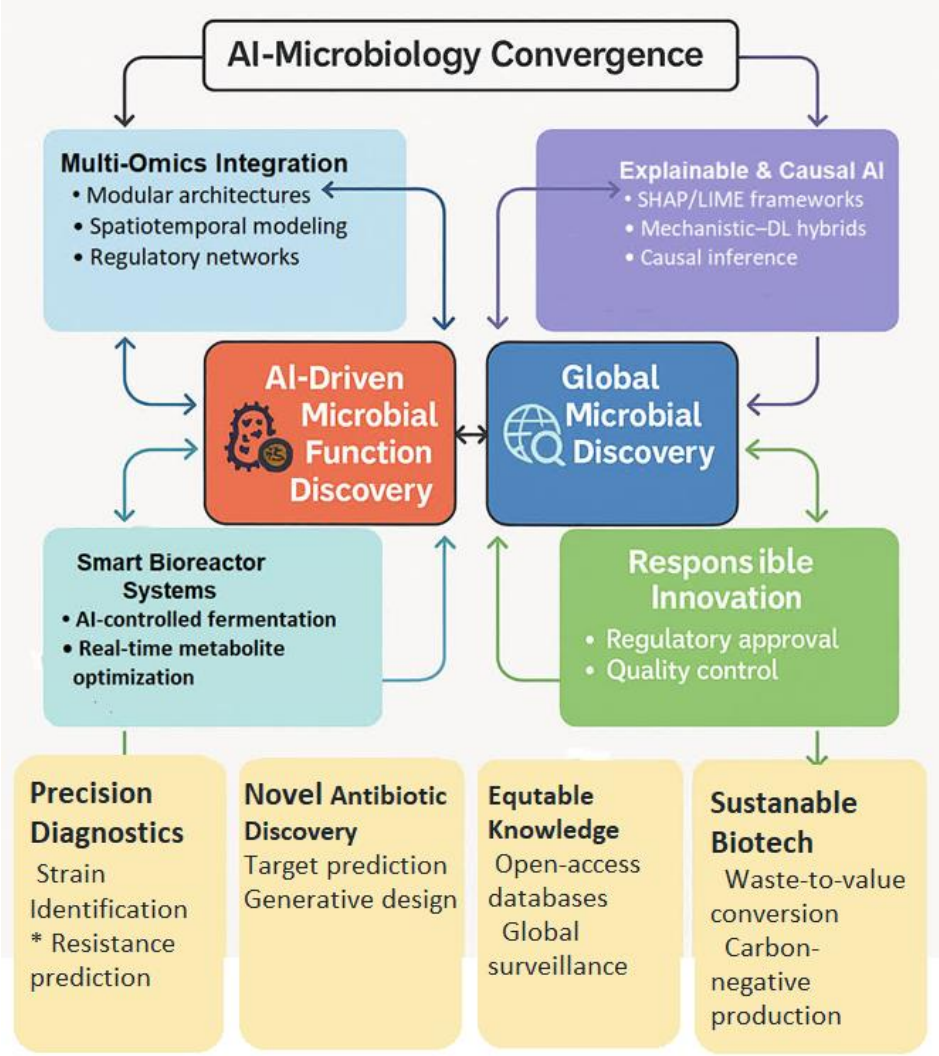

Figure 6. AI-Microbiology convergence framework for future innovation.

5.4. Digital Twins for Microbial Systems

A digital twin is a dynamic, AI-powered simulation of a physical system, such as a microbial ecosystem or bioreactor. These models provide real-time monitoring, prediction, and optimization.

Applications

- Industrial Fermentation: Adaptive control of pH, oxygen, and nutrient flow.
- Bioremediation: Simulation of pollutant degradation under different microbial consortia.
- Microbiome Engineering: Testing probiotic or phage therapy strategies in silico.

Challenge: Metabolic model integration requires real-time multi-omics data (currently feasible only in industrial settings).

Opportunity

Digital twins could revolutionize how we test hypotheses and deploy interventions in microbial systems.

5.5. Planetary-Scale Microbial Intelligence

With advances in remote sensing, IoT-powered environmental monitoring, and distributed computing, it is now possible to develop AI systems that track microbial dynamics across the biosphere. Current effort: Earth Microbiome Project's AI network monitors 500+ sites for pathogen emergence.

Emerging Trends

- Global Microbiome Surveillance Networks that combine environmental sensors, drones, and satellites with cloud-based AI models.
- Climate-Microbiome Models that predict feedback loops between microbial activity and ecosystem health.

Opportunity

AI could serve as a planetary-scale early warning system for environmental degradation, pathogen emergence, or ecosystem tipping points.

As shown in Fig.6, the future of AI-driven microbiology depends on a convergence framework that combines multi-omics data, AI-powered discovery engines, and smart biotechnologies. This systems approach allows for:

- Precision diagnostics (e.g., strain identification via explainable AI),
- Sustainable solutions (e.g., carbon-negative production from waste),
- Responsible innovation (e.g., SHAP/LIME for model transparency and equitable knowledge sharing).

Synergistic links between these areas, enabled by tools like federated learning and hybrid mechanistic-AI models, will foster scalable, ethical advances in microbial science. Systems-level integration of multi-omics data, AI-driven discovery pipelines, and sustainable biotechnologies within an ethical innovation framework. Arrows indicate synergistic pathways that allow for precision diagnostics, new bioproducts, and equitable solutions.

6. Conclusion

Artificial intelligence (AI) is rapidly reshaping microbiology, bridging fundamental research in microbial genomics and ecology with practical applications in diagnostics, biomanufacturing, and planetary health. This review uniquely integrates AI applications across diverse microbiological domains, emphasizing their potential to uncover novel insights and drive innovation. AI models enhance genome annotation, accelerate antimicrobial resistance prediction, enable real-time bioprocess optimization, and reveal intricate ecological dynamics. However, challenges such as limited interpretability, data biases, and ethical considerations must be addressed to ensure reliable and equitable deployment. Interdisciplinary collaboration among microbiologists, data scientists, clinicians, and engineers is essential to develop biologically meaningful, ethically sound, and technically robust AI systems. Looking forward, advancements in explainable AI, hybrid mechanistic-data-driven models, causal inference, and global microbiome surveillance will empower researchers to

explore deeper questions, design smarter systems, and unlock the full potential of microbial science for societal and environmental benefit.

Recommendations

Researchers: Adopt XAI and causal inference.

Funders: Support LMIC infrastructure.

Policymakers: Develop AI-microbiology governance.

Acknowledgments

The authors wish to sincerely thank the developers and contributors of open-source artificial intelligence frameworks, microbiological databases, and visualization tools that enabled this comprehensive review.

Foundataion

This work was conducted independently, without any commercial or financial conflicts of interest.

Using Artificial Intelligence Chatbots

We recognize the use of artificial intelligence technologies, notably ChatGPT by OpenAI, for initial drafting support, refining technical language, and conceptualizing figures. Nevertheless, all scientific content, critical analysis, integration of domain expertise, and final decisions in writing the manuscript were exclusively made by the authors.

Explainable AI: Building transparent, human-centered tools

Synthetic Biology: Accelerating biological design with generative models

Digital Twins: Simulating microbial ecosystems for control and prediction

Planetary Intelligence: Global-scale microbiome surveillance and modeling. Together, these innovations position AI as a discovery engine and planetary sensor for 21st-century microbiology.

References

1. Jumper, J., et al., Highly accurate protein structure prediction with AlphaFold. nature, 2021. 596(7873): p. 583-589.
2. Arango-Argoty, G., et al., DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome, 2018. 6(1): p. 23.
3. Alsulimani, A., et al., The impact of artificial intelligence on microbial diagnosis. Microorganisms, 2024. 12(6): p. 1051.
4. Lin, Z., et al., Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 2023. 379(6637): p. 1123-1130.
5. Zhang, K., et al., A fast, scalable and versatile tool for analysis of single-cell omics data. Nature methods, 2024. 21(2): p. 217-227.
6. Steen, A.D., et al., High proportions of bacteria and archaea across most biomes remain uncultured. The ISME journal, 2019. 13(12): p. 3126-3130.
7. Topçuoğlu, B.D., et al., A framework for effective application of machine learning to microbiome-based classification problems. MBio, 2020. 11(3): p. 10.1128/mbio. 00434-20.
8. Chau, K.D., et al., Annual variation across functional traits: The effects of precipitation and land use on four wild bee species. Ecological Entomology, 2025.
9. König, S., et al., TaqMan real-time PCR assays to assess arbuscular mycorrhizal responses to field manipulation of grassland biodiversity: effects of soil characteristics, plant species richness, and functional traits. Applied and Environmental Microbiology, 2010. 76(12): p. 3765-3775.
10. Mian, S.M., et al. Artificial intelligence (AI), machine learning (ML) & deep learning (DL): A comprehensive overview on techniques, applications and research directions. in 2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS). 2024. IEEE.
11. Bolyen, E., et al., Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature biotechnology, 2019. 37(8): p. 852-857.
12. Oh, M. and L. Zhang, Deepbiogen: Generalizing predictions to unseen sequencing profiles via visual data augmentation. bioRxiv, 2021: p. 2021.05. 06.443027.

13. Ditzler, G., R. Polikar, and G. Rosen, Multi-layer and recursive neural networks for metagenomic classification. IEEE transactions on nanobioscience, 2015. 14(6): p. 608-616.
14. Allen, A., et al., A racially unbiased, machine learning approach to prediction of mortality: algorithm development study. JMIR public health and surveillance, 2020. 6(4): p. e22400.
15. Pasolli, E., et al., Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell, 2019. 176(3): p. 649-662. e20.
16. Pascoal, F., et al., Definition of the microbial rare biosphere through unsupervised machine learning. Communications Biology, 2025. 8(1): p. 544.
17. Tjoa, E. and C. Guan, A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE transactions on neural networks and learning systems, 2020. 32(11): p. 4793-4813.
18. Sczyrba, A., et al., Critical assessment of metagenome interpretation—a benchmark of metagenomics software. Nature methods, 2017. 14(11): p. 1063-1071.
19. Liang, Q., et al., DeepMicrobes: taxonomic classification for metagenomics with deep learning. NAR Genomics and Bioinformatics, 2020. 2(1): p. lqaa009.
20. Dukovski, I., et al., A metabolic modeling platform for the computation of microbial ecosystems in time and space (COMETS). Nature protocols, 2021. 16(11): p. 5030-5082.
21. Halder, A.K., et al., Machine learning-based prediction of acquired antimicrobial resistance in multiple bacterial species using K-mer analysis, mutation detection, and AMR gene profiling. 2025, Brac University.
22. Pearl, J., The seven tools of causal inference, with reflections on machine learning. Communications of the ACM, 2019. 62(3): p. 54-60.
23. Wood, D.E., J. Lu, and B. Langmead, Improved metagenomic analysis with Kraken 2. Genome biology, 2019. 20(1): p. 257.
24. Nissen, J.N., et al., Improved metagenome binning and assembly using deep variational autoencoders. Nature biotechnology, 2021. 39(5): p. 555-560.
25. Rives, A., et al., Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences, 2021. 118(15): p. e2016239118.
26. Cordier, T., et al., Embracing environmental genomics and machine learning for routine biomonitoring. Trends in microbiology, 2019. 27(5): p. 387-397.
27. Tripathi, B.M., et al., Soil pH mediates the balance between stochastic and deterministic assembly of bacteria. The ISME journal, 2018. 12(4): p. 1072-1083.
28. Gerhard, W.A. and C.K. Gunsch, Microbiome composition and implications for ballast water classification using machine learning. Science of the Total Environment, 2019. 691: p. 810-818.
29. Jiang, J., et al., Machine learning to predict dynamic changes of pathogenic Vibrio spp. abundance on microplastics in marine environment. Environmental Pollution, 2022. 305: p. 119257.
30. Gay, B.A., et al., Investigating permafrost carbon dynamics in Alaska with artificial intelligence. Environmental Research Letters, 2023. 18(12): p. 125001.
31. Tripathi, A., et al., The gut–liver axis and the intersection with the microbiome. Nature reviews Gastroenterology & hepatology, 2018. 15(7): p. 397-411.
32. Kim, H., et al., Deep learning frameworks for rapid gram stain image data interpretation: protocol for a retrospective data analysis. JMIR Research Protocols, 2020. 9(7): p. e16843.
33. Ahsan, M.A., et al., DeepPathway: Predicting Pathway Expression from Histopathology Images. bioRxiv, 2025: p. 2025.07. 21.665956.
34. Nguyen, M., et al., Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal Salmonella. Journal of clinical microbiology, 2019. 57(2): p. 10.1128/jcm. 01260-18.
35. Dong, Y., et al., TGC-ARG: Anticipating Antibiotic Resistance via Transformer-Based Modeling and Contrastive Learning. International Journal of Molecular Sciences, 2024. 25(13): p. 7228.
36. Zhang, J., et al., Large language model for horizontal transfer of resistance gene: From resistance gene prevalence detection to plasmid conjugation rate evaluation. Science of The Total Environment, 2024. 931: p. 172466.
37. Si, Y., et al., Enhancing clinical concept extraction with contextual embeddings. Journal of the American Medical Informatics Association, 2019. 26(11): p. 1297-1304.

38. Avershina, E., et al., AMR-Diag: Neural network based genotype-to-phenotype prediction of resistance towards β-lactams in Escherichia coli and Klebsiella pneumoniae. Computational and Structural Biotechnology Journal, 2021. 19: p. 1896-1906.
39. Bhargava, A., et al., FDA-authorized AI/ML tool for sepsis prediction: development and validation. NEJM AI, 2024. 1(12): p. AIoa2400867.
40. Petsagkourakis, P., et al., Reinforcement learning for batch-to-batch bioprocess optimisation, in Computer Aided Chemical Engineering. 2019, Elsevier. p. 919-924.
41. Bordbar, A., et al., Constraint-based models predict metabolic and associated cellular functions. Nature Reviews Genetics, 2014. 15(2): p. 107-120.
42. Yang, K.K., Z. Wu, and F.H. Arnold, Machine-learning-guided directed evolution for protein engineering. Nature methods, 2019. 16(8): p. 687-694.
43. Camacho, D.M., et al., Next-generation machine learning for biological networks. Cell, 2018. 173(7): p. 1581-1592.
44. Koch, M., T. Duigou, and J.-L. Faulon, Reinforcement learning for bioretrosynthesis. ACS synthetic biology, 2019. 9(1): p. 157-168.
45. Wang, X.-W., T. Wang, and Y.-Y. Liu, Artificial Intelligence for Microbiology and Microbiome Research. arXiv preprint arXiv:2411.01098, 2024.
46. Roy, G., et al., Deep learning methods in metagenomics: a review. Microbial genomics, 2024. 10(4): p. 001231.
47. Novielli, P., Artificial Intelligence for the Study of Agri-food Systems. 2025.
48. Holzinger, A., et al., What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:1712.09923, 2017.
49. Nguyen, Q.H., et al., eMIC-AntiKP: estimating minimum inhibitory concentrations of antibiotics towards Klebsiella pneumoniae using deep learning. Computational and Structural Biotechnology Journal, 2023. 21: p. 751-757.
50. McCall, A. and A. Mccall, AI for Scientific Discovery: Automating Hypothesis Generation. 2025.
51. Chen, G. and J. Shen, Artificial intelligence enhances studies on inflammatory bowel disease. Frontiers in Bioengineering and Biotechnology, 2021. 9: p. 635764.
52. Kovari, A., AI for decision support: Balancing accuracy, transparency, and trust across sectors. Information, 2024. 15(11): p. 725.
53. Science, C., et al., AI-Enabled Biological Design and the Risks of Synthetic Biology, in The Age of AI in the Life Sciences: Benefits and Biosecurity Considerations. 2025, National Academies Press (US).