JOURNALS | Journal of Biomedical Systems and Engineering

Volume 1 Issue 1 (September 2026)
J. Biomed. Syst. Eng., 2026, 1(1), 26804
DOI : https://doi.org/10.64189/bse.26804

Review Article

OPEN ACCESS

Integrating Metagenomics and Deep Learning for Accurate Microbial Species Differentiation

David Odiba,* Faith Chidinma Terna, Osuyi Gerard Uyi, Samuel Anzaku and Olukayode Olugbenga Orole

Department of Microbiology, Federal University of Lafia, P.M.B. 146, Makurdi Road, Gandu, Lafia, Nasarawa State, 950101, Nigeria

*Corresponding Author
David Odiba
davidodiba@gmail.com

Received: 24 August 2026
Revised: 24 September 2026
Accepted: 29 September 2026
Published Online: 30 September 2026

Citation

D. Odiba, F. C. Terna, O. G. Uyi, S. Anzaku, O. O. Orole, Integrating metagenomics and Deep Learning for accurate microbial species differentiation, *Journal of Biomedical Systems and Engineering*, 2026, 1(1), 26804, https://doi.org/10.64189/bse.26804.

Open Access

This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, which permits the non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as appropriate credit to the original author(s) and the source is given by providing a link to the Creative Commons License and changes need to be indicated if there are any. The images or other third-party material in this article are included in the article's Creative Commons License, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons License and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this License, visit: https://creativecommons.org/licenses/by-nc/4.0/

© The Author(s) 2026

Abstract

Accurate identification of microbial species is important for clinical diagnosis, environmental monitoring, and evaluations in agriculture and biotechnology. Conventional culture- and alignment-based metagenomic methods have several drawbacks, including inadequate reference databases, the cost of digital compilations, and poor resolution of closely related taxa and strains. This review explores how Deep Learning is integrated into metagenomics to overcome these limitations. The four architecture families, namely convolutional neural networks, recurrent/LSTM networks, transformer-based genomic language models, and graph neural networks, can be applied to resolve taxonomic classification, genome-resolved binning, gene prediction and functional annotation, and microbiome-phenotype association. The review also discusses data preprocessing strategies as critical determinants of model performance. Despite notable gains over traditional methods, persistent challenges include reference dependency, limited interpretability, data leakage, domain shift across platforms and environments, and the absence of standardized benchmarks. The review shows that future progress depends on multimodal, reference-free foundation models validated on taxonomically independent datasets to achieve generalizable, interpretable, and biologically robust species differentiation.

Keywords: Identification; Microbial species; Metagenomics; Deep learning; Transformers.

1. Introduction

Accurate microbial species differentiation is fundamental to clinical diagnostics, environmental monitoring, food safety, agriculture, and biotechnology. Laboratory culturing of microorganisms, although basic, captures only a small fraction of the microbial diversity in most natural and host-associated communities, known as the "great plate count anomaly." Metagenomics enables culture-independent characterization of the collective genetic material recovered from microbial communities, providing access to taxonomic, functional, and ecological information that conventional culture-based approaches cannot readily provide.[1] Metagenomics directly sequences genetic material recovered from the environment or host-associated microbial communities. It has transformed the understanding of microbial ecology, human health, and biotechnology. The falling cost of high-throughput sequencing has produced an explosion of data, and DL complements conventional methods by reducing limitations associated with reference dependency, feature engineering, and complex nonlinear relationships.[2] Because metagenomic data are distinct, conventional techniques face obstacles. Deep learning (DL), a popular area of artificial intelligence, can address these challenges. Deep learning is increasingly applied to metagenomic data because its representation-learning capacity can capture complex sequence, abundance and relational patterns relevant to taxonomic classification, genome reconstruction, functional prediction and host-phenotype association. This review examines how deep learning approaches are being integrated into metagenomic workflows to improve microbial taxonomic resolution, species and strain differentiation, genome reconstruction, and interpretation of complex microbial communities, while critically evaluating their limitations and prospects for broader application.

2. Metagenomic foundations for microbial species differentiation

The branch of metagenomics is built on some fundamental and important principles, which include culture independence, in which DNA (or RNA, in meta transcriptomics) is extracted directly from a sample, bypassing the bias introduced by selective cultivation media and growth conditions; community-level sampling, where metagenomics captures the collective genomic content of all organisms present, including bacteria, archaea, viruses, fungi, and other microeukaryotes. Others are sequence-based inference: Taxonomic and functional identity is inferred computationally by comparing recovered sequences (reads, contigs, or assembled genomes) against reference databases or by de novo classification using sequence composition; depth versus breadth trade-off: Because community DNA is a mixture from many organisms at vastly different abundances, sequencing depth determines whether rare community members are detected, while breadth of coverage determines how completely each genome is represented; and bioinformatic reconstruction: Raw sequence data must pass through quality control, assembly or mapping, binning, and annotation steps before biologically meaningful taxonomic or functional conclusions can be drawn. Microbial species differentiation in metagenomes is achieved by sequencing-based recovery and classification of DNA from environmental, clinical, or host-associated samples, where amplicon and shotgun metagenomics offer complementary taxonomic resolution. Amplicon sequencing, typically targeting conserved marker genes like the 16S rRNA gene, offers an economical way to recover sequences for microbiome taxonomic profiling, but its utility for species- or strain-level resolution is generally limited because many closely related species or strains share similar marker sequences. In contrast, shotgun metagenomics recovers the entire DNA content of microbial communities and thus provides more information for species-level identification, strain-level differentiation, functional analysis, and genome reconstruction.[3,4] Read length can also affect taxonomic resolution. Short-read sequencing yields high accuracy and depth but may fail to resolve repetitive genomic regions and closely related genomes, while long-read sequencing can resolve complex genomic regions, structural variation, and genome assembly, though it has historically had higher error rates and greater computational requirements.[5] After sequencing, reads are assembled into contigs and binned into metagenome-assembled genomes (MAGs) using features such as sequence composition and coverage to obtain genome-level characterization of organisms, including those that cannot be cultured.[6,7] After assembly and binning, reads, contigs, or MAGs can be assigned to specific microbial taxa using taxonomic classifiers that rely on marker-gene similarity, k-mer composition, whole-genome similarity, or reference databases. However, accurate species- and strain-level assignment still faces challenges due to incomplete reference genomes, horizontal gene transfer, closely related taxa, strain heterogeneity, uneven sample abundance, sequencing errors, and chimeric or fragmented assemblies.[8] These limitations provide an important rationale for integrating deep learning with metagenomics, as learned sequence representations may capture complex discriminatory patterns beyond conventional similarity- and reference-based approaches.

a) Sequencing approaches in metagenomics

There are several complementary sequencing strategies which are used depending on the resolution and scope required:

- **Amplicon metagenomics (marker-gene sequencing):** Targeted PCR amplification and sequencing of conserved phylogenetic markers, most commonly the 16S rRNA gene for bacteria/archaea, and the 18S rRNA gene or ITS region for fungi. This approach is cost-effective and well suited to broad community profiling.
- **Shotgun metagenomics:** Random fragmentation and sequencing of all DNA in a sample, generating short reads (typically Illumina platforms) that are mapped to reference genomes or assembled de novo. Shotgun

sequencing provides both taxonomic and functional (gene-level) information and achieves finer taxonomic resolution than amplicon approaches, at greater cost and computational burden. It provides the sequence diversity needed for species-level classification, strain resolution and genome-resolved analysis.

- **Long-read metagenomics:** This uses third-generation platforms to generate reads, spanning kilobases to megabases. Long reads improve genome assembly, resolve repetitive regions, and can span entire operons, supporting strain- and species-level differentiation. However, long-read platforms have historically exhibited higher per-base error rates than short-read platforms, although newer technologies have substantially reduced these errors.
- **Single-cell metagenomics (ScM):** This process involves the physical isolation of individual microbial cells through microfluidics or flow sorting. This approach directly links genomic content to individual organisms, helping to avoid the assembly ambiguity of bulk community sequencing, but it suffers from amplification bias and low throughput (a major limiting factor).

b) Traditional approaches for microbial species identification

Beyond metagenomic sequencing, researchers have developed numerous computational and molecular approaches for species-level microbial identification. Traditional alignment-based methods, such as BLAST, Bowtie2, and BWA, are commonly used for metagenomic sequence classification and taxonomic assignment because they generate easily interpretable matches against curated reference databases. However, these methods are inherently limited by the availability, quality, and taxonomic breadth of the reference genomes included in the database. Reliance on reference databases limits accurate classification of reads from poorly represented or highly divergent species and contributes to uncultured microbial "dark matter." Additionally, alignment-based searches may be computationally expensive for large metagenomic datasets when searching against extensive reference databases. These limitations are especially relevant for closely related microbial taxa, where high genomic similarity hinders species- and strain-level differentiation. Therefore, while alignment-based methods provide valuable reference-guided taxonomic evidence, their dependence on existing databases and computationally intensive searches necessitates complementary alignment-free and deep learning approaches that can learn discriminative sequence representations and potentially improve classification of novel or poorly characterized microbes.

Alignment-based methods: Sequence reads or contigs are aligned against curated reference databases (e.g., BLAST, Bowtie2, BWA) for taxonomic assignment based on best-match identity. While accurate when close references are available, these methods struggle with novel or underrepresented taxa.

Marker gene-based methods: Beyond 16S rRNA, single-copy universal marker genes enable faster, less computationally intensive taxonomic assignment with higher resolution than rRNA alone, but still rely on reference databases for classification.

Phylogenetic methods: Building phylogenetic trees from resolved gene sets or whole-genome alignments can place unknown sequences in an evolutionary framework. This can improve confidence in species boundaries, but it still requires significant computational time and high-quality assemblies.

K-mer-based methods: Methods like Kraken2, Centrifuge, and CLARK break sequences into small k-mers and align them against pre-computed k-mer databases. These methods are highly efficient and can handle large volumes of data, but their performance depends on choosing an appropriate k-mer length and avoiding reference database biases.

3. Deep learning approaches for microbial species differentiation

Deep learning (DL) is a family of machine learning methods built on artificial neural networks with multiple stacked layers, each learning increasingly abstract representations of the input data (Table 1). DL models learn feature extraction and classification jointly, end-to-end, guided only by a loss function and gradient-based optimization.[2,9] Four architecture families dominate current applications; they include:

3.1. Convolutional Neural Networks (CNNs)

In metagenomics, Convolutional Neural Networks (CNNs) apply deep learning to raw DNA sequences, gene prediction, and microbial classification. CNNs apply learnable filters (kernels) that slide across an input, detecting local patterns regardless of their position. This property suits DNA and protein sequences, where short motifs (transcription-factor binding sites, protein domains, k-mer signatures) carry biological meaning regardless of their exact location. In metagenomics, CNNs have been applied to gene prediction; for example, CNN-MGP numerically encodes candidate open reading frames and passes them through stacked convolutional and max-pooling layers to distinguish coding from non-coding regions.[10] It is also applicable to identifying phage-specific proteins directly from metagenomic sequencing reads, as in DeephageTP. DeephageTP is an alignment-free CNN classifier for tail proteins that traditional homology searches often miss because of high phage sequence diversity.[11] CNNs were also among the earliest DL architectures tested for 16S rRNA gene fragment classification and remain a standard baseline against which more elaborate architectures are benchmarked.[12]

3.2 Recurrent Neural Networks (RNNs)

Recurrent neural networks (RNNs) and their variants, particularly long short-term memory (LSTM) networks, model sequential dependencies by processing input elements in order. In metagenomics, this property makes them suitable for analyzing nucleotide sequences, temporal abundance profiles and other ordered representations in which relationships between observations may extend across multiple positions or time points. Unlike conventional machine-learning

Table 1: Comparison of deep learning architectures applied to metagenomics

Architecture	Mechanism	Strengths	Limitation	Applications	Taxonomic resolution	Tools
CNN	Learnable filters slide across one-hot or embedded sequence to detect local motifs; pooling reduces dimensionality.	Computationally efficient; detects position-invariant local motifs (k-mer/domain signatures).	Limited capacity to model long-range dependencies; one-hot-encoded CNNs underperformed on short-read taxonomic classification.[13]	Gene/ORF prediction (CNN-MGP); phage protein identification (DeephageTP); early 16S fragment classification.	Coarse, typically read- or fragment-level detection rather than fine taxonomic assignment.	CNN-MGP; DeephageTP
RNN / LSTM (+ attention)	Processes the sequence step by step via recurrent hidden states; gated LSTM units retain long-range signal; self-attention weighs informative positions.	Captures sequential/contextual dependencies; embedding + attention design outperformed one-hot CNN and hybrid baselines. [13]	Sequential computation limits parallelization; slower training/inference and larger memory footprint than transformers.[14]	Read-level species and genus classification (DeepMicrobes).	Species/genus-level.	DeepMicrobes
Transformer	Self-attention relates every sequence position to every other position in parallel; pre-trained genomic language models use masked or byte-pair-encoding objectives.	Efficient parallel training; strong long-range dependency modelling; transferable pre-trained embeddings reusable across tasks.	Substantial GPU/compute cost for pre-training; general multi-species models may need continued pre-training for microbiome specificity.[15,16]	Read classification (MetaTransformer); metagenome-level embeddings for disease prediction (MetagenBERT); foundational genomic language modelling (DNABERT-2).[17]	Read- to species-level, depending on fine-tuning.	MetaTransformer; MetagenBERT; DNABERT-2
GNN	Message passing propagates and aggregates information between connected nodes (contigs, microbes, diseases) across an explicit graph structure.	Exploits assembly-graph or knowledge-graph connectivity that composition/abundance-only methods discard; improves binning and disease-association prediction.	Requires a well-constructed graph; performance depends on graph quality and completeness.	Genome binning (GraphBin and related GNN binners); knowledge-graph-based pathogen/disease prediction (MetagenomicKG).[17,18]	Contig/genome-bin level rather than individual read level.	GraphBin; MetagenomicKG
Hybrid models	Combine two or more architectures or feature modalities — e.g., variational autoencoders fusing composition and abundance, or recurrent/attention layers atop k-mer embeddings.	Hybrid models can improve performance by integrating complementary sequence, abundance, taxonomic or graph-derived information, although the magnitude of improvement depends on the dataset and task.	Added architectural and training complexity; more hyperparameters to tune.	Genome binning (VAMB, AAMB, TaxVAMB); metagenome-level phenotype prediction combining read embeddings with clustering (MetagenBERT).	Varies by component architecture — spans read-level to genome-bin level.	VAMB; AAMB; TaxVAMB; MetagenBERT

approaches that require predefined features, RNNs can learn representations directly from sequential input data. RNN and LSTM models have been used to classify metagenomic sequences and identify microbes, with nucleotide sequences or encoded k-mers provided as ordered inputs. The ability to maintain information from previous positions can help learn dependencies beyond single-sequence motifs. LSTMs avoid the vanishing gradient problem of standard RNNs by using a gating mechanism to preserve information over longer sequences. This makes them potentially useful for classifying microbial taxa and recognizing sequence patterns associated with specific organisms or genomic functions. Despite these benefits, RNN/LSTMs have some significant drawbacks. Because RNNs process data sequentially, they can be less computationally efficient to train than parallelized transformer architectures, especially on large datasets. They might also struggle to learn very long-range dependencies and could be sensitive to sequence length and training data properties. CNNs tend to perform better for learning local sequence motifs and shorter-range patterns, while RNNs/LSTMs are better suited for learning ordered or temporal dependencies. While RNNs/LSTMs are useful for sequential representations, transformers have largely replaced them for large-scale genomic sequence modelling because they are more parallelizable and better at modelling long-range dependencies.

3.3 Transformer and genomic language models

A major advantage of transformers is their ability to capture long-range contextual relationships while allowing substantial parallelization during model training. This contrasts with CNNs, which are particularly effective at identifying local sequence motifs, and RNN/LSTM models, which capture sequential dependencies but process information recurrently. Graph neural networks (GNNs), in contrast, are particularly appropriate when the biological information is naturally represented as relationships among entities, such as contigs, genomes or nodes in a microbial interaction network. Thus, the choice of architecture should be determined by the structure of the biological information being modelled rather than by the assumption that one architecture is universally superior. The increasing use of transformer-based genomic language models also introduces new possibilities for transfer learning and foundation-model approaches in metagenomics. Instead of training a model independently for every downstream task, pretrained models can provide sequence representations that can be fine-tuned for taxonomic classification, functional prediction or microbial identification. This approach may be particularly valuable for metagenomic datasets with limited labelled examples. However, the effectiveness of transfer learning depends on the similarity between pretraining and target datasets, the diversity of organisms represented during pretraining, sequence length and computational resources. Transformer models can also require substantial memory and computational capacity, particularly when processing long genomic sequences. Transformer architectures have progressed from task-specific sequence classifiers to pretrained genomic foundation models. These models are trained on large collections of largely unlabeled DNA sequences and can subsequently be adapted to different downstream tasks, reducing dependence on task-specific labelled datasets. DNABERT demonstrated the potential of transformer-based language modelling for genomic sequences by learning contextual DNA representations using k-mer tokens and masked-language modelling. These representations could be transferred to several downstream genomic prediction tasks.[19] DNABERT-2 subsequently improved tokenization and computational efficiency using Byte Pair Encoding and was evaluated across multiple species, supporting the potential for more generalizable genomic representations.[20] Similarly, the Nucleotide Transformer family extended this approach through large-scale pretraining on diverse genomic sequences, demonstrating transferability across multiple genomic prediction tasks.[3] For metagenomics, these developments matter because pretrained sequence representations may facilitate taxonomic classification, species differentiation, and functional prediction, particularly where labelled datasets are limited. However, foundation models should not be assumed to eliminate reference or database bias, since their performance depends strongly on the diversity of organisms and sequences represented during pretraining. Future metagenomic applications will therefore require models that are taxonomically diverse, computationally efficient and transferable across microbial communities and environments.

3.4 Graph neural networks

GNNs operate on data with an explicit relational or topological structure, propagating and aggregating information between connected nodes through message passing. Metagenomic assembly naturally produces such a structure: the assembly graph (e.g., a de Bruijn graph) encodes overlap and connectivity relationships between contigs that classical binning tools, which rely only on per-contig composition and abundance, discard. Several recent tools exploit this structure directly; for example, GraphBin, which refines existing bin assignments by propagating labels across the assembly graph.[21] A variational-autoencoder-plus-GNN approach refines contig embeddings using neighborhood information from the assembly graph.[22] Lately, recent frameworks formulate metagenomic binning explicitly as representation learning over the assembly graph using message-passing GNN layers.[23] Beyond binning, knowledge-graph-based approaches apply GNNs to a curated graph linking microbes, genes, functions, diseases, and drugs. This is to generate sample-level embeddings that improve pathogen identification and body-site classification relative to taxonomic profiling alone.[24] GNNs can also be applied to determine relationships between genes, proteins, taxa, and biological functions can be combined for functional prediction and biological reasoning. A further use case is pathogen and phenotype prediction, where genomic

attributes can be modelled as connected nodes and edges. By pooling information from adjacent nodes, GNNs can discover patterns that may not be visible when considering each sequence individually. GNN effectiveness depends on the graph's accuracy and biological relevance, and building informative graphs from diverse metagenomic data remains a significant hurdle. Therefore, while CNNs are best for learning local sequence motifs, RNN/LSTMs for sequential/temporal patterns, and transformers for long-range contextual patterns, GNNs are most useful when the metagenomic problem is fundamentally relational/topological.

4. Data preprocessing and representation

The performance of deep learning models depends strongly on the quality, representation and normalization of metagenomic input data. Shotgun sequencing produces millions of short reads containing sequencing errors, adapter contamination, duplicates, host-derived sequences and sequences from organisms poorly represented in reference databases. Data preprocessing and feature representation in metagenomics involve cleaning raw sequencing reads, converting nucleotide strings into numerical formats, and applying deep learning architectures like convolutional networks (CNNs), autoencoders, and attention-based models to manage high-dimensional microbial profiles. The performance of any deep learning model in metagenomics depends heavily on how raw sequencing output is transformed into a numerical representation the network can consume. Unlike images or natural-language text, DNA sequences are compositional, strand-ambiguous, and high-dimensional at the read level, and derived from mixtures of unknown and often previously uncharacterized organisms. Researchers have developed several complementary encoding strategies to address these properties.

4.1. From raw reads to clean input

Shotgun metagenomic sequencing produces reads that carry several kinds of noise before any biological question is even asked. Sequencing errors accumulate toward the ends of reads on most platforms; adapter sequences left over from library preparation can appear as spurious nucleotides. In host-associated samples, host-derived DNA can add contamination and may constitute a substantial fraction of sequencing reads. A non-trivial share of reads will map back to the host genome rather than to any microbe. None of this is specific to deep learning; however, classical alignment-based pipelines have always had to deal with it too.[2] Hence, quality trimming, host-read removal, and de-duplication remain standard upstream steps regardless of the downstream model. These steps interact with later representation choices in ways that are easy to overlook. Preprocessing choices propagate into downstream feature representations and may therefore influence model performance, particularly when fixed-length k-mers or embeddings are used.

4.2 Sequence-level coding

These methods have the significant benefit of being "end-to-end," avoiding the computational expense of binning techniques, whether alignment-free or not, or using binning as a supplementary information source. The simplest approach, one-hot encoding, represents each nucleotide as a sparse binary vector (e.g., A = [1,0,0,0], C = [0,1,0,0], G = [0,0,1,0], T = [0,0,0,1]), preserving the raw sequence but producing an information-sparse representation that treats each base independently and represents the complementary (reverse) strand as an entirely unrelated matrix.[13] This scheme underlies many CNN-based genomic models, including CNN-MGP and DeephageTP, which feed fixed-length one-hot matrices directly into convolutional layers. [10,11] However, benchmarking on taxonomic classification of short metagenomic reads found that one-hot-encoded CNN and hybrid architectures achieved comparatively low accuracy and confidence, motivating alternative representations.[13] The alternative that has become dominant in taxonomic and phenotype-prediction models is k-mer-based encoding. Sequences are decomposed into overlapping substrings of length k, which are then either counted to produce fixed-length compositional frequency vectors (a natural extension of classical tetranucleotide-frequency features), or treated as discrete tokens, analogous to words in natural language processing, and mapped to dense, learned embedding vectors. DeepMicrobes popularized the k-mer embedding approach for metagenomic reads, drawing a direct analogy between k-mers and words in NLP; replacing one-hot encoding with a learned k-mer embedding layer was shown to substantially improve classification performance, reportedly because embeddings can encode taxonomic similarity between k-mers that one-hot vectors cannot capture.[13] Genomic language models extend this idea further by pre-training transformer-based tokenizers and encoders (e.g., byte-pair-encoding tokenization in DNABERT-2) on very large multi-species corpora, yielding general-purpose sequence embeddings that can be reused across many downstream metagenomic tasks without retraining from scratch.[15,20,25]

4.3. Compositional and abundance features

Tetranucleotide-frequency profiles have long been used as informative compositional signatures in microbial genome binning, although their discriminatory performance depends on genome characteristics and task context. Later tools extended this fusion strategy with adversarial autoencoders, contrastive and metric-learning approaches, and semi-supervised bimodal variational autoencoders that additionally incorporate predicted or refined taxonomic labels.[21,26,27] Sequencing depth varies from sample to sample, so raw counts are not directly comparable across samples; because counts within a sample are constrained to sum to a fixed total, standard statistical assumptions of independence do not hold, a problem generally referred to as compositionality. The conventional recommendation has long been to apply a log-ratio transformation, most commonly the centered log-ratio (CLR), which normalizes each feature

relative to the geometric mean of all features within a sample and is intended to move compositional data into a space where ordinary statistical and machine learning methods behave more reasonably. It is worth noting, though, that this recommendation is not as settled as it might appear: in a systematic comparison across several normalization schemes, Yerke *et al.* (2024) found that simpler proportion-based normalizations (essentially relative abundances, without the log-ratio machinery matched or outperformed CLR and related compositionally aware transformations when the downstream task was a random forest classifier, which complicates the assumption that more mathematically principled normalization automatically translates into better machine learning performance.[28]

4.4 Abundance-based approaches

Deep Learning approaches, through the autoencoding hypothesis, offer an intriguing basis for task-adapted dimensionality reduction.[18] Autoencoders can provide task-adapted dimensionality reduction, although their effectiveness depends strongly on sample size, model complexity, regularization, and the quality of the underlying abundance representation. The optimal type of autoencoder to employ, however, remains poorly understood and is still debated and under active research. For instance, DeepMicro trains various autoencoder types to determine which one extracts the most important information for illness prediction from metagenomic data.[29] Convolutional autoencoders (CAEs), variational autoencoders (VAEs), sparse autoencoders (SAEs), and denoising autoencoders (DAEs) were all tested and produced good results; however, none of them surpassed the other in terms of performance, and the optimal approach varied depending on the six different diseases it was tested on. Ensdeepdp used ensemble learning to obtain the optimal representation while accounting for these particularities.[30,31] An illness score is derived from the distance vector between the original input metagenome and the rebuilt output. The experiment can be replicated using numerous autoencoders, VAEs, and CAEs with different designs and parameters, then select the top k models. When analyzing a new metagenome, add the most informative representations to the original feature space by computing a matrix from the input data and the k best models' representations of that data.

4.5. Handling sparsity, compositionality, and class imbalance

Beyond the choice of encoding, metagenomic data preprocessing must address several domain-specific statistical challenges. Read- and taxon-abundance tables are highly sparse (most taxa are absent or undetected in most samples) and compositional (relative rather than absolute abundances, which sum to a constant and induce spurious correlations if analyzed with methods that assume independence).[2] Quality control steps, including adapter and low-quality base trimming, host-read removal, and de-duplication, are typically performed upstream of any DL model to reduce noise. Class imbalance is also common, since a handful of dominant taxa or a small set of characterized reference genomes can dwarf rare or novel lineages in training data; this can bias models toward already well-studied organisms unless addressed through resampling, class weighting, or self-supervised pre-training on unlabeled sequence, which does not require balanced labelled examples at all. Dimensionality reduction and normalization are frequently used both to make downstream models tractable and to mitigate compositional artefacts.[2] Data leakage is a critical concern in machine-learning benchmarking, particularly when closely related genomes, contigs, or reads from the same microbial species or strain occur in both training and test datasets. This can produce overly optimistic performance metrics, potentially obscuring the model's ability to generalize to novel taxa, strains, or new metagenomes.

5. Integration of deep learning with metagenomic workflows

To support metagenome coherence, users can input data beyond abundance tables, since raw metagenomic data are not always suitable for DL. They vary and may originate from external knowledge or the data itself. Rather than replacing the metagenomic pipeline wholesale, deep learning models are increasingly integrated as specialized modules that plug into, augment, or in some cases replace individual steps in the conventional analysis workflow: quality control, assembly, taxonomic profiling, binning, functional annotation, and downstream statistical or phenotype association analysis. Fig. 1 illustrates this integration end-to-end for microbial species differentiation, from raw sequencing reads through preprocessing, sequence representation, and model inference to species-level prediction and downstream application.

5.1 Taxonomic classification and profiling

Taxonomic classification is one of the most established deep learning applications in metagenomics. When training NNs for classification tasks, this data can be directly combined with abundance data and displayed as a taxonomy tree. Several methods for integrating taxonomy data have been tested: MDeep groups.[32,33] Before employing dense layers, the authors created a three-layer CNN meant to replicate the various levels of phylogeny and their interconnections. TaxoNN uses a similar but distinct method: it groups each abundance unit by phylum and trains a CNN for each phylum to learn characteristics unique to that phylum.[34] Afterwards, the network combines the feature vectors and uses them for the final classification. After that, the issue shifts from the species level to the phylum level, and each phylum is examined independently before the dense layers. Ph-CNN extends this concept by considering taxon proximity using distance metrics from the taxonomic tree. The k-nearest neighbors' abundances are convolved via a bespoke layer. The chosen distance metric significantly affects this procedure. The disadvantage is that, despite considering nearby taxa, it concentrates on local patterns

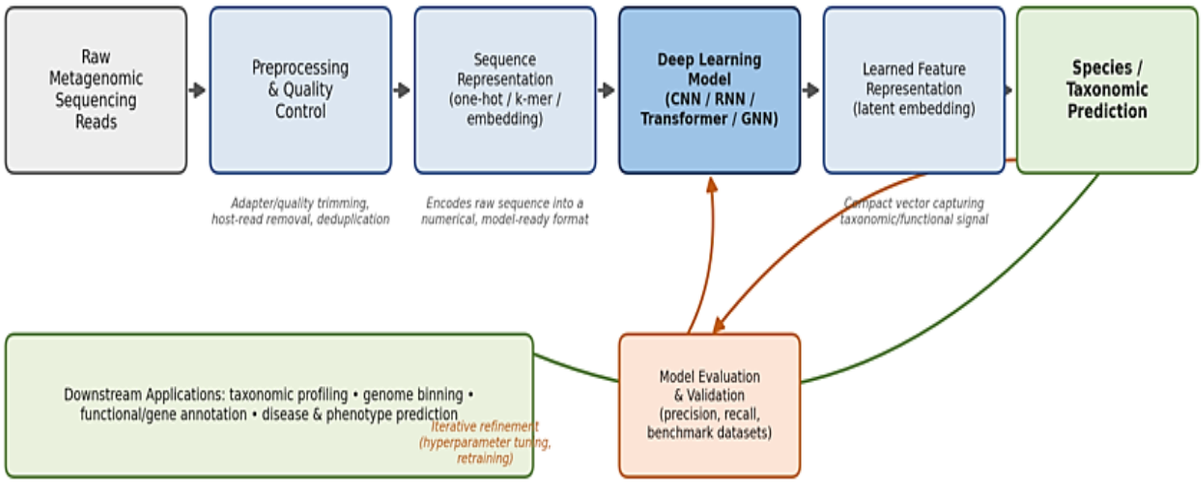


Fig 1: Framework for integrating metagenomics and deep learning in microbial species differentiation

rather than processing the data's global structure. Classical taxonomic classifiers rely on alignment or exact k-mer matching against reference databases, which struggle with reads from organisms poorly represented in those databases. Alignment-free DL classifiers address this by learning sequence-to-taxon mappings directly from labelled training reads. DeepMicrobes established the embedding-plus-recurrent-attention design for read-level species and genus classification, and Meta Transformer subsequently showed that a purely attention-based (transformer) encoder could match or exceed this performance with better training efficiency.[13,14] These models are typically deployed as a drop-in replacement for, or a complement to, conventional k-mer-matching classifiers within a standard shotgun-metagenomics pipeline, taking quality-controlled reads as input and outputting per-read or per-sample taxonomic assignments.[35]

5.2 Genome-resolved metagenomics: binning

Genome-resolved metagenomics enables reconstruction and analysis of metagenome-assembled genomes (MAGs) from complex microbial communities. Metagenomic assembly typically yields a highly fragmented set of contigs that must be grouped ("binned") into individual putative genomes. Contigs from the same genome are arranged into bins, each representing a distinct genome. Metagenomic species identification is not only a matter of accurately classifying reads, but also of de novo reconstructing genomes from complex microbial communities. Reads from multiple species may share substantial sequence similarity, vary in abundance, and contain repetitive or horizontally transferred regions, making it hard to assign them to specific species. To address these challenges, genome-resolved approaches first assemble short or long reads into contigs, and then use properties such as tetranucleotide frequencies, coverage, sequence embeddings and assembly-graph connectivity to group related contigs into metagenome-assembled genomes (MAGs). This reconstruction enables analysis of the complete genomic context, which may facilitate species identification, comparison of genomic similarities, and more robust identification of closely related species that may be difficult to differentiate at the read level alone. Deep learning models are increasingly valuable at this stage because they can integrate multiple complementary signals, including sequence composition, differential abundance, learned sequence embeddings, and contig–contig relationships, to improve genome binning and MAG reconstruction. Thus, accurate microbial species differentiation in complex metagenomes should be viewed as a multistage problem in which read-level classification and genome-level reconstruction are complementary rather than competing strategies. In this context, graph neural networks are particularly relevant because assembly graphs preserve relationships among connected contigs and can provide structural information that treating each sequence independently misses.

5.3 Gene prediction and functional annotation

Downstream of assembly and binning, genes and proteins must be identified and functionally characterized. CNN-based models such as CNN-MGP treat open-reading-frame prediction as a sequence-classification problem over numerically encoded candidate ORFs, while protein-level CNN classifiers such as DeephageTP identify functionally important but poorly conserved protein families (e.g., phage tail proteins) directly from metagenomic assemblies without relying on sequence homology to characterized references.[10,11] These tools are typically inserted after assembly (and, where applicable, binning) and before functional or taxonomic interpretation. They provide alignment-free alternatives or complements to homology-based annotation tools that struggle with the substantial fraction of "dark matter" genes lacking characterized homologs. Prodigal, the most popular and effective tool in this field, has consistently performed well across a variety of settings. Conventional annotation pipelines remain important because most

functional assignments still depend on homology, profile models or curated databases. DL methods should therefore be viewed primarily as complementary approaches, particularly for poorly characterized proteins and sequence "dark matter." Because they offer extensive resources across numerous biological research disciplines, these databases are widely utilized.[36]

5.4 Microbiome-associated phenotype prediction

While focusing on the functions of microorganisms in humans, animals, plants, and ecological niches, recent studies on microbial communities have actively investigated metagenomics of environmental sample diversity.[37] A major downstream application of metagenomics is linking microbiome composition or function to host phenotypes such as disease status. Historically, this has relied on species- or gene-abundance tables derived from reference-based profiling, fed into relatively shallow classifiers. Deep learning has been integrated primarily by improving the front-end representation: MetagenBERT builds phenotype-prediction features directly from foundational read embeddings rather than fixed abundance tables, reporting competitive or superior performance across several gut-microbiome disease benchmarks, including cirrhosis and type 2 diabetes.[16] Knowledge-graph-informed GNN embeddings that incorporate known relationships among microbes, functions, and diseases have likewise been shown to improve pathogen and disease-association prediction beyond what taxonomic profiling alone provides.[24]

5.5. Practical integration challenges

Nonetheless, several practical issues remain that affect the adoption of deep learning in metagenomic pipelines. Computational costs are high: transformer-based genomic LMs require expensive GPUs for pretraining and, in some cases, even inference when processing millions of reads per sample.[14,20] Interpretability is still a major concern in medical and biotech settings.[2] However, this dependency on references remains largely unresolved. Most DL classifiers still rely on extensive reference-based datasets, usually large, curated, and annotated training datasets, which reduces their use for applications in which the environment consists primarily of novel or poorly defined taxa. However, reference-free methods such as VAMB-style binning and self-supervised genomic language models offer some remedy.[2,38] Moreover, extending a trained model to new taxa, samples, or platforms typically requires retraining or transfer learning, and performance gains have not always been consistent across benchmarking datasets, highlighting the need for common benchmarking standards.[21,25] Another concern when assessing deep learning-based algorithms for metagenomic species differentiation is data leakage and lack of external validation. If taxonomic groups overlap between training and test datasets, the model is likely to perform much better at distinguishing between species because it has already encountered similar samples during training. Specifically, training and testing datasets may contain reads, contigs, genomes, or closely related strains from the same species or source, leading to taxonomic overlap between the training and testing datasets. In addition, benchmarks may be contaminated if sequences used to train a model are included in the test datasets or if the databases used for pre-processing overlap with the test data. As a result, a high accuracy score on a test dataset does not guarantee the model's ability to discriminate between truly unseen species or strains. Benchmarking should therefore use taxonomically independent dataset splits, distinct training and testing genomes, and external validation datasets not used during model training or parameter tuning. In addition, deep learning-based species differentiation models may see a significant drop in performance when applied to datasets from a different sequencing platform because of variations in sequencing technology, read length, error profiles, library preparation, or bioinformatics preprocessing. Furthermore, geographic and environmental biases may arise if the training dataset comprises mostly microorganisms from specific geographical areas, clinical samples, hosts, or ecological niches, making the model unable to generalize well to datasets comprising microorganisms from other geographical areas or ecosystems. Furthermore, a standard benchmark for species-level differentiation that includes closely related taxa, new or underrepresented organisms, multiple sequencing platforms, and various environmental and clinical scenarios is lacking. A standard benchmark that incorporates these factors would be necessary to enable meaningful comparisons of different deep learning approaches and to determine whether model improvements translate into real gains in microbial species differentiation in practical metagenomics applications.

6. Applications of deep learning in microbial species differentiation

6.1 Clinical settings

Deep learning tools can accelerate diagnosis through automation and improved laboratory tests. They are mainly used for the following purposes (Table 2):

a. Metagenomic sequence classification

Metagenomic Sequence Classification has the significant advantage of capturing all biological material present, rather than targeting a single pathogen or specific DNA region, as in traditional culture or standard PCR methods. DCiPatho integrates 3-to 7-k-mer frequency features into deep cross-fusion networks combining cross, residual, and deep neural network components, accurately identifying both learned and unlearned pathogenic bacteria from genomics and metagenomics datasets.[39,40] TCINet is a clinical diagnostic tool that uses a Sparse Neural Network and a Hierarchical Reasoning model to process raw sequencing reads and determine the taxonomic (evolutionary) relationships of clinical samples.[41] The sparse neural network forces the network to ignore noise and respect evolutionary tree relationships, while the hierarchical reasoning model passes evidence up and down the evolutionary tree to refine its final assessment.[41]

b. Environmental settings

Environmental microbiology presents different computational challenges, and deep learning offers better solutions than traditional tools. Many environmental microorganisms remain uncultured or difficult to cultivate under standard laboratory conditions. They exhibit extreme diversity and class imbalance, as environmental samples contain a few very common germs alongside numerous rare and new species. This problem is compounded because standard tools miss rare species. Finally, DNA extracted from environmental microorganisms is fragmented and low-coverage.[31,42] Traditional alignment-based bioinformatic software such as BLAST and Kraken2 compares uncultivated organisms' sequence similarity to reference databases. However, this approach often fails because of incomplete databases and fragmented sequencing reads.[43,44] In contrast, alignment-free CNNs and genomic language models such as DeepTaxa, DeepMicroClass, and ICCTax can process raw reads or contigs to assign taxonomic labels across heterogeneous ecological specimens.[40,44,45]

c. Agricultural & plant breeding

In agriculture, deep learning models address the challenge of differentiating beneficial soil or rhizosphere microbes from destructive phytopathogenic species. They support metagenomic and biomarker classification of rhizosphere microbial communities by analyzing them to distinguish between healthy and pathogen-suppressive soils. Models like MDeep and MetaPheno integrate multi-modal features by combining sequence data, metabolic potential, and evolutionary relationships with traditional taxonomy to determine the pathogenic potential of target microbial communities.[46,47] This approach captures functional redundancy among distinct species, decodes unclassified DNA from unknown species (i.e., "microbial dark matter"), and exploits evolutionary relationships between related species.

d. Food & industrial environment

CNNs classify microbial contaminants with 95–100% accuracy, identifying 12 species of Gram-positive bacteria, Gram-negative bacteria, and fungi. Yield predictions are made in minutes per sample, compared with traditional culturing methods that take 1 to 21 days.[47] Real-time predictive fermentation control is achieved with the following tools: A 1D/2D CNN processes multi-sensor data for Lactiplantibacillus plantarum and classifies batches collected in the first 24 hours of fermentation into successful, semi-successful, or failed outcomes with 97.87% accuracy, providing actionable early warnings to prevent batch failures.[48,49] Computer vision titer monitoring uses 1D-CNN models to continuously measure, predict, and optimize the titer of target products like gentamicin C1a, achieving an R^2 of 0.9862.[50] Autoencoders and one-class SVMs are used for unsupervised contaminant detection. They flag contaminated batches, achieving up to 1.0 recall and 0.99 specificity without requiring labelled contamination training samples.[51] Finally, in silico trait screening uses genome-scale deep learning models to perform virtual phenotypic screening across large microbial strain libraries and identify starters with desired industrial traits such as biosafety, pleasant flavor, and texture.

7. Challenges and limitations

Deep learning-based microbial species differentiation remains highly dependent on the quality, diversity, and representativeness of training data and reference databases. Training datasets may be biased toward well-studied organisms, geographic regions, environments,

Table 2: Representative studies applying deep learning to metagenomic microbial species differentiation

Study	Objective	Target Organism(s)	Deep Learning Model	Training Dataset	Performance Metrics	Key Conclusion
[44]	Species-level 16S taxonomic classification	Prokaryotic 16S rRNA sequences	Hybrid CNN-BERT	Greengenes2 2024.09; full-length and V3-V4 checkpoints	Accuracy, F1, calibration error	DeepTaxa reached 92.96% species accuracy and F1 0.9212, outperforming DADA2, QIIME 2, SINTAX, and Kraken 2
[56]	Classify metagenomic reads without a reference database	>3000 bacterial species; 639-species test set	CNN-based DL-TODA	Training on over 3000 bacterial species; test data from 2454 genomes in 639 species	Classification rate, rank-level accuracy, species accuracy	DL-TODA achieved species accuracy 0.97, above Kraken2 and Centrifuge on the same test set, but low-coverage species had poorer precision
[25]	Fast, accurate fungal ITS classification	Fungal ITS barcode sequences	MycoAI-CNN and MycoAI-BERT	>5 million labelled UNITE sequences	Species-level accuracy, speed, independent-test benchmarking	MycoAI-CNN was the fastest and most accurate model, classifying >300,000 sequences in 5 min, while BERT better clustered rare or unidentified taxa

sequencing platforms, and publicly available genomes, causing models to perform poorly on underrepresented or novel taxa. Likewise, incomplete or biased reference databases can skew taxonomic classification and prevent identification of organisms with no close references.[52,53] Another challenge is differentiating closely related species, especially when their genomes are highly similar. HGT exacerbates the problem as genes can move across taxonomic groups, which may lead to models that rely heavily on gene content or sequence features to assign incorrect taxa. These issues are especially relevant for metagenomic datasets that contain novel, rare, or highly diverse species.[2,4] The second main issue concerns model reliability and interpretability. Deep learning models can be black boxes, and we often do not know which genomic features lead the model to assign a particular species, or whether those features reflect a biologically relevant signal. Domain shift also poses a challenge: models trained on data from one laboratory, geographic area, sequencing platform, or community may not perform as well on another. Finally, data leakage can occur when closely related genomes or samples appear in both training and testing, inflating performance estimates. Therefore, external validation on independent datasets that differ taxonomically and/or geographically is important yet infrequently done. A recent review showed that many studies applying deep learning to metagenomics use relatively small datasets, simulations, or closely related validation datasets, which limits our understanding of generalizability.[2,54] Lastly, widespread adoption requires addressing computational cost, reproducibility, and standardization. Large deep learning models can require significant computational power for training and inference, while differences in sequencing protocols, preprocessing steps, databases, model architectures, hyperparameters, and performance metrics can limit reproducibility and comparability between studies. To ensure that models perform well and can be compared across platforms and different biological contexts, we need standardized datasets, benchmarking methods, reporting guidelines, and independent validation. Studies should provide full details of preprocessing and model specifications, avoid data leakage, and assess generalizability on truly independent datasets, with additional attention to explainability and uncertainty estimation. Ultimately, the continued advancement of microbial species differentiation requires not only the development of more accurate models but also the creation of standardized, diverse, reproducible, and biologically representative benchmarking frameworks. [2,55,53] Overall, for species-level taxonomic profiling, Kraken2 and MetaPhlAn remain highly competitive, while DL-based genome binners SemiBin and COMEBin outperform commonly used methods such as MetaBAT2. Therefore, current results suggest that DL-based methods are not always superior to standard methods for all metagenomic applications. Another major challenge is strain-level taxonomic profiling. Differentiating strains remains an unsolved challenge not only for DL-based approaches but also for standard methods, owing to the large number of shared sequences and similarities between strains. CAMI results also reflect inconsistent improvements from applying DL at the strain level. To understand if DL offers any improvement for strain-level profiling rather than just species-level profiling, extensive comparisons with state-of-the-art tools using identical data with similar evaluation measures such as precision, recall, F1-score, and abundance estimation, which unfortunately are lacking in the current studies, are necessary.[57]

Conclusion

Deep learning has moved from a niche method to a mainstream tool for metagenomics and has entered alignment-free architectures for taxonomic classification, joint representation learning approaches for genome binning, sequence-based approaches for predicting genes and functions, and increasingly powerful representations for associating microbiome data with host phenotypes. We expect these applications to continue to benefit from improved data representations (e.g., from one-hot encoding to k-mer embeddings and pretrained genomic language models), as well as improvements to model architectures. As the field moves forward, we see increasing interest in combining multiple model architectures (e.g., graph-aware transformer, taxonomy-informed autoencoder) and multiple types of features (e.g., composition, abundance, assembly graph, and prior biological knowledge) within a single model, while at the same time placing greater emphasis on interpretability, computational cost, and reference-free generalization to the vast amounts of microbial diversity that have yet to be characterized. Despite this progress, notable challenges remain. Although deep learning models show impressive results on benchmark datasets, it is unclear how well these translate to biological generalization because of data leakage, taxonomic overlap, benchmark contamination, reliance on references, sequencing platform differences, and geographic and environmental biases. Additionally, functional and phenotype predictions are not yet mature because the underlying processes that drive these traits are governed by complex biological and environmental factors that cannot always be inferred accurately from sequence alone. No consensus exists on the gold standard for species-level classification, and most studies lack external validation, making direct comparisons across models challenging. Together, these limitations make it difficult to apply these tools in clinical or environmental scenarios, where models must be robust to variation in laboratories, sequencing platforms, populations, and environments while also producing interpretable and reproducible predictions. Other factors that limit the utility of current approaches include the cost of training large models, lack of interpretability, and the need for high-quality training data. Moving forward, we believe the most promising path for improving deep learning approaches in microbial genomics is through the development of multimodal foundation models that combine information from multiple sources (e.g., genomic sequences, k-mer embeddings, genome or assembly graphs, functional and protein information, and relevant clinical or

environmental metadata) in a single framework. To advance the state of the art in species-level differentiation and functional and phenotypic prediction, we encourage future studies to use taxonomically independent datasets, present external validation, adopt standardized benchmarks, and report measures of uncertainty, all of which should be experimentally confirmed wherever possible. In short, we anticipate that the future of deep learning in microbial genomics lies not in maximizing accuracy on specific tasks but in building models that are generalizable, interpretable, and biologically validated, so they can predict accurately beyond the data on which they were trained.

Acknowledgments
Not applicable

CRediT Author Contribution Statement
David Odiba: Conceptualization, Writing - Original draft, Writing- Review and editing. **Faith Chidinma Terna:** Software, Writing - Original draft,. **Osuyi Gerard Uyi:** Writing - Original draft, Writing - Review and editing. **Samuel Anzaku:** Writing - Original draft, Writing-Review and editing. **Olukayode Olugbenga Orole:** Conceptualization, Writing - Original draft, Writing—Review and editing. All authors have read and approved the final version of the manuscript for publication and agree to be accountable for all aspects of the work, ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding Declaration
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors

Data Availability Statement
Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Conflict of Interest
There is no conflict of interest.

Artificial Intelligence (AI) Use Disclosure
The authors declare that artificial intelligence (AI)-assisted tools were used only for language refinement, grammar improvement, and manuscript structuring purposes during the preparation of this work. All technical content, experimental implementation, results, and interpretations were independently developed and verified by the authors.

Supporting Information
Not applicable.

References

[1] J. R. Marchesi, J. Ravel, The vocabulary of microbiome research: A proposal, *Microbiome*, 2015, **3**, 31, doi: 10.1186/s40168-015-0094-5.

[2] G. Roy, E. Prifti, E. Belda, J.-D. Zucker, Deep learning methods in metagenomics: A review, *Microbial Genomics*, 2024, **10**, 001231, doi: 10.1099/mgen.0.001231.

[3] H. Dalla-Torre, L. Gonzalez, J. Mendoza-Revilla, N. Lopez Carranza, A. H. Grzywaczewski, F. Oteri, C. Dallago, E. Trop, B. P. de Almeida, H. Sirelkhatim, G. Richard, M. Skwark, K. Beguir, M. Lopez, T. Pierrot, Nucleotide Transformer: Building and evaluating robust foundation models for human genomics, *Nature Methods*, 2025, **22**, 287–297, doi: 10.1038/s41592-024-02523-z.

[4] C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, N. Segata, Shotgun metagenomics, from sampling to analysis, *Nature Biotechnology*, 2017, **35**, 833–844, doi: 10.1038/nbt.3935.

[5] S. L. Amarasinghe, S. Su, X. Dong, L. Zappia, M. E. Ritchie, Q. Gouil, Opportunities and challenges in long-read sequencing data analysis, *Genome Biology*, 2020, **21**, 30, doi: 10.1186/s13059-020-1935-5.

[6] C. Chen, Y. Xu, O. Jian, X. Xiong, P. Labaj, A. Chmielarczyk, A. Rózanska, H. Zhang, K. Liu, T. Shi, J. Wu, VirulentHunter: Deep learning-based virulence factor predictor illuminates pathogenicity in diverse microbial contexts, *Briefings in Bioinformatics*, 2025, **26**, bbaf271, doi: 10.1093/bib/bbaf271.

[7] L.X. Chen, K. Anantharaman, A. Shaiber, A. M. Eren, J. F. Banfield, Accurate and complete genomes from metagenomes, *Genome Research*, 2020, **30**, 315–333, doi: 10.1101/gr.258640.119.

[8] F. Beghini, L. J. McIver, A. Blanco-Míguez, L. Dubois, F. Asnicar, S. Maharjan, A. Mailyan, P. Manghi, M. Scholz, A. Maltez Thomas, M. Valles-Colomer, G. Weingart, Y. Zhang, M. Zolfo, C. Huttenhower, E. A. Franzosa, N. Segata, Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3, *eLife*, 2021, **10**, e65088, doi: 10.7554/eLife.65088.

[9] J. Lyu, X. Zhang, J.-W. Tang, Y.-H. Zhao, S. Liu, Y. Zhao, N. Zhang, D. Wang, L. Ye, X.-L. Chen, L. Wang, B. Gu, Rapid prediction of multidrug-resistant *Klebsiella pneumoniae* through deep learning analysis of SERS spectra, *Microbiology Spectrum*, 2023, **11**, e04126-22, doi: 10.1128/spectrum.04126-22.

[10] A. Al-Ajlan, A. El Allali, CNN-MGP: Convolutional neural networks for metagenomics gene prediction, *Interdisciplinary Sciences: Computational Life Sciences*, 2019, **11**, 628–635, 10.1007/s12539-018-0313-4.

[11] Y. Chu, S. Guo, D. Cui, X. Fu, Y. Ma, DeephageTP: A convolutional neural network framework for identifying phage-specific proteins from metagenomic sequencing data, *PeerJ*, 2022, **10**, e13404, doi: 10.7717/peerj.13404.

[12] A. Mathieu, M. Leclercq, M. Sanabria, O. Perin, A. Droit, Machine learning and deep learning

applications in metagenomic taxonomy and functional annotation, *Frontiers in Microbiology*, 2022, **13**, 811495, doi: 10.3389/fmicb.2022.811495.

[13] Q. Liang, P. W. Bible, Y. Liu, B. Zou, L. Wei, DeepMicrobes: Taxonomic classification for metagenomics with deep learning, *NAR Genomics and Bioinformatics*, 2020, **2**, lqaa009, doi: 10.1093/nargab/lqaa009.

[14] A. Wichmann, E. Buschong, A. Müller, D. Jünger, A. Hildebrandt, T. Hankeln, B. Schmidt, MetaTransformer: Deep metagenomic sequencing read classification using self-attention models, *NAR Genomics and Bioinformatics*, 2023, **5**, lqad082, doi: 10.1093/nargab/lqad082.

[15] G. Roy, E. Belda, B. Hennecart, Y. Chevaleyre, E. Prifti, J.-D. Zucker, MetagenBERT: A Transformer-based architecture using foundational genomic large language models for novel metagenome representation, arXiv, 2026, doi: 10.48550/arXiv.2601.03295.

[16] G. Roy, E. Prifti, E. Belda, J.-D. Zucker, MetagenBERT: A transformer architecture using foundational DNA read embedding models to enhance disease classification, bioRxiv, 2025, doi: 10.1101/2025.05.06.652444.

[17] T. H. Nguyen, T. T. Phan, C. T. Dao, D. V. P. Ta, T. N. C. Nguyen, N. M. T. Phan, H. N. Pham, Effective disease prediction on gene family abundance using feature selection and binning approach, IT Convergence and Security: Proceedings of ICITCS 2020, 2020, **712**, 19–28, doi: 10.1007/978-981-15-9354-3_2.

[18] D. Wickramaratne, R. Wijesinghe, R. Weerasinghe, Human gut microbiome data analysis for disease likelihood prediction using autoencoders, 2021 21st International Conference on Advances in ICT for Emerging Regions (ICter), 2021, 49–54, doi: 10.1109/ICter53630.2021.9774811.

[19] Y. Ji, Z. Zhou, H. Liu, R. V. Davuluri, DNABERT: Pre-trained bidirectional encoder representations from transformers model for DNA-language in genome, *Bioinformatics*, 2021, **37**, 2112–2120, doi: 10.1093/bioinformatics/btab083.

[20] Z. Zhou, Y. Ji, W. Li, P. Dutta, R. Davuluri, H. Liu, DNABERT-2: Efficient foundation model and benchmark for multi-species genome, International Conference on Learning Representations, 2023, **2024**,41642-41665 doi: 10.48550/arXiv.2306.15006.

[21] V. Mallawaarachchi, A. Wickramarachchi, H. Xue, B. Papudeshi, S. R. Grigson, G. Bouras, R. E. Prahl, A. Kaphle, A. Verich, B. Talamantes-Becerra, E. A. Dinsdale, R. A. Edwards, Solving genomic puzzles: Computational methods for metagenomic binning, *Briefings in Bioinformatics*, 2024, **25**, bbae372, doi: 10.1093/bib/bbae372.

[22] A. Lamurias, M. Sereika, M. Albertsen, K. Hose, T. D. Nielsen, Metagenomic binning with assembly graph embeddings, *Bioinformatics*, 2022, **38**, 4481–4487, doi: 10.1093/bioinformatics/btac557.

[23] A. Lamurias, A. Tibo, K. Hose, M. Albertsen, T. D. Nielsen, Graph neural networks for metagenomic binning, Proceedings of the 2023 ICML Workshop on Computational Biology, 2023, doi: 10.34726/5406.

[24] C. Ma, S. Liu, D. Koslicki, MetagenomicKG: A knowledge graph for metagenomic applications, *bioRxiv*, 2024, doi: 10.1101/2024.03.14.585056.

[25] L. Romeijn, A. Bernatavicius, D. Vu, MycoAI: Fast and accurate taxonomic classification for fungal ITS sequences, *Molecular Ecology Resources*, 2024, **24**, e14006, doi: 10.1111/1755-0998.14006.

[26] S. Kutuzova, P. Piera Líndez, L. S. Danielsen, K. N. Nielsen, N. S. Olsen, L. Riber, A. Gobbi, L. M. Forero-Junco, P. Erdmann Dougherty, J. C. Westergaard, P. D. Browne, S. Christensen, L. Hestbjerg Hansen, M. Nielsen, J. Nybo Andersen, S. Rasmussen, Improving metagenome binning by integrating intrinsic features and taxonomy, *Nature Biotechnology*, 2026, doi: 10.1038/s41587-026-03098-0.

[27] P. P. Líndez, J. Johansen, S. Kutuzova, A. I. Sigurdsson, J. N. Nissen, S. Rasmussen, Adversarial and variational autoencoders improve metagenomic binning, *Communications Biology*, 2023, **6**, 1073, doi: 10.1038/s42003-023-05452-3.

[28] A. Yerke, D. Fry Brumit, A. A. Fodor, Proportion-based normalizations outperform compositional data transformations in machine learning applications, *Microbiome*, 2024, **12**, 45, doi: 10.1186/s40168-023-01747-z.

[29] M. Oh, L. Zhang, DeepMicro: Deep representation learning for disease prediction based on microbiome data, *Scientific Reports*, 2020, **10**, 6026, doi: 10.1038/s41598-020-63159-5.

[30] Y. Shen, J. Zhu, Z. Deng, W. Lu, H. Wang, EnsDeepDP: An ensemble deep learning approach for disease prediction through metagenomics, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2023, **20**, 986–998, doi: 10.1109/TCBB.2022.3201295.

[31] Y. Shen, K. Shi, C. Yu, R. Zhang, Y. Sun, J. Shang, PhageMind: Generalized strain-level phage host range prediction via meta-learning, *Bioinformatics*, 2026, **42**, btag262, doi: 10.1093/bioinformatics/btag262.

[32] H. Y. Wang, T. Hsieh, C. R. Chung, H. C. Chang, J. T. Horng, J. J. Lu, J. H. Huang, Efficiently predicting vancomycin resistance of Enterococcus faecium from MALDI-TOF MS spectra using a deep learning-based approach, *Frontiers in*

Microbiology, 2022, **13**, 821233, doi: 10.3389/fmicb.2022.821233.

[33] Y. Wang, T. Bhattacharya, Y. Jiang, X. Qin, Y. Wang, Y. Liu, A. J. Saykin, L. Chen, A novel deep learning method for predictive modeling of microbiome data, *Briefings in Bioinformatics*, 2021, **22**, bbaa073, doi: 10.1093/bib/bbaa073.

[34] D. Sharma, A. D. Paterson, W. Xu, TaxoNN: Ensemble of neural networks on stratified microbiome data for disease prediction, *Bioinformatics*, 2020, **36**, 4544–4550, doi: 10.1093/bioinformatics/btaa542.

[35] E. Levy Karin, M. Steinegger, Cutting-edge deep-learning based tools for metagenomic research, *National Science Review*, 2025, **12**, nwaf056, doi: 10.1093/nsr/nwaf056.

[36] N. Kim, J. Ma, W. Kim, J. Kim, P. Belenky, I. Lee, Genome-resolved metagenomics: A game changer for microbiome medicine, *Experimental and Molecular Medicine*, 2024, **56**, 1501–1512, doi: 10.1038/s12276-024-01262-7.

[37] N. N. Nam, H. D. Khoa Do, K. T. Loan Trinh, N. Y. Lee, Metagenomics: An effective approach for exploring microbial diversity and functions, *Foods*, 2023, **12**, 2140, doi: 10.3390/foods12112140.

[38] J. N. Nissen, J. Johansen, R. L. Allesøe, C. K. Sønderby, J. J. A. Armenteros, C. H. Grønbech, L. J. Jensen, H. B. Nielsen, T. N. Petersen, O. Winther, S. Rasmussen, Improved metagenome binning and assembly using deep variational autoencoders, *Nature Biotechnology*, 2021, **39**, 555–560, doi: 10.1038/s41587-020-00777-4.

[39] A. Green, C. Yoon, M. Chen, Y. Ektefaie, M. Fina, L. Freschi, M. Gröschel, I. Kohane, A. Beam, M. Farhat, A convolutional neural network highlights mutations relevant to antimicrobial resistance in Mycobacterium tuberculosis, *Nature Communications*, 2022, **13**, 3817, doi: 10.1038/s41467-022-31236-0.

[40] G. Jiang, J. Zhang, Y. Zhang, X. Yang, T. Li, N. Wang, X. Chen, F.-J. Zhao, Z. Wei, Y. Xu, Q. Shen, W. Xue, DCiPatho: Deep cross-fusion networks for genome scale identification of pathogens, *Briefings in Bioinformatics*, 2023, **24**, bbad194, doi: 10.1093/bib/bbad194.

[41] X. Peng, Y. Wei, X. Zhou, Enhancing pathogen identification through AI-assisted metagenomic sequencing, *Frontiers in Microbiology*, 2025, **16**, 1634194, doi: 10.3389/fmicb.2025.1634194.

[42] H. Park, S. J. Lim, J. Cosme, K. O'Connell, J. Sandeep, F. Gayanilo, G. R. Cutter, E. Montes, C. Nitikitpaiboon, S. Fisher, H. Moustahfid, L. R. Thompson, Investigation of machine learning algorithms for taxonomic classification of marine metagenomes, *Microbiology Spectrum*, 2023, **11**, e05237-22, doi: 10.1128/spectrum.05237-22.

[43] K. Vervier, P. Mahé, M. Tournoud, J. B. Veyrieras, J. P. Vert, Large-scale machine learning for metagenomics sequence classification, *Bioinformatics* 2016, **32**, 1023-1032, doi: 10.1093/bioinformatics/btv683.

[44] R. Salah, K. AbdElaal, L. Ghonaim, O. Awe, A. Moustafa, DeepTaxa: A hybrid CNN-BERT framework for 16S rRNA taxonomic classification, *Bioinformatics Advances*, 2026, **6**, vbag166, doi: 10.1093/bioadv/vbag166.

[45] Y. Gao, J. Bai, F. Zhou, Y. He, Y. Wang, X. Huang, ICCTax: A hierarchical taxonomic classifier for metagenomic sequences on a large language model, *Bioinformatics Advances*, 2025, **5**, vbaf257, doi: 10.1093/bioadv/vbaf257.

[46] Z. Deng, J. Zhang, J. Li, X. Zhang, Application of deep learning in plant–microbiota association analysis, *Frontiers in Genetics*, 2021, **12**, 697090, doi: 10.3389/fgene.2021.697090.

[47] Z. Z. Wang, D.W. Zeng, Y. F. Zhu, M. H. Zhou, A. Kondo, T. Hasunuma, X. Q. Zhao, Fermentation design and process optimization strategy based on machine learning, *BioDesign Research*, 2025, **7**, 100002, doi: 10.1016/j.bidere.2025.100002.

[48] A. Escobar-Zepeda, E. E. Godoy-Lozano, L. Raggi, L. Segovia, E. Merino, R. M. Gutiérrez-Rios, K. Juarez, A. F. Licea-Navarro, L. Pardo-Lopez, A. Sanchez-Flores, Analysis of sequencing strategies and tools for taxonomic annotation: Defining standards for progressive metagenomics, *Scientific Reports*, 2018, **8**, 12034, doi: 10.1038/s41598-018-30515-5.

[49] C. C. Wu, J.S. Chen, Y. C. Lu, J. S. Wu, Y. F. Huang, C. S. Liao, Predictive fermentation control of *Lactiplantibacillus plantarum* using deep learning convolutional neural networks, *Microorganisms*, 2025, **13**, 2601, doi: 10.3390/microorganisms13112601.

[50] F. Xu, L. Su, Y. Wang, K. Hu, L. Liu, R. Ben, H. Gao, A. Mohsin, J. Chu, X. Tian, A paradigm of computer vision and deep learning empowers the strain screening and bioprocess detection, *Biotechnology and Bioengineering*, 2025, **122**, 817–832, doi: 10.1002/bit.28926.

[51] X. Dũng, J. Nguyen, Y. Liu, C. McDowell, L. Dooley, Methodology for contamination detection and reduction in fermentation processes using machine learning, *Bioprocess and Biosystems Engineering*, 2025, **48**, 1547–1563, doi: 10.1007/s00449-025-03194-6.

[52] F. Meyer, A. Fritz, Z. L. Deng, D. Koslicki, T. R. Lesker, A. Gurevich, G. Robertson, M. Alser, D. Antipov, F. Beghini, D. Bertrand, Critical assessment of metagenome interpretation: the second round of challenges, *Nature methods,* 2022,19, 429-440, doi: 10.1038/s41592-022-01431-4.

[53] R. J. Wright, A. M. Comeau, M. G. I. Langille, From defaults to databases: Parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools, *Microbial Genomics*, 2023, **9**, 000949, doi: 10.1099/mgen.0.000949.

[54] A. Shrikumar, P. Greenside, A. Kundaje, Reverse-

complement parameter sharing improves deep learning models for genomics, bioRxiv, 2017, doi: 10.1101/103663.

[55] A. M. Schiffer, A. Rahman, W. Sutton, M. L. Putnam, A. J. Weisberg, A comparison of short- and long-read whole-genome sequencing for microbial pathogen epidemiology, *msystems*, 2025, **10**, e01426-e01425, doi: 10.1128/msystems.01426-25.

[56] C. Crès, A. Tritt, K. Bouchard, Y. Zhang, DL-TODA: A deep learning tool for omics data analysis, *Biomolecules*, 2023, **13**, 585, doi: 10.3390/biom13040585.

[57] B. Martin, T. D. Bennett, P. E. DeWitt, S. Russell, L. N. Sanchez-Pinto, Use of the area under the precision-recall curve to evaluate prediction models of rare critical illness events, *Pediatric Critical Care Medicine*, 2025, **26**, e855–e859, doi: 10.1097/PCC.0000000000003752.

Publisher Note: The views, statements, and data in all publications solely belong to the authors and contributors. GR Scholastic is not responsible for any injury resulting from the ideas, methods, or products mentioned. GR Scholastic remains neutral regarding jurisdictional claims in published maps and institutional affiliations.