Artificial Intelligence for Microbiology and Microbiome Research

Xu-Wen Wang[1], Tong Wang[1], and Yang-Yu Liu[1,2,*]

[1]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA
[2]Center for Artificial Intelligence and Modeling, The Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[*]Correspondence: yyl@channing.harvard.edu

SUMMARY

Advancements in artificial intelligence (AI) have transformed many scientific fields, with microbiology and microbiome research now experiencing significant breakthroughs through machine learning and deep learning applications. This review provides a comprehensive overview of AI-driven approaches tailored for microbiology and microbiome studies, emphasizing both technical advancements and biological insights. We begin with an introduction to foundational AI techniques, including primary machine learning paradigms and various deep learning architectures, and offer guidance on choosing between machine learning and deep learning methods based on specific research goals. The primary section on application scenarios spans diverse research areas, from taxonomic profiling, functional annotation & prediction, microbe-X interactions, microbial ecology, metabolic modeling, precision nutrition, clinical microbiology, to prevention & therapeutics. Finally, we discuss challenges unique to this field, including the balance between interpretability and complexity, the ”small n, large p” problem, and the critical need for standardized benchmarking datasets to validate and compare models. Together, this review underscores AI's transformative role in microbiology and microbiome research, paving the way for innovative methodologies and applications that enhance our understanding of microbial life and its impact on our planet and our health.

Contents

Introduction

For over 3.5 billion years, our planet and its inhabitants have been shaped by various microorganisms [1]. For example, Cyanobacteria, through photosynthesis, produced oxygen and contributed to the Great Oxygenation Event around 2.4 billion years ago, making the Earth hospitable for aerobic life [2]. Certain bacteria, like Rhizobium, fix atmospheric nitrogen into forms usable by plants, supporting plant growth and agriculture [3]. Commensal microbes in human and animal guts aid in digestion and nutrient absorption, essential for health and survival [4]. Similarly, some microbes can break down organic matter, recycling nutrients in ecosystems, which is vital for maintaining soil fertility and ecosystem balance [5]. Given the profound influence microorganisms have had on the evolution of life and the functioning of ecosystems, advancing microbiology research is crucial for understanding and harnessing these processes to benefit health, agriculture, and environmental sustainability.

It is not a big surprise that disrupted microbial communities (or microbiomes) can have a huge impact on our planet and ourselves. Indeed, agricultural practices, such as excessive use of chemical fertilizers and pesticides, can disrupt soil microbiomes, leading to reduced soil fertility and increased vulnerability to erosion [6]. Runoff containing pollutants and antibiotics can significantly disrupt the microbiomes of freshwater and marine ecosystems, leading to changes in water quality and impacting the health of aquatic life by altering the natural balance of microbial communities within the environment; this can potentially promote the growth of harmful bacteria and disrupt critical ecological processes like nutrient cycling [7, 8]. Many human diseases have been associated with disrupted microbiomes, including acne, eczema, dental caries, obesity, malnutrition, inflammatory bowel disease, asthma/allergies, hardening of arteries, colorectal cancer, type 2 diabetes, as well as neurological conditions such as autism, anxiety, depression, and post-traumatic stress disorder, etc [9, 10]. Gaining a deeper understanding of the activities of microbial communities, both within and around us, can greatly benefit our health and the health of our planet. This explains why in the past decades the microbiome has been a very active research topic in microbiology.

Artificial Intelligence (AI) focuses on creating intelligent machines that can execute tasks that usually need human intelligence. AI emerged as an academic discipline at the 1956 Dartmouth conference, shaped by pioneering work by Warren McCulloch, Walter Pitts, and Alan Turing on neural networks and machine intelligence. At first, AI research concentrated on symbolic reasoning, including early applications in biomedicine, such as the MYCIN expert system for diagnosing bacterial infections. Meanwhile, machine learning developed, showcasing algorithms that improved through data training. Despite early excitement and positive forecasts, the pace of AI advancement decelerated over the following decades, hindered by hardware constraints and unmet expectations, leading to a period known as ”AI winter.” However, the domain continued to progress, incorporating probabilistic methods to manage uncertainty. In around 2010, a new phase in AI emerged, fueled by breakthroughs in deep learning frameworks, the advent of powerful hardware (e.g., GPUs), open-source software tools, and greater access to extensive datasets (e.g., ImageNet [11]). In 2012, significant breakthroughs occurred when AlexNet (a deep learning architecture based on the convolutional neural network) surpassed preceding machine learning methodologies in visual recognition [12]. The subsequent innovations, particularly the Transformer (a deep learning architecture initially developed for machine translation) introduced in 2017 [13], triggered an ”AI boom” marked by considerable investment.

This surge in investment led to a wide range of AI applications by the 2020s, accompanied by increasing concerns regarding its societal implications and the pressing need for regulatory measures.

In this article, we review the application of various AI techniques in microbiology and microbiome research. We will focus on the applications of machine learning, particularly deep learning techniques. Traditional microbiologists excel in image analysis skills for identifying pathogens in Gram stains, ova and parasite preparations, blood smears, and histopathologic slides. They classify colony growth on agar plates for assessment. AI advances in computer vision can automate these processes, supporting timely and accurate diagnoses [14, 15]. Advances in sequencing technologies, especially next-generation sequencing, enable substantial numbers of samples to be processed rapidly and cost-efficiently [16]. The accessibility of large-scale microbiome datasets propelled the development of numerous AI (especially machine learning or deep learning) approaches in microbiome studies, as reviewed previously [17–51]. However, a comprehensive review of existing applications of AI techniques in microbiology and microbiome research is still lacking. This review article aims to fill this gap. The following sections are organized as follows. We first briefly describe various AI subfields, focusing on machine learning and the three basic machine learning paradigms. Next, we elaborate on the different deep learning techniques categorized under the three primary machine learning paradigms. Then, we systematically review the various applications of AI techniques in microbiology and microbiome research. Finally, we will present an outlook on the future directions of AI for microbiology and microbiome research.

Artificial Intelligence Techniques

The multiple subfields of AI research are focused on specific objectives and the utilization of distinct tools. The conventional objectives of AI research encompass searching, knowledge representation, reasoning, planning, learning, communicating, perceiving, and acting [52]. Most AI applications in microbiology and microbiome research rely on machine learning, which is the focus AI subfield of this Review.

Learning Paradigms

Machine learning is a subfield of AI that employs algorithms and statistical models, enabling machines to learn from data and improve their performance on specific tasks over time [53]. Machine learning is typically categorized into three primary learning paradigms: **supervised learning**, **unsupervised learning**, and **reinforcement learning**. These paradigms differ in the specific tasks they can address as well as in the manner in which data is presented to the computer. Generally, the nature of the task and the data directly influence the selection of the appropriate paradigm.

Supervised learning involves using labeled datasets, where each data point is linked to a class label. The algorithms in this approach aim to create a mathematical function that connects input features to the expected output values, relying on these labeled instances. Common uses include classification and regression. Classical machine learning methods for classification/regression include Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), Random Forest, Extreme Gradient Boosting (XGBoost), etc. Those methods have been heavily used in microbiology and microbiome research.

In unsupervised learning, algorithms analyze unlabeled data to detect patterns and relationships without any defined categories. This process uncovers similarities in the dataset and includes techniques like clustering, dimensionality reduction, and association rules mining. Classical unsupervised learning methods include k-means clustering, Principal Component Analysis (PCA), Principal Coordinate Analysis (PCoA), and t-distributed stochastic neighbor embedding (t-SNE) for dimension reduction, and the Apriori algorithm for association rules mining. Among them, PCoA is a commonly used tool in microbiome data analysis, particularly valuable for visualizing and interpreting the differences in microbial community composition between samples.

Reinforcement learning focuses on enabling intelligent agents to learn through trial-and-error in a dynamic environment to maximize their cumulative rewards [54–56]. Without labeled datasets, these agents make decisions to maximize rewards, engaging in autonomous exploration and knowledge acquisition, which is crucial for tasks that are difficult to program explicitly.

Integrating these paradigms can often lead to better outcomes. For instance, **semi-supervised learning** finds a middle ground by utilizing a small set of labeled data alongside a larger collection of unlabeled data. This method harnesses the strengths of both supervised and unsupervised learning, making it a cost-effective and efficient way to train models when labeled data is scarce. In situations where obtaining high-quality labeled data is difficult, **self-supervised learning** presents a viable alternative [57]. In this framework, models are pre-trained on unlabeled data, with labels generated automatically in subsequent iterations. Self-supervised learning effectively converts unsupervised machine learning challenges into supervised tasks, improving learning efficiency.

Transfer learning is another interesting machine learning technique, which involves taking a pre-trained model on a large dataset and fine-tuning it on a smaller, task-specific dataset [58, 59]. This approach leverages the knowledge acquired by the model during pre-training to improve performance on a new task. Transfer learning can be applied within both supervised and unsupervised learning paradigms, meaning it can utilize knowledge learned from either labeled or unlabeled data depending on the situation; essentially, transfer learning ”transfers” the learned representations from one task to another, regardless of whether the original task was supervised or unsupervised.

Note that both self-supervised learning and transfer learning leverage **pre-trained models** to improve performance on new tasks, but the key difference is that self-supervised learning generates its own labels, often called “pseudo-labels”, from unlabeled data during the pre-training phase, while transfer learning relies on existing labeled or unlabeled data for pre-training. Both self-supervised learning and transfer learning are extensively used in the training of **large language models** (LLMs), with self-supervised learning often being the primary method for pre-training on massive amounts of unlabeled data, while transfer learning allows the pre-trained model to be adapted to specific downstream tasks with fine-tuning on smaller labeled datasets. LLMs tailored for biology, e.g., genomic and protein language models [60–64], have numerous applications in microbiology and microbiome research. These models, trained on vast amounts of biological sequence data, can generate insights and predictions that are valuable across various areas in microbiology and microbiome research, as we discuss later.

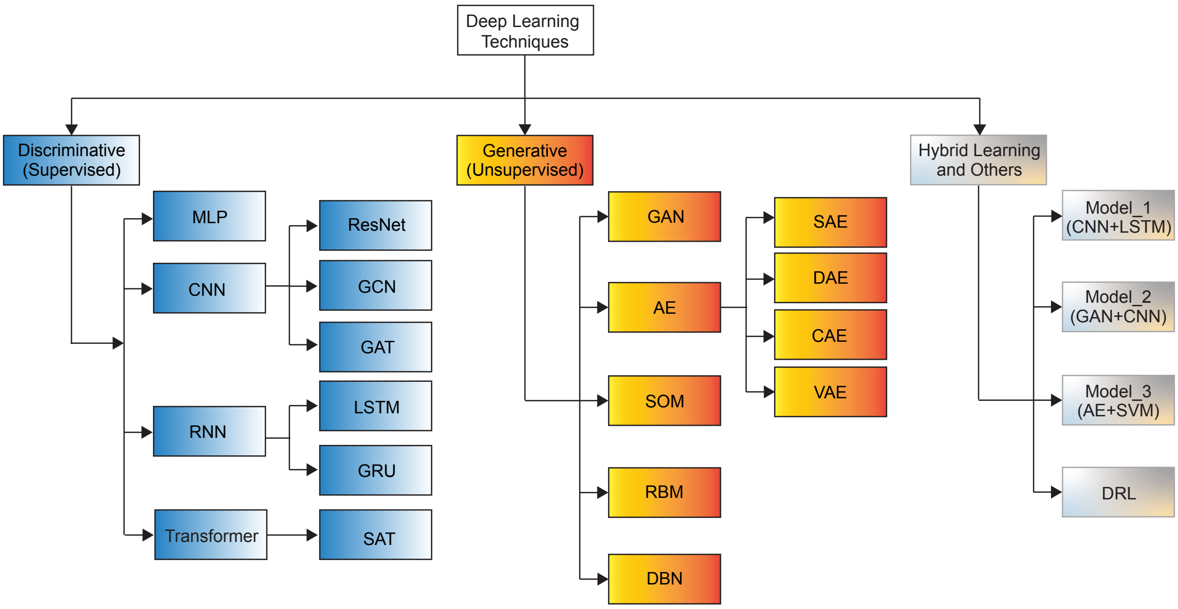

Figure 1. A taxonomy of deep learning techniques. Figure adapted from Ref [70]. MLP: Multi-Layer Perceptron; CNN: Convolutional Neural Network; ResNet: Residual Neural Network; GCN: Graph Convolutional Network; GAT: Graph Attention Network; RNN: Recurrent Neural Network; LSTM: Long Short-Term Memory; GRU: Gated Recurrent Unit; SAT: Structure-Aware Transformer; GAN: Generative Adversarial Network; AE: Auto-Encoder; SAE: Sparse Autoencoder; DAE: Denoising Autoencoder; CAE: Contractive Autoencoder; VAE: Variational Autoencoder; SOM: Self-Organizing Map; RBM: Restricted Boltzmann Machine; DBN: Deep Belief Network; DRL: Deep Reinforcement Learning.

Deep learning techniques

As a subfield of machine learning, deep learning represents a further specialization that utilizes deep neural networks to process and analyze large datasets, allowing for the automatic identification of patterns and the solving of complex problems. The reason why we often need a deeper rather than a wider neural network is that, if we regard a neural network as a function approximator, the complexity of the approximation function will typically grow exponentially with depth (not width). In other words, with the same number of parameters, a deep and narrow network has stronger expressive power than a shallow and wide network [65–69].

Based on the three primary machine learning paradigms, deep learning can be broadly divided into three major categories (Fig.1). The first category includes deep networks for supervised or discriminative learning, such as Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN) and their variants, Recurrent Neural Network (RNN) and their variants, as well as the Transformer. Roughly speaking, RNN propagates information through all hidden states in a sequential way, while CNN takes local information in developing each representation. By contrast, Transformer develops global contextual embedding via self-attention [13], which enables models to dynamically determine the relative importance of various words in a sequence, improving the ability to capture long-range dependencies. Another big advantage of Transformer is its easy parallelism. Unlike RNN, the Transformer can process entire sequences in parallel, which allows us to use GPUs for training. This significantly reduces the training time, and allows the use of very large models, often with hundreds of billions of parameters. These two advantages explain why the Transformer has facilitated so many LLMs, e.g., BERT, T5, GPT, PaLM, Gemini, and has revolutionized AI. As we will see later, all those deep network architectures in the first category (i.e., MLP, CNN, RNN, and Transformer), which were originally used for supervised learning, have been widely used in microbiome research.

The second category includes deep networks for unsupervised or generative learning, such as Generative Adversarial Network (GAN), Autoencoder (AE) and its variants, Self-Organizing Map (SOM), Restricted Boltzmann Machine (RBM), and Deep Belief Network (DBN). GAN is a very popular neural network architecture in recent years [71]. This architecture uses the idea of game theory to train two neural networks to compete with each other, thereby generating more realistic new data from a given training data set. AE is also a very common unsupervised neural network model, which can learn the latent features of the input data (called encoding), and at the same time use the learned features to reconstruct the original input data (called decoding) [72]. There are many variants of AE. Among them, the variational autoencoder (VAE) is probably the most famous one. VAE uses a probabilistic framework. Instead of mapping the input to a single point in the latent space, VAE maps the input to a distribution on the latent space, allowing for more flexible and expressive data representation [73]. As we will see later, both GAN and AE have been widely used in microbiome research. The other three models (SOM, RBM, and DBN) have not.

The third category includes deep networks for hybrid learning and relevant other tasks. There are three kinds of hybrid learning models: (1) An integration of different generative (or discriminative) models to extract more meaningful and robust features, e.g., CNN+LSTM, AE+GAN; (2) An integration of a generative model followed by a discriminative model, e.g., DBN+MLP, GAN+CNN, AE+CNN, etc; (3) An integration of generative or discriminative model followed by a non-deep learning classifier, e.g., AE+SVM, CNN+Random Forest, etc. As we will see later, all three hybrid learning models have been widely used in microbiome research. This category also includes Deep Reinforcement Learning (DRL). DRL is a subfield of machine learning that combines reinforcement learning and deep learning. Reinforcement Learning helps agents learn decision-making through trial and error. DRL improves this by using deep learning to extract decisions from unstructured data without manual state space engineering. DRL algorithms can take in very large inputs (e.g., an image of the raw board state and the history of states) and decide what actions to perform to optimize an objective (e.g., winning the game). A famous DRL algorithm is AlphaGo Zero, learning from playing the ancient Chinese game of Go without using any human knowledge [74]. So far, applications of DRL techniques in microbiome research are still very rare.

When to Use Machine learning vs. Deep learning?

We do not always need fancy deep learning techniques for microbiology and microbiome research. Sometimes we do not need deep learning at all. Logistic Regression or Random Forest might work very well. Choosing between deep learning and traditional machine learning methods depends on data characteristics, the specific problem at hand, available computational resources, and the need for model interpretability. Traditional methods are generally preferred for smaller, structured datasets and scenarios requiring interpretability (such as clinical applications), while deep learning excels with large, unstructured datasets and complex tasks requiring high performance.

If we decide to apply or develop deep learning methods to solve our problem, there is a general procedure [75]. First, we need to choose the appropriate performance metrics (e.g., Accuracy, Precision, Recall, F1-score, AUROC, AUPRC). Second, we need to find the default baseline deep learning models based on the data structure. For supervised learning tasks that involve fixed-size vector inputs, it is advisable to utilize a feedforward network featuring fully connected layers (e.g., MLP). If the input possesses a known topological structure, such as images or graphs, opting for CNN or its variants (e.g., graph convolutional network (GCN)) is recommended. When dealing with inputs or outputs that form sequences, we should consider using RNN and its variant (e.g., LSTM or GRU) or Transformer. 1D CNN or temporal convolutional network (TCN) might also work. Depending on the task, a hybrid deep learning model could also be considered. Third, we need to establish a reasonable end-to-end system, which involves choosing the appropriate optimization algorithm (e.g., SGD with momentum, Adam) and incorporating regularization (via early stop, dropout, or batch normalization). Finally, we need to measure the performance and determine how to improve it. We can either gather more training data or tune hyperparameters (e.g., learning rate, number of hidden units) via grid search or random search.

Application Scenarios

There are numerous applications of AI techniques in microbiome research. We can briefly group those applications into the following scenarios: taxonomic profiling, functional annotation & prediction, microbe-X interactions, microbial ecology, metabolic modeling, precision nutrition, clinical microbiology, prevention & therapeutics. For each application scenario, there are many specific tasks. In the following, we will present each of the specific tasks and the representative AI methods.

Taxonomic Profiling

A fundamental goal of microbiology and microbiome research is determining the compositions of microbial communities, i.e., identifying and quantifying different types of microorganisms (such as bacteria, fungi, viruses, and archaea) present in a given sample. This involves analyzing their relative abundances and diversity, often using DNA sequencing techniques. Currently, three generations of DNA sequencing techniques are available for microbiome research. The first-generation sequencing utilizes the chain termination method, offering read lengths of 500-1000 base pairs [76]. Second-generation sequencing, also known as next-generation sequencing (NGS), includes methods such as pyrosequencing, sequencing by synthesis, and sequencing by ligation, with read lengths ranging between 50 and 500 bp [77]. Two key NGS applications in microbiome research are (1) amplicon sequencing, which targets small fragments of one or two hypervariable regions of the 16S rRNA gene (for archaea and bacteria) or 18S rRNA gene (for fungi); and (2) metagenomic shotgun sequencing, which comprehensively samples all genes in all organisms present in a given community. NGS also offers short reads, with read lengths reaching 50-500 bp [77–79]. The third-generation sequencing performs single-molecule sequencing, offering long reads with lengths reaching tens of kilobases on average [80]. In the following, we discuss applications of AI techniques in various aspects of taxonomic profiling.

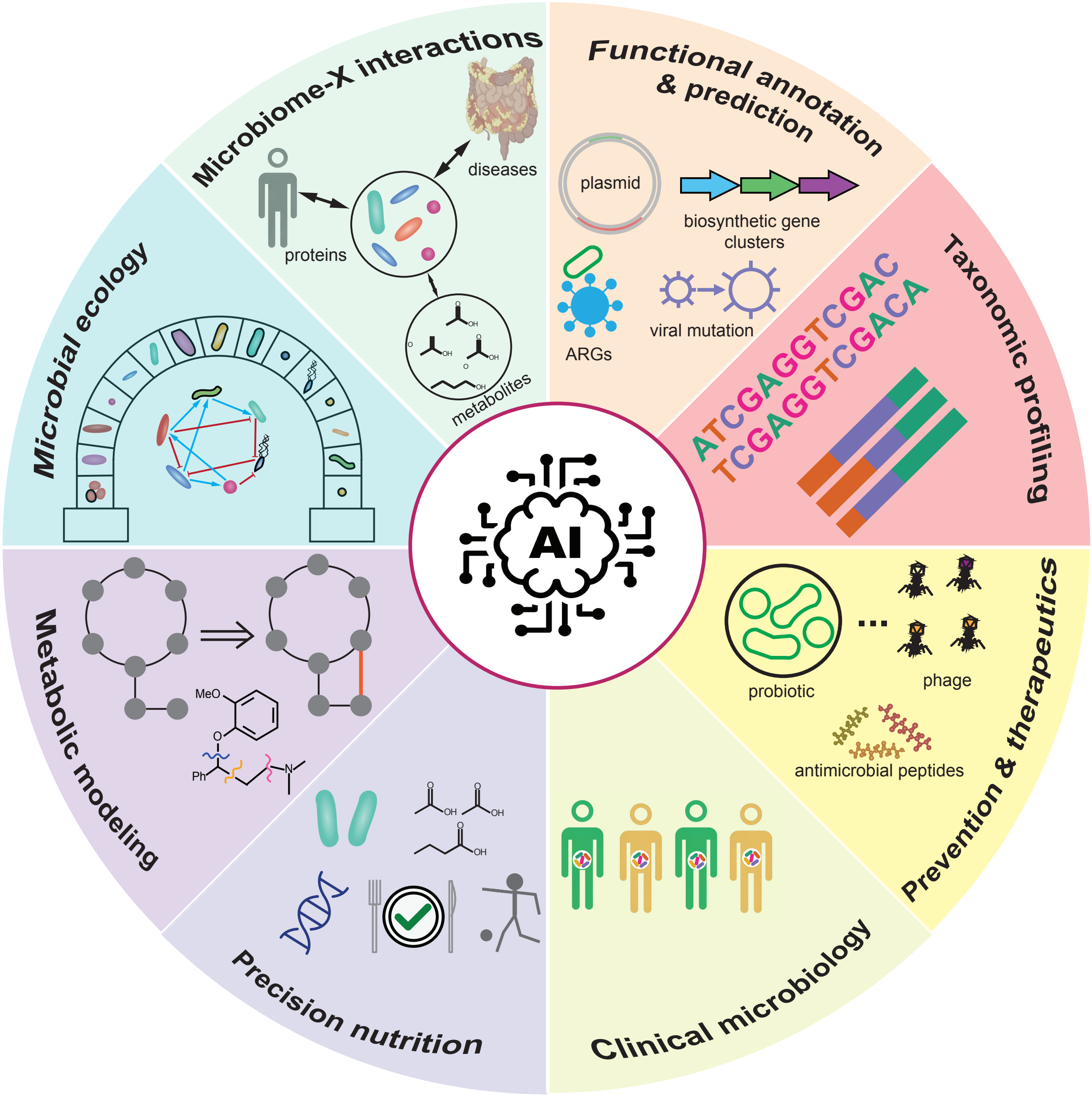

Figure 2. **Application scenarios of AI in microbiology and microbiome research.**

Metagenome assembly

Metagenomics refers to the direct study of the entire genomic information contained in a microbial community. Metagenomics avoids isolating and culturing individual microorganisms in a community and provides a way to study microorganisms that cannot be isolated and cultured. There are two main approaches for processing metagenomic sequencing data: (1) assembly-based and (2) reference database-based. The goal of the assembly-based approach is to construct and annotate the so-called metagenome-assembled genomes (MAGs) [81]. The construction and annotation of MAGs have greatly promoted our understanding of microbial populations and their interactions with the environment. It is worth noting that most MAGs represent new species, which helps to understand the so-called microbial dark matter. The process of constructing MAGs includes two main steps: assembly and binning. Assembly refers to the process of reconstructing longer sequences (contigs) from short DNA reads obtained through sequencing. This involves piecing together overlapping reads to form continuous sequences that represent parts of the genomes present in the microbial community.

Deep learning has been widely used in the quality control of metagenomic assembly. Many factors (e.g., sequencing errors, variable coverage, repetitive genomic regions, etc.) can produce misassemblies. For taxonomically novel genomic data, detecting misassemblies is very challenging due to the lack of closely related reference genomes. Deep learning methods can identify misassembled contigs in a reference-free manner. Representative methods include DeepMAsEd [82] and ResMiCo [83]. DeepMAsEd is based on CNN. Denote a contig as a sequence of nucleotides. At each position in the sequence, the concatenation of two types of information (raw sequence and read-count features) yields the input vector. To train and evaluate DeepMAsEd, one can generate a synthetic dataset of contigs, read counts, and binary assembly quality labels. As an extension of DeepMAsEd, ResMiCo is based on ResNet, a variant of CNN. The key feature of ResNet is the introduction of skip connections, which effectively solves the degradation problem of deep neural networks [84]. Compared to DeepMAsEd, ResMiCo leveraged a much more informative input vector computed from raw reads and contigs. Moreover, ResMiCo was trained on a very large and varied dataset. Through thorough validation, it was demonstrated that ResMiCo significantly outperforms other methods in accuracy, and the model remains robust when faced with novel taxonomic diversity and different assembly methods. We notice that both DeepMAsEd and ResMiCo used a carefully designed input vector. It would be interesting to explore if we can use a more advanced deep learning architecture (e.g., the Transformer) or a hybrid learning approach (e.g., CNN + RNN) to directly deal with the raw sequence data, avoiding the manual design of the input vector.

Metagenome binning

Metagenomic binning involves grouping those assembled sequences into clusters (bins or MAGs) that correspond to different species or genomes. Metagenomic binning helps in identifying and categorizing the different microorganisms present in a metagenomic sample, even if they are not fully assembled into complete genomes. There are many methods for metagenomic binning [85–88]. Several binning methods are based on deep learning, e.g., VAMB [89], CLMB [90], SemiBin [91], GraphMB [92], and COMEBin [93]. VAMB (Variational Autoencoders for Metagenomic Binning) uses VAE to encode sequence coabundance and k-mer distribution information, and clusters the resulting latent representation into genome clusters and sample-specific bins [89]. As an extension of VAMB, CLMB (Contrastive Learning framework for Metagenome Binning) can efficiently eliminate the disturbance of noise and produce more stable and robust results [90]. CLMB is based on contrastive learning, an machine learning approach that focuses on extracting meaningful representations by contrasting positive and negative instances [90]. SemiBin employs deep siamese neural networks to exploit the information in reference genomes, while retaining the capability of reconstructing high-quality bins that are outside the reference dataset [91]. Here, a siamese neural network (a.k.a. twin neural network) is a neural network that uses the same weights while working in tandem on two different input vectors to compute comparable output vectors [94]. GraphMB integrates GCN with assembly graphs to improve binning accuracy [92]. It models each contig using VAE for feature generation and aggregates these features using a GCN. This method accounts for read coverage in its loss function and uses iterative medoid clustering to finalize the binning. COMEBin is the latest metagenomic binning method [93]. This method is based on contrastive multiview representation learning. It introduces a data augmentation approach that generates multiple views for each contig, enabling contrastive learning and yielding high-quality representations of the heterogeneous features. Moreover, it incorporates a “Coverage module” to obtain fixed-dimensional coverage embeddings, which enhances its performance across datasets with varying numbers of sequencing samples. It also adapts an advanced community detection algorithm, Leiden, specifically for the binning task, considering single-copy gene information and contig length. COMEBin outperformed VAME and SemiBin on various simulated and real datasets, especially in recovering near-complete genomes from real environmental samples.

Taxonomic classification

All the methods discussed in the previous section are assembly-based metagenomic analysis methods. There are also many metagenomic analysis methods based on reference databases. In particular, those methods used for microbial classification and abundance estimation are also known as metagenomic profilers, which can be grouped into three categories based on the type of reference data [95]: (1) DNA-to-DNA methods (such as Bracken [96], Kraken [97, 98], and PathSeq [99]), which compare sequence reads with comprehensive genomes; (2) DNA-to-Protein methods (such as Diamond [100], Kaiju [101], and MMSeqs [102, 103]), which compare sequence reads with protein-coding DNA; (3) DNA-to-Marker methods (such as MetaPhlAn [104–107] and mOTUs [108, 109]), whose reference databases only contain specific gene families. It has been pointed out that the output of the first two categories is the sequencing abundance of species (without correction for genome size and copy number), while the output of the third category is the species abundance in a taxonomic or ecological sense [110]. Given these different types of relative abundances, benchmarking metagenomic profilers remains a big challenge [110].

These metagenomic profilers query DNA sequences in reference databases based on the concept of homology, which refers to the similarity between sequences of DNA, RNA, or protein that is due to shared ancestry. Obviously, those methods are largely affected by the quality of the reference database. A rather optimistic estimate suggests that the number of reference genomes in current comprehensive databases (such as RefSeq) may account for less than 5.319% of all species [111]. This explains why homology-based methods sometimes work poorly.

Deep learning techniques provide an alternative solution. These deep learning methods do not rely on similar sequences to exist in the reference database, and they allow for the modeling of complex correspondences between DNA sequences and corresponding species classifications. In these deep learning methods, DNA sequences are usually encoded into numeric matrices first, e.g., converting a sequence into a one-hot matrix or embedding the k-mers into a representative matrix. For example, DeepMicrobes is a deep learning method for taxonomic classification of short metagenomic sequencing reads [112]. In DeepMicrobes, DNA sequences are segmented into substrings, each mapped to a 100-dimensional embedding vector. These vectors are processed by a bidirectional LSTM and a self-attention layer, which prioritizes relevant k-mers (with k = 12) for the classification task. The LSTM outputs are combined with attention scores to produce an output matrix that feeds into a classifier for final species and genus identification. DeepMicrobes outperforms traditional tools like Kraken [97], Kraken2 [98] (where sequences are classified using the taxonomic tree), CLARK (using target-specific k-mer for classification) [113] in accuracy, but requires extensive computational resources and dataset sizes. Moreover, adding new species also necessitates retraining the entire network.

BERTax is another deep learning method for taxonomic classification. It classifies DNA sequences into three different classification levels, namely superkingdom (archaea, bacteria, eukaryotes, and viruses), phylum, and genus [114]. The novelty of BERTax is to assume DNA is a “language” and to classify the taxonomic origin based on this language understanding rather than by local similarity to known genomes in any database (i.e., homology). As its name suggests, BERTax is based on the state-of-the-art NLP architecture BERT (bidirectional encoder representations from transformers), which relies on a transformer employing the mechanism of self-attention. The training process of BERTax consists of two steps. First, BERT is pre-trained in an unsupervised manner, with the goal of learning the general structure of the genomic DNA “language”. Second, the pre-trained BERT model is combined with a classification layer and fine-tuned for the specific task of predicting classification categories. It has been shown that BERTax is at least comparable to state-of-the-art methods when similar species are part of the training data. However, for the classification of new species, BERTax significantly outperforms any existing method. BERTax can also be combined with database approaches to further increase the prediction quality in almost all cases.

Nanopore sequencing basecalling

Nanopore sequencing technology has enabled inexpensive long-read sequencing with reads longer than a few thousand bases [115]. The basic principle of nanopore sequencing is to pass an ionic current through a nanopore and measure the change in current when a biomolecule passes through or approaches the nanopore. Information about the change in current can be used to identify the molecule, a process often referred to as basecalling. There are two challenges in basecalling. First, the current signal level is most dominantly influenced by the several nucleotides that reside inside the pore at any given time, rather than a single base. Second, DNA molecules do not translocate at a constant speed. Basecalling is conceptually similar to speech recognition. Both processes involve interpreting complex signals to extract meaningful sequences—DNA bases in the case of basecalling, and spoken words in the case of speech recognition. Much like the evolution of speech recognition methods, computational methods for basecalling have evolved from statistical tests to hidden Markov models and finally deep learning models. Those methods are often referred to as basecallers.

Various deep learning models have been developed for basecalling. Chiron is the first deep learning model that can translate raw electrical signal directly to nucleotide sequence [116]. It applied a CNN to extract features from the raw signal, an RNN to relate such features in a temporal manner, and a connectionist temporal classification (CTC) decoder to create the nucleotide sequence. Here, CTC enabled us to generate a variant length base sequence for a fixed-length signal window through output-space searching, avoiding explicit segmentation for basecalling from raw signals. Similar to the Chiron architecture, SACall [117] (CATCaller [118] or Bonito [119]) integrated CNN with Transformer (Lite Transformer or LSTM) and CTC. Mincall [120] (or Causalcall [121]) directly integrated ResNet (or causal dilated CNN) with CTC. Halcyon used a different architecture. It combines a novel inception-block-based CNN module, an LSTM-based encoder, and an LSTM-based decoder using an attention mechanism. The inception-block-based CNN module aims to extract local features of input raw signal and reduce the dimension of the input timestep axis. The LSTM-based encoder captures long-time dependencies in the timestep dimension and deals with the variable lengths of inputs. The attention mechanism allows the decoder to focus on specific parts of the input sequence when generating each element of the output sequence.

All those methods mentioned so far treat basecalling as a sequence labeling task. URnano formalized the basecalling as a multi-label segmentation task that splits raw signals and assigns corresponding labels [122]. In particular, URnano used a U-Net with integrated RNNs. Here, U-Net is a u-shaped CNN architecture that was originally designed for biomedical image segmentation [123].

Benchmarking and architecture analysis of these deep learning-based basecallers show that: (1) the conditional random field (CRF) decoder is vastly superior to CTC; (2) complex convolutions are most robust, but simple convolutions are still very competitive; (3) LSTM is superior to Transformer and is depth dependent [124]. The reason why the attention mechanism in Transformer is not beneficial for basecalling could be the temporal relationships in the electric signal are local enough so that LSTM is sufficient for the task.

Functional Annotation & Prediction

Gene prediction

After carefully selecting MAGs from the metagenome assembly, we need to identify and annotate genes by recognizing potential coding sequences within MAGs [86]. This can be achieved by two types of methods: model-based methods (e.g., MetaGeneMark [125], Glimmer-MG [126] and FragGeneScan [127] using hidden Markov models, and Prodigal [128], MetaGene [129], MetaGeneAnnotator [130] using dynamic programming); and deep learning-based methods (e.g., Meta-MFDL [131], CNN-MGP [132], and Balrog [133]). Meta-MFDL generates a representation vector by integrating various features (e.g., single codon usage, mono-amino acid usage, etc.), and subsequently trains a deep stacking network to classify coding and non-coding ORFs. Here, the deep stacking network is composed of a series of modules with the same or similar structure stacked together. For Meta-MFDL, the authors used a simple MLP with only one hidden layer for each module. The “stacking” is completed by combining the outputs of all previous modules with the original input vector to form a new “input” vector as the input of the next module. CNN-MGP utilizes CNNs to automatically learn features of coding and non-coding ORFs from the training dataset and predict the probability of ORFs in MAGs. The authors extracted ORFs from each metagenomics fragment and encoded ORFs numerically. Then they built 10 CNN models for classification. Finally, they used 10 CNN classifiers to approximate the gene probability for the candidate ORFs, and used a greedy algorithm to select the final gene set. Balrog uses a TCN to predict genes based on a large number of diverse microbial genomes. The authors used the state of the last node of the linear output layer of the TCN as representative of the binary classifier, with a value close to 1 predicting a protein-coding gene sequence and 0 predicting an out-of-frame sequence. It is not clear which of those gene prediction methods is the best. Systematic benchmarking is necessary.

Antibiotic resistance genes identification

Antibiotics become less effective as bacterial pathogens develop and spread resistance over time. This has led to the antibiotic resistance crisis, e.g., resistance may involve most or even all the available antimicrobial options [134]. It has been estimated that antibiotic resistance could cause over 10 million deaths annually by 2050 if no significant action is taken. The economic costs associated with these outcomes could also reach approximately 100 trillion USD globally [135]. Some particular ecosystems, for instance, wastewater, have been considered reservoirs and environmental suppliers of antibiotic resistance due to the spreading of antibiotic resistance gene transfer between different bacterial species [136, 137]. Computational methods that can help identify potential resources of novel antibiotic resistance genes (ARGs) are particularly crucial.

DeepARG is a deep learning approach for predicting ARGs from metagenomic data [138]. First, genes in Uniprot were aligned to the CARD and ARDB databases using DIAMOND to obtain the dissimilarity representation, e.g., bit score after normalization so that scores close to 0 represent small distance or high similarity, and scores around 1 represent distant alignments. The final feature matrix indicates the sequence similarity of the Uniprot genes to the ARDB and CARD genes. The feature matrix was fed into four dense fully connected hidden layers and a SoftMax output layer to predict the probability of the input sequence against each ARG category. HMD-ARG is an end-to-end hierarchical multi-task deep learning framework for ARG annotation [139]. HMD-ARG used a CNN model where each sequence composed of 23 characters representing different amino acids was converted into one-hot encoding. Those sequence encodings were fed into six convolutional layers and four pooling layers to detect important motifs and aggregate local and global information across input sequences. The outputs of the last pooling layer were flatted and fed into three fully connected layers and a Softmax layer to predict final labeling [139]. HyperVR is a hybrid deep ensemble learning method that can simultaneously predict virulence factors and ARGs [140].

ARGNet is a two-stage deep learning approach that incorporates an unsupervised deep learning model autoencoder to first identify ARGs from the input genomic sequences and then uses a supervised deep learning model CNN to predict the antibiotic resistance category for sequences determined as ARGs by the autoencoder [141]. This hybrid learning approach enables a more efficient discovery of both known and novel ARGs. It was shown that ARGNet outperformed DeepARG and HMD-ARG in most of the applications and reduced inference runtime by up to 57% relative to DeepARG.

Ground-breaking LLMs initially created for NLP have found success in predicting protein functions. These models, referred to as protein language models (PLMs), excel at generating intricate semantic representations that forge meaningful links between gene sequences and protein functions [62–64]. FunGeneTyper is a PLM-based deep learning framework designed for accurate and scalable prediction of protein-coding gene functions [142]. This framework includes two interconnected deep learning models: FunTrans and FunRep. While these models share a similar architecture, they are tailored for classifying functional genes at type and subtype levels, respectively. Both models utilize modular adapter-based architectures, incorporating a few additional parameters for efficient fine-tuning of extensive PLMs. Specifically, utilizing the ESM-1b model (a large-scale PLM built on a 33-layer transformer architecture [62]), adapters are inserted into each transformer layer, serving as individual modular units that introduce new weights tuned for specific tasks. FunGeneTyper has shown exceptional performance in classifying ARGs and virulence factor genes. More significantly, it is a flexible deep learning framework that can accurately classify general protein-coding gene functions and aid in discovering numerous valuable enzymes.

Plasmid identification

Plasmids are small, typically circular DNA molecules that are found in many microorganisms, e.g., Bactria, Archaea, and Eukaryota, which play an important role in microbial ecology and evolution through horizontal gene transfer, antibiotic resistance, and ecological interaction, etc. Identifying plasmid sequences from microbiome studies can provide a unique opportunity to study the mechanisms of plasmid persistence, transmission, and host specificity [143].

Many classical machine learning methods have been proposed for plasmid identification, e.g., cBar [144] based on sequential minimal optimization, PlasClass [145] using Logistic Regression, PlasmidVerify [146] using Naïve Bayesian classifier, PlasForest [147], Plasmer [148], Plasmidhunter [149], RFPlasmid [150] and SourceFinder [151] using Random Forest. Several deep learning methods have also been developed for plasmid identification. For example, PlasFlow employs MLP for the identification of bacterial plasmid sequences in environmental samples [152]. It can recover plasmid sequences from assembled metagenomes without any prior knowledge of the taxonomical or functional composition of samples with high accuracy. Deeplasmid is another deep learning method for distinguishing plasmids from bacterial chromosomes based on the DNA sequence [143]. It leverages both LSTM and fully connected layers to generate features, which are then concatenated and passed to another block of fully connected layers to generate the final output — the Deeplasmid score $y \in [0, 1]$. The higher the score is for the sequence, the more likely it is to be a true plasmid. plASgraph2 is a new deep learning method for identifying plasmid contigs in fragmented genome assemblies built from short-read data [153]. The innovation of plASgraph2 lies in its use of GCN and the assembly graph to propagate information from neighboring nodes, resulting in more accurate classification. The GCN model consists of a set of graph convolutional layers designed to propagate information from neighboring contigs within the assembly graph. plASgraph2 generates two scores for each graph node: a plasmid score and a chromosomal score, which are used to assess whether a contig is likely derived from a plasmid, chromosome, or both.

Note that both plasmids and viruses are mobile genetic elements — a type of genetic material that can move around within a genome or be transferred from one species to another. Mobile genetic elements are often referred to as selfish genetic elements, because they have the ability to promote their own transmission at the expense of other genes in the genome. Mobile genetic elements are found in all organisms. The set of mobile genetic elements in an organism is called a mobilome, including viruses, plasmids, transposons, integrons, introns, etc. Recently, deep learning methods have been developed to simultaneously identify both viruses and plasmids, the two major components of the mobilome. For example, PPR-Meta is the first tool that can simultaneously identify phage and plasmid fragments from metagenomic assemblies efficiently and reliably [154]. PPR-Meta leveraged a novel architecture, Bi-path CNN, to improve the performance for short fragments. The Bi-path CNN leverages both base and codon information to enhance performance: the "base path" is effective for extracting sequence features of noncoding regions, while the "codon path" is useful for capturing features of coding regions. geNomad is a hybrid framework that combines the strengths of alignment-free and alignment-based models for concurrent identification and annotation of both plasmids and viruses in sequencing data [155]. To achieve that, geNomad processes user-provided nucleotide sequences via two distinct branches. In the sequence branch ("alignment-free"), the inputs are one-hot encoded and passed through an IGLOO neural network, which evaluates them by identifying non-local sequence motifs. In the marker branch ("alignment-based"), the proteins encoded by the input sequences are annotated with markers specific to chromosomes, plasmids, or viruses. Here, the key idea behind the IGLOO neural network is to leverage the relationships between "non-local patches" sliced from feature maps generated by successive convolutions to effectively represent long sequences, allowing it to handle both short and long sequences efficiently, unlike traditional RNNs which struggle with very long sequences [156].

Biosynthetic gene clusters prediction

Natural products are chemical compounds that serve as the foundation for numerous therapeutics in the pharmaceutical industry [157]. In microbes, these natural products are produced by clusters of colocalized genes known as biosynthetic gene clusters (BGCs) [158]. Advances in high-throughput sequencing have led to a surge in the availability of complete microbial isolate genomes and metagenomes, offering a great opportunity to discover a vast number of new BGCs. Deep learning models have been very useful in this genome mining effort [159–162].

For example, DeepBGC and its extension employ (1) Pfam2vec (a word2vec-like word embedding model, which is a shallow neural network with a single hidden layer); (2) a Bidirectional LSTM (a classical RNN), which offers the advantage of capturing short- and long-term dependencies between adjacent and distant genes. e-DeepBGC still leverages those neural networks, but improves DeepBGC in the following aspects [159]. First, e-DeepBGC employs Pfam names, Pfam domain summary, Pfam domain clan information. This additional information is used to create new embedding of each Pfam domain by providing more biological information than that encoded by Pfam2vec which only uses the Pfam names. Second, a novel data augmentation step is introduced to overcome the limited number of BGCs with known functional classes.

BiGCARP is a self-supervised neural network masked language model [161]. It is based on the convolutional autoencoding representations of proteins (CARP), a masked language model of proteins. That's why it is called Biosynthetic Gene CARP (or BiGCARP). The CARP is based on CNN, and has been shown to be competitive with transformer-based models for protein sequence pretraining [163]. SanntiS (Secondary metabolite gene cluster annotations using neural networks trained on InterPro signatures) is a new method for BGC prediction [164]. At the core of SanntiS is the detection model, a neural network with a one-dimensional convolutional layer, plus a bidirectional LSTM. This is quite similar to DeepBGC. The authors claimed that SanntiS outperforms DeepBGC, but it was not compared with BiGCAPR. Therefore, systematic benchmarking work is warranted.

16S rRNA copy number prediction

The 16S rRNA gene is highly conserved across different bacterial species but contains hyper-variable regions that provide species-specific signatures. By sequencing these regions, we can determine the composition and diversity of bacterial communities in various environments. Yet, different bacterial species can have varying numbers of 16S rRNA gene copies (ranging from 1 to 21 copies/genome), which can lead to biases in quantifying microbial communities if not accounted for [165]. To accurately estimate the relative abundance of bacterial species in a microbiome sample, we need to adjust the proportion of 16S rRNA gene read counts by the inverse of the 16S rRNA gene copy number. Experimentally measuring the 16S rRNA gene copy numbers through whole genome sequencing or competitive PCR is expensive and/or culture-dependent. To resolve this limitation, based on the hypothesis that 16S rRNA gene copy number correlates with the phylogenetic proximity of species, many bioinformatics tools have been developed to infer 16S rRNA gene copy numbers from taxonomy or phylogeny [166–169]. Yet, an independent assessment demonstrated that regardless of the method tested, 16S rRNA gene copy numbers could only be accurately predicted for a limited fraction of taxa [170].

Recently, a deep learning-based method ANNA16 was developed to predict 16S rRNA gene copy numbers directly from DNA sequences, avoiding information loss in taxonomy classification and phylogeny [171]. Essentially, ANNA16 treats the 16S GCN prediction problem as a regression problem. A stacked ensemble model (mainly consisting of MLP and SVM) is the core of ANNA16. The 16S rRNA gene sequences were first preprocessed with K-merization. The resulting k-mer counts (with k=6) and the existing 16S rRNA gene copy number data (retrieved from rrnDB database) were used to train the stacked ensemble model. Based on 27,579 16S rRNA gene sequences and copy number data, it has been shown that ANNA16 outperforms previous methods (i.e., rrnDB, CopyRighter, PAPRICA, and PICRUST2). We expect that in the near future more deep learning-based methods will be developed to solve this fundamental problem in microbiology and microbiome research.

Mutation/evolution prediction

Predicting evolution has been a longstanding objective in evolutionary biology, offering significant implications for strategic pathogen management, genome engineering, and synthetic biology. In microbiology, evolution prediction has been studied for several microorganisms. For instance, Wang et al. used the evolutionary histories of *Escherichia coli* to train an ensemble predictor to predict which genes are likely to have mutations given a novel environment [172]. To achieve that, they first created a training dataset consisting of more than 15,000 mutation events for *E. coli* under 178 distinct environmental settings reported in 95 publications. For each mutation event, they recorded its genome position with respect to a reference genome and the mutation event type (e.g., single-nucleotide polymorphisms (SNPs), deletions, insertions, amplifications, inversions). Then, they integrated a deep learning model MLP and two classical machine learning models, Support Vector Machine and Naive Bayes, to build an ensemble predictor to predict the mutation probability of any given gene under a new environment. The input of the ensemble predictor consists of 83 binary variables (features) that capture attributes related to the strain, medium, and stress from experiments. The model output is a binary variable that captures the presence/absence of mutation(s) in any given gene, computed from the predicted probability of this gene's mutation event. This work clearly illustrated how the evolutionary histories of microbes can be utilized to develop predictive models of evolution at the gene level, clarifying the impact of evolutionary mechanisms in specific environments. One limitation of this approach is that those 83 features were manually selected, which relies on domain knowledge.

Another interesting work is EVEscape, a generalizable modular framework that can predict viral mutations based on pre-pandemic data [173]. It has been shown that EVEscape, if trained on sequences available before 2020, is as accurate as high-throughput experimental scans in predicting pandemic variation for SARS-CoV-2 and is generalizable to other viruses (such as influenza, HIV, Lassa, and Nipah). The EVEscape framework is based on the assumption that the probability that a viral mutation will induce immune escape is the joint probability of three independent events: (1) this mutation will maintain viral fitness ('fitness' term); (2) the mutation will occur in an antibody-accessible region ('accessibility' term); and (3) the mutation will disrupt antibody binding ('dissimilarity' term). All three terms can be computed from pre-pandemic data sources, providing early warning time critical for vaccine development. The accessibility and dissimilarity terms are computed using biophysical information. The fitness term is computed via the deep learning of evolutionary sequences. In particular, the authors computed the fitness term using EVE [174], a deep generative model (i.e., VAE) trained on evolutionarily related protein sequences that learn constraints underpinning structure and function for a given protein family.

Long-term and system-level evolution has also been systematically examined. Konno et al. clearly demonstrated that the evolution of gene content in metabolic systems is largely predictable by using ancestral gene content reconstruction and machine learning techniques [175]. They first inferred the gene content of the ancestral species using the genomes of 2894 bacterial species (encompassing 50 phyla) and a reference phylogeny. Then they applied two classical machine learning models (logistic regression and random forest) to predict which genes will be gained or lost in metabolic pathway evolution, using the gene content vector of the parental node in the phylogenetic tree. Their framework, Evodictor, successfully predicted gene gain and loss evolution at the branches of the reference phylogenetic tree, suggesting that evolutionary pressures and constraints on metabolic systems are universally shared. It would be interesting to see if deep learning techniques can be applied to predict metabolic system evolution.

Microbe-X Interactions

Recent advancements in microbiology and microbiome research have significantly deepened our understanding of the complex interactions between the microbes and the host, diseases, and drugs. In this section, we will discuss how deep learning-based methods have facilitated the inference of those complex interactions.

Microbe-host interactions

A disrupted gut microbiome has been linked to a wide variety of diseases, yet the mechanisms by which these microbes affect human health remain largely unclear. Protein-protein interactions (PPIs) are increasingly recognized as a key mechanism through which gut microbiota influence their human hosts [31, 176–178]. A vast and largely unexplored network of microbe-host PPIs may play a significant role in both the prevention and progression of various diseases. Future research is needed to further uncover these interactions and their potential therapeutic implications.

Many machine learning methods have been developed to predict PPIs. Basically, they can be grouped into three categories: sequence-based, structure-based, and network-based [31]. Sequence-based methods utilize amino acid sequences to predict PPIs. For instance, PIPR employs a deep residual recurrent CNN within a siamese architecture to select local features and maintain contextual information without predefined features [179]. Similarly, DeepPPISP integrates global and contextual sequence features by applying a sliding window approach to neighboring amino acids and utilizing a TextCNN architecture to treat the protein sequence as a one-dimensional image for global feature extraction [180]. Additionally, hybrid approaches have been developed for microbe-host PPI prediction, combining a denoising autoencoder (unsupervised learning) with logistic regression (supervised learning) [181]. Another model, DeepViral, enhances performance by incorporating infectious disease phenotypes alongside protein sequences for microbe-host PPI prediction [182].

Structure-based methods leverage the three-dimensional structures of proteins to predict PPIs. For example, DeepInterface is one of the first methods to use 3D CNNs for predicting PPI interfaces at the atomic level [183]. Different from DeepInterface, MaSIF (Molecular Surface Interaction Fingerprints) uses geometric deep learning to process non-Euclidean data, breaking proteins into overlapping patches with specific physicochemical properties to predict PPI interfaces [184]. Graph-based neural network methods, where nodes represent atoms or amino acid residues linked by edges based on spatial proximity or chemical bonds, apply convolutional filters on the graph representation of proteins to predict interactions while being invariant to rotation and translation. PECAN further integrates a graph CNN with an attention mechanism and transfer learning, using sequence-based conservation profiles and spatial distance features to predict antigen-antibody interactions [185].

Network-based methods consider the PPI prediction problem as a link prediction task, using inferring missing links based on existing network knowledge. These methods have been benchmarked across various interactomes, demonstrating that advanced similarity-based methods, which leverage the network characteristics of PPIs, outperform other link prediction methods [186]. These general-purpose methods can be tailored for microbe-host PPI prediction. Moreover, integrating sequence-based, structure-based, and network-based approaches can leverage the strengths of each approach, potentially leading to more accurate and robust PPI predictions.

Of course, the microbe-host interactions are not limited to PPIs. Besides PPIs, microbes can interact with the host through many other mechanisms, including: (1) Gene regulation:

Microbial metabolites can influence host gene expression via epigenetic changes or signaling pathways. (2) Immune modulation: Microbes interact with the host immune system, educating immune cells and promoting tolerance or inflammation. (3) Metabolite production: Gut microbes produce metabolites like short-chain fatty acids (SCFAs), which influence host energy metabolism, immune function, and gut health. (4) Gut barrier function: Microbes can strengthen or disrupt the gut barrier, affecting intestinal permeability.

Machine learning methods have been developed to study some of those mechanisms. For example, Morton et al. developed mmvec, a neural-network-based method to analyze microbe-metabolite interactions [187]. It takes microbial sequence counts and metabolite abundances from various samples as the input and outputs the estimated conditional probabilities of observing a metabolite given the presence of a specific microbe. This method is similar to a popular word embedding method in NLP, i.e., word2vec, which is a shallow neural network with a single hidden layer [188]. Note that in the original application of word2vec, the skip-gram technique (i.e., creating word embeddings that focus on predicting surrounding words based on a specific word or target word) was employed to account for the sequential nature of the text. For microbiome and metabolome data, there is no clear sequential nature. Therefore, in mmvec, the skip-gram was replaced by multinomial sampling, where a single microbe is randomly sampled from a microbiome sample at each gradient descent step. Morton et al. evaluated mmvec's performance against traditional methods like Pearson's, Spearman's, SparCC, and SPIEC-EASI correlations, and found it demonstrated greater specificity and sensitivity, especially when applied to complicated datasets with vast amounts of microbiome and metabolomics information.

Microbe-disease associations

The exploration of microbe-disease associations (MDAs) is crucial for understanding various health conditions and tailoring effective treatments. Traditional studies directly correlate microbial features with disease outcomes, creating MDA databases such as HMDAD [189] and mBodyMap [190]. Advanced deep-learning methods have also been developed to infer new MDAs, including NinimHMDA [191], LGRSH [192], BPNNHMDA [193], and DMFMDA [194].

NinimHMDA uses a multiplex heterogeneous network constructed from HMDAD and other biological databases [191]. By integrating biological knowledge of microbes and diseases represented by various similarity networks and utilizing an end-to-end GCN-based mining model, it predicts different types of HMDAs (elevated or reduced) through a one-time model training. Predicting HMDAs is akin to solving a link-prediction problem within a multiplex heterogeneous network. In terms of predictive performance, NinimHMDA was compared with several existing methods such as DeepWalk [195], metapath2vec [196].

Similar to NinimHMDA, LGRSH [192] and BPNNHMDA [193] were developed for the same predictive task but with different deep-learning architectures. LGRSH applies graph representation techniques to predict associations, using calculated similarities between microbes and diseases [192]. BPNNHMDA uses a back-propagation neural network to predict potential associations [193]. DMFMDA employs deep matrix factorization and Bayesian Personalized Ranking to predict associations [194]. Unfortunately, we haven't seen any benchmark studies that systematically compare those deep learning methods in predicting microbiome-disease associations.

Very recently, thanks to the advancements in large language models, extraction of MDAs directly from biomedical literature has become much easier than before. For example, Karkera et al. demonstrated that pre-trained language models (specifically GPT-3, BioMedLM, and BioLinkBERT), when fine-tuned with domain and problem-specific data, can achieve state-of-the-art results for extracting MDAs from scientific publications [197]. The extracted MDAs will further expand the human MDA database. We expect that those deep learning methods will be more powerful with an expanded human MDA database.

Deep learning techniques have also been leveraged to study the association between microbes and specific diseases. For instance, MICAH is a deep learning method based on a heterogeneous graph transformer to study the relationships between intratumoral microbes and cancer tissues [198]. The inputs of MICAH are the species abundance matrix and sample labels (i.e., cancer types of samples). From the inputs, MICAH constructs a heterogeneous group with two types of nodes (microbes and samples), and three types of edges (species-species metabolic edges based on the NJS16 database [199], species-species phylogenetic edges based on the NCBI Taxonomy database, species-sample edges representing the relative abundance of a species in a sample). Then, MICAH used a two-layer graph transformer to update node embeddings and a fully connected layer based on updated node embeddings to perform sample node (cancer type) classification. Finally, MICAH extracts the attention scores of species to samples from the well-trained model to output subsets of microbial species associated with different cancer types. This framework significantly refines the number of microbes that can be used for follow-up experimental validation, facilitating the study of the relationship between tumors and intratumoral microbiomes.

Microbe-drug associations

Accumulated clinical studies show that microbes living in humans interact closely with human hosts, and get involved in modulating drug efficacy and drug toxicity. Microbes have become novel targets for the development of antibacterial agents. Therefore, screening of microbe–drug associations can benefit greatly drug research and development. With the increase of microbial genomic and pharmacological datasets, we are greatly motivated to develop effective computational methods to identify new microbe–drug associations.

Many deep-learning methods have been recently developed to identify microbe–drug associations, e.g., GARFMDA [200], GCNATMDA [201], LCASPMDA [202], MCHAN [203], MDSVDNV [204], NMGMDA [205], OGNNMDA [206], STNMDA [207], etc. Most of the deep learning methods can be divided into six different categories based on the deep learning model they used [208], e.g., CNN-based, GCN-based autoencoder, Graph Attention Network(GAT)-based autoencoder, Collective Variational Autoencoder (CVAE), Sparse Autoencoder (SAE). A recent method STNMDA is an exception [209]. STNMDA integrates a Structure-Aware Transformer (SAT) with an MLP classifier to infer microbe-drug associations. It begins with a “random walk with a restart” approach to construct a heterogeneous network using Gaussian kernel similarity and functional similarity measures for microorganisms and drugs. This heterogeneous network was then fed into the SAT to extract attribute features and graph structures for each drug and microbe node. Finally, the MLP classifier calculated the probability of associations between microbes and drugs. A systematic comparison of those existing methods using benchmark datasets is warranted.

Microbial Ecology

Deciphering inter-species interactions and assembly rules of microbial communities are fundamental but challenging questions in microbial ecology. Efforts based on population dynamics models have been made. However, parameterizing those dynamics models is very challenging [210]. Deep learning approaches can overcome such challenges by learning the assembly rules implicitly without knowing the population dynamics. Especially, with the prominent progress in metagenomics and next-generation sequence technologies, collecting large-sample size data is feasible, providing sufficient diverse communities to train deep learning models.

Microbial interactions prediction

Microbes interact with each other and influence each other's growth in various ways. The microbial interactions can be represented as a directed, signed, and weighted graph, i.e., the ecological network of the microbial community. Inferring the microbial interactions is important to understand the systems-level properties and dynamics of the microbial communities. Typically, this is achieved by analyzing high-quality longitudinal [211–215], or steady-state data [216], which is hard to obtain for large-scale microbial communities. Recently, the traditional random forest classifier was proposed to tackle this issue [217]. For each species, a trait is represented as a binary code in its trait vector. For each species pair within a community, a composite trait vector is created by concatenating the trait vectors of both species. This composite vector is then related to the observed responses of the interacting species. All interactions observed are utilized to train the classifier, which predicts the results of unobserved interactions. This approach has been evaluated in three case studies: a mapped interaction network of auxotrophic *Escherichia coli* strains, a soil microbial community, and a comprehensive *in silico* network illustrating metabolic interdependencies among 100 human gut bacteria. The results demonstrated that having partial knowledge of a microbial interaction network, combined with trait-level data of individual microbial species, can lead to accurate predictions of missing connections within the network, as well as propose potential mechanisms for these interactions. It would be very interesting to explore if deep learing methods can further improve the prediction of microbial interactions.

Microbial composition prediction

cNODE (compositional neural ordinary differential equation) is a deep learning method that can predict the community compositions from the species assemblages for a given ecological habitat of interest, e.g., the human gut [218]. All microbial species that can inhabit this habitat form a species pool or meta-community. A microbiome sample collected from this habitat can be considered as a local community assembled from the meta-community. The species assemblage of this sample is characterized by a binary vector, where the entry indicates if species-i is present (or absent) in this sample. The community composition is characterized by a compositional vector, where the ith-entry represents the relative abundance of species-i. cNODE aims to implicitly learn the community assembly rules by learning the mapping from species assemblage into community composition. To learn such a mapping, cNODE used Neural ODE [219], which can be interpreted as a continuous limit of the ResNet architecture [84]. Extensive simulations suggest that the sample size in the training data acquired to reach a relatively accurate prediction should be twice the species pool size. cNODE has been successfully applied to predict compositions of the ocean and soil microbiota, Drosophila melanogaster gut microbiota, and the human gut and oral microbiota.

Instead of relying on species assemblage, MicrobeGNN employs a graph neural network-based approach to predict the microbial composition at steady state from the genomes of mixed bacteria, with each species represented by a node [220]. Bacterial genomes are encoded into binary feature vectors that indicate the presence or absence of specific genes. Two types of GNNs, GraphSAGE [221] and MPGNN [222], are utilized for node and edge computations, respectively. Due to the lack of prior knowledge regarding the exact graph topology, fully connected graphs are employed, allowing each node to influence all other nodes within a single message-passing step. The results demonstrate that GNNs can accurately predict the relative abundances of bacteria in communities based on their genomes across various compositions and sizes.

Note that neither cNODE nor MicrobeGNN utilizes environmental or host factors in predicting microbial compositions. Incorporating environmental/host factors into deep learning models might further improve the accuracy of microbial composition predictions.

Keystone species identification

By implicitly learning the community assembly rules, cNODE or its variant enables us to predict the new community compositions after adding or removing any species or any species combinations via thought experiments. In particular, predicting the impact of species' removal facilitates the identification of keystone species that have a disproportionately large effect on the structure or function of their community relative to their abundance [223]. Note that the impact of a species' removal naturally depends on the resident community, i.e., a species may be a keystone in one community but not necessarily a keystone in another community. In other words, the keystoneness of a species can be highly community-specific.

The DKI (Data-driven Keystone species Identification) framework is based on cNODE [223]. In the DKI framework, the keystoneness of species in microbial communities was defined as the product of two components: the impact component and the biomass component. The impact component quantifies the impact of species's removal on the structure of community, while the biomass component captures how disproportionate this impact is.

The DKI framework was validated using synthetic data generated from a classical population dynamics model in community ecology, i.e., the Generalized Lotka-Volterra (GLV) model, and then applied to compute the keystoneness of species in the human gut, oral microbiome, and the soil and coral microbiome. It was found that those taxa with high median keystoneness across different samples display strong community specificity, and some of them have been reported as keystone taxa in literature. Instead of studying the impact of removing a single species, the DKI framework can be extended to study the impact of removing any species combinations, and hence study keystone duos or trios, etc, in complex microbial communities. Instead of studying the impact of removing a single species, the DKI framework can be extended to study the impact of removing any species combinations, and hence study keystone duos or trios, etc, in complex microbial communities.

Colonization outcome prediction

Microbial communities are typically subject to various environmental perturbations, e.g., antibiotic administration and diet, which can impact the balance of the microbial ecosystem and cause or exacerbate disease [224]. Machine learning models can be trained on some observed communities and make predictions for those unobserved communities upon similar perturbations. For instance, MLP has been used to predict the temporal gut community composition of termite perturbed by six different lignocellulose food sources [225]. In addition to predicting the impact of diet change on microbial composition, machine learning methods have also been used to predict the colonization outcomes of exogenous species for complex microbial communities [226]. Those machine learning methods treat the baseline (i.e., pre-invasion) taxonomic profile as inputs and the steady state abundance of the invasive species as output or mathematically, learn the mapping from the baseline taxonomic profile of a community to the steady state abundance of the invading species. Validation of the approach using synthetic data and two commensal gut bacteria species *Enterococcus faecium* and *Akkermansia muciniphila* in hundreds of human stool-derived *in vitro* microbial communities, showed that machine learning models, including random forest, linear regression/logistic regression, and neural ODE can predict not only the binary colonization outcome but also the final abundance of the invading species [226].

Fecal microbiota transplantation (FMT) has shown a high success rate for the treatment of recurrent *Clostridioides difficile* infection (rCDI). However, the mechanisms and dynamics dictating which donor microbiomes can engraft in the recipient are poorly understood. Traditional machine learning models, e.g., random forest, have been applied to predict the post-FMT bacterial species engraftment [227]. We expect that, given high-quality training data, deep learning methods can also be used to predict species engraftment and outperform traditional machine learning methods.

Microbial dynamics prediction

A fundamental question in microbial ecology is whether we can predict the temporal behaviors of complex microbial communities. Traditionally, this problem is addressed using system identification or network reconstruction techniques, which assume specific population dynamics described by a set of ordinary differential equations. For example, the classical GLV model in community ecology, which considers pair-wise interactions, can be represented as a directed, signed, and weighted graph, often referred to as an ecological network. Numerous methods have been developed to infer these dynamics and reconstruct the ecological network using temporal or steady-state data [210]. However, this network-based approach typically assumes that inter-species interactions are exclusively pair-wise, which may not reflect the true nature of complex microbial interactions.

Recently, deep learning techniques have been deployed to predict temporal behaviors of microbiomes. For example, in 2022, Baranwal et al. applied LSTM (a classical variant of RNN) to learn from experimental data on temporal dynamics and functions of microbial communities to predict their future behavior and design new communities with desired functions [228]. Using a significant amount of experimental data, they found that this method outperforms the widely used GLV model in community ecology. In 2023, Thompson et al. proposed the Microbiome

Recurrent Neural Network (MiRNN) architecture. Inputs to the MiRNN at time step t1 include the state of species abundances, metabolite concentrations, control inputs, and a latent vector that stores information from previous steps and whose dimension determines the flexibility of the model.The output from each MiRNN block is the predicted system state and the latent vector at the next time step t. To avoid the physically unrealistic emergence of previously absent species, a constrained feed-forward neural network outputs zero-valued species abundances if species abundances at the previous time step were zero. The authors demonstrated that MiRNN yielded comparable prediction performance to the LSTM model, but with more than a 50,000 fold reduction in the number of model parameters.

These works are of broad interest to those working on microbiome prediction and design to optimize specific target functions. So far, LSTM and MiRNN have been just applied to synthetic communities with 25 diverse and prevalent human gut species and 4 major health-relevant metabolites (acetate, butyrate, lactate, and succinate). Its potential to large systems, e.g., the human gut microbiome, with thousands of species and metabolites would be interesting to explore. The quality of the training data would be crucial.

In addition to methods specifically designed for predicting microbial dynamics, existing methodologies developed for multiple time series forecasting (MTSF) can also be potentially employed. For example, MTSF-DG is a model capable of learning historical relation graphs and predicting future relation graphs to capture dynamic correlations [229]. Evaluating the performance of these general time series prediction methods in the context of microbial dynamics prediction would be very interesting..

Microbiome data simulation and imputation

Often, we need to generate synthetic microbiome data for testing computational methods or imputing missing data points, and there are two primary approaches to achieve this. First, data can be generated from statistical models, such as SparseDOSSA [230], or various population dynamics models using existing software, e.g., miaSim [231]. miaSim is particularly versatile, offering users the ability to simulate data based on specific assumptions and scenarios using four widely recognized population dynamics models: the stochastic logistic model, MacArthur's consumer-resource model, Hubbell's neutral model, and the GLV model, along with several of their derivations. Second, generative deep learning techniques, such as generative adversarial networks (GANs), can be employed to create synthetic data. Recent advancements have introduced several GAN-based methods for generating synthetic microbiome data. For example, MB-GAN [232] learns latent spaces from observed microbial abundances and generates simulated abundances based on these learned distributions. DeepBioGen [233]: This model captures visual patterns of sequencing profiles and generates realistic human gut microbiome profiles. Both MB-GAN and DeepBioGen are designed for data augmentation of single time point microbiome datasets. For longitudinal microbiome data imputation, DeepMicroGen offers a robust solution [234]. This method extracts features that incorporate phylogenetic relationships between taxa using CNN. These features are subsequently processed by a bidirectional RNN-based GAN model, which generates imputed values by learning the temporal dependencies between observations at different time points. These advanced methods enhance our ability to generate high-fidelity synthetic microbiome data, crucial for developing and testing new analytical tools in microbiome research.

Microbial source tracking

Determining the contributions of various environmental sources (“sources”) to a specific microbial community (“sink”) represents a traditional challenge in microbiology, commonly referred to as microbial source tracking (MST). Addressing this MST challenge will not only enhance our understanding of microbial community formation but also has significant implications in areas like pollution management, public health, and forensics. MST techniques are generally categorized into two types: target-based methods, which concentrate on identifying source-specific indicator species or chemicals, and community-based methods, which analyze community structures to assess the similarity between sink samples and potential source environments. With next-generation sequencing becoming standard for community assessment in microbiology, numerous community-based computational methods, known as MST solvers, have been developed and applied to various real-world datasets, showcasing their effectiveness across different scenarios.

Here, we introduce some representative MST solvers. The first solver is based on the classification analysis in machine learning, for example, using the random forest classifier. In this case, each source represents a distinct class, and the classifier will classify the sink into different classes with different probabilities. The probabilities of the sink belonging to the different classes can be naturally interpreted as the mixing proportions or contributions of those sources to the sink. Beyond the simple classification analysis, more advanced statistical methods based on Bayesian modeling have been developed. For example, SourceTracker is a Bayesian MST solver that explicitly models the sink as a convex mixture of sources and infers the mixing proportions via Gibbs sampling [235]. FEAST (fast expectation□maximization for microbial source tracking [236]) is a more recent statistical method. FEAST also assumes each sink is a convex combination of sources. But it infers the model parameters via fast expectation□maximization, which is much more scalable than Markov Chain Monte Carlo used by SourceTracker. STENSL (microbial Source Tracking with ENvironment SeLection) is also based on expectation-maximization [237]. STENSL enhances traditional MST analysis through unsupervised source selection and facilitates the sparse identification of hidden source environments. By integrating sparsity into the estimation of potential source environments, it boosts the accuracy of true source contributions and considerably diminishes the noise from non-contributing sources. ONN4MST is a deep learning method based on the Ontology-aware Neural Network (ONN) to solve large-scale MST problems [238]. The ONN model promotes predictions in line with the “biome ontology.” Essentially, it leverages biome ontology information to represent the relationships among biomes and to estimate the distribution of different biomes within a community sample. The authors demonstrated clear evidence that ONN4MST outperformed other methods (e.g., SourceTracker and FEAST) with near-optimal accuracy when source tracking among 125,823 samples from 114 niches.

Many MST solvers draw inspiration from the analogy between the MST problem and estimating the mixing proportions of conversation topics in a test document. It has been pointed out that this analogy is problematic [239]. In topic modeling [240], a specialized area within NLP, the objective is to uncover the abstract “topics” present in a set of documents, which can be viewed as static or “dead.” In contrast, MST typically involves dynamic, thriving microbial communities where ecological dynamics significantly influence community assembly and their state, that is, the microbial composition. Given these ecological dynamics, a sink community cannot merely be viewed as a convex mixture of known and unknown sources. Indeed, through numerical simulations, analytical calculations, and real data analysis, compelling evidence has been presented that ecological dynamics impose fundamental challenges in community□based MST [239]. Thus, results from current MST solvers require very cautious interpretation.

Metabolic Modeling

Metabolic modeling has become a crucial component in microbiology and microbiome research, significantly enhancing our understanding of microbial interactions and their effects on environments or host well-being. This approach integrates computational methods with biological insights, facilitating the prediction, analysis, and comprehension of metabolic capabilities and interactions within microbial communities.

Gap filling: inferring missing reactions

Genome-scale metabolic models (GEMs) have substantially advanced our understanding of the complex interactions among genes, reactions, and metabolites. These models, integrated with high-throughput data, support applications in metabolic engineering and drug discovery. For instance, AGORA2 (Assembly of Gut Organisms through Reconstruction and Analysis, version 2), representing the cutting-edge GEM resource for human gut microorganisms, comprises 7,302 strains and provides strain-resolved capabilities for drug degradation and biotransformation for 98 drugs [218]. This resource has been meticulously curated using comparative genomics and extensive literature reviews. AGORA2 facilitates personalized, strain-resolved modeling by predicting how patients' gut microbiomes convert drugs. Additionally, AGORA2 acts as a comprehensive knowledge base for the human microbiome, paving the way for personalized and predictive analyses of host–microbiome metabolic interactions. Reconstruction of GEMs typically require extensive manual curation to improve their quality for effective use in biomedical applications. Yet, due to our imperfect knowledge of metabolic processes, even highly curated GEMs could have knowledge gaps (e.g., missing reactions). Various optimization-based gap-filling methods have been developed to identify missing reactions in draft GEMs [241–243].

The existing gap-filling methods often require experimental data, but such experimental data is scarce for non-model organisms, limiting tool utility. If not using any domain knowledge, gap-filling of GEMs or inferring missing reactions in GEMs purely from the topology of the GEM can be treated as a hyperlink prediction problem [244]. As we know, we can always consider a metabolic network or any biochemical reaction network as a hypergraph, where metabolites are nodes, reactions are hyperlinks. For instance, Chen et al. present the Chebyshev spectral hyperlink predictor (CHESHIRE), a deep learning-based method for identifying missing reactions in GEMs based on the topology of metabolic networks [245]. CHISHIRE leverages the Chebyshev spectral GCN on the decomposed graph of a metabolic network to refine the feature vector of each metabolite by incorporating the features of other metabolites from the same reaction. As a variant of GCN, Chebyshev spectral GCN was designed to efficiently process data represented as graphs [246]. It leverages spectral graph theory and Chebyshev polynomials to perform graph convolutions in the spectral domain. It has been shown that CHESHIRE outperforms other topology-based hyperlink rediction methods, e.g., Neural Hyperlink Predictor (NHP) [247] and C3MM Clique Closure-based Coordinated Matrix Minimization (C3MM) [248] in predicting artificially removed reactions over 926 GEMs (including 818 GEMs from AGORA). Furthermore, CHESHIRE is able to improve the phenotypic predictions of 49 draft GEMs for fermentation products and amino acids secretions. Both types of validation suggest that CHESHIRE is a powerful tool for GEM curation..

Retrosynthesis: breaking down a target molecule

Note that gap-filling is the strategy used to complete metabolic networks when certain reactions or pathways are missing. It identifies reactions that need to be added to a metabolic model to ensure the system can produce all required metabolites and metabolic phenotypes. Retrosynthesis is a complementary strategy. Retrosynthesis involves iteratively breaking down a target molecule into simpler molecules that can be combined chemically or enzymatically to produce it. Eventually, all the required compounds are either commercially available or present in the microbial strain of choice. Retrosynthesis is used to map out potential biosynthetic pathways to produce a desired compound by analyzing reaction steps in reverse. While gap-filling aims to ensure the completeness of the metabolic network for overall functionality, retrosynthesis focuses on pathway construction for a specific product. Recently, a reinforcement learning method RetroPath RL was developed for bioretrosynthesis [249]. RetroPath RL is based on the Monte Carlo Tree Search (MCTS), which is a heuristic search algorithm combining the principles of random sampling (Monte Carlo methods) and search trees to balance exploration and exploitation in making optimal decisions [250, 251]. RetroPath RL takes as input a compound of interest, a microbial strain as a sink (i.e., the list of available precursor metabolites) and a set of reaction rules, e.g., RetroRules, a database of reaction rules for metabolic engineering [252].

One interesting application of RetroPath RL is to complete further the metabolism of specific compounds in the human gut microbiota. For instance, Balzerani et al. used RetroPath RL to predict the degradation pathways of phenolic compounds [253]. By leveraging Phenol-Explorer [254], the largest database of phenolic compounds in the literature, and AGREDA [255], an extended metabolic network amenable to analyze the interaction of the human gut microbiota with diet, the authors generated a more complete version of the human gut microbiota metabolic network.

Precision Nutrition

Machine-learning models have shown remarkable accuracy in predicting metabolite profiles from microbial compositions [256–258]. Furthermore, the intersection of computational biology with nutrition science has led to notable strides in personalized nutrition and food quality prediction [259–261]. This emerging field focuses on customizing dietary recommendations to individual biological and physiological profiles, aiming to optimize health outcomes. By employing machine learning algorithms and microbiome data analysis, researchers are able to predict individual responses to various foods and diets, marking a significant advancement in the field of precision nutrition.

Nutrition profile correction

An unhealthy diet is associated with higher risks of various diseases [262, 263]. Measuring dietary intake in large cohort studies is often difficult, so we frequently depend on self-reported tools (like food frequency questionnaires, 24-hour recalls, and diet records) that are established in nutritional epidemiology [264–266]. However, these self-reported instruments can be susceptible to measurement errors [267], resulting in inaccuracies in nutrient profile calculations. Although nutritional epidemiology uses methods such as regression calibrations [268, 269] and cumulative averages [270] to address these inaccuracies, deep-learning approaches have not been leveraged to correct random measurement errors.

Wang et al. introduce a deep-learning method called METRIC (Microbiome-based Nutrient Profile Corrector) that utilizes gut microbial compositions to correct random measurement errors in nutrient profiles derived from self-reported dietary assessments [271]. METRIC draws inspiration from Noise2Noise, a deep learning model for image denoising in computer vision that reconstructs clean images using only corrupted inputs [272]. The core concept of Noise2Noise is training the model on pairs of noisy images as both the input and output, compelling the neural network to estimate the average of these corrupted images. This process leads the predictions to statistically align with the clean image due to the zero-mean property of the noise. In a similar way, METRIC addresses random errors in the assessed nutrient profile generated from self-reported dietary assessments, without using clean data (i.e., the ground truth dietary intake). It's important to note that METRIC targets the correction of the nutrient profile rather than the food profile (or the original dietary assessment), since the high frequency of zero values in the food profile—many food items not consumed—poses significant challenges for machine learning. In contrast, the derived nutrient profile tends to contain predominantly non-zero values. Additionally, METRIC aims to rectify random errors characterized by zero means, instead of systematic biases or errors with non-zero means, as correcting the latter effectively necessitates access to the ground truth dietary intake, which is often unavailable.

Metabolomic profile prediction

Predicting the metabolomic profile (i.e., quantified amount of metabolites within a biological sample) from the composition of a microbial community is an active area in microbiome research. Experimental measurement of metabolites relies on expensive and complex techniques like Liquid Chromatography-Mass Spectrometry, which have incomplete coverage [273, 274]. In contrast, microbial composition measurements are cheaper, more automated, and have better coverage. Therefore, it is desirable to develop computational methods that predict metabolomic profiles based on microbial compositions [257, 258, 275]. Additionally, such a method could facilitate our understanding of the interplay between microorganisms and their metabolites.

Various machine-learning methods have been developed to solve this problem. For example, MelonnPan uses an elastic net linear regression to model the relative abundance of each metabolite using metagenomic features [275]. It simply models each metabolite individually, missing the opportunity to use shared information across metabolomic features to boost prediction performance. Neural encoder-decoder (NED) leverages the constraints of sparsity and non-negative weights for mapping microbiomes to metabolomes [276]. The use of non-negative weights in NED imposes a stringent constraint on the model, which simplifies the model complexity but may limit the learning capacity. MiMeNet (Microbiome-Metabolome Network) is essentially an MLP that models the community metabolome profile using metagenomic taxonomic or functional features obtained from a microbiome sample [257].

Leveraging the state-of-the-art deep-learning method, neural ordinary differential equations (NODE) [219], Wang et al. developed mNODE (metabolomic profile predictor using neural ordinary differential equations) [258]. Since the input dimension (the number of microbial species) is different from the output data (the number of microbial species), mNODE integrates the NODE as a middle module, sandwiched by two densely connected layers to adjust for data dimension variability. The method shows superior performance in both synthetic and real datasets than existing methods. Additionally, mNODE can naturally incorporate dietary information into its analysis of human gut microbiomes, improving metabolomic profile predictions. Its susceptibility analysis uncovers microbe–metabolite interactions, which can be confirmed with both synthetic and real datasets. Overall, these findings highlight mNODE's effectiveness in exploring the microbiome–diet–metabolome connection and advancing research in precision nutrition.

Personalized diet recommendation

In recent years, the intersection of gut microbiome, nutrition science, and machine learning has led to significant advancements in personalized nutrition and food quality prediction. This emerging field aims to tailor dietary recommendations to individual biological and physiological factors (e.g., gut microbial composition), thereby optimizing health outcomes [259–261, 277].

Numerous studies use traditional machine learning methods to predict blood glucose levels based on the time-series data from continuous glucose monitor [278, 279]. Similarly, Kim et al. apply RNN to predict blood glucose levels in hospitalized patients with type-2 diabetes [280]. Recently, Lutsker et al. present GluFormer, a generative foundation model based on the Transformer architecture to predict blood glucose measurements from non-diabetic individuals [281]. However, these models do not incorporate dietary information in their inputs, limiting their ability to generate personalized dietary recommendations. In contrast, leveraging mathematical models and Bayesian statistics, Albers et al. predict an individual's postprandial blood glucose level using the preprandial blood glucose level and carbohydrate intake [282].

Zeevi et al. use the gradient-boosting regressor (GBR) to predict personalized postprandial blood glucose responses (PPGRs) to meals based on individual factors, including dietary habits, physical activity, blood parameters, anthropometric data, and gut microbiome composition [259]. After being trained on a cohort with 800 participants, GBR is validated using an independent cohort, achieving a Pearson correlation coefficient between predicted and measured PPGRs R = 0.70. A similar machine learning method has been used for other cohorts, such as Food & You [277].

Rein et al. extend this personalized approach to a clinical setting, focusing on a randomized dietary intervention pilot trial of 23 individuals with type 2 diabetes mellitus [260]. Based on the well-trained GBR, a personalized postprandial targeting diet is designed for each individual to minimize the individual's PPGR. Using a leave-one-out approach, the well-trained GBR assigns rankings to each participant's meals during the profiling week, where 4–6 distinct isocaloric options represent each meal type.

Neumann et al. predict the future blood glucose levels in type-1 diabetes patients during and after various types of physical activities in real-world conditions [283]. The study employs several machine learning and deep learning regression models, including XGBoost, Random Forest, LSTM, CNN-LSTM, and Dual-encoder models with an attention layer. The models use multiple data types, including continuous glucose monitoring data, insulin pump data, carbohydrate intake, exercise details (like intensity and duration), and physical activity-related information (e.g., number of steps and heart rate). The output is the predicted blood glucose level at future times, specifically at 10, 20, and 30 minutes after the inputs are recorded. Among many employed models, LSTM is the best-performing model for most patients.

Although several machine-learning methods have been proposed to predict short-term postprandial responses of only a few metabolite biomarkers, less is explored over the important long-term responses of a wider range of health-related metabolites following dietary interventions. Wang et al. introduced a deep learning model, McMLP (Metabolic response predictor using coupled Multilayer Perceptrons), to fill this gap. McMLP consists of two coupled MLPs [261]. The first MLP forecasts endpoint (i.e., after dietary interventions) microbial compositions from baseline (i.e., before dietary interventions) microbial and metabolomic profiles, and dietary intervention strategy. The second MLP uses these predicted endpoint microbial compositions, baseline metabolomic profiles as well as dietary intervention strategies to forecast endpoint metabolomic profiles. When McMLP is benchmarked with existing methods on synthetic data and six real data, it consistently yields a much better performance of predicting metabolic response than previous methods like random forest and GBR.

Despite significant advancements in metabolic modeling and the integration of machine learning techniques for predicting metabolomic profiles, several open questions remain that could drive future research. One such question is to explore whether integrating multi-omics data (combining metagenomic, transcriptomic, and proteomic data) could further refine these predictions. Additionally, reinforcement learning could potentially be leveraged to generate better personalized dietary recommendations.

Clinical Microbiology

The earliest applications of AI in microbiology can be traced back to the 1970s when MYCIN was developed at Stanford University. MYCIN was an expert system designed to diagnose bacterial infections and recommend appropriate antibiotics. It used a rule-based approach, drawing on a knowledge base of expert-encoded rules to make decisions about infectious diseases, particularly blood infections. MYCIN was notable for demonstrating that AI could assist in clinical decision-making, setting the stage for later developments in AI for microbiology and medicine. AI pioneer Allen Newell referred to MYCIN as “the granddaddy of expert systems”, stating it was “the one that launched the field.” Nowadays, various AI techniques have been applied in clinical microbiology. Here we briefly discuss those applications.

Microorganism detection, identification and quantification

AI techniques, especially supervised machine learning algorithms, are widely used to detect, identify, or quantify microorganisms using various types of data from cultured bacteria [14]. Here we briefly discuss how AI techniques are applied across three different data types. (1) Microscopic Images: Deep learning models, particularly CNNs, have been highly effective in analyzing microscopic images of bacterial colonies [284, 285]. By training on labeled images, these models can classify bacterial species based on their shapes, sizes, arrangements, and staining characteristics (e.g., Gram staining). This approach aids in automating bacterial identification in clinical labs and research, improving the speed and accuracy of microbial diagnostics. (2) Spectroscopy Data: Supervised machine learning algorithms are also employed to analyze spectroscopy data, such as mass spectrometry or Raman spectroscopy, to identify microorganisms [286, 287]. For instance, MALDI-TOF (Matrix-Assisted Laser Desorption/Ionization Time-of-Flight) mass spectrometry generates unique protein "fingerprints" for bacterial species. Machine learning models trained on these spectra can quickly and accurately classify species based on their spectral profiles. Raman spectroscopy, which provides molecular fingerprints of samples, also benefits from machine learning algorithms to classify bacterial species or detect specific metabolic or pathogenic profiles. (3) Volatile Organic Compounds (VOCs): Many bacteria emit VOCs as metabolic byproducts, which can serve as unique biomarkers for microbial identification [288]. Gas chromatography-mass spectrometry (GC-MS) or electronic noses (e-noses) are often used to capture these VOCs. Machine learning models trained on VOC patterns can distinguish bacterial species based on their unique VOC profiles. This approach has potential in medical diagnostics, food safety, and environmental monitoring.

Machine learning algorithms in these applications often require substantial labeled data for training, so accurate labeling and quality data collection are crucial. As these models learn to detect subtle differences in physical, chemical, and visual features, they contribute significantly to rapid, non-invasive, and automated bacterial identification, offering promising alternatives to traditional microbiological techniques.

Antimicrobial susceptibility evaluation

The evaluation of antimicrobial susceptibility has evolved significantly, especially with advancements in genomics and AI. Early approaches focused on using well-known antibiotic resistance genes to predict phenotypic susceptibility, achieving good accuracy for pathogens like Staphylococcus aureus, Escherichia coli, and Klebsiella pneumoniae. However, challenges arose with pathogens such as Pseudomonas aeruginosa, where resistance is driven by gene expression changes, leading to less reliable phenotype predictions. AI has emerged as a promising tool to address these limitations, especially when mutational knowledge is incomplete. Combining machine learning with gene expression data has improved predictive accuracy, as seen in recent studies on P. aeruginosa, achieving over 90% accuracy for resistance to meropenem and tobramycin [289]. Nonetheless, predictions for other antibiotics, such as ceftazidime, remain suboptimal. Combining phenotypic and genotypic data has further enhanced accuracy in rapid diagnostics, as demonstrated by Bhattacharyya et al., who achieved 94-99% accuracy in predicting susceptibility profiles for several bacterial species within hours [290]. The use of whole-genome sequencing (WGS) data in machine learning systems has been extended to predict minimal inhibitory concentrations (MICs) and antibiotic susceptibility, with mixed results. For example, prediction accuracy for ciprofloxacin MICs in *E. coli* remained relatively low compared to other antibiotics [291]. Similar machine learning approaches have been employed for Mycobacterium tuberculosis [292], viral evolution studies [293], and understanding viral resistance [294], showcasing AI's broad applicability. We emphasize that while AI techniques show great promise in improving antimicrobial susceptibility testing, challenges remain, particularly in achieving consistent accuracy across different pathogens and antibiotic classes.

Disease diagnosis, classification, and clinical outcome prediction

AI can assist in examining novel and intricate data that clinical environments have not fully utilized for diagnostic aims. For instance, for certain diseases involving infections, microbes can generate some VOCs in clinical samples. Hence, we can utilize machine learning to evaluate the odors of those clinical samples to diagnose urinary tract infections [295], active tuberculosis [296], pneumonia [297], and acute exacerbation of chronic obstructive pulmonary disease [298]. For many other diseases associated with disrupted microbiomes, VOCs in clinical samples might not be helpful for disease diagnosis. In this case, we can leverage the microbiome data itself. Indeed, numerous studies have shown microbiome dysbiosis is associated with human diseases [299, 300]. Those diseases include GI disorders, i.e., *Clostridioides difficile* infection [301], inflammatory bowel disease [302], and irritable bowel syndrome [303], and other non-GI disorders, for example, autism [304], obesity [305], multiple sclerosis [306], hepatic encephalopathy [307], and Parkinson's disease [308]. Applying supervised classification analysis to the human microbiome data can help us build classifiers that can accurately classify individuals' disease status, which could assist physicians in designing treatment plans [18].

Classical machine learning classifiers. Classical ML methods (e.g., Random Forest, XGBoost, Elastic Net, and SVM) have been systematically compared in the classification analysis of human microbiome data [309]. It was found that, overall, the XGBoost, Random Forest, and Elastic Net display comparable performance [309]. In case the training data contains microbiome data (features) collected before the disease diagnosis (labels), the well-trained classifiers can act as predictors, which have an even bigger clinical impact in terms of early diagnosis. For example, predicting asthma development at year three from the microbiome and other omics and clinical data collected at and before year one [310].

Phylogenetic tree-based deep learning methods. Classical ML classifiers just treat microbiome data (more specifically, the taxonomic profiles) as regular tabular data, represented as a matrix with rows representing different samples or subjects and columns representing features (i.e., microbial species' relative abundances). In fact, unlike many other omics, microbial features are endowed with a hierarchical structure provided by the phylogenetic tree defining the evolutionary relationships between those microorganisms. We can exploit the phylogenetic structure and leverage the CNN architecture to deal with species abundance data. With this very simple idea, several deep learning methods (e.g., Ph-CNN [311], PopPhy-CNN [312], taxoNN [313], and MDeep [314]) have been developed. Each method exploits the phylogenetic tree in a slightly different way.

Ph-CNN takes the taxa abundances table and the taxa distance matrix as the input, and outputs the class of each sample [311]. Here, the distance between two taxa is defined as their patristic distance, i.e., the sum of the lengths of all branches connecting the two taxa on the phylogenetic tree. The patristic distance is used together with multi-dimensional scaling to embed the phylogenetic tree in an Euclidean space. Each taxon is represented as a point in Euclidean space preserving the tree distance as well as possible. Since the data is endowed with an intrinsic concept of neighborhood in the input space, Ph-CNN can then use CNN to perform classification. PopPhy-CNN represents the phylogenetic tree and species abundances in a matrix format, and then directly applies CNN to perform classification [312]. taxoNN incorporates a stratified approach to group OTUs into phylum clusters and then applies CNNs to train within each cluster individually [313]. Further, through an ensemble learning approach, features obtained from each cluster were concatenated to improve prediction accuracy. Note that with each phylum cluster, the authors proposed two ways (either based on distance to the cluster center or based on taxa correlations) to order and place correlated taxa together to generate matrix or image-like inputs amenable for CNN. MDeep directly incorporates the taxonomic levels of the phylogenetic tree into the CNN architecture [314]. OTUs on the species level are clustered based on the evolutionary model. This clustering step makes convolutional operation capture OTUs highly correlated in the phylogenetic tree. The number of hidden nodes decreases as the convolutional layer moves forward, reflecting the taxonomic grouping.

Other deep learning methods. Besides the above deep learning methods that exploit the phylogenetic structure for microbiome data classification, some other deep learning methods (e.g., DeepMicro [112], GDmicro [315], and a transformer-based microbial "language" model [316]) have been developed. Those methods do not leverage the phylogenetic structure of microbiome data.

DeepMicro incorporated various autoencoders (including SAE, DAE, VAE, and CAE) to learn a low-dimensional embedding for the input microbial compositional feature, and then employed MLP to classify disease status with the learned latent features [317]. GDmicro is a GCN-based method for microbiome feature learning and disease classification [315]. GDmicro formulates the disease classification problem as a semi-supervised learning task, which uses both labeled and unlabeled data for feature learning ([318]). To overcome the domain discrepancy problem (i.e., data from different studies have many differences due to confounding factors, such as region, ethnicity, and diet, which all shape the gut microbiome), GDmicro applies a deep adaptation network [319] to learn transferable latent features from the microbial compositional matrix across different domains/studies with or without disease status labels. Then, GDmicro constructs a similarity graph, where each node represents a host whose label can be either healthy, diseased, or unlabeled, and edges represent the similarity between two hosts' learned latent features. GDmicro then employs GCN to take this microbiome similarity graph as input and incorporate both the structural and node abundance features for disease status classification. Note that this is a very classical application of GCN to solve the semi-supervised node classification problem on graphs, where some nodes have no labels.

Recently, a transformer-based microbial "language" model (MLM) was developed [316]. This MLM was trained in a self-supervised fashion to capture the interactions among different microbial species and the common compositional patterns in microbial communities. The trained MLM can generate robust, context-sensitive representations of microbiome samples to enhance predictive modeling. Note that in this transformer-based MLM, taxa present in each microbiome sample were ranked in decreasing order of abundance to create an ordered list of taxa so that the inputs are analogous to texts. The transformer model then processes these inputs through multiple encoder layers, producing a hidden representation for each taxon. The output of the model includes both sample-level embeddings for classification tasks and context-sensitive embeddings for individual taxon, enabling a nuanced understanding of microbial interactions. By pre-training the transformer using self-supervised learning on large, unlabeled datasets and fine-tuning on specific labeled tasks, this approach leads to improved performance for multiple prediction tasks including predicting IBD and diet patterns.

Note that those three methods (DeepMicro, GDmicro, and the transformer-based MLM) can be applied to any omics data for classification purposes. Their design principles were not based on any unique features of microbiome data.

Despite the development of various methods, a systematical comparison of those deep learning methods and classical machine learning methods on benchmarking datasets is lacking. Since some of those deep learning methods incorporate domain knowledge (i.e., information on the phylogenetic tree, or unlabeled samples), it would be necessary to do that for classical ML methods too, for a fair comparison.

Integration of various feature types. Note that 16S rRNA gene sequencing can only provide taxonomic profiles (in terms of microbial compositions) and cannot directly profile microbial genes/functions. Shotgun metagenome sequencing can provide comprehensive data on both taxonomic and functional profiles. It is quite natural to investigate if combining both taxonomic and functional features will enhance classification performance. MDL4Microbiome is such a deep learning method. It employs MLP and combines three different feature types, i.e., taxonomic profiles, genome-level relative abundance, and metabolic functional characteristics, to enhance classification accuracy [320].

Quite often, we have multi-omics data and clinical data. It would be more insightful to integrate those different data types for better disease status classification or prediction. A straight approach would be to concatenate all datasets into a single view, which is then used as the input to a supervised learning model of choice. A more advanced approach is MOGONET, which jointly explores omics-specific learning using GCNs and cross-omics correlation learning for effective multi-omics data classification [321].

Recently, in a childhood asthma prediction project, 18 methods were evaluated using standard performance metrics for each of the 63 omics combinations of six omics data (including GWAS, miRNA, mRNA, microbiome, metabolome, DNA methylation) collected in The Vitamin D Antenatal Asthma Reduction Trial cohort [310]. It turns out that, surprisingly, Logistic Regression, MLP, and MOGONET display superior performance than other methods. Overall, the combination of transcriptional, genomic, and microbiome data achieves the best prediction for childhood asthma prediction. In addition, including the clinical data (such as the father and mother's asthma status, race, as well as vitamin D level in the prediction model) can further improve the prediction performance for some but not all the omics combinations. Results from this study imply that deep learning classifiers do not always outperform traditional classifiers.

So far, the integration of various data types discussed above is often referred to as early fusion. It begins by transforming all datasets into a single representation, which is then used as the input to a supervised learning model of choice. There is another approach called late fusion, which works by developing first-level models from individual data types and then combining the predictions by training a second-level model as the final predictor. Recently, encompassing early and late fusions, cooperative learning combines the usual squared error loss of predictions with an agreement penalty term to encourage the predictions from different data views to align [322]. It would be interesting to explore this idea of cooperative learning in disease classification using multi-omics data [323, 324] (including microbiome data).

Prevention & Therapeutics

Peptides identification & generation

Bacterial resistance to antibiotics is a growing concern. Antimicrobial peptides (AMPs), natural components of innate immunity, are popular targets for developing new drugs. We can divide the AMP activities into different categories, e.g., antibacterial, antiviral, antifungal, antiparasitic, anti-tumor peptides, etc. [325]. Deep learning methods are now commonly adopted by wet-laboratory researchers to screen for promising AMPs. The first work that used neural networks to identify AMPs dates back to 2007, where Lata et al. used a very simple MLP with only one hidden layer [326]. In this work, the authors predicted AMPs based on their N-terminal residues or C-terminal residues, because it has been observed that certain types of residues are preferred at the N-terminal (or C-terminal) regions of the AMPs. In another work published in 2010, Torrent et al. still used a simple MLP with one hidden layer to identify AMPs [327]. In this work, they used the physicochemical properties of AMPs as their features. In total, the authors chosen eight features, including isoelectric point (pI), peptide length, a-helix, b-sheet and turn structure propensity, in vivo and in vitro aggregation propensity and hydrophobicity.

Those early works apparently require quite a lot of domain knowledge and manual feature selection. This effort can be avoided or mitigated by using deep learning models that can automatically learn complex representations and features from raw data, reducing the need for manual feature engineering. For example, in 2018 Veltri et al. proposed a deep neural network model with convolutional and recurrent layers that leverage primary sequence composition [328]. Apparently, it is a hybrid deep learning model. By combining CNN and RNN, the model can extract more meaningful and robust features, avoid the burden of a priori feature construction, and consequently reduce our reliance on domain experts. In 2022, Tang et al. proposed a similar hybrid deep learning model that integrated CNN and RNN [329]. This model is called MLBP: multi-label deep learning approach for determining the multi-functionalities of bioactive peptides. It can predict multi-function, e.g., anti-cancer peptides, anti-diabetic peptides, anti-hypertensive peptides, anti-inflammatory peptides, and anti-microbial peptides, simultaneously. Firstly, the amino acids were converted into natural numbers, and the sequences of all peptides were set to be fixed by using the zero-filled method. Then, an embedding layer was used to learn the embedding matrix of the representation of peptide sequences. The embedding matrix was fed into a CNN to extract the features from the peptide. Then, an RNN is used to analyze streams of the sequence by means of hidden units. Finally, a fully connected layer is applied to the final classification.

The hybrid deep learning approach has been extended further in Ref [330]. The authors started by collecting sequences to build training and test sets and then built and optimized deep learning models to form the AMP prediction pipeline. In particular, the authors included five deep learning models for testing and building the prediction pipeline, including (1) Two CNN + LSTM models; (2) Two CNN + Attention models; and (3) One BERT model. Because the prediction biases were independent of each other, the authors eventually tested the intersection of predictions from various combinations of models (2–5 models). This is a very robust approach. Then they mined metagenomic and metaproteomic data of the human gut microbiome for potential AMPs, further filtering using correlation network analysis between candidate AMPs and bacteria. Finally, they selected promising candidates AMPs from initial screening and further subjected them to efficacy tests against multi-drug resistant (MDR) bacteria, and then in vivo experiments in an animal model. This is a very comprehensive work, clearly demonstrating the power of deep learning models in the identification of AMPs from microbiome data.

Besides identifying natural AMPs, deep learning approaches have also been developed to generate synthetic AMPs. These approaches include GAN and VAE, as well as their conditional variants cGAN and cVAE. The conditional variants enable the generation of peptides satisfying a given condition. For example, AMPGANv2 is based on a bidirectional conditional GAN [331]. It uses generator-discriminator dynamics to learn data-driven priors and control generation using conditioning variables [331]. The bidirectional component, implemented using a learned encoder to map data samples into the latent space of the generator, aids iterative manipulation of candidate peptides. These elements allow AMPGANv2 to generate candidates that are novel, diverse, and tailored for specific applications. Training of GANs was reported to face substantial technical obstacles, such as training instabilities and mode collapse. Hence, VAE-based AMP generations could be an alternative solution. For example, Peptide VAE is based on a VAE, where both encoder and decoder are single-layer LSTMs [332]. The authors also proposed Conditional Latent (attribute) Space Sampling (CLaSS) for controlled sequence generation, aimed at controlling a set of binary (yes/no) attributes of interest, such as antimicrobial function and/or toxicity. HydrAMP is based on a conditional VAE to generate novel peptide sequences satisfying given antimicrobial activity conditions [333]. This method is suitable not only for the generation of AMPs de novo, but also for the generation starting off from a prototype sequence (either known AMPs or non-AMPs).

Probiotic mining

The discovery and experimental validation of probiotics demand significant time and effort. Developing efficient screening methods for identifying probiotics is therefore of great importance. Recent advances in sequencing technology have produced vast amounts of genomic data, allowing us to design machine learning-based computational approaches for probiotic mining. For example, Sun et al. developed iProbiotics, which utilizes k-mer frequencies to characterize complete bacterial genomes and employs the support vector machine for probiotic identification [334]. iProbiotics conducted a k-mer compositional analysis (with k ranging from 2 to 8) on a comprehensive probiotic genome dataset, which was built using the PROBIO database and literature reviews. This analysis revealed significant diversity in oligonucleotide composition among strain genomes, showing that probiotic genomes exhibit more probiotic-related features compared to non-probiotic genomes. A total of 87,376 k-mers were further refined using an incremental feature selection method, with iProbiotics achieving peak accuracy using 184 core features. This study demonstrated that the probiotic role is not determined by a single gene but rather by a composition of k-mer genomic elements.

Although iProbiotics has been validated using complete bacterial genomes, its effectiveness on draft genomes derived from metagenomes remains uncertain. Additionally, while the k-mer frequency model has been applied in various bioinformatics tasks, it primarily captures the occurrence frequencies of oligonucleotides and may not fully represent sequence function. Recent advancements in NLP have introduced novel methods for representing biological sequences. In these models, oligonucleotides or oligo-amino acids are treated as ’words,’ and DNA or protein sequences as ’sentences.’ By using unsupervised pretraining on large datasets, each word is mapped to a context-based feature vector, potentially offering more informative representations than k-mer frequencies. Building on this concept, Wu et al. developed metaProbiotics, a method designed to mine probiotics from metagenomic binning data [335]. It represents DNA sequences in metagenomic bins using word vectors and employs random forests to identify probiotics from the metagenomic binned data.

Technically speaking, both iProbiotics and metaProbiotics are not based on deep learning techniques. In particular, the classification analysis still relies on traditional machine learning methods, e.g., SVM and RF. We expect that soon more deep learning-based methods will be developed to solve this very important task.

Antibiotic discovery

Compared with probiotic discovery, deep learning has been extensively used in antibiotic discovery. This thanks to the success of GCNs, which have been repeatedly shown to have robust capacities for modeling graph data such as small molecules. In particular, message-passing neural networks (or MPNNs) are a group of GCN variants that can learn and aggregate local information of molecules through iterative message-passing iterations [336]. MPNNs have exhibited advancements in molecular modeling and property prediction.

The original MPNN operates on undirected graphs. It is trivial to extend MPNN to directed multigraphs. This yields Directed MPNN, which translates the graph representation of a molecule into a continuous vector via a directed bond-based message passing approach [337]. This builds a molecular representation by iteratively aggregating the features of individual atoms and bonds. The model operates by passing "messages" along bonds that encode information about neighboring atoms and bonds. By applying this message passing operation multiple times, the model constructs higher-level bond messages that contain information about larger chemical substructures. The highest-level bond messages are then combined into a single continuous vector representing the entire molecule.

Stokes et al. discovered a drug halicin by drug repurposing using deep neural networks Chemprop [338, 339] to predict molecules with antibacterial activity. Halicin can against a wide phylogenetic spectrum of pathogens, including Mycobacterium tuberculosis, carbapenem-resistant *Enterobacteriaceae*, and *Clostridioides difficile* and pan-resistant *Acinetobacter baumannii* infections in Murine models [340]. The first module of Chemprop is a local feature encoding function. A molecule's molecular SMILES string (simplified molecular-input line-entry system) is used as input and transformed into a molecular graph with nodes representing atoms and edges representing bonds using RDKit [341]. The molecular embedding was learned by GCN and was fed into a feed-forward neural networks for classification.

Jame Collins' lab at MIT recently published two papers on antibiotic discovery [340, 342]. In both papers, they utilized a Direc-MPNN. In principle, their results can be further improved by incorporating a new variant of MPNN, i.e., atom-bond transformer-based MPNN (or ABT-MPNN), which combines the self-attention mechanism in Transformer with MPNNs for better molecular representation and better molecular property predictions. By designing corresponding attention mechanisms in the message-passing and readout phases of the MPNN, ABT-MPNN provides a novel architecture that integrates molecular representations at the bond, atom and molecule levels in an end-to-end way. This model also has a visualization modality of attention at the atomic level, which could be an insightful way to investigate molecular atoms or functional groups associated with desired biological properties, and hence serve as a valuable way to investigate the mechanism of action of drugs (including, but limited to antibiotics).

Phage therapy

As the most abundant organisms in the biosphere, bacteriophages (a.k.a. phages) are viruses that specifically target bacteria and archaea. They play a significant role in microbial ecology by influencing bacterial populations, gene transfer, and nutrient cycles. Moreover, they can be an alternative to antibiotics and hold the potential therapeutic ability for bacterial infections [343–346].

Phage identification. Many computational tools have been developed to identify bacteriophage sequences in metagenomic datasets [347]. They can be roughly grouped into two classes: (1) alignment-based (or database-based) methods, e.g., MetaPhinder [348], VIBRANT [349], and VirSorter2 [350]; (2) alignment-free (or learning-based) methods, e.g., VirFinder [351], PPR-meta [154], Seeker [352], DeepVirFinder [353], and PhaMer [354]. Alignment-based methods typically use a large number of sequences of references and utilize DNA or protein sequence similarity as the main feature to distinguish phages from other sequences. Their limitations are evident. Firstly, bacterial contigs may align with multiple phage genomes, potentially resulting in false-positive phage predictions. Secondly, novel or highly diverged phages may not have significant alignments with the selected phage protein families, which can lead to low sensitivity in identifying new phages. Alignment-free methods can overcome those limitations via machine learning or deep learning techniques. Those methods learn the features of the sequence data and are mainly classification models with training data consisting of both phages and bacteria. Some classification models use manually extracted sequence features such as k-mers, while others use deep learning techniques to automatically learn features. For example, VirFinder uses k-mers to train a logistic regression model for phage identification. Seeker (or DeepVirFinder) uses one-hot encoding to represent the sequence data and trains an LSTM (or CNN) to identify phages, respectively. PhaMer leverages the start-of-the-art language model, the Transformer, to conduct contextual embedding for phage contigs. It feeds both the protein composition and protein positions from each contig into the Transformer, which learns the protein organization and associations to predict the label for test contigs. It has been shown that PhaMer outperforms VirSorter, Seeker, VirFinder, DeepVirfinder, and PPR-meta.

Recently, a hybrid method called INHERIT was developed. INHERIT (IdentificatioN of bacteriopHagEs using deep RepresentatIon model with pre-Training) naturally ‘inherits’ the characteristics from both alignment-based and alignment-free methods [355]. In particular, INHERIT uses pre-training as an alternative way of acquiring knowledge representations from existing databases, and then uses a BERT-style deep learning framework to retain the advantage of alignment-free methods. The independent pre-training strategy can effectively deal with the data imbalance issue of bacteria and phages, helping the deep learning framework make more accurate predictions for both bacteria and phages. The deep learning framework in INHERIT is based on a novel DNA sequence language model: DNABERT [60], a pre-trained bidirectional encoder representation model, which can capture global and transferrable understanding of genomic DNA sequences based on up and downstream nucleotide contexts. It has been demonstrated that INHERIT outperforms four existing state-of-the-art approaches: VIBRANT, VirSorter2, Seeker, and DeepVirFinder. It would be interesting to compare the performance of

INHERIT and PhaMer.

Phage lifestyle prediction. Besides phage identification, machine learning techniques can also be used to predict the phage lifestyle (virulent or temperate), which is crucial to enhance our understanding of the phage-host interactions. For example, PHACTS used an RF classifier on protein similarities to classify phage lifestyles [356]. BACPHLIP also used an RF classifier on a set of lysogeny-associated protein domains to classify phage lifestyles [357]. Those two methods do not work well for metagenomic data. By contrast, DeePhage can directly classify the lifestyle for contigs assembled from metagenomic data [358]. DeePhage uses one-hot encoding to represent DNA sequences and trains a CNN to obtain valuable local features. PhaTYP further improved the accuracy of phage lifestyle prediction on short contigs by adopting BERT to learn the protein composition and associations from phage genomes [359]. In particular, PhaTYP solved two tasks: a self-supervised learning task and a fine-tuning task. In the first task, PhaTYP applies self-supervised learning to pre-train BERT to learn protein association features from all the phage genomes, regardless of the available lifestyle annotations. In the second task, PhaTYP fine-tunes BERT on phages with known lifestyle annotations for classification. It has been shown that PhaTYP outperforms DeePhage and three other machine learning methods PHACTS (based on RF), BACPHLIP (based on RF), and PhagePred (based on Markov model). DeePhafier is another deep learning method for phage lifestyle classification [360]. Based on a multilayer self-attention neural network combining protein information, DeePhafier directly extracts high-level features from a sequence by combining global self-attention and local attention and combines the protein features from genes to improve the performance of phage lifestyle classification. It has been shown that DeePhafier outperforms DeePhage and PhagePred. It would be interesting to compare the performance of DeePhafier and PhaTYP.

Phage-host interaction prediction. Phages can specifically recognize and kill bacteria, which leads to important applications in many fields. Screening suitable therapeutic phages that are capable of infecting pathogens from massive databases has been a principal step in phage therapy design. Experimental methods to identify phage-host interactions (PHIs) are time-consuming and expensive; using high-throughput computational methods to predict PHIs is therefore a potential substitute. There are two types of computational methods for PHI prediction. One is alignment-based. We explicitly align the viral and bacterial whole-genome sequences and acquire matched sequences to indicate PHI. The other is alignment-free. We compare nucleotide features and/or protein features extracted from viral and bacterial genomes, and predict PHI using machine learning. Each type of method has its pros and cons. A benchmark study ([361]) of those alignment-free machine learning methods demonstrated that GSPHI [362] and PHIAF [363] are the two best deep learning-based methods for PHI prediction. PHIAF is a deep learning method based on date augmentation, feature fusion, and the attention mechanism. It first applies a GAN-based data augmentation module, which generates pseudo-PHIs to alleviate the data scarcity issue. Then it fuses the features originating from DNA and protein sequences for better performance. Finally, it incorporates an attention mechanism into CNN to consider different contributions of DNA/protein sequence-derived features, which provides interpretability of the predictions. GSPHI is a novel deep learning method for PHI prediction with complementing multiple information. It first initializes the node representations of phages and target bacterial hosts via a word embedding algorithm (word2vec). Then it uses a graph embedding algorithm (structural deep network embedding: SDNE) to extract local and global information from the interaction network. Finally, it uses a multi-layer perceptron (MLP) with two hidden layers to detect PHIs.

Recently, a deep learning-based method SpikeHunter was developed to perform a large-scale characterization of phage receptor-binding proteins (i.e., tailspike proteins), which are essential for determining the host range of phages [364]. SpikeHunter uses the ESM-2 protein language model [365] to embed a protein sequence into a representative vector. Then it predicts the probability of that protein being a tailspike protein using a fully connected 3-layer neural network. A reference set of 1,912 tailspike protein sequences and 200,732 non-tailspike protein sequences was curated from the INPHARED database [366]. SpikeHunter identified 231,965 diverse tailspike proteins encoded by phages across 787,566 bacterial genomes from five virulent, antibiotic-resistant pathogens. Remarkably, 86.60% (143,200) of these proteins demonstrated strong correlations with specific bacterial polysaccharides. The authors found that phages with identical tailspike proteins can infect various bacterial species that possess similar polysaccharide receptors, highlighting the essential role of tailspike proteins in determining host range. This work significantly enhances the understanding of phage specificity determinants at the strain level and provides a useful framework for guiding phage selection in therapeutic applications.

Phage virion protein annotation. Phage virion proteins (PVPs) determine many biological properties of phages. In particular, they are effective at recognizing and binding to their host cell receptors without having deleterious effects on human or animal cells [367]. Due to the very time-consuming and labor-intensive nature of experimental methods, PVP annotation remains a big challenge, which affects various areas of viral research, including viral phylogenetic analysis, viral host identification, and antibacterial drug development. Various ML methods have been developed to solve the PVP annotation problem [367]. Those methods can be roughly classified into three groups: (1) traditional machine learning-based methods (using NB: naive bayes, RF: random forest, SCM: scoring card matrix, or SVM: support vector machine); (2) ensemble-based methods (using multiple machine learning models or training datasets), and (3) deep learning-based methods. Representative deep learning-based PVP classification methods are PhANNs [368], VirionFinder [369], DeePVP [370], PhaVIP [371], ESM-PVP [372], and a PLM-based classifier [373]. PhANNs used k-mer frequency encoding and 12 MLPs as the classifiers. Both VirionFinder and DeePVP used CNN as classifiers. In VirionFinder, each protein sequence is represented by a "one-hot" matrix and a biochemical property matrix, while DeePVP only used one-hot encoding to characterize the protein sequence. PhaVIP adapted a novel image classifier, Vision Transformer (ViT) [374, 375], to conduct PVP classification. In particular, PhaVIP employed the chaos game representation (CGR) to encode k-mer frequency of protein sequence into images, and then leveraged ViT to learn both local and global features from sequence "images". The self-attention mechanism in ViT helps PhaVIP learn the importance of different subimages and their associations for PVP classification. ESM-PVP integrated a large pre-trained protein language model (PLM), i.e., ESM-2 [365], and an MLP to perform PVP identification and classification. A similar approach was proposed in [373], where various pretrained PLMs [63, 64, 376]) were used.

Phage lysins mining. Phage lysins are enzymes produced by bacteriophages to degrade bacterial cell walls, allowing newly replicated phages to burst out of the host cell [377]. These enzymes specifically target and break down peptidoglycan, a major component of bacterial cell walls, causing rapid bacterial cell lysis and death. Phage lysins have garnered interest as potential therapeutic agents, especially given the rise of antibiotic-resistant bacteria. Unlike traditional antibiotics, lysins have a unique mechanism of action and can target specific bacterial species, reducing the risk of off-target effects on beneficial microbiota. However, experimental lysin screening methods pose significant challenges due to heavy workload.

Very recently, AI techniques have been applied to discover novel phage lysins [378, 379]. DeepLysin is a unified software package to employ AI for mining the vast genome reservoirs for novel antibacterial phage lysins [378]. DeepLysin consists of two modules: the lysin mining module and the antibacterial activity prediction module. The input of the lysin mining module is assembled contigs. This module utilizes traditional blastP/protein sequence alignment-based methods to identify putative lysins. The second module estimates the antibacterial activity of the putative lysins identified by the first module. This module utilizes multiple AI techniques, such as Word2vec and an ensemble classifier that integrates five common classifiers to differentiate diverse and complex protein features. It ultimately applies Logistic Regression as a non-linear activation function to produce final activity predictions as scores ranging from 0 to 1, with higher scores indicating increased antibacterial activity. One limitation of DeepLysin is that four types of manually selected features (i.e., composition-based feature, binary profile-based feature, position-based feature, physiochemical based feature) need to be provided to the classifier. The feature selection procedure apparently heavily relies on domain knowledge.

DeepMineLys is a deep learning method based on CNN to identify phage lysins from human microbiome datasets [379]. DeepMineLys started from collecting phage protein sequences to build training and test datasets. These protein sequences were then processed using two distinct embedding methods (TAPE [380] and PHY [381]). Each of the two embeddings was fed into a CNN to learn sequence information and generate representations separately. The two representations of TAPE and PHY were then concatenated into a final representation and fed into a densely connected layer for the final prediction. DeepMineLys leverages existing methods for processing protein sequence features. To some extent, it alleviates the burden of manual feature selection.

Vaccine design

Vaccines work by stimulating the immune system to produce antibodies, offering protection against future infections. Traditional vaccine development, known as vaccinology, involves isolating a pathogen, identifying its antigenic components, and testing them for immune response. Reverse vaccinology (RV), a more modern and computational approach, begins by analyzing the pathogen's genome to identify potential antigenic proteins, which are then synthesized and evaluated as vaccine candidates. RV accelerates vaccine discovery and can reveal novel targets that traditional methods might overlook [382, 383].

Current RV approaches can be classified into two categories: (1) rule-based filtering methods, e.g., NERVE [384] and Vaxign [385]; and (2) Machine learning-based methods, e.g., VaxiJen [386], ANTIGENpro [387], Antigenic [388], and Vaxign-ML [389, 390]. The rule-based filtering method narrows down potential vaccine candidates from the large number of antigenic proteins identified through genome analysis. This process involves applying predefined biological rules or criteria (e.g., protein localization, the absence of similarity to host proteins to reduce the risk of autoimmune responses, immunogenicity potential, etc.). These rules help prioritize proteins most likely to elicit a protective immune response, speeding up vaccine candidate identification. Note that all these currently available rule-based filtering methods use only biological features as the data input. Machine learning-based RV methods predict potential vaccine candidates by training classifiers on known antigenic proteins and non-antigenic proteins. These machine learning methods can analyze physicochemical or biological features of the input proteins, and then classify new proteins based on the learned patterns. These machine learning methods can identify vaccine candidates with higher accuracy and efficiency compared to traditional methods, leveraging vast datasets and complex patterns that may not be evident through rule-based filtering alone. For example, Vaxign-ML, the successor to Vaxign, utilized XGBoost as the classifier and emerged as the top-performing Machine learning-based RV methods [389, 390].

Recently, deep learning techniques have also been developed for RV. For example, Vaxi-DL is a web-based deep learning software that evaluates the potential of protein sequences to serve as vaccine target antigens [391]. Vaxi-DL consists of four different deep learning pathogen models trained to predict target antigens in bacteria, protozoa, fungi, and viruses, respectively. All the four pathogen models are based on MLPs. For each pathogen model, a particular training dataset consisting of antigenic (positive samples) and non-antigenic (negative samples) sequences was derived from known vaccine candidates and the Protegen database. Vaxign-DL is another deep learning-based method to predict viable vaccine candidates from protein sequences [392]. Vaxign-DL is also based on MLP. It has been shown that Vaxign-DL achieved comparable results with Vaxign-ML in most cases, and outperformed Vaxi-DL in the prediction of bacterial protective antigens.

In the future, it would be interesting to test if other deep learning models (e.g., 1D CNN, RNN, and its variants, or Transformer) can also be used to predict target antigens.

Outlook

In this review article, we introduced the applications of AI techniques in various application scenarios in microbiology and microbiome research. There are some common challenges in those applications. Here we summarize those challenges and offer tentative solutions to inform future research.

Tradeoff between interpretability and complexity

Machine learning models, especially deep learning models, often suffer from high complexity and low interpretability, hindering their application in clinical decision-making. In addition, deep learning models typically have more than thousands of neural weights whose training requires large sample sizes and high computational resources. We anticipate that those deep learning models can reach better performance than traditional machine learning models as long as the sample size is enough. However, in most clinic-related studies, traditional models (e.g., Random Forest) are still widely used due to their ease of implementation, smaller sample size requirement, and better interpretability.

To address the interpretability issue, two different approaches can be employed. One approach is to employ methods such as SHAP (SHapley Additive exPlanations) [393], LIME (Local Interpretable Model-agnostic Explanations) [394] to enhance the interpretability of black-box models. SHAP is a game-theoretic method used to explain the output of any machine learning model. It links optimal credit allocation to local explanations by leveraging Shapley values from game theory and their related extensions. LIME is a technique that approximates any black box machine learning model with a local, interpretable model to explain each individual prediction. By applying SHAP and LIME, we can gain insights into complex deep learning models, identify biases, and improve transparency, crucial for applications in microbiome research.

The other approach is to employ "white-box" models. For instance, ReduNet [395] is a white-box deep network based on the principle of maximizing rate reduction. The authors argued that, at least in classification tasks, a key objective for a deep network is to learn a low-dimensional, linearly discriminative representation of the data. The effectiveness of this representation can be assessed by a principled measure from (lossy) data compression, i.e., rate reduction. Appropriately structured deep networks can then be naturally interpreted as optimization schemes designed to maximize this measure. The resulting multi-layer deep network shares key characteristics with modern deep learning architectures, but each component of ReduNet has a well-defined optimization, statistical, and geometric interpretation. Applying ReduNet to microbiome data would be an interesting attempt. Unlike ReduNet, MDITRE is a supervised deep learning method specifically designed for microbiome research. It takes a phylogenetic tree, microbiome time-series data, and host status labels to learn human-interpretable rules for predicting host status [396]. The model consists of five hidden layers that can be directly interpreted in terms of if-then rule statements. The first layer focuses on phylogenetic relationships by selecting taxa relevant to predicting host status. The second layer focuses on time by identifying relevant time windows for prediction. The following layers determine whether the data from selected taxa and time windows exceed specific learned thresholds, and subsequently combine these conditions to generate the final rules for prediction.

The "Small n, Large p" issue

Similar to many other omics studies, statistical or machine learning methods for microbiome research typically face the "small n, large p" issue, i.e., the number of parameters or microbial features (p) is much larger than the sample size (n). This issue may result in overfitting, models behaving unexpectedly, providing misleading results, or failing completely. There are several classical strategies to deal with the "small n, large p" issue, e.g., feature selection, projection methods, and regularization algorithms.

Feature selection involves selecting a subset of features to use as input to predictive models. Although the selection of an optimal subset of features is an NP-hard problem [397], many compromised feature selection methods have been proposed. Those methods are often grouped into filtering, wrapped, and embedded methods [398]. For instance, GRACES is a GCN-based feature selection method [399]. It exploits latent relations between samples with various overfitting-reducing techniques to iteratively find a set of optimal features which gives rise to the greatest decreases in the optimization loss. It has been demonstrated that GRACES significantly outperforms other feature selection methods on both synthetic and real-world gene expression datasets. It would be interesting to apply GRACES to microbiome data analysis.

Projection methods generate lower-dimensional representations of data while preserving the original relationships between samples. These techniques are often employed for visualization but can also serve as data transformations to reduce the number of predictors. Examples include linear algebra methods like SVD, PCA, and PCoA, as well as manifold learning algorithms, such as t-SNE, commonly used for visualization.

In standard machine learning models, regularization can be introduced during training to penalize the use or weighting of multiple features, promoting models that both perform well and minimize the number of predictors. This acts as an automatic feature selection process, and can involve augmenting existing models (e.g., regularized linear and logistic regression) or employing specialized methods like LASSO or multivariate nonlinear regression [400]. Since no single regularization method is universally optimal, it's advisable to conduct controlled experiments to evaluate various approaches.

Recently, it has been proposed to use promising deep learning techniques (e.g., transfer learning, self-supervised learning, semi-supervised learning, few-shot learning, zero-shot learning, etc.) to deal with the "small n, large p" issue [401]. For example, transfer learning involves pre-training a model on a large dataset and then fine-tuning it on a smaller, task-specific dataset [58]. By leveraging knowledge from a related but larger dataset, the pre-trained model can transfer learned representations to the small dataset, helping mitigate the issue of insufficient data. Self-supervised learning is an approach to creating supervisory signals from the data itself, eliminating the need for labeled data [57]. This approach can effectively learn useful representations even with limited labeled data, as the model can train on unlabeled data, which is usually more abundant. In microbiome research, self-supervised techniques can use metagenomics sequences without annotations to learn meaningful patterns, later applied to the small labeled subset. Semi-supervised learning leverages a small amount of labeled data and a large amount of unlabeled data to train the model. Since the labeled data is small (small n), semi-supervised learning helps by learning from both labeled and unlabeled data to improve generalization. Few-shot learning enables models to generalize from very few examples [402]. Few-shot learning techniques are specifically designed to handle scenarios with limited training data. They can quickly adapt to new tasks with only a handful of training samples. In personalized medicine, few-shot learning can help tailor models to individual patient data even when there is limited patient-specific training data. Zero-shot learning enables models to make predictions for classes they have not been explicitly trained on by learning from related classes or tasks [403]. This approach is especially useful when the data for certain categories or conditions is entirely missing (n = 0), allowing models to generalize from related categories or contexts. Deep learning models, especially those trained using self-supervised and transfer learning methods, can handle the high-dimensional feature space (large p) because they are adept at extracting useful features or representations from complex data. These approaches mitigate the problem of small sample sizes by either leveraging external data (e.g., transfer learning) or creating more efficient learning algorithms (e.g., few-shot and zero-shot learning). Applying those promising deep learning techniques to microbiome research to deal with the "small n, large p" issue would be very interesting. Some of the deep learning methods (especially those methods based on LLMs) discussed in this Review have already leveraged some of those techniques (e.g., transfer learning).

Benchmarking evaluations

As we mentioned in previous sections several times, benchmarking evaluations are typically lacking in microbiology and microbiome research. Currently, there is no standardized pipeline for benchmarking machine learning or deep learning methods in microbiology and microbiome research. To ensure reproducibility across studies, it's critical to standardize data preprocessing, which includes consistent methods for data collection, bioinformatics pipelines, and the profiling of microbiome taxonomies. Additionally, if feature dimension reduction is needed, it must be unbiased, using standardized methods for feature selection or reduction that apply uniformly across studies. Importantly, feature engineering should only be applied to training data and later evaluated on test data to avoid data leakage or overfitting. Furthermore, the creation of publicly available, well-annotated benchmarking datasets (analogous to MNIST or ImageNet in computer science) would provide the microbiome research community with reliable tools to assess and compare different machine learning models. Such datasets would accelerate progress and provide a framework for objective evaluation of new computational methods. Some attempts have been made in this regard. For example, MicrobiomeHD is a standardized database that compiles human gut microbiome studies related to health and disease [404]. It contains publicly available 16S data from published case-control studies, along with associated patient metadata. The raw sequencing data for each study was obtained and processed using a standardized pipeline. The curatedMetagenomicData package is another excellent example of benchmark microbiome datasets. It offers uniformly processed human microbiome data, including bacterial, fungal, archaeal, and viral taxonomic abundances, as well as quantitative metabolic functional profiles and standardized participant metadata [405]. This comprehensive, curated collection of metagenomic data is well-documented and easily accessible, making it suitable for benchmarking machine learning methods.

Establishing benchmark datasets is critical for advancing AI application in microbiology and microbiome research. Such datasets enable consistent, unbiased comparisons of algorithms and promote the development of robust predictive models. By providing standardized data, the research community can evaluate AI methods on a level playing field, ensuring reproducibility and transparency. Similar to the successful DREAM challenges in genomics, a community-driven effort to create public benchmarking datasets will foster collaboration, accelerate discovery, and establish best practices for AI approaches in microbiology and microbiome research. Collaborative input is vital for making this a reality.

Acknowledgments

We thank Dr. Yiyan Yang for valuable suggestions and for carefully examining the manuscript. Y.-Y.L. acknowledges funding support from the National Institutes of Health (R01AI141529, R01HD093761, RF1AG067744, UH3OD023268, U19AI095219 and U01HL089856) as well as the Office of the Assistant Secretary of Defense for Health Affairs, through the Traumatic Brain Injury and Psychological Health Research Program (Focused Program Award) under award no. W81XWH-22-S-TBIPH2, endorsed by the Department of Defense. X.-W.W. acknowledges funding support from the National Institutes of Health (K25HL166208).

Declaration of interests

The authors declare no competing interests.

References

1. Blaser, M. J., Cardon, Z. G., Cho, M. K., Dangl, J. L., Donohue, T. J., Green, J. L., Knight, R., Maxon, M. E., Northen, T. R., Pollard, K. S., et al. (2016). *Toward a predictive understanding of Earth's microbiomes to address 21st century challenges*. https://doi.org/10.1128/mbio.00714-16.

2. Lyons, T. W., Reinhard, C. T., and Planavsky, N. J. (2014). The rise of oxygen in Earth's early ocean and atmosphere. Nature *506*, 307–315. https://doi.org/10.1038/nature13068.

3. Oldroyd, G. E. and Dixon, R. (2014). Biotechnological solutions to the nitrogen problem. Current Opinion in Biotechnology *26*, 19–24. https://doi.org/10.1016/j.copbio.2013.08.006.

4. Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. Nature *449*, 804–810. https://doi.org/10.1038/nature06244.

5. Gadd, G. M. (2010). Metals, minerals and microbes: geomicrobiology and bioremediation. Microbiology *156*, 609–643. https://doi.org/10.1099/mic.0.037143-0.

6. Geisseler, D. and Scow, K. M. (2014). Long-term effects of mineral fertilizers on soil microorganisms–A review. Soil Biology and Biochemistry *75*, 54–63. https://doi.org/10.1016/j.soilbio.2014.03.023.

7. Grenni, P., Ancona, V., and Caracciolo, A. B. (2018). Ecological effects of antibiotics on natural ecosystems: A review. Microchemical Journal *136*, 25–39.

8. Martinez, J. L. (2009). Environmental Pollution by antibiotics and by antibiotic resistance determinants. Environmental Pollution *157*, 2893–2902.

9. Young, V. B. (2017). The role of the microbiome in human health and disease: an introduction for clinicians. Bmj *356*. https://doi.org/10.1136/bmj.j831.

10. Afzaal, M., Saeed, F., Shah, Y. A., Hussain, M., Rabail, R., Socol, C. T., Hassoun, A., Pateiro, M., Lorenzo, J. M., Rusu, A. V., et al. (2022). Human gut microbiota in health and disease: Unveiling the relationship. Frontiers in microbiology *13*, 999001. https://doi.org/10.3389/fmicb.2022.999001.

11. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "ImageNet: A large-scale hierarchical image database". *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

12. Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems *25*.

13. Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems.

14. Peiffer-Smadja, N., Dellière, S., Rodriguez, C., Birgand, G., Lescure, F.-X., Fourati, S., and Ruppé, E. (2020). Machine learning in the clinical microbiology laboratory: has the

time come for routine practice? Clinical Microbiology and Infection *26*, 1300–1309. https://doi.org/10.1016/j.cmi.2020.02.006.

15. Burns, B. L., Rhoads, D. D., and Misra, A. (2023). The use of machine learning for image analysis artificial intelligence in clinical microbiology. Journal of clinical microbiology *61*, e02336–21. https://doi.org/10.1128/jcm.02336-21.

16. Cox, M. J., Cookson, W. O., and Moffatt, M. F. (2013). Sequencing the human microbiome in health and disease. Human Molecular Genetics *22*, R88–R94. https://doi.org/10.1093/hmg/ddt398.

17. Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., Berland, M., Gruca, A., Hasic, J., Hron, K., et al. (2021). Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. Frontiers in Microbiology *12*, 313. https://doi.org/10.3410/f.739778223.793587742.

18. Wu, S., Chen, Y., Li, Z., Li, J., Zhao, F., and Su, X. (2021a). Towards multi-label classification: Next step of machine learning for microbiome research. Computational and Structural Biotechnology Journal *19*, 2742–2749. https://doi.org/10.1016/j.csbj.2021.04.054.

19. Qu, K., Guo, F., Liu, X., Lin, Y., and Zou, Q. (2019). Application of machine learning in microbiology. Frontiers in Microbiology *10*, 827. https://doi.org/10.3389/fmicb.2019.00827.

20. Ghannam, R. B. and Techtmann, S. M. (2021). Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. Computational and Structural Biotechnology Journal *19*, 1092–1107. https://doi.org/10.1016/j.csbj.2021.01.028.

21. Cammarota, G., Ianiro, G., Ahern, A., Carbone, C., Temko, A., Claesson, M. J., Gasbarrini, A., and Tortora, G. (2020). Gut microbiome, big data and machine learning to promote precision medicine for cancer. Nature reviews gastroenterology & hepatology *17*, 635–648. https://doi.org/10.1038/s41575-020-0327-3.

22. Moreno-Indias, I., Lahti, L., Nedyalkova, M., Elbere, I., Roshchupkin, G., Adilovic, M., Aydemir, O., Bakir-Gungor, B., Santa Pau, E. C.-d., D'Elia, D., et al. (2021). Statistical and machine learning techniques in human microbiome studies: contemporary challenges and solutions. Frontiers in Microbiology *12*, 277. https://doi.org/10.3389/fmicb.2021.635781.

23. Namkung, J. (2020). Machine learning methods for microbiome studies. Journal of Microbiology *58*, 206–216. https://doi.org/10.1007/s12275-020-0066-8.

24. Li, P., Luo, H., Ji, B., and Nielsen, J. (2022). Machine learning for data integration in human gut microbiome. Microbial Cell Factories *21*, 1–16. https://doi.org/10.1186/s12934-022-01973-4.

25. Yeşilyurt, N., Yılmaz, B., Ağagündüz, D., and Capasso, R. (2022). Microbiome-based personalized nutrition as a result of the 4.0 technological revolution: A mini literature review. Process Biochemistry. https://doi.org/10.1016/j.procbio.2022.07.012.

26. Metcalf, J. L., Xu, Z. Z., Bouslimani, A., Dorrestein, P., Carter, D. O., and Knight, R. (2017). Microbiome tools for forensic science. Trends in Biotechnology *35*, 814–823. https://doi.org/10.1016/j.tibtech.2017.03.006.

27. Goodswen, S. J., Barratt, J. L., Kennedy, P. J., Kaufer, A., Calarco, L., and Ellis, J. T. (2021). Machine learning and applications in microbiology. FEMS Microbiology Reviews *45*, fuab015. https://doi.org/10.1093/femsre/fuab015.

28. Soueidan, H. and Nikolski, M. (2015). Machine learning for metagenomics: methods and tools. arXiv preprint arXiv:1510.06621. https://doi.org/10.1515/metgen-2016-0001.

29. Roy, G., Prifti, E., Belda, E., and Zucker, J.-D. (2024). Deep learning methods in metagenomics: a review. Microbial Genomics *10*, 001231. https://doi.org/10.1099/mgen.0.001231.

30. Gerber, G. K. (2024). AI in microbiome research: Where have we been, where are we going? Cell Host & Microbe *32*, 1230–1234. https://doi.org/10.1016/j.chom.2024.07.021.

31. Lim, H., Cankara, F., Tsai, C.-J., Keskin, O., Nussinov, R., and Gursoy, A. (2022). Artificial intelligence approaches to human-microbiome protein–protein interactions. Current Opinion in Structural Biology *73*, 102328. https://doi.org/10.1016/j.sbi.2022.102328.

32. Zhu, Q., Huo, B., Sun, H., Li, B., and Jiang, X. (2020). Application of deep learning in microbiome. Journal of Artificial Intelligence for Medical Sciences *1*, 23–29. https://doi.org/10.2991/jaims.d.201028.001.

33. Zeng, T., Yu, X., and Chen, Z. (2021). Applying artificial intelligence in the microbiome for gastrointestinal diseases: A review. Journal of Gastroenterology and Hepatology *36*, 832–840. https://doi.org/10.1111/jgh.15503.

34. McCoubrey, L. E., Elbadawi, M., Orlu, M., Gaisford, S., and Basit, A. W. (2021). Harnessing machine learning for development of microbiome therapeutics. Gut Microbes *13*, 1872323. https://doi.org/10.1080/19490976.2021.1872323.

35. Loganathan, T. and Priya Doss C, G. (2022). The influence of machine learning technologies in gut microbiome research and cancer studies - A review. Life Sciences *311*, 121118. https://doi.org/10.1016/j.lfs.2022.121118.

36. Knights, D., Costello, E. K., and Knight, R. (2011). Supervised classification of human microbiota. FEMS Microbiology Reviews *35*, 343–359. https://doi.org/10.1111/j.1574-6976.2010.00251.x.

37. Hernández Medina, R., Kutuzova, S., Nielsen, K. N., Johansen, J., Hansen, L. H., Nielsen, M., and Rasmussen, S. (2022). Machine learning and deep learning applications in microbiome research. ISME Communications *2*, 98. https://doi.org/10.1038/s43705-022-00182-9.

38. Asnicar, F., Thomas, A. M., Passerini, A., Waldron, L., and Segata, N. (2023). Machine learning for microbiologists. Nature Reviews Microbiology, 1–15. https://doi.org/10.1038/s41579-023-00984-1.

39. Malakar, S., Sutaoney, P., Madhyastha, H., Shah, K., Chauhan, N. S., and Banerjee, P. (2024). Understanding gut microbiome-based machine learning platforms: A review on therapeutic approaches using deep learning. Chemical Biology & Drug Design *103*, e14505. https://doi.org/10.1111/cbdd.14505.

40. Lin, Y., Wang, G., Yu, J., and Sung, J. J. (2021). Artificial intelligence and metagenomics in intestinal diseases. Journal of Gastroenterology and Hepatology *36*, 841–847. https://doi.org/10.1111/jgh.15501.

41. Jiang, Y., Luo, J., Huang, D., Liu, Y., and Li, D.-d. (2022). Machine learning advances in microbiology: A review of methods and applications. Frontiers in Microbiology *13*, 925454. https://doi.org/10.3389/fmicb.2022.925454.

42. Iadanza, E., Fabbri, R., Bašić-ČiČak, D., Amedei, A., and Telalovic, J. H. (2020). Gut microbiota and artificial intelligence approaches: a scoping review. Health and Technology *10*, 1343–1358. https://doi.org/10.1007/s12553-020-00486-7.

43. Tonkovic, P., Kalajdziski, S., Zdravevski, E., Lameski, P., Corizzo, R., Pires, I. M., Garcia, N. M., Loncar-Turukalo, T., and Trajkovik, V. (2020). Literature on applied machine learning in metagenomic classification: a scoping review. Biology *9*, 453. https://doi.org/10.3390/biology9120453.

44. Loganathan, T. and George Priya Doss, C. (2022). The influence of machine learning technologies in gut microbiome research and cancer studies-A review. Life Sciences *311*, 121118. https://doi.org/10.1016/j.lfs.2022.121118.

45. Mathieu, A., Leclercq, M., Sanabria, M., Perin, O., and Droit, A. (2022). Machine learning and deep learning applications in metagenomic taxonomy and functional annotation. Frontiers in Microbiology *13*, 811495. https://doi.org/10.3389/fmicb.2022.811495.

46. Krause, T., Wassan, J. T., Mc Kevitt, P., Wang, H., Zheng, H., and Hemmje, M. (2021). Analyzing large microbiome datasets using machine learning and big data. BioMedInformatics *1*, 138–165. https://doi.org/10.3390/biomedinformatics1030010.

47. Abavisani, M., Foroushan, S. K., Ebadpour, N., and Sahebkar, A. (2024). Deciphering the gut microbiome: The revolution of artificial intelligence in microbiota analysis and intervention. Current Research in Biotechnology, 100211. https://doi.org/10.1016/j.crbiot.2024.100211.

48. He, Q., Niu, X., Qi, R.-Q., and Liu, M. (2022). Advances in microbial metagenomics and artificial intelligence analysis in forensic identification. Frontiers in Microbiology *13*, 1046733. https://doi.org/10.3389/fmicb.2022.1046733.

49. Kumar, P., Sinha, R., and Shukla, P. (2022). Artificial intelligence and synthetic biology approaches for human gut microbiome. Critical Reviews in Food Science and Nutrition *62*, 2103–2121. https://doi.org/10.1080/10408398.2020.1850415.

50. Wu, J., Singleton, S. S., Bhuiyan, U., Krammer, L., and Mazumder, R. (2024a). Multi-omics approaches to studying gastrointestinal microbiome in the context of precision medicine and machine learning. Frontiers in molecular biosciences *10*, 1337373. https://doi.org/10.3389/fmolb.2023.1337373.

51. Yan, B., Nam, Y., Li, L., Deek, R. A., Li, H., and Ma, S. (2024). Recent advances in deep learning and language models for studying the microbiome. arXiv preprint arXiv:2409.10579. https://doi.org/10.48550/arXiv.2409.10579.

52. Russell, S. and Norvig, P. (2021). Artificial intelligence: a modern approach, 4th US ed. aima: сайт. URL: https://aima. cs. berkeley. edu/(дата обращения: 26.02. 2023).

53. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag. ISBN: 0387310738.

54. Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning: A survey. Journal of Artificial Intelligence Research *4*, 237–285. https://doi.org/10.1613/jair.301.

55. Mnih, V. (2013). Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602. https://doi.org/10.48550/arXiv.1312.5602.

56. Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. (2014). “Deterministic policy gradient algorithms”. *International conference on machine learning*. Pmlr, 387–395. https://doi.org/10.1109/caibda53561.2021.00025.

57. Geiping, J., Garrido, Q., Fernandez, P., Bar, A., Pirsiavash, H., LeCun, Y., and Goldblum, M. (2023). A Cookbook of Self-Supervised Learning. arXiv preprint arXiv:2304.12210. https://doi.org/10.48550/arXiv.2304.12210.

58. Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. IEEE Transactions on knowledge and data engineering *22*, 1345–1359. https://doi.org/10.1109/TKDE.2009.191.

59. Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C. (2018). “A survey on deep transfer learning”. *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*. Springer, 270–279. https://doi.org/10.1007/978-3-030-01424-7_27.

60. Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics *37*. Ed. by J. Kelso, 2112–2120. https://doi.org/10.1093/bioinformatics/btab083.

61. Hwang, Y., Cornman, A. L., Kellogg, E. H., Ovchinnikov, S., and Girguis, P. R. (2024). Genomic language model predicts protein co-regulation and function. Nature Communications *15*, 2880. https://doi.org/10.1038/s41467-024-46947-9.

62. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences *118*, e2016239118. https://doi.org/10.3410/f.739876259.793585293.

63. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. (2021). Prottrans: Toward understanding the

language of life through self-supervised learning. IEEE transactions on pattern analysis and machine intelligence *44*, 7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381.

64. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. (2022). ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics *38*, 2102–2110. https://doi.org/10.1093/bioinformatics/btac020.

65. Montufar, G. F., Pascanu, R., Cho, K., and Bengio, Y. (2014). On the number of linear regions of deep neural networks. Advances in neural information processing systems *27*.

66. Pascanu, R., Gulcehre, C., Cho, K., and Bengio, Y. (2014). "How to construct deep recurrent neural networks". *International Conference on Learning Representations*.

67. Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., and Sohl-Dickstein, J. (2017). "On the expressive power of deep neural networks". *International Conference on Machine Learning*. PMLR, 2847–2854.

68. Serra, T., Tjandraatmadja, C., and Ramalingam, S. (2018). "Bounding and counting linear regions of deep neural networks". *International conference on Machine Learning*. PMLR, 4558–4566.

69. Arora, R., Basu, A., Mianjy, P., and Mukherjee, A. (2018). "Understanding Deep Neural Networks with Rectified Linear Units". *International Conference on Learning Representations*.

70. Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN Computer Science *2*, 160. https://doi.org/10.1007/s42979-021-00592-x.

71. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems *27*.

72. Rumelhart, D. E., McClelland, J. L., and Group, P. R. (1986). *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. The MIT press.

73. Kingma, D. P. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114. https://doi.org/10.48550/arXiv.1312.6114.

74. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. nature *550*, 354–359. https://doi.org/10.1038/nature24270.

75. Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*. Vol. 1. MIT press Cambridge.

76. Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., and Waterston, R. H. (2017). DNA sequencing at 40: past, present and future. Nature *550*, 345–353. https://doi.org/10.1038/nature24286.

77. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. Nature Biotechnology *35*, 833–844. https://doi.org/10.1038/nbt.3935.

78. Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L.-I., McDonald, D., et al. (2018). Best practices for analysing microbiomes. Nature Reviews Microbiology *16*, 410–422. https://doi.org/10.1038/s41579-018-0029-9.

79. Pinto, Y. and Bhatt, A. S. (2024). Sequencing-based analysis of microbiomes. Nature Reviews Genetics, 1–17. https://doi.org/10.1038/s41576-024-00746-6.

80. Schadt, E. E., Turner, S., and Kasarskis, A. (2010). A window into third-generation sequencing. Human Molecular Genetics *19*, R227–R240. https://doi.org/10.1093/hmg/ddq416.

81. Setubal, J. C. (2021). Metagenome-assembled genomes: concepts, analogies, and challenges. Biophysical Reviews *13*, 905–909. https://doi.org/10.1007/s12551-021-00865-y.

82. Mineeva, O., Rojas-Carulla, M., Ley, R. E., Schölkopf, B., and Youngblut, N. D. (2020). DeepMAsED: evaluating the quality of metagenomic assemblies. Bioinformatics *36*, 3011–3017. https://doi.org/10.1093/bioinformatics/btaa124.

83. Mineeva, O., Danciu, D., Schölkopf, B., Ley, R. E., Rätsch, G., and Youngblut, N. D. (2023). ResMiCo: Increasing the quality of metagenome-assembled genomes with deep learning. PLoS Computational Biology *19*, e1011001. https://doi.org/10.1371/journal.pcbi.1011001.

84. He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition". *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

85. Sedlar, K., Kupkova, K., and Provaznik, I. (2017). Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. Computational and structural biotechnology journal *15*, 48–55. https://doi.org/10.1016/j.csbj.2016.11.005.

86. Yang, C., Chowdhury, D., Zhang, Z., Cheung, W. K., Lu, A., Bian, Z., and Zhang, L. (2021). A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. Computational and Structural Biotechnology Journal *19*, 6301–6314. https://doi.org/10.1016/j.csbj.2021.11.028.

87. Lettich, R., Egan, R., Riley, R., Wang, Z., Tritt, A., Oliker, L., Yelick, K., and Buluç, A. (2024). GenomeFace: a deep learning-based metagenome binner trained on 43,000 microbial genomes. bioRxiv, 2024–02. https://doi.org/10.1101/2024.02.07.579326.

88. Lamurias, A., Tibo, A., Hose, K., Albertsen, M., and Nielsen, T. D. (2023). "Graph Neural Networks for Metagenomic Binning". *#PLACEHOLDER_PARENT_METADATA_VALUE#*. ICML compbio workshop.

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

90. Zhang, P., Jiang, Z., Wang, Y., and Li, Y. (2022). “CLMB: Deep contrastive learning for robust metagenomic binning”. *International Conference on Research in Computational Molecular Biology*. Springer, 326–348. https://doi.org/10.1101/2021.11.15.468566.

91. Pan, S., Zhu, C., Zhao, X.-M., and Coelho, L. P. (2022). A deep siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments. Nature Communications *13*, 2326. https://doi.org/10.1038/s41467-022-29843-y.

92. Lamurias, A., Sereika, M., Albertsen, M., Hose, K., and Nielsen, T. D. (2022). Metagenomic binning with assembly graph embeddings. Bioinformatics *38*. Ed. by I. Birol, 4481–4487. https://doi.org/10.1093/bioinformatics/btac557.

93. Wang, Z., You, R., Han, H., Liu, W., Sun, F., and Zhu, S. (2024a). Effective binning of metagenomic contigs using contrastive multi-view representation learning. Nature Communications *15*, 585. https://doi.org/10.1038/s41467-023-44290-z.

94. Chicco, D. (2021). Siamese neural networks: An overview. Artificial neural networks, 73–94. https://doi.org/10.1007/978-1-0716-0826-5_3.

95. Simon, H. Y., Siddle, K. J., Park, D. J., and Sabeti, P. C. (2019). Benchmarking metagenomics tools for taxonomic classification. Cell *178*, 779–794. https://doi.org/10.1016/j.cell.2019.07.010.

96. Lu, J., Breitwieser, F. P., Thielen, P., and Salzberg, S. L. (2017). Bracken: estimating species abundance in metagenomics data. PeerJ Computer Science *3*, e104. https://doi.org/10.7717/peerj-cs.104.

97. Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology *15*, 1–12. https://doi.org/10.1186/gb-2014-15-3-r46.

98. Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biology *20*, 1–13. https://doi.org/10.1186/s13059-019-1891-0.

99. Kostic, A. D., Ojesina, A. I., Pedamallu, C. S., Jung, J., Verhaak, R. G., Getz, G., and Meyerson, M. (2011). PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nature Biotechnology *29*, 393–396. https://doi.org/10.1038/nbt.1868.

100. Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. Nature Methods *12*, 59–60. https://doi.org/10.1038/nmeth.3176.

101. Menzel, P., Ng, K. L., and Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nature Communications *7*, 11257. https://doi.org/10.1038/ncomms11257.

102. Hauser, M., Steinegger, M., and Söding, J. (2016). MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. Bioinformatics *32*, 1323–1330. https://doi.org/10.1093/bioinformatics/btw006.

103. Steinegger, M. and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology *35*, 1026–1028. https://doi.org/10.1038/nbt.3988.

104. Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. Nature Methods *9*, 811–814. https://doi.org/10.1038/nmeth.2066.

105. Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods *12*, 902–903. https://doi.org/10.1038/nmeth.3589.

106. Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A. M., et al. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. eLife *10*, e65088. https://doi.org/10.7554/eLife.65088.

107. Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., Manghi, P., Dubois, L., Huang, K. D., Thomas, A. M., et al. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nature Biotechnology *41*, 1633–1644. https://doi.org/10.1038/s41587-023-01688-w.

108. Sunagawa, S., Mende, D. R., Zeller, G., Izquierdo-Carrasco, F., Berger, S. A., Kultima, J. R., Coelho, L. P., Arumugam, M., Tap, J., Nielsen, H. B., et al. (2013). Metagenomic species profiling using universal phylogenetic marker genes. Nature Methods *10*, 1196–1199. https://doi.org/10.1038/nmeth.2693.

109. Milanese, A., Mende, D. R., Paoli, L., Salazar, G., Ruscheweyh, H.-J., Cuenca, M., Hingamp, P., Alves, R., Costea, P. I., Coelho, L. P., et al. (2019). Microbial abundance, activity and population genomic profiling with mOTUs2. Nature Communications *10*, 1014. https://doi.org/10.1038/s41467-019-08844-4.

110. Sun, Z., Huang, S., Zhang, M., Zhu, Q., Haiminen, N., Carrieri, A. P., Vázquez-Baeza, Y., Parida, L., Kim, H.-C., Knight, R., et al. (2021). Challenges in benchmarking metagenomic profilers. Nature Methods *18*, 618–626. https://doi.org/10.1038/s41592-021-01141-3.

111. Louca, S., Mazel, F., Doebeli, M., and Parfrey, L. W. (2019). A census-based estimate of Earth’s bacterial and archaeal diversity. PLoS Biology *17*, e3000106. https://doi.org/10.1371/journal.pbio.3000106.

112. Liang, Q., Bible, P. W., Liu, Y., Zou, B., and Wei, L. (2020). DeepMicrobes: taxonomic classification for metagenomics with deep learning. NAR Genomics and Bioinformatics *2*, lqaa009. https://doi.org/10.1093/nargab/lqaa009.

113. Ounit, R., Wanamaker, S., Close, T. J., and Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics *16*, 1–13. https://doi.org/10.1186/s12864-015-1419-2.

114. Mock, F., Kretschmer, F., Kriese, A., Böcker, S., and Marz, M. (2022). Taxonomic classification of DNA sequences beyond sequence similarity using deep neural networks. Proceedings of the National Academy of Sciences *119*, e2122636119. https://doi.org/10.1073/pnas.2122636119.

115. Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., et al. (2008). The potential and challenges of nanopore sequencing. Nature Biotechnology *26*, 1146–1153. https://doi.org/10.1038/nbt.1495.

116. Teng, H., Cao, M. D., Hall, M. B., Duarte, T., Wang, S., and Coin, L. J. (2018). Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. GigaScience *7*, giy037. https://doi.org/10.1093/gigascience/giy037.

117. Huang, N., Nie, F., Ni, P., Luo, F., and Wang, J. (2020). Sacall: a neural network basecaller for oxford nanopore sequencing data based on self-attention mechanism. IEEE/ACM transactions on computational biology and bioinformatics *19*, 614–623. https://doi.org/10.1109/TCBB.2020.3039244.

118. Lv, X., Chen, Z., Lu, Y., and Yang, Y. (2020). “An end-to-end Oxford Nanopore basecaller using convolution-augmented transformer”. *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 337–342. https://doi.org/10.1109/BIBM49941.2020.9313290.

119. Xu, Z., Mai, Y., Liu, D., He, W., Lin, X., Xu, C., Zhang, L., Meng, X., Mafofo, J., Zaher, W. A., et al. (2021). Fast-bonito: A faster deep learning based basecaller for nanopore sequencing. Artificial Intelligence in the Life Sciences *1*, 100011. https://doi.org/10.1016/j.ailsci.2021.100011.

120. Miculinić, N., Ratković, M., and Šikić, M. (2019). MinCall-MinION end2end convolutional deep learning basecaller. arXiv preprint arXiv:1904.10337. https://doi.org/10.48550/arXiv.1904.10337.

121. Zeng, J., Cai, H., Peng, H., Wang, H., Zhang, Y., and Akutsu, T. (2020a). Causalcall: Nanopore basecalling using a temporal convolutional network. Frontiers in Genetics *10*, 1332. https://doi.org/10.3389/fgene.2019.01332.

122. Zhang, Y.-z., Akdemir, A., Tremmel, G., Imoto, S., Miyano, S., Shibuya, T., and Yamaguchi, R. (2020). Nanopore basecalling from a perspective of instance segmentation. BMC Bioinformatics *21*, 1–9. https://doi.org/10.1186/s12859-020-3459-0.

123. Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. CoRR *abs/1505.04597*. arXiv: 1505.04597.

124. Pagès-Gallego, M. and Ridder, J. de (2023). Comprehensive benchmark and architectural analysis of deep learning models for nanopore sequencing basecalling. Genome Biology *24*, 71. https://doi.org/10.1186/s13059-023-02903-2.

125. Zhu, W., Lomsadze, A., and Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. Nucleic Acids Research *38*, e132–e132. https://doi.org/10.1093/nar/gkq275.

126. Kelley, D. R., Liu, B., Delcher, A. L., Pop, M., and Salzberg, S. L. (2012). Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. Nucleic Acids Research *40*, e9–e9. https://doi.org/10.1093/nar/gkr1067.

127. Rho, M., Tang, H., and Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Research *38*, e191–e191. https://doi.org/10.1093/nar/gkq747.

128. Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 1–11. https://doi.org/10.1186/1471-2105-11-119.

129. Noguchi, H., Park, J., and Takagi, T. (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Research *34*, 5623–5630. https://doi.org/10.1093/nar/gkl723.

130. Noguchi, H., Taniguchi, T., and Itoh, T. (2008). MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA research *15*, 387–396. https://doi.org/10.1093/dnares/dsn027.

131. Zhang, S.-W., Jin, X.-Y., and Zhang, T. (2017). Gene prediction in metagenomic fragments with deep learning. BioMed Research International *2017*, 4740354. https://doi.org/10.1155/2017/4740354.

132. Al-Ajlan, A. and El Allali, A. (2019). CNN-MGP: convolutional neural networks for metagenomics gene prediction. Interdisciplinary Sciences: Computational Life Sciences *11*, 628–635. https://doi.org/10.1007/s12539-018-0313-4.

133. Sommer, M. J. and Salzberg, S. L. (2021). Balrog: a universal protein model for prokaryotic gene prediction. PLoS Computational Biology *17*, e1008727. https://doi.org/10.1371/journal.pcbi.1008727.

134. Rossolini, G. M., Arena, F., Pecile, P., and Pollini, S. (2014). Update on the antibiotic resistance crisis. Current Opinion in Pharmacology *18*, 56–60. https://doi.org/10.1016/j.coph.2014.09.006.

135. Kraker, M. E. de, Stewardson, A. J., and Harbarth, S. (2016). Will 10 million people die a year due to antimicrobial resistance by 2050? PLoS Medicine *13*, e1002184. https://doi.org/10.1371/journal.pmed.1002184.

136. Karkman, A., Do, T. T., Walsh, F., and Virta, M. P. (2018). Antibiotic-resistance genes in waste water. Trends in Microbiology *26*, 220–228. https://doi.org/10.1007/978-3-031-44618-4_6.

137. Zhang, X.-X., Zhang, T., and Fang, H. H. (2009). Antibiotic resistance genes in water environment. Applied Microbiology and Biotechnology *82*, 397–414. https://doi.org/10.1007/s00253-008-1829-z.

138. Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., and Zhang, L. (2018). DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome *6*, 1–15. https://doi.org/10.1186/s40168-018-0401-z.

139. Li, Y., Xu, Z., Han, W., Cao, H., Umarov, R., Yan, A., Fan, M., Chen, H., Duarte, C. M., Li, L., et al. (2021). HMD-ARG: hierarchical multi-task deep learning for annotating antibiotic resistance genes. Microbiome *9*, 1–12. https://doi.org/10.1186/s40168-021-01002-3.

140. Ji, B., Pi, W., Liu, W., Liu, Y., Cui, Y., Zhang, X., and Peng, S. (2023). HyperVR: a hybrid deep ensemble learning approach for simultaneously predicting virulence factors and antibiotic resistance genes. NAR Genomics and Bioinformatics *5*, lqad012. https://doi.org/10.1093/nargab/lqad012.

141. Pei, Y., Shum, M. H.-H., Liao, Y., Leung, V. W., Gong, Y.-N., Smith, D. K., Yin, X., Guan, Y., Luo, R., Zhang, T., et al. (2024). ARGNet: using deep neural networks for robust identification and classification of antibiotic resistance genes from sequences. Microbiome *12*, 1–17. https://doi.org/10.1186/s40168-024-01805-0.

142. Zhang, G., Wang, H., Zhang, Z., Zhang, L., Guo, G., Yang, J., Yuan, F., and Ju, F. (2024a). Highly accurate classification and discovery of microbial protein-coding gene functions using FunGeneTyper: an extensible deep learning framework. Briefings in Bioinformatics *25*, bbae319. https://doi.org/10.1093/bib/bbae319.

143. Andreopoulos, W. B., Geller, A. M., Lucke, M., Balewski, J., Clum, A., Ivanova, N. N., and Levy, A. (2022). Deeplasmid: deep learning accurately separates plasmids from bacterial chromosomes. Nucleic Acids Research *50*, e17–e17. https://doi.org/10.1093/nar/gkab1115.

144. Zhou, F. and Xu, Y. (2010). cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. Bioinformatics *26*, 2051–2052. https://doi.org/10.1093/bioinformatics/btq299.

145. Pellow, D., Mizrahi, I., and Shamir, R. (2020). PlasClass improves plasmid sequence classification. PLoS Computational Biology *16*, e1007781. https://doi.org/10.1371/journal.pcbi.1007781.

146. Antipov, D., Raiko, M., Lapidus, A., and Pevzner, P. A. (2019). Plasmid detection and assembly in genomic and metagenomic data sets. Genome Research *29*, 961–968. https://doi.org/10.1101/gr.241299.118.

147. Pradier, L., Tissot, T., Fiston-Lavier, A.-S., and Bedhomme, S. (2021). PlasForest: a homology-based random forest classifier for plasmid detection in genomic datasets. BMC Bioinformatics *22*, 349.

148. Zhu, Q., Gao, S., Xiao, B., He, Z., and Hu, S. (2023). Plasmer: an accurate and sensitive bacterial plasmid prediction Tool Based on Machine Learning of Shared k-mers

and genomic features. Microbiology Spectrum *11*, e04645–22. https://doi.org/10.1128/spectrum.04645-22.

149. Tian, R., Zhou, J., and Imanian, B. (2024). PlasmidHunter: Accurate and fast prediction of plasmid sequences using gene content profile and machine learning. Briefings in Bioinformatics *25*, bbae322. https://doi.org/10.1093/bib/bbae322.

150. Graaf-Van Bloois, L. van der, Wagenaar, J. A., and Zomer, A. L. (2021). RFPlasmid: predicting plasmid sequences from short-read assembly data using machine learning. Microbial Genomics *7*, 000683. https://doi.org/10.1099/mgen.0.000683.

151. Aytan-Aktug, D., Grigorjev, V., Szarvas, J., Clausen, P. T., Munk, P., Nguyen, M., Davis, J. J., Aarestrup, F. M., and Lund, O. (2022). SourceFinder: A machine-learning-based tool for identification of chromosomal, plasmid, and bacteriophage sequences from assemblies. Microbiology Spectrum *10*, e02641–22. https://doi.org//10.1128/spectrum.02641-22.

152. Krawczyk, P. S., Lipinski, L., and Dziembowski, A. (2018). PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. Nucleic Acids Research *46*, e35–e35. https://doi.org/10.1093/nar/gkx1321.

153. Sielemann, J., Sielemann, K., Brejová, B., Vinař, T., and Chauve, C. (2023). plASgraph2: using graph neural networks to detect plasmid contigs from an assembly graph. Frontiers in Microbiology *14*, 1267695. https://doi.org/10.3389/fmicb.2023.1267695.

154. Fang, Z., Tan, J., Wu, S., Li, M., Xu, C., Xie, Z., and Zhu, H. (2019). PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. Gigascience *8*, giz066. https://doi.org/10.1093/gigascience/giz066.

155. Camargo, A. P., Roux, S., Schulz, F., Babinski, M., Xu, Y., Hu, B., Chain, P. S., Nayfach, S., and Kyrpides, N. C. (2023). Identification of mobile genetic elements with geNomad. Nature Biotechnology, 1–10. https://doi.org/10.1038/s41587-023-01953-y.

156. Sourkov, V. (2018). Igloo: Slicing the features space to represent sequences. arXiv preprint arXiv:1807.03402. https://doi.org/10.48550/arXiv.1807.03402.

157. Dias, D. A., Urban, S., and Roessner, U. (2012). A historical overview of natural products in drug discovery. Metabolites *2*, 303–336. https://doi.org/10.3390/metabo2020303.

158. Wang, S., Li, N., Zou, H., and Wu, M. (2019). Gut microbiome-based secondary metabolite biosynthetic gene clusters detection in Parkinson's disease. Neuroscience Letters *696*, 93–98. https://doi.org/10.1016/j.neulet.2018.12.021.

159. Hannigan, G. D., Prihoda, D., Palicka, A., Soukup, J., Klempir, O., Rampula, L., Durcak, J., Wurst, M., Kotowski, J., Chang, D., et al. (2019). A deep learning genome-mining strategy for biosynthetic gene cluster prediction. Nucleic Acids Research *47*, e110–e110. https://doi.org/10.1093/nar/gkz654.

160. Liu, M., Li, Y., and Li, H. (2022). Deep learning to predict the biosynthetic gene clusters in bacterial genomes. Journal of Molecular Biology *434*, 167597. https://doi.org/10.1016/j.jmb.2022.167597.

161. Rios-Martinez, C., Bhattacharya, N., Amini, A. P., Crawford, L., and Yang, K. K. (2023). Deep self-supervised learning for biosynthetic gene cluster detection and product classification. PLoS Computational Biology *19*, e1011162. https://doi.org/10.1371/journal.pcbi.1011162.

162. Qilong, L., Shuai, Y., Yuguo, Z., Hong, B., and Kang, N. (2023). Microbiome-based biosynthetic gene cluster data mining techniques and application potentials. Synthetic Biology Journal *4*, 611. https://doi.org/10.12211/2096-8280.2022-075.

163. Yang, K. K., Fusi, N., and Lu, A. X. (2024). Convolutions are competitive with transformers for protein sequence pretraining. Cell Systems *15*, 286–294. https://doi.org/10.1016/j.cels.2024.01.008.

164. Sanchez, S., Rogers, J. D., Rogers, A. B., Nassar, M., McEntyre, J., Welch, M., Hollfelder, F., and Finn, R. D. (2023). Expansion of novel biosynthetic gene clusters from diverse environments using SanntiS. bioRxiv, 2023–05. https://doi.org/10.1101/2023.05.23.540769.

165. Klappenbach, J. A., Dunbar, J. M., and Schmidt, T. M. (2000). rRNA operon copy number reflects ecological strategies of bacteria. Applied and Environmental Microbiology *66*, 1328–1333. https://doi.org/10.1128/AEM.66.4.1328-1333.2000.

166. Kembel, S. W., Wu, M., Eisen, J. A., and Green, J. L. (2012). Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. PLoS Computational Biology *8*, e1002743. https://doi.org/10.1371/journal.pcbi.1002743.

167. Angly, F. E., Dennis, P. G., Skarshewski, A., Vanwonterghem, I., Hugenholtz, P., and Tyson, G. W. (2014). CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. Microbiome *2*, 1–13. https://doi.org/10.1186/2049-2618-2-11.

168. Stoddard, S. F., Smith, B. J., Hein, R., Roller, B. R., and Schmidt, T. M. (2015). rrn DB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nucleic Acids Research *43*, D593–D598. https://doi.org/10.1093/nar/gku1201.

169. Douglas, G. M., Maffei, V. J., Zaneveld, J. R., Yurgel, S. N., Brown, J. R., Taylor, C. M., Huttenhower, C., and Langille, M. G. (2020). PICRUSt2 for prediction of metagenome functions. Nature Biotechnology *38*, 685–688. https://doi.org/10.1038/s41587-020-0548-6.

170. Louca, S., Doebeli, M., and Parfrey, L. W. (2018). Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. Microbiome *6*, 1–12. https://doi.org/10.1186/s40168-018-0420-9.

171. Miao, J., Chen, T., Misir, M., and Lin, Y. (2024a). Deep learning for predicting 16S rRNA gene copy number. Scientific Reports *14*, 14282. https://doi.org/10.1038/s41598-024-64658-5.

172. Wang, X., Zorraquino, V., Kim, M., Tsoukalas, A., and Tagkopoulos, I. (2018). Predicting the evolution of Escherichia coli by a data-driven approach. Nature Communications *9*, 3562. https://doi.org/10.1038/s41467-018-05807-z.

173. Thadani, N. N., Gurev, S., Notin, P., Youssef, N., Rollins, N. J., Ritter, D., Sander, C., Gal, Y., and Marks, D. S. (2023). Learning from prepandemic data to forecast viral escape. Nature *622*, 818–825. https://doi.org/10.1038/s41586-023-06617-0.

174. Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., Gal, Y., and Marks, D. S. (2021). Disease variant prediction with deep generative models of evolutionary data. Nature *599*, 91–95. https://doi.org/10.1038/s41586-021-04043-8.

175. Konno, N. and Iwasaki, W. (2023). Machine learning enables prediction of metabolic system evolution in bacteria. Science Advances *9*, eadc9130. https://doi.org/10.1126/sciadv.adc9.

176. Post, S. E. and Brito, I. L. (2022). Structural insight into protein–protein interactions between intestinal microbiome and host. Current Opinion in Structural Biology *74*, 102354. https://doi.org/10.1016/j.sbi.2022.102354.

177. Balint, D. and Brito, I. L. (2024). Human–gut bacterial protein–protein interactions: understudied but impactful to human health. Trends in Microbiology *32*, 325–332. https://doi.org/10.1016/j.tim.2023.09.009.

178. Pan, J., Zhang, Z., Li, Y., Yu, J., You, Z., Li, C., Wang, S., Zhu, M., Ren, F., Zhang, X., et al. (2024). A microbial knowledge graph-based deep learning model for predicting candidate microbes for target hosts. Briefings in Bioinformatics *25*, bbae119. https://doi.org/10.1093/bib/bbae119.

179. Chen, M., Ju, C. J.-T., Zhou, G., Chen, X., Zhang, T., Chang, K.-W., Zaniolo, C., and Wang, W. (2019). Multifaceted protein–protein interaction prediction based on Siamese residual RCNN. Bioinformatics *35*, i305–i314. https://doi.org/10.1093/bioinformatics/btz328.

180. Zeng, M., Zhang, F., Wu, F.-X., Li, Y., Wang, J., and Li, M. (2020b). Protein–protein interaction site prediction through combining local and global features with deep neural networks. Bioinformatics *36*, 1114–1120. https://doi.org/10.1093/bioinformatics/btz699.

181. Chen, H., Shen, J., Wang, L., and Song, J. (2020). A framework towards data analytics on host–pathogen protein–protein interactions. Journal of Ambient Intelligence and Humanized Computing *11*, 4667–4679. https://doi.org/10.1007/s12652-020-01715-7.

182. Liu-Wei, W., Kafkas, Ş., Chen, J., Dimonaco, N. J., Tegnér, J., and Hoehndorf, R. (2021). DeepViral: prediction of novel virus–host interactions from protein sequences and infectious disease phenotypes. Bioinformatics *37*, 2722–2729. https://doi.org/10.1093/bioinformatics/btab147.

183. Balci, A., Gumeli, C., Hakouz, A., Yuret, D., Keskin, O., and Gursoy, A. (2019). Deep-Interface: protein-protein interface validation using 3D convolutional neural networks. BiorXiv, 617506. https://doi.org/10.1101/617506.

184. Gainza, P., Sverrisson, F., Monti, F., Rodola, E., Boscaini, D., Bronstein, M., and Correia, B. (2020). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nature Methods *17*, 184–192. https://doi.org/10.1038/s41592-019-0666-6.

185. Pittala, S. and Bailey-Kellogg, C. (2020). Learning context-aware structural representations to predict antigen and antibody binding interfaces. Bioinformatics *36*, 3996–4003. https://doi.org/10.1093/bioinformatics/btaa263.

186. Wang, X.-W., Madeddu, L., Spirohn, K., Martini, L., Fazzone, A., Becchetti, L., Wytock, T. P., Kovács, I. A., Balogh, O. M., Benczik, B., et al. (2023a). Assessment of community efforts to advance network-based prediction of protein–protein interactions. Nature Communications *14*, 1582. https://doi.org/10.1038/s41467-023-37079-7.

187. Morton, J. T., Aksenov, A. A., Nothias, L. F., Foulds, J. R., Quinn, R. A., Badri, M. H., Swenson, T. L., Van Goethem, M. W., Northen, T. R., Vazquez-Baeza, Y., et al. (2019). Learning representations of microbe–metabolite interactions. Nature Methods *16*, 1306–1314. https://doi.org/10.1038/s41592-019-0616-3.

188. Mikolov, T. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. https://doi.org/10.48550/arXiv.1301.3781.

189. Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., Geng, B., Yang, J., Kong, W., Zhou, X., and Cui, Q. (2017). An analysis of human microbe–disease associations. Briefings in Bioinformatics *18*, 85–97. https://doi.org/10.1093/bib/bbw005.

190. Jin, H., Hu, G., Sun, C., Duan, Y., Zhang, Z., Liu, Z., Zhao, X.-M., and Chen, W.-H. (2022). mBodyMap: a curated database for microbes across human body and their associations with health and diseases. Nucleic Acids Research *50*, D808–D816. https://doi.org/10.1093/nar/gkab973.

191. Ma, Y. and Jiang, H. (2020). NinimHMDA: neural integration of neighborhood information on a multiplex heterogeneous network for multiple types of human microbe–disease association. Bioinformatics *36*, 5665–5671. https://doi.org/10.1093/bioinformatics/btaa1080.

192. Lei, X. and Wang, Y. (2020). Predicting microbe-disease association by learning graph representations and rule-based inference on the heterogeneous network. Frontiers in Microbiology *11*, 579. https://doi.org/10.3389/fmicb.2020.00579.

193. Li, H., Wang, Y., Zhang, Z., Tan, Y., Chen, Z., Wang, X., Pei, T., and Wang, L. (2020). Identifying microbe-disease association based on a novel back-propagation neural network model. IEEE/ACM transactions on computational biology and bioinformatics *18*, 2502–2513. https://doi.org/10.1109/tcbb.2020.2986459.

194. Liu, Y., Wang, S.-L., Zhang, J.-F., Zhang, W., Zhou, S., and Li, W. (2020). Dmfmda: Prediction of microbe-disease associations based on deep matrix factorization using bayesian personalized ranking. IEEE/ACM Transactions on Computational Biology and Bioinformatics *18*, 1763–1772. https://doi.org/10.1109/tcbb.2020.3018138.

195. Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). “Deepwalk: Online learning of social representations”. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 701–710. https://doi.org/10.1145/2623330.2623732.

196. Dong, Y., Chawla, N. V., and Swami, A. (2017). “metapath2vec: Scalable representation learning for heterogeneous networks”. *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 135–144. https://doi.org/10.1145/3097983.3098036.

197. Karkera, N., Acharya, S., and Palaniappan, S. K. (2023). Leveraging pre-trained language models for mining microbiome-disease relationships. BMC Bioinformatics *24*, 290. https://doi.org/10.1186/s12859-023-05411-z.

198. Liu, Z., Sun, Y., Li, Y., Ma, A., Willaims, N. F., Jahanbahkshi, S., Hoyd, R., Wang, X., Zhang, S., Zhu, J., et al. (2023). An Explainable Graph Neural Framework to Identify Cancer-Associated Intratumoral Microbial Communities. Advanced Science, 2403393. https://doi.org/10.1002/advs.202403393.

199. Sung, J., Kim, S., Cabatbat, J. J. T., Jang, S., Jin, Y.-S., Jung, G. Y., Chia, N., and Kim, P.-J. (2017). Global metabolic interaction network of the human gut microbiota for context-specific community-scale analysis. Nature Communications *8*, 15393. https://doi.org/10.1038/ncomms15393.

200. Kuang, H., Zhang, Z., Zeng, B., Liu, X., Zuo, H., Xu, X., and Wang, L. (2024). A novel microbe-drug association prediction model based on graph attention networks and bilayer random forest. BMC Bioinformatics *25*, 78. https://doi.org/10.1186/s12859-024-05687-9.

201. Wang, B., Wang, T., Du, X., Li, J., Wang, J., and Wu, P. (2024b). Microbe-drug association prediction model based on graph convolution and attention networks. Scientific Reports *14*, 22327. https://doi.org/10.1038/s41598-024-71834-0.

202. Yang, Z., Wang, L., Zhang, X., Zeng, B., Zhang, Z., and Liu, X. (2024a). LCASPMDA: a computational model for predicting potential microbe-drug associations based on learnable graph convolutional attention networks and self-paced iterative sampling ensemble. Frontiers in Microbiology *15*, 1366272. https://doi.org/10.3389/fmicb.2024.1366272.

203. Li, G., Cao, Z., Liang, C., Xiao, Q., and Luo, J. (2024). MCHAN: Prediction of Human Microbe-drug Associations Based on Multiview Contrastive Hypergraph Attention Network. CURRENT BIOINFORMATICS. https://doi.org/10.2174/0115748936288616240212073805.

204. Tan, H., Zhang, Z., Liu, X., Chen, Y., Yang, Z., and Wang, L. (2024). MDSVDNV: predicting microbe–drug associations by singular value decomposition and Node2vec. Frontiers in Microbiology *14*, 1303585. https://doi.org/10.3389/fmicb.2023.1303585.

205. Liang, M., Liu, X., Chen, Q., Zeng, B., and Wang, L. (2024). NMGMDA: a computational model for predicting potential microbe–drug associations based on minimize matrix nuclear norm and graph attention network. Scientific Reports *14*, 650. https://doi.org/10.1038/s41598-023-50793-y.

206. Zhao, J., Kuang, L., Hu, A., Zhang, Q., Yang, D., and Wang, C. (2024). OGNNMDA: a computational model for microbe-drug association prediction based on ordered message-passing graph neural networks. Frontiers in Genetics *15*, 1370013. https://doi.org/10.3389/fgene.2024.1370013.

207. Liu, F., Xiaoyu, Y., Lei, W., and Xianyou, Z. (2024). STNMDA: A Novel Model for Predicting Potential Microbe-Drug Associations with Structure-Aware Transformer. Current Bioinformatics *19*, 919–932. https://doi.org/10.2174/0115748936272939231212102627.

208. Wang, L., Tan, Y., Yang, X., Kuang, L., and Ping, P. (2022a). Review on predicting pairwise relationships between human microbes, drugs and diseases: from biological data to computational models. Briefings in Bioinformatics *23*, bbac080. https://doi.org/10.1093/bib/bbac080.

209. Fan, L., Yang, X., LeiWang, and Zhu, X. (2024). STNMDA: A Novel Model for Predicting Potential Microbe-Drug Associations with Structure-Aware Transformer. Current Bioinformatics *19*, 919–932. https://doi.org/10.2174/0115748936272939231212102627.

210. Liu, Y.-Y. (2023). Controlling the human microbiome. Cell Systems *14*, 135–159. https://doi.org/10.1016/j.cels.2022.12.010.

211. Cao, H.-T., Gibson, T. E., Bashan, A., and Liu, Y.-Y. (2017). Inferring human microbial dynamics from temporal metagenomics data: Pitfalls and lessons. BioEssays *39*, 1600188. https://doi.org/10.1002/bies.201600188.

212. Gerber, G. K., Onderdonk, A. B., and Bry, L. (2012). Inferring dynamic signatures of microbes in complex host ecosystems. PLoS Computational Biology. https://doi.org/10.1371/journal.pcbi.1002624.

213. Stein, R. R., Bucci, V., Toussaint, N. C., Buffie, C. G., Rätsch, G., Pamer, E. G., Sander, C., and Xavier, J. B. (2013). Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. PLoS Computational Biology *9*, e1003388. https://doi.org/10.1371/journal.pcbi.1003388.

214. Steinway, S. N., Biggs, M. B., Loughran Jr, T. P., Papin, J. A., and Albert, R. (2015). Inference of network dynamics and metabolic interactions in the gut microbiome. PLoS Computational Biology *11*, e1004338. https://doi.org/10.1371/journal.pcbi.1004338.

215. Bucci, V., Tzen, B., Li, N., Simmons, M., Tanoue, T., Bogart, E., Deng, L., Yeliseyev, V., Delaney, M. L., Liu, Q., et al. (2016). MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. Genome biology *17*, 1–17. https://doi.org/10.1186/s13059-016-0980-6.

216. Xiao, Y., Angulo, M. T., Friedman, J., Waldor, M. K., Weiss, S. T., and Liu, Y.-Y. (2017). Mapping the ecological networks of microbial communities. Nature Communications *8*, 2042. https://doi.org/10.1038/s41467-017-02090-2.

217. DiMucci, D., Kon, M., and Segrè, D. (2018). Machine learning reveals missing edges and putative interaction mechanisms in microbial ecosystem networks. Msystems *3*, 10–1128. https://doi.org/10.1128/msystems.00181-18.

218. Michel Mata, S., Wang, X.-W., Liu, Y.-Y., and Angulo, M. T. (2022). Predicting microbiome compositions from species assemblages through deep learning. iMeta *1*, e3. https://doi.org/10.1002/imt2.3.

219. Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. (2018). Neural ordinary differential equations. Advances in neural information processing systems *31*.

220. Ruaud, A., Sancaktar, C., Bagatella, M., Ratzke, C., and Martius, G. (2024). *Modelling Microbial Communities with Graph Neural Networks*.

221. Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. Advances in neural information processing systems *30*.

222. Kipf, T. N. and Welling, M. (2017). *Semi-Supervised Classification with Graph Convolutional Networks*. arXiv:1609.02907 [cs, stat].

223. Wang, X.-W., Sun, Z., Jia, H., Michel-Mata, S., Angulo, M. T., Dai, L., He, X., Weiss, S. T., and Liu, Y.-Y. (2024c). Identifying keystone species in microbial communities using deep learning. Nature Ecology & Evolution *8*, 22–31. https://doi.org/10.1038/s41559-023-02250-2.

224. Schwartz, D. J., Langdon, A. E., and Dantas, G. (2020). Understanding the impact of antibiotic perturbation on the human microbiome. Genome Medicine *12*, 82. https://doi.org/10.1186/s13073-020-00782-x.

225. Benjamino, J., Lincoln, S., Srivastava, R., and Graf, J. (2018). Low-abundant bacteria drive compositional changes in the gut microbiota after dietary alteration. Microbiome *6*, 1–13. https://doi.org/10.1186/s40168-018-0469-5.

226. Wu, L., Wang, X.-W., Tao, Z., Wang, T., Zuo, W., Zeng, Y., Liu, Y.-Y., and Dai, L. (2024b). Data-driven prediction of colonization outcomes for complex microbial communities. Nature Communications *15*, 2406. https://doi.org/10.52843/cassyni.r4c572.

227. Ianiro, G., Punčochář, M., Karcher, N., Porcari, S., Armanini, F., Asnicar, F., Beghini, F., Blanco-Míguez, A., Cumbo, F., Manghi, P., et al. (2022). Variability of strain engraftment and predictability of microbiome composition after fecal microbiota transplantation across different diseases. Nature Medicine *28*, 1913–1923. https://doi.org/10.1038/s41591-022-01964-3.

228. Baranwal, M., Clark, R. L., Thompson, J., Sun, Z., Hero, A. O., and Venturelli, O. S. (2022). Recurrent neural networks enable design of multifunctional synthetic human gut microbiome dynamics. eLife *11*, e73870. https://doi.org/10.7554/elife.73870.sa0.

229. Zhao, K., Guo, C., Cheng, Y., Han, P., Zhang, M., and Yang, B. (2023). Multiple time series forecasting with dynamic graph modeling. Proceedings of the VLDB Endowment *17*, 753–765.

230. Ma, S., Ren, B., Mallick, H., Moon, Y. S., Schwager, E., Maharjan, S., Tickle, T. L., Lu, Y., Carmody, R. N., Franzosa, E. A., et al. (2021). A statistical model for describing and simulating microbial community profiles. PLoS Computational Biology *17*, e1008913. https://doi.org/10.1371/journal.pcbi.1008913.

231. Gao, Y., Şimşek, Y., Gheysen, E., Borman, T., Li, Y., Lahti, L., Faust, K., and Garza, D. R. (2023). miaSim: an R/Bioconductor package to easily simulate microbial community dynamics. Methods in Ecology and Evolution *14*, 1967–1980. https://doi.org/10.1111/2041-210x.14129.

232. Rong, R., Jiang, S., Xu, L., Xiao, G., Xie, Y., Liu, D. J., Li, Q., and Zhan, X. (2021). MB-GAN: Microbiome Simulation via Generative Adversarial Network. GigaScience *10*, giab005. https://doi.org/10.1093/gigascience/giab005.

233. Oh, M. and Zhang, L. (2022). Generalizing predictions to unseen sequencing profiles via deep generative models. Scientific Reports *12*, 7151. https://doi.org/10.1038/s41598-022-11363-w.

234. Choi, J. M., Ji, M., Watson, L. T., and Zhang, L. (2023). DeepMicroGen: a generative adversarial network-based method for longitudinal microbiome data imputation. Bioinformatics *39*. Ed. by V. Boeva, btad286. https://doi.org/10.1093/bioinformatics/btad286.

235. Knights, D., Kuczynski, J., Charlson, E. S., Zaneveld, J., Mozer, M. C., Collman, R. G., Bushman, F. D., Knight, R., and Kelley, S. T. (2011). Bayesian community-wide culture-independent microbial source tracking. Nature Methods *8*, 761–763. https://doi.org/10.1038/nmeth.1650.

236. Shenhav, L., Thompson, M., Joseph, T. A., Briscoe, L., Furman, O., Bogumil, D., Mizrahi, I., Pe'er, I., and Halperin, E. (2019). FEAST: fast expectation-maximization for microbial source tracking. Nature Methods *16*, 627–632. https://doi.org/10.1038/s41592-019-0431-x.

237. An, U., Shenhav, L., Olson, C. A., Hsiao, E. Y., Halperin, E., and Sankararaman, S. (2022). STENSL: Microbial Source Tracking with ENvironment SeLection. Msystems *7*, e00995–21. https://doi.org/10.1128/msystems.00995-21.

238. Zha, Y., Chong, H., Qiu, H., Kang, K., Dun, Y., Chen, Z., Cui, X., and Ning, K. (2022). Ontology-aware deep learning enables ultrafast and interpretable source tracking among sub-million microbial community samples from hundreds of niches. Genome Medicine *14*, 43. https://doi.org/10.1186/s13073-022-01047-5.

239. Wang, X.-W., Wu, L., Dai, L., Yin, X., Zhang, T., Weiss, S. T., and Liu, Y.-Y. (2023b). Ecological dynamics imposes fundamental challenges in community-based microbial source tracking. iMeta *2*, e75. https://doi.org/10.1002/imt2.145.

240. Griffiths, T. L. (2004). Finding Scientific Topics. PNAS. https://doi.org/10.1073/pnas.0307752101.

241. Orth, J. D. and Palsson, B. Ø. (2010). Systematizing the generation of missing metabolic knowledge. Biotechnology and bioengineering *107*, 403–412. https://doi.org/10.1002/bit.22844.

242. Pan, S. and Reed, J. L. (2018). Advances in gap-filling genome-scale metabolic models and model-driven experiments lead to novel metabolic discoveries. Current Opinion in Biotechnology *51*, 103–108. https://doi.org/10.1016/j.copbio.2017.12.012.

243. Rana, P., Berry, C., Ghosh, P., and Fong, S. S. (2020). Recent advances on constraint-based models by integrating machine learning. Current Opinion in Biotechnology *64*, 85–91. https://doi.org/10.1016/j.copbio.2019.11.007.

244. Chen, C. and Liu, Y.-Y. (2023). A survey on hyperlink prediction. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS.2023.3286280.

245. Chen, C., Liao, C., and Liu, Y.-Y. (2023). Teasing out missing reactions in genome-scale metabolic networks through hypergraph learning. Nature Communications *14*, 2375. https://doi.org/10.1038/s41467-023-38110-7.

246. Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. Advances in neural information processing systems *29*. https://doi.org/10.1109/access.2020.2999520.

247. Yadati, N., Nitin, V., Nimishakavi, M., Yadav, P., Louis, A., and Talukdar, P. (2020). “Nhp: Neural hypergraph link prediction”. *Proceedings of the 29th ACM international conference on information & knowledge management*, 1705–1714. https://doi.org/10.1145/3340531.3411870.

248. Sharma, G., Patil, P., and Murty, M. N. (2021). “C3mm: clique-closure based hyperlink prediction”. *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 3364–3370. https://doi.org/10.24963/ijcai.2020/465.

249. Koch, M., Duigou, T., and Faulon, J.-L. (2019). Reinforcement learning for bioretrosynthesis. ACS synthetic biology *9*, 157–168. https://doi.org/10.1021/acssynbio.9b00447.

250. Coulom, R. (2006). “Efficient selectivity and backup operators in Monte-Carlo tree search”. *International conference on computers and games*. Springer, 72–83. https://doi.org/10.1007/978-3-540-75538-8_7.

251. Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of Go with deep neural networks and tree search. Nature *529*, 484–489. https://doi.org/10.1038/nature16961.

252. Duigou, T., Du Lac, M., Carbonell, P., and Faulon, J.-L. (2019). RetroRules: a database of reaction rules for engineering biology. Nucleic Acids Research *47*, D1229–D1235. https://doi.org/10.1093/nar/gky940.

253. Balzerani, F., Hinojosa-Nogueira, D., Cendoya, X., Blasco, T., Pérez-Burillo, S., Apaolaza, I., Francino, M. P., Rufián-Henares, J. Á., and Planes, F. J. (2022). Prediction of degradation pathways of phenolic compounds in the human gut microbiota through enzyme promiscuity methods. NPJ systems biology and applications *8*, 24. https://doi.org/10.1038/s41540-022-00234-9.

254. Rothwell, J. A., Perez-Jimenez, J., Neveu, V., Medina-Remon, A., M’hiri, N., García-Lobato, P., Manach, C., Knox, C., Eisner, R., Wishart, D. S., et al. (2013). Phenol-Explorer 3.0: a major update of the Phenol-Explorer database to incorporate data on

the effects of food processing on polyphenol content. Database *2013*, bat070. https://doi.org/10.1093/database/bat070.

255. Blasco, T., Pérez-Burillo, S., Balzerani, F., Hinojosa-Nogueira, D., Lerma-Aguilera, A., Pastoriza, S., Cendoya, X., Rubio, Á., Gosalbes, M. J., Jiménez-Hernández, N., et al. (2021). An extended reconstruction of human gut microbiota metabolism of dietary compounds. Nature Communications *12*, 4728. https://doi.org/10.1038/s41467-021-25056-x.

256. Bar, N., Korem, T., Weissbrod, O., Zeevi, D., Rothschild, D., Leviatan, S., Kosower, N., Lotan-Pompan, M., Weinberger, A., Le Roy, C. I., et al. (2020). A reference map of potential determinants for the human serum metabolome. Nature *588*, 135–140. https://doi.org/10.1038/s41586-020-2896-2.

257. Reiman, D., Layden, B. T., and Dai, Y. (2021). MiMeNet: Exploring microbiome-metabolome relationships using neural networks. PLoS Computational Biology *17*, e1009021. https://doi.org/10.1371/journal.pcbi.1009021.

258. Wang, T., Wang, X.-W., Lee-Sarwar, K. A., Litonjua, A. A., Weiss, S. T., Sun, Y., Maslov, S., and Liu, Y.-Y. (2023c). Predicting metabolomic profiles from microbial composition through neural ordinary differential equations. Nature Machine Intelligence *5*, 284–293. https://doi.org/10.1038/s42256-023-00627-3.

259. Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., et al. (2015). Personalized Nutrition by Prediction of Glycemic Responses. Cell *163*, 1079–1094. https://doi.org/10.1016/j.cell.2015.11.001.

260. Rein, M., Ben-Yacov, O., Godneva, A., Shilo, S., Zmora, N., Kolobkov, D., Cohen-Dolev, N., Wolf, B.-C., Kosower, N., Lotan-Pompan, M., et al. (2022). Effects of personalized diets by prediction of glycemic responses on glycemic control and metabolic health in newly diagnosed T2DM: a randomized dietary intervention pilot trial. BMC Medicine *20*, 56. https://doi.org/10.1186/s12916-022-02254-y.

261. Wang, T., Holscher, H. D., Maslov, S., Hu, F. B., Weiss, S. T., and Liu, Y.-Y. (2023d). Predicting metabolic response to dietary intervention using deep learning. bioRxiv, 2023–03. https://doi.org/10.1101/2023.03.14.532589.

262. Hu, F. B. and Willett, W. C. (2002). Optimal diets for prevention of coronary heart disease. JAMA *288*, 2569–2578. https://doi.org/10.1001/jama.288.20.2569.

263. Afshin, A., Sur, P. J., Fay, K. A., Cornaby, L., Ferrara, G., Salama, J. S., Mullany, E. C., Abate, K. H., Abbafati, C., Abebe, Z., et al. (2019). Health effects of dietary risks in 195 countries, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet *393*, 1958–1972. https://doi.org/10.1016/S0140-6736(19)30041-8.

264. McNutt, S., Zimmerman, T. P., and Hull, S. G. (2008). Development of food composition databases for food frequency questionnaires (FFQ). Journal of Food Composition and Analysis *21*, S20–S26. https://doi.org/10.1016/j.jfca.2007.05.007.

265. Sharpe, I., Kirkpatrick, S. I., Smith, B. T., Keown-Stoneman, C. D., Omand, J., Vanderhout, S., Maguire, J. L., Birken, C. S., Anderson, L. N., and collaboration, T. K. (2021). Automated Self-Administered 24-H Dietary Assessment Tool (ASA24) recalls for parent proxy-reporting of children's intake (> 4 years of age): a feasibility study. Pilot and Feasibility Studies *7*, 1–10. https://doi.org/10.21203/rs.3.rs-332425/v1.

266. Hebert, J. R., Ockene, I. S., Hurley, T. G., Luippold, R., Well, A. D., Harmatz, M. G., et al. (1997). Development and testing of a seven-day dietary recall. Journal of Clinical Epidemiology *50*, 925–937. https://doi.org/10.1016/s0895-4356(97)00098-x.

267. Westerterp, K. R. and Goris, A. H. (2002). Validity of the assessment of dietary intake: problems of misreporting. Current Opinion in Clinical Nutrition & Metabolic Care *5*, 489–493. https://doi.org/10.1097/00075197-200209000-00006.

268. Rosner, B., Willett, W., and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Statistics in Medicine *8*, 1051–1069. https://doi.org/10.1002/sim.4780080905.

269. Spiegelman, D., McDermott, A., and Rosner, B. (1997). Regression calibration method for correcting measurement-error bias in nutritional epidemiology. The American Journal of Clinical Nutrition *65*, 1179S–1186S. https://doi.org/10.1093/ajcn/65.4.1179s.

270. Hu, F. B., Stampfer, M. J., Rimm, E., Ascherio, A., Rosner, B. A., Spiegelman, D., and Willett, W. C. (1999). Dietary fat and coronary heart disease: a comparison of approaches for adjusting for total energy intake and modeling repeated dietary measurements. American Journal of Epidemiology *149*, 531–540. https://doi.org/10.1093/oxfordjournals.aje.a009849.

271. Wang, T., Fu, Y., Shuai, M., Zheng, J.-S., Zhu, L., Chan, A. T., Sun, Q., Hu, F. B., Weiss, S. T., and Liu, Y.-Y. (2024d). Microbiome-based correction for random errors in nutrient profiles derived from self-reported dietary assessments. Nature Communications *15*, 9112. https://doi.org/10.1101/2023.11.21.568102.

272. Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., and Aila, T. (2018). Noise2Noise: learning image restoration without clean data. Proc. 35th International Conference on Machine Learning, 2965–2974.

273. Letertre, M. P., Dervilly, G., and Giraudeau, P. (2020). Combined nuclear magnetic resonance spectroscopy and mass spectrometry approaches for metabolomics. Analytical Chemistry *93*, 500–518. https://doi.org/10.1021/acs.analchem.0c04371.

274. Alseekh, S., Aharoni, A., Brotman, Y., Contrepois, K., D'Auria, J., Ewald, J., C. Ewald, J., Fraser, P. D., Giavalisco, P., Hall, R. D., et al. (2021). Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. Nature Methods *18*, 747–756. https://doi.org/10.1038/s41592-021-01197-1.

275. Mallick, H., Franzosa, E. A., Mclver, L. J., Banerjee, S., Sirota-Madi, A., Kostic, A. D., Clish, C. B., Vlamakis, H., Xavier, R. J., and Huttenhower, C. (2019). Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. Nature Communications *10*, 3136. https://doi.org/10.1038/s41467-019-10927-1.

276. Le, V., Quinn, T. P., Tran, T., and Venkatesh, S. (2020). Deep in the bowel: highly interpretable neural encoder-decoder networks predict gut metabolites from gut microbiome. BMC genomics *21*, 1–15. https://doi.org/10.1186/s12864-020-6652-7.

277. Salathé, M., Singh, R., and Toumi, M. (2024). Personalized glucose prediction using in situ data only. https://doi.org/10.21203/rs.3.rs-4252145/v1.

278. Li, J. and Fernando, C. (2016). Smartphone-based personalized blood glucose prediction. ICT Express *2*, 150–154. https://doi.org/10.1016/j.icte.2016.10.001.

279. Cheng, M., Diao, X., Zhou, Z., Cui, Y., Liu, W., and Cheng, S. (2024). Toward Short-Term Glucose Prediction Solely Based on CGM Time Series. arXiv preprint arXiv:2404.11924. https://doi.org/10.48550/arXiv.2404.11924.

280. Kim, D.-Y., Choi, D.-S., Kim, J., Chun, S. W., Gil, H.-W., Cho, N.-J., Kang, A. R., and Woo, J. (2020). Developing an individual glucose prediction model using recurrent neural network. Sensors *20*, 6460. https://doi.org/10.3390/s20226460.

281. Lutsker, G., Sapir, G., Godneva, A., Shilo, S., Greenfield, J. R., Samocha-Bonet, D., Mannor, S., Meirom, E., Chechik, G., Rossman, H., et al. (2024). From glucose patterns to health outcomes: A generalizable foundation model for continuous glucose monitor data analysis. arXiv preprint arXiv:2408.11876. https://doi.org/10.48550/arXiv.2408.11876.

282. Albers, D. J., Levine, M., Gluckman, B., Ginsberg, H., Hripcsak, G., and Mamykina, L. (2017). Personalized glucose forecasting for type 2 diabetes using data assimilation. PLoS Computational Biology *13*, e1005232. https://doi.org/10.1371/journal.pcbi.1005232.

283. Neumann, A., Zghal, Y., Cremona, M. A., Hajji, A., Morin, M., and Rekik, M. (2024). A Data-Driven Personalized Approach to Predict Blood Glucose Levels in Type-1 Diabetes Patients Exercising in Free-Living Conditions. Available at SSRN 4777350. https://doi.org/10.2139/ssrn.4777350.

284. Ramesh, H., Elshinawy, A., Ahmed, A., Kassoumeh, M. A., Khan, M., and Mounsef, J. (2024). “BIOINTEL: Real-Time Bacteria Identification Using Microscopy Imaging”. *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 1–4. https://doi.org/10.1109/ISBI56570.2024.10635473.

285. Hallström, E., Kandavalli, V., Ranefall, P., Elf, J., and Wählby, C. (2023). Label-free deep learning-based species classification of bacteria imaged by phase-contrast microscopy. PLoS Computational Biology *19*, e1011181. https://doi.org/10.1371/journal.pcbi.1011181.

286. Wang, L., Tang, J.-W., Li, F., Usman, M., Wu, C.-Y., Liu, Q.-H., Kang, H.-Q., Liu, W., and Gu, B. (2022b). Identification of bacterial pathogens at genus and species levels through combination of Raman spectrometry and deep-learning algorithms. Microbiology Spectrum *10*, e02580–22. https://doi.org/10.1128/spectrum.02580-22.

287. Rahman, M. H.-U., Sikder, R., Tripathi, M., Zahan, M., Ye, T., Gnimpieba Z, E., Jasthi, B. K., Dalton, A. B., and Gadhamshetty, V. (2024). Machine learning-assisted raman

spectroscopy and SERS for bacterial pathogen detection: clinical, food safety, and environmental applications. Chemosensors *12*, 140. https://doi.org/10.3390/chemosensors12070140.

288. Fend, R., Kolk, A. H., Bessant, C., Buijtels, P., Klatser, P. R., and Woodman, A. C. (2006). Prospects for clinical application of electronic-nose technology to early detection of Mycobacterium tuberculosis in culture and sputum. Journal of Clinical Microbiology *44*, 2039–2045. https://doi.org/10.1128/JCM.01591-05.

289. Khaledi, A., Weimann, A., Schniederjans, M., Asgari, E., Kuo, T.-H., Oliver, A., Cabot, G., Kola, A., Gastmeier, P., Hogardt, M., et al. (2020). Predicting antimicrobial resistance in Pseudomonas aeruginosa with machine learning-enabled molecular diagnostics. EMBO Molecular Medicine *12*, e10264. https://doi.org/10.15252/emmm.201910264.

290. Bhattacharyya, R. P., Bandyopadhyay, N., Ma, P., Son, S. S., Liu, J., He, L. L., Wu, L., Khafizov, R., Boykin, R., Cerqueira, G. C., et al. (2019). Simultaneous detection of genotype and phenotype enables rapid and accurate antibiotic susceptibility determination. Nature Medicine *25*, 1858–1864. https://doi.org/10.1038/s41591-019-0650-9.

291. Pataki, B. Á., Matamoros, S., Putten, B. C. van der, Remondini, D., Giampieri, E., Aytan-Aktug, D., Hendriksen, R. S., Lund, O., Csabai, I., Schultsz, C., et al. (2020). Understanding and predicting ciprofloxacin minimum inhibitory concentration in Escherichia coli with machine learning. Scientific Reports *10*, 15026. https://doi.org/10.1038/s41598-020-71693-5.

292. Gumbo, T., Chigutsa, E., Pasipanodya, J., Visser, M., Helden, P. D. van, Sirgel, F. A., and McIlleron, H. (2014). The pyrazinamide susceptibility breakpoint above which combination therapy fails. Journal of Antimicrobial Chemotherapy *69*, 2420–2425. https://doi.org/10.1093/jac/dku136.

293. Shim, H. (2019). Feature learning of virus genome evolution with the nucleotide skip-gram neural network. Evolutionary Bioinformatics *15*, 1176934318821072. https://doi.org/10.1177/1176934318821072.

294. Wang, D. and Larder, B. (2003). Enhanced prediction of lopinavir resistance from genotype by use of artificial neural networks. The Journal of Infectious Diseases *188*, 653–660. https://doi.org/10.1086/377453.

295. Kodogiannis, V. S., Lygouras, J. N., Tarczynski, A., and Chowdrey, H. S. (2008). Artificial odor discrimination system using electronic nose and neural networks for the identification of urinary tract infection. IEEE Transactions on Information Technology in Biomedicine *12*, 707–713. https://doi.org/10.1109/TITB.2008.917928.

296. Mohamed, E., Mohamed, M., Moustafa, M., Abdel-Mageed, S., Moro, A., Baess, A., and El-Kholy, S. (2017). Qualitative analysis of biological tuberculosis samples by an electronic nose-based artificial neural network. The International Journal of Tuberculosis and Lung Disease *21*, 810–817. https://doi.org/10.5588/ijtld.16.0677.

297. He, J., Zhong, R., Xue, L., Wang, Y., Chen, Y., Xiong, Z., Yang, Z., Chen, S., Liang, W., and He, J. (2024). Exhaled Volatile Organic Compounds Detection in Pneumonia Screening: A Comprehensive Meta-analysis. Lung *202*, 501–511. https://doi.org/10.1007/s00408-024-00737-8.

298. Geffen, W. H. van, Bruins, M., and Kerstjens, H. A. (2016). Diagnosing viral and bacterial respiratory infections in acute COPD exacerbations by an electronic nose: a pilot study. Journal of breath research *10*, 036001. https://doi.org/10.1088/1752-7155/10/3/036001.

299. Lynch, S. V. and Pedersen, O. (2016). The human intestinal microbiome in health and disease. New England Journal of Medicine *375*, 2369–2379. https://doi.org/10.1056/NEJMra1600266.

300. Cryan, J. F., O'Riordan, K. J., Cowan, C. S., Sandhu, K. V., Bastiaanssen, T. F., Boehme, M., Codagnone, M. G., Cussotto, S., Fulling, C., Golubeva, A. V., et al. (2019). The microbiota-gut-brain axis. Physiological Reviews. https://doi.org/10.1152/physrev.00018.2018.

301. Schubert, A. M., Rogers, M. A., Ring, C., Mogle, J., Petrosino, J. P., Young, V. B., Aronoff, D. M., and Schloss, P. D. (2014). Microbiome data distinguish patients with Clostridium difficile infection and non-C. difficile-associated diarrhea from healthy controls. MBio *5*, 10–1128. https://doi.org/10.1128/mbio.01021-14.

302. Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., Reyes, J. A., Shah, S. A., LeLeiko, N., Snapper, S. B., et al. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biology *13*, 1–18. https://doi.org/10.1186/gb-2012-13-9-r79.

303. Enck, P., Aziz, Q., Barbara, G., Farmer, A. D., Fukudo, S., Mayer, E. A., Niesler, B., Quigley, E. M. M., Rajilić-Stojanović, M., Schemann, M., et al. (24, 2016). Irritable bowel syndrome. Nature Reviews Disease Primers *2*, 16014. https://doi.org/10.1038/nrdp.2016.14.

304. Kang, D.-W., Park, J. G., Ilhan, Z. E., Wallstrom, G., LaBaer, J., Adams, J. B., and Krajmalnik-Brown, R. (2013). Reduced incidence of Prevotella and other fermenters in intestinal microflora of autistic children. PLoS One *8*, e68322. https://doi.org/10.1371/journal.pone.0068322.

305. Liu, J., Lee, J., Hernandez, M. A. S., Mazitschek, R., and Ozcan, U. (2015). Treatment of obesity with celastrol. Cell *161*, 999–1011. https://doi.org/10.1016/j.cell.2015.05.011.

306. Jangi, S., Gandhi, R., Cox, L. M., Li, N., Von Glehn, F., Yan, R., Patel, B., Mazzola, M. A., Liu, S., Glanz, B. L., et al. (2016). Alterations of the human gut microbiome in multiple sclerosis. Nature Communications *7*, 12015. https://doi.org/10.1038/ncomms12015.

307. Kindt, A., Liebisch, G., Clavel, T., Haller, D., Hörmannsperger, G., Yoon, H., Kolmeder, D., Sigruener, A., Krautbauer, S., Seeliger, C., et al. (2018). The gut microbiota promotes hepatic fatty acid desaturation and elongation in mice. Nature Communications *9*, 3760. https://doi.org/10.1038/s41467-018-05767-4.

308. Scheperjans, F., Aho, V., Pereira, P. A., Koskinen, K., Paulin, L., Pekkonen, E., Haapaniemi, E., Kaakkola, S., Eerola-Rautio, J., Pohja, M., et al. (2015). Gut microbiota are related to Parkinson's disease and clinical phenotype. Movement Disorders *30*, 350–358. https://doi.org/10.1002/mds.26069.

309. Wang, X.-W. and Liu, Y.-Y. (2020). Comparative study of classifiers for human microbiome data. Medicine in Microecology *4*, 100013. https://doi.org/10.1016/j.medmic.2020.100013.

310. Wang, X.-W., Wang, T., Schaub, D. P., Chen, C., Sun, Z., Ke, S., Hecker, J., Maaser-Hecker, A., Zeleznik, O. A., Zeleznik, R., et al. (2023e). Benchmarking omics-based prediction of asthma development in children. Respiratory Research *24*, 63. https://doi.org/10.1186/s12931-023-02368-8.

311. Fioravanti, D., Giarratano, Y., Maggio, V., Agostinelli, C., Chierici, M., Jurman, G., and Furlanello, C. (2018). Phylogenetic convolutional neural networks in metagenomics. BMC Bioinformatics *19*, 49. https://doi.org/10.1186/s12859-018-2033-5.

312. Reiman, D., Metwally, A. A., Sun, J., and Dai, Y. (2020). PopPhy-CNN: a phylogenetic tree embedded architecture for convolutional neural networks to predict host phenotype from metagenomic data. IEEE journal of biomedical and health informatics *24*, 2993–3001. https://doi.org/10.1109/JBHI.2020.2993761.

313. Sharma, D., Paterson, A. D., and Xu, W. (2020). TaxoNN: ensemble of neural networks on stratified microbiome data for disease prediction. Bioinformatics *36*, 4544–4550. https://doi.org/10.1093/bioinformatics/btaa542.

314. Wang, Y., Bhattacharya, T., Jiang, Y., Qin, X., Wang, Y., Liu, Y., Saykin, A. J., and Chen, L. (2021a). A novel deep learning method for predictive modeling of microbiome data. Briefings in Bioinformatics *22*, bbaa073. https://doi.org/10.1093/bib/bbaa073.

315. Liao, H., Shang, J., and Sun, Y. (2023). GDmicro: classifying host disease status with GCN and deep adaptation network based on the human gut microbiome data. Bioinformatics *39*, btad747. https://doi.org/10.1093/bioinformatics/btad747.

316. Pope, Q., Varma, R., Tataru, C., David, M., and Fern, X. (2023). Learning a deep language model for microbiomes: the power of large scale unlabeled microbiome data. bioRxiv, 2023–07. https://doi.org/10.1101/2023.07.17.549267.

317. Oh, M. and Zhang, L. (2020). DeepMicro: deep representation learning for disease prediction based on microbiome data. Scientific Reports *10*, 6026. https://doi.org/10.1038/s41598-020-63159-5.

318. Van Engelen, J. E. and Hoos, H. H. (2020). A survey on semi-supervised learning. Machine Learning *109*, 373–440. https://doi.org/10.1007/s10994-019-05855-6.

319. Long, M., Cao, Y., Wang, J., and Jordan, M. (2015). “Learning transferable features with deep adaptation networks”. *International conference on Machine Learning*. PMLR, 97–105.

320. Lee, S. J. and Rho, M. (2022). Multimodal deep learning applied to classify healthy and disease states of human microbiome. Scientific Reports *12*, 824. https://doi.org/10.1038/s41598-022-04773-3.

321. Wang, T., Shao, W., Huang, Z., Tang, H., Zhang, J., Ding, Z., and Huang, K. (2021b). MOGONET integrates multi-omics data using graph convolutional networks allowing pa-

tient classification and biomarker identification. Nature Communications *12*, 3445. https://doi.org/10.1038/s41467-021-23774-w.

322. Ding, D. Y., Li, S., Narasimhan, B., and Tibshirani, R. (2022). Cooperative learning for multiview analysis. Proceedings of the National Academy of Sciences *119*, e2202113119. https://doi.org/10.1073/pnas.2202113119.

323. Meqdad, M. N., Husain, S. O., Jawad, A. M., Kadry, S., and Khekan, A. R. (2023). Classification of electroencephalography using cooperative learning based on participating client balancing. International Journal of Electrical & Computer Engineering (2088-8708) *13*. https://doi.org/10.11591/ijece.v13i4.pp4692-4699.

324. Ferjani, R., Rejeb, L., and Said, L. B. (2020). "Cooperative reinforcement multi-agent learning system for sleep stages classification". *2020 International Multi-Conference on:"Organization of Knowledge and Advanced Technologies"(OCTA)*. IEEE, 1–8. https://doi.org/10.1109/octa49274.2020.9151700.

325. Huan, Y., Kong, Q., Mou, H., and Yi, H. (2020). Antimicrobial peptides: classification, design, application and research progress in multiple fields. Frontiers in Microbiology *11*, 582779. https://doi.org/10.3389/fmicb.2020.582779.

326. Lata, S., Sharma, B., and Raghava, G. P. (2007). Analysis and prediction of antibacterial peptides. BMC Bioinformatics *8*, 1–10. https://doi.org/10.1186/1471-2105-8-263.

327. Torrent, M., Andreu, D., Nogués, V. M., and Boix, E. (2011). Connecting peptide physicochemical and antimicrobial properties by a rational prediction model. PLoS One *6*, e16968. https://doi.org/10.1371/journal.pone.0016968.

328. Veltri, D., Kamath, U., and Shehu, A. (2018). Deep learning improves antimicrobial peptide recognition. Bioinformatics *34*, 2740–2747. https://doi.org/10.1093/bioinformatics/bty179.

329. Tang, W., Dai, R., Yan, W., Zhang, W., Bin, Y., Xia, E., and Xia, J. (2022). Identifying multi-functional bioactive peptide functions using multi-label deep learning. Briefings in Bioinformatics *23*, bbab414. https://doi.org/10.1093/bib/bbab414.

330. Ma, Y., Guo, Z., Xia, B., Zhang, Y., Liu, X., Yu, Y., Tang, N., Tong, X., Wang, M., Ye, X., et al. (2022). Identification of antimicrobial peptides from the human gut microbiome using deep learning. Nature Biotechnology *40*, 921–931. https://doi.org/10.1038/s41587-022-01230-4.

331. Van Oort, C. M., Ferrell, J. B., Remington, J. M., Wshah, S., and Li, J. (2021). AMPGAN v2: machine learning-guided design of antimicrobial peptides. Journal of chemical information and modeling *61*, 2198–2207. https://doi.org/10.1021/acs.jcim.0c01441.

332. Dean, S. N., Alvarez, J. A. E., Zabetakis, D., Walper, S. A., and Malanoski, A. P. (2021). PepVAE: variational autoencoder framework for antimicrobial peptide generation and activity prediction. Frontiers in Microbiology *12*, 725727. https://doi.org/10.3389/fmicb.2021.725727.

333. Szymczak, P., Możejko, M., Grzegorzek, T., Jurczak, R., Bauer, M., Neubauer, D., Sikora, K., Michalski, M., Sroka, J., Setny, P., et al. (2023). Discovering highly potent antimicrobial peptides with deep generative model HydrAMP. Nature Communications *14*, 1453. https://doi.org/10.1038/s41467-023-36994-z.

334. Sun, Y., Li, H., Zheng, L., Li, J., Hong, Y., Liang, P., Kwok, L.-Y., Zuo, Y., Zhang, W., and Zhang, H. (2022). iProbiotics: a machine learning platform for rapid identification of probiotic properties from whole-genome primary sequences. Briefings in Bioinformatics *23*, bbab477. https://doi.org/10.1093/bib/bbab477.

335. Wu, S., Feng, T., Tang, W., Qi, C., Gao, J., He, X., Wang, J., Zhou, H., and Fang, Z. (2024c). metaProbiotics: a tool for mining probiotic from metagenomic binning data based on a language model. Briefings in Bioinformatics *25*, bbae085. https://doi.org/10.1093/bib/bbae085.

336. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). “Neural message passing for quantum chemistry”. *International conference on machine learning*. PMLR, 1263–1272.

337. Dai, H., Dai, B., and Song, L. (2016). “Discriminative embeddings of latent variable models for structured data”. *International conference on machine learning*. PMLR, 2702–2711.

338. Heid, E., Greenman, K. P., Chung, Y., Li, S.-C., Graff, D. E., Vermeire, F. H., Wu, H., Green, W. H., and McGill, C. J. (2023). Chemprop: A machine learning package for chemical property prediction. Journal of Chemical Information and Modeling *64*, 9–17. https://doi.org/10.1021/acs.jcim.3c01250.

339. Wong, F., Zheng, E. J., Valeri, J. A., Donghia, N. M., Anahtar, M. N., Omori, S., Li, A., Cubillos-Ruiz, A., Krishnan, A., Jin, W., et al. (2023). Discovery of a structural class of antibiotics with explainable deep learning. Nature. https://doi.org/10.1038/s41586-023-06887-8.

340. Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., et al. (2020). A deep learning approach to antibiotic discovery. Cell *180*, 688–702. https://doi.org/10.1016/j.cell.2020.01.02.

341. Bento, A. P., Hersey, A., Félix, E., Landrum, G., Gaulton, A., Atkinson, F., Bellis, L. J., De Veij, M., and Leach, A. R. (2020). An open source chemical structure curation pipeline using RDKit. Journal of Cheminformatics *12*, 1–16. https://doi.org/10.21203/rs.3.rs-34715/v2.

342. Wong, F., Zheng, E. J., Valeri, J. A., Donghia, N. M., Anahtar, M. N., Omori, S., Li, A., Cubillos-Ruiz, A., Krishnan, A., Jin, W., et al. (2024). Discovery of a structural class of antibiotics with explainable deep learning. Nature *626*, 177–185. https://doi.org/10.1038/s41586-023-06887-8.

343. Schooley, R. T., Biswas, B., Gill, J. J., Hernandez-Morales, A., Lancaster, J., Lessor, L., Barr, J. J., Reed, S. L., Rohwer, F., Benler, S., et al. (2017). Development and use of personalized bacteriophage-based therapeutic cocktails to treat a patient with a dissemi-

nated resistant Acinetobacter baumannii infection. Antimicrobial Agents and Chemotherapy *61*, 10–1128. https://doi.org/10.1128/AAC.00954-17.

344. Pirnay, J.-P., Djebara, S., Steurs, G., Griselain, J., Cochez, C., De Soir, S., Glonti, T., Spiessens, A., Vanden Berghe, E., Green, S., et al. (2024). Personalized bacteriophage therapy outcomes for 100 consecutive cases: a multicentre, multinational, retrospective observational study. Nature Microbiology, 1–20. https://doi.org/10.1038/s41564-024-01705-x.

345. Green, S. I., Clark, J. R., Santos, H. H., Weesner, K. E., Salazar, K. C., Aslam, S., Campbell, J. W., Doernberg, S. B., Blodget, E., Morris, M. I., et al. (2023). A retrospective, observational study of 12 cases of expanded-access customized phage therapy: production, characteristics, and clinical outcomes. Clinical Infectious Diseases *77*, 1079–1091. https://doi.org/10.1093/cid/ciad335.

346. Chen, G., Tang, X., Shi, M., and Sun, Y. (2023). VirBot: an RNA viral contig detector for metagenomic data. Bioinformatics *39*, btad093. https://doi.org/10.1093/bioinformatics/btad093.

347. Ho, S. F. S., Wheeler, N. E., Millard, A. D., and Schaik, W. van (2023). Gauge your phage: benchmarking of bacteriophage identification tools in metagenomic sequencing data. Microbiome *11*, 84. https://doi.org/10.1186/s40168-023-01533-x.

348. Jurtz, V. I., Villarroel, J., Lund, O., Voldby Larsen, M., and Nielsen, M. (2016). MetaPhinder—identifying bacteriophage sequences in metagenomic data sets. PLoS One *11*, e0163111. https://doi.org/10.1371/journal.pone.0163111.

349. Kieft, K., Zhou, Z., and Anantharaman, K. (2020). VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome *8*, 1–23. https://doi.org/10.1186/s40168-020-00867-0.

350. Guo, J., Bolduc, B., Zayed, A. A., Varsani, A., Dominguez-Huerta, G., Delmont, T. O., Pratama, A. A., Gazitúa, M. C., Vik, D., Sullivan, M. B., et al. (2021). VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome *9*, 1–13. https://doi.org/10.1186/s40168-020-00990-y.

351. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., and Sun, F. (2017). VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome *5*, 1–20. https://doi.org/10.1186/s40168-017-0283-5.

352. Auslander, N., Gussow, A. B., Benler, S., Wolf, Y. I., and Koonin, E. V. (2020). Seeker: alignment-free identification of bacteriophage genomes by deep learning. Nucleic Acids Research *48*, e121–e121. https://doi.org/10.1093/nar/gkaa856.

353. Ren, J., Song, K., Deng, C., Ahlgren, N. A., Fuhrman, J. A., Li, Y., Xie, X., Poplin, R., and Sun, F. (2020). Identifying viruses from metagenomic data using deep learning. Quantitative Biology *8*, 64–77. https://doi.org/10.1007/s40484-019-0187-4.

354. Shang, J., Tang, X., Guo, R., and Sun, Y. (2022). Accurate identification of bacteriophages from metagenomic data using Transformer. Briefings in Bioinformatics *23*, bbac258. https://doi.org/10.1093/bib/bbac258.

355. Bai, Z., Zhang, Y.-z., Miyano, S., Yamaguchi, R., Fujimoto, K., Uematsu, S., and Imoto, S. (2022). Identification of bacteriophage genome sequences with representation learning. Bioinformatics *38*, 4264–4270. https://doi.org/10.1093/bioinformatics/btac509.

356. McNair, K., Bailey, B. A., and Edwards, R. A. (2012). PHACTS, a computational approach to classifying the lifestyle of phages. Bioinformatics *28*, 614–618. https://doi.org/10.1093/bioinformatics/bts014.

357. Hockenberry, A. J. and Wilke, C. O. (2021). BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. PeerJ *9*, e11396. https://doi.org/10.7717/peerj.11396.

358. Wu, S., Fang, Z., Tan, J., Li, M., Wang, C., Guo, Q., Xu, C., Jiang, X., and Zhu, H. (2021b). DeePhage: distinguishing virulent and temperate phage-derived sequences in metavirome data with a deep learning approach. GigaScience *10*, giab056. https://doi.org/10.1093/gigascience/giab056.

359. Shang, J., Tang, X., and Sun, Y. (2023). PhaTYP: predicting the lifestyle for bacteriophages using BERT. Briefings in Bioinformatics *24*, bbac487. https://doi.org/10.1093/bib/bbac487.

360. Miao, Y., Sun, Z., Lin, C., Gu, H., Ma, C., Liang, Y., and Wang, G. (2024b). DeePhafier: a phage lifestyle classifier using a multilayer self-attention neural network combining protein information. Briefings in Bioinformatics *25*. https://doi.org/10.1093/bib/bbae377.

361. Nie, W., Qiu, T., Wei, Y., Ding, H., Guo, Z., and Qiu, J. (2024). Advances in phage–host interaction prediction: in silico method enhances the development of phage therapies. Briefings in Bioinformatics *25*, bbae117. https://doi.org/10.1093/bib/bbae117.

362. Swan, B. K., Tupper, B., Sczyrba, A., Lauro, F. M., Martinez-Garcia, M., González, J. M., Luo, H., Wright, J. J., Landry, Z. C., Hanson, N. W., et al. (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. Proceedings of the National Academy of Sciences *110*, 11463–11468. https://doi.org/10.1073/pnas.1304246110.

363. Li, M. and Zhang, W. (2022). PHIAF: prediction of phage-host interactions with GAN-based data augmentation and sequence-based feature fusion. Briefings in Bioinformatics *23*, bbab348. https://doi.org/10.1093/bib/bbab348.

364. Yang, Y., Dufault-Thompson, K., Yan, W., Cai, T., Xie, L., and Jiang, X. (2024b). Large-scale genomic survey with deep learning-based method reveals strain-level phage specificity determinants. GigaScience *13*, giae017. https://doi.org/10.1093/gigascience/giae017.

365. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science *379*, 1123–1130. https://doi.org/10.1126/science.ade2574.

366. Cook, R., Brown, N., Redgwell, T., Rihtman, B., Barnes, M., Clokie, M., Stekel, D. J., Hobman, J., Jones, M. A., and Millard, A. (2021). INfrastructure for a PHAge REference

database: identification of large-scale biases in the current collection of cultured phage genomes. Phage *2*, 214–223. https://doi.org/10.1089/phage.2021.0007.

367. Kabir, M., Nantasenamat, C., Kanthawong, S., Charoenkwan, P., and Shoombuatong, W. (2022). Large-scale comparative review and assessment of computational methods for phage virion proteins identification. EXCLI journal *21*, 11. https://doi.org/10.17179/excli2021-4411.

368. Cantu, V. A., Salamon, P., Seguritan, V., Redfield, J., Salamon, D., Edwards, R. A., and Segall, A. M. (2020). PhANNs, a fast and accurate tool and web server to classify phage structural proteins. PLoS Computational Biology *16*, e1007845. https://doi.org/10.1371/journal.pcbi.1007845.

369. Fang, Z. and Zhou, H. (2021). VirionFinder: identification of complete and partial prokaryote virus virion protein from virome data using the sequence and biochemical properties of amino acids. Frontiers in Microbiology *12*, 615711. https://doi.org/10.3389/fmicb.2021.615711.

370. Fang, Z., Feng, T., Zhou, H., and Chen, M. (2022). DeePVP: Identification and classification of phage virion proteins using deep learning. Gigascience *11*, giac076. https://doi.org/10.1093/gigascience/giac076.

371. Shang, J., Peng, C., Tang, X., and Sun, Y. (2023). PhaVIP: Phage VIrion Protein classification based on chaos game representation and Vision Transformer. Bioinformatics *39*, i30–i39. https://doi.org/10.1093/bioinformatics/btad229.

372. Li, B. and Liang, G. (2023). ESM-PVP: Identification and classification of phage virion proteins with a large pretrained protein language model and an MLP neural network. bioRxiv, 2023–12. https://doi.org/10.1101/2023.12.29.573676.

373. Flamholz, Z. N., Biller, S. J., and Kelly, L. (2024). Large language models improve annotation of prokaryotic viral proteins. Nature Microbiology *9*, 537–549. https://doi.org/10.1038/s41564-023-01584-8.

374. Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929.

375. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., and Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks? Advances in neural information processing systems *34*, 12116–12128.

376. Robson, E., Xu, C., and Wills, L. W. (2022). “ProSE: the architecture and design of a protein discovery engine”. *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, 655–668. https://doi.org/10.1145/3503222.3507722.

377. Schmelcher, M. and Loessner, M. J. (2021). Bacteriophage endolysins—extending their application to tissues and the bloodstream. Current Opinion in Biotechnology *68*, 51–59. https://doi.org/10.1016/j.copbio.2020.09.012.

378. Zhang, Y., Li, R., Zou, G., Guo, Y., Wu, R., Zhou, Y., Chen, H., Zhou, R., Lavigne, R., Bergen, P. J., et al. (2024b). Discovery of Antimicrobial Lysins from the "Dark Matter" of Uncharacterized Phages Using Artificial Intelligence. Advanced Science, 2404049. https://doi.org/10.1002/advs.202404049.

379. Fu, Y., Yu, S., Li, J., Lao, Z., Yang, X., and Lin, Z. (2024). DeepMineLys: Deep mining of phage lysins from human microbiome. Cell Reports *43*. https://doi.org/10.1016/j.celrep.2024.114583.

380. Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., and Song, Y. (2019). Evaluating protein transfer learning with TAPE. Advances in neural information processing systems *32*.

381. Vázquez, R., Blanco-Gañán, S., Ruiz, S., and García, P. (2021). Mining of Gram-negative surface-active enzybiotic candidates by sequence-based calculation of physicochemical properties. Frontiers in Microbiology *12*, 660403. https://doi.org/10.3389/fmicb.2021.660403.

382. Pizza, M., Scarlato, V., Masignani, V., Giuliani, M. M., Arico, B., Comanducci, M., Jennings, G. T., Baldi, L., Bartolini, E., Capecchi, B., et al. (2000). Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. Science *287*, 1816–1820. https://doi.org/10.1126/science.287.5459.1816.

383. Dalsass, M., Brozzi, A., Medini, D., and Rappuoli, R. (2019). Comparison of open-source reverse vaccinology programs for bacterial vaccine antigen discovery. Frontiers in Immunology *10*, 113. https://doi.org/10.3389/fimmu.2019.00113.

384. Vivona, S., Bernante, F., and Filippini, F. (2006). NERVE: new enhanced reverse vaccinology environment. BMC Biotechnology *6*, 1–8. https://doi.org/10.1186/1472-6750-6-35.

385. He, Y., Xiang, Z., and Mobley, H. L. (2010). Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. BioMed Research International *2010*, 297505.

386. Doytchinova, I. A. and Flower, D. R. (2007). VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinformatics *8*, 1–7. https://doi.org/10.1186/1471-2105-8-4.

387. Magnan, C. N., Zeller, M., Kayala, M. A., Vigil, A., Randall, A., Felgner, P. L., and Baldi, P. (2010). High-throughput prediction of protein antigenicity using protein microarray data. Bioinformatics *26*, 2936–2943. https://doi.org/10.1093/bioinformatics/btq551.

388. Rahman, M. S., Rahman, M. K., Saha, S., Kaykobad, M., and Rahman, M. S. (2019). Antigenic: an improved prediction model of protective antigens. Artificial intelligence in medicine *94*, 28–41. https://doi.org/10.1016/j.artmed.2018.12.010.

389. Ong, E., Wang, H., Wong, M. U., Seetharaman, M., Valdez, N., and He, Y. (2020). Vaxign-ML: supervised machine learning reverse vaccinology model for improved prediction of bacterial protective antigens. Bioinformatics *36*, 3185–3191. https://doi.org/10.1093/bioinformatics/btaa119.

390. Ong, E., Cooke, M. F., Huffman, A., Xiang, Z., Wong, M. U., Wang, H., Seetharaman, M., Valdez, N., and He, Y. (2021). Vaxign2: the second generation of the first Web-based vaccine design program using reverse vaccinology and machine learning. Nucleic Acids Research *49*, W671–W678. https://doi.org/10.1093/nar/gkab279.

391. Rawal, K., Sinha, R., Nath, S. K., Preeti, P., Kumari, P., Gupta, S., Sharma, T., Strych, U., Hotez, P., and Bottazzi, M. E. (2022). Vaxi-DL: A web-based deep learning server to identify potential vaccine candidates. Computers in Biology and Medicine *145*, 105401. https://doi.org/10.1016/j.compbiomed.2022.105401.

392. Zhang, Y., Huffman, A., Johnson, J., and He, Y. (2023). Vaxign-DL: A Deep Learning-based Method for Vaccine Design and its Evaluation. Biorxiv. https://doi.org/10.1101/2023.11.29.569096.

393. Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems *30*.

394. Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "" Why should i trust you?" Explaining the predictions of any classifier". *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144. https://doi.org/10.18653/v1/n16-3020.

395. Chan, K. H. R., Yu, Y., You, C., Qi, H., Wright, J., and Ma, Y. (2022). ReduNet: A white-box deep network from the principle of maximizing rate reduction. Journal of Machine Learning Research *23*, 1–103.

396. Maringanti, V. S., Bucci, V., and Gerber, G. K. (2022). MDITRE: scalable and interpretable machine learning for predicting host status from temporal microbiome dynamics. Msystems *7*, e00132–22. https://doi.org/10.1128/msystems.00132-22.

397. Chen, B., Hong, J., and Wang, Y. (1997). The minimum feature subset selection problem. Journal of Computer Science and Technology *12*, 145–153. https://doi.org/10.1007/BF02951333.

398. Stańczyk, U. (2015). Feature evaluation by filter, wrapper, and embedded approaches. Feature selection for data and pattern recognition, 29–44. https://doi.org/10.1007/978-3-662-45620-0_3.

399. Chen, C., Weiss, S. T., and Liu, Y.-Y. (2023). Graph convolutional network-based feature selection for high-dimensional and low-sample size data. Bioinformatics *39*, btad135. https://doi.org/10.1093/bioinformatics/btad135.

400. Chakraborty, S., Ghosh, M., and Mallick, B. K. (2012). Bayesian nonlinear regression for large p small n problems. Journal of Multivariate Analysis *108*, 28–40. https://doi.org/10.1016/j.jmva.2012.01.015.

401. Safonova, A., Ghazaryan, G., Stiller, S., Main-Knorn, M., Nendel, C., and Ryo, M. (2023). Ten deep learning techniques to address small data problems with remote sensing. International Journal of Applied Earth Observation and Geoinformation *125*, 103569. https://doi.org/10.1016/j.jag.2023.103569.

402. Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. ACM computing surveys (csur) *53*, 1–34. https://doi.org/10.1145/3386252.

403. Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. (2018). Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE transactions on pattern analysis and machine intelligence *41*, 2251–2265. https://doi.org/10.1109/TPAMI.2018.2857768.

404. Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., and Alm, E. J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nature Communications *8*, 1784. https://doi.org/10.1038/s41467-017-01973-8.

405. Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., Beghini, F., Malik, F., Ramos, M., Dowd, J. B., et al. (2017). Accessible, curated metagenomic data through ExperimentHub. Nature Methods *14*, 1023–1024. https://doi.org/10.1038/nmeth.4468.