Journal Title Here, 2023, pp. 1–12

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

MIOSTONE: Modeling microbiome-trait associations with taxonomy-adaptive neural networks

Yifan Jiang,[1] Matthew Aton,[2] Qiyun Zhu[2,*] and Yang Young Lu[1,*]

[1]Cheriton School of Computer Science, University of Waterloo and [2]School of Life Sciences, Arizona State University
[*]Corresponding author. Qiyun.Zhu@asu.edu and yanglu@uwaterloo.ca

Abstract

Motivation: The human microbiome, a complex ecosystem of microorganisms inhabiting the body, plays a critical role in human health. Investigating its association with host traits is essential for understanding its impact on various diseases. Although shotgun metagenomic sequencing technologies have produced vast amounts of microbiome data, analyzing such data is highly challenging due to its sparsity, noisiness, and high feature dimensionality.
Results: Here we propose MIOSTONE, a novel machine learning method that leverages the intercorrelation of microbiome features due to their phylogeny-based taxonomic relationships. MIOSTONEemploys a novel taxonomy-encoded deep neural network (DNN) architecture that harnesses the capabilities of DNNs with mitigated concerns of overfitting. In addition, MIOSTONE has the ability to determine whether taxa within the corresponding taxonomic group provide a better explanation in a data-driven manner. We empirically assessed MIOSTONE's accuracy and interpretability on seven real microbiome datasets, demonstrating its superior performance and biological interpretability.
Availability: TheApache licensed source code is available at (`https://github.com/batmen-lab/MIOSTONE`).

Introduction

The human microbiome characterizes the complex communities of microorganisms living in and on our bodies, with bacteria alone encoding 100 times more unique genes than humans [30]. As the microbiome influences the impact of host genes, microbiome genomes are often referred to as the "second genome" [9]. Subsequently, microbiomes have been found to play pivotal roles in various aspects of human health and disease [35], including diabetes [31], obesity [41], inflammatory bowel disease [24], Alzheimer's disease [42], cancers [29], and more. For example, the microbiome influences cancer development and therapeutic responses in cancer patients [33]. the association of the microbiome with host traits will provide insights into the underlying mechanisms governing the microbiome's impact on human health and disease and facilitate the development of novel therapeutic strategies.

To explore microbiome-trait associations, various methods have been developed to identify and analyze predictive microbial biomarker signatures [23]. These methods rely on microbiome features, which are represented by the relative abundance profiled from specific microbial taxa across various samples [27]. Nevertheless, these features exhibit unique characteristics that pose challenges in thoroughly exploring the relationships among the taxa in predictive analysis. Firstly, sequencing produces millions of short fragments from a mixture of taxa rather than individual taxa, making accurate profiling challenging due to the presence of low-abundance taxa and ambiguous assignments of fragments to taxa [47]. Secondly, the microbiome features, representing the abundance of hundreds to thousands of taxa, are sparse, high-dimensional, and include inflated zero counts [23]. The excessive number of features presents challenges in analysis due to the curse of dimensionality [17]. Specifically, the relatively low number of samples lead to overfitting during training, thereby limiting generalization to other datasets. For example, the small sample size might lead to conflicting results when inferring the association between microbes and disease states [13, 32]. Hence, there is a need for novel methods to analyze microbiome data, taking into account factors such as data imperfections, sparsity, low signal-to-noise ratio, and the curse of dimensionality.

© The Author 2023. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

To tackle these challenges, recent studies have leveraged the intrinsic phylogenetic correlation among taxa as an informative prior, aiming to improve predictive performance. For example, phylogenetic relationships can serve as a smoothness regularizer to enhance linear regression models [46]. Furthermore, leveraging phylogenetic relationships can assist in weighting and prioritizing the most relevant taxa, thereby enhancing the prediction accuracy of random forest [2]. More recently, several methods have incorporated phylogenetic relationships into the preprocessing stage of employing advanced analysis tools like deep neural networks (DNN) [36, 32, 16, 38, 43]. Specifically, these preprocessing steps involve aggregating taxa into distinct taxonomic clusters [36, 43], reordering the taxa spatially based on their phylogenetic structure [32, 38], and assigning varying weights according to phylogenetic distances [16]. However, these methods may suboptimally utilize phylogenetic relationships by limiting its power to the preprocessing step while relegating the DNN to a black box model. While black box modeling remains useful, it proves insufficient for reasoning about the mechanisms governing dynamic interactions among taxa – an aspect that is critical for a scientific understanding of microbiome-trait associations. From a practitioner's perspective, an ideal predictive model should not only accurately predict microbiome-trait associations but also be interpretable. Both aspects are crucial for facilitating *in silico* investigations into the biological mechanisms underlying such associations among microbial taxa, even in the presence of data imperfections and curse of dimensionality. However, among existing methods, we are unaware of prior attempts to fully harness phylogenetic relationships for both prediction and interpretation.

In this paper, we introduce MIOSTONE (MIcrObiome-trait aSsociations with TaxONomy-adaptivE neural networks), an accurate and interpretable machine learning method for investigating microbiome-trait associations. At its core, MIOSTONE leverages the intercorrelation of microbiome features based on a phylogeny-based taxonomy system, which organizes microbial taxa into a hierarchical tree structure based on their phylogenetic relationships. This taxonomy system spans from the three domains (Bacteria, Archaea, and Eukarya) down to species, using a well-established and widely accepted naming system [26]. MIOSTONE introduces a novel DNN framework that explicitly encodes taxonomic relationships into its architecture, drawing inspiration from biologically-informed DNNs [22, 6]. Specifically, MIOSTONE organizes the neural network into layers to emulate the taxonomic hierarchy, as illustrated in Figure 1. In this network, each neuron represents a specific taxonomic group, with connections between neurons symbolizing the subordination relationships among these groups. More importantly, the novelty of MIOSTONE lies in the unique capability of its internal neurons to determine whether taxa within the corresponding taxonomic group offer a more effective explanation of the trait when considered holistically as a group or individually as distinct taxa. This taxonomy-adaptive strategy significantly reduces the model's complexity, thereby mitigating the curse of dimensionality and overfitting, while also providing a natural interpretation of the model's internal mechanisms. We have applied MIOSTONE to various tasks, demonstrating its empirical utility in three aspects: (1) providing accurate predictions of the host's disease status, (2) improving predictive performance in sample-limited tasks through knowledge transfer, and (3) identifying significant microbiome-disease associations with high interpretability.

Approach

Problem setup

Consider a prediction task involving n sample-label pairs (X, Y), where $X = \{x_i\}_{i=1}^n \in \mathbb{R}^{n\times p}$ represents the data matrix with p-dimensional microbiome features, and $Y = \{y_i\}_{i=1}^n \in \mathbb{R}^{n\times 1}$ denotes the corresponding trait labels. In this context, the feature profiles correspond to the relative abundance of microbial taxa in each sample, which are typically sparse and high-dimensional. The trait labels can be binary (*e.g.,* disease status) or continuous (*e.g.,* age). In this paper, we focus on binary classification tasks, where the trait labels are binary. Additionally, we have phylogenetic relationships among microbial taxa represented as a taxonomy tree, which mathematically corresponds to a hierarchical graph $G = (V_L \cup V_I, E)$. Here, V_L denotes the set of microbial taxa, i.e., the tree's leaf nodes, with $|V_L| = p$. V_I consists of the internal nodes within the taxonomic tree, representing taxonomic groups ranging from domains to species. For each node in the taxonomic tree, denoted as $v \in V_I \cup V_L$, its parent is denoted as $\text{parent}(v) \in V_I$. An undirected edge $e = (u, v) \in E$ connects two nodes $u, v \in V_I \cup V_L$ if and only if $u = \text{parent}(v)$ or $v = \text{parent}(u)$. Similarly, the children of v are defined as $\text{children}(v) = \{u : \text{parent}(u) = v \text{ for } u \in V_I \cup V_L\}$. Since the tree G is acyclic, each taxa $v \in V_L$ can trace its taxonomic lineage from species to domain along a unique path $v, \text{parent}(v), \text{parent}(\text{parent}(v)), \cdots$.

Given X, Y, and G as input, our goal is to learn a predictive function $f : \mathbb{R}^p \mapsto \mathbb{R}$, parameterized by a DNN, that maps a sample $x \in \mathbb{R}^p$ to the trait label $y \in \mathbb{R}$. This task is particularly challenging due to DNN's tendency to overfit when the number of samples is low relative to the high dimensionality of features.

MIOSTONE design

MIOSTONE introduces a novel DNN framework to model microbiome-trait associations, guided by phylogenetic relationships among taxa encoded in a taxonomy tree, as illustrated in Figure 1(A). Specifically, MIOSTONE trains a DNN with an architecture that mirrors the taxonomic hierarchy to model the microbiome-trait associations. Each neuron in the network corresponds to a specific taxonomic group, and the connections between neurons represent the subordination relationships between these groups. Unlike fully connected DNNs, MIOSTONE only

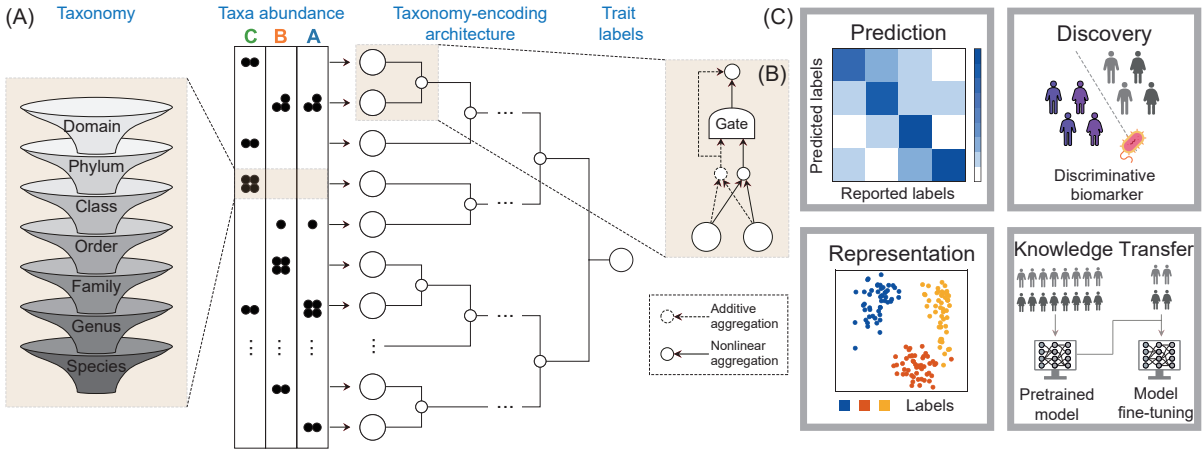

Fig. 1. MIOSTONE employs a DNN encoded with taxonomic relationships to model microbiome-trait associations. (A) A graphical illustration of MIOSTONE architecture. The input microbiome data represents the abundance of microbial taxa, each categorized into hierarchical taxonomic groups ranging from domains to species. The trait is predicted from the microbiome features (*i.e.,* taxa abundance) using a neural network that connects a set of neurons organized in layers to emulate the taxonomic hierarchy. (B) Each internal neuron is able to determine whether taxa within the corresponding taxonomic group provide a better explanation for the label when considered holistically (additively) or individually (non-linearly). (C) MIOSTONE establishes a versatile microbiome data analysis pipeline, applicable to a variety of tasks including disease status prediction, microbiome representation learning, identification of microbiome-disease associations, and enhancement of predictive performance in tasks with limited samples through knowledge transfer.

connects neurons that are directly related in the taxonomic hierarchy. This design choice significantly reduces the model's complexity, thereby mitigating the overfitting problem, while also enhancing the model's interpretability. We implement this design by leveraging pruning techniques [10] to eliminate redundant connections between neurons while preserving its parallelizability.

One challenge in modeling microbiome-trait associations is the ambiguity in fragment-to-taxon assignments. For example, a sequenced viral fragment from the Omicron variant may be mistakenly assigned to the Delta variant, both belonging to the SARS-CoV-2 lineage, but it is unlikely to be assigned to SARS-CoV-1. To address this issue, MIOSTONE employs a data-driven strategy to determine whether taxa within the corresponding taxonomic group provide a better explanation for the label when considered holistically or individually. This strategy aims to balance the reduction of ambiguity in fragment-to-taxon assignments with the effective explanation of the trait of interest. The underlying rationale is that each taxonomic group may exhibit different levels of ambiguity in these assignments, necessitating distinct treatment. High ambiguity suggests that taxa within a group may not be individually meaningful and should be considered collectively, while low ambiguity implies that each taxon may have an individual impact on the trait of interest.

We implement this strategy by introducing a stochastic gate [19] for each internal neuron, as illustrated in Figure 1(B). Specifically, an internal neuron $v \in V_I$ is characterized by two multi-dimensional representations: an additive representation $I_v^A \in \mathbb{R}^{d_v}$ and a nonlinear representation $I_v^N \in \mathbb{R}^{d_v}$, where d_v is the representation dimension. The additive representation I_v^A is obtained by concatenating the additive representations of all children of v and then applying a linear transformation:

$$I_v^A = \text{Linear}\left(\left[I_{u_1}^A, I_{u_2}^A, \cdots\right]\right) \tag{1}$$

where $u_1, u_2, \cdots \in \text{children}(v)$ and $\text{Linear}(\cdot)$ is a linear transformation function. To obtain the nonlinear representation I_v^N, we first concatenate the nonlinear representations of all children of v and then apply a multi-layer perceptron (MLP) to transform it into an intermediate non-linear representation $I_v^M \in \mathbb{R}^{d_v}$:

$$I_v^M = \text{MLP}\left(\left[I_{u_1}^N, I_{u_2}^N, \cdots\right]\right) \tag{2}$$

Then we apply a stochastic gate $m_v \in (0, 1)$ to combine I_v^A and I_v^M into the final nonlinear representation I_v^N:

$$I_v^N = (1 - m_v) \cdot I_v^A + m_v \cdot I_v^M \tag{3}$$

where the gate m_v is based on the hard concrete distribution [19], which is a differentiable relaxation of the Bernoulli distribution:

$$m_v = \sigma\left(\frac{1}{\beta}\left(\log U_v - \log(1 - U_v) + \log \alpha_v\right)\right) \tag{4}$$

where $\sigma(x) = (1 + \exp(-x))^{-1}$ is the sigmoid function and $U_v \sim \text{Uniform}(0, 1)$ an independent random variable following a continuous uniform distribution. This relaxation is parameterized by a trainable parameter α_v and a temperature coefficient $\beta \in (0, 1)$ controlling the degree of approximation. As $\beta \to 0$, the gate m_v converge to a

Bernoulli random variable. We set $\beta = 0.3$ in our experiments. When the gate m_v has a value close to 1 (*i.e.,* in "on" state), all taxa within the group (*e.g.,* X_1 and X_2 in Figure 1(B)) will be selected to contribute to the prediction individually. When the gate m_v has a value close to 0 (*i.e.,* in "off" state), all taxa within the group may not be as individually meaningful as when considered holistically as a group.

Intuitively, larger taxonomic groups should possess a greater representation dimension to capture potentially more complex biological patterns. However, the dimension should not become excessively large, as this might lead each taxonomic group to merely memorize information from its descendants, rather than distill and learn new patterns. Thus, we determine the representation dimension d_v for an internal neuron $v \in V_I$ recursively as $d_v = \alpha^{L_v} \cdot \sum_{u \in \text{children}(v)} d_u$. Here, α is a hyperparameter controlling the shrinkage of the representation dimension and L_v is the taxonomic level of v, starting from $L = 1$ for species and increasing by 1 for each level up to $L = 7$ for domains. We set $\alpha = 0.6$ in our experiments.

Tracing the taxonomy tree up from the leaf nodes to the root, we obtain the nonlinear representations of the root node I^N_{root}. We then apply a batch normalization layer [12] to I^N_{root}, and feed it into a MLP classifier to predict the trait label y: Batch normalization (BN) [12] assists in mitigating the influence of internal covariate shift caused by different taxonomic groups. The training objective is to minimize the cross-entropy loss between the predicted label and the ground truth label:

$$\text{CrossEntropy}\left(Y, \text{Softmax}(\text{MLP}(\text{BN}(I^N_{\text{root}})))\right) \quad (5)$$

Materials and methods

Datasets

We conducted a comprehensive evaluation of MIOSTONE using seven publicly available microbiome datasets with varying sample sizes and feature dimensionality. The AlzBiom dataset [14] explored the relationship between the gut microbiome and Alzheimer's disease (AD). It comprises 75 amyloid-positive AD samples and 100 cognitively healthy control samples from the AlzBiom study, profiled with $p = 8,350$ taxa. The ASD dataset [4] investigated the connection between the gut microbiome and abnormal metabolic activity in Autism Spectrum Disorder (ASD). It comprises 30 typically developing (TD) and 30 constipated ASD (C-ASD) samples, profiled with $p = 7,287$ taxa. The GD dataset [48] explored the relationship between the gut microbiome and Graves' disease (GD). It comprises 100 GD samples and 62 healthy control samples, profiled with $p = 8,487$ taxa. The TBC and RUMC datasets [3] are two cohort studies investigating the connection between the gut microbiome and the Parkinson's disease (PD). The TBC cohort includes 46 PD samples and 67 healthy control samples, profiled with $p = 6,227$ taxa. The RUMC cohort comprises 42 PD samples and 72 healthy control samples, profiled with $p = 7,256$ taxa. The IBD dataset [8] investigated the relationship between the gut microbiome with two primary subtypes of inflammatory bowel disease (IBD): Crohn's disease (CD) and ulcerative colitis (UC). It comprises 108 CD and 66 UC samples, profiled with $p = 5,287$ taxa. The HMP2 dataset [18] in the Integrative Human Microbiome Project (iHMP) [11] also investigated the relationship between the gut microbiome and two IBD subtypes: CD and UC. Compared to the IBD dataset, this dataset expands the sample size to $n = 1,158$ (728 CD and 430 UC samples), with an expanded taxa set of size $p = 10,614$. Full details about these datasets are provided in Supplementary Table A.1.

The abundance features in these datasets were profiled using Metagenomic Shotgun (MGS) sequencing, following the standard operating procedure for shotgun metagenomics implemented in the Qiita platform [7]. Specifically, the raw sequencing reads were classified using Woltka v0.1.4 [49] against the Web of Life (WoL) v2 database [50], which includes 15,953 microbial genomes. WoL offers taxonomic annotations, which are based on the Genome Taxonomy Database (GTDB) [26], for these microbial genomes, spanning 124 phyla, 320 classes, 914 orders, 2,057 families, 6,811 genera, and 12,258 species. Because each dataset only profiles a subset of taxa in the WoL taxonomy, we build the MIOSTONE model using a pruned taxonomy tree specific to each dataset.

Benchmark settings

We systematically assessed the performance of MIOSTONE in comparison to five baseline methods: random forest (RF), support vector machine (SVM) with linear kernel, multi-layer perceptron (MLP), TaxoNN [36], and PopPhy-CNN [32]. For the RF and SVM classifiers, we employed the implementation provided by Scikit-learn [28] with default settings. For the MLP classifier, we configured a pyramid-shaped architecture with one hidden layer of size half of the input dimensionality. For TaxoNN, we utilized the recommended settings, *i.e.,* clustering the taxa based on their phylum information and ordering them according to Spearman correlation, as reported to yield optimal performance. For PopPhy-CNN, we followed its built-in preprocessing pipeline, which involved log-scaling the normalized relative abundance of each taxon followed by subsequent min-max normalization. For all methods, except PopPhy-CNN, we preprocessed microbiome features using centered log-ratio transformation [1].

Results

MIOSTONE provides accurate predictions of the host's disease status

To quantify the predictive performance of different methods, we employed two metrics: the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area Under the Precision-Recall Curve (AUPR). Both metrics are widely used in evaluating the performance of binary classifiers, with AUPR being more suitable for imbalanced datasets [5]. We assessed the performance through 5-fold cross-validation, training individual models

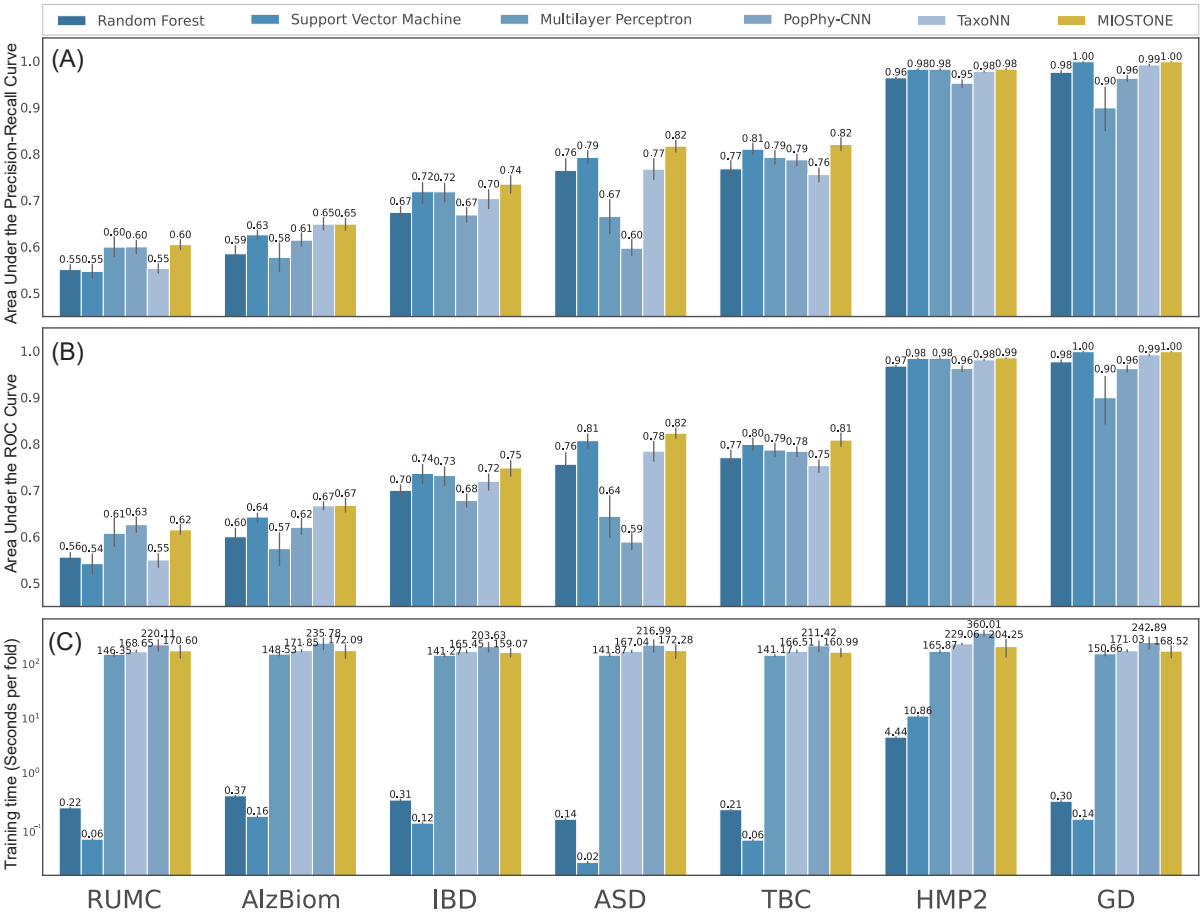

Fig. 2. Evaluation of MIOSTONE in predicting the host's disease status. The evaluation was conducted on seven publicly available microbiome datasets, comparing MIOSTONE to five baseline methods: RF, SVM, MLP, PopPhy-CNN, and TaxoNN. We trained each model ten times, on the same environment, with 10 different random seeds and reported the mean performance with 95% confidence intervals. The model performance is measured by (A) AUPR, (B) AUROC, and (C) The training time in seconds.

on each of the training splits and evaluating them on the corresponding test splits. Predictions from each model were concatenated across all test splits, and the AUROC and AUPR scores were computed on the full dataset. For robustness, we repeated this process 10 times with different random seeds and reported the mean performance with 95% confidence intervals.

Our analysis shows that MIOSTONE considerably outperforms other baseline methods across all datasets, as illustrated in Figure 2(A) and 2(B). The superior performance across diverse microbiome datasets indicates its robustness and generalizability. In contrast, the other two DNN-based baseline methods, PopPhy-CNN and TaxoNN, exhibited varied performance across different datasets. For example, PopPhy-CNN performed among the best in the RUMC dataset, but among the worst in the ASD dataset. Similarly, TaxoNN achieved nearly perfect AUROC and AUPR scores in the GD dataset but performed among the worst in the RUMC dataset, contrasting with PopPhy-CNN's performance. Additionally, we observed that model complexity, a key characteristic of DNN-based methods, plays a critical role in the predictive performance. For example, in datasets like RUMC, higher model complexity leads to better predictive performance in DNN-based methods compared to RF and SVM. However, in other datasets, such as ASD, this complexity inversely hampers the predictive performance in contrast to RF and SVM. Notably, in all examined scenarios, MIOSTONE consistently provides the most accurate predictions of the host's disease status, striking an optimal balance between model complexity and predictive power.

Finally, we assessed the computational efficiency of MIOSTONE compared to the other baseline methods. For a fair comparison, all models were trained on the same environment: AMD EPYC 7302 16-Core Processor, NVIDIA RTX A6000, with 32GB DDR4 RAM. As illustrated in Figure 2(c), training the MIOSTONE model is comparably efficient to training an MLP classifier, which is the simplest DNN-based method. It is worth noting that the RF and SVM classifiers, as provided by Scikit-learn,

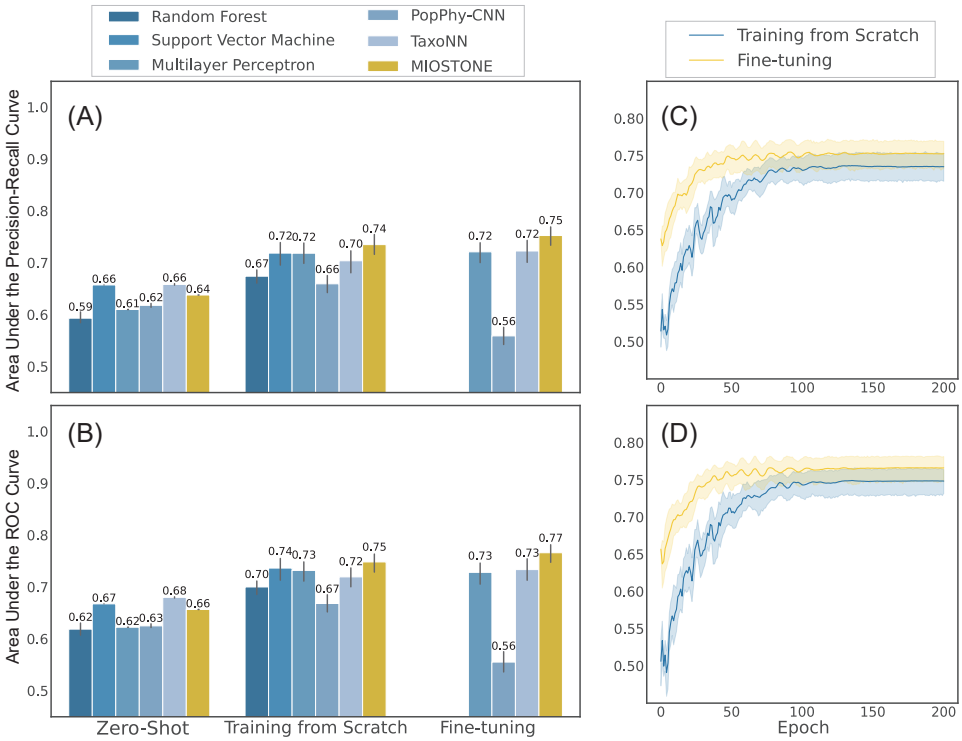

Fig. 3. Evaluation of MIOSTONE's use of knowledge from pre-trained models to improve disease prediction. We pre-trained a model on the large HMP2 dataset and then employed it for the smaller IBD dataset prediction task in three settings: direct prediction on IBD, fine-tuning on IBD, and training IBD from scratch. Two baseline methods, RF and SVM, cannot undergo fine-tuning, so we left them blank. The performance is measured by (A) AUPR and (B) AUROC. Furthermore, we examined the MIOSTONE training dynamics by comparing fine-tuning and training from scratch based on (C) AUPR and (D) AUROC on test splits as a function of the training epochs. Fine-tuning required significantly fewer training epochs to achieve better performance compared to training from scratch.

remain the most efficient models compared to any DNN-based models. However, considering the relatively small size of microbiome datasets and the need for improved predictive models, we assert that computational cost will not be a bottleneck to hinder the applicability of MIOSTONE.

MIOSTONE improves predictive performance in sample-limited tasks through knowledge transfer

In microbiome data analysis, the transfer learning paradigm [44] can be highly beneficial. It proves to be particularly useful when dealing with small datasets for prediction tasks; leveraging the knowledge accumulated from existing models can significantly enhance predictive performance. To evaluate MIOSTONE's ability to transfer knowledge from existing models, we selected two datasets (HMP2 and IBD) with the common objective of exploring the relationship between the gut microbiome and two inflammatory bowel disease subtypes: Crohn's disease (CD) and ulcerative colitis (UC). We then investigated whether the knowledge acquired from the large HMP2 dataset with $1,158$ samples could enhance the predictive performance on the smaller IBD dataset with 174 samples.

We pre-trained a model on the HMP2 dataset and then employed it for the IBD dataset prediction task in three settings: (1) directly using the pre-trained HMP2 model for predictions on the IBD dataset (zero-shot); (2) initializing the IBD model with the pre-trained HMP2 model and fine-tuning it on the IBD dataset (fine-tuning); and (3) training a new model from scratch using only the IBD dataset (train-from-scratch). Note that the HMP2 dataset has a higher feature dimensionality ($10,614$) than the IBD dataset ($5,287$). To ensure compatibility with the pre-trained model, we truncated the HMP features to match the dimensionality of the IBD dataset. It would be interesting to explore the direct use of the pre-trained, incompatible model in future research.

We then evaluated MIOSTONE's performance using knowledge from pre-trained models, as shown in Figure 3(A) and 3(B). Our analysis demonstrates that fine-tuning enhances predictive performance in MIOSTONE compared to training from scratch. In contrast, the other three DNN-based baseline methods, MLP, PopPhy-CNN, and TaxoNN, showed varied ability to transfer knowledge from existing models. Like MIOSTONE, TaxoNN showed enhanced predictive performance, though it still did not surpass MIOSTONE's performance. However, MLP did not show a noticeable improvement with the pre-trained model. Furthermore, PopPhy-CNN exhibited a reverse

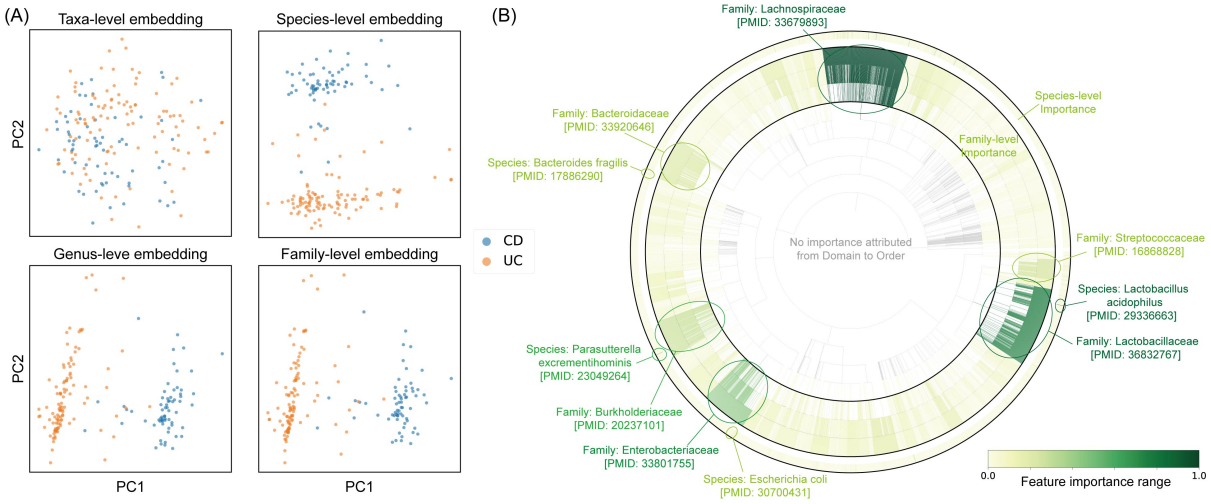

Fig. 4. Using the MIOSTONE model for discovering microbiome-disease associations. (A) PCA embeddings of input features based on taxa profiling cannot effectively distinguish two IBD disease subtypes, while embeddings based on internal neuron representations exhibit improved separation. (B) Feature importance attribution on family-level and species-level taxonomic groups. The importance of a family-level taxonomic group will also be assigned to its children taxonomic groups. The important family-level and species-level taxonomic groups are labeled with their respective group names, supported by literature evidence with accompanying PubMed identifiers.

effect, experiencing a 16.4% decline in AUPR scores with fine-tuning compared to training from scratch.

Finally, we examined the MIOSTONE training dynamics by comparing fine-tuning with training from scratch, focusing on the reported performance on test splits. As illustrated in Figure 3(C) and 3(D), fine-tuning reaches a plateau and attains optimal performance within 40 out of 200 epochs, requiring significantly fewer training epochs for superior performance compared to training from scratch. In other words, leveraging knowledge from pre-trained models through fine-tuning empowers MIOSTONE, which is already computationally efficient, can reduce the training time even further. We conclude that MIOSTONE effectively improves disease prediction through knowledge transfer via fine-tuning.

MIOSTONE identifies microbiome-disease associations with high interpretability

While MIOSTONE was primarily developed for prediction, its exceptional performance in distinguishing disease status suggests that understanding the model's internal mechanisms could provide valuable insights for scientific discovery. To this end, we began by investigating whether internal neuron representations within the MIOSTONE model encode disease-specific signatures. We projected the internal neuron representations of samples into a two-dimensional embedding space using Principal Components Analysis (PCA) and assessed their effectiveness in differentiating two IBD disease subtypes: Crohn's disease (CD) and ulcerative colitis (UC). As shown in Figure 4(A), the PCA embeddings of microbiome features based on taxa profiling fail to distinguish IBD disease subtypes, as samples from CD and UC patients are mixed together. However, when we represented each patient sample by concatenating internal neuron representations at specific taxonomic levels (e.g., Species, Genus, or Family), the resulting embeddings showed markedly better separation, where there was a clear distinction between the different IBD subtypes, indicating that the model's internal representations capture rich disease-specific signatures.

Recognizing the model's potential in capturing disease-specific signatures, we further explored the MIOSTONE model to discover significant microbiome-disease associations. Important associations were scored using feature attribution methods, which assign importance scores to taxonomic groups, with higher scores denoting greater importance to the model's prediction. In this study, we used DeepLIFT [37] as a representative feature attribution method, but in principle, any off-the-shelf feature attribution methods [39, 21, 20] can be used. We focused on microbiome-disease associations at two taxonomic levels: the species level for finer taxonomic resolution and the family level for a coarser resolution. As shown in Figure 4(B), at the family level, significant groups, supported by literature evidence, have their corresponding PubMed identifiers presented alongside their group names. For example, the Lachnospiraceae family, predominantly found in the gut microbiota of mammals and humans, is associated with IBD disease. [15, 34]. At the species level, while significant groups appear more sparsely, they align well with significant family-level groups. For example, Lactobacillus acidophilus, a species in the Lactobacillus family, has been recognized as a potential novel treatment for IBD due to its role in modulating the balance between Th17 and Treg cells [25]. We conclude that MIOSTONE effectively identifies microbiome-disease

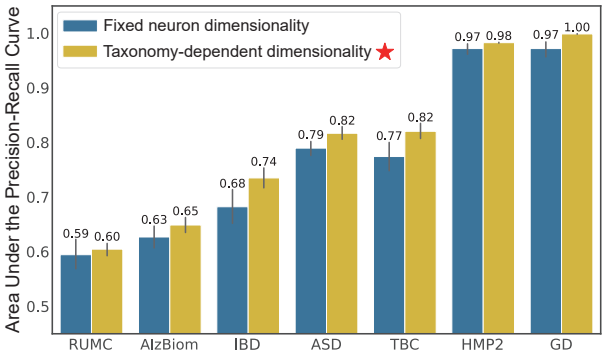

Fig. 5. Evaluating the internal neuron dimensionality in MIOSTONE through an ablation study. We evaluated the necessity of the taxonomy-dependent internal neuron dimensionality (marked by ★) by comparing performance changes when replaced with a fixed solution. The taxonomy-dependent internal neuron dimensionality consistently outperforms the fixed dimensionality approach.

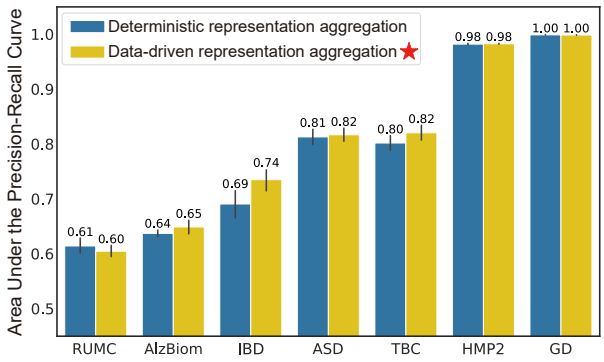

Fig. 6. Evaluating the data-driven aggregation of neuron representation in MIOSTONE through an ablation study. We evaluated the necessity of the data-driven aggregation of neuron representation (marked by ★) by comparing performance changes when replaced with an alternative solution. the data-driven aggregation of neuron representations outperforms or matches the deterministic selection of nonlinear representation on most of the datasets.

associations across different taxonomic levels, providing valuable insights for scientific discovery.

Necessity of MIOSTONE and its components

The design of MIOSTONE incorporates several key components, including the taxonomy-encoded DNN architecture, the data-driven aggregation of additive and nonlinear neuron representations through stochastic gating, and the taxonomy-dependent internal neuron dimensionality. To assess the impact of each component on disease prediction, we conducted an ablation study, in which we modified MIOSTONE to replace its components with alternative solutions. In particular, we considered 2 variants of MIOSTONE: (1) the taxonomy-dependent internal neuron dimensionality was set to a fixed dimension of 2; and (2) the data-driven aggregation of neuron representations was made deterministic by disabling stochastic gating and opting for nonlinear representation only. For each variant, we applied MIOSTONE to seven real microbiome datasets using the same settings described in Section 4.1. Note that we do not explicitly evaluate the necessity of the taxonomy-encoded DNN architecture because its superior performance against other DNN methods has been demonstrated in Section 4.1.

The results of this analysis indicate that all key components positively contribute to MIOSTONE's performance. In the first study, as depicted in Figure 5, the taxonomy-dependent internal neuron dimensionality consistently outperforms the fixed dimensionality approach across all seven microbiome datasets, irrespective of sample sizes and feature dimensionality. In the second study, as depicted in Figure 6, the data-driven aggregation of neuron representations outperforms or matches the deterministic selection of nonlinear representation on six out of seven datasets, with RUMC being the only exception.

Finally, we examined the necessity of MIOSTONE by addressing the question, "Can we mitigate the curse of dimensionality by only utilizing a subset of highly informative features?" To do this, we chose the top-k highly variable taxa that contribute strongly to sample-to-sample variation, where k ranges from 1%, 5%, 10%, 20%, 50%, and 100% among all taxa. The selection of highly variable taxa is inspired by the identification of highly variable genes [40] in single-cell RNA-seq, as implemented by the Scanpy analysis platform [45]. Given the selected taxa subset, we then trained and evaluated MIOSTONE on seven real microbiome datasets using the same settings described in Section 4.1. As shown in Figure 7, MIOSTONE trained with all microbiome features outperforms or matches the performance of the one trained with the highly variable feature subset on six out of seven datasets, with ASD being the only exceptions. It is crucial to note that the ASD dataset contains only 60 samples, profiled with 7,287 taxa. In such an extreme case of a low number of samples, we reasoned that proper feature selection tends to be beneficial in mitigating potential overfitting problems. Exploring the integration of automatic feature selection into MIOSTONE could be an intriguing avenue for future research.

Discussion and conclusion

In this paper, we propose MIOSTONE, an accurate and interpretable machine learning method for investigating microbiome-trait associations. At its core, MIOSTONE leverages the intercorrelation ofmicrobiome features based on their taxonomic relationships. The key novelties of MIOSTONE are threefold: (1) the taxonomy-encoded architecture to harnesses the capabilities of DNN with mitigated concerns of overfitting; (2) the ability to determine whether taxa within the corresponding taxonomic group provide a better explanation in a data-driven manner; and (3) the interpretable architecture to facilitate the understanding of microbiome-trait associations. We

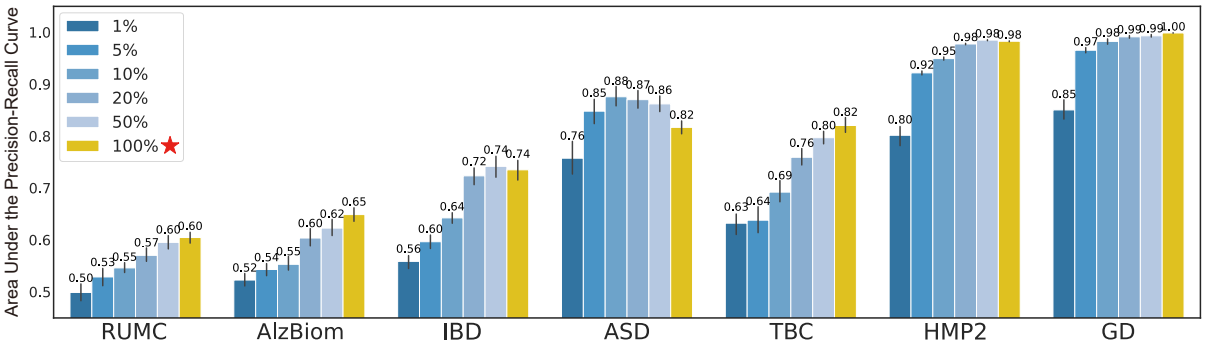

Fig. 7. Evaluating the necessity of MIOSTONE by mitigating the curse of dimensionality using feature selection. We evaluated the robustness of MIOSTONE by utilizing a subset of highly variable taxa, and found that MIOSTONE trained with all microbiome features outperforms or matches the performance of the one trained with the taxa subset on most of the datasets.

validated its performance on seven real datasets, demonstrating its superiority in predictive performance and biological interpretability. Beyond disease status prediction, it can discover significant microbiome-disease associations and transfer knowledge to enhance predictive performance in tasks with limited samples. In conclusion, MIOSTONE adeptly navigates the analysis of microbiome data, effectively addressing issues such as data imperfections, sparsity, low signal-to-noise ratio, and the curse of dimensionality. We believe that this powerful analytical tool will enhance our understanding of the microbiome's impact on human health and disease and will be instrumental in advancing novel microbiome-based therapeutics.

References

1. J. Aitchison. The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological), 44(2):139–160, 1982.
2. D. Albanese, C. De Filippo, D. Cavalieri, and C. Donati. Explaining diversity in metagenomic datasets by phylogenetic-based feature weighting. PLoS Computational Biology, 11(3):e1004186, 2015.
3. J. Boktor, G. Sharon, L. Verhagen Metman, D. A. Hall, P. A. Engen, Z. Zreloff, D. J. Hakim, J. W. Bostick, J. Ousey, D. Lange, et al. Integrated multi-cohort analysis of the Parkinson's disease gut metagenome. Movement Disorders, 38(3):399–409, 2023.
4. Z. Dan, X. Mao, Q. Liu, M. Guo, Y. Zhuang, Z. Liu, K. Chen, J. Chen, R. Xu, J. Tang, et al. Altered gut microbial profile is associated with abnormal metabolism activity of Autism Spectrum Disorder. Gut Microbes, 11(5):1246–1267, 2020.
5. J. Davis and M. Goadrich. The relationship between precision-recall and ROC curves. In International Conference on Machine Learning, pages 233–240, 2006.
6. H. A. Elmarakeby, J. Hwang, R. Arafeh, J. Crowdis, S. Gang, D. Liu, S. H. AlDubayan, K. Salari, S. Kregel, C. Richter, et al. Biologically informed deep neural network for prostate cancer discovery. Nature, 598(7880):348–352, 2021.
7. A. Gonzalez, J. A. Navas-Molina, T. Kosciolek, D. McDonald, Y. Vázquez-Baeza, G. Ackermann, J. DeReus, S. Janssen, A. D. Swafford, S. B. Orchanian, et al. Qiita: rapid, web-enabled microbiome meta-analysis. Nature Methods, 15(10):796–798, 2018.
8. C. G. Gonzalez, R. H. Mills, Q. Zhu, C. Sauceda, R. Knight, P. S. Dulai, and D. J. Gonzalez. Location-specific signatures of Crohn's disease at a multi-omics scale. Microbiome, 10(1):133, 2022.
9. E. A. Grice and J. A. Segre. The human microbiome: Our second genome. Annual Review of Genomics and Human Genetics, 13:151–170, 2012.
10. S. Han, J. Pool, J. Tran, and W. Dally. Learning both weights and connections for efficient neural network. In Advances in Neural Information Processing Systems, volume 28, 2015.
11. The Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell Host & Microbe, 16(3):276–289, 2014.
12. S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. International Conference on Machine Learning, pages 448–456, 2015.
13. D. Knights, L. W. Parfrey, J. Zaneveld, C. Lozupone, and R. Knight. Human-associated microbial signatures: examining their predictive value. Cell Host & Microbe, 10(4):292–296, 2011.
14. C. Laske, S. Müller, O. Preische, V. Ruschil, M. Munk, I. Honold, S. Peter, U. Schoppmeier, and M. Willmann. Signature of Alzheimer's disease in intestinal microbiome: Results from the AlzBiom study. Frontiers in Neuroscience, 16:792996, 2022.
15. A. A. Lee, K. Rao, J. Limsrivilai, M. Gillilland III, B. Malamet, E. Briggs, V. B. Young, and P. D. Higgins. Temporal gut microbial changes predict recurrent clostridiodes difficile infection in patients with and without ulcerative colitis. Inflammatory Bowel Diseases, 26(11):1748–1758, 2020.

16. B. Li, D. Zhong, X. Jiang, and T. He. TopoPhyCNN: integrating topological information of phylogenetic tree for host phenotype prediction from metagenomic data. In IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 456–461. IEEE, 2021.
17. B. Liu, Y. Wei, Y. Zhang, and Q. Yang. Deep neural networks for high dimension, low sample size data. In International Joint Conference on Artificial Intelligence, pages 2287–2293, 2017.
18. J. Lloyd-Price, C. Arze, A. N. Ananthakrishnan, M. Schirmer, J. Avila-Pacheco, T. W. Poon, E. Andrews, N. J. Ajami, K. S. Bonham, C. J. Brislawn, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature, 569(7758):655–662, 2019.
19. C. Louizos, M. Welling, and D. P. Kingma. Learning sparse neural networks through l_0 regularization. International Conference on Learning Representations, 2018.
20. Y. Y. Lu, W. Guo, X. Xing, and W. S. Noble. DANCE: Enhancing saliency maps using decoys. In International Conference on Machine Learning, 2021.
21. S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 2017.
22. J. Ma, M. K. Yu, S. Fong, K. Ono, E. Sage, B. Demchak, R. Sharan, and T. Ideker. Using deep learning to model the hierarchical structure and function of a cell. Nature Methods, 15(4):290–298, 2018.
23. R. Hernandez Medina, S. Kutuzova, K. N. Nielsen, J. Johansen, L. H. Hansen, M. Nielsen, and S. Rasmussen. Machine learning and deep learning applications in microbiome research. ISME Communications, 2(1):98, 2022.
24. R. H. Mills, P. S. Dulai, Y. Vázquez-Baeza, C. Sauceda, N. Daniel, R. R. Gerner, L. E. Batachari, M. Malfavon, Q. Zhu, K. Weldon, et al. Multi-omics analyses of the ulcerative colitis gut microbiome link bacteroides vulgatus proteases with disease severity. Nature Microbiology, 7(2):262–276, 2022.
25. J.-S. Park, J. W. Choi, J. Jhun, J. Y. Kwon, B.-I. Lee, C. W. Yang, S.-H. Park, and M.-L. Cho. Lactobacillus acidophilus improves intestinal inflammation in an acute colitis mouse model by regulation of Th17 and Treg cell balance and fibrosis development. Journal of Medicinal Food, 21(3):215–224, 2018.
26. D. H. Parks, M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil, and P. Hugenholtz. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature Biotechnology, 36(10):996–1004, 2018.
27. E. Pasolli, D. T. Truong, F. Malik, L. Waldron, and N. Segata. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Computational Biology, 12(7):e1004977, 2016.
28. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.
29. G. D. Poore, E. Kopylova, Q. Zhu, C. Carpenter, S. Fraraccio, S. Wandro, T. Kosciolek, S. Janssen, J. Metcalf, S. J. Song, et al. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature, 579(7800):567–574, 2020.
30. J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 464(7285):59–65, 2010.
31. J. Qin, Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature, 490(7418):55–60, 2012.
32. D. Reiman, A. A. Metwally, J. Sun, and Y. Dai. 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(10):2993–3001, 2020.
33. S. Roy and G. Trinchieri. Microbiota: a key orchestrator of cancer therapy. Nature Reviews Cancer, 17(5):271–285, 2017.
34. K. Sasaki, J. Inoue, D. Sasaki, N. Hoshi, T. Shirai, I. Fukuda, T. Azuma, A. Kondo, and R. Osawa. Construction of a model culture system of human colonic microbiota to detect decreased Lachnospiraceae abundance and butyrogenesis in the feces of ulcerative colitis patients. Biotechnology Journal, 14(5):1800555, 2019.
35. I. Sekirov, S. L. Russell, C. M. Antunes, and B. B. Finlay. Gut microbiota in health and disease. Physiological Reviews, 2010.
36. D. Sharma, A. D. Paterson, and W. Xu. TaxoNN: ensemble of neural networks on stratified microbiome data for disease prediction. Bioinformatics, 36(17):4544–4550, 2020.
37. A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje. Learning important features through propagating activation differences. In International Conference on Machine Learning, 2017.
38. O. Shtossel, H. Isakov, S. Turjeman, O. Koren, and Y. Louzoun. Ordering taxa in image convolution networks improves microbiome-based machine learning accuracy. Gut Microbes, 15(1):2224474, 2023.
39. K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034, 2013.
40. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papelexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive integration of single-cell data. Cell, 77(7):1888–1902, 2019.

41. P. J. Turnbaugh, R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature, 444(7122):1027–1031, 2006.
42. N. M. Vogt, R. L. Kerby, K. A. Dill-McFarland, S. J. Harding, A. P. Merluzzi, S. C. Johnson, C. M. Carlsson, S. Asthana, H. Zetterberg, K. Blennow, et al. Gut microbiome alterations in Alzheimer's disease. Scientific Reports, 7(1):13537, 2017.
43. M. Wang, X. Zhou, F. Li, J. Huckins, R. W. King, and S. T. C. Wong. Novel cell segmentation and online SVM for cell cycle phase identification in automated microscopy. Bioinformatics, 24(1):94–101, 2008.
44. K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. Journal of Big data, 3(1):1–40, 2016.
45. F. A. Wolf, P. Angerer, and F. J. Theis. SCANPY: large-scale single-cell gene expression data analysis. Genome Biology, 19:1–5, 2018.
46. J. Xiao, L. Chen, S. Johnson, Y. Yu, X. Zhang, and J. Chen. Predictive modeling of microbiome data using a phylogeny-regularized generalized linear mixed model. Frontiers in Microbiology, 9:1391, 2018.
47. S. H. Ye, K. J. Siddle, D. J. Park, and P. C. Sabeti. Benchmarking metagenomics tools for taxonomic classification. Cell, 178(4):779–794, 2019.
48. Q. Zhu, Q. Hou, S. Huang, Q. Ou, D. Huo, Y. Vázquez-Baeza, C. Cen, V. Cantu, M. Estaki, H. Chang, et al. Compositional and genetic alterations in Graves' disease gut microbiome reveal specific diagnostic biomarkers. The ISME journal, 15(11):3399–3411, 2021.
49. Q. Zhu, S. Huang, A. Gonzalez, I. McGrath, D. McDonald, N. Haiminen, G. Armstrong, Y. Vazquez-Baeza, J. Yu, J. Kuczynski, et al. Phylogeny-aware analysis of metagenome community ecology based on matched reference genomes while bypassing taxonomy. mSystems, 7(2):e00167–22, 2022.
50. Q. Zhu, U. Mai, W. Pfeiffer, S. Janssen, F. Asnicar, J. G. Sanders, P. Belda-Ferre, G. A. Al-Ghalith, E. Kopylova, D. McDonald, et al. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. Nature Communications, 10(1):5477, 2019.

Appendix

Table A.1. The details of the datasets investigated by MIOSTONE.

Dataset	Data source	Notes
AlzBiom [14]	EBI-ENA ID: PRJEB47976	The sequencing data is clean (QC'ed, host-filtered).
ASD [4]	EBI-ENA ID: PRJNA451479	The sequencing data is raw (non-QC'ed). In the original study, sequences were filtered against human reference genome hg19.
GD [48]	EBI-ENA ID: PRJNA602729, PRJNA602731, PRJNA602732, PRJNA638403, PRJNA638404, PRJNA638405	Most samples have a pair of FASTQ files. However, 4 samples (three.lst) have a third, unpaired FASTQ file that is very small, and it should be excluded from the analysis. 12 samples have only one FASTQ file, which appears to be single-end sequences. Two samples: GA61 (SRR12000211) and GA89 (SRR12005695) are missing from the metadata. Therefore they were dropped from the data.
TBC & RUMC [3]	Qiita ID: 14476 (TBC) and 12975 (RUMC)	For TBC, 5 samples in BIOM are missing in metadata. For RUMC, 20 samples in BIOM are missing in metadata. These samples were dropped.
IBD [8]	Qiita ID: 12675	The dataset contains metagenomic, 16S rRNA gene sequencing data, and associated metadata. More details can be found at: `https://qiita.ucsd.edu/study/description/12675`
HMP2 [18]	Qiita ID: 11484	The dataset contains metagenomic, 16S rRNA gene sequencing data, and associated metadata from the Human Microbiome Project. More details can be found at: `https://hmpdacc.org/ihmp`