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

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

[1]Cheriton School of Computer Science, University of Waterloo
[2]School of Life Sciences, Arizona State University

Abstract

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 dimensionality. In this paper, we address these challenges by introducing MIOSTONE, a novel machine learning method that leverages the intercorrelation of microbiome features due to their taxonomic relationships. MIOSTONE integrates prior knowledge from the "Web of Life" (WoL), a comprehensive reference phylogeny with taxonomic annotations, into a deep neural network (DNN) framework. It processes microbiome data by feeding it into a neural network organized in layers to emulate the hierarchy of the WoL taxonomy, reducing overparameterization, and offering natural explanations for microbiome-trait associations. We empirically assessed MIOSTONE's accuracy and interpretability on real datasets, demonstrating its superior performance compared to other methods. MIOSTONE offers a powerful tool for robust analysis of microbiome data, shedding light on the microbiome's impact on human health and disease, and facilitating disease diagnosis and treatment. [2]

1 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 [1]. As the microbiome influences the impact of host genes, microbiome genomes are often referred to as the "second genome" [2]. Subsequently, microbiomes have been found to play pivotal roles in various aspects of human health and disease [3], including diabetes [4], obesity [5], inflammatory bowel disease [6], Alzheimer's disease [7], cancers [8], and more. For example, the microbiome influences cancer development and therapeutic responses in cancer patients [9]. Thus, investigating the microbiome's association with host traits provides valuable insights into its impact on health, offering valuable information for disease diagnosis and treatment.

In recent years, high-throughput shotgun metagenomic sequencing technologies have enabled biologists to gather vast amounts of publicly available microbiome data. Microbiome features are typically profiled from taxa with varying taxonomic and phylogenetic specificity, measuring the

*Corresponding Author. Email: Qiyun.Zhu@asu.edu, yanglu@uwaterloo.ca

[2]The Apache licensed source code of MIOSTONE will be available at https://github.com/batmen-lab/MIOSTONE.

abundance or presence of specific taxa across samples. However, these features possess unique characteristics that present challenges in downstream analyses. 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 [10]. Secondly, the profiling data, represented as the abundance of hundreds to thousands of taxa as features across samples, is sparse, high-dimensional, and includes inflated zero counts [11]. The excessive number of features poses challenges in analysis due to the curse of dimensionality from two perspectives [12]: (1) The relatively low number of samples can lead to overfitting during training, which limits generalization to other datasets. (2) Overfitting concerns can hinder the use of powerful analysis tools like deep neural networks (DNNs). Therefore, new approaches are required to analyze microbiome data, accounting for data imperfections, sparsity, low signal-to-noise ratio, and the curse of dimensionality.

In this paper, to address the inherent challenges of sparse, noisy, and high-dimensional microbiome data, we propose MIOSTONE (MIcrObiome-trait aSociations with TaxONomy-adaptivE neural networks), an accurate and interpretable machine learning method. At its core, MIOSTONE capitalizes on the intercorrelation of microbiome features due to their inherent phylogenetic relationships, which represent the evolutionary relationships among taxa. In a phylogenetic tree, the branch points indicate divergence events where two lineages split from a common ancestor to form distinct species, while the branch length measures the extent of genetic changes [13]. While phylogeny offers a wealth of information for characterizing microbial communities and understanding their evolutionary relationships, its encyclopedia-like resolution can be challenging to interpret and computationally expensive to analyze. As an alternative, taxonomic relationships categorize taxa into hierarchical groups, spanning from the three domains (Bacteria, Archaea, and Eukarya) down to species[14], using a well-established and widely accepted naming system. Since taxonomy offers a standardized framework for interpretation across studies, MIOSTONE's primary concept involves integrating taxonomic prior knowledge to enhance microbiome data analysis.

Concretely, MIOSTONE utilizes "Web of Life" (WoL) [15], a comprehensive reference that contains $15,953$ microbial genomes based on 380 single-copy marker genes. WoL also provides taxonomic annotations for these microbial genomes, covering 124 phyla, 320 classes, 914 orders, $2,057$ families, $6,811$ genera, and $12,258$ species. WoL has been extensively used in multiple microbiome studies and has discovered novel insights into various biological questions [16, 8]. MIOSTONE introduces a novel deep neural network (DNN) framework inspired by WoL, whose design adopts principles from biologically-informed DNNs [17, 18]. Specifically, MIOSTONE processes microbiome data by feeding it into a neural network organized in layers to emulate the hierarchy of the WoL taxonomy, as illustrated in Figure 1. Each neuron in the network represents a specific taxonomic group, and connections between neurons symbolize subordination relations between these groups. Notably, the taxonomy-encoded DNN mitigates overfitting by reducing over-parameterization, as it involves fewer connections between taxonomic groups with subordination relations compared to fully connected layers. More importantly, the novelty of MIOSTONE resides in the ability of each internal neuron to determine, in a data-driven manner, whether taxa within the corresponding taxonomic group provide a better explanation for the label when considered holistically or individually. Consequently, the incorporated taxonomic relationships offer natural explanations for understanding microbiome-trait associations. We have applied MIOSTONE to various tasks, demonstrating its empirical utility in predicting disease status, discovering significant microbiome-disease associations, and transferring knowledge to enhance performance in tasks with limited samples.

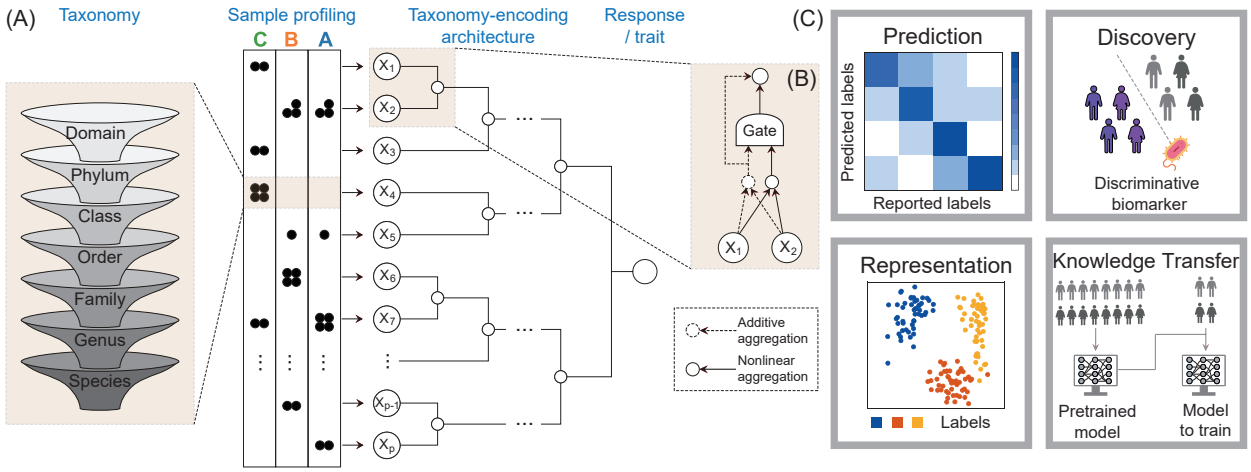

Figure 1: **MIOSTONE utilizes deep neural networks encoded with taxonomic prior knowledge to model microbiome data.** (A) A graphical illustration of a taxonomy-encoded deep neural network. The input microbiome data represents the abundance of microbial taxa, each categorized into hierarchical taxonomic groups ranging from domains to species. The label of interest is predicted from the input data using a neural network that connects a set of neurons organized in layers to emulate the hierarchy of the "Web of Life" taxonomy. (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.

2 Methods

2.1 Problem setup

Consider a prediction task where we have n sample-label pairs (X, Y), where $X = \{x_i\}_{i=1}^n \in \mathbb{R}^{n\times p}$ and $Y = \{y_i\}_{i=1}^n \in \mathbb{R}^{n\times 1}$, denoting the data matrix with p-dimensional microbiome features and the corresponding labels, respectively. In our context, the feature profiles can represent the relative abundance of hundreds to thousands of taxa across various samples, which have undergone centered log-ratio transformation [19]. Usually, the number of microbiome samples n is relatively low when compared to the high feature dimensionality p. Furthermore, we consider additional prior knowledge in the form of hierarchical taxonomic relationships among the observed microbiome taxa, which reflects traditional naming and organization based on their similarities. It is worth noting that in principle, MIOSTONE can be extended to other off-the-shelf relationships, such as phylogeny, which describes the evolutionary relationships among taxa. Mathematically, we represent the taxonomic tree as a hierarchical graph $G = (V_L \cup V_I, E)$. Specifically, we denote V_L as the set of observed microbiome taxa, or equivalently, the tree's leaf nodes with $|V_L| = p$. We further denote V_I as the internal nodes within the tree, representing the taxonomic groups at different ranks, and these nodes are also considered as ancestors to the leaf nodes in V_L. For each node within the taxonomic tree $v \in V_I \cup V_L$, we denote its parent as $\text{parent}(v) \in V_I$. For two nodes $u, v \in V_I \cup V_L$, an undirected

edge $e = (u, v) \in E$ connects u and v if and only if $u = \text{parent}(v)$ or $v = \text{parent}(u)$. Analogously, we denote the children of v 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 microbiome taxon $v \in V_L$ can trace its taxonomic groups from the species to the domain through a unique path $v, \text{parent}(v), \text{parent}(\text{parent}(v)), \cdots$. For each internal node $v \in V_I$, we denote its descendant size, $S(v)$, as the number of observed microbiome taxa that can be traced back to v along the path. 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 deep neural network (DNN), that maps from the input feature $x \in \mathbb{R}^p$ to the label $y \in \mathbb{R}$. Learning such a function is challenging because DNNs are prone to overfitting, especially when the number of samples is relatively low compared to the high feature dimensionality. To address this problem, MIOSTONE incorporates G into the DNN architecture to mitigate overfitting and improve the model's generalization ability for unseen data.

2.2 MIOSTONE architecture

MIOSTONE introduces a novel biologically informed deep neural network (DNN) framework designed with guidance from additional prior knowledge in the form of taxonomic trees, as illustrated in Figure 1(A). Specifically, MIOSTONE trains a DNN to predict the label of interest from the input microbiome data through an architecture that precisely mirrors the hierarchy of the WoL taxonomy. Each neuron in the network corresponds to a specific taxonomic group, and the connections between neurons represent subordination relations within these groups. Compared to fully connected DNNs, the taxonomy-encoded architecture only establishes links between taxonomic groups with subordination relationships, leading to fewer connections. This helps mitigate overfitting by reducing over-parameterization.

Furthermore, it is important to note that a substantial number of taxa may display misleading signals due to the ambiguous assignment of fragments to taxa. For example, a sequenced viral fragment from the Omicron variant may be erroneously assigned to the Delta variant, both belonging to the SARS-CoV-2 lineage, but it is unlikely to be assigned to SARS-CoV-1. Given that taxonomic relationships among microbiome features are encoded within the DNN architecture, an ideal approach would include a systematic strategy for each taxonomic group. Such a strategy would strike a balance between reducing the ambiguity of fragment-to-taxon assignments and effectively explaining the label of interest. Intuitively, a high degree of ambiguity in fragment-to-taxon assignments within a taxonomic group suggests that aggregating taxa may help reduce this ambiguity. Conversely, low ambiguity indicates that each taxon may be individually significant for explaining the label. Specifically, our goal is for the internal neuron to possess the ability, in a data-driven manner, to determine whether taxa within the corresponding taxonomic group provide a more effective explanation of the label when considered holistically as a linear aggregation or individually by leveraging their unique and synergistic nonlinear effects.

This data-driven strategy is accomplished using a neural network-based approach with stochastic gating [20], 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 $I_v^N \in \mathbb{R}^{d_v}$, which combine the taxa within the corresponding taxonomic group either holistically or individually, where d_v is the representation dimension. The additive representation I_v^A is defined by a linear function of the concatenated additive representations of all children of v:

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

where $u_1, u_2, \cdots \in \text{children}(v)$ and $\text{Linear}(\cdot)$ is a linear transformation function. Similarly, the

nonlinear representations of all children of v are combined non-linearly through a fully connected multi-layer perceptron (MLP) with one hidden layer, aiming to map the concatenated nonlinear representations into an aggregated representation $I_v^M \in \mathbb{R}^{d_v}$. The MLP network consists of a linear transformation and a nonlinear activation function. The aggregation of the nonlinear representations of all children of v is represented as:

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

The nonlinear representation I_v^N combines I_v^A and I_v^M using a stochastic gate $m_v \in (0, 1)$:

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

where the gate m_v approximates a relaxed and differentiable Bernoulli distribution based on the re-parameterization trick [20]:

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

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. In this paper we fix $\beta = 0.1$. 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 as a whole 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 may 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 $v \in V_I$ by $\alpha^{\text{L}} \cdot S(v)$, where $\alpha = 0.95$ is a decay factor, $L > 0$ denotes the taxonomic level relative to the species, and $S(v)$ represents the number of observed microbiome taxa within the taxonomic group v. For example, the representation dimension for v_1 at the species level would be $\alpha \cdot S(v_1)$, and for v_2 at the genus level would be $\alpha^2 \cdot S(v_2)$.

Tracing the taxonomy upward, we arrive at three domain-level taxonomic groups: Bacteria, Archaea, and Eukarya. Their nonlinear representations are concatenated, subjected to batch normalization, and then fed into a one-layer fully connected MLP to predict the label of interest. Batch normalization [21] assists in mitigating the influence of internal covariate shift caused by different taxonomic groups. The training objective is to minimize the loss between the predicted labels and the ground truth labels:

$$\text{Loss}\left(Y, \text{Softmax}(\text{MLP}(\text{BatchNorm}(\left[I_{\text{Bacteria}}^N, I_{\text{Archaea}}^N, I_{\text{Eukarya}}^N\right])))\right) \tag{5}$$

where we use the cross-entropy loss function for $\text{Loss}(\cdot, \cdot)$ in the experiments.

3 Results

3.1 Datasets

We evaluated the performance of MIOSTONE on a variety of datasets, including four real gut microbiome datasets with varying sample sizes and feature dimensionality. The features in these

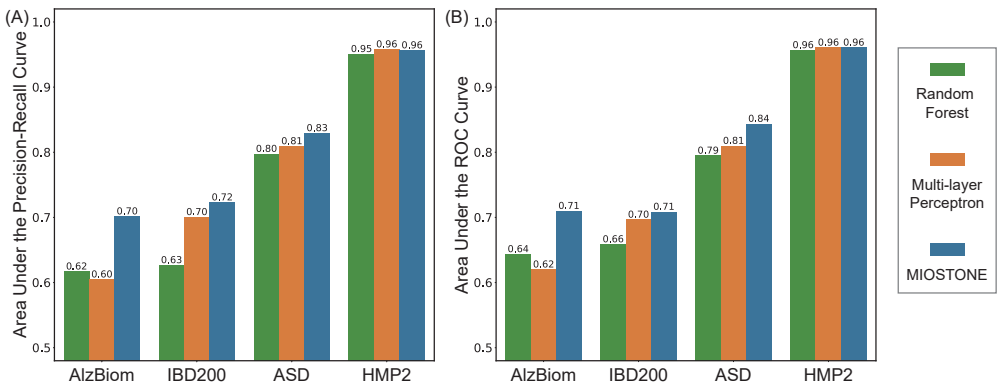

Figure 2: **Evaluation of MIOSTONE in predicting the host's disease status.** The evaluation was conducted on four datasets, comparing MIOSTONE to baseline methods: random forest and multi-layer perceptron. We trained each model three times with different random seeds and reported the median performance. The performance is measured by (A) AUPR and (B) AUROC.

datasets were profiled from shotgun metagenomic sequencing data. The first dataset [22], referred to as AlzBiom, explored the relationship between the gut microbiome and Alzheimer's disease (AD). It contains $n = 175$ samples with 75 amyloid-positive AD patients and 100 cognitively healthy controls from the AlzBiom study, profiled with $p = 15,939$ features. The second dataset [23], referred to as IBD200, investigated the gut microbiome's correlation with two primary subtypes of inflammatory bowel disease (IBD): Crohn's disease (CD) and ulcerative colitis (UC). This dataset consists of 117 CD and 73 UC patient samples, totaling $n = 190$ samples, characterized by $p = 5,300$ features that are present in at least one of the samples. The third dataset [24], referred to as ASD, linked the gut microbiome to abnormal metabolic activity in Autism Spectrum Disorder (ASD). Including 60 samples from children aged 2-13 years, it comprises 30 typically developing (TD) and 30 constipated ASD (C-ASD) cases, each profiled with $p = 7,837$ features. The final dataset [25], referred to as HMP2, is part of a comprehensive study of human microbial communities, namely the Integrative Human Microbiome Project (iHMP) [26]. In our analysis, we specifically examined the subset of gut microbiome data related to two subtypes of IBD: CD and UC. Compared to the IBD200 dataset, this dataset expands the sample size to $n = 1,158$, including 728 CD patient samples and 430 UC patient samples, each represented by an expanded feature set with $p = 10,614$.

3.2 MIOSTONE provides more accurate predictions of the host's disease status

We systematically evaluated the performance of MIOSTONE in comparison to two baseline methods: random forest (RF) and multi-layer perceptron (MLP). For the RF classifier, we used the implementation provided by Scikit-learn [27] with its default settings. For the MLP classifier, we configured it with a pyramid-shaped architecture comprising three hidden layers, with the number of neurons in each layer being p, $p/2$, and $p/4$ respectively. Classification performance was measured by the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area Under the Precision-Recall Curve (AUPR). We utilized both metrics due to their different advantages: AUROC's common adoption and AUPR's suitability for imbalanced datasets. The performance was assessed via 3-fold cross-validation, where individual models were trained on each training split and then used to predict labels for the corresponding test split. Predictions from these three test

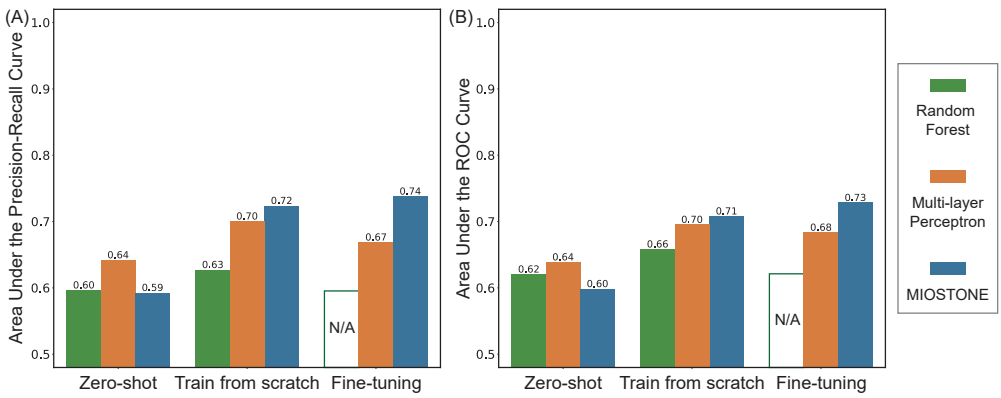

Figure 3: **Evaluation of MIOSTONE's use of knowledge from pre-trained models to improve disease prediction.** We pre-trained a model on the HMP2 dataset and applied it to the IBD dataset prediction task in three settings: direct prediction on IBD, fine-tuning on IBD, and training IBD from scratch. Among the two baseline methods, random forest cannot be fine-tuned, so we left it blank. The performance is measured by (A) AUPR and (B) AUROC.

splits were aggregated to calculate AUROC and AUPR scores. For robustness, each model was trained three times with varying random seeds, and the median performance was reported.

Upon investigating the performance of MIOSTONE on four real microbiome datasets (as illustrated in Figure 2), we noted that in certain scenarios (*e.g.,* IBD200), overparameterization can enable MLP to outperform RF. Conversely, in other cases (e.g., AlzBiom dataset), overparameterization can impair MLP's performance, highlighting the curse of dimensionality issue in microbiome data analysis [12]. Notable, in all examined scenarios, MIOSTONE consistently delivers the most accurate predictions of the host's disease status, as reflected by both AUROC and AUPR scores. This indicates that MIOSTONE successfully harnesses the capabilities of a deep neural network while mitigating overparameterization concerns. Particularly in the AlzBiom dataset, as shown in Figure 2, where MLP's performance is adversely affected by overparameterization, MIOSTONE shows significant improvements over both RF and MLP. On average, MIOSTONE achieved a 14.78% increase in AUPR and a 12.73% increase in AUROC compared to the baseline methods. We conclude that MIOSTONE effectively predicts the host's disease status, irrespective of the dataset's complexity.

3.3 MIOSTONE enhances disease prediction through knowledge transfer from pre-trained models

In microbiome data analysis, the transfer learning paradigm [28], is highly valuable. It proves particularly useful when dealing with small datasets for prediction tasks. Leveraging the knowledge accumulated from existing models can significantly enhance prediction performance. We pre-trained a model on the HMP2 dataset and leveraged it for the IBD dataset prediction task from three perspectives: (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 HMP dataset has a higher feature dimensionality ($10,614$) than the IBD dataset ($5,300$). To ensure compatibility with the pre-training model, we truncated the

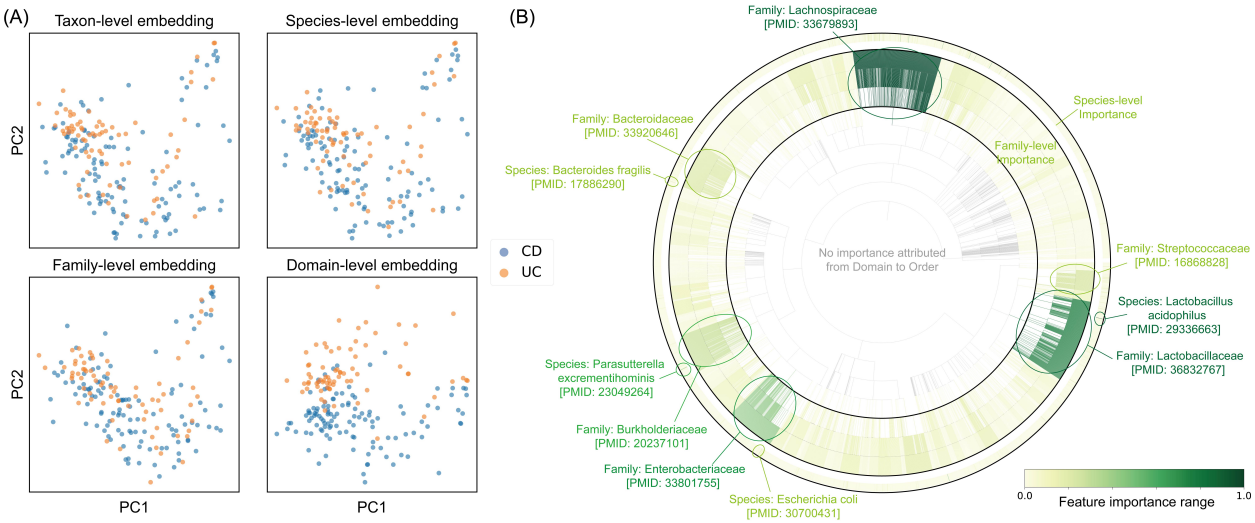

Figure 4: **Using the MIOSTONE model for discovering microbiome-disease associations.** (A) PCA embeddings of input features based on taxa profiling cannot effectively distinguish IBD disease subtypes, while embeddings based on internal neuron representations exhibit improved separation. (B) Feature 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.

HMP features to match the dimensionality of the IBD dataset. It would be interesting in the future to directly use the pre-trained, incompatible model.

We then evaluated MIOSTONE's performance using knowledge from pre-trained models, as shown in Figure 3. The results show that fine-tuning improves disease prediction performance in MIOSTONE compared to training from scratch. Surprisingly, we observe that fine-tuning a pre-trained MLP model results in a decline in performance. Understanding the reasons behind this phenomenon would be an interesting avenue for future research. Furthermore, using the pre-trained model directly consistently leads to worse predictions than training a model, suggesting that the two datasets may not be perfectly aligned. We conclude that MIOSTONE effectively improves disease prediction through knowledge transfer via fine-tuning.

3.4 MIOSTONE discovers microbiome-disease associations through model explanation

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 patient samples into a two-dimensional embedding space using Principal Components Analysis (PCA) and assessed their effectiveness in differentiating between disease subtypes within the IBD dataset. As shown in Figure 4(A), the PCA embeddings of input features based on taxa profiling fail to distinguish IBD disease subtypes, as samples with

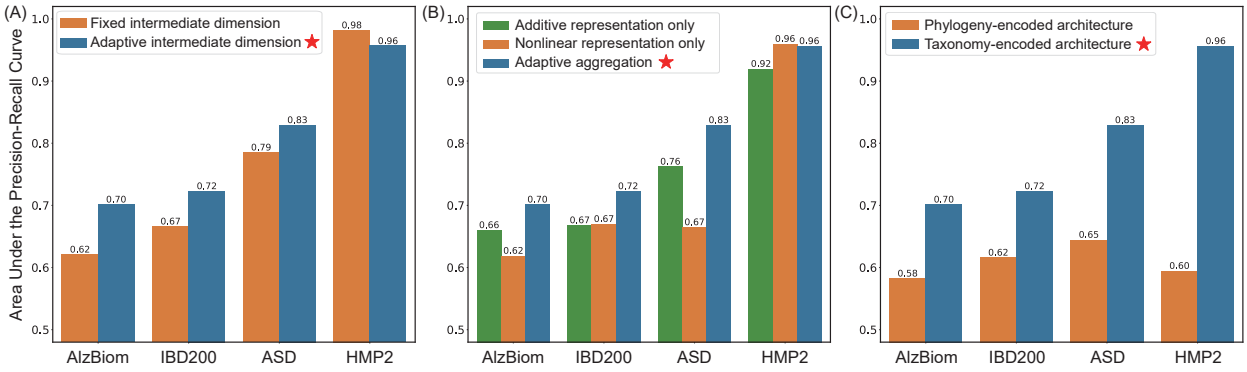

Figure 5: **Evaluation of each key component of MIOSTONE in an ablation study.** We assess the necessity of each component (marked by ★) by comparing performance changes when replaced with an alternative solution. (A) We change the dimension of internal neuron representation from taxonomically dependent to fixed. (B) We disable stochastic gating, which automatically aggregates additive and nonlinear representations, by choosing either one representation exclusively. (C) We replace the taxonomy-encoded DNN architecture with a phylogeny-encoded alternative.

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 or Family), the resulting embeddings showed markedly better separation. This was particularly evident at the Domain level, 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 by employing feature attribution methods, which assign importance scores to taxonomic groups, with higher scores denoting greater importance to the output prediction. In this study, we used DeepLIFT [29] as a representative feature attribution method, but in principle, we can utilize any off-the-shelf feature attribution methods [30, 31, 32]. 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 family-level groups are supported by literature evidence, and their corresponding PubMed identifiers are presented alongside their group names. For example, the Lachnospiraceae family is predominantly found in the gut microbiota of mammals and humans and is associated with IBD disease. [33, 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 regarded as a potential novel treatment for IBD due to its role in modulating the balance between Th17 and Treg cells [35]. We conclude that MIOSTONE effectively identifies microbiome-disease associations across different taxonomic levels, providing valuable insights for scientific discovery.

3.5 Necessity of each MIOSTONE component

The design of MIOSTONE incorporates several key components, including the taxonomy-encoded DNN architecture, the data-driven aggregation of additive and nonlinear representations through

stochastic gating, and the taxonomically dependent internal neuron representation dimension. 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 3 variants of MIOSTONE: (1) the internal neuron representation was set to a fixed dimension of 32; (2) the trainable stochastic gating was made deterministic, choosing either additive or nonlinear representations exclusively; and (3) the taxonomy-encoded DNN architecture was replaced with a phylogeny-encoded one. For each variant, we applied MIOSTONE to four real microbiome datasets using the settings described in Section 3.2.

The results of this analysis indicate that all three key components positively contribute to MIOSTONE's performance. In the first study, as shown in Figure 5(A), the taxonomically dependent representation dimension outperforms the fixed dimension approach on three out of four datasets, with HMP2 being the only exception. It is important to note, however, that HMP2 is such that nearly perfect performance is achieved by all methods, likely due to its larger sample size. The pronounced improvement on datasets with limited samples underscores the value of MIOSTONE. In the second study, as shown in Figure 5(B), depending on the dataset's complexity, simple additive representations work better in some scenarios (*e.g.,* AlzBiom), while nonlinear representations are more effective for others (*e.g.,* HMP2). Notably, in all cases, the data-driven aggregation through stochastic gating yields the most accurate predictions. In the final study, as shown in Figure 5(C), the phylogeny-encoded DNN architecture performs significantly worse than the taxonomy-encoded one. This suggests that the detailed resolution required by phylogeny may pose additional challenges in training compared to taxonomy.

4 Discussion and conclusion

In this paper, we propose MIOSTONE, a novel machine learning method that leverages the intercorrelation of microbiome features, as informed by their taxonomic relationships extracted from the "Web of Life". 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 explainable architecture to facilitate the understanding of microbiome-trait associations. We validated its performance on 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 disease diagnosis and treatment.

References

[1] Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature **464**, 59–65 (2010).

[2] Grice, E. A. & Segre, J. A. The human microbiome: Our second genome. Annual Review of Genomics and Human Genetics **13**, 151–170 (2012).

[3] Sekirov, I., Russell, S. L., Antunes, C. M. & Finlay, B. B. Gut microbiota in health and disease. Physiological Reviews (2010).

[4] Qin, J. et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature **490**, 55–60 (2012).

[5] Turnbaugh, P. J. et al. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature **444**, 1027–1031 (2006).

[6] Mills, R. H. et al. Multi-omics analyses of the ulcerative colitis gut microbiome link bacteroides vulgatus proteases with disease severity. Nature Microbiology **7**, 262–276 (2022).

[7] Vogt, N. M. et al. Gut microbiome alterations in Alzheimer's disease. Scientific Reports **7**, 13537 (2017).

[8] Poore, G. D. et al. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature **579**, 567–574 (2020).

[9] Roy, S. & Trinchieri, G. Microbiota: a key orchestrator of cancer therapy. Nature Reviews Cancer **17**, 271–285 (2017).

[10] Ye, S. H., Siddle, K. J., Park, D. J. & Sabeti, P. C. Benchmarking metagenomics tools for taxonomic classification. Cell **178**, 779–794 (2019).

[11] Medina, R. H. et al. Machine learning and deep learning applications in microbiome research. ISME Communications **2**, 98 (2022).

[12] Liu, B., Wei, Y., Zhang, Y. & Yang, Q. Deep neural networks for high dimension, low sample size data. In IJCAI, 2287–2293 (2017).

[13] Washburne, A. D. et al. Methods for phylogenetic analysis of microbiome data. Nature Microbiology **3**, 652–661 (2018).

[14] Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature Biotechnology **36**, 996–1004 (2018).

[15] Zhu, Q. et al. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. Nature Communications **10**, 5477 (2019).

[16] Zhu, Q. et al. Phylogeny-aware analysis of metagenome community ecology based on matched reference genomes while bypassing taxonomy. mSystems **7**, e00167–22 (2022).

[17] Ma, J. et al. Using deep learning to model the hierarchical structure and function of a cell. Nature Methods **15**, 290–298 (2018).

[18] Elmarakeby, H. A. et al. Biologically informed deep neural network for prostate cancer discovery. Nature **598**, 348–352 (2021).

[19] Aitchison, J. The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological) **44**, 139–160 (1982).

[20] Louizos, C., Welling, M. & Kingma, D. P. Learning sparse neural networks through l_0 regularization. International Conference on Learning Representations (2018).

[21] Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. International Conference on Machine Learning 448–456 (2015).

[22] Laske, C. et al. Signature of Alzheimer's disease in intestinal microbiome: Results from the AlzBiom study. Frontiers in Neuroscience **16**, 792996 (2022).

[23] Gonzalez, C. G. et al. Location-specific signatures of Crohn's disease at a multi-omics scale. Microbiome **10**, 133 (2022).

[24] Dan, Z. et al. Altered gut microbial profile is associated with abnormal metabolism activity of Autism Spectrum Disorder. Gut Microbes **11**, 1246–1267 (2020).

[25] Lloyd-Price, J. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature **569**, 655–662 (2019).

[26] (iHMP) Research Network Consortium, T. I. H. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell Host & Microbe **16**, 276–289 (2014).

[27] Pedregosa, F. et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011).

[28] Weiss, K., Khoshgoftaar, T. M. & Wang, D. A survey of transfer learning. Journal of Big data **3**, 1–40 (2016).

[29] Shrikumar, A., Greenside, P., Shcherbina, A. & Kundaje, A. Learning important features through propagating activation differences. In International Conference on Machine Learning (2017).

[30] Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013).

[31] Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (2017).

[32] Lu, Y. Y., Guo, W., Xing, X. & Noble, W. S. DANCE: Enhancing saliency maps using decoys. In International Conference on Machine Learning (2021).

[33] Lee, A. A. et al. Temporal gut microbial changes predict recurrent clostridiodes difficile infection in patients with and without ulcerative colitis. Inflammatory Bowel Diseases **26**, 1748–1758 (2020).

[34] Sasaki, K. et al. 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**, 1800555 (2019).

[35] Park, J.-S. et al. 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**, 215–224 (2018).