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

[2]School of Life Sciences, Arizona State University

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. Here we develop MIOSTONE, an accurate and interpretable neural network model that simulates a real taxonomy by encoding the relationships among microbial features. The taxonomy-encoding architecture provides a natural bridge from variations in microbial taxa abundance to variations in traits, encompassing increasingly coarse scales from species to domains. MIOSTONE has the ability to determine whether taxa within the corresponding taxonomic group provide a better explanation in a data-driven manner. MIOSTONE serves as an effective predictive model, as it not only accurately predicts microbiome-trait associations across extensive real datasets but also offers interpretability for scientific discovery. Both attributes are crucial for facilitating *in silico* investigations into the biological mechanisms underlying such associations among microbial taxa.

Microbiome | Microbiome-disease association | Biomarker detection | Interpretable AI

Correspondence: *Qiyun.Zhu@asu.edu, yanglu@uwaterloo.ca*

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 (Qin et al., 2010). As the microbiome influences the impact of host genes, microbiome genomes are often referred to as the "second genome" (Grice and Segre, 2012). Subsequently, microbiomes have been found to play pivotal roles in various aspects of human health and diseases (Sekirov et al., 2010), including diabetes (Qin et al., 2012), obesity (Turnbaugh et al., 2006), inflammatory bowel disease (Mills et al., 2022), Alzheimer's disease (Vogt et al., 2017), cancers (Battaglia et al., 2024), and more. The association of the microbiome with host traits will provide insights into the underlying mechanisms governing the microbiome's impact on human health and diseases and facilitate the development of novel therapeutic strategies.

To explore microbiome-trait associations, a primary focus has been on identifying predictive microbial markers for disease prediction from microbial samples (Medina et al., 2022). Here, a microbial sample is typically characterized by its taxonomic profile, which includes the abundance of microbial taxa at certain taxonomic levels (Pasolli et al., 2016), such as species, genus, family, and so on. However, the unique characteristics of microbiome data pose challenges in thoroughly exploring the relationships among the taxa. For example, sample-wise sequencing generates millions of short fragments from a mixture of taxa rather than an individual taxon. The dynamic and complex nature of microbial communities can lead to inaccurate taxonomic profiling, potentially resulting in inaccuracies in downstream microbiome-trait association analyses (Ye et al., 2019).

Another difficulty in analyzing microbiome data stems from its sparsity, with a substantial portion of data entries being zeros (Medina et al., 2022). These zeros can indicate either the true absence of the taxa in the environmental sample (*i.e.,* biological zeros) or the failure to detect the taxa due to low sequencing depth and sampling variation (*i.e.,* technical zeros) (Jiang et al., 2021). To address the sparsity issue, existing imputation methods are typically employed to distinguish technical zeros from biological zeros and replace technical zeros with nonzero values (Jiang et al., 2021; Zeng et al., 2022; Linderman et al., 2022). However, these methods require subjective user decisions to choose the threshold that decide which zeros require imputation and which do not (Zeng et al., 2022), inevitably diminishing the reliability and reproducibility of the downstream analyses. Moreover, imputation methods can lead to data misinterpretation by introducing the risk of bias and potentially yielding false signals (Jiang et al., 2022; Andrews and Hemberg, 2018).

Additionally, microbiome data is typically high-dimensional and noisy, with a much larger number of taxa than sample size (Kharchenko, 2021). The excessive number of taxa as features not only increases computational costs but also presents challenges in analysis due to the curse of dimensionality. (Liu et al., 2017). 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 microbiome and disease states (Knights et al., 2011; Finucane et al., 2014). To alleviate the curse of dimensionality, feature selection methods have been developed to choose highly variable taxa (Stuart et al., 2019; Hao et al., 2021; Ditzler et al., 2015). However, these methods often result in a significant loss of information from non-selected taxa (Kharchenko, 2021). Altogether, the nature of microbiome data necessitates novel analytic methods that consider factors such as data imperfections, sparsity, and the curse of dimensionality.

To address these challenges, recent studies have leveraged the inherent correlation structure among taxa as an informative prior, with the aim of enhancing disease prediction performance. Two fundamental correlation structures among taxa are taxonomy and phylogeny (Washburne et al., 2018). The taxonomy categorizes taxa into hierarchical groups, spanning from the three domains (Bacteria, Archaea, and Eukarya) down to species (Parks et al., 2018), using a well-established and widely accepted naming system. The phylogeny aims to encode the evolutionary relationships among taxa and classifies taxa by a series of splits, corresponding to estimated events in which two lineages split from a common ancestor to form distinct species (Washburne et al., 2018). The distinction between taxonomy and phylogeny lies in the fact that taxonomy is coarser, with taxonomic labels categorizing only a small fraction of the branches in the phylogeny, whereas the phylogeny provides a more detailed scaffold. Both taxonomy and phylogeny have been utilized to incorporate relevant structural knowledge among taxa into existing predictive models.

For example, phylogeny can serve as a smoothness regularizer to enhance linear regression models (Xiao et al., 2018). Utilizing phylogeny can also aid in weighting and prioritizing the most relevant taxa, thus enhancing the prediction accuracy of random forest models (Albanese et al., 2015). More recently, several methods have incorporated phylogeny or taxonomy into the preprocessing stage of employing advanced analysis tools like deep neural networks (DNN) (Sharma et al., 2020; Reiman et al., 2020; Li et al., 2021; Shtossel et al., 2023; Wang et al., 2008). Specifically, these preprocessing steps involve aggregating taxa into distinct taxonomic clusters (Sharma et al., 2020; Wang et al., 2008), reordering the taxa spatially based on their phylogenetic structure (Reiman et al., 2020; Shtossel et al., 2023), and assigning varying weights according to phylogenetic distances (Li et al., 2021). However, these methods may underutilize the relationships inherent in phylogeny or taxonomy by confining their 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 (Ma et al., 2018) – an aspect that is critical for a scientific understanding of microbiome-trait associations.

In this study, we introduce MIOSTONE (MIcrObiome-trait aSsociations with TaxONomy-adaptivE neural networks), an accurate and interpretable neural network model that simulates a real taxonomy by encoding the relationships among microbial features (Fig. 1(A)). Drawing inspiration from biologically-informed DNNs (Ma et al., 2018; Elmarakeby et al., 2021), the model organizes the neural network into layers to explicitly emulate the taxonomic hierarchy within its architecture, based on the Genome Taxonomy Database (GTDB) (Parks et al., 2018), spanning 124 phyla, 320 classes, 914 orders, $2,057$ families, $6,811$ genera, and $12,258$ species. In this taxonomy-encoding network, each neuron represents a specific taxonomic group, with connections between neurons symbolizing the hierarchical subordination relationships among these groups. These hierarchies provide a natural bridge from variations in microbial taxa abundance to variations in traits, encompassing increasingly coarse scales from species to domains.

The key 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. Such taxonomy-adaptive strategy is achieved during the training phase, where variations in microbial taxa propagate individually through the hierarchy to impact the parent taxonomic group that contain them, competing against the aggregation of the parent as a whole for better trait prediction (Fig. 1(B)). The taxonomy-encoding design 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 to demonstrate its empirical utility (Fig. 1(C)). From a practitioner's perspective, MIOSTONE serves as an effective predictive model, as it not only accurately predict microbiome-trait associations but also be interpretable. Both attributes are crucial for facilitating *in silico* investigations into the biological mechanisms underlying such associations among microbial taxa.

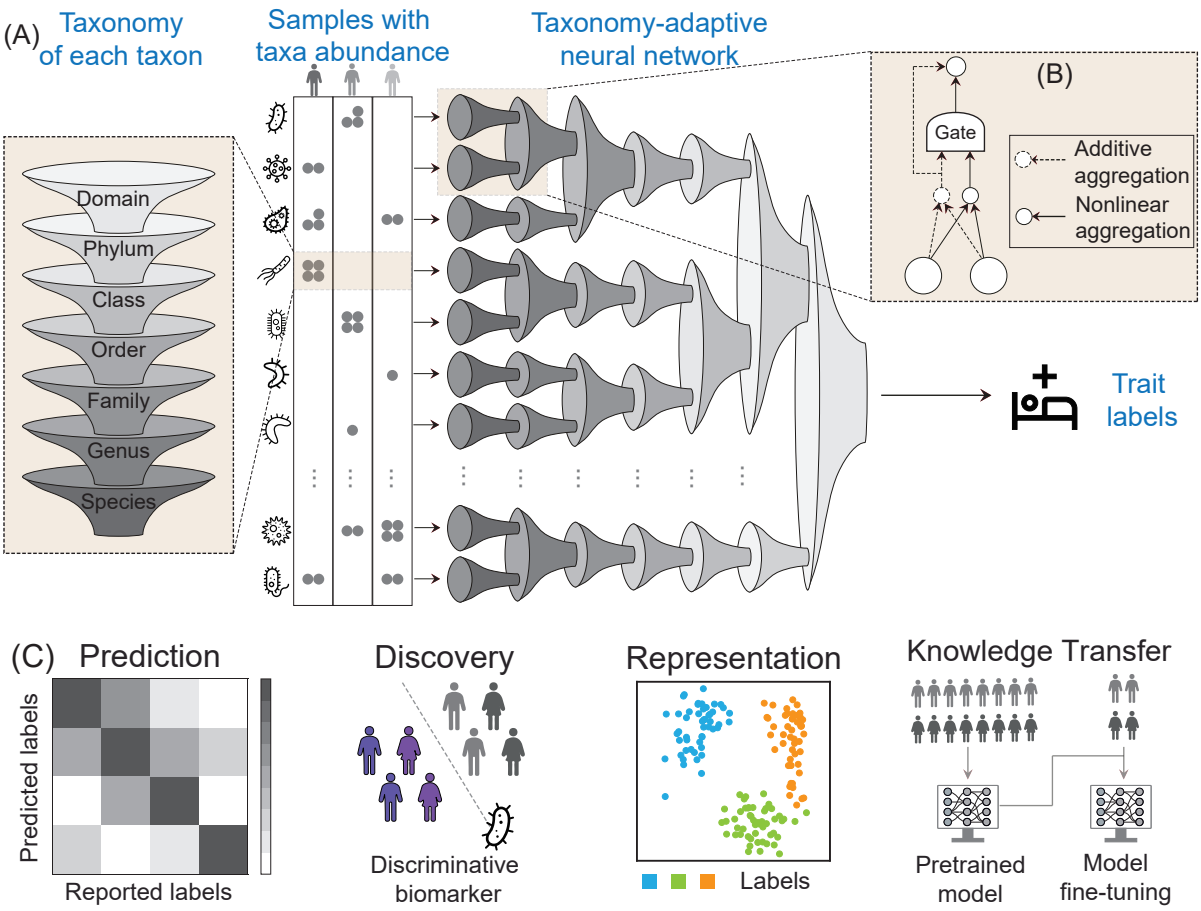

Figure 1. Overview of MIOSTONE. (A) MIOSTONE designs the neural network with layered architecture, explicitly mirroring the taxonomic hierarchy. Each neuron represents a distinct taxonomic group, while connections between neurons signify hierarchical subordination relationships among these groups. By contrast, a conventional neural network functions as a black box, lacking explicit encoding of knowledge within its architecture. (B) Each internal neuron in MIOSTONE has the capability to discern whether taxa within the corresponding taxonomic group provide a more effective explanation of the trait when assessed either holistically (*i.e.,* additively) as a group or individually (*i.e.,* non-linearly) as distinct taxa. (C) MIOSTONE establishes a versatile microbiome data analysis pipeline, applicable to a variety of tasks including disease status prediction, microbiome representation learning, microbiome-disease association identification, and enhancement of predictive performance in tasks with limited samples through knowledge transfer.

Figure 2. MIOSTONE provides accurate predictions the host's disease status. (A) The evaluation was performed on seven publicly available microbiome datasets with varying microbial taxa sizes, covering different proportions of taxonomic levels. (B-C) MIOSTONE is compared against five baseline methods: random forest (RF), support vector machine (SVM) with linear kernel, multi-layer perceptron (MLP), TaxoNN, and PopPhy-CNN. Each model was trained by times using different train-test splits, and reported by the average performance along with 95% confidence intervals. The models' performances are measured by the Area Under the Precision-Recall Curve (AUPRC). For scientific rigor, the performance comparison between MIOSTONE and any other baseline method is quantified using one-tailed two-sample t-tests to calculate p-values: $****$p-value ≤ 0.0001; $***$p-value ≤ 0.001; $**$p-value ≤ 0.01; $*$p-value ≤ 0.05. This performance comparison is further bolstered by qualitative visualization. (D-E) The same setting as (B) and (C), but using the Area Under the Receiver Operating Characteristic Curve (AUROC) as the metric.

Results

MIOSTONE provides accurate predictions of the host's disease status

We benchmarked MIOSTONE with the other five baseline methods (See Baselines for details). Among these, machine learning methods such as Random Forest (RF) and Support Vector Machine (SVM) with linear kernel are popular choices for disease prediction (Medina et al., 2022). Notably, PopPhy-CNN (Reiman et al., 2020) and TaxoNN (Sharma et al., 2020) are also DNN-based methods adept at leveraging phylogenetic or taxonomic structure in microbial taxa for improved disease prediction.

We evaluated the performance of MIOSTONE on seven publicly available microbiome datasets (See Datasets for details), each characterized by different sample sizes and feature dimensionality. These datasets vary in microbial taxa sizes, encompassing notably different proportions of taxonomic levels (Fig. 2(A)). To quantify the predictive performance of different methods, we utilized two metrics: the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area Under the Precision-Recall Curve (AUPRC). Both metrics are commonly employed in assessing the performance of binary classifiers, with AUPRC being particularly suitable for imbalanced datasets (Davis and Goadrich, 2006). We assessed the performance through 5-fold cross-validation, training individual models 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 AUPRC scores were computed on the full dataset. For robustness, we repeated this process 20 times with different random seeds and reported the mean performance with 95% confidence intervals.

Our analysis shows that MIOSTONE considerably outperformed other baseline methods across all datasets, in terms of AUPRC (Fig. 2(B)) and AUROC scores (Fig. 2(D)). For scientific rigor, the performance comparison between MIOSTONE and any other baseline method is quantified using one-tailed two-sample t-tests to calculate p-values. These quantitative p-values, along with qualitative visualizations (Fig. 2(C) for AUPRC and Fig. 2(E) for AUROC), confirm that the performance superiority of MIOSTONE is statistically significant and qualitatively discernible. The superior performance across diverse microbiome datasets and disease models indicates its robustness and generalizability. In contrast, the other two DNN-based baseline methods, PopPhy-CNN and TaxoNN, showed varying performance across different datasets. For example, PopPhy-CNN performed nearly as well as the best in the RUMC dataset but was among the worst performers in the ASD dataset. On the contrary, TaxoNN achieved nearly perfect AUROC and AUPRC scores in the GD dataset but performed among the worst in the RUMC dataset, in contrast to PopPhy-CNN's performance.

Finally, 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.

Dissecting the performance of MIOSTONE

We first assessed the computational efficiency of MIOSTONE in comparison to the other baseline methods (Fig. 3(A)). For a fair comparison, all models were trained and tested under the same environment: AMD EPYC 7302 16-Core Processor, NVIDIA RTX A6000, with 32GB DDR4 RAM. We found that MIOSTONE is relatively efficient to train, comparable to training an MLP classifier, which is the simplest DNN-based method. It's worth noting that RF and SVM classifiers, as provided by Scikit-learn, remain the most efficient models. However, their training times escalate much more rapidly than those of DNN-based methods when the number of samples becomes large. While DNN-based methods typically exhibit slower training times compared to non-DNN methods, the relatively small size of microbiome datasets ensures that even the slowest DNN-based methods require only minutes for training. Given the enhanced predictive performance, we assert that computational cost will not impede the applicability of MIOSTONE (For optimized implementation details, see Implementation and Fig. 3(F).).

One major challenge in analyzing microbiome data arises from the curse of dimensionality (Liu et al., 2017), where the number of samples is relatively low, and the number of microbiome taxa is very large. One may naturally wonder if we can mitigate the curse of dimensionality by only utilizing a subset of highly informative features. To answer this question, 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 (Stuart et al., 2019) in single-cell RNA-seq, as implemented by the Scanpy python package (Wolf et al., 2018). For each taxa subset, we trained and evaluated MIOSTONE on seven real microbiome datasets using the same settings as for the full set. We found that MIOSTONE trained with all microbiome features either outperforms or matches the performance of the one trained with a subset of highly variable taxa on most of the datasets, with ASD being the only exceptions (Fig. 3(B)). 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.

Finally, the design of MIOSTONE incorporates several key components, including the taxonomy-encoding DNN architecture, the data-driven aggregation of neuron representations, and the taxonomy-dependent

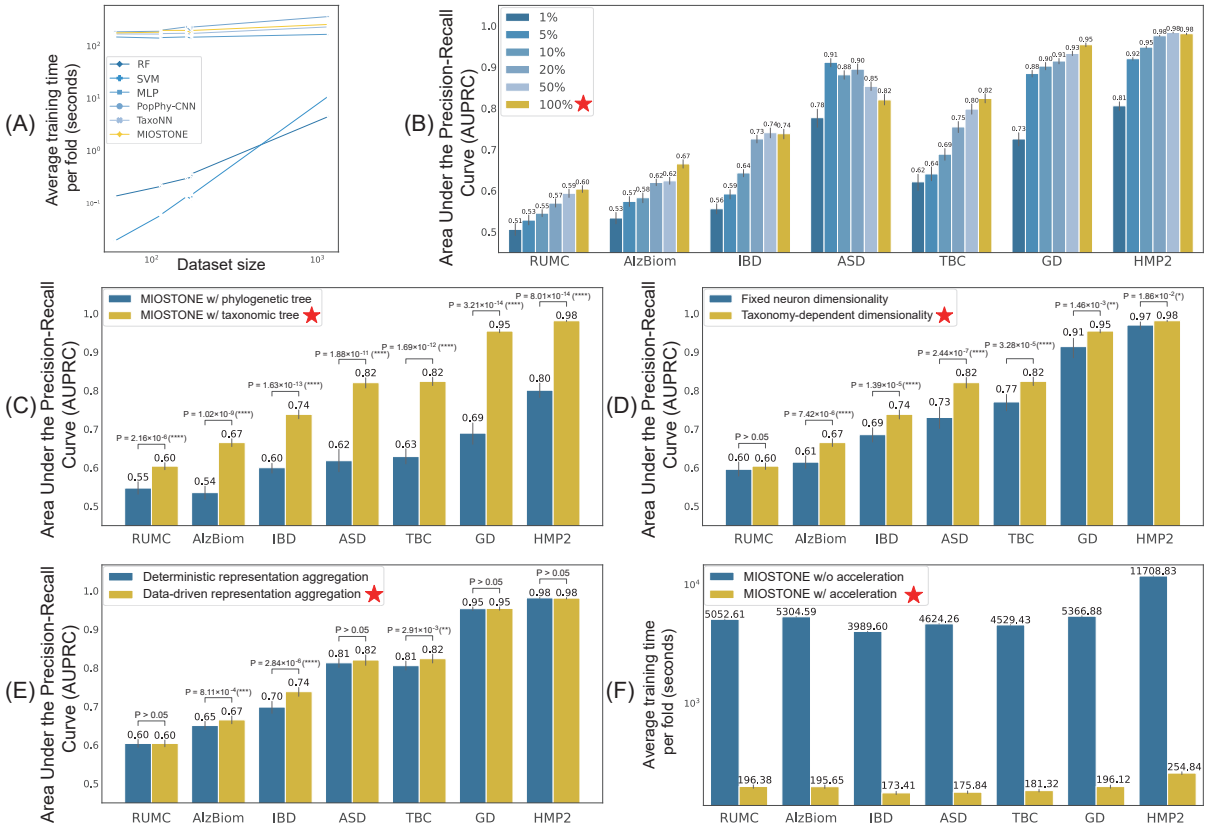

Figure 3. Dissecting the performance of MIOSTONE through control studies (The settings used by MIOSTONE are marked by ★). (A) MIOSTONE is efficient to train, comparable to training other DNN-based methods. While RF and SVM remain the most efficient models compared to any DNN-based methods when microbiome datasets are small, their training times increase much more rapidly than those of DNN-based methods when the number of samples becomes large. (B) The curse of dimensionality cannot simply be mitigated using feature selection. MIOSTONE trained with all microbiome features either outperforms or matches the performance of the one trained with a subset of highly variable taxa on most of the datasets. (C) While MIOSTONE can emulate any hierarchical correlation among taxa within its architecture, alternatives such as taxonomy-encoding DNN architecture perform significantly worse than the phylogeny-encoding one. (D) MIOSTONE's assigning larger taxonomic groups with greater representation dimensionality can aid in capturing more complex biological patterns to predict traits, compared to using fixed representation dimensionality. (E) MIOSTONE's data-driven aggregation of neuron representations outperforms or matches the deterministic selection of nonlinear representation on most of the datasets. This underscores MIOSTONE's key novelty in discerning whether taxa within the corresponding taxonomic group provide a more effective explanation of the trait when evaluated holistically as a group or individually as distinct taxa. (F) MIOSTONE is implemented by employing a fully connected DNN with additional pruning, which is significantly more efficient than naively modeling with customized taxonomic connections.

internal neuron dimensionality. To assess the impact of each component on disease prediction, we conducted a control study in which we modified MIOSTONE by replacing its components with alternative solutions. Specifically, we considered 3 variants of MIOSTONE: (1) Replacing the taxonomy-encoding DNN architecture with a phylogeny-encoding alternative; (2) Setting the taxonomy-dependent internal neuron dimensionality to a fixed dimension of 2 (Fig. 3(D)); and (3) Making the data-driven aggregation of neuron representations deterministic by disabling stochastic gating and opting for nonlinear representation only (Fig. 3(E)). For each variant, we applied MIOSTONE to seven real microbiome datasets using the same settings.

The results indicate that all key components positively contribute to MIOSTONE's performance. In the first study (Fig. 3(C)), while MIOSTONE can emulate any hierarchical correlation among taxa within its architecture, alternatives such as phylogeny-encoding DNN architecture perform significantly worse than the taxonomy-encoding one. This suggests that phylogeny, as a more detailed scaffold for microbial classification, may present additional challenges during training compared to taxonomy. In the second study (Fig. 3(D)), the taxonomy-dependent internal neuron dimensionality consistently outperforms the fixed dimensionality approach across all seven microbiome datasets, irrespective of sample sizes and feature dimensionality. This suggests that MIOSTONE's assigning larger taxonomic groups with greater representation dimensionality can aid in capturing more complex biological patterns to predict traits, compared to using fixed representation dimensionality. In the last study (Fig. 3(E)), the data-driven aggregation of neuron representations outperforms or matches the deterministic selection of nonlinear representation across all seven microbiome datasets. This underscores MIOSTONE's key novelty in discerning whether taxa within the corresponding taxonomic group provide a more effective explanation of the trait when evaluated holistically as a group or individually as distinct taxa.

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

In microbiome data analysis, the transfer learning paradigm (Weiss et al., 2016) 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 large HMP2 dataset and then utilized it for the smaller IBD dataset prediction task in three settings: (1) Directly employing 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). It's worth noting 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, in terms of AUPRC (Fig. 4(A)) and AUROC scores (Fig. 4(C)). Our analysis demonstrates that fine-tuning enhances predictive performance in MIOSTONE compared to training from scratch. For scientific rigor, the performance between fine-tuning and training from scratch is quantified using one-tailed two-sample t-tests to calculate p-values. These quantitative p-values, along with qualitative visualizations (Fig. 4(B) for AUPRC and Fig. 4(D) for AUROC), confirm that the performance superiority of fine-tuning is statistically significant and qualitatively discernible than training from scratch.

In contrast, the other three DNN-based baseline methods, MLP, PopPhy-CNN, and TaxoNN, exhibited varying degrees of 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 exhibit a noticeable improvement with the pre-trained model, suggesting that the knowledge learned

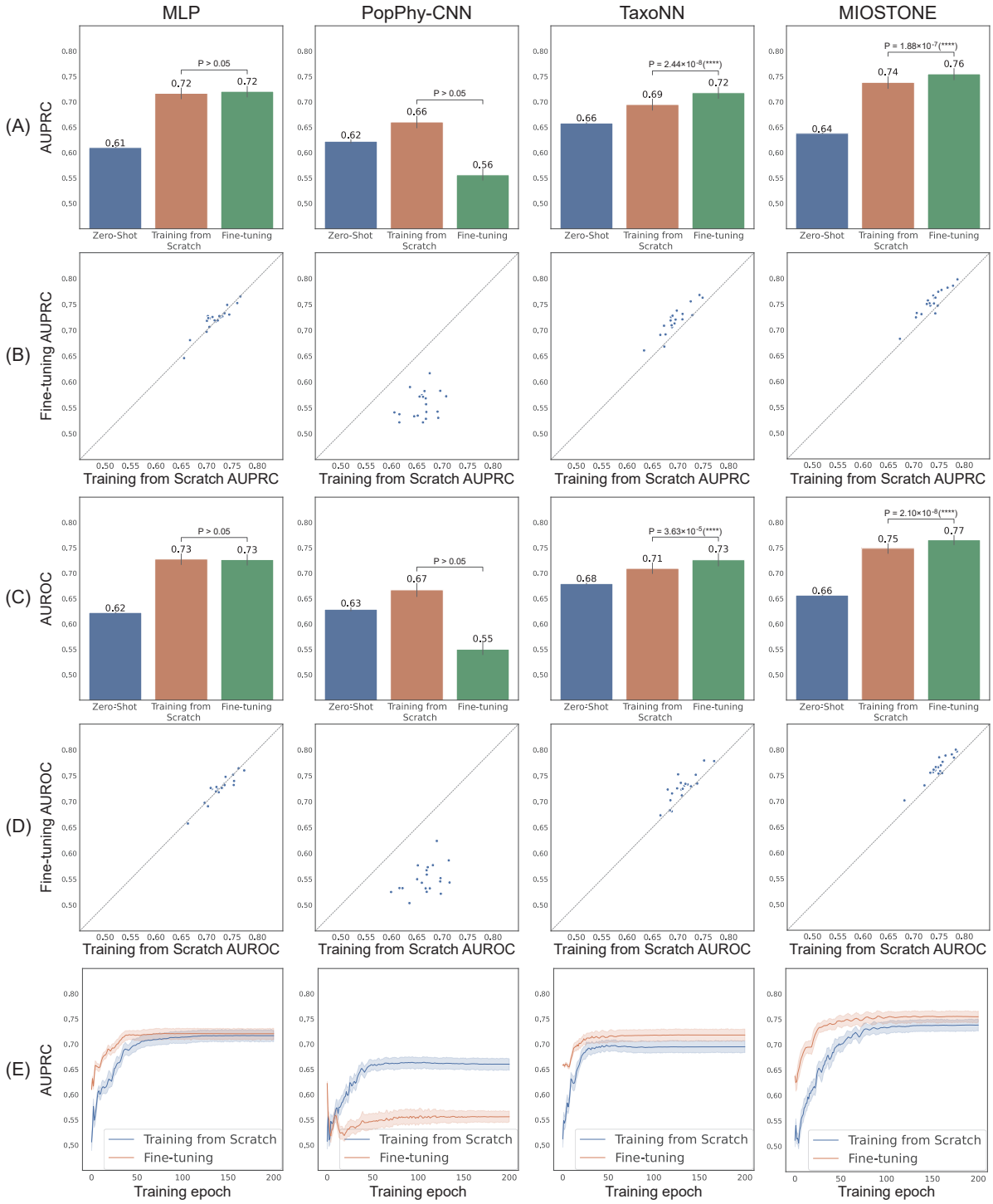

Figure 4. MIOSTONE enhances disease prediction by transferring knowledge from pre-trained models. A model on the large HMP2 dataset is pre-trained and then employed for the smaller IBD dataset in three settings: direct prediction on IBD (*i.e.,* zero-shot), fine-tuning on IBD, and training IBD from scratch. Only DNN-based methods are included for comparison because RF and SVM are not well-suited for fine-tuning. (A-B) The prediction is conducted across three settings 20 times with varied train-test splits, and reported by the average performance assessed by the Area Under the Precision-Recall Curve (AUPRC), along with 95% confidence intervals. For scientific rigor, the performance between fine-tuning and training from scratch is quantified using one-tailed two-sample t-tests to calculate p-values. This performance comparison is further bolstered by qualitative visualization. (C-D) The same setting as (A) and (B), but using the Area Under the Receiver Operating Characteristic Curve (AUROC) as the metric. (E) The training dynamics of various models by comparing fine-tuning and training from scratch, analyzing AUPRC on test splits across different training epochs. Fine-tuning required significantly fewer training epochs to achieve better performance compared to training from scratch in MIOSTONE.

by the pre-trained model was completely overwritten by fine-tuning. Furthermore, PopPhy-CNN exhibited a reverse effect, experiencing a 15.2% decline in AUPRC and a 17.9% decline in AUROC scores with fine-tuning compared to training from scratch. This suggests that the knowledge learned by the pre-trained model may even hinder learning from the new data.

Finally, we examined the MIOSTONE training dynamics by comparing fine-tuning with training from scratch, focusing on the reported performance on test splits across different training epochs (Fig. 4(E)). We found that MIOSTONE's 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. Since MIOSTONE is already computationally efficient, this approach can further reduce training time. Although fine-tuning generally achieves better training dynamics than training from scratch, PopPhy-CNN's degraded performance can be discerned from its training dynamics. We conclude that MIOSTONE effectively improves disease prediction through knowledge transfer via fine-tuning.

MIOSTONE learns meaningful and discriminative sample representations

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. Representation learning is renowned for extracting meaningful high-level semantics from raw data and has been widely employed to uncover hidden patterns in biological and biomedical data (Bengio et al., 2013; Iuchi et al., 2021).

We extracted MIOSTONE's internal neuron representations from three taxonomic levels (species, genus, and order), comparing them against other DNN-based methods. Given the drastically different DNN architectures of these methods, for a fair comparison, we extracted the last-layer latent representations, believed to encode the maximum semantic meanings, of these DNN-based methods. We projected these extracted representations from all methods, encoding the semantic meanings of input microbial samples, into a two-dimensional embedding space using Principal Component Analysis (PCA). We subsequently assessed the effectiveness of these representations in differentiating between different disease statuses.

The patient samples with different disease statuses cannot be distinguished initially, considering that microbiome data is typically high-dimensional and noisy. For example, the PCA visualization of microbiome features based on taxa profiling fails to distinguish IBD disease subtypes, as samples from Crohn's disease (CD) and ulcerative colitis (UC) patients are mixed together. Other DNN-based methods, such as MLP, PopPhy-CNN, and TaxoNN, demonstrated varying degrees of improvement, albeit marginal, in distinguishing between the two IBD disease subtypes (Fig. 5). However, when we represented each patient sample by MIOSTONE's internal neuron representations from even the bottom taxonomic level with least encoded semantics (*i.e.,* species), the resulting representations exhibited significantly improved separation among disease subtypes, suggesting that the model's internal representations effectively capture diverse disease-specific signatures. The separation between disease subtypes is quantitatively measured by the silhouette value (Rousseeuw, 1987). MIOSTONE's internal neuron representations exhibit higher silhouette values, suggesting greater similarity of each sample to its own disease subtype compared to other subtypes.

MIOSTONE identifies microbiome-disease associations with high interpretability

Recognizing the model's potential in capturing disease-specific semantics, we further delved into the MIOSTONE model to uncover significant microbiome-disease associations. Important associations were scored using feature attribution methods, which assign importance scores to taxonomic groups, with higher scores indicating greater importance to the model's prediction. In this study, we employed three representative model-agnostic feature attribution methods, DeepLIFT, integrated gradient, and SHAP, to elucidate the relationship between microbiome taxa and disease trait without assuming any specific model architecture (See Interpretation for details). We discovered that three representative feature attribution methods demonstrate strong consistency in quantifying crucial microbiome-disease associations from the

Figure 5. MIOSTONE learns meaningful and discriminative representations. The MIOSTONE internal neuron representations of samples are projected onto a two-dimensional Principal Component space and evaluated their efficacy in distinguishing between different disease subtypes. A qualitative evaluation of the MIOSTONE representations is conducted at three taxonomic levels (species, genus, and order), compared against the last-layer latent representations of other DNN-based methods. MIOSTONE representations exhibited significantly improved separation between disease subtypes, suggesting that the model's internal representations effectively capture diverse disease-specific signatures. The separation between disease subtypes is quantitatively measured by the silhouette value.

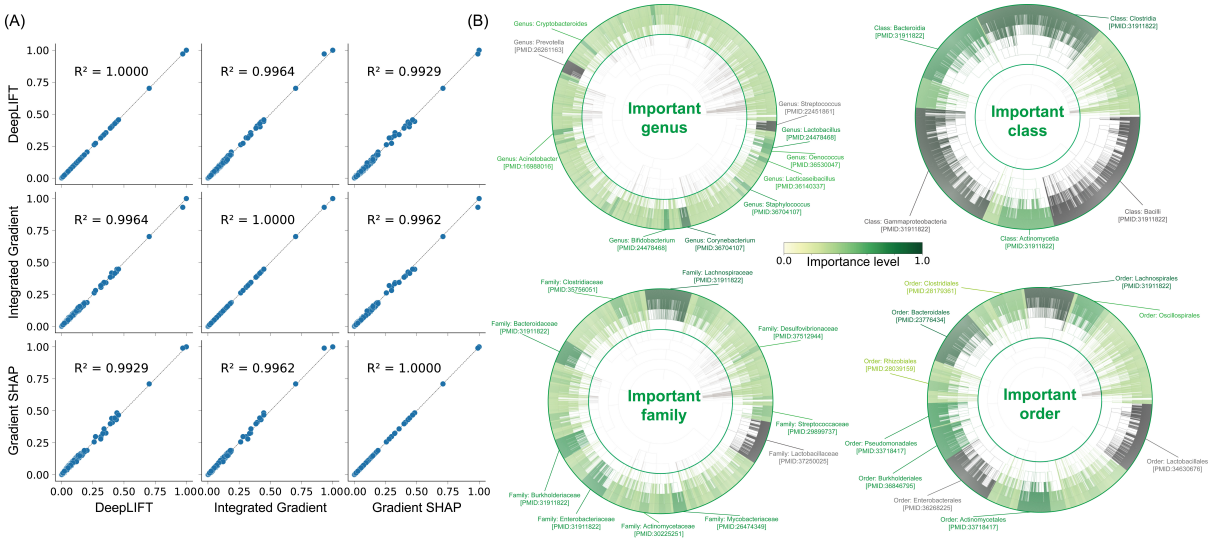

Figure 6. MIOSTONE discovers microbiome-disease associations across different taxonomic levels. The relationships between microbiome taxa and disease traits are measured using feature attribution methods to interpret the MIOSTONE model. (A) Three representative feature attribution methods, DeepLIFT, Integrated Gradient, and SHAP, exhibit strong consistency in quantifying crucial microbiome-disease associations from the trained MIOSTONE model. (B) Feature importance attribution by DeepLIFT at the genus-level, family-level, order-level, and class-level taxonomic groups, respectively. When highlighting an important taxonomic group, the taxonomic subtree rooted at that group, including group-specific taxa, is also highlighted for better visualization. The top-ranked taxonomic groups are highlighted with their respective names, supported by literature evidence with accompanying PubMed identifiers.

trained MIOSTONE model (Fig. 6(A)). Therefore, the consistency justifies presenting only the DeepLIFT results hereafter.

We initially focused on identifying important microbiome-disease associations in differentiating between two IBD disease subtypes: Crohn's disease (CD) and ulcerative colitis (UC), at the genus level. We highlighted top-ranked genera reported by DeepLIFT with their respective names, supported by literature evidence with accompanying PubMed identifiers (Fig. 6(B)). For example, the microbial community associated with *Prevotella* has been reported to have significantly different abundance in UC compared to controls as well as CD (Kabeerdoss et al., 2015). Furthermore, studies have reported that the detection frequency of *Streptococcus* in UC patients was significantly higher than in healthy subjects. Infection with highly virulent specific types of Streptococcus might be a potential risk factor in the aggravation of UC (Kojima et al., 2012).

The identification of significant microbiome-disease associations in distinguishing between IBD disease subtypes can be further extended to coarser resolutions such as family-level, order-level, and class-level. For example, the *Lachnospiraceae* family, predominantly found in the gut microbiota of mammals and humans, has been reported to have significantly different abundance between IBD disease and health controls (Lee et al., 2020; Sasaki et al., 2019). Moreover, studies have reported increased levels of the *Bacilli* and *Clostridia* classes in UC patients, while the levels of *Clostridia* and *Bacteroidia* are decreased in CD patients (Alam et al., 2020). We conclude that MIOSTONE effectively identifies microbiome-disease associations across different taxonomic levels, providing valuable insights for scientific discovery.

Discussion

In this study, we propose MIOSTONE, an accurate and interpretable machine learning method for investigating microbiome-trait associations. At its core, MIOSTONE leverages the intercorrelation of microbiome features based on their taxonomic relationships. The key novelties of SCOT are threefold: (1) the taxonomy-encoding architecture harnesses the capabilities of DNNs 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 facilitates the understanding of microbiome-trait associations. We 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.

Methodologically, our approach provides a systematic way to circumvent the fundamental computational challenges in the conventional analysis of microbiome data. First of all, the curse of dimensionality has long been a dilemma in computational modeling. Specifically, the relatively low number of samples may lead to overfitting during training, thereby impeding the use of powerful analysis tools like DNNs at the expense of prediction accuracy. The hierarchical neural network framework adopted by MIOSTONE significantly reduces the model's complexity, thereby mitigating the curse of dimensionality and overfitting. Moreover, the biologically-informed neural networks are highly generic and expressive, allowing for the representation of any hierarchical relationships or functional dependencies among microbial taxa. By incorporating this knowledge into biologically-informed neural networks, we can attribute the information encoded by the data to these pre-specified biological concepts, offering a natural interpretation of the model's internal mechanisms.

Numerous studies have focused on accurately differentiating disease states and understanding the differences in microbiome profiles between healthy and ill individuals. Most of them primarily focus on various statistical approaches, without explicitly modeling the underlying molecular mechanisms that give rise to nonlinearity and microbe-microbe interactions among a large number of microbial taxa, which in principle drive microbiome dynamics. We hypothesized that this might be due to the fact that the curse of dimensionality already makes first-order association identification highly challenging, let alone the detection of higher-order interactions, which is a much more difficult task. Given that the curse of dimensionality has been systematically mitigated, a potential research direction for future studies is to quantify important higher-order microbe-microbe interactions instead of focusing solely on an individual taxon. This could involve using feature interaction detection methods developed in the interpretable machine learning community (Tsang et al., 2018; Chen et al., 2023).

Lastly, the evaluation of detected important microbiome-disease associations or microbe-microbe interactions relies solely on literature support. While this approach is reasonable for evaluation purposes, it might limit the credibility for less studied taxa. A potential research direction for future studies is to provide confidence estimation for the top-ranked microbiome-disease associations to complement the literature support, using measures such as q-values (Storey, 2003), with the assistance of the recently proposed knockoffs framework (Barber and Candès, 2015; Lu et al., 2018).

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.

Methods

Datasets used for evaluations

We collected seven publicly available microbiome datasets generated using whole-metagenome shotgun sequencing (WMS), with varying sample sizes and feature dimensionality. The AlzBiom dataset (Laske et al., 2022) 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 $8,350$ taxa. The ASD dataset (Dan et al., 2020) 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 $7,287$ taxa. The GD dataset (Zhu et al., 2021) explored the relationship between the gut microbiome and Graves' disease (GD). It comprises 100 GD samples and 62 healthy control samples, profiled with $8,487$ taxa. The TBC and RUMC datasets (Boktor et al., 2023) 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 $6,227$ taxa. The RUMC cohort comprises 42 PD samples and 72 healthy control samples, profiled with $7,256$ taxa. The IBD dataset (Gonzalez et al., 2022) 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 $5,287$ taxa. The HMP2 dataset (Lloyd-Price et al., 2019) in the Integrative Human Microbiome Project (iHMP) 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 $1,158$ (728 CD and 430 UC samples), with an expanded taxa set of size $10,614$. Full details regarding these datasets are available in the Supplementary Table S.1.

Dataset preprocessing

The abundance features in these datasets were profiled using the standard operating procedure for shotgun metagenomics implemented in the Qiita platform (Gonzalez et al., 2018). Specifically, the raw sequencing reads were processed using fastp and Minimap2 to remove low-quality sequences, adapter sequences and sequences that are susceptible to host contamination (Armstrong et al., 2022). The processed reads were classified using Woltka v0.1.4 (Zhu et al., 2022) against the Web of Life (WoL) v2 database (Zhu et al., 2019), which includes $15,953$ microbial genomes. WoL offers taxonomic annotations, which are based on the Genome Taxonomy Database (GTDB) (Parks et al., 2018), for these microbial genomes, spanning 124 phyla, 320 classes, 914 orders, $2,057$ families, $6,811$ genera, and $12,258$ species. Given that each dataset only profiles a subset of taxa within the WoL taxonomy, the MIOSTONE model is constructed using a pruned taxonomy tree tailored for each dataset.

MIOSTONE design

MIOSTONE trains a deep neural network to predict disease traits from microbial taxa abundance profiles, with its architecture that precisely mirrors the taxonomic hierarchy based on the Genome Taxonomy Database (GTDB) (Parks et al., 2018). Each neuron in the network corresponds to a specific taxonomic group, and the connections between neurons represent the subordination relationships between these groups, such as "species A belongs to genus B" or "genus B belongs to family C" relationships. Unlike fully connected DNNs, MIOSTONE only connects neurons that are directly related in the taxonomic hierarchy. This design choice significantly reduces the model's complexity, effectively mitigating the overfitting problem, while simultaneously enhancing its interpretability.

We denote our input training data set as $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$, where n is the number of samples. For each sample i, $x_i \in \mathbb{R}^p$ represents the p-dimensional profiled abundance of microbial taxa as features, and $y_i \in \mathbb{R}$ denotes the corresponding trait label, which can be either binary (*e.g.,* disease status) or continuous (*e.g.,* age). Our goal is to learn a predictive function $\mathbb{R}^p \mapsto \mathbb{R}$, parameterized by a DNN, that accurately predict the trait label $y \in \mathbb{R}$ for the microbiome sample $x \in \mathbb{R}^p$.

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 tackle this issue, MIOSTONE employs a data-driven strategy to determine whether taxa within the corresponding

taxonomic group provide a better explanation for the disease traits 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 (Louizos et al., 2018) for each internal neuron (Fig. 1(B)). Specifically, an internal neuron v 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 $u_1, u_2, \cdots$ are the children of v and 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:

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

During the training phase, the additive representation and the nonlinear representation compete against each other for improved trait prediction through a stochastic gate $m_v \in (0,1)$ that combines 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 (Louizos et al., 2018), 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 each internal neuron v recursively as $d_v = \alpha^{L_v} \cdot \sum_{u \in \text{children}(v)} d_u$, where children$(v)$ denotes the children of v. 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_{root}^N. We then apply a batch normalization layer (Ioffe and Szegedy, 2015) to I_{root}^N, and feed it into a MLP classifier to predict the trait label y: Batch normalization (BN) (Ioffe and Szegedy, 2015) 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:

$$\text{CrossEntropy}\left(y, \text{Softmax}(\text{MLP}(\text{BN}(I_{\text{root}}^N)))\right) \tag{5}$$

MIOSTONE implementation and training

The MIOSTONE model is designed to only connect neurons that are directly related in the taxonomic hierarchy. Unlike fully connected DNNs, which are efficiently optimized for parallel training by off-the-shelf deep learning libraries, the MIOSTONE model with customized hierarchical connections lacks similar optimization. In essence, pursuing interpretability in model architecture may come at the cost of computational efficiency. This is due to the need to calculate and sequentially backpropagate gradients with respect to model parameters layer-by-layer, without leveraging the speed-up provided by parallelism. To attain both interpretability and computational efficiency concurrently, we construct an equivalent MIOSTONE model by employing a fully connected DNN with additional pruning. Specifically, we initially construct a 7-layer fully connected DNN, with each layer corresponding to a specific taxonomic level. We subsequently utilize pruning techniques (Han et al., 2015) to remove connections between neurons that are not relevant in the taxonomic hierarchy, retaining only the taxonomy-encoding connections. This approach ensures that the pruned DNN not only maintains interpretability in terms of taxonomic hierarchy but is also efficiently optimized for parallel training by off-the-shelf deep learning libraries (Fig. 3(F)).

To train the MIOSTONE model, we initialize all weights uniformly at random. We optimize the objective function (Eq. 5) using ADAM with decoupled weight decay (Loshchilov and Hutter, 2019), a commonly-used stochastic gradient descent algorithm, with an initial learning rate of 0.001. Given the small size of microbiome datasets, we process the entire dataset in each batch and train for 200 epochs. We implemented MIOSTONE using the PyTorch library (Paszke et al., 2019) on NVIDIA RTX A6000 GPUs.

Alternative microbiome-trait prediction methods

We evaluate 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 (Sharma et al., 2020), and PopPhy-CNN (Reiman et al., 2020). For the RF and SVM classifiers, we employed the implementation provided by Scikit-learn (Pedregosa et al., 2011) 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 (Aitchison, 1982).

Model interpretation

We evaluate the potential of MIOSTONE in uncovering significant microbiome-disease associations. We quantify important associations using feature attribution methods, which assign importance scores to taxonomic groups, with higher scores indicating greater importance to the model's prediction. In this study, we employ three representative model-agnostic feature attribution methods to elucidate the relationship between microbiome taxa and disease trait without assuming any specific model architecture. In brief, the first method, DeepLIFT (Shrikumar et al., 2017), compares the activation of each neuron to its reference activation and assigns contribution scores based on the disparity. The second method, integrated gradient (Sundararajan et al., 2017), involves summing over the gradient with respect to different scaled versions of the input. And the third method, SHAP (Lundberg and Lee, 2017), explains the contribution of each feature to the prediction based on the game-theoretically optimal Shapley values. However, in principle, any off-the-shelf feature attribution methods can be utilized.

Code and data availability

The software implementation and all models described in this study are available at `https://github.com/batmen-lab/MIOSTONE`. All public datasets used for evaluating the models are available at `https://github.com/batmen-lab/MIOSTONE/data`.

Bibliography

Aitchison, J. (1982). The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological), 44(2):139–160.

Alam, M. T., Amos, G. C., Murphy, A. R., Murch, S., Wellington, E. M., and Arasaradnam, R. P. (2020). Microbial imbalance in inflammatory bowel disease patients at different taxonomic levels. Gut Pathogens, 12:1–8.

Albanese, D., Filippo, C. D., Cavalieri, D., and Donati, C. (2015). Explaining diversity in metagenomic datasets by phylogenetic-based feature weighting. PLoS Computational Biology, 11(3):e1004186.

Andrews, T. S. and Hemberg, M. (2018). False signals induced by single-cell imputation. F1000Research, 7.

Armstrong, G., Martino, C., Morris, J., Khaleghi, B., Kang, J., DeReus, J., Zhu, Q., Roush, D., McDonald, D., Gonazlez, A., et al. (2022). Swapping metagenomics preprocessing pipeline components offers speed and sensitivity increases. mSystems, 7(2):e01378–21.

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. The Annals of Statistics, 43(5): 2055–2085.

Battaglia, T. W., Mimpen, I. L., Traets, J. J., van Hoeck, A., Zeverijn, L. J., Geurts, B. S., de Wit, G. F., Noë, M., Hofland, I., Vos, J. L., et al. (2024). A pan-cancer analysis of the microbiome in metastatic cancer. Cell, 187(9):2324–2335.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8):1798–1828.

Boktor, J., Sharon, G., Metman, L. V., Hall, D. A., Engen, P. A., Zreloff, Z., Hakim, D. J., Bostick, J. W., Ousey, J., Lange, D., et al. (2023). Integrated multi-cohort analysis of the Parkinson's disease gut metagenome. Movement Disorders, 38(3):399–409.

Chen, W., Noble, W. S., and Lu, Y. Y. (2023). DeepROCK: Error-controlled interaction detection in deep neural networks. arXiv preprint arXiv:2309.15319.

Dan, Z., Mao, X., Liu, Q., Guo, M., Zhuang, Y., Liu, Z., Chen, K., Chen, J., Xu, R., Tang, J., et al. (2020). Altered gut microbial profile is associated with abnormal metabolism activity of Autism Spectrum Disorder. Gut Microbes, 11(5): 1246–1267.

Davis, J. and Goadrich, M. The relationship between precision-recall and ROC curves. In International Conference on Machine Learning, pages 233–240, (2006).

Ditzler, G., Morrison, J. C., Lan, Y., and Rosen, G. L. (2015). Fizzy: feature subset selection for metagenomics. BMC Bioinformatics, 16:1–8.

Elmarakeby, H. A., Hwang, J., Arafeh, R., Crowdis, J., Gang, S., Liu, D., AlDubayan, S. H., Salari, K., Kregel, S., Richter, C., et al. (2021). Biologically informed deep neural network for prostate cancer discovery. Nature, 598(7880):348–352.

Finucane, M. M., Sharpton, T. J., Laurent, T. J., and Pollard, K. S. (2014). A taxonomic signature of obesity in the microbiome? Getting to the guts of the matter. PloS One, 9(1):e84689.

Gonzalez, A., Navas-Molina, J. A., Kosciolek, T., McDonald, D., Vázquez-Baeza, Y., Ackermann, G., DeReus, J., Janssen, S., Swafford, A. D., Orchanian, S. B., et al. (2018). Qiita: rapid, web-enabled microbiome meta-analysis. Nature Methods, 15(10):796–798.

Gonzalez, C. G., Mills, R. H., Zhu, Q., Sauceda, C., Knight, R., Dulai, P. S., and Gonzalez, D. J. (2022). Location-specific signatures of Crohn's disease at a multi-omics scale. Microbiome, 10(1):133.

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

Han, S., Pool, J., Tran, J., and Dally, W. Learning both weights and connections for efficient neural network. In Advances in Neural Information Processing Systems, volume 28, (2015).

Hao, Y., Hao, S., Andersen-Nissen, E., III, W. M. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell.

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

Iuchi, H., Matsutani, T., Yamada, K., Iwano, N., Sumi, S., Hosoda, S., Zhao, S., Fukunaga, T., and Hamada, M. (2021). Representation learning applications in biological sequence analysis. Computational and Structural Biotechnology Journal, 19:3198–3208.

Jiang, R., Li, W. V., and Li, J. J. (2021). mbImpute: an accurate and robust imputation method for microbiome data. Genome Biology, 22(1):192.

Jiang, R., Sun, T., Song, D., and Li, J. J. (2022). Statistics or biology: the zero-inflation controversy about scRNA-seq data. Genome Biology, 23(1):31.

Kabeerdoss, J., Jayakanthan, P., Pugazhendhi, S., and Ramakrishna, B. (2015). Alterations of mucosal microbiota in the colon of patients with inflammatory bowel disease revealed by real time polymerase chain reaction amplification of 16s ribosomal ribonucleic acid. Indian Journal of Medical Research, 142(1):23–32.

Kharchenko, P. V. (2021). The triumphs and limitations of computational methods for scRNA-seq. Nature Methods, 18 (7):723–732.

Knights, D., Parfrey, L. W., Zaneveld, J., Lozupone, C., and Knight, R. (2011). Human-associated microbial signatures: examining their predictive value. Cell Host & Microbe, 10(4):292–296.

Kojima, A., Nakano, K., Wada, K., Takahashi, H., Katayama, K., Yoneda, M., Higurashi, T., Nomura, R., Hokamura, K., Muranaka, Y., et al. (2012). Infection of specific strains of streptococcus mutans, oral bacteria, confers a risk of ulcerative colitis. Scientific Reports, 2(1):332.

Laske, C., Müller, S., Preische, O., Ruschil, V., Munk, M., Honold, I., Peter, S., Schoppmeier, U., and Willmann, M. (2022). Signature of Alzheimer's disease in intestinal microbiome: Results from the AlzBiom study. Frontiers in Neuroscience, 16:792996.

Lee, A. A., Rao, K., Limsrivilai, J., III, M. G., Malamet, B., Briggs, E., Young, V. B., and Higgins, P. D. (2020). Temporal gut microbial changes predict recurrent clostridiodes difficile infection in patients with and without ulcerative colitis. Inflammatory Bowel Diseases, 26(11):1748–1758.

Li, B., Zhong, D., Jiang, X., and He, T. TopoPhy-CNN: 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).

Linderman, G. C., Zhao, J., Roulis, M., Bielecki, P., Flavell, R. A., Nadler, B., and Kluger, Y. (2022). Zero-preserving imputation of single-cell RNA-seq data. Nature Communications, 13(1):192.

Liu, B., Wei, Y., Zhang, Y., and Yang, Q. Deep neural networks for high dimension, low sample size data. In International Joint Conference on Artificial Intelligence, pages 2287–2293, (2017).

Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., Andrews, E., Ajami, N. J., Bonham, K. S., Brislawn, C. J., et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature, 569(7758):655–662.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. International Conference on Learning Representations.

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

Lu, Y. Y., Fan, Y., Lv, J., and Noble, W. S. DeepPINK: reproducible feature selection in deep neural networks. In Advances in Neural Information Processing Systems, (2018).

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

Ma, J., Yu, M. K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R., and Ideker, T. (2018). Using deep learning to model the hierarchical structure and function of a cell. Nature Methods, 15(4):290–298.

Medina, R. H., 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(1):98.

Mills, R. H., Dulai, P. S., Vázquez-Baeza, Y., Sauceda, C., Daniel, N., Gerner, R. R., Batachari, L. E., Malfavon, M., Zhu, Q., Weldon, K., et al. (2022). Multi-omics analyses of the ulcerative colitis gut microbiome link bacteroides vulgatus proteases with disease severity. Nature Microbiology, 7(2):262–276.

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., and Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature Biotechnology, 36(10):996–1004.

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., and Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Computational Biology, 12(7):e1004977.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems, pages 8024–8035. Curran Associates, Inc., Vancouver, Canada, (2019).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 464(7285): 59–65.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature, 490(7418):55–60.

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(10):2993–3001.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20:53–65.

Sasaki, K., Inoue, J., Sasaki, D., Hoshi, N., Shirai, T., Fukuda, I., Azuma, T., Kondo, A., and Osawa, R. (2019). 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.

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

Sharma, D., Paterson, A. D., and Xu, W. (2020). TaxoNN: ensemble of neural networks on stratified microbiome data for disease prediction. Bioinformatics, 36(17):4544–4550.

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

Shtossel, O., Isakov, H., Turjeman, S., Koren, O., and Louzoun, Y. (2023). Ordering taxa in image convolution networks improves microbiome-based machine learning accuracy. Gut Microbes, 15(1):2224474.

Storey, J. D. (2003). The positive false discovery rate: A bayesian interpretation and the q-value. The Annals of Statistics, 31(6):2013–2035.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papelexi, E., III, W. M. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. Cell, 77(7):1888–1902.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In International Conference on Machine Learning, (2017).

Tsang, M., Cheng, D., and Liu, Y. (2018). Detecting statistical interactions from neural network weights. International Conference on Learning Representations.

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. Nature, 444(7122):1027–1031.

Vogt, N. M., Kerby, R. L., Dill-McFarland, K. A., Harding, S. J., Merluzzi, A. P., Johnson, S. C., Carlsson, C. M., Asthana, S., Zetterberg, H., Blennow, K., et al. (2017). Gut microbiome alterations in Alzheimer's disease. Scientific Reports, 7 (1):13537.

Wang, M., Zhou, X., Li, F., Huckins, J., King, R. W., and Wong, S. T. C. (2008). Novel cell segmentation and online SVM for cell cycle phase identification in automated microscopy. Bioinformatics, 24(1):94–101.

Washburne, A. D., Morton, J. T., Sanders, J., McDonald, D., Zhu, Q., Oliverio, A. M., and Knight, R. (2018). Methods for phylogenetic analysis of microbiome data. Nature Microbiology, 3(6):652–661.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. Journal of Big data, 3(1):1–40.

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biology, 19:1–5.

Xiao, J., Chen, L., Johnson, S., Yu, Y., Zhang, X., and Chen, J. (2018). Predictive modeling of microbiome data using a phylogeny-regularized generalized linear mixed model. Frontiers in Microbiology, 9:1391.

Ye, S. H., Siddle, K. J., Park, D. J., and Sabeti, P. C. (2019). Benchmarking metagenomics tools for taxonomic classification. Cell, 178(4):779–794.

Zeng, Y., Li, J., Wei, C., Zhao, H., and Wang, T. (2022). mbDenoise: microbiome data denoising using zero-inflated probabilistic principal components analysis. Genome Biology, 23(1):94.

Zhu, Q., Mai, U., Pfeiffer, W., Janssen, S., Asnicar, F., Sanders, J. G., Belda-Ferre, P., Al-Ghalith, G. A., Kopylova, E., McDonald, D., et al. (2019). Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. Nature Communications, 10(1):5477.

Zhu, Q., Hou, Q., Huang, S., Ou, Q., Huo, D., Vázquez-Baeza, Y., Cen, C., Cantu, V., Estaki, M., Chang, H., et al. (2021). Compositional and genetic alterations in Graves' disease gut microbiome reveal specific diagnostic biomarkers. The ISME journal, 15(11):3399–3411.

Zhu, Q., Huang, S., Gonzalez, A., McGrath, I., McDonald, D., Haiminen, N., Armstrong, G., Vazquez-Baeza, Y., Yu, J., Kuczynski, J., et al. (2022). Phylogeny-aware analysis of metagenome community ecology based on matched reference genomes while bypassing taxonomy. mSystems, 7(2):e00167–22.

Supplementary Information

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

Dataset	Data source	Notes
AlzBiom	EBI-ENA ID: PRJEB47976	The sequencing data is clean (QC'ed, host-filtered).
ASD	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	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	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	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	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