Journal Title Here, 2023, pp. 1–11

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

PAPER

PhyloMix: Enhancing microbiome-trait association prediction through phylogeny-mixing augmentation

Yifan Jiang,[1,*] Disen Liao,[1,*] Qiyun Zhu[2] and Yang Young Lu[1]

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

Abstract

Motivation: Understanding the associations between traits and microbial composition is a fundamental objective in microbiome research. Recently, researchers have turned to machine learning (ML) models to achieve this goal with promising results. However, the effectiveness of advanced ML models is often limited by the unique characteristics of microbiome data, which are typically high-dimensional, compositional, and imbalanced. These characteristics can hinder the models' ability to fully explore the relationships among taxa in predictive analyses. To address this challenge, data augmentation has become crucial. It involves generating synthetic samples with artificial labels based on existing data and incorporating these samples into the training set to improve ML model performance.
Results: Here we propose PhyloMix, a novel data augmentation method specifically designed for microbiome data to enhance predictive analyses. PhyloMix leverages the phylogenetic relationships among microbiome taxa as an informative prior to guide the generation of synthetic microbial samples. Leveraging phylogeny, PhyloMix creates new samples by removing a subtree from one sample and combining it with the corresponding subtree from another sample. Notably, PhyloMix is designed to address the compositional nature of microbiome data, effectively handling both raw counts and relative abundances. This approach introduces sufficient diversity into the augmented samples, leading to improved predictive performance. We empirically evaluated PhyloMix on six real microbiome datasets across five commonly used ML models. PhyloMix significantly outperforms distinct baseline methods including sample-mixing-based data augmentation techniques like vanilla mixup and compositional cutmix, as well as the phylogeny-based method TADA. We also demonstrated the wide applicability of PhyloMix in both supervised learning and contrastive representation learning.
Availability: The Apache licensed source code is available at (`https://github.com/batmen-lab/phylomix`).

Key words: phylogeny, data augmentation, microbiome, mixup, representation learning

Introduction

The human microbiome represents the intricate communities of microorganisms residing in and on our bodies, with bacteria alone possessing 100 times more unique genes than the human genome [26]. These microbiomes play a crucial role in human physiology, significantly affecting both health and disease [30]. For example, the microbiome influences a wide range of diseases, including diabetes [27], obesity [32], inflammatory bowel disease [23], and even neurodegenerative conditions such as Parkinson's disease [10] and Alzheimer's disease [33]. Exploring the association between the microbiome and host traits will offer insights

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

into the mechanisms by which the microbiome influences human health and disease, paving the way for the development of innovative therapeutic strategies.

To investigate microbiome-trait associations, researchers have employed machine learning (ML) models to uncover complex patterns within microbial samples under various conditions and thereby facilitate microbiome-based disease detection [21]. Disease prediction from microbial samples is typically framed as a supervised learning task. In this approach, an ML model is trained on microbial samples–often characterized by the abundance of selected microbial taxa–to predict a host trait of interest, which may be binary (*e.g.,* disease status) or continuous (*e.g.,* age). Extensive studies have shown that ML models hold significant potential for accurately classifying clinically relevant traits across various diseases using microbial sample profiles [35, 22]. These ML models range from classic approaches like logistic regression and support vector machines to more advanced techniques like random forests and deep neural networks.

Despite increasing interest, the effectiveness of advanced ML models has been constrained by the unique characteristics of microbiome data, limiting their ability to fully explore the relationships among taxa in predictive analyses. Specifically, microbiome features, which represent the abundance of hundreds to thousands of taxa, are inherently high-dimensional. This high dimensionality presents analytical challenges, as the relatively small number of samples can result in overfitting during training, thus limiting the model's ability to generalize to unseen samples or even leading to conflicting research conclusions [17, 28]. Furthermore, microbiome data is inherently compositional because the profiling instrument imposes an arbitrary total on the profiled taxa abundances [11]. This can lead to issues when compositional data is analyzed using non-compositional methods. Lastly, the host traits of interest are often unevenly distributed, with training data either over-representing diseased states and lacking sufficient healthy samples, or vice versa, resulting in a distribution that does not accurately reflect the general population. Thus, new methods are required to adapt machine learning models to complex data challenges, including low sample sizes, high feature dimensionality, compositionality, and unbalanced trait distribution.

Ideally, ML models could be better trained for predictive analyses with more microbial samples, but in reality, collecting these samples is laborious and costly. From a practitioner's perspective, a more cost-effective alternative is to computationally augment the data, an approach that has been increasingly studied by the ML community in recent years [24]. The core concept of data augmentation is to generate synthetic samples with artificial labels based on existing data, and incorporate these into the training data to enhance ML model performance (Fig. 1**a**). For example, one of the pioneering methods, SMOTE [6], generates synthetic samples by oversampling minority classes using k-nearest intra-class neighbors. More recently, mixup [38] and its variants [5] create synthetic samples by linearly combining two samples and their corresponding labels, which are drawn randomly from the training data. However, most previous works are general-purpose and not tailored for microbiome data, overlooking domain-specific knowledge unique to the microbiome. To our knowledge, the only exception is TADA [29], but this algorithm is limited to mimicking sample-wise variations and fails to fully explore the potential of advanced data augmentation techniques (See detailed comparison in Sec. 4).

In this study, we introduce a novel data augmentation method specifically designed for microbiome data to enhance predictive analyses, called PhyloMix (Phylogeny-Mixing augmentation). The key ideas behind PhyloMix is to leverage the phylogenetic relationships among microbiome taxa, using them as an informative prior to guide the generation of synthetic microbial samples. Here, phylogeny serves as a tool to organize and understand microbial taxa by encoding their evolutionary relationships. Leveraging phylogeny, PhyloMix generates new samples by removing a subtree from one sample and recombining the counterpart from another sample, rather than interpolating between two whole microbial samples. This approach considers that variations among microbiome taxa are constrained by their evolutionary relationships as encoded by the phylogeny, thereby producing more realistic synthetic samples. Importantly, PhyloMix encodes phylogenetic relationships using a novel phylogenetic probabilistic profiling to address the compositional nature of microbiome data, allowing it to effectively handle both raw counts and relative abundances.

We have evaluated PhyloMix for supervised learning tasks using six publicly available microbiome datasets, each with varying sample sizes and feature dimensionality. We also benchmarked the evaluation using five commonly used ML models, ranging from classic logistic regression and support vector machines to more advanced methods like random forests and deep neural networks. Overall, the augmentation generated by PhyloMix preserves or notably enhances predictive performance, regardless of the dataset complexity or the sophistication of the ML models used. In addition to supervised learning, PhyloMix enables representation learning by using randomly augmented training samples to define a self-supervised optimization objective, pulling random augmentations of the same sample closer together while pushing those generated from different samples further apart [7]. We observe that PhyloMix outperforms previous representation learning approaches by identifying more distinguishable representations of samples with different host trait labels.

Approach

Problem setup

Consider a supervised learning task involving a training data of n microbial samples, represented as a data

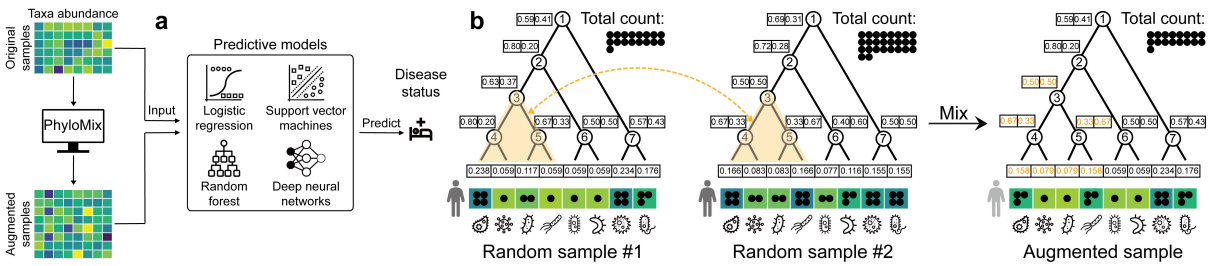

Fig. 1. Overview of PhyloMix. (**a**) PhyloMix generates synthetic samples with artificial labels based on existing data and incorporates them into the training data to enhance the performance of general ML models. (**b**) PhyloMix leverages phylogeny to generate synthetic samples by removing a subtree from one sample and recombining it with the counterpart from another sample. The involved subtree, along with the affected probabilistic phylogenetic profile, is highlighted in orange.

matrix. The rows of the data matrix correspond to the set of samples, denoted as $\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$ where $x_i \in \mathbb{R}^p$ represents the p-dimensional features of the i-th sample. The columns of the data matrix represent the set of features corresponding to the abundance of microbial taxa, denoted as $\mathcal{T} = \{t_1, t_2, \cdots, t_p\}$ where $t_j \in \mathbb{R}^n$ represents the abundance of the j-th taxon observed across all samples. The abundance of microbial taxa can be measured either by count or by normalized probability. The training data also includes trait labels, denoted as $\mathcal{Y} = \{y_1, y_2, \cdots, y_n\}$. Each sample x_i is paired with a corresponding trait label $y_i \in \mathbb{R}$, which is binary(*e.g.,* disease status) or continuous (*e.g.,* age).

Additionally, we have phylogenetic relationships among microbial taxa that encode their evolutionary dependencies. The phylogeny classifies taxa through a series of splits, representing estimated events where two lineages diverge from a common ancestor to form distinct species [34]. Two taxa with close phylogenetic relationships are likely to have similar functional roles, exhibiting more similar abundance variations compared to two randomly selected taxa. PhyloMix uses an inferred binary phylogenetic tree (See Sec. 3.2), denoted as $\mathcal{P}$, with the tree's leaf nodes labeled by the set of taxa $\mathcal{T}$. The tree's internal nodes are indexed from 1 to $p-1$ in a top-down fashion, denoted as $\mathcal{V} = \{v_1, v_2, \cdots, v_{p-1}\}$.

Given the samples $\mathcal{X}$, the corresponding trait labels $\mathcal{X}$, as well as the phylogenetic tree $\mathcal{P}$, our goal is to generate synthetic samples $\bar{\mathcal{X}}$ with corresponding synthetic trait labels $\bar{\mathcal{Y}}$, where $|\bar{\mathcal{X}}| = |\bar{\mathcal{Y}}| = \beta \cdot n$ and $\beta > 0$ is the augmentation ratio specified by the user. Empowered by the augmented samples $\mathcal{X} \cup \bar{\mathcal{X}}$ with labels $\mathcal{Y} \cup \bar{\mathcal{Y}}$, a more effective ML model $f : \mathbb{R}^p \mapsto \mathbb{R}$ should be learned (Fig. 1**a**), mapping a sample $x \in \mathbb{R}^p$ to the trait label $y \in \mathbb{R}$, which may no longer remains binary after the data augmentation.

Sample mixup

PhyloMix draws inspiration from mixup [38], a widely-used data augmentation technique that creates synthetic samples by linearly combining two randomly drawn samples and their corresponding labels. Specifically, given two random drawn sample-label pairs (x_i, y_i) and (x_j, y_j), where $x_i, x_j \in \mathcal{X}$ denotes the input samples and $y_i, y_j \in \mathcal{Y}$ are their corresponding trait labels, mixup a new synthetic sample $(\bar{x}, \bar{y})$ by:

$$\begin{aligned} \bar{x} &= \lambda \cdot x_i + (1-\lambda) \cdot x_j, \\ \bar{y} &= \lambda \cdot y_i + (1-\lambda) \cdot y_j \end{aligned} \tag{1}$$

where $\lambda \in [0, 1]$ is the mixing weight drawn from a beta distribution. It is worth noting that mixup is designed to be generic, treating the features corresponding to the abundance of microbial taxa equally. Considering that recent studies have leveraged the inherent correlation structure among taxa as an informative prior to improve disease prediction performance [36, 2, 31, 28, 16], we hypothesize that a data augmentation technique specifically tailored for microbiome data by incorporating this information can further enhance predictive accuracy.

Constructing sample-specific phylogenetic profiles

PhyloMix aims to enhance mixup-based data augmentation by incorporating phylogenetic relationships among microbiome taxa to guide the generation of synthetic samples. We are aware of the mixup variation that uses constituency parsing trees to generate synthetic sentences in the natural language processing community [39]. However, this method cannot be directly applied to microbiome data due to its unique characteristics. Specifically, each microbial sample is represented by a set of features corresponding to the abundance of microbial taxa, measured either by count or normalized probability. In the case of count data, two samples may be profiled under different sequencing depths. Thus, a higher count measured for taxon A in sample 1 compared to taxon B in sample 2 may result either from taxon A being more highly expressed than B or from sample 1 having a much higher overall sequencing depth than sample 2. In the case of normalized probability, where counts across taxa are scaled to sum to 1, substituting the probability of taxon A in sample 1 with that of sample 2 may violate the sum-to-one constraint in sample 1, rendering the resultant features of the synthetic sample invalid.

To address this challenge, we construct a probabilistic phylogenetic profile for each sample, ensuring that it remains valid to remove features corresponding to an

arbitrary subtree from one sample and recombine them with the counterpart from another sample (Fig. 1**b**). Specifically, the probabilistic phylogenetic profile of a sample x includes the following information: (1) the total abundance $S(x) = \sum_{j=1}^{p} t_j(x)$, where $t_j(x)$ indicates the abundance of the j-th taxon in sample x, and (2) the phylogenetic probability $(l_v(x), r_v(x))$ for each internal node in the phylogenetic tree, where $l_v(x)$ and $r_v(x)$ indicate the relative abundance of the descendants of its left and right children, respectively, with $l_v(x)+r_v(x) = 1$:

$$\begin{aligned} l_v(x) &= \frac{\sum_{j \in \mathrm{dleaf}(\mathrm{lchild}(v))} t_j(x)}{\sum_{j \in \mathrm{dleaf}(v)} t_j(x)}, \\ r_v(x) &= \frac{\sum_{j \in \mathrm{dleaf}(\mathrm{rchild}(v))} t_j(x)}{\sum_{j \in \mathrm{dleaf}(v)} t_j(x)} \end{aligned} \tag{2}$$

where $\mathrm{lchild}(v) \subset \mathcal{V} \cup \mathcal{T}$ and $\mathrm{rchild}(v) \subset \mathcal{V} \cup \mathcal{T}$ denote the left and right children of v, and $\mathrm{dleaf}(v) \subset \mathcal{T}$ denotes the descendent leaf nodes of v.

With the probabilistic phylogenetic profile of x, we can reconstruct the relative and absolute abundance for each microbiome taxon by traversing the phylogenetic tree in a top-down fashion. Specifically, for the j-th taxon, we trace the path from the root to the leaf node t_j and calculate its relative abundance by multiplying the phylogenetic probabilities of the internal nodes encountered along the path. For example, as illustrated in Fig. 1**b**, the relative abundance of the first taxon in sample 1 is calculated as $p_1(x) = 0.59 \times 0.80 \times 0.63 \times 0.80 \approx 0.238$. Similarly, the relative abundance of the last taxon in sample 1 is calculated as $p_8(x) = 0.41 \times 0.43 \approx 0.176$. Given the total abundance of x across all taxa, the absolute abundance for the j-th taxon is simply the allocation of this total abundance to that taxon. For example, the absolute abundance of the first and last taxa in sample 1 is calculated as $t_1(x) = p_1(x) \times S(x) = 0.238 \times 17 \approx 4$ and $t_2(x) = p_2(x) \times S(x) = 0.176 \times 17 \approx 3$, respectively.

Generating synthetic samples

With probabilistic phylogenetic profiles enabling valid exchange of arbitrary subtree between samples, we can now leverage phylogenetic information to guide the generation of synthetic microbial samples. Concretely, for a given sample-label pair (x_i, y_i), we randomly sample another sample-label pair (x_j, y_j) from the training data. First, we construct probabilistic phylogenetic profiles for both samples, allowing us to sample subtrees for exchange. Next, we draw a mixing weight $\lambda \in [0, 1]$ from a beta distribution $\mathrm{Beta}(\alpha, \alpha)$, where $\alpha \in (0, \infty)$ determines whether the synthetic sample combines evenly from the two original samples or is biased toward one of them. Intuitively, we want the portion of taxa exchanged to be reasonable. If the portion exchanged is too small or too large, the exchange may closely resemble the original samples and fail to introduce sufficient diversity into the augmented sample. Therefore, we use a $\mathrm{Beta}(2, 2)$ distribution in this study, ensuring that the mixing weights drawn are centered around 0.5 (See Sec. 4.3 for the robustness of different Beta distribution options).

To ensure that the taxa exchanged are constrained by the phylogenetic tree, we employ a two-phase sampling strategy. Specifically, We randomly draw an internal node $v \in \mathcal{V}$ from the phylogenetic tree and exchange the phylogenetic profiles between $(l_v(x_i), r_v(x_i))$ and $(l_v(x_j), r_v(x_j))$ for all descendent internal nodes of v, provided they have not already been exchanged. We repeat this draw-and-exchange process the number of exchanged internal nodes approximately reaches $(1 - \lambda) \cdot p$. Once we have the mixed probabilistic phylogenetic profiles, with $\lambda \cdot p$ internal nodes originally belonging to sample x_i and $(1 - \lambda) \cdot p$ to sample x_j, we can reconstruct the relative and absolute abundance for each microbiome taxon based on the total abundance $S(x_i)$, as described in Sec. 2.3. We denote the synthetic sample with the generated absolute abundance values as $\bar{x}_i$, and create its label as $\bar{y}_i = \lambda \cdot y_i + (1 - \lambda) \cdot y_j$. We repeat the process of generating synthetic samples until we have $\beta \cdot n$ new samples, where $\beta > 0$ is the augmentation ratio specified by the user. We use $\beta = 3$ in this study (See Sec. 4.3 for the robustness of different β options).

Materials and methods

Datasets

We conducted a comprehensive evaluation of PhyloMix using six publicly available microbiome datasets with varying sample sizes and feature dimensionality (Details in Tab. S.1). The AlzBiom dataset [18] investigated the connection between the gut microbiome and Alzheimer's disease (AD). It includes 75 amyloid-positive AD samples and 100 cognitively healthy control samples from the AlzBiom study, with profiles containing $p = 8,350$ taxa. The ASD dataset [8] examined the relationship between the gut microbiome and abnormal metabolic activity in individuals with Autism Spectrum Disorder (ASD). It consists of 30 typically developing (TD) and 30 constipated ASD (C-ASD) samples, with profiles containing $p = 7,287$ taxa. The GD dataset [40] investigated the link between the gut microbiome and Graves' disease (GD). It includes 100 GD samples and 62 healthy control samples, with profiles containing $p = 8,487$ taxa. The RUMC dataset [4] is part of a cohort study exploring the relationship between the gut microbiome and the Parkinson's disease (PD). The dataset includes 42 PD samples and 72 healthy control samples, with profiles containing $p = 7,256$ taxa. The IBD dataset [13] studied the relationship between the gut microbiome and two main subtypes of inflammatory bowel disease (IBD): Crohn's disease (CD) and ulcerative colitis (UC). It includes 108 CD and 66 UC samples, with profiles containing $p = 5,287$ taxa. The HMP2 dataset [19], part of the Integrative Human Microbiome Project (iHMP) [15], also explored the relationship between the gut microbiome and two IBD subtypes: CD and UC. Compared to the IBD dataset, this dataset increases the sample size to $n = 1,158$ (728 CD and 430 UC samples) and expands the taxa set to $p = 10,614$.

Phylogenetic tree

The abundance features in these datasets were profiled using Metagenomic Shotgun (MGS) sequencing, adhering to the standard operating procedures for shotgun metagenomics as implemented on the Qiita platform [12]. Specifically, the raw sequencing reads were classified using Woltka v0.1.4 [41] against the Web of Life (WoL) v2 database [42], which includes 15,953 microbial genomes. Note that a small number of internal nodes in the WoL phylogeny have more than two lineages (*i.e.,* polytomies). To address this, we resolved the issue by creating an arbitrary dichotomous structure among the affected nodes. Furthermore, because each dataset only profiles a subset of taxa in the WoL phylogeny, we used a pruned phytogenetic tree having the dataset-specific taxa as leaf nodes only.

Benchmark settings

We assessed the effectiveness of PhyloMix using five ML models with varying predictive capabilities: logistic regression (LR), support vector machine (SVM) with a linear kernel, random forest (RF), multi-layer perceptron (MLP), and MIOSTONE, a state-of-the-art deep learning model that encodes taxonomy [16]. We used the Scikit-learn implementation [25] with default settings for the LR, RF, and SVM models. The MLP model was configured with a pyramid-shaped architecture featuring two hidden layers of size 256 and 128, respectively. For the MIOSTONE model, we employed the default implementation settings. For all methods, we preprocessed microbiome features using centered log-ratio transformation [1] prior to data augmentation.

It is important to note that while the trait labels for the training samples are binary, they become continuous for the generated synthetic samples. For deep learning models like MLP and MIOSTONE, training is based on cross-entropy loss, so the continuous labels will not affect the learning process. Continuous labels provide more robust predictions, similar to the role of label smoothing [37]. For other models like LR, SVM, and RF, which are designed for discrete labels, we instead use their regression counterparts–linear regression, support vector regression, and random forest regression–to train on the augmented data with continuous labels. We also utilized the Scikit-learn implementation [25] with default settings for these regression models.

We compared PhyloMix's performance against three distinct baseline methods, each generating an equivalent amount of synthetic samples to train the ML models alongside the original training data. The vanilla mixup [38], a classic data augmentation method, creates synthetic samples by linearly combining two samples along with their corresponding labels. The compositional cutmix [14] extends the vanilla mixup to accommodate compositional data. TADA [29] generates synthetic samples by explicitly modeling biological and sampling variations, guided by the evolutionary relationships among taxa.

Results

PhyloMix enhances supervised learning

We evaluated data augmentation performance on six publicly available microbiome datasets with varying sample and microbial taxa sizes. These datasets cover a range of disease prediction tasks, including Autism Spectrum Disorder, Alzheimer's disease, Graves' disease, Parkinson's disease, and inflammatory bowel disease. Thus, they provide a comprehensive evaluation for predictive models of the human gut microbiome.

To assess the predictive performance of different methods, we employed two metrics: the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area Under the Precision-Recall Curve (AUPRC). Both metrics are widely used to evaluate predictive performance, with AUPRC being particularly well-suited for imbalanced datasets [9]. We evaluated performance using 5-fold cross-validation, training individual models on each training split and testing them on the corresponding test splits. The predictions from all test splits were concatenated, and the AUROC and AUPRC scores were calculated on the full dataset. For robustness, we repeated this process 20 times using different random seeds and reported the mean performance along with the 95% confidence intervals.

Our analysis shows that the augmentation generated by PhyloMix preserves or notably enhances predictive performance (Fig. 2**a** and Fig. S.1), regardless of the dataset complexity or the sophistication of the ML models used. On average across all ML models, PhyloMix achieves a 2.7%, 8.7%, 2.2%, and 3.9% improvement in test AUPRC on the RUMC, AlzBiom, IBD, and ASD datasets, respectively. Importantly, PhyloMix rarely hurt model performance, with drops observed in only 2 out of 20 cases, and the average decrease is typically just 0.3%. In comparison, vanilla mixup, the second-best data augmentation method, yields only a 1.8%, 6.0%, 1.3%, and 2.3% improvement in test AUPRC on the RUMC, AlzBiom, IBD, and ASD datasets, respectively. This highlights the effectiveness of leveraging phylogeny as an informative prior to enhance disease prediction performance. In the remaining GD and HMP2 datasets, PhyloMix achieves only a modest improvement of 0.3% and 0.1% in test AUPRC, as these datasets already exhibit near-perfect predictive performance, regardless of data augmentation. A similar conclusion can be drawn based on AUROC performance (Fig. S.2).

Finally, in addition to the quantitative comparison, we conducted a qualitative evaluation of PhyloMix and other baseline methods on the AlzBiom dataset (Fig. 2**b**). We directly visualize the improvement in predictive performance, measured by test AUPRC and AUROC, after applying each data augmentation method to generate synthetic training samples. We observed that TADA, which relies on modeling biological and sampling variations, fails to noticeably improve model performance, despite using the same phylogenetic information as PhyloMix. This underscores the importance of employing

Fig. 2. Data augmentation performance in the supervised learning setting. (**a**) The evaluation was performed on six publicly available microbiome datasets with varying microbial taxa sizes. PhyloMix is evaluated alongside five ML models with varying predictive capabilities and compared against three distinct baseline methods. The augmentation generated by PhyloMix preserves or notably enhances predictive performance, regardless of the dataset complexity or the sophistication of the ML models used. For scientific rigor, the performance comparison between PhyloMix and other baseline methods 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. (**b**) A qualitative evaluation of PhyloMix and other baseline methods on the AlzBiom dataset. Each dot corresponds to the performance with or without data augmentation for a given method under a random seed.

sample and label mixing for effective synthetic sample generation. Compositional cutmix and vanilla mixup can fail to enhance model performance when applied to LR and MIOSTONE models, highlighting their limitations in applicability. Importantly, PhyloMix consistently improves the performance of all ML models in both metrics across different random seeds, regardless of the model's complexity.

PhyloMix enhances representation learning

Having demonstrated improvements in supervised learning, we next evaluated PhyloMix's performance in representation learning. Representation learning, which extracts meaningful high-level features from raw data, is widely used to uncover hidden patterns in biological and biomedical contexts [3]. Specifically, a low-dimensional representation is learned through a deep autoencoder by minimizing reconstruction loss. This learned representation can then be used as input for a downstream classifier to perform predictive tasks. We use a deep autoencoder as the first baseline method, where the encoder consists of three hidden layers with sizes of 512, 256, and 128, respectively. The decoder mirrors this architecture in reverse.

Unlike the deep autoencoder, PhyloMix adopts a contrastive learning-based representation learning approach. Specifically, PhyloMix uses the same encoder architecture as the deep autoencoder, mapping representations to a space where the contrastive loss is applied [7]. The contrastive loss function is designed to bring the representations of positive pairs of samples closer together while pushing the representations of negative pairs further apart. Here, positive pairs are defined as two synthetic samples generated by applying random data augmentations to the same training sample pairs, whereas negative pairs are synthetic samples generated from different sample pairs. Similarly, we compared PhyloMix's performance against three alternative data augmentation methods evaluated in the supervised setting: the vanilla mixup, the compositional cutmix, and TADA. In addition, since the UniFrac distance is a widely used metric for comparing microbial samples [20], we included an additional contrastive learning-based baseline, where positive and negative sample pairs were selected based on their UniFrac distance, with positive pairs having a smaller UniFrac distance compared to negative pairs.

We assessed the quality of the learned representations using the following evaluation criteria. The learned representation from all methods is input into a logistic regression model to predict the trait label, using the same procedure as in the supervised learning setting. For robustness, we repeated this process 20 times using

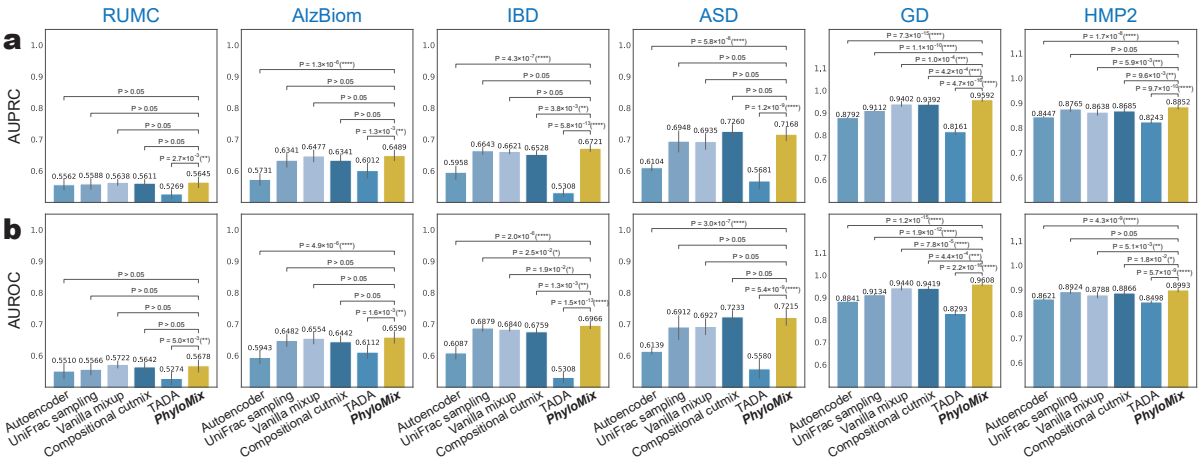

Fig. 3. Data augmentation performance in the representation learning setting. The evaluation was conducted on six publicly available microbiome datasets and compared against five distinct baseline methods. The representations obtained through PhyloMix consistently outperform those generated by other baseline methods in predicting disease status, measured by both (**a**) AUPRC and (**b**) AUROC. The performance comparison between PhyloMix and other baseline methods is quantified using one-tailed two-sample t-tests to calculate p-values.

different random seeds and reported the mean performance along with the 95% confidence intervals. Each repetition involved data splitting, representation learning, and ML model training.

Our analysis reveals that PhyloMix's contrastive model consistently outperforms both the deep autoencoder and the contrastive models that use alternative data augmentation methods in terms of learned representation quality (Fig. 3). On average, PhyloMix achieves a 9.8% and 9.8% improvement in test AUPRC and AUROC compared to using the representation learned by deep autoencoder, respectively. Importantly, PhyloMix enhances the quality of learned representations across all datasets, with the smallest improvement observed being 1.5% in test AUPRC and 3.0% in AUROC on the RUMC dataset. In comparison, compositional cutmix, the second-best data augmentation method in representation learning, shows only an 8.3% improvement in test AUPRC and an 8.2% improvement in AUROC over the deep autoencoder. Although UniFrac distance is widely used for comparing microbial samples based on phylogeny, PhyloMix proves to be more effective in leveraging phylogeny as an informative prior to enhance the quality of representation learning.

Dissecting the performance of PhyloMix

We first evaluated the computational cost of using PhyloMix in conjunction with five ML models with different levels of complexities. For a fair comparison, all models were trained and tested in the same environment: AMD EPYC 7302 16-Core Processor, NVIDIA RTX A6000 GPU, with 32GB DDR4 RAM. PhyloMix's data augmentation involves two primary components: generating synthetic samples and the additional training time necessary to integrate these synthetic samples. We found that PhyloMix's sample generation incurs only moderate computational overhead, measured in seconds (Fig. 4**a** and Fig. S.3). The additional training time required to incorporate these synthetic samples is marginal on the AlzBiom dataset, except for the HMP2 dataset, which has a larger sample size of $n = 1,158$ and higher dimensionality with $p = 10,614$ taxa. Note that when the sample size is large, data augmentation becomes less critical, as the dataset already exhibits near-perfect predictive performance, with or without augmentation. However, when the sample size is small, the computational time is negligible compared to the improvement in predictive accuracy gained through augmentation.

PhyloMix's data augmentation relies on two factors: the choice of the sample-mixing distribution and the augmentation level. To evaluate the impact of each component on predictive performance, we conducted a control study using the AlzBiom dataset by modifying PhyloMix, altering its components with different variations Specifically, we considered two variants of PhyloMix: (1) Altering the Beta distributions that control sample mixing weights; (2) Adjusting the number of augmentation samples added.

The results indicate that PhyloMix is robust to the choice of these two factors. In the first study (Fig. 4**b**), we compared three Beta distribution options, each representing a distinct sample mixing strategy. Specifically, $\text{Beta}(0.5, 0.5)$ tends to bias the synthetic sample toward one of the original samples, $\text{Beta}(2.0, 2.0)$ encourages a more even combination from both original samples, while $\text{Beta}(1.0, 1.0)$ results in a more arbitrary combination. In general, we observed that the Beta distribution choice does not negatively impact data augmentation performance. The only exception is the logistic regression model, where data augmentation transforms logistic regression into linear regression,

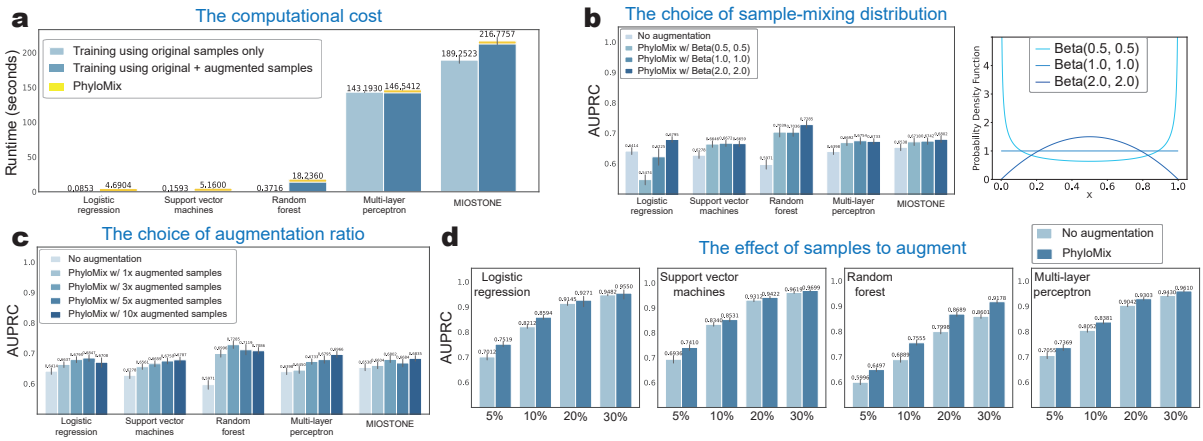

Fig. 4. Dissecting the performance of PhyloMix through control studies. (**a**) The data augmentation by PhyloMix incurs only moderate computational overhead, consisting of two main components: the generation of synthetic samples and the additional training time required to incorporate these synthetic samples. (**b**) The choice of the Beta distribution is crucial in determining the sample mixing weights. In general, this choice does not negatively impact data augmentation performance. The only exception is the logistic regression model, where data augmentation transforms logistic regression into linear regression, leading to a performance downgrade if the Beta distribution is not chosen properly. (**c**) PhyloMix is robust to the number of augmented samples. In general, data augmentation consistently improves predictive performance. However, increasing the number of augmented samples inevitably leads to higher computational overhead. (**d**) PhyloMix demonstrates more significant improvements when the training data size is small.

leading to a performance downgrade if the Beta distribution is not chosen properly so that the synthetic samples may closely resemble the original ones and fail to introduce sufficient diversity. In the second study (Fig. 4**c**), we compared four different augmentation levels, ranging from adding the same number of synthetic samples as the original samples to adding up to ten times as many synthetic samples. In general, we observed that data augmentation consistently improves predictive performance. To achieve a significant performance boost, we expect the number of synthetic samples added to be at least three times the number of original samples. It's important not to include excessive synthetic samples. On one hand, performance may diminish or even drop if too many synthetic samples are added; on the other hand, increasing the number of augmented samples leads to higher computational overhead.

Finally, we investigate the conditions under which PhyloMix's data augmentation provides the greatest improvement in predictive performance. To address this question, we randomly selected subsets of training samples from the HMP2 dataset, with subset sizes ranging from 5%, 10%, 20%, and 30% of the total training samples, to train the ML models. For robustness, we repeated this process 20 times with different random seeds. For each taxa subset, we trained and evaluated PhyloMix using the same settings as for the full dataset. We found that while training PhyloMix with more samples generally leads to better predictive performance, the improvement provided by PhyloMix's data augmentation is notably more significant across all ML models (Fig. 4**d**). It is important to note that data augmentation becomes more critical for performance improvement when the sample size is small, highlighting the necessity of PhyloMix.

Discussion and conclusion

In this study, we propose PhyloMix, a novel data augmentation method specifically designed for microbiome data to enhance predictive analyses. At its core, PhyloMix leverages the phylogenetic relationships among microbiome taxa as an informative prior to guide the generation of synthetic microbial samples. Leveraging phylogeny, PhyloMix generates new samples by removing a subtree from one sample and recombining the counterpart from another sample, rather than interpolating between two whole microbial samples. The key novelties of PhyloMix are threefold: (1) We enable valid phylogeny exchange through the use of sample-specific probabilistic phylogenetic profiles, effectively handling both raw counts and relative abundances. (2) We demonstrate that phylogeny-based sample mixing achieves superior performance compared to relying solely on phylogeny or using sample mixing alone. (3) We showcase the wide applicability of PhyloMix in both supervised learning and contrastive representation learning.

We evaluated PhyloMix's performance on six real microbiome datasets, each with varying sample sizes and feature dimensionality. We benchmarked the evaluation using five commonly used ML models, ranging from classic logistic regression and support vector machines to more advanced methods like random forests and deep neural networks. Overall, the augmentation generated by PhyloMix preserves or notably enhances predictive performance, regardless of the dataset complexity or

the sophistication of the ML models used. Furthermore, PhyloMix significantly outperforms distinct baseline methods in supervised learning and contrastive representation learning settings, including sample-mixing-based data augmentation techniques like vanilla mixup and compositional cutmix, as well as the phylogeny-based method TADA. Importantly, our comprehensive control studies demonstrate that PhyloMix is robust with respect to the choice of sample-mixing distribution and the level of augmentation.

Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions. The research is supported by the Canadian NSERC Discovery Grant RGPIN-03270-2023.

Competing interests

No competing interest is declared.

Author contributions statement

Q.Z. provided the datasets, Y.J. and D.L implemented the code and did the analysis, D.L conducted the experiments, Y.L. and Y.J. prepared the figures. Y.L. wrote the main manuscript. All authors participated the discussion. All authors reviewed the manuscript.

References

1. J. Aitchison. The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological), 44(2):139–160, 1982.
2. D. Albanese, C. De Filippo, D. Cavalieri, and C. Donati. Explaining diversity in metagenomic datasets by phylogenetic-based feature weighting. PLoS Computational Biology, 11(3):e1004186, 2015.
3. Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8):1798–1828, 2013.
4. J. Boktor, G. Sharon, L. Verhagen Metman, D. A. Hall, P. A. Engen, Z. Zreloff, D. J. Hakim, J. W. Bostick, J. Ousey, D. Lange, et al. Integrated multi-cohort analysis of the Parkinson's disease gut metagenome. Movement Disorders, 38(3):399–409, 2023.
5. C. Cao, F. Zhou, Y. Dai, and J. Wang. A survey of mix-based data augmentation: Taxonomy, methods, applications, and explainability. arXiv preprint arXiv:2212.10888, 2022.
6. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16:321–357, 2002.
7. T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. pages 1597–1607, 2020.
8. Z. Dan, X. Mao, Q. Liu, M. Guo, Y. Zhuang, Z. Liu, K. Chen, J. Chen, R. Xu, J. Tang, et al. Altered gut microbial profile is associated with abnormal metabolism activity of Autism Spectrum Disorder. Gut Microbes, 11(5):1246–1267, 2020.
9. J. Davis and M. Goadrich. The relationship between precision-recall and ROC curves. In International Conference on Machine Learning, pages 233–240, 2006.
10. S. K. Dutta, S. Verma, V. Jain, B. K. Surapaneni, R. Vinayek, L. Phillips, and P. Nair. Parkinson's disease: the emerging role of gut dysbiosis, antibiotics, probiotics, and fecal microbiota transplantation. Journal of Neurogastroenterology and Motility, 25(3):363, 2019.
11. G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. Microbiome datasets are compositional: and this is not optional. Frontiers in Microbiology, 8:2224, 2017.
12. A. Gonzalez, J. A. Navas-Molina, T. Kosciolek, D. McDonald, Y. Vázquez-Baeza, G. Ackermann, J. DeReus, S. Janssen, A. D. Swafford, S. B. Orchanian, et al. Qiita: rapid, web-enabled microbiome meta-analysis. Nature Methods, 15(10):796–798, 2018.
13. C. G. Gonzalez, R. H. Mills, Q. Zhu, C. Sauceda, R. Knight, P. S. Dulai, and D. J. Gonzalez. Location-specific signatures of Crohn's disease at a multi-omics scale. Microbiome, 10(1):133, 2022.
14. E. Gordon-Rodriguez, T. Quinn, and J. P. Cunningham. Data augmentation for compositional data: Advancing predictive models of the microbiome. Advances in Neural Information Processing Systems, 35:20551–20565, 2022.
15. The Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell Host & Microbe, 16(3):276–289, 2014.
16. Y. Jiang, M. Atton, Q. Zhu, and Y. Y. Lu. MIOSTONE: Modeling microbiome-trait associations with taxonomy-adaptive neural networks. bioRxiv, pages 2023–11, 2023.
17. D. Knights, L. W. Parfrey, J. Zaneveld, C. Lozupone, and R. Knight. Human-associated microbial signatures: examining their predictive value. Cell Host & Microbe, 10(4):292–296, 2011.
18. C. Laske, S. Müller, O. Preische, V. Ruschil, M. Munk, I. Honold, S. Peter, U. Schoppmeier, and M. Willmann. Signature of Alzheimer's disease in intestinal microbiome: Results from the AlzBiom study. Frontiers in Neuroscience, 16:792996, 2022.
19. J. Lloyd-Price, C. Arze, A. N. Ananthakrishnan, M. Schirmer, J. Avila-Pacheco, T. W. Poon, E. Andrews, N. J. Ajami, K. S. Bonham, C. J. Brislawn, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature, 569(7758):655–662, 2019.

20. C. Lozupone and R. Knight. UniFrac: a new phylogenetic method for comparing microbial communities. Applied and Environmental Microbiology, 71(12):8228–8235, 2005.
21. L. E. McCoubrey, M. Elbadawi, M. Orlu, S. Gaisford, and A. W. Basit. Harnessing machine learning for development of microbiome therapeutics. Gut Microbes, 13(1):1872323, 2021.
22. R. Hernandez Medina, S. Kutuzova, K. N. Nielsen, J. Johansen, L. H. Hansen, M. Nielsen, and S. Rasmussen. Machine learning and deep learning applications in microbiome research. ISME Communications, 2(1):98, 2022.
23. R. H. Mills, P. S. Dulai, Y. Vázquez-Baeza, C. Sauceda, N. Daniel, R. R. Gerner, L. E. Batachari, M. Malfavon, Q. Zhu, K. Weldon, et al. Multi-omics analyses of the ulcerative colitis gut microbiome link bacteroides vulgatus proteases with disease severity. Nature Microbiology, 7(2):262–276, 2022.
24. A. Mumuni and F. Mumuni. Data augmentation: A comprehensive survey of modern approaches. Array, 16:100258, 2022.
25. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.
26. J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 464(7285):59–65, 2010.
27. J. Qin, Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature, 490(7418):55–60, 2012.
28. D. Reiman, A. A. Metwally, J. Sun, and Y. Dai. PopPhy-CNN: a phylogenetic tree embedded architecture for convolutional neural networks to predict host phenotype from metagenomic data. IEEE Journal of Biomedical and Health Informatics, 24(10):2993–3001, 2020.
29. E. Sayyari, B. Kawas, and S. Mirarab. TADA: phylogenetic augmentation of microbiome samples enhances phenotype classification. Bioinformatics, 35(14):i31–i40, 2019.
30. I. Sekirov, S. L. Russell, C. M. Antunes, and B. B. Finlay. Gut microbiota in health and disease. Physiological Reviews, 2010.
31. D. Sharma, A. D. Paterson, and W. Xu. TaxoNN: ensemble of neural networks on stratified microbiome data for disease prediction. Bioinformatics, 36(17):4544–4550, 2020.
32. P. J. Turnbaugh, R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature, 444(7122):1027–1031, 2006.
33. N. M. Vogt, R. L. Kerby, K. A. Dill-McFarland, S. J. Harding, A. P. Merluzzi, S. C. Johnson, C. M. Carlsson, S. Asthana, H. Zetterberg, K. Blennow, et al. Gut microbiome alterations in Alzheimer's disease. Scientific Reports, 7(1):13537, 2017.
34. A. D. Washburne, J. T. Morton, J. Sanders, D. McDonald, Q. Zhu, A. M. Oliverio, and R. Knight. Methods for phylogenetic analysis of microbiome data. Nature Microbiology, 3(6):652–661, 2018.
35. S. Wu, Y. Chen, Z. Li, J. Li, F. Zhao, and X. Su. Towards multi-label classification: Next step of machine learning for microbiome research. Computational and Structural Biotechnology Journal, 19:2742–2749, 2021.
36. J. Xiao, L. Chen, S. Johnson, Y. Yu, X. Zhang, and J. Chen. Predictive modeling of microbiome data using a phylogeny-regularized generalized linear mixed model. Frontiers in Microbiology, 9:1391, 2018.
37. C.-B. Zhang, P.-T. Jiang, Q. Hou, Y. Wei, Q. Han, Z. Li, and M.-M. Cheng. Delving deep into label smoothing. IEEE Transactions on Image Processing, 30:5984–5996, 2021.
38. H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. International Conference on Learning Representations, 2018.
39. L. Zhang, Z. Yang, and D. Yang. TreeMix: Compositional constituency-based data augmentation for natural language understanding. Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2022.
40. Q. Zhu, Q. Hou, S. Huang, Q. Ou, D. Huo, Y. Vázquez-Baeza, C. Cen, V. Cantu, M. Estaki, H. Chang, et al. Compositional and genetic alterations in Graves' disease gut microbiome reveal specific diagnostic biomarkers. The ISME journal, 15(11):3399–3411, 2021.
41. Q. Zhu, S. Huang, A. Gonzalez, I. McGrath, D. McDonald, N. Haiminen, G. Armstrong, Y. Vazquez-Baeza, J. Yu, J. Kuczynski, et al. Phylogeny-aware analysis of metagenome community ecology based on matched reference genomes while bypassing taxonomy. mSystems, 7(2):e00167–22, 2022.
42. Q. Zhu, U. Mai, W. Pfeiffer, S. Janssen, F. Asnicar, J. G. Sanders, P. Belda-Ferre, G. A. Al-Ghalith, E. Kopylova, D. McDonald, et al. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. Nature Communications, 10(1):5477, 2019.

Appendix

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

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).
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.
RUMC	Qiita ID: 12975	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`

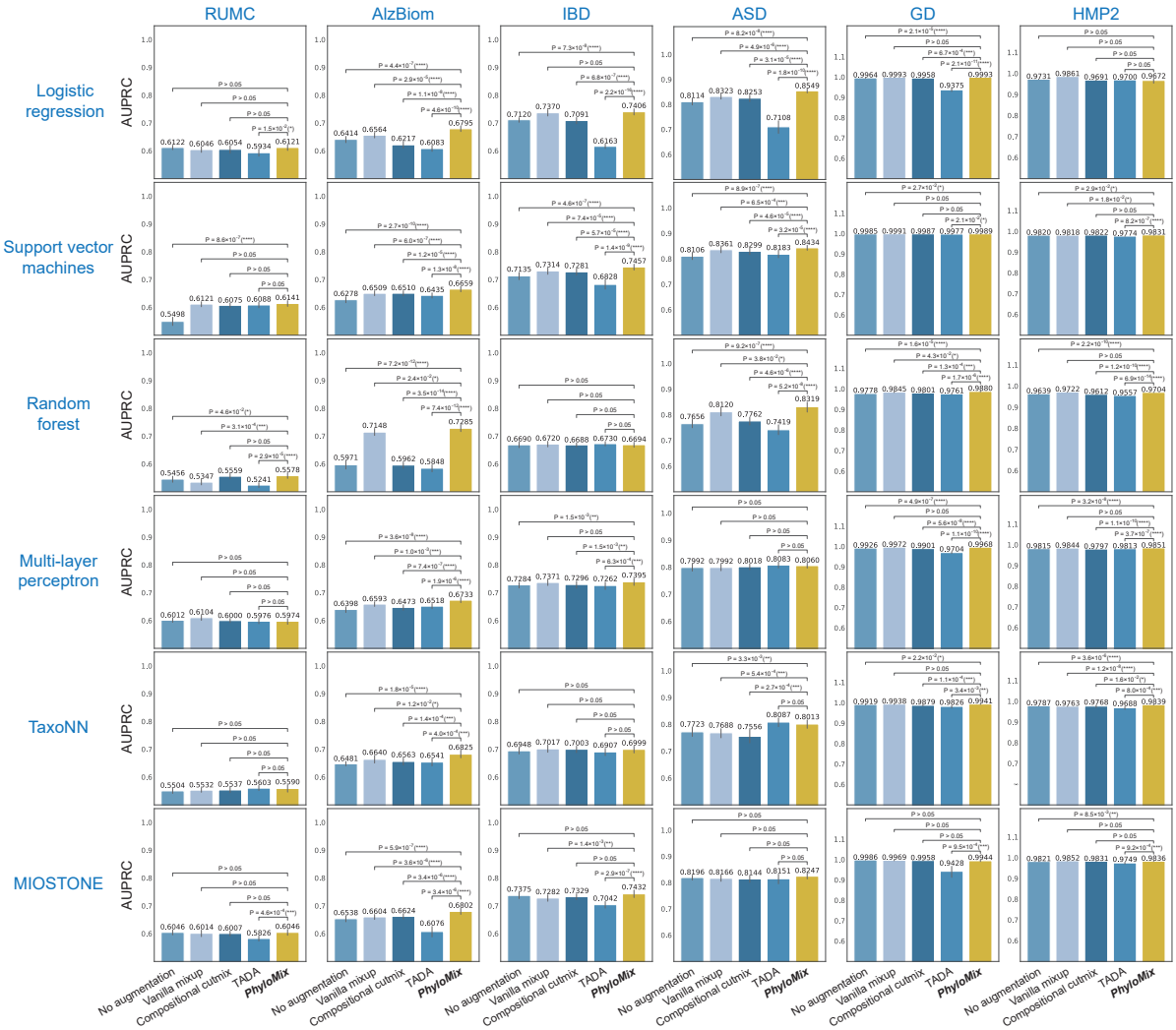

Fig. S.1. Data augmentation performance in the supervised learning setting measured by AUPRC. (**a**) The evaluation was performed on six publicly available microbiome datasets with varying microbial taxa sizes. PhyloMix is evaluated alongside five ML models with varying predictive capabilities and compared against three distinct baseline methods.

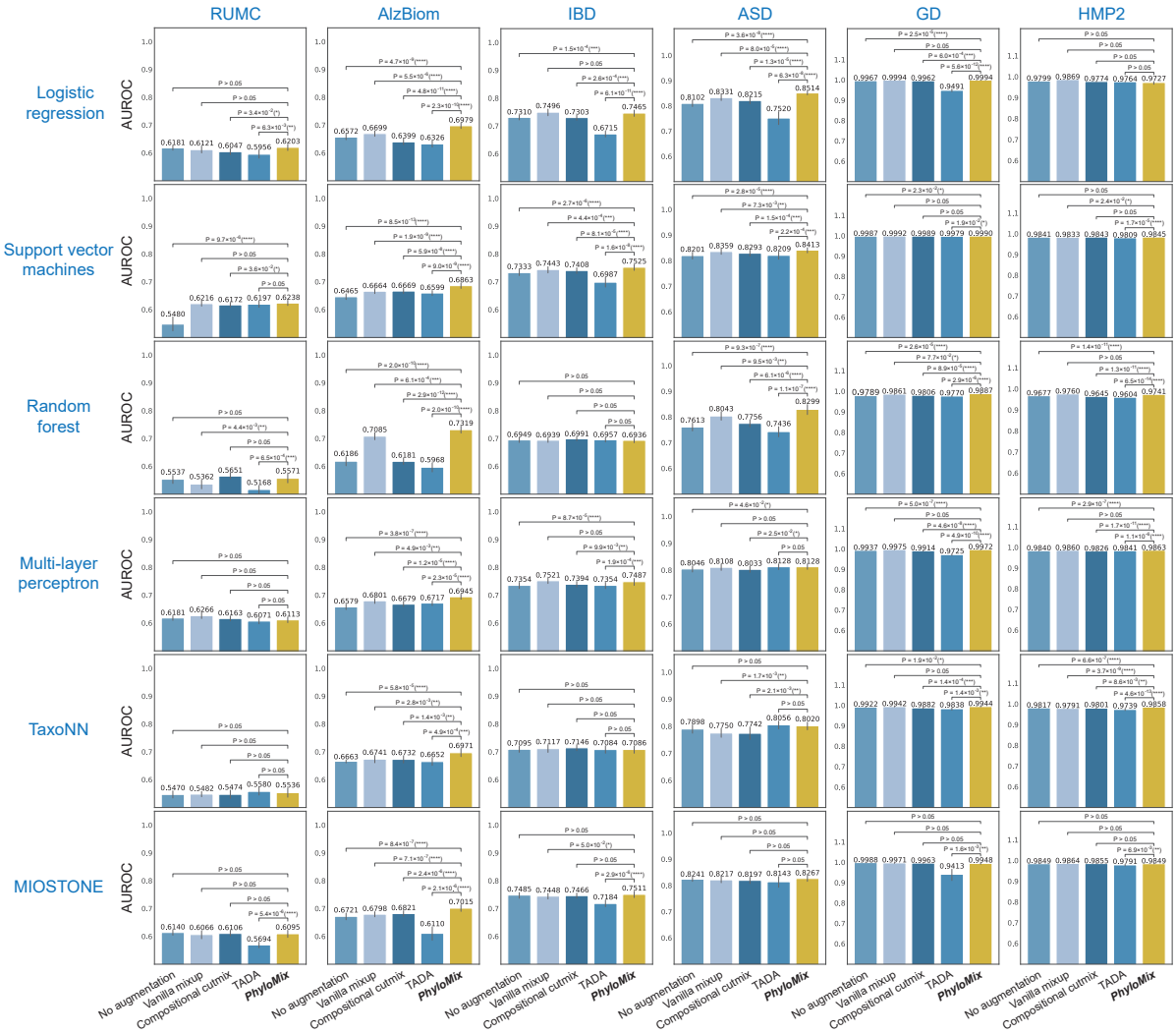

Fig. S.2. Data augmentation performance in the supervised learning setting measured by AUROC. (**a**) The evaluation was performed on six publicly available microbiome datasets with varying microbial taxa sizes. PhyloMix is evaluated alongside five ML models with varying predictive capabilities and compared against three distinct baseline methods.

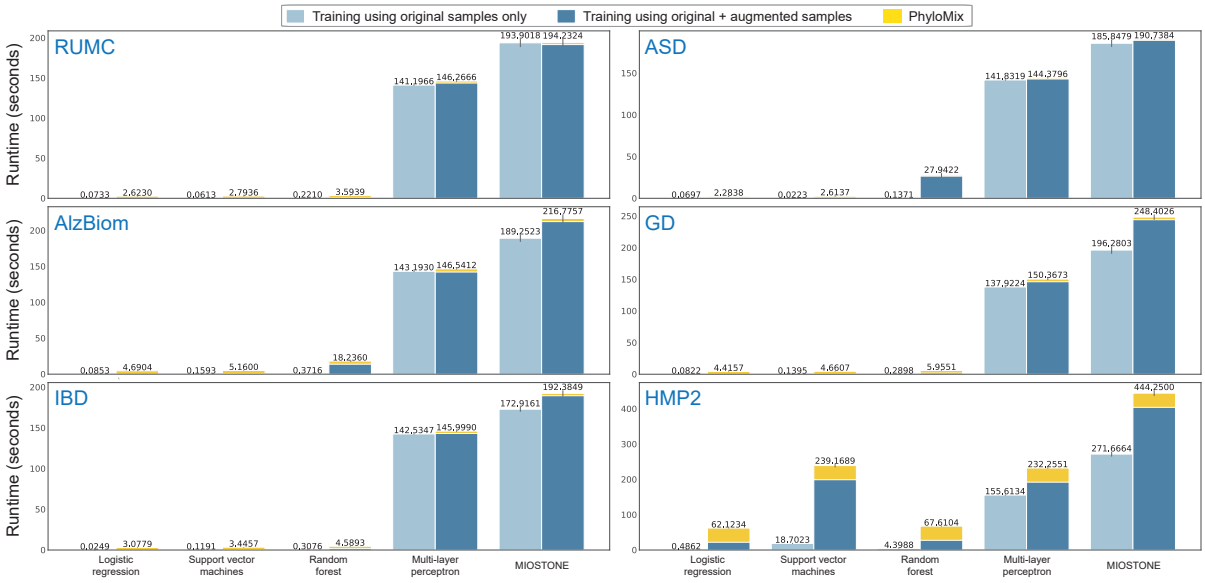

Fig. S.3. The computational cost of PhyloMix.

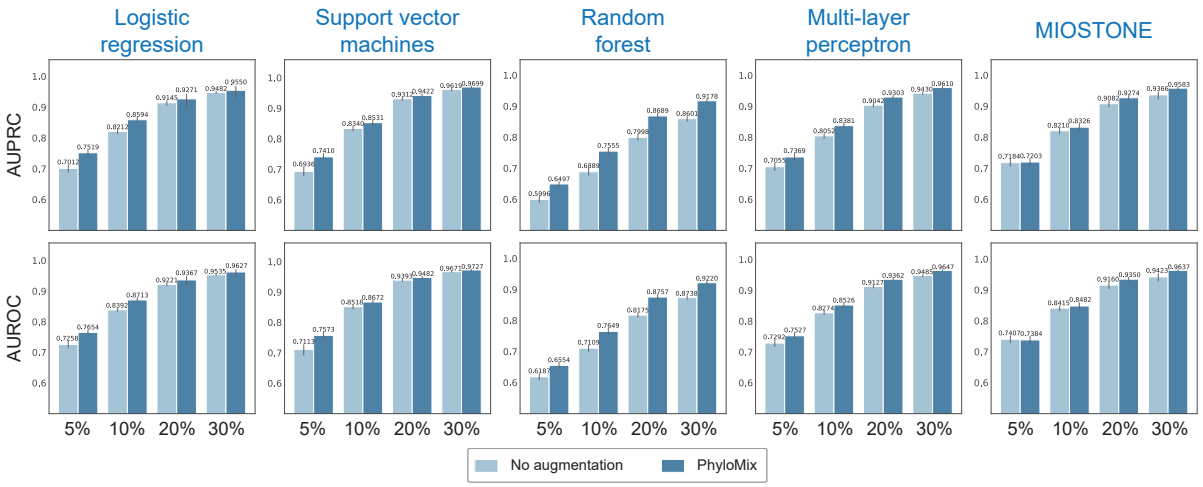

Fig. S.4. PhyloMix exhibits more significant improvements when the training data size is small.