DeepBiome: a phylogenetic tree informed deep neural network for microbiome data analysis

Jing Zhai

University of Arizona

Youngwon Choi

Seoul national university

Yin Chen

University of Arizona

Kenneth Knox

University of Arizona

Homer Twigg III

Indiana University

Won Joong-Ho

Seoul National university

Hua Zhou

University of California Los Angeles

Jin Zhou (✉ jzhou@email.arizona.edu)

University of Arizona <https://orcid.org/0000-0001-7983-0274>

Software article

Keywords: Metagenomics, Phylogenetic Tree, Neural Networks, Prediction, Mixed Taxonomic Levels

Posted Date: June 10th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-33860/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

1 DeepBiome: a phylogenetic tree informed deep neural network for
2 microbiome data analysis

3 Jing Zhai^{1†}, Youngwon Choi^{2†}, Yin Chen³, Kenneth Knox⁴, Homer L. Twigg III⁵,
4 Joong-Ho Won², Hua Zhou⁶ and Jin J. Zhou^{1*}

June 6, 2020

5 ¹Department of Epidemiology and Biostatistics, College of Public Health, University of Arizona,
6 Tucson, AZ 85724

7 ²Department of Statistics, Seoul National University, Seoul 08826, Korea

8 ³Department of Pharmacology and Toxicology, College of Pharmacology, University of Arizona,
9 Tucson, AZ 85724

10 ⁴Division of Pulmonary, Allergy, Critical Care, Sleep Medicine, Department of Medicine,
11 University of Arizona, Tucson, AZ 85724

12 ⁵Division of Pulmonary, Critical Care, Sleep, and Occupational Medicine, Indiana University
13 Medical Center, Indianapolis, IN 46202

14 ⁶Department of Biostatistics, University of California, Los Angeles, CA 90095

15 * Correspondence: Jin J. Zhou (jzhou@email.arizona.edu), Department of Epidemiology and Biostatistics,
16 College of Public Health, University of Arizona, 1295 N Martin Ave, Tucson, AZ 85724

17 [†] Contribute equally.

Abstract

Background: Evidence linking microbiome and human health is rapidly growing. Microbiome profile can be a novel predictive biomarker for many diseases. However, bacteria count tables are typically sparse and bacteria are classified at a hierarchy of taxonomic levels, ranging from species to phylum. Existing analysis tools focus on identifying microbiome associations either at the community level or at a specific pre-defined taxonomic level. They fail to incorporate the evolutionary relationship between bacteria and cannot learn from the data to aggregate microbiome contribution, thus leading to less accurate and less interpretable results in prediction, classification or selection.

Results: We present **DeepBiome**, a phylogeny-informed neural network architecture for predicting phenotypes from microbiome counts and uncovering the microbiome-phenotype association network. It takes microbiome abundance as the input and let the phylogenetic taxonomy guide the neural network architecture. Commonly used neural network architectures are targeted towards image and text analysis and typically require huge amount of training data, which is scarce in biomedical applications. By leveraging the phylogenetic information, **DeepBiome** relieves the heavy burden of tuning for the optimal deep learning architecture, avoids overfitting, and more importantly enables visualizing the path from microbiome counts to disease. It is applicable to both regression and classification problems. The simulation study and real-life data analysis demonstrate that **DeepBiome** is highly accurate and efficient and provides a deep understanding of complex microbiome-phenotype associations even using small to moderate training sample sizes.

Conclusions: In practice, it is unknown at which taxonomic level that microbiome clusters tag the association. Therefore, the central advantage of the presented method over other analytical methods is that it offers an ecological and evolutionary understanding of host-microbe interactions which is important for microbiome-based medicine. **DeepBiome** is implemented using Python packages Keras and the Tensorflow. It is an open-source tool available at (<https://github.com/Young-won/DeepBiome>).

Contact: jzhou@email.arizona.edu

Keywords: Metagenomics; Phylogenetic Tree; Neural Networks; Prediction; Mixed Taxonomic Levels

1 Introduction

Emerging high-throughput sequencing technologies have vastly improved our understanding of the role of human microbiome in many diseases (Sartor, 2008; Gilbert et al., 2016; Ni et al., 2017; Franzosa et al., 2019). Despite these achievements, knowledge about the mechanism of microbes’ involvement in the disease aggravation is still limited. One highlighted obstacle is to infer and visualize the bacteria-to-disease path across taxonomic levels.

In the 16S ribosomal ribonucleic acid (RNA) sequencing, data are grouped into Operational Taxonomic Units (OTUs) according to a similarity threshold, e.g., 97%. These OTUs are then clustered at different phylogenetic depths to build a phylogenetic tree portraying their evolutionary relationships (Schloss and Handelsman, 2005; Turnbaugh et al., 2007). The phylogenetic tree contains critical knowledge about how OTUs are related to each other and was leveraged in statistical models to boost the power of identifying the microbiome associated with host phenotypes (Xiao et al., 2017; Zhai et al., 2018, 2019). In fact, closely related microbial taxa may form clusters with similar biological functions. An inherent challenge is to determine at which taxonomic level the association signals are clustered. At a shallow phylogenetic depth (e.g., genus), at a deeper depth (e.g., phylum), or even at a mixture taxonomic ranks?

Multiple approaches have been proposed to incorporate the phylogenetic structure into analysis. For example, Chen et al. (2012, 2015) imposed a Laplacian penalty constructed using the sum of branch lengths linking any two OTUs on the evolutionary tree. They are not designed to detect signals at different evolutionary depths. Garcia et al. (2013) proposed a sparse regression model using ℓ_1 and ℓ_2 regularizations to achieve sparsity at multiple taxonomic levels. However, this method used three tuning parameters and needed an exhaustive grid search. Moreover, this approach can only select taxa at up to three levels therefore cannot cover the entire range of phylogenetic depths. Wang and Zhao (2017) used a tree-guided variable fusion method that is capable of building predictive models using bacteria at different taxonomic levels. Their method was built upon the assumption that closely-related bacteria may have similar biological function. However Zhou et al. (2019) provided an example that violated this assumption. In their study, *Corynebacterium* and *Rothia* were shown to have opposite effects on the change of lung functions in an Human Immunodeficiency Virus (HIV) positive population, even though they belong to the same order of *Actinomycetales*. Xiao et al. (2018) have developed a generalized linear mixed model that used the evolutionary rate as a tuning parameter. This approach can identify clustered signals without a prior knowledge of phylogenetic depth. However, it is not desirable when microbiome effects are clustered at a mixture of taxonomic levels.

Deep neural network (DNN), an area of growing interest in biomedical research, is a promising approach for analyzing complex microbiome data. Lu et al. (2018) applied a DNN model to

gut microbiome and identified important bacteria that were associated with body mass index (BMI). Reiman et al. (2017) embedded the phylogenetic into a convolutional neural network (CNN) architecture to predict outcome of interest. In that study, the embedding algorithm translated taxa abundances at every phylogenetic level to an abundance matrix. This abundance matrix captured the spatial information of taxa in the phylogenetic tree. However, after embedding, only a fixed number of layers (three) were used and the taxa lineage information is lost. The identified neurons do not have any biological meaning and the result lacks interpretability.

We present **DeepBiome**, a DNN-based predictive model, for capturing microbiome signals at different phylogenetic depths. This model is applicable to both regression and classification problems. It takes microbiome taxonomic abundance data as input. By regularizing the neural network architecture towards the phylogenetic structure, **DeepBiome** greatly reduces the number of parameters and tuning burden compared to the conventional neural networks. It is able to identify important taxa that are associated with outcomes at all taxonomic levels. Phylogeny regularization is achieved by weight decay, a popular technique (Mundie and Massengill, 1991; Krogh and Hertz, 1992; Gupta and Lam, 1998) to prevent overfitting and boost performances of DNNs (Zhang et al., 2018). All existing weight decay schemes assume a global rate of decay. **DeepBiome** instead incorporates the bacteria evolutionary relationship into a differential weight decay regularization matrix, thus generating an interpretable effect transfer network in modeling and analyzing microbiome data. Simulation studies and analyses of the datasets from the American Gut Project (AGP) and a lung microbiome study demonstrate the superior performance of **DeepBiome** over commonly used tools such as support vector machine (SVM), regression with ℓ_1 (lasso) or $\ell_1 + \ell_2$ (elastic net) penalties, DNN without tree regularization, and DNN with ℓ_1 penalty.

2 Methods and materials

2.1 Microbiome data structure

Suppose we have p OTUs from a total of n microbiome samples and a phylogenetic tree that depicts the evolutionary relationship among microbes. Each OTU is a tip node on the phylogenetic tree and each internal node is a taxonomic unit representing a common ancestor of its descendent taxa. In this article, we aggregate p OTUs to m genus level taxa as the basic analyzing units. However, basic analyzing unit can start at finer levels. Figure ?? displays an example phylogenetic tree used throughout the method and data simulation sections. Let $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, where $\mathbf{x}_i = (x_{i1}, \dots, x_{im})^T$ indicates the abundance of m genera of the i th subject, be the input data and $\mathbf{y} = (y_1, \dots, y_n)$ be the outcome of interest. Outcome variables

can be continuous, binary, or categorical.

2.2 Architecture of DeepBiome

DeepBiome is a neural network architecture that associates input vectors $\mathbf{x}$ (representing microbiome abundance) with a clinical outcome $\mathbf{y}$. A major challenge in constructing a neural network is to make decisions about the optimal number of layers and neurons in the network. The conventional wisdom of going deep (many layers) and wide (many neurons per layer) finds great success in many artificial intelligence tasks such as image pattern recognition and natural language processing but requires a huge amount of training data (Bergstra et al., 2011; Snoek et al., 2012), which is rarely affordable in biomedical studies due to resource constraints. DeepBiome prespecifies the network architecture according to the phylogenetic tree. The number of taxonomic levels decides the number of hidden layers and the number of taxa at each taxonomic level decides the number of neurons in the corresponding layer. Figure 1 illustrates an example DeepBiome architecture. The input layer receive microbiome abundances as inputs. The information is then propagated through multiple layers of the DeepBiome network to the outcome of interest $\mathbf{y}$. The input vector $\mathbf{x}$, e.g., the abundances of $m^{(0)}$ genera, is propagated to the first hidden layer vector $\mathbf{z}^{(1)}$ with a total of $m^{(1)}$ neurons, e.g., the number of family level taxa, using an $m^{(1)} \times m^{(0)}$ weight matrix $\mathbf{w}^{(1)}$ and an $m^{(1)}$ bias vector $\mathbf{b}^{(1)}$

$$\mathbf{z}^{(1)} = v(\mathbf{w}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}). \quad (1)$$

Each weight parameter $w_{jk}^{(1)}$ represents the effect of the k th input unit on the j th hidden neuron and $v(\cdot)$ is an activation function. Throughout this paper we use the rectified linear unit (ReLU) activations (Haykin, 1994; Maas et al., 2013)

$$\text{ReLU: } v(a) = a^+ = \max(0, a), \quad (2)$$

but it can be easily changed to other activation functions in our software. In the same manner, $\mathbf{z}^{(\ell)} = v(\mathbf{w}^{(\ell)}\mathbf{z}^{(\ell-1)} + \mathbf{b}^{(\ell)})$, $\ell = 2, \dots, L$, where L is the total number of hidden layers in the neural network. The last hidden layer $\mathbf{z}^{(L)}$ is linked to outcome using either an identity link or a softmax link. Specifically, we use identity link to predict continuous outcome, $\mathbf{y} = \mathbf{w}^{(L+1)}\mathbf{z}^{(L)} + \mathbf{b}$. For categorical outcome with K categories, softmax function is adopted to predict the probability of i th subject belongs to c th category,

$$\Pr(y_i = c) = \frac{e^{(\mathbf{w}\mathbf{z}^{(L)} + \mathbf{b})_c}}{\sum_{q=1}^K e^{(\mathbf{w}\mathbf{z}^{(L)} + \mathbf{b})_q}}. \quad (3)$$

Finally, we use $f_{\boldsymbol{\theta}}(\mathbf{x})$ with parameters $\boldsymbol{\theta} = \{\mathbf{w}, \mathbf{b}\}$ to represent the whole neural network that maps an input $\mathbf{x}$ to an output $\mathbf{y}$, where $\mathbf{w} = (\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(L)}, \mathbf{w}^{(L+1)})$ and $\mathbf{b} = (\mathbf{b}^{(1)}, \dots, \mathbf{b}^{(L)}, \mathbf{b}^{(L+1)})$.

2.3 Phylogeny regularization via weight decay

We introduce phylogeny regularization through weight decay. We assume that if taxa j and k have ancestor-descendent relationship, the associations between the corresponding neurons are stronger, i.e., larger weight value w_{jk} . When taxa j and k do not have this ancestral relationship, we assume $w_{j,k}$ to be a small value, i.e., weight decay. Thus, we construct a weight decay matrix ω to regularize weights in the neural network using evolutionary relationship carried by the phylogenetic tree. If nodes j and k are ancestor-descendent related, $\omega_{jk} = 1$; if not, ω_{jk} is a small value, e.g., 0.01. See Figure ?? as an illustration.

2.4 Model training

Given a training set consisting of training pairs $\{\mathbf{x}, \mathbf{y}\}$ and a neural network $f_{\boldsymbol{\theta}}(\mathbf{x})$ with parameters $\boldsymbol{\theta}$, a supervised training procedure is implemented to learn neural network function by minimizing the empirical loss. We define loss to be Mean Squared Error (MSE) for continuous outcome and standard Cross-Entropy (CE) for categorical outcome:

$$\begin{aligned} \text{MSE} &= \frac{1}{n} \sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{y}}_i)^2 \\ \text{CE} &= -\frac{1}{n} \sum_{i=1}^n \sum_{q=1}^K y_{i,q} \log \hat{y}_{i,q}, \end{aligned} \quad (4)$$

where $y_{i,k}$ is a binary indicator (0 or 1) indicating whether observation i belongs to class k , $c \in \{1, \dots, K\}$ (i.e., one hot encoding). $y_{i,c} = \Pr(y_i = c)$ is the probability that observation i belongs to class c which is defined by equation (3). In this study, the holdout validation method is used to determine the training stopping criterion. The training process stops when the holdout validation error achieves the minimum.

Adam optimizer, an adaptive gradient algorithm (Kingma and Ba, 2014), is used to train DeepBiome. Algorithm 1 describes the parameter estimation in Adam with the proposed phylogeny regularization. Specifically, at each update t , the estimated weight w_t is elementwisely multiplied by the corresponding regularization factor ω and $\omega \circ w_t$ is used to make the prediction of $\hat{\mathbf{y}}$ in the loss function (see equation [4]). Once the stopping criteria is met, the Algorithm 1 outputs the parameter estimation set of each layer $(\hat{\mathbf{w}}, \hat{\mathbf{b}}) = (\{\omega^{(1)} \circ \hat{\mathbf{w}}^{(1)}, \dots, \omega^{(L)} \circ \hat{\mathbf{w}}^{(L)}\}, \{\hat{\mathbf{b}}^{(1)}, \dots, \hat{\mathbf{b}}^{(L)}\})$. This phylogeny regularization effectively uses biological meaningful prior knowledge to limit the number of free parameters in the model, therefore avoids overfitting.

Data: $\mathbf{y}$, $\mathbf{x}_i = (x_1, \dots, x_m)$, phylogenetic tree designed matrix ω , learning rate $\alpha \in \mathcal{R}$

Result: $\hat{\boldsymbol{\theta}}$ such that the lost function $\text{Loss}(\boldsymbol{\theta})$ is minimized.

- 1 **Initialize:** parameter $(\mathbf{w}, \mathbf{b})_{t=0} \in \mathcal{R}$, 1st and 2nd moment vector $m_{t=0}^1 \leftarrow \mathbf{0}$ and $m_{t=0}^2 \leftarrow \mathbf{0}$, and step index $t \leftarrow 0$,
- 2 **repeat**
- 3 $t \leftarrow t + 1$
- 4 $l'_t \leftarrow \nabla_{\boldsymbol{\theta}} \text{Loss}(\omega \circ \mathbf{w}_{t-1}, \mathbf{b}_{t-1})$ (Get gradients w.r.t. stochastic objective at t)
- 5 $m_t^1 \leftarrow \beta_1 m_{t-1}^1 + (1 - \beta_1) l'_t$ (Update biased 1st moment estimate)
- 6 $m_t^2 \leftarrow \beta_2 m_{t-1}^2 + (1 - \beta_2) l''_t$ (Update biased 2nd moment estimate)
- 7 $\hat{m}_t^1 \leftarrow m_t^1 / (1 - \beta_1^t)$ (Update bias-corrected 1st moment estimate)
- 8 $\hat{m}_t^2 \leftarrow m_t^2 / (1 - \beta_2^t)$ (Update bias-corrected 2nd moment estimate)
- 9 $(\mathbf{w}, \mathbf{b})_t \leftarrow (\mathbf{w}, \mathbf{b})_{t-1} - \alpha \cdot \hat{m}_t^1 / (\sqrt{\hat{m}_t^2} + \epsilon)$ (Update parameters)
- 10 **until** *stopping criterion is met and return* $\hat{\boldsymbol{\theta}} = (\omega \circ \hat{\mathbf{w}}, \hat{\mathbf{b}})$;

Algorithm 1: Phylogeny regularized weight decay in Adam. β_1, β_2 refer to the exponential decay rates for the moment estimates in Adam. $\epsilon = 10^{-8}$ is used to prevent division from zero error (Kingma and Ba, 2014).

2.5 Performance metrics

We employ several statistical metrics to evaluate the performance of **DeepBiome** for its prediction, classification and taxa selection performances. For a quantitative outcome, the primary metric is the mean square error (MSE). Additionally, we report the Pearson correlation coefficient ρ between predicted $\hat{y}_i$ and true y_i . For categorical outcomes, i.e., classification problems, we measure their performance using sensitivity (true positive rate [TPR]), specificity, g -measure, accuracy (ACC), precision (PPV), and the F1 score:

$$\begin{aligned}
 \text{Sensitivity} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\
 \text{Specificity} &= \frac{\text{TN}}{\text{TN} + \text{FP}}, \\
 g\text{-Measure} &= (\text{Sensitivity} \times \text{Specificity})^{\frac{1}{2}}, \\
 \text{ACC} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \\
 \text{PPV} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\
 \text{F1 score} &= 2 \times \frac{\text{PPV} \times \text{TPR}}{\text{PPV} + \text{TPR}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}},
 \end{aligned}$$

where TP is true positive (or recall), TN is true negative, FP is false positive, and FN is false negative. F1 score is the harmonic mean of precision and sensitivity. An F1 score reaches its best value at one when the prediction has perfect precision and recall and the worst at 0. Note that F1 score does not take the true negative into account. We use the g -measure, which is the geometric mean of sensitivity and specificity, to assess the performance of a binary classifier. Same as F1 score, a g -Measure reaches its best value at one when the sensitivity and specificity are both perfect (one) while the worst at zero if any of sensitivity and specificity is zero. We also report AUC (area under the receiver operating characteristics), which reports the capability of a model to distinguish between classes. Sensitivity, specificity, g -measure, and ACC across all hidden layers (see Table 1) are used to report the selection accuracies.

3 Results

3.1 Simulation studies

We conducted extensive simulation studies to evaluate the performance of **DeepBiome** and compared it with conventional methods in three different schemes, i.e, linear regression, binary, and multiclass ($K \geq 3$) classification design. Throughout the simulation experiments, we used sample size $n = 1000$ and split them into a training set (75%, $n_{\text{training}} = 750$) and a test set (25%, $n_{\text{test}} = 250$). Different proportions of the split give qualitatively similar results (not shown). All results were obtained based on 1000 replicates. Simulation scenario 1 covers continuous outcome models; simulation scenario 2 is for binary outcome cases; and simulation scenario 3 considers the situation when outcome variable is categorical. Model robustness is evaluated in simulation scenario 4, when tree structure is mis-specified and when sequencing abundances contain measurement errors.

Generation of microbiome abundance data was described in detail by Zhai et al. (2018, 2019). Briefly, to generate an $n \times p$ OTU count matrix, we use a dirichlet multinomial (DM) distribution with the mean proportion vector and the dispersion parameter estimated from a real pulmonary microbiome dataset, which contains $p = 2964$ OTUs and a phylogenetic tree (Twigg III et al., 2016).

The association network between the microbiome $\mathbf{x}$ and the outcome $\mathbf{y}$ can be extremely complex. In this session, we use a forward propagation approach described below to generate $\mathbf{y}$. We start with 2964 OTUs and aggregate them as 48 genus. The following steps were used to generate outcome $\mathbf{y}$.

1. Read the phylogenetic tree to obtain the number of phylogenetic levels and nodes at each level. The microbiome data are then summarized at genus, family, order, class, and phylum

level as shown in Figure 1. Based on a real lung microbiome dataset, the number of nodes are $m^{(0)} = 48$, $m^{(1)} = 40$, $m^{(2)} = 23$, $m^{(3)} = 17$, and $m^{(4)} = 9$, respectively.

- Construct the weight matrix $\mathbf{w}^{(1)} \in \mathbb{R}^{m^{(1)} \times m^{(0)}}$ to propagate the input layer $\mathbf{x}_{\text{genus}}$ to the 1st hidden layer by

$$\mathbf{h}^{(1)} = \mathbf{w}^{(1)} \mathbf{x}_{\text{genus}} + \mathbf{b}^{(1)}.$$

The bias vector $\mathbf{b}^{(1)} \in \mathbb{R}^{m^{(1)} \times 1}$ follows a standard normal distribution $\mathcal{N}(0, \sigma_e^2)$ with $\sigma_e^2 = 4$. Suppose we have node j at the genus level and k at the family level, then

$$w_{j,k}^{(1)} \sim \begin{cases} \text{Uniform}(-0.5, 1) & \text{associated with output} \\ \mathcal{N}(0, 0.01) & \text{not associated with output.} \end{cases} \quad (5)$$

- Multiply the $w_{j,k}^{(1)}$ by a small value $\omega_{j,k} = 0.01$, if family taxa k is not a direct ancestor of genus taxa j ; otherwise, $w_{j,k}^{(1)}$ stays the same.
- Activate the neurons using ReLU in equation 2,

$$\mathbf{x}_{\text{family}} = \mathbf{z}^{(1)} = v(\mathbf{w}^{(1)} \mathbf{x}_{\text{genus}} + \mathbf{b}^{(1)}).$$

- Repeating step 2 - 4 to compute the $\mathbf{x}_{\text{order}}$, $\mathbf{x}_{\text{class}}$, and $\mathbf{x}_{\text{phylum}}$.
- Simulate the continuous or categorical output layer $\mathbf{y}$ as follow

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{w}^{(4)} \mathbf{x}_{\text{phylum}} + \mathbf{b}^{(4)} \\ \hat{\text{Pr}}(y_i = c) &= \frac{(e^{\mathbf{w}^{(4)} \mathbf{x}_{\text{phylum}} + \mathbf{b}^{(4)}})_c}{\sum_{q=1}^K (e^{\mathbf{w}^{(4)} \mathbf{x}_{\text{phylum}} + \mathbf{b}^{(4)}})_q} \end{aligned} \quad (6)$$

where $K = 2$ for binary classification and $K \geq 3$ for multicategorical classification.

Scenario 1: Regression design

In this section, two simulation strategies illustrated in Figure 2 were used. In strategy (1) microbiome taxa associated with outcome y (red and blue nodes) are clustered at the phylum level. In strategy (2), the associated taxa are clustered at phylum (red nodes) and order levels (blue nodes). Taxa represented by blue are negatively associated with continuous outcome $\mathbf{y}$, whereas taxa indicated as red are positively associated with $\mathbf{y}$. The gray taxa have no contribution.

We compare **DeepBiome** to linear regression as well as penalized regression with ℓ_1 norm (Lasso), ℓ_2 norm (Ridge), and $\ell_1 + \ell_2$ norm (Elastic-Net) penalties. We also compare **DeepBiome** to conventional DNN and ℓ_1 -regularized DNN. DNN and ℓ_1 -DNN use the same number of hidden layers and neurons on each layer as **DeepBiome** without phylogenetics tree regularization.

We use five-fold cross-validation to choose the tuning parameters for regularized linear regression models. For the deep learning models (i.e., DNN, ℓ_1 -DNN, and **DeepBiome**), we use a

holdout validation set and stop training when the validation error achieves the lowest point and starts to increase during the model training process. Once the optimal model is achieved, we evaluate the prediction performance on the test set. Twenty percent of the training data was used for the holdout validation. We trained the models until validation error increases or 5000 epochs using Adam (Kingma and Ba, 2014) optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$, learning rate $lr = 0.01$ and the mini-batch gradient descent with a mini-batch size of 50. The learning rate decayed for each epoch with $lr_{epoch+1} = lr_{epoch} \frac{1}{1+0.0001}$.

Table 2 displays the predictive performance by two metrics, MSE and Pearson’s correlation ρ , under a case that the outcome associated taxa are only clustered at one phylogenetic level (i.e., phylum). The outcome y predicted by DeepBiome has higher Pearson correlation and lower MSE on the test set than the regression methods and other deep learning models, which shows that DeepBiome has improved prediction performance. Overall, the penalized linear regression models, Lasso and Elastic Net, have slightly larger correlation coefficients on the test set compared to linear regression and Ridge regression. All deep learning models perform better than regression models with lower MSE and higher ρ . DeepBiome performs the best among all deep learning models. Table 3 shows the prediction performance under a more complex case, where the outcome-associated taxa are clustered at different phylogenetic levels (phylum and order). It is obvious that all regression schemes perform poorly in this case; the correlation values ρ are only around 0.6. DeepBiome has over 70% reduction in MSE compared to regression based methods. The deep learning models DNN and ℓ_1 -DNN improve ρ to 0.91 and 0.90 respectively. However both show hint of overfitting with lower testing performance. DeepBiome consistently achieves the best performances on the test set.

Identifying associated taxa at precise levels is critical for downstream biological validation. Figure 3, Supplementary Tables S1 and S2 use the metrics sensitivity, specificity, g-Measure and ACC to compare the selection performance of different methods. Regular regression methods do not discriminate associated taxa; therefore only the results of penalized regressions were included in the selection performance comparisons. The penalized regression schemes, Lasso and Elastic Net, can only select the taxa at one phylogenetic level. Here, we compute the performance metrics based on their phylogeny relationship. For example, if genus *Prevotella* is selected, we assume that its corresponding ancestor, phylum *Bacteroidetes*, is also selected. In contrast, the selection performance of regularized deep learning models are based on the weights estimated at each hidden layer. DeepBiome offers outstanding performances not only in terms of sensitivity and specificity, but also g-measure and ACC (Figure 3, the 1st row). Its g-Measure ranges from 0.80 to 0.91, while that of Lasso regression (the second best method) ranges from 0.54 to 0.72 (Supplementary Tables S1 and S2). Interestingly, despite being the second best method regarding prediction (see Table 3), ℓ_1 -DNN fails to identify the true microbiome taxa

across all phylogenetic levels.

Scenario 2: Binary classification

We consider the case that outcome associated taxa are clustered at a mixture of phylogenetic levels. For a binary outcome, we suppose

- (1) the higher the abundance of blue node taxa, the higher the probability of $\mathbf{y}$ belong to the disease group;
- (2) the higher the abundance of red node taxa, the higher the probability of $\mathbf{y}$ belong to the healthy control group.

We compare **DeepBiome** to logistic regression, three penalized logistic regression models, and two conventional deep learning networks. The same learning rate, stopping criteria, and mini-batch size (100) are used for **DeepBiome**, DNN and ℓ_1 -DNN. In Table 4, we present the metrics for evaluating the classification performance of binary outcome, including sensitivity, specificity, g-measure, ACC, and AUC. Logistic models have satisfying sensitivity values, but other metrics are not competitive compared to **DeepBiome**. They tend to have many false positives which lead to poor specificity. In contrast, **DeepBiome** achieves the best classification performance with the highest specificity, g-measure, ACC, and AUC reaching 0.84, 0.87, 0.89 and 0.94, respectively. Figure 3 (2nd row) and Supplementary Table S3 displays the performance of identifying the associated taxa. Although Lasso and ℓ_1 -DNN show good sensitivity at some phylogenetic levels, g-measure and ACC are much worse compared to Elastic-Net and **DeepBiome**. This suggests that Lasso and ℓ_1 -DNN tend to select more taxa (false positive). Using the order level as an example, the g-measure value of **DeepBiome** is 0.91, while the ℓ_1 -DNN is 0.15 (Supplementary Table S3).

Scenario 3: Multiclass classification

We simulated multi-categorical outcomes, e.g., the severity of illness, which may be categorized as “mild”, “moderate”, or “severe”. Consistent with previous simulations, we assume that the blue node taxa contribute to the “severe” group, reds contribute to the “mild” group, and part of the gray node taxa contribute to the neutral “moderate” group.

Table 5 presents the evaluation metrics for **DeepBiome**, DNN, ℓ_1 -DNN, and the support vector machine (SVM) with different kernels (Friedman et al., 2001). For SVM, both linear and non-linear kernels such as the radial or polynomial kernels are included. The default parameter setting is used for training SVM (Meyer et al., 2019). The same learning rate, stopping criteria, and mini-batch size (200) are used for **DeepBiome**, DNN and ℓ_1 -DNN. Among the SVMs, the linear SVM has the highest accuracy and AUC, while the SVM with radial kernel yields better

recall and F1 score. However, all SVMs are inferior to deep learning models: **DeepBiome** exhibits the highest values on all evaluation metrics, with an AUC of 0.90; AUCs of DNN and ℓ_1 -DNN are around 0.86. The F1 score of **DeepBiome** is 0.71, which is 14% higher than the second best, ℓ_1 -DNN. We find that **DeepBiome** offers the best and most balanced performance with precision and recall, which are 0.72 and 0.71, respectively. Since SVM models cannot perform selection on the microbiome taxa, we only compare **DeepBiome** to ℓ_1 -DNN in (Figure 3, 4th row and Supplementary Table S4). **DeepBiome** surpasses ℓ_1 -DNN in all of the evaluation metrics at all phylogenetic levels. For instance, the g-measure of ℓ_1 -DNN in selecting genus level taxa is only 0.19 while that of **DeepBiome** is 0.82 (see Supplementary Table S4).

Scenario 4: Robustness under tree mis-specification and measurement errors of microbiome abundance

To examine the robustness of **DeepBiome**, we consider two sources of model mis-specifications,

- (1) Abundance contain measurement errors at genus levels. We assume that 10% of the associated genus reads are mis-classified to one randomly selected genus from the same phylum. The microbiome abundance data with measurement errors is then used for training models.
- (2) The phylogenetic tree for training models is mis-specified (see Figure ??).
 - At class level, the genera that belong to *Clostridia* and *Flavobacteria* are mis-classified to *Bacilli* and *Bacteroidia*.
 - At order level, the genera that belong to *Coriobacteriales* and *Flavobacteriales* are mis-classified to *Actinomycetales* and *Bacteroidales*.

The same learning rate, stopping criteria, and mini-batch size (100) are used for **DeepBiome**, DNN and ℓ_1 -DNN. Tables 6, Figure 3, and Supplementary Table S5 present the results with measurement errors. Table 7 and Supplementary Table S6 show the results when using a mis-specified phylogenetic tree. Like Scenario 1, we compared **DeepBiome** to linear regression, penalized regression, conventional DNN, and ℓ_1 -regularized DNN. When the model is trained using data with measurement errors (case 1), performance of **DeepBiome** and ℓ_1 -DNN drops, i.e., higher MSE and lower Pearson’s ρ , compared to scenario 1 using data without errors (Table 6; see also Table 3). **DeepBiome** has the best prediction performance among all methods. The average Pearson’s ρ of **DeepBiome** is 0.95, while those of DNN and ℓ_1 -DNN are 0.87 and 0.91 respectively. Table 7 displays the predictive performance under case 2 (mis-specified phylogenetic tree). **DeepBiome** outperforms other methods in both MSE and Pearson’s ρ , demonstrating its robustness to tree mis-specifications. Figure 3 (5th and 6th row), Supplementary Tables S5 and S6 show the ability of identifying associated microbiome taxa. When the abundance data contain measurement errors, the sensitivity decreases in both penalized regression and deep

learning methods. For example, the specificity of **DeepBiome** at genus level is 0.67 (compared to 0.95 in Supplementary Table S2), leading to slightly decreases of g-Measure from 0.84 to 0.80. **DeepBiome** tends to select less associated taxa when input abundance data contains measurement errors. For the second case when the phylogenetic tree has taxonomic classification errors, the Lasso, Elastic-Net, and ℓ_1 -DNN have similar performance compare to scenario 1 (Figure 3, 6th row). Even if **DeepBiome** used a wrong tree structure to guide model, it still has a decent performance with g-measure of 0.80 at the finest level (Figure 3, 6th).

3.2 Application to the American Gut Project

The American Gut Project (AGP), launched in 2012, is an open platform for citizen microbiome research. Microbiome sample in AGP were collected from stools using dry swabs. Participants' self-reported meta data were entered through a web portal. Detailed data quality control (QC) and sequence processing procedures are reported in (McDonald et al., 2018). Resulting metadata and the operational taxonomic unit (OTU) table are available for download at <https://github.com/biocore/American-Gut/tree/master/data>. The phylogenetic tree was extracted from the *.biom* file that contains OTU tables and the taxonomic information. For this analysis, we further excluded microbiome samples with less than 10,000 sequence reads and genera with abundance less than 2% in all available samples. Samples without metadata or missing demographic information, i.e., age, gender, and ethnicity, were also excluded.

3.2.1 Type 2 Diabetes

Type 2 diabetes (T2D) is a metabolic disorder with a combination of risk factors such as family history, lifestyle and genetic factors. It has imposed a huge and growing burden on the public health (Cho et al., 2018). In the past decade, multiple studies have indicated that the risk of developing type 2 diabetes (T2D) may also involve factors from gut microbiome (Larsen et al., 2010; Musso et al., 2011; Qin et al., 2012). In this application, using **DeepBiome**, we aim to further understand the association between gut microbiome and T2D as well as provide a complex microbiome-phenotype association network at each phylogenetic level. We compare the performance of **DeepBiome** with logistic regression, logistic regression with Ridge, Lasso, and Elastic-Net penalization, DNN, and ℓ_1 -DNN. Subjects who reported T2D diagnosed by medical professionals (doctor or physician assistant) were included as T2D cases ($n = 154$). We randomly select 154 subjects without T2D to form a control group. We use their demographic information, age, gender, and ethnicity (Supplementary Table S7), 373 genus level taxa, and a phylogenetic tree to classify T2D as well as to select associated microbiome taxa.

Table 8 shows the performance of classifying T2D based on 5-fold cross-validation. Although Elastic-Net and Lasso regression have the highest sensitivity, their low specificities suggest that

these methods intend to predict healthy subjects as T2D (false positive). DNN and ℓ_1 -DNN show signs of over training. DeepBiome performs the best among all methods with highest g-Measure, accuracy, and AUC, which are 0.65, 0.64, and 0.66, respectively.

Figure 5 summarizes the taxa selected by DeepBiome using the 85th percentile of the weight coefficient estimated at each phylogenetic level averaged over 5-fold cross validation. In total, we have selected 56 genera, 15 families, 8 orders, and 6 classes. Among these taxa, 33 genera, 7 families, 2 orders, and 1 class are positively associated with T2D, indicating that the higher the abundance of those taxa, the higher the probability of subjects having T2D. Compared with the analysis carried out by Qin et al. (2012), we also selected T2D enriched genus *Alistipes* and *Lachnospira*, and healthy control enriched genus *Haemophilus* and family *Erysipelotrichaceae*. In an analysis of a Denmark T2D cohort, Larsen et al. (2010) have pointed out that T2D is associated with gut microbiome dysbiosis, such as the proportion of *Clostridia* in diabetics is significantly lower than that in controls and class *Betaproteobacteria* is highly enriched in subjects with diabetes. However, DeepBiome suggests that class *Betaproteobacteria* is negatively associated with T2D and a group of genera positively associated with T2D are from class *Clostridia*, which also disagrees with Larsen et al. (2010). Indeed, results from Larsen et al. (2010) can not be validated in other cohorts with different ethnicity (Sircana et al., 2018; Qin et al., 2012; Karlsson et al., 2013). This results suggest that the contribution of gut microbiome to T2D may be different across ethnicity groups and environmental factors. Beside T2D, the selected *Haemophilus* from AGP has been demonstrated to be associated with prediabetes (Zhang et al., 2013).

3.2.2 Body Mass Index (BMI)

Turnbaugh et al. (2006) have identified the gut microbiota as an additional contributing factor to the pathophysiology of obesity. Although microbiome are by no means the sole factor in the obesity epidemic, multiple studies have uncovered profound changes in the composition and metabolic function of the gut microbiota among obese individuals (Turnbaugh et al., 2009; Tilg and Kaser, 2011; Wu et al., 2011; Verdam et al., 2013). In this analysis, we apply DeepBiome to further uncover the association between BMI or obesity and microbiome taxa. Subjects who are 18-80 years old with BMI or weight and height are included for analysis ($n = 9292$). The continuous BMI is either available in AGP metadata or is calculated from their reported weight and height ($\text{weight (kg)} / [\text{height (m)}]^2$) (Supplementary Table S8). The BMI categories are defined according to adult BMI index chart from U.S. Centers for Disease Control and Prevention (CDC, <https://www.cdc.gov/obesity/adult/defining.html>). Specifically, we have 5607 subjects whose BMI are normal (i.e., $18.5 \leq \text{BMI} < 25$), 399 underweight ($\text{BMI} < 18.5$), 2346 overweight ($25 \leq \text{BMI} < 30$), and 940 obese ($\text{BMI} \geq 30$). We used demographic information,

i.e., age, gender, and ethnicity (Supplementary Table S8), and gut microbiome to predict BMI and obesity categories.

Tables 9 and 10 present the performance of predicting continuous and categorical BMI, respectively. **DeepBiome** has the highest testing Pearson correlation for predicting continuous BMI (Table 9). To classify BMI categories (Table 10), all deep learning methods have higher accuracies and **DeepBiome** has the highest AUC. The selected taxa were reported by using the 85th percentile of the weight coefficient estimated by **DeepBiome** at each taxonomic level averaged across 5-fold cross validation. In total, we identified 25 genera, 8 families, 5 orders, and 3 classes that are positively associated with BMI. The selected taxa are highly consistent with previous findings. For example, Wu et al. (2011) reported that *veillonellaceae* and *Bacteroides* were positively correlated to BMI. Patil et al. (2012) also observed long-term high-fat diet shaped the genus *Bacteroides* abundances in the obese subjects. Similarly, we show that higher abundance of unclassified genus *veillonellaceae* and *Bacteroides* predict higher BMI. Turnbaugh et al. (2009) stated that the microbiome community among obese population have reduced diversity, which shift to higher *Firmicutes* abundance in obese subjects. In this analysis, 10 out of the 25 selected genera with positive association with BMI are from phylum *Firmicutes*. Furthermore, *Firmicutes* shows directly positive association with higher BMI. Consistent with Tims et al. (2013) and Ferrer et al. (2013), genus *Eubacterium* with increased abundance in obese category is also from *Firmicutes*. Previous studies have shown occasionally contradictory enrichment for specific taxa, e.g., *Lactobacillus* Collado et al. (2007); Bervoets et al. (2013). **DeepBiome** enables microbiome-phenotype association analysis and provide a deep understanding of their relationships.

3.3 Pulmonary microbiome in HIV infected population

Recent research has uncovered the relationship between pulmonary function and the respiratory microbiota (Zhai et al., 2018, 2019). Evidence suggests that two distinct microbiome communities, defined as the supraglottic (SPT) and background (BPT) predominant taxa (Segal et al., 2016), have different characteristics in relation to lung function. Zhou et al. (2019) present a co-occurrence network of genus-level taxa, where the SPT taxa derived from the oral cavity and supraglottic region are associated with reduced lung function and the BPT taxa may act as a neutral or protective factor. Therefore, in this application, we aim to further understand the microbiome and visualize the complex microbiome-phenotype network at each phylogenetic level. We compare the performance of **DeepBiome** with Ridge, Lasso, ElasticNet, DNN, and ℓ_1 -DNN in analysis of this pulmonary microbiome dataset. The microbiome samples were collected from bronchoalveolar lavage (BAL) fluid of $n = 30$ HIV infected patients whose CD4 count was less than 500 cells/mm³. Most of the 30 subjects are white ($n=20$, 66.7%), male ($n=24$, 80%), and

smoker (n=17, 56.7%). All subjects are adults with age between 18-53 (mean±sd, 36.6±9.0). Detailed description of the study design and data collection method has been published by Twigg III et al. (2016); Zhai et al. (2018, 2019).

Forced expiratory volume in one second, or FEV1, is a lung function biomarker that monitors the chronic obstructive pulmonary disease (COPD) or other lung diseases over time. FEV1 is usually converted to a percentage of normal through an established equation with input of height, weight, and race. In this application, we use their demographic information (age, gender, ethnicity, and smoking history), 48 genus level taxa, and a phylogenetic tree to predict FEV1_{% predicted} (96.3%±17.3%, range: 69%–140%) and to discover the associated microbiome taxa. Due to the limited sample size, we apply the leave-one-out cross-validation, which trains the model on all samples except observation i ($i = 1, \dots, 30$), and then computes the test error on the held out point.

Table 11 shows the MSE and Pearson’s ρ . Among the deep learning models, DeepBiome performs the best with smallest MSE. Figure 6 depicts the microbiome taxa selected via penalized regression methods (Lasso and Elastic-Net) and deep learning models (ℓ_1 -DNN and DeepBiome). No genus is selected by Lasso associated with FEV1_{%predicted}, while all genera except *Filifactor* and *Bradyrhizobium* are selected by Elastic-Net. Elastic-Net only identifies three genera with negative coefficient estimation, and other risk SPT-characteristic taxa are selected as protective factors for lung function, which is contradictory with previous studies (Segal et al., 2016; Zhou et al., 2019). Because the penalized regression models can only perform the selection at one phylogenetic level, it is unknown how taxa contribute to the outcome FEV1_{%predicted} at the other phylogenetic levels. Therefore, in Figure 6 (a)-(b), we use lines with grey color to express this unknown status. Since ℓ_1 -DNN ignores the evolutionary relationship carried by the tree in the training process, no biological meaningful genus can be selected. We use grey lines and black nodes indicating the 85th percentile genera ranked by their absolute weight coefficient estimation in Figure 6(c). DeepBiome successfully locates the selection at each phylogenetic level. Specifically, DeepBiome identifies five genera that are negative associated with FEV1_{%predicted}, including *TM7 genera incertae sedis* (SPT), *Capnocytophaga* (NDT), *Veillonella* (SPT), *Staphylococcus* (BPT), and *Moryella* (SPT). Except *Staphylococcus*, the other three all have strong positive clustering effects with SPT-characteristic taxa, which are risk factors for lung function suggested by (Zhou et al., 2019). In addition, genus *Sphingomonas*(BPT) and *Bradyrhizobium*(BPT) are selected as protective factors for outcome FEV1. Segal et al. (2016); Zhou et al. (2019) have shown that the BPT-characteristic taxa are related with better lung function and lower level of inflammation. An increase in lower respiratory tract inflammation links to the lung disease, especially in HIV-infected subjects with a vulnerable immune system. *Lachnospiraceae*, the ancestor of genus *Moryella* and *Oribacterium*, is identified at the family

level. Interestingly, in the same study, both genera show significant negative correlation with the spirometry measurement FEV1. **DeepBiome** reveals *Spirochaetales* and *Bacillales* with negative association at the order level. The results are consistent with previous work that shows *Treponema*, which belongs to *Spirochaetales*, has significant positive correlation with an inflammatory marker, interleukin-8 (IL-8) (Zhou et al., 2019) and negative correlation with another spirometry measurement FEF. Furthermore, genus *Selenomonas* from *Negativicutes*, which is the only taxon selected at class level, are also positively correlated with interleukin-6 (IL-6). Caution need to be taken in this analysis with a very small sample size. Even, with slightly higher MSE, **DeepBiome** show clear advantages over other deep learning methods. Regression methods show smaller MSE indicating better prediction performance than **DeepBiome** but they cannot identify meaningful taxa and sacrifice interpretability.

3.4 Computational Efficiency

DeepBiome is implemented in Python 3.6 based the TensorFlow (Abadi et al., 2015, 2016) and Keras (Chollet et al., 2015) framework. It can be built on Python 3.4, 3.5, and 3.6. All simulations are performed using a workstation equipped with Intel(R) Xeon(R) CPU E5-2650 v4 processor with 24 cores @ 2.20GHz and one NVIDIA GeForce GTX TITAN X GPU with 3072 CUDA cores @ 1 GHz and 12GB memory. **DeepBiome** required 290 ± 69 seconds to fully train the network for one replicate with 1000 samples, 50 mini-batches and 5000 epochs. For the same data, DNN took 282 ± 67 second and ℓ_1 -DNN took 282 ± 67 second. **DeepBiome** and all other deep learning approaches took less than 0.004 seconds for prediction. All real data analysis is performed on a MacBook Pro with 2.8GHz Intel Core i7 processor and 16GB 2133MHz LPDDR3 memory.

4 Discussion and conclusions

The proposed **DeepBiome**, a phylogenetic tree-regularized deep learning model, can be used for both prediction and classification tasks. We provide comprehensive simulation experiments and real data applications to demonstrate the superiority of **DeepBiome**. For regression tasks, our results suggest that, compared to sparse regression and other deep learning models, **DeepBiome** has a competitive performance particularly when microbiome taxa associated with outcome are clustered at different phylogenetic levels. **DeepBiome** also excels in complex classification tasks with higher accuracy and AUC. More importantly, **DeepBiome** enables an intuitive visualization of microbiome-phenotype association network.

Deep learning models are gaining much popularity due to their supremacy in imaging and natural language analysis. However typical biomedical studies can rarely afford the huge amount

of training data required for hyperparameter tuning (Shen et al., 2017; De Fauw et al., 2018). **DeepBiome** regularizes the neural network structure towards the phylogenetic structure inherent in the microbiome data through weight decay. This way it greatly reduces the number of parameters, including the architecture itself, to be tuned and trained, avoids overfitting, and allows visualization of the pathway from microbiome counts to phenotypes.

The limitations of **DeepBiome** include the possibility of violation of the assumptions: (1) microbiome classified in the same cluster have similar effects to outcome of interests and (2) phylogenetic tree structure translates to effects aggregation structure. The extension of **DeepBiome** to longitudinal microbiome data is also needed for many studies with repeated measures of microbiome.

Availability and Requirements

Project name: DeepBiome

Project home page: <https://github.com/Young-won/DeepBiome>

Operating system(s): Platform independent

Programming language: Python

Other requirements: None

License: Open source

Any restrictions to use by non-academics: None

List of abbreviations

DeepBiome: A phylogenetic tree informed deep neural network for microbiome data analysis; RNA: Ribosomal ribonucleic acid; OTUs: Operational Taxonomic Units; HIV: Human immunodeficiency Virus; DNN: Deep neural network; BMI: Body mass index; CNN: Convolutional neural network; AGP: American gut project; SVM: Support vector machine; Adam: An adaptive gradient algorithm; MSE: mean square error; AUC: area under the receiver operating characteristics; TPR: True positive rate; ACC: accuracy; PPV: precision; TP: true positive (recall); TN is true negative; FP: False positive; FN: False negative; T2D: Type 2 diabetes; COPD: Chronic obstructive pulmonary disease; FEV1: Forced expiratory volume in one second; SPT: Supraglottic predominant taxa; BPT: Background predominant taxa.

Declarations

Ethics approval and consent to participate

All utilized microbiome datasets are de-identified secondary data analysis. No ethics approval or consent to participate was required for this study.

Consent for publication

All utilized microbiome datasets are de-identified secondary data analysis. No consent for publication was required for this study.

Availability of data and material

AGP metadata and the operational taxonomic unit (OTU) table were downloaded from <https://github.com/biocore/American-Gut/tree/master/data>. The phylogenetic tree was extracted from the *.biom* file that contains OTU tables and the taxonomic information. Lung microbiome data is available upon request.

Competing interests

The authors declare that they have no competing interests.

Funding

This work has been supported by NIH grant (K01DK106116, HG006139, GM105785, GM53275, U01 HL121831, U01 HL098960), NSF grant (DMS-1645093), and Arizona Biomedical Research Commission (ABRC) grant.

Authors' contributions

Jing Zhai, Youngwon Choi, Hua Zhou and Jin J. Zhou designed the model and the computational framework. Jing Zhai and Youngwon Choi implemented the method and carried out the analysis. Kenneth Knox and Homer L. Twigg III provided lung microbiome datasets. Yin Chen, Kenneth Knox, Homer L. Twigg III, and Joong-Ho Won helped to interpret the results. Joong-Ho Won supported computing recourse. All authors provided critical feedback and helped shape the research, analysis and manuscript. Jin J. Zhou supervised project.

Acknowledgements

The authors are grateful to the editors and reviewers for their insightful observations and comments.

References

- Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL <http://tensorflow.org/>. Software available from tensorflow.org.
- Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, 2016. URL <https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf>.
- James S. Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyperparameter optimization. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2546–2554. Curran Associates, Inc., 2011. URL <http://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf>.
- Liene Bervoets, Kim Van Hoorenbeeck, Ineke Kortleven, Caroline Van Noten, Niel Hens, Carl Vael, Herman Goossens, Kristine N Desager, and Vanessa Vankerckhoven. Differences in gut microbiota composition between obese and lean children: a cross-sectional study. *Gut Pathogens*, 5(1):10, 2013.
- Jun Chen, Frederic D Bushman, James D Lewis, Gary D Wu, and Hongzhe Li. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, 14(2):244–258, 2012.
- Li Chen, Han Liu, Jean-Pierre A Kocher, Hongzhe Li, and Jun Chen. glmgraph: an r package for variable selection and predictive modeling of structured genomic data. *Bioinformatics*, 31(24):3991–3993, 2015.

NH1 Cho, JE Shaw, S Karuranga, Y Huang, JD da Rocha Fernandes, AW Ohlrogge, and B Ma-
landa. Idf diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections
for 2045. *Diabetes Research and Clinical Practice*, 138:271–281, 2018.

François Chollet et al. Keras. <https://keras.io>, 2015.

M Carmen Collado, Muriel Derrien, Erika Isolauri, Willem M de Vos, and Seppo Salminen.
Intestinal integrity and akkermansia muciniphila, a mucin-degrading member of the intestinal
microbiota present in infants, adults, and the elderly. *Appl. Environ. Microbiol.*, 73(23):7767–
7770, 2007.

Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad
Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O’Donoghue, Daniel
Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease.
Nature Medicine, 24(9):1342, 2018.

Manuel Ferrer, Alicia Ruiz, Francesca Lanza, Sven-Bastiaan Haange, Andreas Oberbach, Holger
Till, Rafael Bargiela, Cristina Campoy, Maria Teresa Segura, Michael Richter, et al. Micro-
biota from the distal guts of lean and obese adolescents exhibit partial functional redundancy
besides clear differences in community structure. *Environmental Microbiology*, 15(1):211–226,
2013.

Eric A Franzosa, Alexandra Sirota-Madi, Julian Avila-Pacheco, Nadine Fornelos, Henry J
Haiser, Stefan Reinker, Tommi Vatanen, A Brantley Hall, Himel Mallick, Lauren J McIver,
et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature*
Microbiology, 4(2):293, 2019.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*,
volume 1. Springer series in statistics New York, 2001.

Tanya P Garcia, Samuel Müller, Raymond J Carroll, and Rosemary L Walzem. Identifica-
tion of important regressor groups, subgroups and individuals via regularization methods:
application to gut microbiome data. *Bioinformatics*, 30(6):831–837, 2013.

Jack A Gilbert, Robert A Quinn, Justine Debelius, Zhenjiang Z Xu, James Morton, Neha Garg,
Janet K Jansson, Pieter C Dorrestein, and Rob Knight. Microbiome-wide association studies
link dynamic microbial consortia to disease. *Nature*, 535(7610):94, 2016.

Amit Gupta and Siuwa M Lam. Weight decay backpropagation for noisy data. *Neural Networks*,
11(6):1127–1138, 1998.

Simon Haykin. *Neural Networks*, volume 2. Prentice hall New York, 1994.

- Fredrik H Karlsson, Valentina Tremaroli, Intawat Nookaew, Göran Bergström, Carl Johan Behre, Björn Fagerberg, Jens Nielsen, and Fredrik Bäckhed. Gut metagenome in european women with normal, impaired and diabetic glucose control. *Nature*, 498(7452):99, 2013.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.
- Anders Krogh and John A Hertz. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems*, pages 950–957, 1992.
- Nadja Larsen, Finn K Vogensen, Frans WJ Van Den Berg, Dennis Sandris Nielsen, Anne Sofie Andreassen, Bente K Pedersen, Waleed Abu Al-Soud, Søren J Sørensen, Lars H Hansen, and Mogens Jakobsen. Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PloS One*, 5(2):e9085, 2010.
- Yang Young Lu, Yingying Fan, Jinchi Lv, and William Stafford Noble. Deeppink: reproducible feature selection in deep neural networks. *CoRR*, abs/1809.01185, 2018. URL <http://arxiv.org/abs/1809.01185>.
- Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. *Proc. ICML*, 30(1):3, 2013.
- Daniel McDonald, Embriette Hyde, Justine W Debelius, James T Morton, Antonio Gonzalez, Gail Ackermann, Alexander A Aksenov, Bahar Behsaz, Caitriona Brennan, Yingfeng Chen, et al. American gut: an open platform for citizen science microbiome research. *mSystems*, 3(3):e00031–18, 2018.
- David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, Friedrich Leisch, Chih-Chung Chang, Chih-Chen Lin, and Maintainer David Meyer. Package ‘e1071’. *R Journal*, 2019.
- David Bryan Mundie and Lloyd W Massengill. Weight decay and resolution effects in feedforward artificial neural networks. *IEEE Transactions on Neural Networks*, 2(1):168–170, 1991.
- Giovanni Musso, Roberto Gambino, and Maurizio Cassader. Interactions between gut microbiota and host metabolism predisposing to obesity and diabetes. *Annual Review of Medicine*, 62:361–380, 2011.
- Josephine Ni, Ting-Chin David Shen, Eric Z Chen, Kyle Bittinger, Aubrey Bailey, Manuela Roggiani, Alexandra Sirota-Madi, Elliot S Friedman, Lillian Chau, Andrew Lin, et al. A role for bacterial urease in gut dysbiosis and crohn’s disease. *Science Translational Medicine*, 9(416):eaah6888, 2017.

- Deepak P Patil, Dhiraj P Dhotre, Sachin G Chavan, Armiya Sultan, Dhawal S Jain, Vikram B Lanjekar, Jayshree Gangawani, Poonam S Shah, Jayshree S Todkar, Shashank Shah, et al. Molecular analysis of gut microbiota in obesity among indian individuals. *Journal of Biosciences*, 37(4):647–657, 2012.
- Junjie Qin, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, Wenwei Zhang, Yuanlin Guan, Dongqian Shen, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55, 2012.
- Derek Reiman, Ahmed Metwally, and Yang Dai. Using convolutional neural networks to explore the microbiome. In *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*, pages 4269–4272. IEEE, 2017.
- R Balfour Sartor. Microbial influences in inflammatory bowel diseases. *Gastroenterology*, 134(2):577–594, 2008.
- Patrick D Schloss and Jo Handelsman. Introducing dotur, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.*, 71(3):1501–1506, 2005.
- Leopoldo N Segal, Jose C Clemente, Jun-Chieh J Tsay, Sergei B Koralov, Brian C Keller, Benjamin G Wu, Yonghua Li, Nan Shen, Elodie Ghedin, Alison Morris, et al. Enrichment of the lung microbiome with oral taxa is associated with lung inflammation of a th17 phenotype. *Nature Microbiology*, 1(5):16031, 2016.
- Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19:221–248, 2017.
- Antonio Sircana, Luciana Framarin, Nicola Leone, Mara Berrutti, Francesca Castellino, Renato Parente, Franco De Michieli, Elena Paschetta, and Giovanni Musso. Altered gut microbiota in type 2 diabetes: Just a coincidence? *Current Diabetes Reports*, 18(10):98, 2018.
- Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2951–2959. Curran Associates, Inc., 2012. URL <http://papers.nips.cc/paper/4522-practical-bayesian-optimization-of-machine-learning-algorithms.pdf>.
- Herbert Tilg and Arthur Kaser. Gut microbiome, obesity, and metabolic dysfunction. *The Journal of Clinical Investigation*, 121(6):2126–2132, 2011.

- Sebastian Tims, Catherine Derom, Daisy M Jonkers, Robert Vlietinck, Wim H Saris, Michiel Kleerebezem, Willem M De Vos, and Erwin G Zoetendal. Microbiota conservation and bmi signatures in adult monozygotic twins. *The ISME journal*, 7(4):707, 2013.
- Peter J Turnbaugh, Ruth E Ley, Michael A Mahowald, Vincent Magrini, Elaine R Mardis, and Jeffrey I Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122):1027, 2006.
- Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project. *Nature*, 449(7164):804, 2007.
- Peter J Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L Cantarel, Alexis Duncan, Ruth E Ley, Mitchell L Sogin, William J Jones, Bruce A Roe, Jason P Affourtit, et al. A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480, 2009.
- Homer L Twigg III, Kenneth S Knox, Jin Zhou, Kristina A Crothers, David E Nelson, Evelyn Toh, Richard B Day, Huaiying Lin, Xiang Gao, Qunfeng Dong, et al. Effect of advanced HIV infection on the respiratory microbiome. *American Journal of Respiratory and Critical Care Medicine*, 194(2):226–235, 2016. doi: 10.1164/rccm.201509-1875OC.
- Froukje J Verdam, Susana Fuentes, Charlotte de Jonge, Erwin G Zoetendal, Runi Erbil, Jan Willem Greve, Wim A Buurman, Willem M de Vos, and Sander S Rensen. Human intestinal microbiota composition is associated with local and systemic inflammation in obesity. *Obesity*, 21(12):E607–E615, 2013.
- Tao Wang and Hongyu Zhao. Constructing predictive microbial signatures at multiple taxonomic levels. *Journal of the American Statistical Association*, 112(519):1022–1031, 2017.
- Gary D Wu, Jun Chen, Christian Hoffmann, Kyle Bittinger, Ying-Yu Chen, Sue A Keilbaugh, Meenakshi Bewtra, Dan Knights, William A Walters, Rob Knight, et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108, 2011.
- Jian Xiao, Hongyuan Cao, and Jun Chen. False discovery rate control incorporating phylogenetic tree increases detection power in microbiome-wide multiple testing. *Bioinformatics*, 33(18):2873–2881, 2017.
- Jian Xiao, Li Chen, Stephen Johnson, Xianyang Zhang, and Jun Chen. Predictive modeling of microbiome data using a phylogeny-regularized generalized linear mixed model. *Frontiers in Microbiology*, 9:1391, 2018.

- Jing Zhai, Juhyun Kim, Kenneth S Knox, Homer L Twigg III, Hua Zhou, and Jin J Zhou. Variance component selection with applications to microbiome taxonomic data. *Frontiers in Microbiology*, 9:509, 2018.
- Jing Zhai, Kenneth Knox, Homer L Twigg III, Hua Zhou, and Jin J Zhou. Exact variance component tests for longitudinal microbiome studies. *Genetic Epidemiology*, 43(3):250–262, 2019.
- Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse. Three mechanisms of weight decay regularization, 2018.
- Xiuying Zhang, Dongqian Shen, Zhiwei Fang, Zhuye Jie, Xinmin Qiu, Chunfang Zhang, Yingli Chen, and Linong Ji. Human gut microbiota changes reveal the progression of glucose intolerance. *PloS One*, 8(8):e71108, 2013.
- Jin J Zhou, Jing Zhai, Hua Zhou, Yin Chen, Ian Robey, George Weinstock, Erica Weinstock, Kenneth S Knox, and Homer L Twigg III. The lung microbiome is associated with lung function and inflammation in an hiv infected population treated with antiretroviral therapy. *In submission*, 2019.

Figure and Table Legend

Figure 1: DeepBiome architecture. (a) A phylogenetic tree with 48 genera as tip nodes. Color represents phylum types. (b) Network layout of DeepBiome architecture. The input layer is genus level microbiome abundance. Each hidden layer represents one phylogenetic level, e.g., family, order, class, and phylum. The dark lines represent relationship defined by a phylogenetics tree. The gray lines represent association between layers. (c) Phylogenetic tree regularized weight decay. Suppose we have a simple tree as shown in the left panel, which has 6 genera (taxa 1-6) and 2 classes (7-8). Genera 1-3 belong to class 7 and genera 4-6 belong to class 8. The ancestor-descendent information is embedded into a 6×2 matrix. Without loss of generality, we use ω to indicate a regularization factor with small value (e.g., 0.01). For tree regularized weight decay, the weight estimation matrix $w_{6 \times 2}$ is multiplied with this phylogenetic embedded matrix $\Omega_{6 \times 2}$ elementwisely, denoted by $\Omega_{6 \times 2} \circ w_{6 \times 2}$.

Figure 2: Simulation specifications. The outcome associated taxa (blue and red) are specified at (a) the phylum level and (b) a mixture of phylum & order levels. The blue nodes represent “bad” taxa which result in a disease status or are negatively associated with continuous phenotype, e.g., FEV1. The red nodes represent “good” taxa which result in a healthy status. In simulation scenario 4, we evaluate the impact of mis-specified phylogenetic tree: (c) indicates the true phylogenetic tree used in simulation scenario 4 and (d) indicates the phylogenetic tree used in model learning (same as the tree shown in Figure (b)).

Figure 3: Taxa selection performance under 4 simulation schemes at each phylogenetic level. Sensitivity, specificity, g-measure, and accuracy (ACC) were used to evaluate taxa selection performance. Vertical bar represents standard deviation over 1000 simulation replicates.

Figure 4: DeepBiome selected T2D associated taxa from phylum *Proteobacteria* among the American Gut Project (AGP). Estimated weights were overlaid on *Proteobacteria* branch of the phylogenetic tree. One hundred and fifty-five self-reported T2D was extracted from survey and 154 non-T2D participants were randomly sampled as controls. The red and blue nodes indicate taxa have positive and negative weights, respectively. The size of colored nodes represent the magnitudes of the weights. Black nodes represent non-selected taxa.

Figure 5: DeepBiome selected T2D associated taxa from all phylums overlaid on the phylogenetic tree. Microbiome and phenotype data were downloaded from American Gut Project (AGP). One hundred and fifty-five self-reported T2D was extracted from survey. A

random selected 154 non-T2D were then sampled to serve as controls. The blue and red nodes indicate taxa have negative and positive weights, respectively. The size of colored nodes represent the magnitudes of the weights. Black nodes represent non-selected taxa.

Figure 6: Selection of FEV1_{% predicted} associated microbiome taxa using Lasso, Elastic-Net, ℓ_1 -DNN, and DeepBiome. The red nodes indicate taxa have positive association with FEV1_{% predicted} and blue nodes indicate taxa have negative association with FEV1_{% predicted}.

Table 1: Metrics used to assess the performance of outcome prediction and microbiome taxa selection.

Table 2: Scenario 1. Mean squared error (MSE) and Pearson correlation coefficient between predicted and true outcomes for continuous outcome. The associated taxa are clustered at the phylum level.

Table 3: Scenario 1. Mean squared error (MSE) and Pearson correlation coefficient between predicted and true outcomes for continuous outcome. The associated taxa are clustered at the phylum and order levels.

Table 4: Scenario 2. Classification performance for binary outcomes.

Table 5: Scenario 3. Classification performance for multi-categorical outcomes.

Table 6: Scenario 4. Mean squared error (MSE) and Pearson correlation coefficient between predicted and true outcome (continuous), when the input microbiome abundance data contain measurement errors. The associated taxa are clustered at the phylum and order levels.

Table 7: Scenario 4. Mean squared error (MSE) and Pearson correlation coefficient between predicted and true outcome (continuous), when using an mis-specified phylogenetic tree. The associated taxa are clustered at the phylum and order levels.

Table 8: Type 2 Diabetes (T2D) prediction performance using microbiome composition in the American Gut Project (AGP).

Table 9: Body Mass Index (BMI, continuous) prediction using microbiome composition in the American Gut Project (AGP).

Table 10: Body Mass Index (BMI, multi-categorical) classification using microbiome composi-

800 tion in the American Gut Project (AGP).

801 Table 11: Predicting FEV1_{%predicted} from lung microbiome composition.

(a) An example phylogenetic tree with 48 genera (b) Network layout of DeepBiome architecture

(c) Phylogenetic tree regularized weight decay procedure

Figure 1: DeepBiome architecture. (a) A phylogenetic tree with 48 genera as tip nodes. Color represents phylum types. (b) Network layout of DeepBiome architecture. The input layer is genus level microbiome abundance. Each hidden layer represents one phylogenetic level, e.g., family, order, class, and phylum. The dark lines represent relationship defined by a phylogenetics tree. The gray lines represent association between layers. (c) Phylogenetic tree regularized weight decay. Suppose we have a simple tree as shown in the left panel, which has 6 genera (taxa 1-6) and 2 classes (7-8). Genera 1-3 belong to class 7 and genera 4-6 belong to class 8. The ancestor-descendent information is embedded into a 6×2 matrix. Without loss of generality, we use ω to indicate a regularization factor with small value (e.g., 0.01). For tree regularized weight decay, the weight estimation matrix $w_{6 \times 2}$ is multiplied with this phylogenetic embedded matrix $\Omega_{6 \times 2}$ elementwisely, denoted by $\Omega_{6 \times 2} \circ w_{6 \times 2}$.

(a) Phylum Level

(b) Phylum & Order Levels

(c) Simulation scenario 4:
Tree in network simulation

(d) Simulation scenario 4:
Tree in model learning

Figure 2: Simulation specifications. The outcome associated taxa (blue and red) are specified at (a) the phylum level and (b) a mixture of phylum & order levels. The blue nodes represent “bad” taxa which result in a disease status or are negatively associated with continuous phenotype, e.g., FEV1. The red nodes represent “good” taxa which result in a healthy status. In simulation scenario 4, we evaluate the impact of mis-specified phylogenetic tree: (c) indicates the true phylogenetic tree used in simulation scenario 4 and (d) indicates the phylogenetic tree used in model learning (same as the tree shown in Figure (b)).

Figure 3: Taxa selection performance under 4 simulation schemes at each phylogenetic level. Sensitivity, specificity, g-measure, and accuracy (ACC) were used to evaluate taxa selection performance. Vertical bar represents standard deviation over 1000 simulation replicates.

Figure 4: DeepBiome selected T2D associated taxa from phylum *Proteobacteria* among the American Gut Project (AGP). Estimated weights were overlaid on *Proteobacteria* branch of the phylogenetic tree. One hundred and fifty-five self-reported T2D was extracted from survey and 154 non-T2D participants were randomly sampled as controls. The red and blue nodes indicate taxa have positive and negative weights, respectively. The size of colored nodes represent the magnitudes of the weights. Black nodes represent non-selected taxa.

Figure 5: DeepBiome selected T2D associated taxa from all phylums overlaid on the phylogenetic tree. Microbiome and phenotype data were downloaded from American Gut Project (AGP). One hundred and fifty-five self-reported T2D was extracted from survey. A random selected 154 non-T2D were then sampled to serve as controls. The blue and red nodes indicate taxa have negative and positive weights, respectively. The size of colored nodes represent the magnitudes of the weights. Black nodes represent non-selected taxa.

(a) Lasso

(b) Elastic-Net

(c) ℓ_1 -DNN

(d) DeepBiome

Figure 6: **Selection of FEV1% predicted associated microbiome taxa using Lasso, Elastic-Net, ℓ_1 -DNN, and DeepBiome.** The red nodes indicate taxa have positive association with FEV1% predicted and blue nodes indicate taxa have negative association with FEV1% predicted.

Table 1: Metrics used to assess the performance of outcome prediction and microbiome taxa selection.

	Prediction			Selection
	Regression	Binary	Multiclass	
Metrics	MSE	Sensitivity	Sensitivity	Sensitivity
	Pearson's ρ	Specificity	PPV	Specificity
		g-Measure	F1 score	g-Measure
		ACC	ACC	ACC
		AUC	AUC	

MSE: mean squared error; PPV: positive predictive value; ACC: accuracy; AUC: area under the receiver operating characteristic (ROC) curve.

Table 2: Scenario 1. Mean squared error (MSE) and Pearson correlation coefficient between predicted and true outcomes for continuous outcome. The associated taxa are clustered at the phylum level.

Method	Testing				Training			
	MSE		Correlation		MSE		Correlation	
	mean	sd	mean	sd	mean	sd	mean	sd
Linear Regression	0.104	0.024	0.824	0.049	0.087	0.011	0.851	0.023
Ridge	0.104	0.022	0.824	0.049	0.09	0.012	0.851	0.023
Lasso	0.100	0.023	0.833	0.048	0.092	0.013	0.843	0.025
Elastic Net	0.100	0.023	0.833	0.048	0.092	0.012	0.844	0.025
DNN	0.076	0.040	0.874	0.077	0.032	0.034	0.947	0.067
DNN+ ℓ_1	0.075	0.040	0.875	0.073	0.034	0.039	0.945	0.068
DeepBiome	0.071	0.036	0.882	0.069	0.043	0.034	0.929	0.061

DNN: Deep neural network; DNN- ℓ_1 : Lasso (least absolute shrinkage and selection operator) penalized deep neural network.

Table 3: Scenario 1. Mean squared error (MSE) and Pearson correlation coefficient between predicted and true outcomes for continuous outcome. The associated taxa are clustered at the phylum and order levels.

Method	Testing				Training			
	MSE		Correlation		MSE		Correlation	
	mean	sd	mean	sd	mean	sd	mean	sd
Linear Regression	1.561	0.146	0.639	0.035	1.337	0.068	0.694	0.018
Lasso	1.479	0.115	0.662	0.034	1.411	0.075	0.678	0.02
Ridge	1.546	0.121	0.639	0.034	1.361	0.075	0.694	0.018
Elastic-net	1.481	0.117	0.662	0.034	1.405	0.076	0.68	0.02
DNN	0.457	0.522	0.905	0.118	0.164	0.337	0.964	0.091
DNN+ ℓ_1	0.456	0.516	0.904	0.122	0.176	0.362	0.963	0.085
DeepBiome	0.423	1.474	0.916	0.139	0.256	0.463	0.944	0.110

DNN: Deep neural network; DNN- ℓ_1 : Lasso (least absolute shrinkage and selection operator) penalized deep neural network.

Table 4: Scenario 2. Classification performance for binary outcomes.

Method	Testing										Training									
	Sensitivity		Specificity		g-Measure		ACC		AUC		Sensitivity		Specificity		g-Measure		ACC		AUC	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
Logistic	0.92	0.03	0.725	0.066	0.815	0.038	0.86	0.026	0.822	0.033	0.95	0.021	0.785	0.049	0.863	0.03	0.9	0.02	0.867	0.026
Lasso	0.965	0.016	0.583	0.086	0.747	0.056	0.848	0.027	0.774	0.041	0.973	0.006	0.619	0.081	0.774	0.053	0.866	0.023	0.796	0.04
Ridge	0.957	0.018	0.461	0.077	0.661	0.055	0.805	0.027	0.709	0.037	0.97	0.007	0.522	0.062	0.711	0.043	0.835	0.016	0.746	0.03
ElasticNet	0.998	0.004	0.022	0.02	0.121	0.082	0.699	0.03	0.51	0.01	0.998	0.002	0.022	0.013	0.14	0.046	0.702	0.015	0.51	0.006
DNN	0.887	0.049	0.725	0.148	0.788	0.133	0.837	0.038	0.897	0.039	0.932	0.032	0.869	0.158	0.887	0.146	0.913	0.043	0.958	0.033
DNN+ ℓ_1	0.887	0.048	0.727	0.146	0.790	0.129	0.838	0.038	0.898	0.036	0.931	0.031	0.868	0.154	0.887	0.141	0.912	0.042	0.958	0.030
DeepBiome	0.918	0.042	0.835	0.111	0.870	0.093	0.892	0.044	0.941	0.051	0.952	0.033	0.922	0.106	0.932	0.094	0.943	0.041	0.974	0.048

ACC: accuracy; AUC: area under the receiver operating characteristic (ROC) curve; DNN: deep neural network; DNN- ℓ_1 : Lasso (least absolute shrinkage and selection operator) penalized deep neural network.

Table 5: Scenario 3. Classification performance for multi-categorical outcomes.

Method	Testing										Training									
	ACC		Precision		Recall		F1		AUC		ACC		Precision		Recall		F1		AUC	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
SVM-Linear	0.667	0.032	0.525	0.06	0.493	0.027	0.508	0.039	0.667	0.027	0.694	0.017	0.571	0.034	0.52	0.019	0.544	0.024	0.679	0.018
SVM-Radial	0.659	0.033	0.527	0.092	0.506	0.024	0.514	0.05	0.665	0.027	0.821	0.012	0.807	0.035	0.663	0.017	0.728	0.02	0.809	0.015
SVM-Polynomial	0.576	0.034	0.443	0.09	0.394	0.026	0.414	0.048	0.566	0.03	0.767	0.014	0.837	0.012	0.622	0.025	0.714	0.019	0.742	0.026
DNN	0.752	0.045	0.620	0.136	0.599	0.105	0.599	0.115	0.856	0.041	0.854	0.063	0.741	0.188	0.718	0.149	0.718	0.166	0.941	0.044
DNN+ ℓ_1	0.760	0.042	0.668	0.139	0.612	0.090	0.618	0.104	0.863	0.043	0.864	0.058	0.794	0.160	0.731	0.130	0.739	0.146	0.948	0.037
DeepBiome	0.815	0.066	0.720	0.151	0.714	0.115	0.711	0.135	0.900	0.080	0.880	0.072	0.804	0.170	0.793	0.135	0.791	0.157	0.942	0.084

ACC: accuracy; AUC: area under the receiver operating characteristic (ROC) curve; DNN: Deep neural network; DNN- ℓ_1 : Lasso (least absolute shrinkage and selection operator) penalized deep neural network; SVM: support vector machine.

Table 6: Scenario 4. Mean squared error (MSE) and Pearson correlation coefficient between predicted and true outcome (continuous), when the input microbiome abundance data contain measurement errors. The associated taxa are clustered at the phylum and order levels.

Method	Testing				Training			
	MSE		Correlation		MSE		Correlation	
	mean	sd	mean	sd	mean	sd	mean	sd
Linear Regression	1.569	0.154	0.639	0.036	1.336	0.066	0.694	0.018
Ridge	1.551	0.128	0.639	0.036	1.358	0.073	0.694	0.018
Lasso	1.488	0.119	0.661	0.034	1.408	0.073	0.679	0.02
Elastic-net	1.49	0.121	0.66	0.034	1.402	0.075	0.681	0.019
DNN	0.619	0.682	0.873	0.137	0.188	0.317	0.961	0.068
DNN+ ℓ_1	0.445	0.351	0.909	0.081	0.129	0.234	0.974	0.050
DeepBiome	0.243	0.400	0.950	0.087	0.117	0.244	0.976	0.052

DNN: Deep neural network; DNN- ℓ_1 : Lasso (least absolute shrinkage and selection operator) penalized deep neural network.

Table 7: Scenario 4. Mean squared error (MSE) and Pearson correlation coefficient between predicted and true outcome (continuous), when using an mis-specified phylogenetic tree. The associated taxa are clustered at the phylum and order levels.

Method	Testing				Training			
	MSE		Correlation		MSE		Correlation	
	mean	sd	mean	sd	mean	sd	mean	sd
Linear Regression	0.872	0.163	0.683	0.046	0.737	0.074	0.726	0.027
Lasso	0.826	0.144	0.706	0.047	0.78	0.08	0.71	0.03
Ridge	0.866	0.138	0.683	0.046	0.752	0.081	0.726	0.027
Elastic-net	0.826	0.144	0.706	0.047	0.779	0.079	0.711	0.028
DNN	0.437	0.208	0.849	0.077	0.167	0.166	0.944	0.058
DNN+ ℓ_1	0.434	0.214	0.850	0.080	0.166	0.171	0.944	0.065
DeepBiome	0.316	0.261	0.892	0.094	0.195	0.207	0.933	0.075

DNN: Deep neural network; DNN- ℓ_1 : Lasso (least absolute shrinkage and selection operator) penalized deep neural network.

Table 8: Type 2 Diabetes (T2D) prediction performance using microbiome composition in the American Gut Project (AGP).

Method	Testing										Training									
	Sensitivity		Specificity		g-Measure		ACC		AUC		Sensitivity		Specificity		g-Measure		ACC		AUC	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
Logistic	0.499	0.117	0.414	0.074	0.448	0.047	0.464	0.045	0.513	0.065	0.952	0.087	0.937	0.127	0.944	0.107	0.944	0.107	0.945	0.107
Lasso	0.712	0.423	0.386	0.416	0.238	0.326	0.531	0.116	0.549	0.072	0.726	0.42	0.365	0.41	0.264	0.282	0.549	0.052	0.545	0.062
ridge	0.369	0.507	0.669	0.453	0.16	0.219	0.477	0.077	0.519	0.033	0.398	0.546	0.691	0.424	0.19	0.261	0.555	0.062	0.545	0.061
ElasticNet	0.956	0.071	0.067	0.113	0.146	0.213	0.501	0.087	0.511	0.021	0.956	0.022	0.067	0.164	0.202	0.258	0.544	0.06	0.542	0.072
DNN	0.584	0.037	0.501	0.044	0.54	0.036	0.575	0.029	0.604	0.028	0.915	0.013	0.899	0.016	0.907	0.013	0.915	0.013	0.941	0.014
DNN+ ℓ_1	0.578	0.052	0.555	0.08	0.566	0.061	0.577	0.05	0.604	0.052	0.92	0.021	0.907	0.028	0.913	0.023	0.919	0.022	0.942	0.02
DeepBiome	0.653	0.057	0.659	0.082	0.653	0.03	0.641	0.167	0.655	0.079	0.643	0.026	0.656	0.057	0.649	0.021	0.629	0.134	0.647	0.067

ACC: accuracy; AUC: area under the receiver operating characteristic (ROC) curve; DNN: Deep neural network; DNN- ℓ_1 : Lasso (least absolute shrinkage and selection operator) penalized deep neural network.

Table 9: Body Mass Index (BMI, continuous) prediction using microbiome composition in the American Gut Project (AGP).

Method	Testing				Training			
	MSE		Correlation		MSE		Correlation	
	mean	sd	mean	sd	mean	sd	mean	sd
Linear Regression	5.375	5.597	0.071	0.054	0.896	0.049	0.323	0.008
Lasso	1.036	0.281	0.140	0.049	0.974	0.062	0.180	0.038
Ridge	1.176	0.531	0.027	0.038	0.995	0.053	0.107	0.099
Elastic Net	0.996	0.198	0.067	0.005	0.990	0.053	0.145	0.018
DNN	1.005	0.206	0.175	0.070	0.903	0.038	0.314	0.021
DNN+ ℓ_1	1.032	0.215	0.180	0.074	0.907	0.046	0.299	0.083
DeepBiome	1.011	0.186	0.200	0.017	1.010	0.142	0.201	0.016

DNN: Deep neural network; DNN- ℓ_1 : Lasso (least absolute shrinkage and selection operator) penalized deep neural network.

Table 10: Body Mass Index (BMI, multi-categorical) classification using microbiome composition in the American Gut Project (AGP).

Method	Testing										Training									
	Sensitivity		Specificity		g-Measure		ACC		AUC		Sensitivity		Specificity		g-Measure		ACC		AUC	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
SVM-Linear	0.549	0.016	0.28	0.025	0.249	0.01	0.263	0.011	0.532	0.016	0.576	0.008	0.53	0.033	0.284	0.003	0.37	0.006	0.548	0.007
SVM-Radial	0.475	0.011	0.332	0.074	0.248	0.013	0.281	0.033	0.558	0.015	0.579	0.007	0.613	0.005	0.326	0.004	0.426	0.005	0.613	0.007
SVM-Polynomial	0.578	0.01	0.273	0.042	0.252	0.008	0.261	0.023	0.506	0.011	0.643	0.004	0.731	0.008	0.34	0.002	0.464	0.003	0.539	0.004
DNN	0.593	0.01	0.291	0.121	0.224	0.01	0.196	0.011	0.623	0.033	0.609	0.009	0.411	0.1	0.233	0.019	0.21	0.022	0.656	0.022
DNN+ ℓ_1	0.596	0.01	0.37	0.116	0.212	0.01	0.194	0.011	0.623	0.03	0.609	0.003	0.456	0.147	0.221	0.01	0.207	0.011	0.658	0.022
DeepBiome	0.600	0.011	0.231	0.112	0.225	0.003	0.185	0.003	0.632	0.008	0.600	0.002	0.239	0.078	0.225	0.002	0.185	0.001	0.634	0.004

ACC: accuracy; AUC: area under the receiver operating characteristic (ROC) curve; DNN: deep neural network; DNN- ℓ_1 : lasso (least absolute shrinkage and selection operator) penalized deep neural network; SVM: support vector machine.

Table 11: Predicting FEV1_{%predicted} from lung microbiome composition.

Method	Testing		Training	
	MSE		MSE	
	mean	sd	mean	sd
Ridge	0.916	0.373	0.933	0.011
Lasso	0.452	0.313	0.446	0.046
ElasticNet	0.690	0.606	0.602	0.414
DNN	28.122	66.843	16.544	31.448
DNN+ ℓ_1	29.662	75.867	16.034	25.984
DeepBiome	2.105	4.526	2.881	5.28

DNN: Deep neural network; DNN- ℓ_1 : Lasso (least absolute shrinkage and selection operator) penalized deep neural network.

Figures

(a) An example phylogenetic tree with 48 genera (b) Network layout of DeepBiome architecture

(c) Phylogenetic tree regularized weight decay procedure

Figure 1

DeepBiome architecture. (a) A phylogenetic tree with 48 genera as tip nodes. Color represents phylum types. (b) Network layout of DeepBiome architecture. The input layer is genus level microbiome abundance. Each hidden layer represents one phylogenetic level, e.g., family, order, class, and phylum. The dark lines represent relationship defined by a phylogenetics tree. The gray lines represent association between layers. (c) Phylogenetic tree regularized weight decay. Suppose we have a simple tree as shown in the left panel, which has 6 genera (taxa 1-6) and 2 classes (7-8). Genera 1-3 belong to class 7 and genera 4-6 belong to class 8. The ancestor-descendent information is embedded into a 6×2 matrix. Without loss of generality, we use ω to indicate a regularization factor with small value (e.g., 0.01). For

tree regularized weight decay, the weight estimation matrix $w_{6 \times 2}$ is multiplied with this phylogenetic embedded matrix $\Omega_{6 \times 2}$ elementwisely, denoted by $\Omega_{6 \times 2} \otimes w_{6 \times 2}$.

Figure 2

Simulation specifications. The outcome associated taxa (blue and red) are specified at (a) the phylum level and (b) a mixture of phylum & order levels. The blue nodes represent “bad” taxa which result in a disease status or are negatively associated with continuous phenotype, e.g., FEV1. The red nodes represent “good” taxa which result in a healthy status. In simulation scenario 4, we evaluate the impact of mis-specified phylogenetic tree: (c) indicates the true phylogenetic tree used in simulation scenario 4 and (d) indicates the phylogenetic tree used in model learning (same as the tree shown in Figure (b)).

Figure 3

Taxa selection performance under 4 simulation schemes at each phylogenetic level. Sensitivity, specificity, g-measure, and accuracy (ACC) were used to evaluate taxa selection performance. Vertical bar represents standard deviation over 1000 simulation replicates.

Figure 4

DeepBiome selected T2D associated taxa from phylum Proteobacteria among the American Gut Project (AGP). Estimated weights were overlaid on Proteobacteria branch of the phylogenetic tree. One hundred and fifty-five self-reported T2D was extracted from survey and 154 non-T2D participants were randomly sampled as controls. The red and blue nodes indicate tax3a2have positive and negative weights,

controls. The blue and red nodes indicate taxa have negative and positive weights, respectively. The size of colored nodes represent the magnitudes of the weights. Black nodes represent non-selected taxa.

Figure 6

Selection of FEV1% predicted associated microbiome taxa using Lasso, Elastic-Net, ℓ_1 -DNN, and DeepBiome. The red nodes indicate taxa have positive association with FEV1% predicted and blue nodes indicate taxa have negative association with FEV1% predicted.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [DeepMicrobiomeSupplementaryMaterials.pdf](#)