Supplemental Materials: Phylogeny-aware data augmentation for enhancing machine learning applied to microbiome data

A Supplementary Lemmas and proofs

A.1 Main Lemmas

Lemma S1. *Consider the phylogeny $\mathcal{T}$ on the OTU set $\mathcal{S}$. For each internal node u of $\mathcal{T}$, let C^u be random variable giving the number of observed sequences contained under the node u and let c^u be the observed value of C^u for one sample. Assume $C^{l(u)} \sim Bin(p_u, C^u)$ where where $l(u)$ is the left child of u and let $C^{r(u)} = C^u - C^{l(u)}$ where $r(u)$ is the right child of u. The joint maximum likelihood estimate of all p_u values is given by*

$$(\hat{p}_1, \dots, \hat{p}_u, \dots, \hat{p}_{n-1}) = (\frac{c^{l(1)}}{c^1}, \dots, \frac{c^{l(u)}}{c^u}, \dots \frac{c^{l(n-1)}}{c^{n-1}}) \ . \tag{1}$$

Proof. Consider a sample with the count vector $X = (x_1, x_2, \dots, x_n)$ at the leaves of the tree. Recall the root is indexed 1 and thus $\sum_{i=1}^n x_i = c^1$. Let $path_v^u$ indicate the path from the leaf node v to the node u. The likelihood of observing the count vector $x = (x_1, x_2, \dots, x_n)$ given the phylogeny $\mathcal{T}$, and the conditional probability vector $p = (p_1, \dots p_u, \dots, p_{n-1})$ equals

$$\mathcal{L}(X = (x_1, x_2 \dots, x_n); \mathcal{P} = \{(p_1, \dots, p_u \dots, p_{n-1}, p_n, \dots, p_{u+n-1}, \dots, p_{2n-2}, \mathcal{T}) = \Gamma(X, c^1) \prod_{i=1}^{n} (\prod_{k \in path_i^{root}} p_k)^{x_i}$$

where $p_{n+e-1} = 1 - p_e$, and $\Gamma(X, c^1)$ is a normalization function that doesn't depend on the conditional probability vector p. Consider the left child of the root, $l(1)$, and recall the probability of sequences falling below it is p_1. All the root-to-leaf paths descending from $l(1)$ have the branch connecting the root to $l(1)$; thus, for these, the probability p_1 is multiplied each time. A similar argument works for the right child of the root with the probability $1 - p_1$. So the likelihood could be written as

$$\mathcal{L}(X; \mathcal{P}, \mathcal{T}) = \Gamma(X, c^1) [\prod_{i=1}^{n} (\prod_{k \in path_i^{lr(1,i)}} p_k)^{x_i}] . (p_1)^{(\sum_{i \in a(l(1))} x_i)} . (1 - p_1)^{(\sum_{i \in a(r(1))} x_i)} =$$

$$\mathcal{L}(x; \mathcal{P}, \mathcal{T}^l) . \mathcal{L}(x; \mathcal{P}, \mathcal{T}^r) . (p_1)^{c^{l(1)}} . (1 - p_1)^{c^1 - c^{l(1)}}$$

where $lr(1, i) = l(1)$ for leaves under the left child of 1 and $lr(1, i) = r(1)$ for leaves under the right child of 1, $a(u)$ is the set of leaves under the node u, and $\mathcal{P}^l$ and $\mathcal{P}^r$ indicate the left and right subtrees of the root. This means we could compute the $\mathcal{L}(x; \mathcal{P}, \mathcal{T})$ as the product of the likelihood of the left subtree, the likelihood of the right subtree of the root, and the probability of observing a total of $c^{l(1)}$ counts for the $\mathcal{T}^l$.

Note that p_1 (same is true for $1 - p_1$) does not contribute to the likelihood of $\mathcal{T}^l$ or $\mathcal{T}^r$ and, hence, to find p_1 we could consider $\mathcal{L}(x; \mathcal{P}, \mathcal{T}^r) . \mathcal{L}(x; \mathcal{P}, \mathcal{T}^l)$ as a constant and ignore. Instead of maximizing the $(p_1)^{c^{l(1)}} . (1 - p_1)^{c^1 - c^{l(1)}}$, we could maximize log of this function

$$f(c^{l(1)}, c^1; p_1) = c^{l(1)} \log(p_1) + (c^1 - c^{l(1)}) \log(1 - p_1)$$

and, hence

$$\hat{p}_1 = \frac{c^{l(1)}}{c^1} .$$

This gives us $\hat{p}_1$ for the edges descending from the root. We can use the same argument recursively and compute other probabilities as

$$\hat{p}_u = \frac{c^{l(u)}}{c^u} \ .$$

Lemma S2. *Consider the phylogeny $\mathcal{T}$ on the OTU set $\mathcal{S}$ and samples $s_1, \ldots, s_w$. Let the total number of sequences in each sample be $\mathcal{C} = \{c_1^1, c_2^1 \ldots, c_w^1\}$. Assume that the probability of observing a sequence from a species under the left subtree of the node u follows a beta distribution $p_u^l \sim Beta(\mu_u, \nu_u)$ where ν_u is a fixed parameter which depends only on the phylogeny, $\mathcal{T}$, and is therefore given, and $\mu_u \in \mathcal{M}$ is a parameter shared between all samples. Assume that the number of observed sequences contained under the left subtree of u given p_u^l follows a binomial distribution $C^{l(u)} \sim Bin(p_u^l, c^u)$. Then, the method of moments estimate for μ_u is*

$$\mu_u = \frac{\sum_{j=1}^{w} c_j^{l(u)}}{\sum_{j=1}^{w} c_j^u} \tag{2}$$

where $l(u)$ is the left subtree of u, and c_j^u is the number of observed sequences contained under u in the sample s_j.

Proof. Consider the new random variable $\sum_{j=1}^{w} C_j^{l(u)}$

$$\mathbb{E}[\sum_{j=1}^{w} C_j^{l(u)}] = \sum_{j=1}^{w} \mathbb{E}[C_j^{l(u)}] = \sum_{j=1}^{w} \mathbb{E}_{p_u^l}[\mathbb{E}[C_j^{l(u)} | p_u^l]] =$$
$$\sum_{j=1}^{w} \mathbb{E}_{p_u^l}[c_j^u p_u^l] = \mu_u \sum_{j=1}^{w} c_j^u$$

and hence having

$$\mu_u = \frac{\sum_{j=1}^{w} c_j^{l(u)}}{\sum_{j=1}^{w} c_j^u}$$

A.2 Using a Beta-Binomial distribution with two parameters

Following lemma S2, instead of using a ν_u which only depends on the phylogeny $\mathcal{T}$, we could use a model where (μ_u, ν_u) both depend only on class/cluster label y. In this model for each internal node u of the phylogeny, the probability of observing a species on the left subtree of u follows a beta distribution, (i.e. $p_u^l \sim Beta(\mu_u, \nu_u)$), and the number of observed sequences contained under the left subtree of node u follows a binomial distribution (i.e. $C^{l(u)} \sim Bin(p_u, c^u)$, where c^u is the total number of sequences under the node u). In this section, for the ease of calculation, we use the other formulation of the beta distribution, i.e. $p_u^l \sim Beta(\alpha_u, \beta_u)$, where $\mu_u = \frac{\alpha_u}{\alpha_u+\beta_u}$, and the relationship between ν_u, μ_u, and the variance of the beta distribution is given in Section 2.2.1; and hence these notations are interchangeable. Now, using method of moments, we could estimate α_u and β_u from the class-y samples as follow

$$\hat{\alpha}_u^{(MOM)} = \frac{c_u^{(2)}\mathcal{M}_u - \mathcal{Q}_u c_u^{(1)}}{\frac{\mathcal{Q}_u}{\mathcal{M}_u}(c_u^{(1)})^2 - \mathcal{M}_u c_u^{(2)} + \mathcal{M}_u c_u^{(1)} - (c_u^{(1)})^2} \tag{3}$$

$$\hat{\beta}_u^{(MOM)} = \frac{(\frac{\mathcal{Q}_u}{\mathcal{M}_u}c_u^{(1)} - c_u^{(2)})(c_u^{(1)} - \mathcal{M}_u)}{(c_u^{(1)})^2 - \mathcal{M}_u c_u^{(1)} + \mathcal{M}_u c_u^{(2)} - \frac{\mathcal{Q}_u}{\mathcal{M}_u}(c_u^{(1)})^2} \tag{4}$$

$$\mathcal{M}_u = \frac{1}{w}\sum_{i=1}^{w} c_i^{l(u)} \tag{5}$$

$$\mathcal{Q}_u = \frac{1}{w}\sum_{i=1}^{w}(c_i^{l(u)})^2 \tag{6}$$

$$c_u^{(1)} = \frac{1}{w}\sum_{i=1}^{w} c_i^u \tag{7}$$

$$c_u^{(2)} = \frac{1}{w}\sum_{i=1}^{w} (c_i^u)^2 \tag{8}$$

where $\mathcal{M}_u$ and $\mathcal{Q}_u$ are the first and second moments of the observed sequences contained under the left child of the node u in $\mathcal{T}$. In a special case where the total number of observed sequences contained under the node u are all the same (i.e. $c^u = c_1^u = c_2^u = \ldots = c_w^u$), the distribution becomes the beta-binomial distribution, $c^u = c_u^{(1)}$, and $(c^u)^2 = c_u^{(2)}$. In this case, the $\alpha^{(MOM)}$ and $\beta^{(MOM)}$ equal (Tripathi *et al.*, 1994)

$$\hat{\alpha}_u^{(MOM)} = \frac{c^u\mathcal{M}_u - \mathcal{Q}_u}{c^u(\frac{\mathcal{Q}_u}{\mathcal{M}_u} - \mathcal{M}_u - 1) + \mathcal{M}_u} \tag{9}$$

$$\hat{\beta}_u^{(MOM)} = \frac{(c^u - \mathcal{M}_u)(c^u - \frac{S}{\mathcal{M}_u})}{c^u(\frac{\mathcal{Q}_u}{\mathcal{M}_u} - \mathcal{M}_u - 1) + \mathcal{M}_u} \tag{10}$$

Proof. Consider the internal node u of the phylogeny $\mathcal{T}$. Based on our model, $C_i^{l(u)} \sim Bin(c_i^u, p_u^l)$, where $p_u^l \sim Beta(\alpha_u, \beta_u)$

$$p_u^l \sim Beta(\alpha_u, \beta_u) \tag{11}$$

$$C_i^{l(u)} \sim Bin(c_i^u, p_u) \tag{12}$$

We will compute the expected value for the weighted average of random variables $C_i^{l(u)}$s.

$$\mathbb{E}[\frac{1}{w}\sum_{i=1}^{w} C_i^{l(u)}] = \frac{1}{w}\sum_{i=1}^{w}\mathbb{E}[C^{l(u)}] = \frac{1}{w}\sum_{i=1}^{w}\mathbb{E}[\mathbb{E}[C^{l(u)}|p_u^l]] =$$

$$\frac{1}{w}\sum_{i=1}^{w} c_i^u\mathbb{E}[p_u^l] = \frac{\alpha_u}{\alpha_u+\beta_u}\frac{1}{w}\sum_{i=1}^{w} c^u = \frac{\alpha_u}{\alpha_u+\beta_u}c_u^{(1)}$$

We can name the empirical mean of $C^{l(u)}$s', as $\mathcal{M}_u = \frac{1}{w}\sum_{i=1}^{w} c_i^{l(u)}$. Next we write the expected value of $(C_i^{l(u)})^2$s'

$$\mathbb{E}[\frac{1}{w}\sum_{i=1}^{w}(C_i^{l(u)})^2] = \frac{1}{w}\sum_{i=1}^{w}\mathbb{E}[(C_i^{l(u)})^2] = \frac{1}{w}\sum_{i=1}^{w}\mathbb{E}[\mathbb{E}[(C_i^{l(u)})^2|p_u^l]]$$

$$\mathbb{E}[(C_i^{l(u)})^2|p_u^l] = c_i^u p_u^l(1 + (c_i^u - 1)p_u^l)$$

$$\mathbb{E}[\mathbb{E}[(C_i^{l(u)})^2|p_u^l]] = c_i^u\mathbb{E}[p_u^l] + c_i^u(c_i^u - 1)\mathbb{E}[(p_u^l)^2] =$$

$$c_i^u\frac{\alpha_u}{\alpha_u+\beta_u} + c_i^u(c_i^u - 1)(\frac{\alpha_u\beta_u}{(\alpha_u+\beta_u)^2(\alpha_u+\beta_u+1)} + \frac{\alpha_u^2}{(\alpha_u+\beta_u)^2}) = c_i^u\alpha_u(\frac{c_i^u(\alpha_u+1)+\beta_u}{(\alpha_u+\beta_u)(\alpha_u+\beta_u+1)})$$

hence

$$\mathbb{E}[\frac{1}{w}\sum_{i=1}^{w}(C_i^{l(u)})^2] = \sum_{i=1}^{w}\frac{1}{w}(c_i^u\alpha_u(\frac{c_i^u(\alpha_u+1)+\beta_u}{(\alpha_u+\beta_u)(\alpha_u+\beta_u+1)})) = \tag{13}$$

$$\frac{\alpha_u}{(\alpha_u+\beta_u)(\alpha_u+\beta_u+1)}((\alpha_u+1)\sum_{i=1}^{w}\frac{(c_i^u)^2}{w}+\beta_u\sum_{i=1}^{w}\frac{c_i^u}{w}) = \frac{\alpha_u}{(\alpha_u+\beta_u)(\alpha_u+\beta_u+1)}((\alpha_u+1)c_u^{(2)}+\beta_u c_u^{(1)}) \tag{14}$$

where we call the empirical value for the $\mathbb{E}[\frac{1}{w}\sum_{i=1}^{w} C_i^{l(u)}]$ as $\mathcal{Q}_u = \sum_{i=1}^{w}\frac{1}{w}(c_i^{l(u)})^2$. Using the method of moments and using the following equations we could compute the estimates for α and β

$$\mathcal{M}_u = \frac{\alpha_u}{\alpha_u+\beta_u}c_u^{(1)} \tag{15}$$

$$\mathcal{Q}_u = \frac{\alpha_u}{(\alpha_u+\beta_u)(\alpha_u+\beta_u+1)}((\alpha_u+1)c_u^{(2)}+\beta_u c_u^{(1)}) = \frac{\mathcal{M}_u c_u^{(2)}}{c_u^{(1)}} + (c_u^{(1)}-c_u^{(2)})\frac{\mathcal{M}_u}{c_u^{(1)}}\frac{\beta_u}{\beta_u+\alpha_u+1} \tag{16}$$

$$c_u^{(1)} = \frac{1}{w}\sum_{i=1}^{w}c_i^u \tag{17}$$

$$c_u^{(2)} = \frac{1}{w}\sum_{i=1}^{w}(c_i^u)^2 \tag{18}$$

In order to evaluate the performance of these estimators, we used simulations, where μ's and counts (c^u's) are generated from known beta and binomial distributions respectively (Figure S1). Figure S4 shows results of these estimators on the IBD dataset.

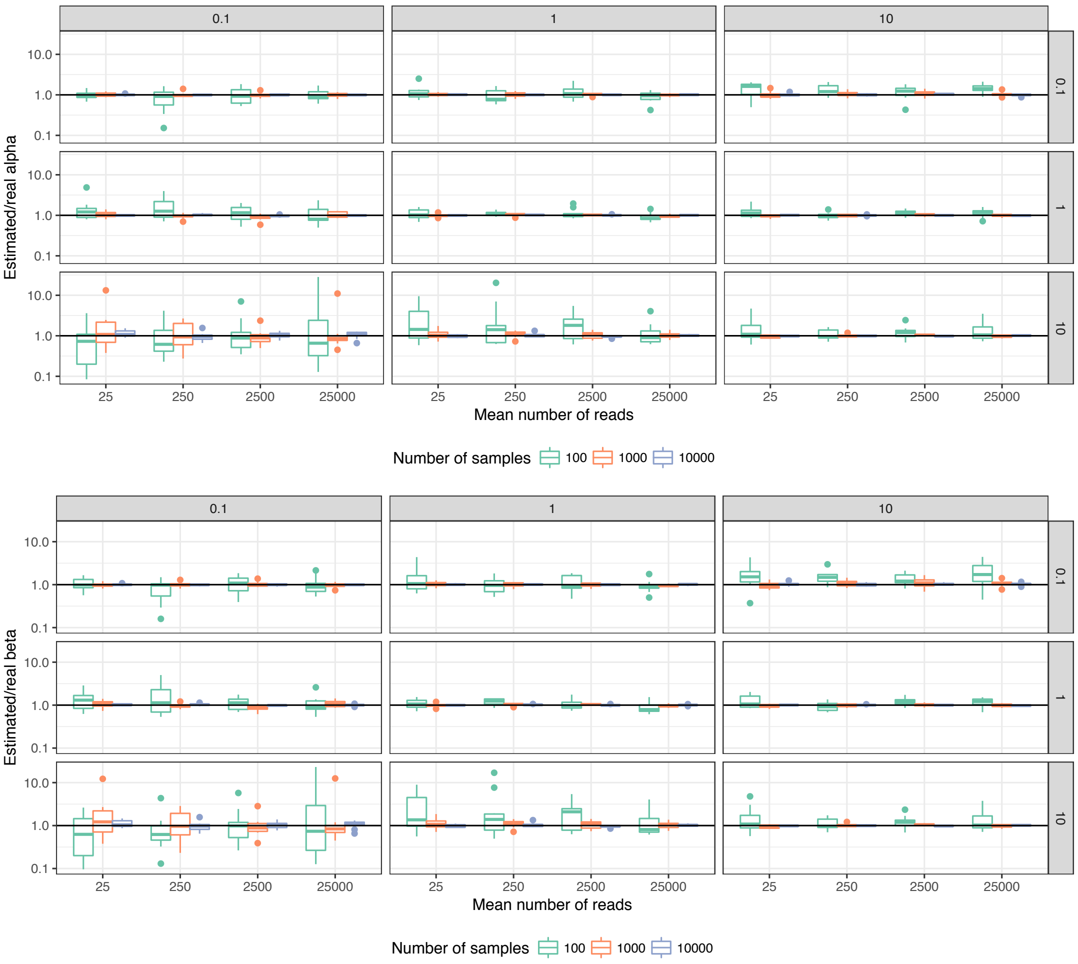

Figure S1. The α estimation error using the method of moments in simulations. 10,000 points are drawn from the hierarchical model and accuracy of the method of moments estimator is shown. The x-axis shows the average number of reads, the y-axis shows the ratio between estimated and the real α (top) and β (bottom). Each row corresponds to the true α values and each column corresponds to the true β values.

B Supplementary Figures and Algorithms

Algorithm S1 Algorithm to compute the average tip-to-tip distances of tree. Length of each node is defined as the length of the edge above it.

```
procedure Average_tip_to_tip($\mathcal{T}$)
    for $u \in$ postorder traversal of $\mathcal{T}$ do
        if $u$ is a leaf then
            $u$.num = 1
            $u$.avg = 0
            $u$.sum = 0
        else
            $v, w$ = left and right children of $u$
            $u$.num = $v$.num + $w$.num
            $u$.avg = ($w$.sum $\times$ $v$.num + $v$.sum $\times$ $w$.num)$/$($v$.num $\times w$.num) $+t_v + t_w$
            $u$.sum = $v$.sum + $w$.sum + $t_v$ $\times$ $v$.num + $t_w$ $\times$ $w$.num
    return $\mathcal{T}$
```

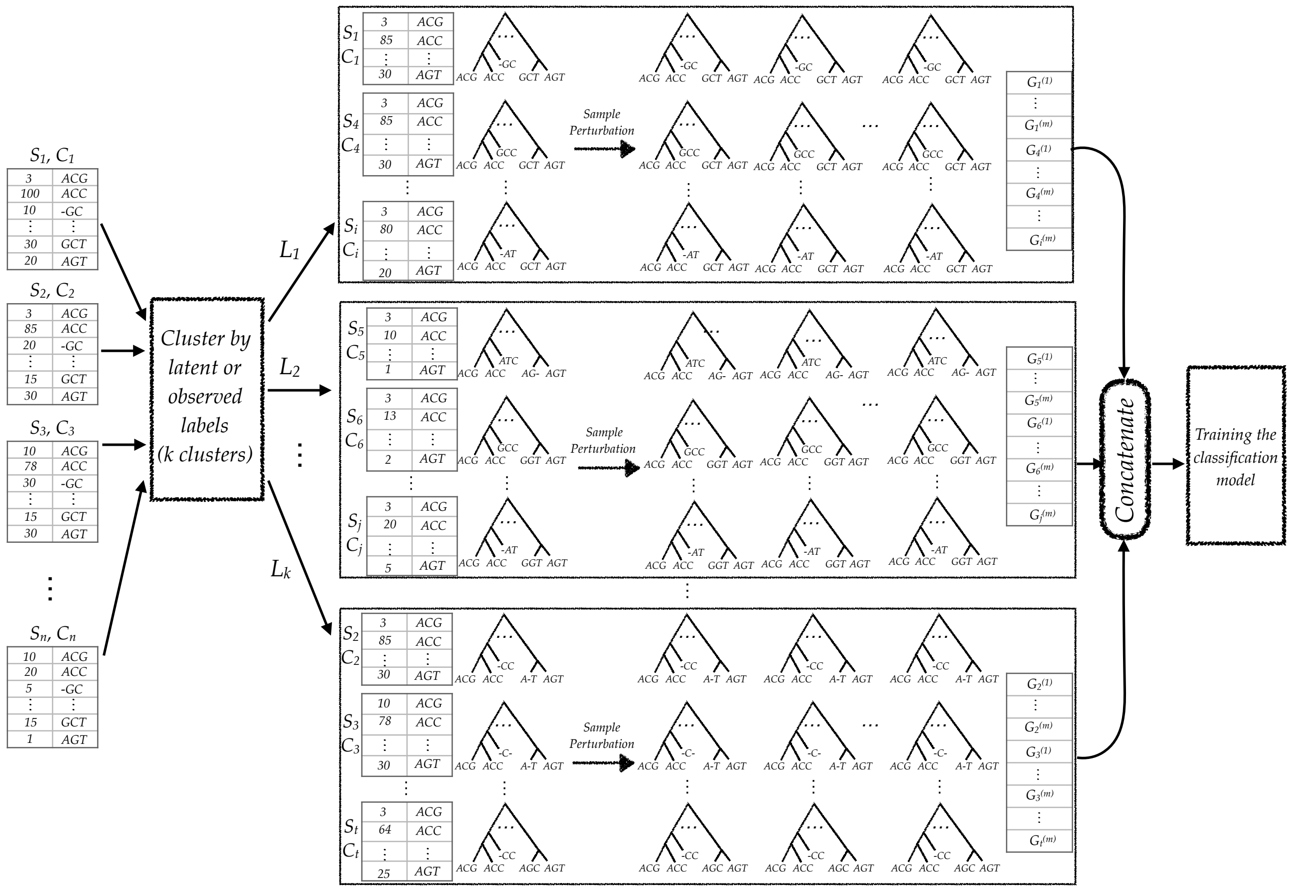

Figure S2. In the clustering scenario, we first group samples using k-means clustering applied to the Bray-curtis distances between samples. We then generate new samples for each group separately.

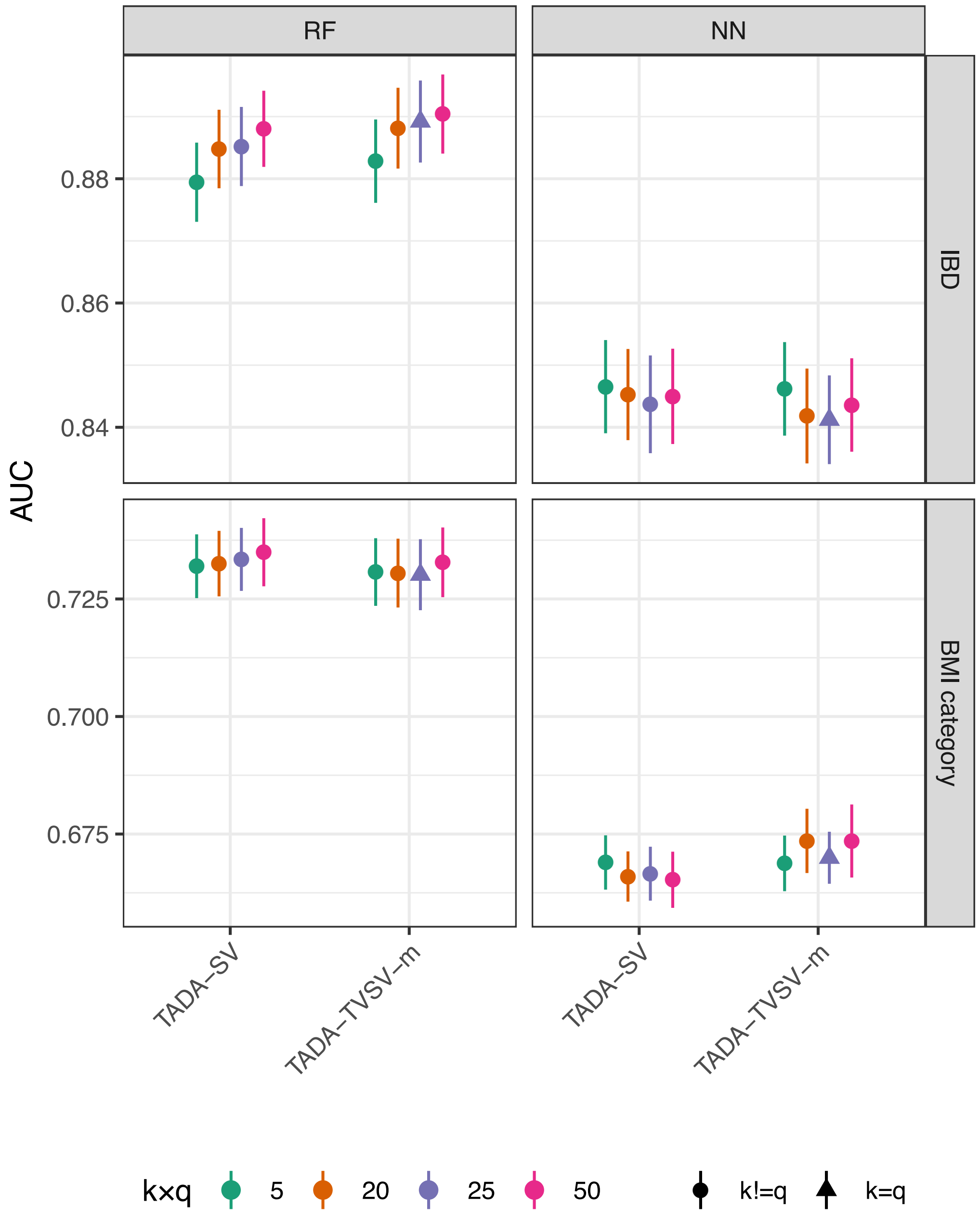

Figure S3. Impact of the choice of k **and** q. On E1, we show results with different augmentation levels. Area Under Curve (AUC) is shown for both Neural Networks(NN) and Random Forest (RF) classifiers and on both Gevers IBD dataset and AGP BMI dataset. Colors show $k \times q$ (set to 1, 5, 20, or 50). For SV, $q = 1$ in all cases. For TVSV, $k = 1$, except for the case where $k \times q = 25$, where, $k = q = 5$.

.

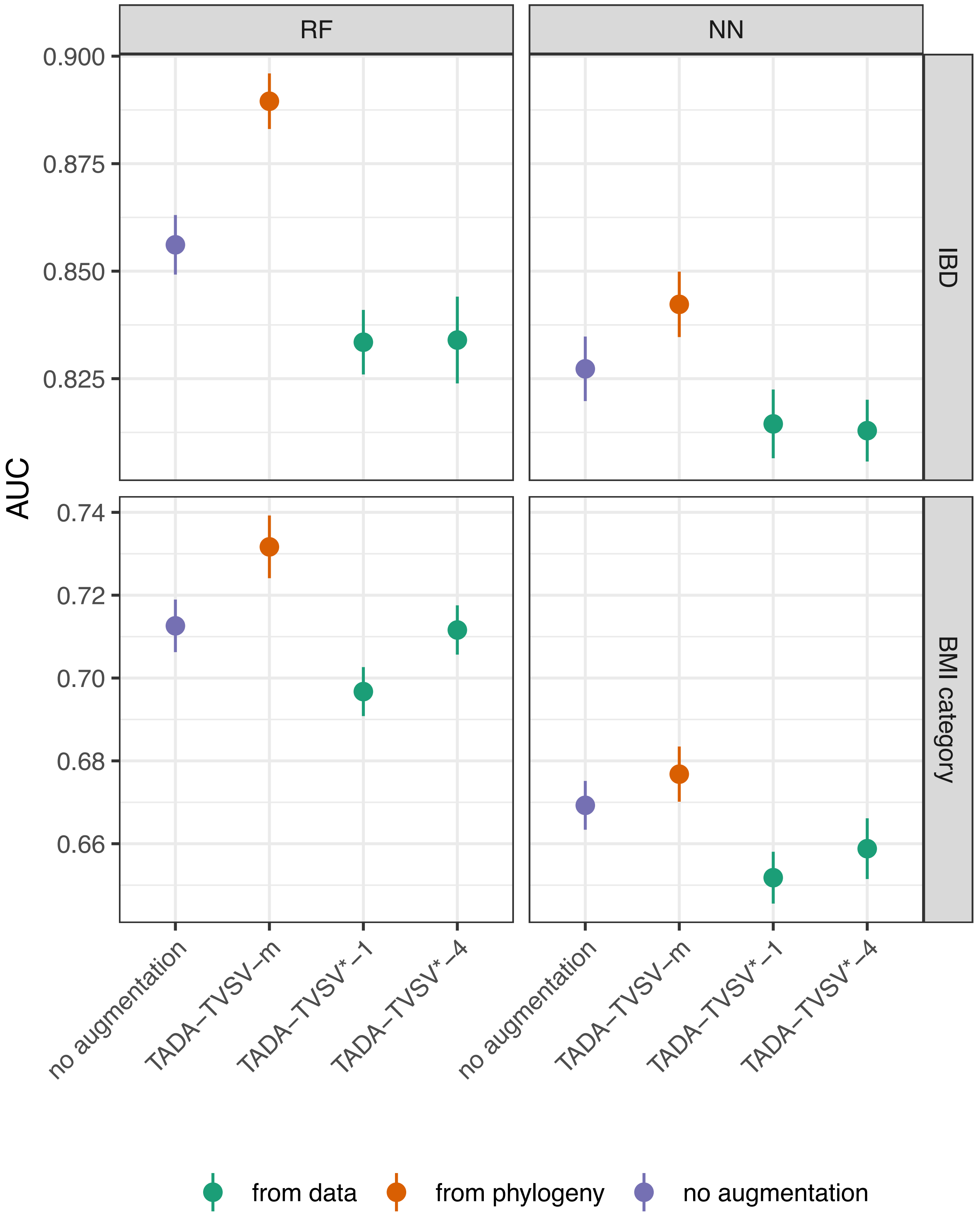

Figure S4. Computing ν from data versus fixing it using the phylogeny. Using the method of moments described in Section A.2, we estimate both μ_u and ν_u from data (instead of fixing ν_u from phylogeny as in our main results). We call the resulting method TADA-TVSV*. Results on the E1 dataset indicate that computing ν_u from variance of the data not only fails to improve accuracy, but can even reduce it.