Submitted to the Annals of Applied Statistics

PHYLOGENETICALLY INFORMED BAYESIAN TRUNCATED COPULA GRAPHICAL MODELS FOR MICROBIAL ASSOCIATION NETWORKS

BY HEE CHEOL CHUNG[*], IRINA GAYNANOVA[†] AND YANG NI[‡]

Department of Statistics, Texas A&M University, [*]*hcchung@stat.tamu.edu;* [†]*irinag@stat.tamu.edu;* [‡]*yni@stat.tamu.edu*

Microorganisms play critical roles in host health. The advancement of high-throughput sequencing technology provides opportunities for a deeper understanding of microbial interactions. However, due to the technological limitations of 16S ribosomal RNA sequencing, microbiome data are zero-inflated, and a quantitative comparison of microbial abundances cannot be made across subjects. By leveraging a recent microbiome profiling technique that quantifies 16S ribosomal RNA microbial counts, we propose a novel Bayesian graphical model that incorporates microorganisms' evolutionary history through a phylogenetic tree prior and explicitly accounts for zero-inflation using the truncated Gaussian copula. Our simulation study reveals that the evolutionary information substantially improves the network estimation accuracy. We apply the proposed model to the quantitative gut microbiome data of 106 healthy subjects, and identify three distinct microbial communities that are not found by existing microbial network estimation models. We further find that these communities are discriminated based on microorganisms' ability to utilize oxygen as an energy source.

1. Introduction.

1.1. *Microbial association network and its importance.* The gut is one of the most significant habitats of a myriad of microbial communities that play critical roles in their host's health. Well-balanced gut microbial communities provide many health benefits, such as maintaining metabolic homeostasis and high functioning immune system (Martinez, Pierre and Chang, 2016; Kim et al., 2016; Cani et al., 2019). The imbalance of the gut microbiome (dysbiosis) has been related to a variety of human diseases (Cho and Blaser, 2012; Lynch and Pedersen, 2016). The gut microbial balance is maintained by complex microbial interactions such as metabolites consumption, production, and exchange. Microbiome dysbiosis occurs when these interactions are interrupted by environmental alterations such as diet change, antibiotic consumption, and chemical exposure. These changes may deplete nutrients for beneficial microbes and create favorable surroundings for disease-causing bacteria to flourish. Nevertheless, some microbes help the microbial communities to maintain their stability under the environmental changes by providing energy sources and necessary metabolites (Zhang and Chen, 2019). Because of the complexity of the functional roles of microbes, identifying microbial association networks, that is, microbe-microbe interaction networks, is crucial for fundamental understanding of the gut microbiome, a key contributor to the host's health.

1.2. *Motivating application: quantitative microbiome profiling data.* Microbiome data collected from 16S ribosomal RNA (rRNA) sequencing are compositional in that each subject has an arbitrary total microbial count determined by the sequencing instrument (Gloor et al., 2017). Hence, a quantitative comparison of microbial abundances cannot be made across subjects as only the information on relative abundances within a subject are available

Keywords and phrases: Gaussian copula, Markov random field, Phylogenetic tree, Zero-inflation.

from such data. Furthermore, compositional data raise a concern for biased estimates of association since a change in absolute abundance of one microbe affects the relative abundance of all the microbes (Vandeputte et al., 2017). The recently developed quantitative microbiome profiling (QMP) techniques account for these compositional limitations by adjusting microbial counts from 16S rRNA sequencing using cell counts and sequencing depths. In this work, we utilize this recent development by considering the QMP data of Vandeputte et al. (2017) with $n = 106$ healthy subjects' gut microbiome. We focus on estimating genus-level association networks with the aim of understanding the overall configurations of healthy gut microbial communities and their interactions.

1.3. *Graphical models and network estimation.* The Gaussian graphical model is a popular tool for modeling an association network via an undirected graph, where an edge between two nodes generally represents conditional dependence, and an absence of an edge represents conditional independence. Under the Gaussian assumption, this graph structure is fully encoded in the concentration matrix (also known as the inverse covariance matrix) as a zero off-diagonal entry is equivalent to the conditional independence between the corresponding variables. Thus, multiple methods focus on sparse estimation of concentration matrices. Neighborhood selection (Meinshausen et al., 2006) recovers the sparse graph structure by performing L_1-regularized regression of each node on the rest. Yuan and Lin (2007); Banerjee, El Ghaoui and d'Aspremont (2008); Dahl, Vandenberghe and Roychowdhury (2008); Friedman, Hastie and Tibshirani (2008) directly estimate the sparse concentration matrix by optimizing the L_1-penalized log-likelihood function, the so-called graphical lasso. Wang (2012) propose a Bayesian counterpart of the graphical lasso using the Laplace prior on the off-diagonal elements of the concentration matrix. Roverato (2002); Dobra, Lenkoski and Rodriguez (2011); Lenkoski and Dobra (2011) consider a G-Wishart prior for the concentration matrix, of which the posterior inference is computationally more expensive than Wang (2012). For better scalability, Wang (2015) develop a continuous spike-and-slab prior for the off-diagonal entries of the concentration matrix. Furthermore, Gaussian graphical models can be extended to non-Gaussian data via latent Gaussian copula models. Liu et al. (2012) consider Gaussian copula model for skewed continuous distributions. Fan et al. (2017) consider extension to mixed binary-continuous variables via latent Gaussian copula. Dobra et al. (2011) consider a Bayesian latent Gaussian copula for graph estimation with binary and ordinal variables, where they approximate the likelihood function using the extended rank likelihood (Hoff et al., 2007).

Despite the significant advancements in Gaussian graphical models, they are not appropriate for estimating microbial association networks. Microbiome data obtained from high-throughput sequencing are heavily right skewed and zero-inflated. The zeros are not necessarily absolute; they are often due to the limited sequencing depth. Thus, direct application of Gaussian graphical models to zero-inflated sequencing data leads to inaccurate estimation and inference. To address these challenges, several graphical models for zero-inflated data have been proposed. Osborne, Peterson and Vannucci (2021) model microbial counts using Dirichlet-multinomial distribution with a latent Gaussian graphical model. Zhou et al. (2020) consider a zero-inflated latent Ising model for microbial association network estimation. McDavid et al. (2019) propose a multivariate hurdle model, which is a mixture of degenerate (at 0) and Gaussian distributions. SPIEC-EASI (Kurtz et al., 2015) is a two-stage inference procedure specifically designed for compositional microbiome data. Yoon et al. (2019) propose Semi-Parametric Rank-based approach for INference in Graphical model (SPRING) based on truncated latent Gaussian copula (Yoon, Carroll and Gaynanova, 2020). Ma (2020) proposes truncated Gaussian graphical model.

1.4. *The major limitation of existing network estimation models and our proposal.* The aforementioned microbial network estimation models share a common limitation: they do not take advantage of additionally available evolutionary information for identifying the graph structure. The information on microbes' genetic similarities is available in the form of a phylogenetic tree. However, to our knowledge, the phylogenetic tree is not taken into account by existing methods for the estimation of microbial networks. Since microbial interactions, positive (e.g., mutualism) or negative (e.g., competition), increase with the microbes' genetic similarity (Rohr and Bascompte, 2014; Peralta, 2016), evolutionary information encoded in a phylogenetic tree has great potential in improving the accuracy of microbial associations network estimation.

In this work, we propose a Bayesian truncated Gaussian copula graphical model for microbial association networks, which takes advantage of available evolutionary information. Our major contributions are four-fold. First, we provide a general framework for incorporating evolutionary history into the estimation of microbe-microbe association networks. We model the phylogenetic tree as a Gaussian diffusion process in the latent space, which allows us to represent the microbes and their ancestors as (correlated) Gaussian vectors. Our framework is not limited to the phylogenetic tree and can accommodate any prior knowledge that is expressed in a tree, e.g., a taxonomic rank tree. We formulate the prior probability model on graph so that the microbes that are closer to each other on the tree have a higher edge inclusion probability. Our simulation study reveals that our approach significantly improves the graph estimation accuracy compared to the methods that do not take advantage of the tree structure (Section 4). Second, the proposed model effectively handles zero-inflation resulting from limited sequencing depth. We consider the observed zeros as truncated realizations of unobserved random quantities that are below certain thresholds. In particular, we establish a Bayesian formulation of the truncated Gaussian copula model (Yoon, Carroll and Gaynanova, 2020) and develop an efficient Gibbs sampling algorithm. Third, the proposed approach facilitates the statistical inference on the estimated network. For each pair of nodes, an edge connectivity is immediately available from the posterior sample, which provides a convenient way to control the posterior expected FDR (Mitra et al., 2013; Peterson, Stingo and Vannucci, 2015). Finally, while our model is designed for quantitative microbiome data, it can also be applied to compositional data using modified centered log-ratio transformation (Yoon et al., 2019).

The rest of this paper is organized as follows. In Section 2, we introduce the proposed graphical model. In Section 3, we discuss posterior inference. In Section 4, we evaluate the graph estimation accuracy of the proposed model on simulated datasets. In Section 5, we analyze quantitative microbiome profiling data of Vandeputte et al. (2017), and compare our results to SPRING (Yoon et al., 2019) and SPIEC-EASI (Kurtz et al., 2015). In Section 6, we provide a brief discussion and the link to download the R code that implements our method.

2. Bayesian truncated Gaussian copula graphical model. In Section 2.1, we discuss the semiparametric modeling of conditional dependencies for zero-inflated data through a truncated Gaussian copula model with a sparse concentration matrix. In Section 2.2, we introduce a prior model that incorporates the phylogenetic tree to facilitate posterior inference of microbial associations. The complete hierarchical model is summarized in Figure 1.

2.1. *Truncated Gaussian copula graphical model.* Let $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$ denote the zero-inflated abundances of p microbes. This is either directly the counts resulting from quantitative microbiome profiling, e.g. motivating data from Vandeputte et al. (2017), or transformed compositional microbiome data via the modified central log-ratio transformation (Yoon et al., 2019). We propose to model $\boldsymbol{x}$ such that only the microbial abundances that

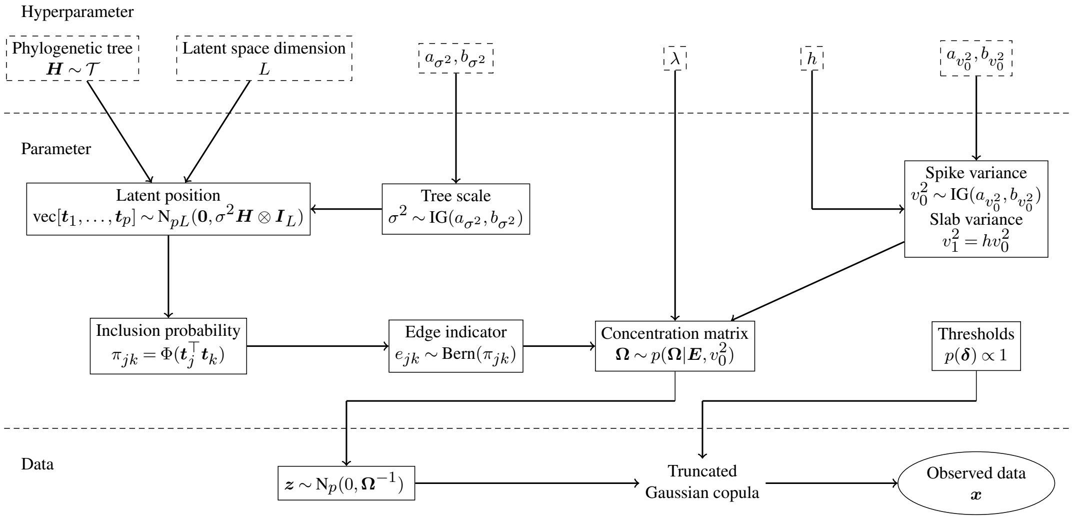
Hyperparameter
Phylogenetic tree $\boldsymbol{H} \sim \mathcal{T}$
Latent space dimension L
$a_{\sigma^2}, b_{\sigma^2}$
λ
h
$a_{v_0^2}, b_{v_0^2}$
Parameter
Latent position $\mathrm{vec}[\boldsymbol{t}_1, \dots, \boldsymbol{t}_p] \sim \mathrm{N}_{pL}(\boldsymbol{0}, \sigma^2 \boldsymbol{H} \otimes \boldsymbol{I}_L)$
Tree scale $\sigma^2 \sim \mathrm{IG}(a_{\sigma^2}, b_{\sigma^2})$
Spike variance $v_0^2 \sim \mathrm{IG}(a_{v_0^2}, b_{v_0^2})$
Slab variance $v_1^2 = h v_0^2$
Inclusion probability $\pi_{jk} = \Phi(\boldsymbol{t}_j^\top \boldsymbol{t}_k)$
Edge indicator $e_{jk} \sim \mathrm{Bern}(\pi_{jk})$
Concentration matrix $\boldsymbol{\Omega} \sim p(\boldsymbol{\Omega} | \boldsymbol{E}, v_0^2)$
Thresholds $p(\boldsymbol{\delta}) \propto 1$
Data
$\boldsymbol{z} \sim \mathrm{N}_p(0, \boldsymbol{\Omega}^{-1})$
Truncated Gaussian copula
Observed data $\boldsymbol{x}$

FIG 1. *Schematic illustration of the phylogenetically informed Bayesian truncated Gaussian copula graphical model. Hyperparameters that are held constant are given in boxes with dashed-line. The quantities that need posterior inference are illustrated in boxes with solid-line. The ellipse with solid line represents the observed data.*

are larger than certain thresholds can be observed. Specifically, we assume that there exist latent $\boldsymbol{x}^* = (x_1^*, \dots, x_p^*)^\top$ representing the true abundances such that

$$x_j = 1(x_j^* > c_j)x_j^*, \quad j = 1, \dots, p, \tag{1}$$

where $1(\cdot)$ is the indicator function and c_j's are unknown thresholds, which allow x_j's to have different levels of zero-inflation. We call x_j a *truncated* variable if x_j^* is less than c_j, and an *observed* variable, otherwise. Let F_j be the marginal cumulative distribution function (cdf) of the jth latent variable x_j^*, Φ be the cdf of standard Gaussian, and $f_j = \Phi^{-1} \circ F_j$, where we assume that F_j's are continuous. The truncated Gaussian copula model (Yoon, Carroll and Gaynanova, 2020) assumes

$$z_j = f_j(x_j^*), \quad j = 1 \dots, p, \tag{2}$$

$$\boldsymbol{z} = (z_1, \dots, z_p)^\top \sim \mathrm{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{3}$$

where $\boldsymbol{\Sigma} = \mathrm{Corr}\{\boldsymbol{\Omega}^{-1}\} \succ 0$ is the positive definite correlation matrix and Corr converts a positive definite matrix to a correlation matrix.

Since f_j's are monotone continuous, we can write (1) as $x_j = 1\{f_j(x_j^*) > f_j(c_j)\}x_j^* = 1(z_j > \delta_j)x_j^*$, where $\delta_j = f_j(c_j)$. We denote the truncated and the observed sub-vectors of $\boldsymbol{x}$ by $\boldsymbol{x}_t \in \mathbb{R}^{p_t}$ and $\boldsymbol{x}_o \in \mathbb{R}^{p_o}$, respectively, where $p_t + p_o = p$. Likewise, let $\boldsymbol{z}_t$ and $\boldsymbol{z}_o$ be the corresponding latent Gaussian vectors, and let $\boldsymbol{\delta}_t$ and $\boldsymbol{\delta}_o$ be their thresholds. Given the thresholds, the conditional distribution of $\boldsymbol{z}$ is given by

$$p(\boldsymbol{z}_o, \boldsymbol{z}_t | \boldsymbol{z}_o > \boldsymbol{\delta}_o, \boldsymbol{z}_t < \boldsymbol{\delta}_t, \boldsymbol{\Omega}) = \frac{\mathrm{N}_p(\boldsymbol{z}_o, \boldsymbol{z}_t | \boldsymbol{0}, \boldsymbol{\Sigma})}{\mathbb{P}(\boldsymbol{z}_o > \boldsymbol{\delta}_o, \boldsymbol{z}_t < \boldsymbol{\delta}_t | \boldsymbol{\delta}, \boldsymbol{\Omega})} 1(\boldsymbol{z}_o > \boldsymbol{\delta}_o) 1(\boldsymbol{z}_t < \boldsymbol{\delta}_t). \tag{4}$$

Unlike the approach in Dobra et al. (2011) that only uses the relative ranks of the observed data, we condition on the observed value of $\boldsymbol{z}_o$, which subsequently allows us to sample truncated variables $\boldsymbol{z}_t$ from the posterior distribution as discussed in Section 3.

Because of the multivariate normality of $\boldsymbol{z}$, zero entries of $\boldsymbol{\Omega} = [\omega_{jk}]_{1 \le j,k \le p}$ imply the conditional independence between the corresponding variables. The dependency structure of $\boldsymbol{z}$ can be graphically summarized as an undirected graph $G = (\boldsymbol{z}, \boldsymbol{E})$ with an adjacency matrix $\boldsymbol{E} = [e_{jk}]_{1\le j,k\le p}$, where nodes z_j and z_k are connected (denoted by $e_{jk} = 1$) if $\omega_{jk} \neq 0$. Consequently, learning the graph structure (adjacency matrix) $\boldsymbol{E}$ is equivalent to finding the sparse pattern of $\boldsymbol{\Omega}$. To encourage sparsity, we follow a similar strategy as in Wang (2015) by assigning a spike-and-slab prior on the off-diagonal elements of $\boldsymbol{\Omega}$ and an exponential prior on the diagonal elements,

$$p(\boldsymbol{\Omega}|\boldsymbol{E}, v_0^2) = C(\boldsymbol{E}, v_0^2)^{-1} 1(\boldsymbol{\Omega} \succ 0) \prod_{j<k} \left\{(1 - e_{jk})\mathrm{N}(\omega_{jk}|0, v_0^2) + e_{jk}\mathrm{N}(\omega_{jk}|0, hv_0^2)\right\} \times \prod_{j=1}^{p} \mathrm{Exp}(\omega_{jj}|\frac{\lambda}{2}), \tag{5}$$

where $\mathrm{Exp}(\cdot|\lambda)$ is the exponential density function with rate parameter λ, v_0^2 is the spike variance, $h \gg 1$ is a large constant such that the slab variance $hv_0^2 \gg v_0^2$, and $C(\boldsymbol{E}, v_0^2)$ is the normalizing constant.

We impose an improper uniform prior on the thresholds $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)^\top \sim p(\boldsymbol{\delta}) \propto 1$, a conjugate inverse-gamma prior on the spike variance $v_0^2 \sim \mathrm{IG}(a_{v_0}, b_{v_0})$, and Bernoulli-like priors on the edge indicators

$$p(\boldsymbol{E}) \propto C(\boldsymbol{E}, v_0^2) \prod_{j<k} \mathrm{Ber}(e_{jk}|\pi_{jk}). \tag{6}$$

Including the normalizing constant $C(\boldsymbol{E}, v_0^2)$ in (6) serves to cancel out that in (5), facilitating the posterior computation in updating $\boldsymbol{E}$ (Wang, 2015).

2.2. *Incorporating phylogenetic tree.* Evolution plays an important role in shaping the interaction patterns of microbes (Peralta, 2016). We will exploit the evolution footprints in identifying microbial association networks through a novel phylogenetic tree prior. The proposed prior is a distribution on edge inclusion probabilities $\boldsymbol{\Pi} = [\pi_{jk}]_{1\le j,k\le p}$ that encourages the interactions, positive (e.g., mutualism) or negative (e.g., competition), of phylogenetically similar microbes as they tend to be phenotypically/functionally correlated (Martiny et al., 2015; Xiao et al., 2018; Zhou et al., 2021). The prior is constructed by first embedding the network in L-dimensional Euclidean space through the latent position model (Hoff, Raftery and Handcock, 2002), and then arranging the latent positions according to the phylogenetic tree.

Latent position model. We introduce a latent position $\boldsymbol{t}_j = (t_{1j}, \ldots, t_{Lj})^\top \in \mathbb{R}^L$ for each node z_j, and link it to the edge inclusion probability π_{jk} through the probit link function

$$\pi_{jk} = \Phi(\boldsymbol{t}_j^\top \boldsymbol{t}_k), \quad j < k.$$

The inner product $\boldsymbol{t}_j^\top \boldsymbol{t}_k$ measures the similarity between $\boldsymbol{t}_j$ and $\boldsymbol{t}_k$, with larger inner product leading to higher prior inclusion probability. We assign a prior on $\boldsymbol{t}_j$'s to encourage the interactions between phylogenetically similar microbes.

Phylogenetic tree. Let $\boldsymbol{T} = [\boldsymbol{t}_1, \ldots, \boldsymbol{t}_p] \in \mathbb{R}^{L\times p}$ and let $\boldsymbol{t}^\ell = (t_{1\ell}, \ldots, t_{p\ell})$ be the ℓth row of $\boldsymbol{T}$. We assume $\boldsymbol{t}^\ell \overset{iid}{\sim} \mathrm{N}_p(\boldsymbol{0}, \sigma^2\boldsymbol{H})$ for $\ell = 1, \ldots, L$, where $\boldsymbol{H}$ is a correlation matrix that reflects the phylogenetic similarity. Our specific choice of $\boldsymbol{H}$ is motivated by the following diffusion process. Let $\mathcal{T}$ be a phylogenetic tree with terminal nodes representing the p microbes under investigation and internal nodes representing their common ancestors. Starting from time 0 at the origin (root), the first branch $\boldsymbol{t}_1$ follows a Brownian motion with variance σ^2 until the divergence time $s_1 \in [0, 1]$. Then it splits into two branches, $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$, each following the same Brownian motion independently before they split at times $s_2, s_3 \in [0, 1]$

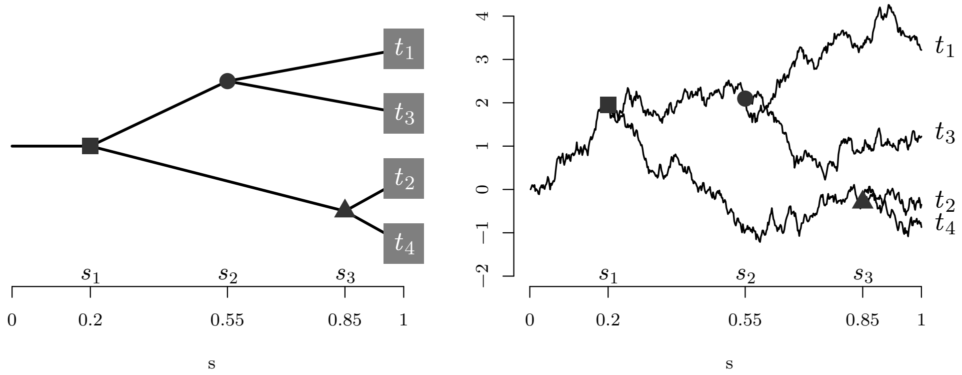

FIG 2. *An illustrative example of a phylgenetic tree with $p = 4$ microbes (left) and the corresponding diffusion process in $\mathbb{R}^1$ with $\sigma^2 = 3$ (right). The new branch t_2 is split from t_1 at the first divergence time $s_1 = 0.2$. Then new branches t_3 and t_4 are split from t_1 and t_2 at the divergence time $s_2 = 0.55$ and $s_3 = 0.85$, respectively. The green circle and blue triangle are the most common ancestors of the pairs (t_1, t_3) and (t_2, t_4), whose heights, $s_2 = 0.55$ and $s_3 = 0.85$, are the correlations of the pairs, respectively. The color figure is available in the online version.*

resulting in $\boldsymbol{t}_3$ and $\boldsymbol{t}_4$. This process repeats until the p terminal nodes are reached at time 1. An illustrative example of the diffusion process is provided in Figure 2.

This diffusion process defines a centered multivariate Gaussian distribution on the terminal nodes of $\mathcal{T}$ with covariance matrix $\sigma^2\boldsymbol{H}$. We define the *height* of each node (split for internal nodes) as its distance from the root (time from 0). The correlation of two terminal nodes equals the height of their most recent common ancestor, which is large for a phylogenetically similar microbes. This multivariate Gaussian prior, together with the latent position model, achieves the desired prior distribution of π_{jk} that encourages interactions between phylogenetically similar microbes. Lastly, we assign a conjugate inverse-gamma prior for the variance (tree scale) parameter $\sigma^2 \sim \text{IG}(a_{\sigma^2}, b_{\sigma^2})$.

3. Posterior inference.

The proposed model is parameterized by $\{\boldsymbol{z}, \boldsymbol{\delta}, \boldsymbol{\Omega}, \boldsymbol{E}, v_0^2, \boldsymbol{T}, \sigma^2\}$ of which the posterior distribution is not available in closed form. We use a Markov chain Monte Carlo (MCMC) algorithm to draw posterior samples from the intractable posterior distribution. Section 3.1 discusses Gibbs steps for sampling $\boldsymbol{z}$ and $\boldsymbol{\delta}$ from their full conditional distributions. Section 3.2 describes Gibbs steps for $\boldsymbol{T}$ and σ^2. In Section S1.2 of the Supplementary Materials, we provide the Gibbs steps for $\boldsymbol{\Omega}, \boldsymbol{E}, v_0^2$ by following Wang (2015). An outline of the Gibbs sampling steps is provided in Algorithm 1.

3.1. *Full conditionals of truncated observations and thresholds.*

Let $\boldsymbol{x}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, be a sample from the model (1)–(3) and let $\boldsymbol{z}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, be the corresponding latent Gaussian vectors. As before, we use subscripts t and o, respectively, to denote the truncated and observed sub-vectors (e.g., $\boldsymbol{x}_{i,t}$ and $\boldsymbol{x}_{i,o}$ are the truncated and observed sub-vectors of $\boldsymbol{x}_i$).

For observed $\boldsymbol{x}_{i,o}$, the corresponding Gaussian variables are defined as $z_{ij,o} = f_j(x_{ij,o}) = \Phi^{-1} \circ F_j(x_{ij,o})$. A natural estimator for the unknown function F_j is the scaled empirical cdf $\widehat{F}_j(c) = \{n/(n+1)\}\sum_{i=1}^n n^{-1}1(x_{ij} \leq c)$ (Klaassen et al., 1997), where the constant term $n/(n+1)$ is needed to make Φ^{-1} finite. Thus, the estimator of f_j is given by $\widehat{f}_j = \Phi^{-1} \circ \widehat{F}_j$, and we set $\widehat{z}_{ij,o} = \widehat{f}_j(x_{ij,o})$. An alternative approach to estimate f_j using B-spline

Algorithm 1 Outline of the Gibbs sampling steps.

```
**Data:** $\widehat{\boldsymbol{z}}_{1,o}, \dots, \widehat{\boldsymbol{z}}_{n,o}$, where $\widehat{z}_{ij,o} = \widehat{f}_j(x_{ij,o})$.
Initialize $\boldsymbol{\Sigma}$, $\boldsymbol{\delta}$, $\boldsymbol{E}$, $\boldsymbol{\Pi}$, $v_0^2$, and $\sigma^2$ and
**for** $s = 1, \dots, S$ **do**
    Draw truncated Gaussian vectors $\boldsymbol{z}^*_{i,t} \sim p(\boldsymbol{z}_{i,t} | \widehat{\boldsymbol{z}}_{i,o}, \boldsymbol{\Sigma}, \boldsymbol{\delta})$, $i = 1, \dots, n$
    Set $\boldsymbol{Z}^* = [\boldsymbol{z}^*_1, \dots, \boldsymbol{z}^*_n]^\top$, where components of $\boldsymbol{z}^*_i$ are properly arranged with $\boldsymbol{z}^*_{i,t}$ and $\widehat{\boldsymbol{z}}_{i,o}$.
    Draw threshold parameter $\boldsymbol{\delta}^* \sim p(\boldsymbol{\delta} | \boldsymbol{Z}^*)$
    Draw concentration matrix $\boldsymbol{\Omega}^* \sim p(\boldsymbol{\Omega} | \boldsymbol{Z}^*, \boldsymbol{\Sigma}, \boldsymbol{E}, v_0^2)$
    Set correlation matrix $\boldsymbol{\Sigma}^* = \mathrm{Corr}\{\boldsymbol{\Omega}^{*-1}\}$
    Draw graph $\boldsymbol{E}^* \sim p(\boldsymbol{E} | \boldsymbol{\Omega}^*, v_0^2, \boldsymbol{\Pi})$;
    Draw spike variance $v_0^{2*} \sim p(v_0^2 | \boldsymbol{\Omega}^*, \boldsymbol{E}^*)$.
    Draw latent positions $\boldsymbol{t}^*_j \sim p(\boldsymbol{t}_j | \boldsymbol{T}^*_{-j}, \sigma^2)$ $j = 1, \dots, p$
    Set edge inclusion probabilities $\boldsymbol{\Pi}^* = [\pi^*_{jk}]_{1 \le j,k \le p}$, where $\Phi(\boldsymbol{t}_j^{*\top} \boldsymbol{t}^*_k)$.
    Draw tree scale parameter $\sigma^{2*} \sim p(\sigma^2 | \boldsymbol{T}^*)$
**end for**
```

basis functions has been considered in Mulgrave et al. (2020); however, we use the empirical cdf for computational simplicity.

Given $\boldsymbol{z}_{i,o} = \widehat{\boldsymbol{z}}_{i,o}$, we sample $\boldsymbol{z}_{i,t}$ from a truncated multivariate Gaussian distribution derived from (4),

$$p(\boldsymbol{z}_{i,t} | \widehat{\boldsymbol{z}}_{i,o}, \boldsymbol{z}_{i,t} < \boldsymbol{\delta}_t, \boldsymbol{\Omega}) = \frac{\mathrm{N}_{p_{i,t}}(\boldsymbol{z}_{i,t} | \boldsymbol{\mu}_i, \boldsymbol{\Delta}_i)}{\mathbb{P}(\boldsymbol{z}_{i,t} < \boldsymbol{\delta}_t | \widehat{\boldsymbol{z}}_{i,o}, \boldsymbol{\Omega}, \boldsymbol{\delta})} 1(\boldsymbol{z}_{i,t} < \boldsymbol{\delta}_{i,t}),$$

where $p_{i,t}$ is the number of truncated variables of $\boldsymbol{x}_i$, and $\boldsymbol{\mu}_i$, $\boldsymbol{\Delta}_i$ are the mean and covariance matrix of $\boldsymbol{z}_{i,t}$ given $\boldsymbol{z}_{i,o} = \widehat{\boldsymbol{z}}_{i,o}$ (detailed expressions are provided in Section S1.1 of the Supplementary Materials). The conditional pdf of $\boldsymbol{\delta}$ given $\boldsymbol{z}_1, \dots, \boldsymbol{z}_n$ is proportional to the n-product of indicator functions of (4). That is, we have the independent uniform full conditional distributions of δ_j's as

$$\begin{aligned} p(\boldsymbol{\delta} | \boldsymbol{z}_1, \dots, \boldsymbol{z}_n, \boldsymbol{\Omega}) &\propto \prod_{i=1}^{n} \mathrm{N}_p(\boldsymbol{z}_{i,t}, \widehat{\boldsymbol{z}}_{i,o}, | \boldsymbol{\Omega}) 1(\boldsymbol{z}_{i,t} < \boldsymbol{\delta}_{i,t}) 1(\widehat{\boldsymbol{z}}_{i,o} > \boldsymbol{\delta}_{i,o}), \\ &\propto \prod_{j=1}^{p} 1(z^{\max}_{j,t} < \delta_j < \widehat{z}^{\min}_{j,o}), \end{aligned}$$

where $z^{\max}_{j,t} = \max_i z_{ij,t}$ and $\widehat{z}^{\min}_{j,o} = \min_i \widehat{z}_{ij,o}$ are the maximum and minimum of the truncated and observed components of the jth Gaussian variable, respectively.

3.2. *Full conditionals of latent positions and the tree scale.* Let $\boldsymbol{T}_{-j}$ be the submatrix of $\boldsymbol{T}$ without the jth column. The full conditional distribution of $\boldsymbol{t}_j$ is given by,

$$p(\boldsymbol{t}_j | \boldsymbol{T}_{-j}, \boldsymbol{E}, \sigma^2) \propto \left[\prod_{k \neq j} \{\Phi(\boldsymbol{t}_k^\top \boldsymbol{t}_j)\}^{e_{kj}} \{1 - \Phi(\boldsymbol{t}_k^\top \boldsymbol{t}_j)\}^{1 - e_{kj}} \right] \mathrm{N}_L(\boldsymbol{t}_j | \boldsymbol{\theta}_j, \boldsymbol{\Psi}_j), \quad j = 1, \dots, p,$$

where $\boldsymbol{\theta}_j$ and $\boldsymbol{\Psi}_j$ are mean and covariance matrix of $\boldsymbol{t}_j$ given $\boldsymbol{T}_{-j}$ (detailed expressions are provided in Section S1.3 of the Supplementary Materials). We update $\boldsymbol{t}_j$ using the data augmentation technique of Albert and Chib (1993) by introducing the auxiliary data $\boldsymbol{y}_j \in$

$\mathbb{R}^{p-1}$. Let $\mathrm{TN}(\mu, \sigma^2, e)$ be $\mathrm{N}(\mu, \sigma^2)$ truncated to be positive if $e = 1$ and negative if $e = 0$. Conditining on $\boldsymbol{T}$, each component y_{kj} of $\boldsymbol{y}_j$ follows $\mathrm{TN}(\boldsymbol{t}_k^\top \boldsymbol{t}_j, 1, e_{kj})$, and the resulting augmented pdf of $\boldsymbol{t}_j$ and $\boldsymbol{y}_j$ is

$$p(\boldsymbol{t}_j, \boldsymbol{y}_j | \boldsymbol{T}_{-j}, \boldsymbol{E}, \sigma^2) \propto \prod_{k \neq j} \{1(y_{kj} > 0, e_{kj} = 1) + 1(y_{kj} < 0, e_{kj} = 0)\}\, \mathrm{N}(y_{kj} | \boldsymbol{t}_k^\top \boldsymbol{t}_j, 1) \mathrm{N}_L(\boldsymbol{t}_j | \boldsymbol{\theta}_j, \boldsymbol{\Psi}_j).$$

We obtain a posterior sample of $\boldsymbol{T}$ by alternately sampling $\boldsymbol{y}_j$ and $\boldsymbol{t}_j$ for $j = 1, \ldots, p$. Conditional on $\boldsymbol{T}$, we independently sample y_{kj} from $\mathrm{TN}(\boldsymbol{t}_k^\top \boldsymbol{t}_j, 1, e_{kj})$ for $k \neq j$. Then, conditional on $\boldsymbol{y}_j$, we have the Gaussian full conditional of $\boldsymbol{t}_j$ as

$$p(\boldsymbol{t}_j | \boldsymbol{y}_j, \boldsymbol{T}_{-j}, \boldsymbol{E}, \sigma^2) \propto \mathrm{N}_{p-1}(\boldsymbol{y}_j | \boldsymbol{T}_{-j}^\top \boldsymbol{t}_j, \boldsymbol{I}_{p-1}) \mathrm{N}_L(\boldsymbol{t}_j | \boldsymbol{\theta}_j, \boldsymbol{\Psi}_j).$$

Accordingly, we draw $\boldsymbol{t}_j$ from $\mathrm{N}_L(\boldsymbol{\gamma}_j, \boldsymbol{\Gamma}_j)$, where

$$\boldsymbol{\Gamma}_j = \left(\boldsymbol{T}_{-j} \boldsymbol{T}_{-j}^\top + \boldsymbol{\Psi}_j^{-1}\right)^{-1}, \quad \boldsymbol{\gamma}_j = \boldsymbol{\Gamma}_j \left(\boldsymbol{T}_{-j} \boldsymbol{y}_j + \boldsymbol{\Psi}_j^{-1} \boldsymbol{\theta}_j\right).$$

The edge inclusion probabilities are updated as $\pi_{jk} = \Phi(\boldsymbol{t}_j^\top \boldsymbol{t}_k)$, $1 \leq j < k \leq p$. Conditional on $\boldsymbol{T}$, we sample the tree scale parameter σ^2 from

$$\sigma^2 | \boldsymbol{T} \sim \mathrm{IG}\left(pL/2 + a_{\sigma^2}, \mathrm{vec}(\boldsymbol{T})^\top (\boldsymbol{H} \otimes \boldsymbol{I}_L)^{-1} \mathrm{vec}(\boldsymbol{T})/2 + b_{\sigma^2}\right),$$

where $\mathrm{vec}(\boldsymbol{T})$ is the vector obtained by stacking the columns of $\boldsymbol{T}$ and $\otimes$ is the Kronecker product.

For sampling concentration matrix and graph, we follow the block Gibbs sampler of Wang (2015). For completeness, we provide the block Gibbs sampling algorithm in Section S1.2 of the Supplementary Materials.

3.3. *Posterior point estimation and prediction.*

Upon the completion of the MCMC, we compute the posterior mean of the edge inclusion probabilities for each pair of nodes, $\widehat{\pi}_{jk} = \sum_{s=1}^{S} e_{jk}^{(s)}/S$, where the superscript indexes posterior samples. We obtain the estimated graph by selecting edges for which $\widehat{\pi}_{jk}$ is larger than some cutoff. We choose the cutoff to control the posterior expected FDR (Mitra et al., 2013; Peterson, Stingo and Vannucci, 2015) at prespecified level α, where the posterior expected FDR is a decreasing function of cutoff c defined as

$$\mathrm{E}(\mathrm{FDR}_c | \mathrm{data}) = \frac{\sum_{j<k} (1 - \widehat{\pi}_{jk}) 1(\widehat{\pi}_{jk} > c)}{\sum_{j<k} 1(\widehat{\pi}_{jk} > c)}. \tag{7}$$

The posterior samples generated from the MCMC also allows for posterior prediction of microbial abundance via the posterior predictive distribution. We describe the posterior predictive sampling procedure in Supplementary Materials S3 with the goodness-of-fit assessment of the proposed model on the real data considered in Section 5. Supplementary Materials S1.4 describes a conditional posterior prediction procedure for x_j^{new} when a new data point $\boldsymbol{x}^{\mathrm{new}}$ is observed without the jth variable. We also discuss the inherent difficulties of microbial abundance prediction.

4. Simulation.

We simulate microbiome data following the data generation mechanism proposed in Yoon et al. (2019), which allows us to obtain synthetic samples that exactly follow the empirical marginal cumulative distributions of measured microbiome count data while respecting user-specified microbial dependencies via $\boldsymbol{\Omega}$.

Specifically, we randomly generate 10 phylogenetic trees $\mathcal{T}_1, \ldots, \mathcal{T}_{10}$ with $p = 50$ terminal nodes using the function rcoal of the ape R package (Paradis and Schliep, 2019), where the coalescent times, the distance between two descendants and the merge of the two branches, are exponentially distributed (Paradis, 2012, Chapter 7). For completeness, we provide detailed tree generation algorithm in Section S8 of Supplementary Materials. For each tree, the latent positions of terminal nodes, $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_p$, are generated from the diffusion process as described in Section 2.2 with $\sigma^2 = 3$ and $L = 2$. The true graph adjacency matrix $\boldsymbol{E}_0$ is obtained by independently generating $e_{jk} \sim \text{Bernoulli}(\pi_{jk})$ for $1 \leq j < k \leq p$ with $\pi_{jk} = \Phi(\boldsymbol{t}_j^\top \boldsymbol{t}_k)$. The trees and the true graphs are plotted in Section S9 of the Supplementary Materials. Given $\boldsymbol{E}_0$, the concentration matrix $\boldsymbol{\Omega}$ is drawn from G-Wishart$(\boldsymbol{I}_p, 4)$ (Roverato, 2002). To obtain empirical cdfs, we use the quantitative microbiome profiling (QMP) data of Vandeputte et al. (2017) from $n = 106$ subjects, more detailed description of QMP data is provided in Section 5. We select $p = 50$ genera (variables) of which 6 genera have no observed zero counts, and 44 genera have 20% to 70% zero counts across samples.

Given the empirical cdf $\widehat{F}_j$ of each selected genus, $j = 1, \ldots, p$, and the correlation matrix $\boldsymbol{\Sigma} = \text{Corr}\{\boldsymbol{\Omega}^{-1}\}$, we generate $n = 106$ independent latent Gaussian vectors $\boldsymbol{z}_i \sim \text{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$. The final data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are obtained as

$$x_{ij} = \widehat{F}_j^- \circ \Phi(z_{ij}), \quad i = 1, \ldots, n, \quad j = 1, \ldots, p,$$

where $\widehat{F}_j^-$ is the pseudo inverse of $\widehat{F}_j$ defined as $\widehat{F}_j^-(u) = \min\{x_{ij} | \widehat{F}_j(x_{ij}) \geq u\}$, $j = 1, \ldots, p$. Marginally, this is equivalent to uniform sampling of real observations with replacement but the joint association structure is induced by $\boldsymbol{\Sigma}$. The comparison with the actual QMP data (in Supplementary Material S2) indicates that simulated data well represent the reality. We consider 50 independent replications of this data generating process for each scenario.

We compare the performance of the the proposed phylogenetically-informed Bayesian Copula Graphical model (PhyloBCG) with SPIEC-EASI (Kurtz et al., 2015) and SPRING (Yoon et al., 2019). Additionally, we also consider three special cases of PhyloBCG with the following simplification to the prior model of graph:

$$\begin{aligned}
\text{Oracle}: \ & \pi_{jk} = \pi = \binom{p}{2}^{-1} |\boldsymbol{E}_0|; \\
\text{Uniform}: \ & \pi_{jk} = \pi \sim \text{Beta}(1, 1); \\
\text{Distance}: \ & \pi_{jk} = \exp(-\gamma d_{jk}), \quad \gamma \sim \text{Exp}(1);
\end{aligned}$$

where $|\boldsymbol{E}_0|$ is the number of true edges in the underlying graph, and d_{jk} is the tree distance between terminal nodes j and k, defined as the sum of the branch lengths to their most recent common ancestor. We refer to the first and second models as “Oracle” and “Uniform”. While Oracle does not explicitly use the phylogenetic information, it utilizes this information indirectly through the true graph sparsity level. On the contrary, Uniform is completely non-informative. We refer to the third model as “Distance”, as it directly incorporates tree distances between the terminal nodes: the edge inclusion probability is higher (smaller) when the corresponding nodes are closer (farther) to each other on the tree. Distance model thus takes into account the information from the tree, however, the tree information is deterministically incorporated to the model in contrast to the stochastic incorporation of PhyloBCG.

To implement SPIEC-EASI and SPRING, we use their corresponding R packages (Kurtz et al., 2021; Yoon, Gaynanova and Müller, 2019b) with sparsity parameters tuned over 100 values. For PhyloBCG, Oracle, Uniform and Distance, the hyperparameters are fixed as $a_{\sigma^2} = b_{\sigma^2} = a_{v_0^2} = b_{v_0^2} = 0.001$, $h = 2500$, $\lambda = 1$. We also fix the latent space dimension at $L = 2$ to facilitate visual interpretation. Our sensitivity analysis (Supplementary Material S4) indicates that PhyloBCG is relatively robust to hyperparameter settings. For the latent

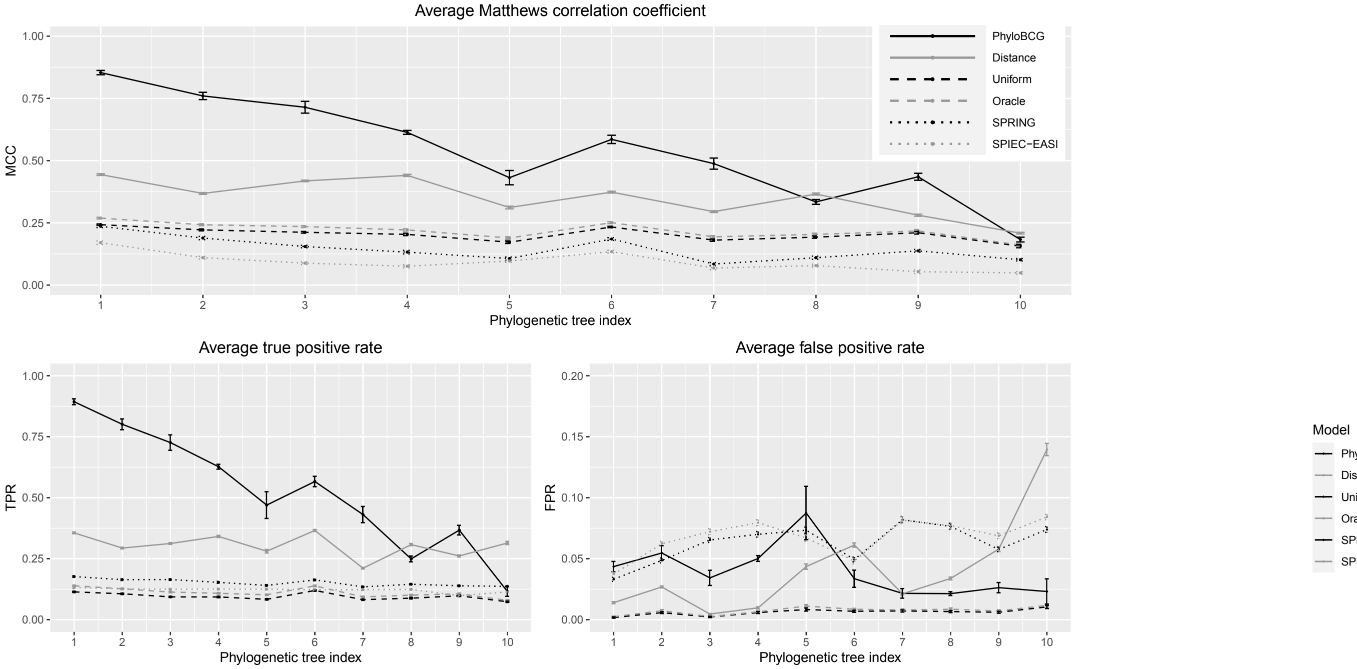

FIG 3. *Averages of Matthews correlation coefficient (MCC), true positive rates (TPR) and false positive rates (FPR) with 2 standard error bars. The phylogenetic trees are indexed in decreasing order of the global clustering coefficients of the true graphs* . The models utilizing phylogenetic information (PhyloBCG and Distance) are depicted with solid lines; the models without phylogenetic information are depicted with dashed-lines (Uniform and Oracle), and dotted-lines (SPRING and SPIEC-EASI). Averages are taken over 50 replicated datasets. The color figure is available in the online version.

space dimension, the sensitivity analysis indicates that, as the latent space dimension increases, both true and false positive rates monotonically increase, resulting in denser network estimates. Our general recommendation is to use $L = 2$ for ease of visual interpretation; see Figure 7 as an example where clustered microbes' latent positions are visualized in $\mathbb{R}^2$ and the microbial cluster with distinct characteristics is located away from the rest of the clusters. However, depending on the study, one may consider to increase L if having a larger true positive rate is more important than controlling false positive rate. We obtain an MCMC posterior sample of size $S = 5000$ after 500 burn-in iterations. The graph estimate is obtained by thresholding the mean of posterior inclusion probability $\widehat{\pi}_{jk}$ at $c_{0.05}$, the smallest c that controls the posterior expected FDR (7) at level 0.05.

We assess the graph recovery performance (accuracy in estimating $\boldsymbol{E}_0$) of each method using Matthews correlation coefficient (Matthews, 1975), true positive rate, and false positive rate, which will be denoted by MCC, TPR, and FPR, respectively. Ranging from -1 to 1, a larger value of MCC represents a better network estimation accuracy, where the two boundary values indicate completely correct (+1) and wrong (-1) edge selection, respectively. Figure 3 summarizes the mean values of these metrics for each of the 10 phylogenetic trees based on 50 replications. The phylogenetic trees are indexed in terms of the global clustering coefficient (Wasserman et al., 1994) of the true graph from largest value ($\mathcal{T}_1$) to lowest value ($\mathcal{T}_{10}$). A large value of the global clustering coefficient indicates a presence of microbial communities, with dense interactions within the same community and sparse interactions across communities. A small value of the global clustering coefficient indicates a random interaction pattern close to what will be expected with Erdős-Rényi random graph. Thus,

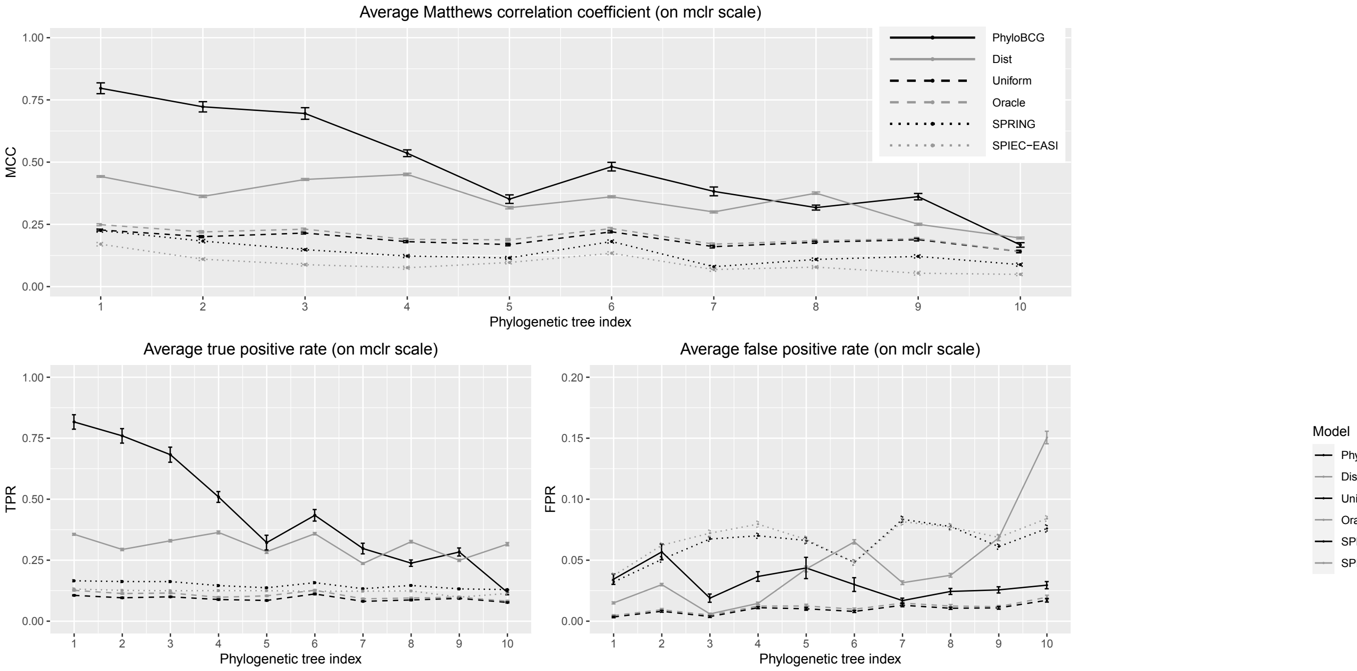

FIG 4. *Compositional data with modified centered log-ratio transformation. Averages of Matthews correlation coefficient (MCC), true positive rates (TPR) and false positive rates (FPR) with 2 standard error bars. The phylogenetic trees are indexed in decreasing order of the global clustering coefficients of the true graphs* . Averages are taken over 50 replicated datasets. The color figure is available in the online version.

we anticipate the phylogenetic tree to be more informative for network estimation when the global clustering coefficient is larger.

Figure 3 supports that incorporating phylogenetic tree information improves the network estimation accuracy, with PhyloBCG and Distance having higher MCC, higher TPR, and similar FPR values when compared to other methods. As expected, the value of MCC for the tree-based methods decreases as the phylogenetic tree becomes less informative (larger tree index). For PhyloBCG, this trend is driven by the decreasing TPR, whereas, for Distance, it is the increasing FPR. Although PhyloBCG and Distance both use the evolutionary information, PhyloBCG shows significantly better performances than Distance possibly due to the flexibility of the latent space embedding. Although Oracle and Uniform show similar performance across all the settings, Oracle slightly outperforms Uniform in terms of MCC, TPR, and FPR values due to the utilization of the true graph sparsity level.

Note that the FPR of PhyloBCG has larger variability for trees $\mathcal{T}_5$ than others, with $\mathcal{T}_5$ leading to PhyloBCG's largest mean FPR. The reason for increased FPR in this setting is the discrepancy between the true graph and the phylogenetic tree, i.e., tree prior misspecification. Recall that the true edge inclusion probabilities π_{jk} are obtained from the Gaussian latent positions rather than directly from the tree, thus allowing the true graph to deviate, sometimes significantly, from the phylogenetic tree. Figure S13 in the Supplementary Materials shows the upper triangular part of the true tree correlation matrix $\boldsymbol{H}$ defined in Section 2.2 against the edge indicators for $\mathcal{T}_1 - \mathcal{T}_{10}$. It can be seen that $\mathcal{T}_5$ shows a large number of disconnected edges, $e_{jk} = 0$, with high tree correlation values which may contribute to the increase in FPR. Nevertheless, for $\mathcal{T}_5$, PhyloBCG still shows favorable performance, having much larger TPRs than SPIEC-EASI and SPRING. Additional simulations are conducted to further investigate the robustness of the proposed method to the tree prior misspecification. In particular, we consider two cases of tree misspecification. In the first case, we randomly permuted the

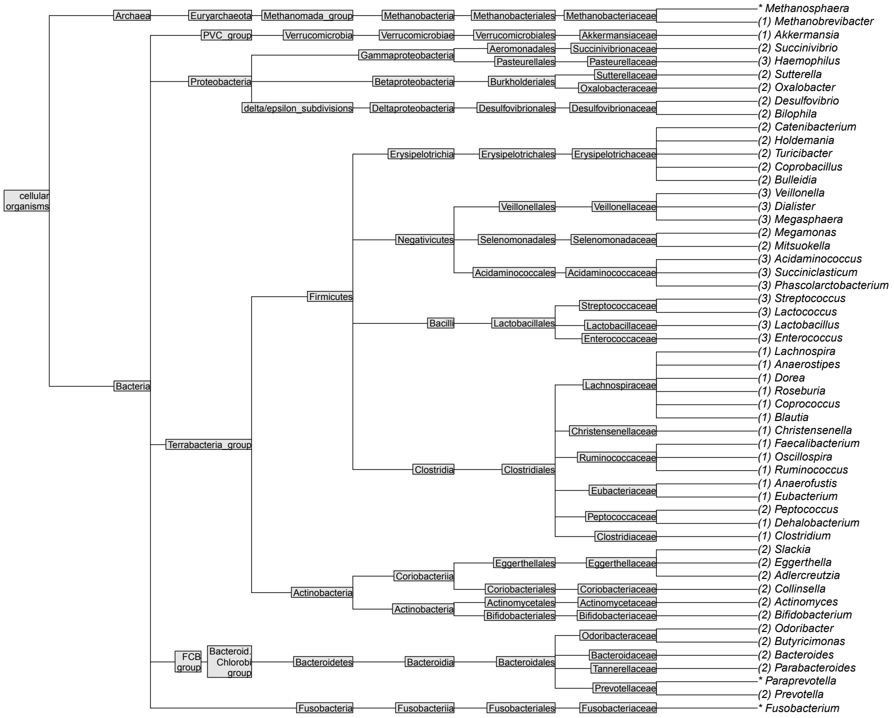

FIG 5. *The phylogenetic tree of 54 genera. Different numbers (colors) represent three microbial communities derived from the graph estimated by the proposed PhyloBCG; the asterisk indicates a stand-alone node. The color figure is available in the online version.*

leaves of the simulation true tree. In the second case, we mimic the later real data analysis by ignoring the simulation true divergence times and assuming the divergence times to be equally spaced. The results summarized in Supplementary Materials S5 indicate that PhyloBCG is robust to tree prior misspecification in that the posterior network estimates are not significantly affected by the tree misspecifications.

Under the same simulation settings, we also conduct additional simulations with the compositional version of the simulated data. Specifically, compositional data are obtained by row normalization, and then each method is applied to modified centered log-ratio transformed (Yoon et al., 2019) compositional data. For SPIEC-EASI, we use centered log-ratio transformed data as initially suggested by the authors. The average MCC, TPR, and FPR summarized in Figure 4 show almost identical trend as in Figure 3. Despite some mild deterioration on compositional data, PhyloBCG still outperforms all the other competing models, supporting the utility of phylogenetic information. Detailed simulation results are provided in Section S6 of Supplementary Materials.

5. Application to quantitative gut microbiome profiling data.

5.1. *Data and phylogenetic tree.* We focus on estimating genus-level association network of the QMP data (Vandeputte et al., 2017) that consists of $n = 106$ healthy subjects' gut microbiome. We use the data as processed in Yoon et al. (2019), which can be obtained from the R-package SPRING (Yoon, Gaynanova and Müller, 2019b). Among the 91 genera, there are 33 genera with missing names and 4 genera without available phylogenetic information on the National Center for Biotechnology Information (NCBI) database on which we based the phylogenetic tree construction. Consequently, we consider $p = 54$ genera and obtain their phylogenetic tree based on the NCBI taxonomy database using the platform PhyloT[1]. As the database does not provide divergence times of branches, we match the branch lengths with the taxonomic ranks so that each major taxonomic rank is equally spaced with the length of 1. We locate intermediate minor ranks in between corresponding major taxonomic ranks. For example, FCB group (superphylum) and Bacteroidetes-Chlorobi group (unranked) are located in between domain Bacteria and phylum Bacteroidetes as illustrated in Figure 5. This branch length specification is supported by our sensitivity analysis with respect to divergence times (provided in Section S5 of Supplementary Materials). These genera have up to 70% zeros (14 have more than 50% zeros).

5.2. *Analysis.* For PhyloBCG, we use the same hyperparameter values as in Section 4 and run 4 parallel Markov chains for 100,000 iterations after 25,000 burn-in iterations. On Apple M1 with 3.2 GHz maximum clock speed, the computation time of a single chain was 5 hours, 9 minutes, and 30 seconds, where $n = 106$ and $p = 54$. We then concatenate the 4 chains and obtain a posterior sample of size 10,000 by retaining every 40th iteration, from which we compute the posterior means $\widehat{\pi}_{jk}$ and $\widehat{\mathbf{\Omega}}$. The point estimate and the upper 97.5% confidence limit of the potential scale reduction factor (Gelman and Rubin, 1992) are 1.025 and 1.075, both within the acceptable range, below 1.1. (Gelman et al., 2013). The trace plot, provided in Section S7 of Supplementary Materials, also shows no evidence for lack of convergence. The goodness-of-fit on QMP data has been assessed by comparing marginal and joint posterior predictive densities with the observed data, where the proposed model shows a good fit (Supplementary Materials S3).

The microbial association network is estimated by controlling the posterior expected FDR at 0.1 which results in $c_{0.1} = 0.719$. SPIEC-EASI and SPRING are tuned using 100 sparsity parameter values. The default stability threshold (Liu, Roeder and Wasserman, 2010) is changed from 0.1 to 0.2 to avoid overly sparse network estimates. The recovered networks are shown in Figure 6.

5.2.1. *Overall network summary and interpretation.* We first compare estimated networks in terms of their density and community structure. The estimated network from PhyloBCG appears to be denser and have much more definitive communities than those from SPIEC-EASI and SPRING. While we do not know the true network and community structure of the gut microbiota in the study population, the following reasons support our belief that the additional findings of microbial interactions and communities from PhyloBCG are biologically meaningful.

First, microbes are known to form communities (Pflughoeft and Versalovic, 2012). Applying the edge proximity measure of Newman and Girvan (2004) to the estimated network by PhyloBCG, we find three evident microbial communities which are marked with different colors in Figures 5 and 6. Posterior mean latent positions, illustrated in Figure 7, also form three distinct clusters that consistently match the estimated microbial communities. Interestingly, the genera within each of these communities tend to share unique characteristics. On the one hand, most of the genera from the top two communities in Figure 5 are obligate anaerobes which only survive in the absence of oxygen. On the other hand, the community located at the bottom of Figure 5 contains genera with species that need or at least can tolerate oxygen. For example, *Streptococcus* and *Enterococcus* contain facultative anaerobic species that

[1] https://phylot.biobyte.de

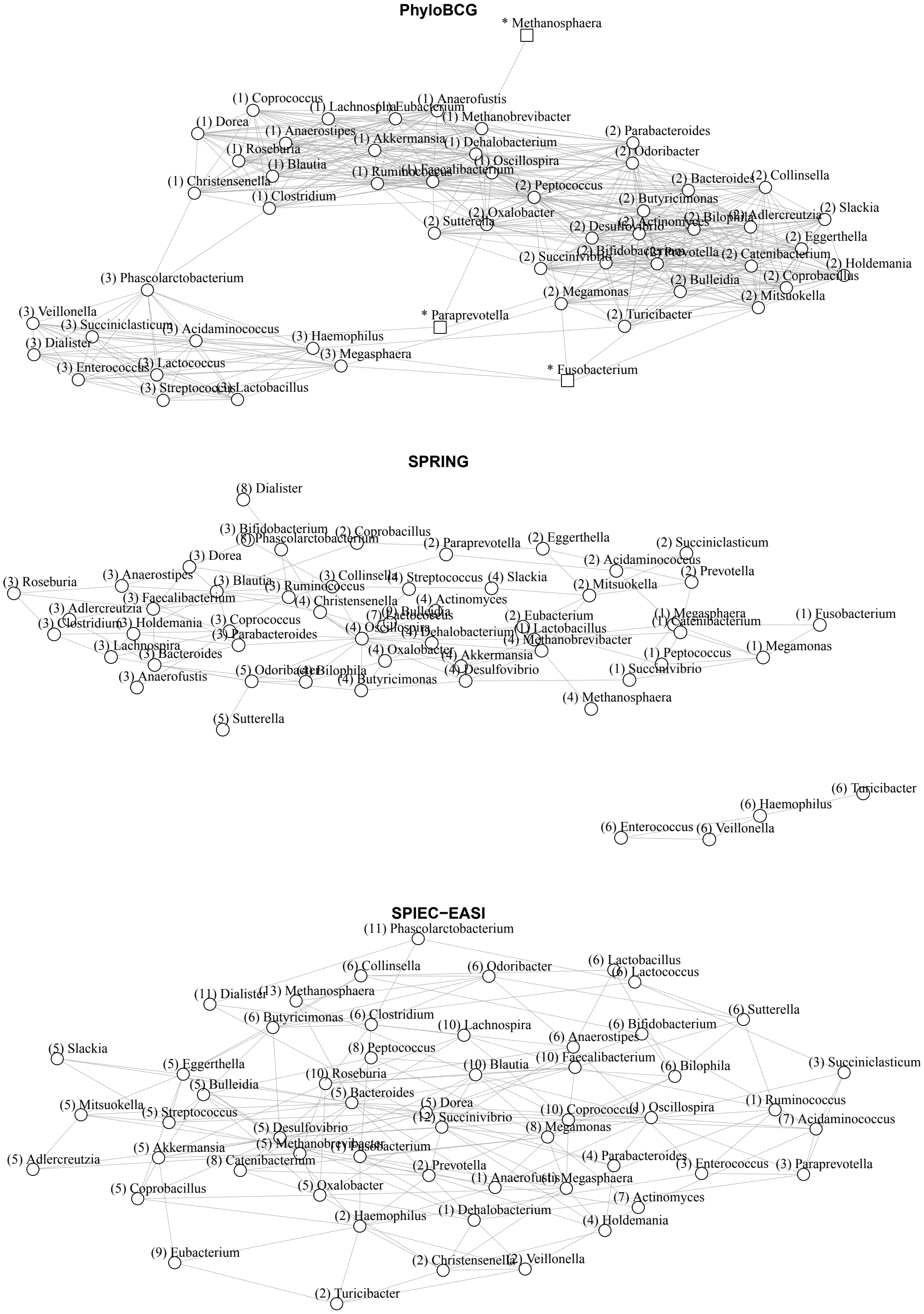

FIG 6. *Estimated microbial association networks of PhyloBCG, SPRING, and SPIEC-EASI with 54 genera in QMP data (Vandeputte et al., 2017). Communities are separately estimated with the graph estimate of each model, and the numbers represent the community memberships. The color figure is available in the online version.*

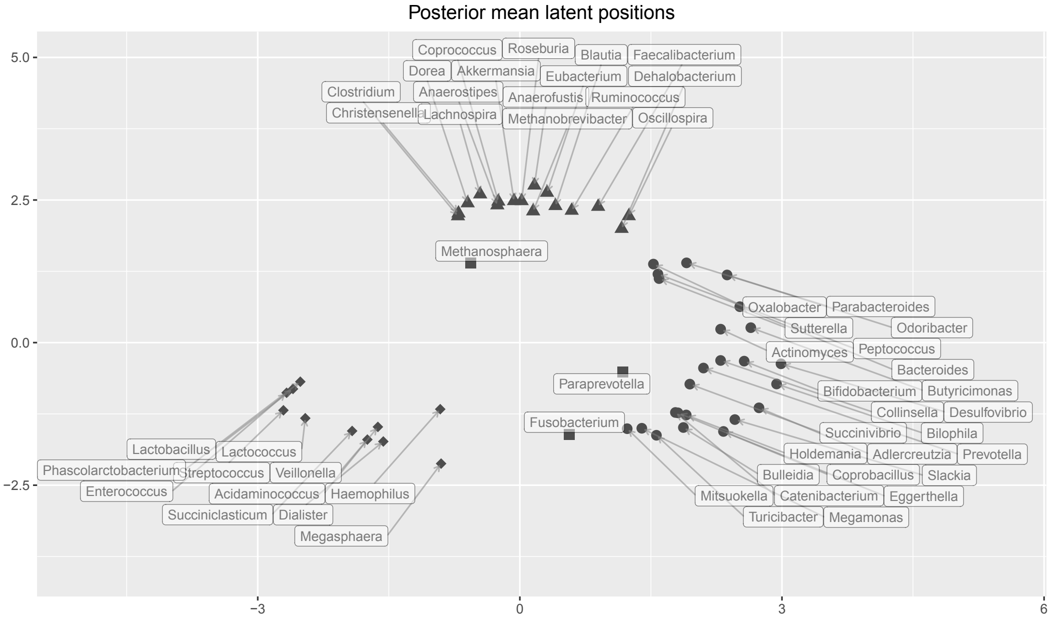

Fig 7. *Posterior mean latent positions of 54 genera in QMP data (Vandeputte et al., 2017). Community memberships estimated with the graph estimate of PhyloBCG are marked by symbols. Black squares represent stand alone genera. The color figure is available in the online version.*

are able to utilize oxygen as a source of energy, but can also generate energy anaerobically in an oxygen-deficient environment (Fisher and Phillips, 2009; Clewell, 1981). All the members of *Haemophilus* are facultatively anaerobic species or aerobic species, where aerobic species need oxygen to survive (Cooke and Slack, 2017). *Lactobacillus* contains aerotolerant and microaerophilic species (Zheng et al., 2020). Aerotolerant species do not need oxygen and use anaerobic fermentation to generate energy, but oxygen is not toxic to them. Microaerophiles need oxygen to survive but require low oxygen concentration to thrive and are damaged by high oxygen level, e.g., atmospheric oxygen level. Furthermore, the bottom community has very few interactions with the two top communities, possibly because of their distinct living environments. By contrast, the estimated microbial interaction networks from SPIEC-EASI and SPRING do not present obvious communities; this lack of community structure is, based on existing literature (Rohr and Bascompte, 2014; Peralta, 2016), unlikely.

Second, while the phylogenetic tree prior helps identify additional interactions and communities, it does not dictate the posterior inference. Some of the genera from the top right community in Figure 5 (e.g., *Succinivibrio* and *Prevotella*) are not phylogenetically similar to each other. Similarly, the other two communities also contains phylogenetically distant genera. This suggests that the phylogenetic tree prior does not override the information of associations that are strongly supported by the data. The comparison of the prior and posterior means of edge inclusion probabilities also indicates that the phylogenetic tree prior does not dominate the posterior inference (Supplement Materials S5).

Third, from the simulation study, we have seen that PhyloBCG is much more powerful in detecting interactions than SPIEC-EASI and SPRING especially when there is a clear community structure, while also having comparable FPR. Given that the communities of the estimated network by PhyloBCG seem biologically plausible, the additional interactions found by PhyloBCG are more likely to be true positives than false discoveries, some of which will be explained in detail in the next section.

TABLE 1

Selected pairs of genera with strong association as identified by PhyloBCG. "–" indicates that no significant partial correlation is found by the corresponding.

Pairs of Microbial Genera		Partial Correlations			Reference
		PhyloBCG	SPRING	SPIEC-EASI	
Dialister	Phascolarctobacterium	-0.384	-0.185	-0.252	Vandeputte et al. (2017), Naderpoor et al. (2019)
Oscillospira	Ruminococcus	0.129	0.320	0.335	Chen et al. (2020a)
Mitsuokella	Prevotella	0.155	0.152	–	Ramayo-Caldas et al. (2016)
Ruminococcus	Blautia	0.115	0.048	–	Ramayo-Caldas et al. (2016)
Oscillospira	Butyricimonas	0.155	0.335	–	Garcia-Mantrana et al. (2018), Thomaz et al. (2021)
Eubacterium	Peptococcus	0.178	0.083	–	Oh et al. (2020)
Bacteroides	Bilophila	0.137	0.221	–	Vandeputte et al. (2017)
Akkermansia	Methanobrevibacter	0.202	0.354	0.054	Vandeputte et al. (2016)
Blautia	Methanobrevibacter	-0.062	–	-0.329	Garcia-Mantrana et al. (2018), Müller et al. (2019)
Prevotella	Bacteroides	-0.107	–	-0.019	Vandeputte et al. (2017)
Veillonella	Streptococcus	0.149	–	–	Ley (2016), Johnson et al. (2017) Anbalagan et al. (2017), Chen et al. (2020b) van den Bogert et al. (2013), Egland, Palmer and Kolenbrander (2004) Zoetendal et al. (2012), van den Bogert et al. (2014)
Bifidobacterium	Holdemania	-0.167	–	–	Liu et al. (2017), Yang et al. (2018) Wang et al. (2020)

5.2.2. *Detailed explanation of interactions.* **All models** identify strong negative partial correlations between *Dialister* and *Phascolarctobacterium*. This finding is in agreement with multiple published results. The original QMP study (Vandeputte et al., 2017) finds a strong negative correlation between these two genera. Naderpoor et al. (2019) report that *Phascolarctobacterium* (*Dialister*) is positively (negatively) correlated with insulin sensitivity. Consistently, Pedrogo et al. (2018) indirectly observe a trade-off relationship between the two genera from obese groups. Besides, strong positive partial correlations between *Oscillospira* and *Ruminococcus* are found by all methods as well, which is consistent with the finding in Chen et al. (2020a).

PhyloBCG and SPRING detect relatively strong partial correlations for the pairs (*Mitsuokella*, *Prevotella*) and (*Ruminococcus*, *Blautia*). In the phylogenetic tree displayed in Figure 5, *Ruminococcus* and *Blautia* are relatively close to each other, being the members of the same order, Clostridales. *Mitsuokella* and *Prevotella* are, however, phylogenetically distant from each other, indicating that the QMP data present a strong association between them and that the tree prior of the proposed PhyloBCG does not dominate the inference. These findings agree with the network analysis of Ramayo-Caldas et al. (2016), where they also detect positive partial correlations in the pairs (*Mitsuokella*, *Prevotella*) and (*Ruminococcus*, *Blautia*).

Additionally, the pairs (*Oscillospira*, *Butyricimonas*) and (*Eubacterium*, *Peptococcus*) also show relatively strong partial correlations under the two models. Both pairs are known to be related to diet and leanness. Oh et al. (2020) discover positive correlations between body weight and each of *Eubacterium* and *Peptococcus*. Garcia-Mantrana et al. (2018) find negative correlations between unhealthy diet (high intake of saturated fat and refined carbohydrates) and each of *Oscillospira* and *Butyricimonas*. Furthermore, positive partial correlations for the pairs (*Bacteroides*, *Bilophila*) and (*Akkermansia*, *Methanobrevibacter*) are observed, where these results match the analyses in Vandeputte et al. (2016) and Vandeputte et al. (2017), respectively.

PhyloBCG and SPIEC-EASI find negative partial correlations for the pairs (*Blautia*, *Methanobrevibacter*) and (*Prevotella*, *Bacteroides*). *Blautia* and *Methanobrevibacter* are known to be positively and negatively related to dietary fiber intake, respectively (Garcia-Mantrana et al., 2018). Müller et al. (2019) suggest that their inverse relationship is possibly due to substrate competition as both use hydrogen as energy source.

For *Prevotella* and *Bacteroides*, Lozupone et al. (2012) find the trade-off between these two genera – carbohydrates (including simple sugars) focused diet increases *Prevotella* and decreases *Bacteroides* whereas protein and fat focused diet has the opposite effects on them. Their trade-off relationship is also discussed in Ley (2016) and Johnson et al. (2017). On the contrary, Vandeputte et al. (2017) claim that their negative association is an artifact of using compositional rather than quantitative microbiome data for analyses .

PhyloBCG uniquely discovers positive partial correlations for the pairs (*Veillonella*, *Streptoccocus*) and (*Bifidobacterium*, *Holdemania*). The estimated positive partial correlation between *Veillonella* and *Streptoccocus* is consistent with that of the gut microbial network analysis of Chen et al. (2020b). Anbalagan et al. (2017) demonstrate their mutualistic relationship: *Streptoccocus* uses glucose as a source of carbon and release lactic acid, whereas *Veillonella* utilizes lactic acid as carbon and energy source for growth. There are also many studies reporting their co-occurrence and mutualism (van den Bogert et al., 2013; Egland, Palmer and Kolenbrander, 2004; Zoetendal et al., 2012; van den Bogert et al., 2014). For *Bifidobacterium* and *Holdemania*, Liu et al. (2017) report that prebiotic supplement significantly increases relative abundance of beneficial *Bifidobacterium* and decreases *Holdemania*, where *Holdemania* is reported to be associated with unhealthy gut and antibiotic use (Yang et al., 2018). Wang et al. (2020) discuss the underlying mechanism of the trade-off relationship. In summary, we find these uniquely identified pairs by the proposed PhyloBCG to be well supported by existing literature. All the genera pairs discussed above are summarized in Table 1 with their estimated partial correlations and supporting references.

6. Discussion.

In this work, we propose a phylogenetically informed Bayesian truncated copula graphical model for estimating microbial association networks with QMP data. The proposed method explicitly accounts for the zero-inflated nature of the QMP data and incorporates the microbial evolutionary information through the diffusion process and latent position model.

Simulation study with various phylogenetic tree structures reveals that the phylogenetic prior significantly improves network estimation accuracy. In particular, the proposed model shows much larger true positive rates while having comparable false positive rates to existing microbial network estimation models. In application to the QMP data analysis, the proposed model identifies several unique genus-level conditional dependencies that are missed by existing microbial network estimation methods.

Our sensitivity analysis shows reasonably stable performance under various hyperparameter settings. Also, we find that the proposed model is robust to tree misspecification in that the phylogenetic tree does not override conditional dependence supported by data. Furthermore, the proposed model shows better network estimation accuracy than competing models, even with the misspecified tree. This supports the usefulness of the phylogenetic prior even when the information on divergence times is unavailable.

Although the proposed model is developed for microbial association network estimation, the established formulation of the truncated Gaussian copula can be directly applied to other zero-inflated data, such as single-cell RNA sequencing data. Furthermore, the framework for incorporating evolutionary information is not limited to undirected graph estimation and can be extended to the directed graph estimation models. The R code implementing the method and QMP data are available at `https://github.com/heech31/phyloBCG`.

Acknowledgements. The authors thank Grace Yoon for constructive discussions on QMP data analysis. This work has been partially supported by the Texas A&M Institute of Data Science (TAMIDS) and the Texas A&M Strategic Transformative Research Program.

Gaynanova's research was partially supported by the National Science Foundation (NSF CAREER DMS-2044823). Ni's research was partially supported by the National Science Foundation (NSF DMS-1918851, NSF DMS-2112943). Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing.

REFERENCES

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88** 669–679.

ANBALAGAN, R., SRIKANTH, P., MANI, M., BARANI, R., SESHADRI, K. G. and JANARTHANAN, R. (2017). Next generation sequencing of oral microbiota in Type 2 diabetes mellitus prior to and after neem stick usage and correlation with serum monocyte chemoattractant-1. *Diabetes Research and Clinical Practice* **130** 204–210.

BANERJEE, O., EL GHAOUI, L. and D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research* **9** 485–516.

CANI, P. D., VAN HUL, M., LEFORT, C., DEPOMMIER, C., RASTELLI, M. and EVERARD, A. (2019). Microbial regulation of organismal energy homeostasis. *Nature Metabolism* **1** 34–46.

CHEN, Y. R., ZHENG, H. M., ZHANG, G. X., CHEN, F. L., CHEN, L. D. and YANG, Z. C. (2020a). High Oscillospira abundance indicates constipation and low BMI in the Guangdong Gut Microbiome Project. *Scientific Reports* **10** 1–8.

CHEN, L., COLLIJ, V., JAEGER, M., VAN DEN MUNCKHOF, I. C., VILA, A. V., KURILSHIKOV, A., GACESA, R., SINHA, T., OOSTING, M., JOOSTEN, L. A. et al. (2020b). Gut microbial co-abundance networks show specificity in inflammatory bowel disease and obesity. *Nature Communications* **11** 1–12.

CHO, I. and BLASER, M. J. (2012). The human microbiome: at the interface of health and disease. *Nature Reviews Genetics* **13** 260–270.

CLEWELL, D. B. (1981). Plasmids, drug resistance, and gene transfer in the genus Streptococcus. *Microbiological Reviews* **45** 409.

COOKE, F. J. and SLACK, M. P. E. (2017). 183 - Gram-Negative Coccobacilli. In *Infectious Diseases* 4 ed. (J. Cohen, W. G. Powderly and S. M. Opal, eds.) 1611-1627.e1. Elsevier.

DAHL, J., VANDENBERGHE, L. and ROYCHOWDHURY, V. (2008). Covariance selection for nonchordal graphs via chordal embedding. *Optimization Methods & Software* **23** 501–520.

DOBRA, A., LENKOSKI, A. et al. (2011). Copula Gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics* **5** 969–993.

DOBRA, A., LENKOSKI, A. and RODRIGUEZ, A. (2011). Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association* **106** 1418–1433.

EGLAND, P. G., PALMER, R. J. and KOLENBRANDER, P. E. (2004). Interspecies communication in Streptococcus gordonii–Veillonella atypica biofilms: signaling in flow conditions requires juxtaposition. *Proceedings of the National Academy of Sciences* **101** 16917–16922.

FAN, J., LIU, H., NING, Y. and ZOU, H. (2017). High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **79** 405–421.

FISHER, K. and PHILLIPS, C. (2009). The ecology, epidemiology and virulence of Enterococcus. *Microbiology* **155** 1749–1757.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

GARCIA-MANTRANA, I., SELMA-ROYO, M., ALCANTARA, C. and COLLADO, M. C. (2018). Shifts on gut microbiota associated to mediterranean diet adherence and specific dietary intakes on general adult population. *Frontiers in Microbiology* **9** 890.

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science* **7** 457–472.

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2013). *Bayesian Data Analysis*, 3 ed. CRC Press.

GLOOR, G. B., MACKLAIM, J. M., PAWLOWSKY-GLAHN, V. and EGOZCUE, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Frontiers in Microbiology* **8** 2224.

HOFF, P. D. et al. (2007). Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics* **1** 265–283.

HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association* **97** 1090–1098.

JOHNSON, E. L., HEAVER, S. L., WALTERS, W. A. and LEY, R. E. (2017). Microbiome and metabolic disease: revisiting the bacterial phylum Bacteroidetes. *Journal of Molecular Medicine* **95** 1–8.

KIM, M., QIE, Y., PARK, J. and KIM, C. H. (2016). Gut microbial metabolites fuel host antibody responses. *Cell Host & Microbe* **20** 202–214.

KLAASSEN, C. A., WELLNER, J. A. et al. (1997). Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli* **3** 55–77.

KURTZ, Z. D., MÜLLER, C. L., MIRALDI, E. R., LITTMAN, D. R., BLASER, M. J. and BONNEAU, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Computational Biology* **11** e1004226.

KURTZ, Z. D., MÜLLER, C. L., MIRALDI, E. R., LITTMAN, D. R., BLASER, M. J. and BONNEAU, R. A. (2021). SpiecEasi: Sparse Inverse Covariance for Ecological Statistical Inference R package version 1.1.1.

LENKOSKI, A. and DOBRA, A. (2011). Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior. *Journal of Computational and Graphical Statistics* **20** 140–157.

LEY, R. E. (2016). Prevotella in the gut: choose carefully. *Nature Reviews Gastroenterology & Hepatology* **13** 69–70.

LIU, H., ROEDER, K. and WASSERMAN, L. (2010). Stability approach to regularization selection (StARS) for high dimensional graphical models. *Advances in Neural Information Processing Systems* **24** 1432.

LIU, H., HAN, F., YUAN, M., LAFFERTY, J., WASSERMAN, L. et al. (2012). High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics* **40** 2293–2326.

LIU, F., LI, P., CHEN, M., LUO, Y., PRABHAKAR, M., ZHENG, H., HE, Y., QI, Q., LONG, H., ZHANG, Y. et al. (2017). Fructooligosaccharide (FOS) and galactooligosaccharide (GOS) increase Bifidobacterium but reduce butyrate producing bacteria with adverse glycemic metabolism in healthy young population. *Scientific Reports* **7** 1–12.

LOZUPONE, C. A., STOMBAUGH, J. I., GORDON, J. I., JANSSON, J. K. and KNIGHT, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature* **489** 220–230.

LYNCH, S. V. and PEDERSEN, O. (2016). The human intestinal microbiome in health and disease. *New England Journal of Medicine* **375** 2369–2379.

MA, J. (2020). Joint Microbial and Metabolomic Network Estimation with the Censored Gaussian Graphical Model. *Statistics in Biosciences* 1–22.

MARTINEZ, K. B., PIERRE, J. F. and CHANG, E. B. (2016). The gut microbiota: the gateway to improved metabolism. *Gastroenterology Clinics* **45** 601–614.

MARTINY, J. B., JONES, S. E., LENNON, J. T. and MARTINY, A. C. (2015). Microbiomes in light of traits: a phylogenetic perspective. *Science* **350** 9323.

MATTHEWS, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure* **405** 442–451.

MCDAVID, A., GOTTARDO, R., SIMON, N. and DRTON, M. (2019). Graphical models for zero-inflated single cell gene expression. *The Annals of Applied Statistics* **13** 848–873.

MEINSHAUSEN, N., BÜHLMANN, P. et al. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34** 1436–1462.

MITRA, R., MÜLLER, P., LIANG, S., YUE, L. and JI, Y. (2013). A Bayesian graphical model for chip-seq data on histone modifications. *Journal of the American Statistical Association* **108** 69–80.

MULGRAVE, J. J., GHOSAL, S. et al. (2020). Bayesian inference in nonparanormal graphical models. *Bayesian Analysis* **15** 449–475.

MÜLLER, M., HERMES, G. D., CANFORA, E. E., SMIDT, H., MASCLEE, A. A., ZOETENDAL, E. G. and BLAAK, E. E. (2019). Distal colonic transit is linked to gut microbiota diversity and microbial fermentation in humans with slow colonic transit. *American Journal of Physiology-Gastrointestinal and Liver Physiology* **318** G361–G369.

NADERPOOR, N., MOUSA, A., GOMEZ-ARANGO, L. F., BARRETT, H. L., DEKKER NITERT, M. and DE COURTEN, B. (2019). Faecal microbiota are related to insulin sensitivity and secretion in overweight or obese adults. *Journal of Clinical Medicine* **8** 452.

NEWMAN, M. E. and GIRVAN, M. (2004). Finding and evaluating community structure in networks. *Physical Review E* **69** 026113.

OH, J. K., CHAE, J. P., PAJARILLO, E. A. B., KIM, S. H., KWAK, M.-J., EUN, J.-S., CHEE, S. W., WHANG, K.-Y., KIM, S.-H. and KANG, D.-K. (2020). Association between the body weight of growing pigs and the functional capacity of their gut microbiota. *Animal Science Journal* **91** e13418.

OSBORNE, N., PETERSON, C. B. and VANNUCCI, M. (2021). Latent nettwork estimation and variable selection for compositional data via variational EM. *Journal of Computational and Graphical Statistics* Ahead-of-print.

Paradis, E. (2012). *Analysis of Phylogenetics and Evolution with R* **2**. Springer.

Paradis, E. and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R R package version 5.4.1.

Pedrogo, D. A. M., Jensen, M. D., Van Dyke, C. T., Murray, J. A., Woods, J. A., Chen, J., Kashyap, P. C. and Nehra, V. (2018). Gut microbial carbohydrate metabolism hinders weight loss in overweight adults undergoing lifestyle intervention with a volumetric diet. In *Mayo Clinic Proceedings* **93** 1104–1110. Elsevier.

Peralta, G. (2016). Merging evolutionary history into species interaction networks. *Functional Ecology* **30** 1917–1925.

Peterson, C., Stingo, F. C. and Vannucci, M. (2015). Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association* **110** 159–174.

Pflughoeft, K. J. and Versalovic, J. (2012). Human microbiome in health and disease. *Annual Review of Pathology: Mechanisms of Disease* **7** 99–122.

Ramayo-Caldas, Y., Mach, N., Lepage, P., Levenez, F., Denis, C., Lemonnier, G., Leplat, J.-J., Billon, Y., Berri, M., Doré, J. et al. (2016). Phylogenetic network analysis applied to pig gut microbiota identifies an ecosystem structure linked with growth traits. *The ISME Journal* **10** 2973–2977.

Rohr, R. P. and Bascompte, J. (2014). Components of phylogenetic signal in antagonistic and mutualistic networks. *The American Naturalist* **184** 556-564.

Roverato, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics* **29** 391–411.

Thomaz, F. S., Altemani, F., Panchal, S. K., Worrall, S. and Nitert, M. D. (2021). The influence of wasabi on the gut microbiota of high-carbohydrate, high-fat diet-induced hypertensive Wistar rats. *Journal of Human Hypertension* **35** 170–180.

van den Bogert, B., Erkus, O., Boekhorst, J., Goffau, d. M., Smid, E. J., Zoetendal, E. G. and Kleerebezem, M. (2013). Diversity of human small intestinal Streptococcus and Veillonella populations. *FEMS Microbiology Ecology* **85** 376–388.

van den Bogert, B., Meijerink, M., Zoetendal, E. G., Wells, J. M. and Kleerebezem, M. (2014). Immunomodulatory properties of Streptococcus and Veillonella isolates from the human small intestine microbiota. *PloS One* **9** e114277.

Vandeputte, D., Falony, G., Vieira-Silva, S., Tito, R. Y., Joossens, M. and Raes, J. (2016). Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut* **65** 57–62.

Vandeputte, D., Kathagen, G., D'hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., Wang, J., Tito, R. Y., De Commer, L., Darzi, Y. et al. (2017). Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551** 507–511.

Wang, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis* **7** 867–886.

Wang, H. (2015). Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Analysis* **10** 351–377.

Wang, S., Xiao, Y., Tian, F., Zhao, J., Zhang, H., Zhai, Q. and Chen, W. (2020). Rational use of prebiotics for gut microbiota alterations: Specific bacterial phylotypes and related mechanisms. *Journal of Functional Foods* **66** 103838.

Wasserman, S., Faust, K. et al. (1994). Social network analysis: Methods and applications.

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

Yang, X., Yin, F., Yang, Y., Lepp, D., Yu, H., Ruan, Z., Yang, C., Yin, Y., Hou, Y., Leeson, S. et al. (2018). Dietary butyrate glycerides modulate intestinal microbiota composition and serum metabolites in broilers. *Scientific Reports* **8** 1–12.

Yoon, G., Carroll, R. J. and Gaynanova, I. (2020). Sparse semiparametric canonical correlation analysis for data of mixed types. *Biometrika* **107** 609–625.

Yoon, G., Gaynanova, I. and Müller, C. L. (2019a). Microbial Networks in SPRING - Semi-parametric Rank-Based Correlation and Partial Correlation Estimation for Quantitative Microbiome Data. *Frontiers in Genetics* **10** 516.

Yoon, G., Gaynanova, I. and Müller, C. L. (2019b). Semi-parametric Rank-Based Correlation and Partial Correlation Estimation for Quantitative Microbiome Data R package version 1.0.4.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35.

Zhang, S. and Chen, D.-C. (2019). Facing a new challenge: the adverse effects of antibiotics on gut microbiota and host immunity. *Chinese Medical Journal* **132** 1135.

ZHENG, J., WITTOUCK, S., SALVETTI, E., FRANZ, C. M., HARRIS, H. M., MATTARELLI, P., O'TOOLE, P. W., POT, B., VANDAMME, P., WALTER, J. et al. (2020). A taxonomic note on the genus Lactobacillus: Description of 23 novel genera, emended description of the genus Lactobacillus Beijerinck 1901, and union of Lactobacillaceae and Leuconostocaceae. *International Journal of Systematic and Evolutionary Microbiology* **70** 2782–2858.

ZHOU, J., VILES, W. D., LU, B., LI, Z., MADAN, J. C., KARAGAS, M. R., GUI, J. and HOEN, A. G. (2020). Identification of microbial interaction network: zero-inflated latent Ising model based approach. *BioData Mining* **13** 1–15.

ZHOU, F., HE, K., LI, Q., CHAPKIN, R. S. and NI, Y. (2021). Bayesian biclustering for microbial metagenomic sequencing data via multinomial matrix factorization. *Biostatistics* **00** 1–19.

ZOETENDAL, E. G., RAES, J., VAN DEN BOGERT, B., ARUMUGAM, M., BOOIJINK, C. C., TROOST, F. J., BORK, P., WELS, M., DE VOS, W. M. and KLEEREBEZEM, M. (2012). The human small intestinal microbiota is driven by rapid uptake and conversion of simple carbohydrates. *The ISME Journal* **6** 1415–1426.