Aggregating statistically correlated metabolic pathways into groups to improve pathway prediction outcomes

Abdur Rahman M. A. Basher[1][a] and Steven J. Hallam[1,2][b]

[1]*Graduate Program in Bioinformatics, University of British Columbia, Vancouver, BC V5Z 4S6, Canada*

[2]*Department of Microbiology & Immunology, University of British Columbia, Vancouver, BC V6T 1Z3, Canada*

arbasher@student.ubc.ca, shallam@mail.ubc.ca

Keywords: pathway group, relabeling, data augmentation, correlated models, metabolic pathway prediction, MetaCyc

Abstract: Given unprecedented amounts of (meta)genomic data, generated by NGS technologies, the computational tools for reproducible data analysis became widely developed and applied. However, many significant challenges are raised for accurate data elucidation due to noises inherited from the upstream pipelines. Especially for metabolic pathway prediction, the current methods are limited as they depend critically on clean data while, in practice, the pathways are usually error-prone. In this paper, we propose two novel hierarchical mixture frameworks, SOAP (sparse correlated pathway group) and SPREAT (distributed sparse correlated pathway group), to characterize pathways. The generic idea is to incorporate pathway abundance to encode each organismal genome as mixture distributions of groups, and each group, in turn, is a mixture of pathways. Moreover, both models deal with missing potential pathways in the training set by augmenting supplementary pathways into the learning framework as part of noise reduction efforts. However, introducing supplementary pathways may lead to the overestimation of some pathways, therefore, dual sparseness is applied for group distribution over data and pathway distribution on groups. Empirical studies with regard to groups correlations and pathway predictions show promising results, hence, providing insights to adapting this novel approach for the metabolic pathway prediction problem.

1 INTRODUCTION

The advances in the high-throughput sequencing technologies have enabled fast and cost-effective generation of petabytes of metatranscriptomics (Bashiardes et al., 2016), metagenomics (Wang et al., 2015), metaproteomics (Hassa et al., 2018), and metabolomics (Aguiar-Pulido et al., 2016) data with exquisite resolution and accuracy (Loh et al., 2012). With the ubiquitous availability of large omics datasets, the development of more sophisticated bioinformatics tools to interpret and discern knowledge from these data has seen to be rapidly evolving. In particular, computational tools for predicting metabolic pathways have seen substantial developments. A metabolic pathway is a chain of chemical reactions occurring in a cell, often catalyzed and coordinated by a group of enzymes, where metabolites are built-up or broken down for cellular processes. This is very useful for elucidating the biological process of an unknown organism. As such, over the past two decades many pathway prediction tools were developed to recover pathways (Mascher et al., 2019) (Baranwal et al., 2020) (Yamanishi et al., 2015) (Tabei et al., 2016) (Ye and Doak, 2009) (Dale et al., 2010) (Karp et al., 2016) (M. A. Basher et al., 2020) (M. A. Basher et al., 2021b). While these tools rely on reference metabolic pathway databases (e.g., MetaCyc (Caspi et al., 2019) and KEGG (Kanehisa et al., 2017)) to reconstruct pathways, other computational methods ignore the use of reference database or follow an agnostic approach through neglecting the pathway boundary into the reconstruction process (Zhao et al., 2012) (Qi et al., 2014) (Shafiei et al., 2014) (Jiao et al., 2013). Moreover, many of these tools solve the pathway prediction problem from the perspective of the metabolome while few works consider enzymes, encoded in genomic sequence information, as inputs to pathway prediction algorithms.

Among recently devolved pathway prediction tools is triUMPF (M. A. Basher et al., 2021b) which uses several layers of interactions among pathways and enzymes within a network to improve the

[a] https://orcid.org/0000-0002-3407-1187

[b] https://orcid.org/0000-0002-4889-6876

precision of pathway predictions in terms of communities represented by a cluster of nodes (pathways and enzymes). Despite triUMPF's predictive gains, its performance on metabolic pathway datasets (M. A. Basher et al., 2020) left extensive room for improvement. This is because triUMPF's prediction process depends on the quality of communities detected from both pathway and enzyme networks that are learned from pathway datasets which contain many missing pathways for organisms. This calls for developing a more flexible data-driven approach to reconstructing pathways that can reduce noise to a greater extent.

Previously, Shafiei and colleagues developed BiomeNet (Shafiei et al., 2014), an extension of MetaNetSim (Jiao et al., 2013), which is a hierarchical Bayesian network to reconstruct metabolic networks purely in a data-driven manner through leveraging enzymes abundances present in metagenomic data. Instead of relying on the arbitrary boundary of pathways, BiomeNet discovers functions that are referred to as subnetworks, where a subnetwork constitutes a group of connected reactions. Applications of BiomeNet to the human gut microbiome revealed distinct subnetworks associated with inflammatory bowel disease (IBD) patients. Moreover, BiomeNet was able to detect subnetworks that are common among healthy and IBD microbiome patients. These promising results suggest that one may perform functional assessments across communities at the reactome level and then project reactions with abundances onto reference pathways to get "pathway abundances" that help gain insights into the differences in metabolic pathways of those communities.

Inspired by BiomeNet, in this paper, we present a package CHAP (correlated pathway-group) comprise of three correlated mixed-membership hierarchical Bayesian models, CTM (Blei and Lafferty, 2006), SOAP, and SPREAT, to capture mixed components given pathway abundance data. The component is referred to as a "pathway group", which is comprised of a set of correlated pathways while pathways are permitted to be inter-mixed across groups with different proportions, resulting in overlapping pathways on groups. Modeling explicitly correlations among pathways, using a Gaussian covariance matrix, is fundamental as functions of similar organisms or communities are shared. Moreover, due to noise or missing probable pathways in pathway abundance data, both SOAP and SPREAT incorporate "background" pathways to supplement missing pathways in the modeling process. Also, both models apply dual sparseness, where each example in the pathway abundance data is represented by a few focused mixing groups and

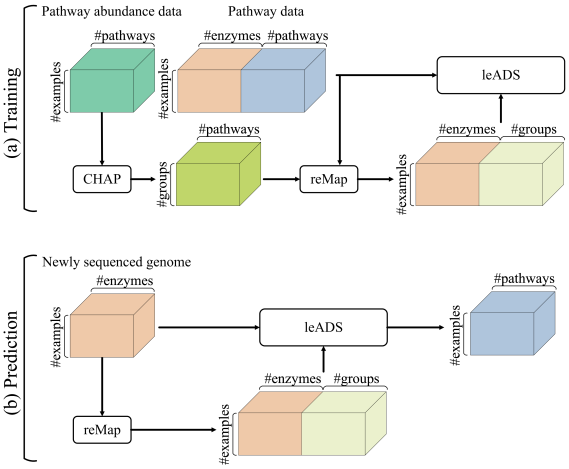


Figure 1: Group-based pathway prediction workflow. The training phase (a) takes pathway abundance data to discover groups using any correlated models in the CHAP package. Then, groups are used to map examples in the pathway abundance data to groups using reMap. Then, the results of this mapping are used in leADS, along with pathway data, to learn the model. After training, pathways can be predicted for a newly sequenced genome (b), by first inferring groups using reMap, and then apply the pretrained leADS to predict pathways from groups.

each pathway group consists of a few relevant pathways. These last two properties were not included in CTM. By modeling examples as mixing groups, one may use results from correlated models for the downstream group-based pathway prediction (Fig. 1).

Using SOAP, SPREAT, and CTM, we evaluated groups performances for the metabolic pathway prediction, and results were compared with two heuristic or rule-based algorithms: MinPath (Ye and Doak, 2009) and PathoLogic (Karp et al., 2016), and two machine learning methods: mlLGPR (M. A. Basher et al., 2020) and triUMPF (M. A. Basher et al., 2021a) on a set of Tier 1 (T1) pathway genome databases (PGDBs) and genomes used in the Critical Assessment of Metagenome Interpretation (CAMI) initiative (Sczyrba et al., 2017) following the genomic information hierarchy benchmarks initially developed for mlLGPR enabling more robust comparison between pathway prediction methods (M. A. Basher et al., 2020). Our novel group-based prediction strategy demonstrated that this approach improves pathway prediction outcomes.

2 CORRELATED MODELS

In this section, we present three correlated pathway models: i)-CTM (correlated topic model) (Blei and Lafferty, 2006), ii)- SOAP (sparse correlated

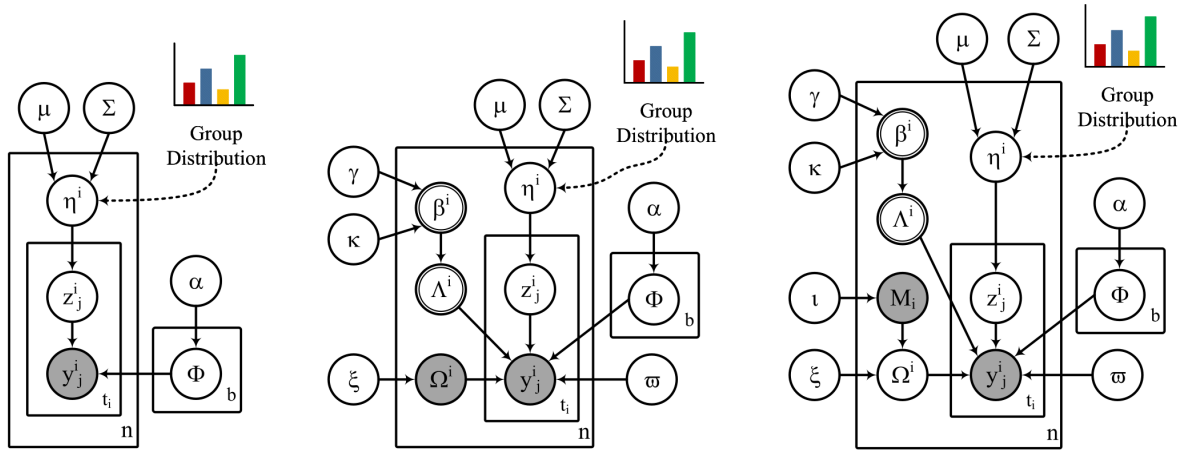


Figure 2: Graphical model representation of the correlated group models. The boxes are "plates" representing replicates. The outer plate represents examples, while the inner plate represents the repeated choice of pathways for an example. The logistic normal distribution, used to model the latent group proportions for an example, captures correlations among groups that are impossible to capture using a single Dirichlet. The observed data for each example i are a set of annotated pathways $\mathbf{y}^{(i)}$ and a set of hypothetical pathways $\mathbf{M}_i$. The hidden variables are: per-example group proportions $\eta^{(i)}$, per-example group selection parameters $\Lambda^{(i)}$, per-example hypothetical pathway distributions $\Omega^{(i)}$, per-pathway group assignment parameter $z_j^{(i)}$, and per-group distribution over pathways Φ_a.

pathway group) and iii)- SPREAT (distributed sparse correlated pathway group). These models incorporate pathway abundance information to encode each example as a mixture distribution of groups, and each pathway group, in turn, is a mixture of pathways with different mixing proportions. The pathway abundance information can be obtained by mapping enzyme –with abundances– onto a reference pathway database (e.g. MetaCyc (Caspi et al., 2019)). Before we discuss these three models, let us first provide some definitions and notations.

Pathway Abundance Data. Let $\mathcal{P} = \{\mathbf{y}^{(i)} : 1 < i \leqslant n\}$ be a collection of n examples corresponding organismal or multi-organismal genomes (e.g. Escherichia coli K-12), where each example $\mathbf{y}^{(i)} = (y_1^{(i)}, y_1^{(i)}, \ldots, y_t^{(i)})$ is a vector encoding the unnormalized abundance information of pathways and t is the pathway size. Let $\mathcal{Y} = \{h_1, h_2, \ldots, h_t\}$ be a set of all known metabolic pathways obtained from a reference database (e.g., MetaCyc (Caspi et al., 2019)), and $\mathcal{Y}_i \subseteq \mathcal{Y}$ corresponds to a subset of true pathways associated with the ith example.

Group Modeling. Given $\mathcal{P}$, a pathway group distribution for the ith example is a multinomial distribution vector, denoted by $\eta^{(i)}$ of size b groups, i.e., $\{p(\Phi_a|\eta^{(i)})\}_{a=1}^{a=b}$, where Φ_j in a multinomial pathway distribution over the group j, i.e., $\{p(y_k|\Phi_j)\}_{k=1}^{k=t}$. The overall goal of group modeling is to discover b hidden groups for each example.

The above definition states that pathways are distributed over groups, which entails correlation between groups.

Group Correlation. Given $\mathcal{P}$, the pairwise group-correlation is defined by a Gaussian covariance matrix, denoted by Σ. Each entry $s_{i,j}$ in Σ characterizes the i-th pathway group associated with the group j, where a larger score indicates both pathway groups are highly correlated.

Missing pathways in $\mathcal{P}$ are common for genomes representing simple or complex microbial communities. Previously, Hanson and colleagues (Hanson et al., 2014) reported missing a set of potential pathways for the Hawaii Ocean Time-series data (Stewart et al., 2011), such as *tricarboxylic acid cycle (TCA)*. These missing pathways have negative implications in group modeling as $\mathcal{P}$, in this case, would be exposed to extreme noise. Although manually incorporating missing pathways to $\mathcal{P}$ may provide a solution to model groups, given the dynamic nature of pathway prediction and discovery this treatment may exacerbate false discovery. Because there exist situations where a set of pathways that were identified earlier as a putative set, may be triggered as spurious in the subsequent stage of experimental studies (e.g. falsely annotating *ascorbate and aldarate metabolism pathway* for human in the KEGG database (Kanehisa et al., 2017) (Ye and Doak, 2009)). A good compromise would be to record those missing pathways

for $a \in \{1,\ldots,b\}$ **do**
  Sample a distribution over pathways
  $\Phi_a \sim \text{Dir}(.|\alpha)$;
for $i \in \{1,\ldots,n\}$ **do**
  Draw per example group weight
  $\eta^{(i)} \sim \mathcal{N}(.|\mu,\Sigma)$;
  Draw group proportions $\theta^{(i)} = \text{softmax}(\eta^{(i)})$;
  for $j \in \{1,\ldots,t^{(i)}\}$ **do**
    Sample a group assignment
    $z_j^{(i)} \sim \text{Mult}(.|\theta^{(i)})$;
    Sample a pathway $y_j^{(i)} \sim \text{Mult}(.|\Phi_{z_j^{(i)}})$;

Algorithm 1: The generative process for CTM given a collection of examples

in a separate list while keeping the original pathway abundance data intact for further investigation. Lets us denote $\mathbf{M} \in \mathbb{Z}_{\geq 0}^{n \times t}$ a matrix holding a set of missing pathways where each entry is an integer value indicating the abundance of a pathway for an example. This matrix is called the *background* or the *supplementary* matrix. Now, with these definitions, let us state our research problem.

Problem Statement. Given $\mathcal{P}$ and $\mathbf{M}$, the goal is to recover the group distribution η for each example such that applying group based metabolic pathway prediction would recover more accurate pathways for single or multi-organismal genomes.

2.1 Correlated Topic Model

The correlated topic model (CTM) is a probabilistic graphical model that extends the generative story of latent Dirichlet allocation (LDA) (Blei et al., 2003) to incorporate correlation among groups (or topics in the original paper). Fig. 2a shows the Bayesian graphical model for CTM using plate notation. Like LDA, CTM is composed of a hierarchical Bayesian mixture model, where features (words in the original paper) are mixed to constitute groups that are assumed to be correlated modeled by a Gaussian covariance matrix. Note that in this paper, we use the terms *feature* and *pathway* interchangeably.

Formally, let n be the total number of examples in $\mathcal{P}$, where each example i consists of features, i.e., $\mathbf{y}^{(i)}$. Then, the generative process for CTM is described as follows. First, we draw a multinomial feature distribution Φ_a from a Dirichlet prior $\alpha > \mathbb{R}_{>0}$ for each group $a \in \{1,\ldots,b\}$. Then, for each example i, a Gaussian random variable is drawn $\eta^{(i)} \sim \mathcal{N}(\mu,\Sigma)$, where μ is a b dimensional mean and $\Sigma \in \mathbb{R}^{b \times b}$ is the covariance matrix. The random variable $\eta^{(i)}$ is projected onto the probability simplex to obtain the group

for $a \in \{1,\ldots,b\}$ **do**
  Sample a distribution over pathways
  $\Phi_a \sim \text{Dir}(.|\alpha)$;
for $i \in \{1,\ldots,n\}$ **do**
  Draw per example group weight
  $\eta^{(i)} \sim \mathcal{N}(.|\mu,\Sigma)$;
  Draw group proportions $\theta^{(i)} = \text{softmax}(\eta^{(i)})$;
  Draw beta distribution $\beta^{(i)} \sim \text{Beta}(.|\gamma,\kappa)$;
  Draw a sparsity indicator vector
  $\Lambda^{(i)} \sim \text{Bernoulli}(.|\beta^{(i)})$;
  if *SPREAT* **then**
    Sample a vector $\mathbf{M}_i \sim \text{Prior}(.|\iota)$;
    Sample background distribution
    $\Omega^{(i)}|\mathbf{M}_i \sim \text{Dir}(.|\xi)$;
  else
    Draw background feature proportions
    $\Omega^{(i)} \sim \text{Dir}(.|\xi)$;
  for $j \in \{1,\ldots,t^{(i)}\}$ **do**
    Sample a group assignment
    $z_j^{(i)} \sim \text{Mult}(.|\Lambda^{(i)} \odot \theta^{(i)})$;
    Sample a pathway
    $y_j^{(i)} \sim \text{Mult}(.|(1-\Omega^{(i)}_{z_j^{(i)}}) \odot \Phi_{z_j^{(i)}})$;

Algorithm 2: The generative process for SOAP and SPREAT

distributions $\theta^{(i)} = \text{softmax}(\eta^{(i)})$, corresponding the logistic-normal distribution, from which a group indicator $z_j^{(i)} \in \{1,\ldots,b\}$ is sampled. Finally, each observed feature $j \in \{1,\ldots,t^i\}$ is drawn from the associated feature distribution, indicated by it's group assignment z_j, i.e., $y_j^{(i)} \sim \Phi_{z_j^{(i)}}$. This generative process (Algorithm 1) is identical to LDA except that the group distributions is sampled from the logistic normal instead from a Dirichlet prior as in LDA.

2.2 Correlated Pathway-Group Model

Correlated pathway group models are extensions to CTM: i)- SOAP (Fig. 2b) and ii)- SPREAT (Fig. 2c). Both models incorporate dual sparseness and supplementary pathways in modeling group proportions. The two properties were not implemented in CTM. Let us discuss these two models.

Analogous to CTM, given n number of examples in $\mathcal{P}$ and a matrix encoding missing pathways $\mathbf{M}$, the generative process for SOAP and SPREAT can be described as follows. First, we draw a multinomial pathway distribution Φ_a from asymmetric Dirichlet prior $\alpha \in \mathbb{R}_{>0}$ for each group $a \in \{1,...,b\}$, where b is assumed to be known and fixed in advance. The symmetric assumption is appropriate, in such a scenario,

because our prior knowledge, associated with these pathways, is inaccessible. For each example i, a group proportion is drawn $\theta^{(i)} = \text{softmax}(\eta^{(i)})$, where $\eta^{(i)}$ is a Gaussian random variable with mean and covariance are denoted by μ and Σ, respectability.

To sample a group, it is reasonable to expect that: i)- each example is usually explained with a handful set of a mixed proportion of groups and ii)- a group should consist only of a few related pathways. Therefore, we apply dual sparsity (Lin et al., 2014) (Airoldi et al., 2008) (Bien and Tibshirani, 2011) (He et al., 2017) to retain those relevant focused groups and pathways by: i)- introducing an auxiliary Bernoulli variable $\Lambda^{(i)}$ of size b to determine whether a group is selected for the ith example or ignored and ii)- applying a cutoff threshold to keep top $k \ll t$ pathways, based on their probabilities, for each group. Instead of sampling each entry in $\Lambda^{(i)}$ directly from a Bernoulli coin toss, we assume that each entry is sampled from a Beta distribution $\beta^{(i)}$, parameterized by two hyperparameters $\gamma \in \mathbb{R}_{>0}$ and $\kappa \in \mathbb{R}_{>0}$. Applying this dual sparsity, we aim to enhance the interpretability of the learned pathway groups while reducing the negative correlation among groups on Σ.

Next, a group indicator $z_j^{(i)} \in \{1, ..., b\}$ is drawn according to the example-specific mixture proportion $\Lambda^{(i)} \odot \theta^{(i)}$, where $\odot$ represents the Hadamard product. Now each pathway $y_j^{(i)}$ for the ith example is generated from a weighted distribution $\Omega^{(i)}_{z_j^{(i)}} \odot \Phi_{z_j^{(i)}}$ using a smoothing prior $\varpi \in \mathbb{R}_{>0}$. The parameter $\Omega^{(i)} \in \mathbb{R}^t$, derived from $\mathbf{M}_i$, represents a normalized supplementary pathway of size t, which is assumed to be drawn from a symmetric Dirichlet prior $\xi \in \mathbb{R}_{>0}$. For SPREAT, this parameter encodes distribution, where each element of $\Omega_j^{(i)}$ corresponds to the pathway probability $y_j^{(i)} \in \mathbf{M}_i$ for ith example. Here, the background pathway is assumed to be drawn from a sparse binary vector prior $\iota \in \mathbb{R}_{>0}$ that is included for completeness because pathways in $\mathbf{M}$ for each example are known.

Representing SOAP and SPREAT as layer-wise mixing components supports the hierarchical modularity of metabolic pathway generation, where the components of one level (e.g., pathways) permit to contribute to other structures with different degrees of granularity. The generative process of SOAP and SPREAT models is summarized in Algorithm 2. Notice that by setting all entries in Ω, Λ, and ϖ to 1, SOAP and SPREAT are reduced to CTM ("collapse2ctm" or c2m), reflecting the flexibility of these models.

Table 1: Correspondence between variational and original parameters.

Original parameter	Φ	μ	Σ	Λ	Ω	z
Variational parameter	ϕ	ν	ζ^2	λ	ω	ς

3 INFERENCE AND PARAMETER ESTIMATION FOR SPREAT

Here, we discuss the inference for the SPREAT model. Similar expression is straightforward to derive for SOAP. Given $\mathcal{P}$, the goal of inference is to compute the posterior distribution of per-example group proportions (η), per-example group selection parameters (Λ) and the associated beta distributions (β), per-example background pathway distributions (Ω), per-pathway group assignment (z), and per-group distribution over pathways (Φ). By denoting all parameters as Θ and variables as $\mathbf{V}$ while omitting hyperparameters, we apply the Jensen's inequality on a variational distribution over hidden variables $q(\Theta, \mathbf{V})$ to obtain the evidence lower bound (ELBO) as:

$$\mathcal{L}(q) = \mathbb{E}_q[\log p(\mathbf{Y}, \mathbf{M}, \Theta, \mathbf{V})] + \mathbb{H}(q) \tag{3.1}$$

where $p(\mathbf{Y}, \mathbf{M}, \Theta, \mathbf{V})$ represents the joint distribution of all observed and latent variables of the model. The ELBO contains two terms. The first term, $\mathbb{E}_q[\log p(\mathbf{Y}, \mathbf{M}, \Theta, \mathbf{V})]$, captures how well $q(\Theta, \mathbf{V})$ describes a distribution of the model. The second term is the entropy of the variational distribution, $\mathbb{E}_q[-\log q(\Theta, \mathbf{V})]$, which protects the variational distribution from "overfitting". The two terms depends on $q(\Theta, \mathbf{V})$ which is defined as:

$$\begin{aligned} q(\Theta, \mathbf{V}) = \prod_{a=1}^{b} q(\Phi_a|\phi_a) \Bigg[\prod_{i=1}^{n} q(\eta^{(i)}|\nu, \zeta^2) \\ \times q(\Lambda^{(i)}|\lambda^{(i)}) q(\Omega^{(i)}|\omega^{(i)}) \prod_{j=1}^{j=t_i} q(z_j^{(i)}|\varsigma_j^{(i)}) \Bigg] \end{aligned} \tag{3.2}$$

where $\phi, \nu, \zeta^2, \lambda, \omega$ and ς are variational free parameters. Table 1 shows the correspondence between variational and the original parameters. Now, the first term in Eq. 3.1 is decomposed into:

$$\begin{aligned} \mathbb{E}_q[\log p(\mathbf{Y}, \mathbf{M}, \Theta, V)] = \sum_{a=1}^{a=b} \mathbb{E}_q[\log p(\Phi_a|\alpha)] \\ + \sum_{i=1}^{i=n} \Bigg(\mathbb{E}_q[\log p(\eta|\mu, \Sigma)] + \mathbb{E}_q[\log p(\Lambda^{(i)}|\beta^i)] \\ + \mathbb{E}_q[\log p(\beta^i|\gamma, \kappa)] + \mathbb{E}_q[\log p(\Omega^{(i)}|\mathbf{M}^{(i)}, \xi)] \\ + \sum_{j=1}^{j=t_i} \Big(\mathbb{E}_q[\log p(y_j^{(i)}|z_j^{(i)}, \Omega_j^{(i)}, \Lambda^{(i)}, \Phi, \varpi)] \\ + \mathbb{E}_q[p(z_j^{(i)}|\eta)] \Big) \Bigg) \end{aligned} \tag{3.3}$$

where,

$$\mathbb{E}_q[\log p(\Phi_a|\alpha)] = \log\Gamma\Big(\sum_{j=1}^{j=t}\alpha_j\Big) - \sum_{j=1}^{j=t}\log\Gamma(\alpha_j) + \sum_{j=1}^{j=t}(\alpha_j - 1)\mathbb{E}_q[\log\Phi_{a,j}]$$

$$\mathbb{E}_q[\log p(\eta|\mu,\Sigma)] = \frac{1}{2}\log|\Sigma^{-1}| - \frac{b}{2}\log 2\pi - \frac{1}{2}\Big(\mathrm{tr}(\mathrm{diag}(\zeta^2)\Sigma^{-1}) + (\nu-\mu)^\top\Sigma^{-1}(\nu-\mu)\Big)$$

$$\mathbb{E}_q[\log p(\Lambda^{(i)}|\beta^{(i)})] = \sum_{a=1}^{a=b}\Big(\lambda_a^{(i)}\log\beta_a^{(i)} + (1-\lambda_a^{(i)}) \times \log(1-\beta_a^{(i)})\Big)$$

$$\mathbb{E}_q[\log p(\beta^{(i)}|\gamma,\kappa)] = \sum_{a=1}^{a=b}\Big((\gamma-1)\log(\beta_a^{(i)}) + (\kappa-1)\log(1-\beta_a^{(i)}) - \log(B(\gamma,\kappa)\Big)$$

$$\mathbb{E}_q[\log p(\Omega_i|\mathbf{M}^{(i)},\xi)] = \log\Gamma\Big(\sum_{j=1}^{j=t}\xi_j + \mathbf{M}_j^{(i)}\Big) - \sum_{j=1}^{j=t}\log\Gamma(\xi_j + \mathbf{M}_j^{(i)}) + \sum_{j=1}^{j=t}(\xi_j + \mathbf{M}_j^{(i)} - 1)\mathbb{E}_q[\log\Omega_j^{(i)}]$$

$$\mathbb{E}_q[\log p(y_j^{(i)}|z_j^{(i)},\Omega_j^{(i)},\Lambda^{(i)},\Phi,\varpi)] = \log\varpi + \sum_{c=1}^{c=t}\sum_{a=1}^{a=b}\Big(y_{j,c}^{(i)}\varsigma_{a,j}^{(i)}\lambda_a^{(i)}\mathbb{E}_q[(1-\Omega_c^{(i)})]\mathbb{E}_q[\log\Phi_{a,j}]\Big)$$

$$\mathbb{E}_q[\log p(z_j^{(i)}|\eta)] \approx 1 - \log\rho + \sum_{a=1}^{a=b}\nu_a\varsigma_{a,j}^{(i)} - \Big(\sum_{k=1}^{k=b}\mathbb{E}_q[\exp(\eta_k)]\Big)\rho^{-1}$$

The second term $\mathbb{H}(q)$ in Eq. 3.1 has the following parametric forms (see Eq. 3.2):

$$\mathbb{H}(q) = -\sum_{a=1}^{a=b}\mathbb{E}_q[\log q(\Phi_a|\phi_a)] - \sum_{i=1}^{i=n}\Bigg(\mathbb{E}_q[\log q(\eta^{(i)}|\nu,\zeta^2)] + \mathbb{E}_q[\log q(\Lambda^{(i)}|\lambda^{(i)})] + \mathbb{E}_q[\log q(\Omega^{(i)}|\omega^{(i)})] + \sum_{j=1}^{j=t_i}\mathbb{E}_q[\log q(z_j^{(i)}|\varsigma_j^{(i)})]\Bigg) \tag{3.4}$$

where,

$$\mathbb{E}_q[\log q(\Phi_a|\phi_a)] = \log\Gamma\Big(\sum_{j=1}^{j=t}\phi_{a,j}\Big) - \sum_{j=1}^{j=t}\log\Gamma(\phi_{a,j}) + \sum_{j=1}^{j=t}(\phi_{a,j}-1)\mathbb{E}_q[\log\Phi_{a,j}]$$

$$\mathbb{E}_q[\log q(\eta^{(i)}|\nu,\zeta^2)] = -\sum_{a=1}^{a=b}\frac{1}{2}\Big(\log\zeta_a^2 + \log(2\pi) + 1\Big)$$

$$\mathbb{E}_q[\log q(\Lambda^{(i)}|\lambda^{(i)})] = \sum_{a=1}^{a=b}\Big(\lambda_a^{(i)}\log\lambda_a^{(i)} + (1-\lambda_a^{(i)})\log(1-\lambda_a^{(i)})\Big)$$

$$\mathbb{E}_q[\log q(\Omega^{(i)}|\omega^{(i)})] = \log\Gamma\Big(\sum_{j=1}^{j=t}\omega_j^{(i)}\Big) - \sum_{j=1}^{j=t}\log\Gamma(\omega_j^{(i)}) + \sum_{j=1}^{j=t}(\omega_j^{(i)} - 1)\mathbb{E}_q[\log\Omega_j^{(i)}]$$

$$\mathbb{E}_q[\log q(z_j^{(i)}|\varsigma_j^{(i)})] = \mathbb{E}_q\Big[\log\prod_{a=1}^{a=b}(\varsigma_{a,j}^{(i)})^{z_{a,j}^{(i)}}\Big] = \sum_{a=1}^{a=b}\varsigma_{a,j}^{(i)}\log\varsigma_{a,j}^{(i)}$$

The exceptions that correspond to the above equations are:

$$\mathbb{E}_q[\log\Phi_{a,j}] = \Big(\Psi(\phi_{a,j}) - \Psi(\sum_{k=1}^{k=t}\phi_{a,k})\Big)$$

$$\mathbb{E}_q[\log\Omega_j^{(i)}] = \Big(\Psi(\omega_j^{(i)}) - \Psi(\sum_{k=1}^{k=t}\omega_k^{(i)})\Big)$$

$$\mathbb{E}_q[(1-\Omega_c^{(i)})] = \frac{1-\omega_c^{(i)}}{\sum_{k=1}^{k=t}(1-\omega_k^{(i)})}$$

$$\mathbb{E}_q[\exp(\eta_k)] = \exp(\nu_a + \frac{1}{2}\zeta_a^2)$$

$$B(\gamma,\kappa) = \frac{\Gamma(\gamma)\Gamma(\kappa)}{\Gamma(\gamma+\kappa)}$$

where Γ denotes the Gamma function while Ψ is the logarithmic derivative of the Gamma function.

After expanding both terms in Eq. 3.1, we can now maximize the bound in Eq. 3.1 with respect to each variational parameters using mini-batch coordinate ascent updates (Hoffman et al., 2013) as:

Optimize ς. The analytical expression of the variational group assignment $q(\varsigma)$ for each pathway j and group a for the ith example is not amenable due to the non-conjugacy of logistic-normal with latent variables. Instead, we approximate the solution as:

$$\varsigma_{a,j}^{(i)} \propto \exp\Bigg(\sum_{c=1}^{c=t}y_{j,c}^{(i)}\lambda_a^{(i)}\frac{1-\omega_c^{(i)}}{\sum_{k=1}^{k=t}(1-\omega_k^{(i)})}\Big(\Psi(\phi_{a,j}) - \Psi(\sum_{k=1}^{k=t}\phi_{a,k})\Big) + \nu_a - 1\Bigg) \tag{3.5}$$

where $\Psi(.)$ is the digamma function. Notice that the variational parameter $\omega_*^{(i)}$ acts as an smoothing parameter to selecting groups for each pathways, either from $\mathbf{M}_i$ or from $\mathcal{P}$.

Optimize ν**.** Collecting terms in the ELBO bound that contain only ν and taking derivatives w.r.t. ν_a for each group a, we obtain:

$$\frac{\partial \mathcal{L}(q)_{[\nu]}}{\partial \nu_a} = -\Sigma^{-1}(\nu - \mu) + \sum_{j=1}^{j=t_i} \varsigma_{a,j}^{(i)} - \Big(\exp(\nu_a + \frac{1}{2}\zeta_a^2)\Big) t_i \rho^{-1} \tag{3.6}$$

where ρ is another variational parameter, as in CTM (Blei and Lafferty, 2006). The above equation in hard to optimize, instead, we use a conjugate gradient algorithm.

Optimize ζ^2**.** By symmetry, we gather all the terms that has ζ^2 from Eq. 3.1, and take derivatives w.r.t. ζ_a^2 for each group a to obtain:

$$\frac{\partial \mathcal{L}(q)_{[\zeta^2]}}{\partial \zeta_a^2} = -\frac{1}{2}\left(\Sigma_{a,a}^{-1} + t_i \rho^{-1} \exp\left(\nu_a + \frac{1}{2}\zeta_a^2\right) - \frac{1}{\zeta_a^2}\right) \tag{3.7}$$

Again, there is no analytical solution to the above formula. Instead, we use Newton's method for each coordinate such that $\zeta_a \in \mathbb{R}_{>0}$.

Optimize ρ**.** We extract terms involved with the variational parameter ρ, and equating it's derivative to zero, we get:

$$\rho = \sum_{k=1}^{k=b} \exp(\nu_k + \frac{1}{2}\zeta_k^2) \tag{3.8}$$

Optimize ω**.** We next isolate only the terms in the bound that contain variational background pathway distributions $q(\omega)$. However, setting it's derivatives to zero does not lead to a closed-form solution, instead, we approximate $\omega_c^{(i)}$ for each example i according to:

$$\begin{aligned}\omega_c^{(i)} \propto & \xi_c + \mathbf{M}_c^{(i)} - \left(\frac{1 - \omega_c^{(i)} - \sum_{k=1}^{k=t}(1 - \omega_k^{(i)})}{(\sum_{k=1}^{k=t}(1 - \omega_k^{(i)}))^2}\right) \\ & \times \sum_{j=1}^{j=t_i}\sum_{a=1}^{a=b} y_{j,c}^{(i)} \varsigma_{a,j}^{(i)} \lambda_a^{(i)} \Big(\Psi(\phi_{a,j}) - \Psi(\sum_{k=1}^{k=t} \phi_{a,k})\Big)\end{aligned} \tag{3.9}$$

Optimize λ**.** To optimize λ, we use the canonical parameterisation of the Bernoulli distribution to get the following updates for each group a for each example:

$$\lambda_a^{(i)} = \frac{1}{1 + \exp^{-(\log(\beta_a^{(i)}) - \log(1 - \beta_a^{(i)}))}} \tag{3.10}$$

Optimize ϕ**.** Finally, the optimal solution of the variational pathway distribution $q(\Phi_a|\phi_a)$ for each group a is obtained by isolating terms involved in the ELBO bound in Eq. 3.1 and setting it's gradient to zero:

$$\phi_{a,c} = \alpha_c + \sum_{i=1}^{i=n}\sum_{j=1}^{j=t_i} y_{j,c}^{(i)} \varsigma_{a,j}^{(i)} \lambda_a^{(i)} \frac{1 - \omega_c^{(i)}}{\sum_{k=1}^{k=t}(1 - \omega_k^{(i)})} \tag{3.11}$$

The variational inference algorithm samples a mini-batch from a collection, and use it to compute the local latent parameters in Eqs 3.5, 3.6, 3.7, 3.8, 3.9, and 3.10 until the evidence lower bound in Eq. 3.1 converges. Then, the global variational parameter ϕ is updated in Eq. 3.11 using the posteriors (β, Λ, η, z, Ω) collected from the previous step after being scaled according to the learning rate $\tau = (s + l)^{-g}$, where s is the current step, $l \geq 0$ is the delay factor, and $g \in (0.5, 1]$ is the forgetting rate. This process for SPREAT is summarized in Algorithm 3.

Initialize ϕ, ν, ζ^2, λ, ω, ς, γ, κ, ξ, α, ϖ, ι, $s = 0$, $l \geq 0$, $g \in (0.5, 1]$
repeat
  $s = s + 1$;
  example a minibatch randomly $\mathcal{B} \subset \mathcal{P}$;
  for $i \in \mathcal{B}$ **do**
    repeat
      Update $\varsigma^{(i)}$ with Eq. 3.5;
      Update $\nu^{(i)}$ with Eq. 3.6 using conjugate gradient algorithm;
      Update $\zeta^{2,(i)}$ with Eq. 3.7 using Newton's method;
      Update $\rho^{(i)}$ with Eq. 3.8;
      Update $\omega^{(i)}$ with Eq. 3.9;
      Update $\lambda^{(i)}$ with Eq. 3.10;
    until *local variational parameters converge*;
    Compute optimal values $\mu = \frac{\nu}{|\mathcal{B}|}$, $\Sigma = \text{diag}(\frac{\zeta^2}{|\mathcal{B}|}) + \mu\mu^\top$;
    Compute global optimal values ϕ with Eq. 3.11;
    Update the current estimate of the global variational paramters, $x = (1 - \tau)x + \tau x$, where $x \in \{\phi, \mu, \Sigma\}$;
  Update the learning rate $\tau = (s + l)^{-g}$;
until *global convergence criterion is satisfied*;

Algorithm 3: Minibatch variational inference for SPREAT

3.1 Posterior Predictive Distribution

The posterior predictive distribution estimates the distribution of an unobserved value ($\tilde{\mathbf{y}}$) given the observed values ($\mathbf{Y}_{obs}$) and parameters (Θ and $\mathbf{V}$) that are trained on a held-out training set (Hoffman et al., 2013). The predictive distribution for SPREAT is:

$$\begin{aligned} p(\tilde{\mathbf{y}}|\mathbf{Y}_{obs}, \tilde{\mathbf{M}}, \mathbf{M}_{obs}) &= \int p(\tilde{\mathbf{y}}|\Theta, \tilde{\mathbf{M}}) p(\Theta|\mathbf{Y}_{obs}, \mathbf{M}_{obs}) d\Theta \\ &\approx \sum_{a=1}^{a=b} \left(\eta_a^{(i)} \times \sum_{j=1}^{j=t} \left(\Phi_{a,j} \times \tilde{\mathbf{y}}_j^{(i)}\right)\right) q(\Theta, \mathbf{V}) \end{aligned} \tag{3.12}$$

where $\tilde{\mathbf{M}}$ is $\tilde{\mathbf{y}}$'s background pathways and $q(\Theta, \mathbf{V})$ corresponds to Eq. 3.2, trained on $\mathbf{Y}_{obs}$ and $\mathbf{M}_{obs}$.

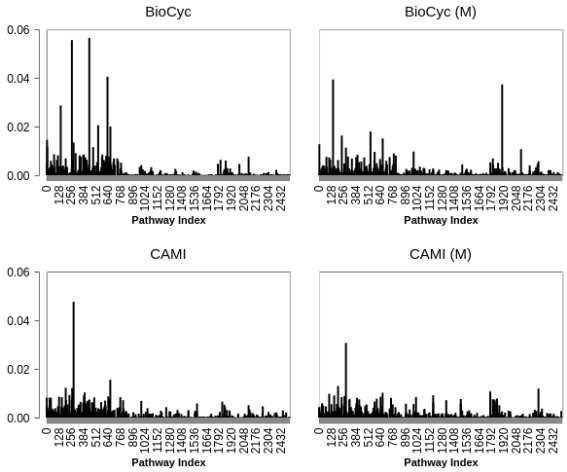


Figure 3: Pathway frequency (averaged on all examples) in BioCyc (v20.5 T2 &3) and CAMI data, and their background pathways, indicated by **M**.

4 EXPERIMENTAL SETTINGS

In this section, we describe the experimental datasets and settings used to validate the performance of the three correlated models. The CHAP package was written in Python v3 and is available under the GNU license at github.com/hallamlab/chap. All tests were conducted on a Linux server using 10 cores of Intel Xeon CPU E5-2650.

4.1 Description of Datasets

The three models were evaluated on diverse pathway datasets traversing the genomic information hierarchy (M. A. Basher et al., 2020): i)- T1 golden consisting of EcoCyc, HumanCyc, AraCyc, YeastCyc, LeishCyc, and TrypanoCyc; ii)- BioCyc (v20.5 T2 & 3) (Caspi et al., 2016); iii)- Critical Assessment of Metagenome Interpretation (CAMI) dataset composed of 40 genomes (Sczyrba et al., 2017); and iv)- Synset-2, a noisy training dataset, introduced in (M. A. Basher et al., 2020). The preprocessed experimental datasets can be obtained from zenodo.org/record/5630322#.YYXur2DMK3B while information about these data is provided in (M. A. Basher et al., 2020).

4.2 Parameter Settings

Three experiments were conducted: i)- parameter sensitivity analysis, ii)- groups visualization, and iii)- metabolic pathway prediction. Unless otherwise mentioned, we applied the following default configurations: the pathway distribution over groups Φ were initialized using gamma distribution (with shape and scale parameters were fixed to 100 and $1/100$, respectively), the forgetting rate was $g = 0.9$, the delay rate was $l = 1$, the batch size was 100, the number of epochs was 3, the number of groups was $b = 200$, top k pathways was 100 (only for SOAP and SPREAT), the Dirichlet hyperparameters α and ξ were 0.0001, and the beta hyperparameters γ and κ were 2 and 3, respectively. The supplementary pathways **M** for BioCyc, CAMI, and golden T1 datasets were obtained using mlLGPR (elastic-net with enzymatic reaction and pathway evidence features)(M. A. Basher et al., 2020) trained on Synset-2. A schematic view of pathway frequency for BioCyc T2 &3 and CAMI data with their background pathways is depicted in Fig. 3.

After obtaining groups, we first applied reMap (M. A. Basher and Hallam, 2021c) to map examples to groups, then, we trained the resulted data using leADS (M. A. Basher and Hallam, 2021b) by setting "factorization" option which enables training pathway groups and pathways, simultaneously, for 10 epochs using "nPSP" as the acquisition function and "pref-voting" as the prediction strategy with cutoff threshold 0.5. All hyperparameters in leADS (M. A. Basher et al., 2021c), reMap (M. A. Basher and Hallam, 2021a), and mlLGPR (M. A. Basher et al., 2020), were fixed to their default values.

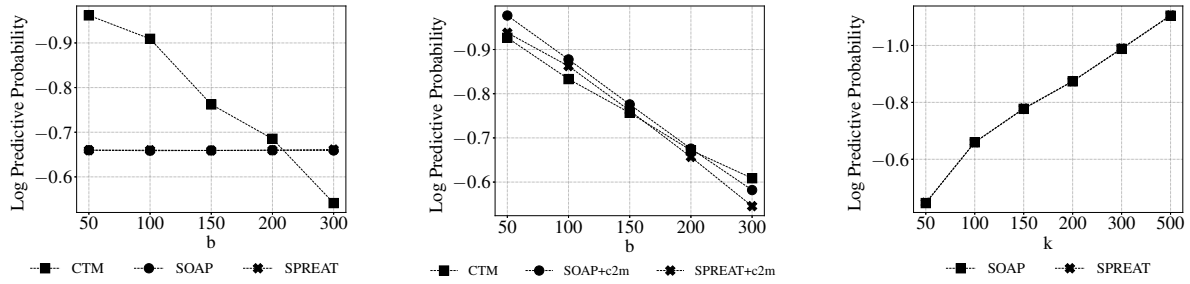
5 EXPERIMENTAL RESULTS

This section analyzes the three models using the settings explained in the previous section.

5.1 Sensitivity Analysis

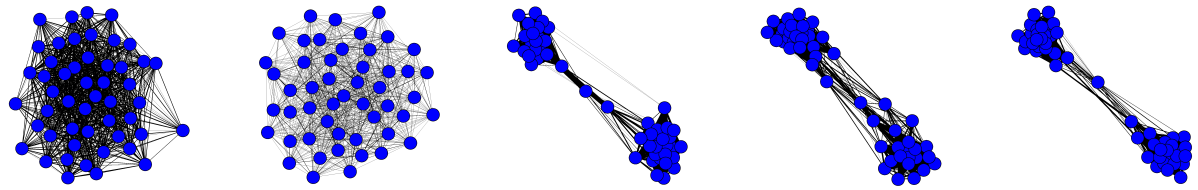
Experimental setup. Following the common practice, here we study the effect of hyperparameters on the performance of correlated models. First, we compare the sensitivity of SOAP and SPREAT against CTM by incorporating the background pathways **M** while varying the number of groups according to $b \in \{50, 100, 150, 200, 300\}$. Next, we examine the SOAP and SPREAT with collapsed option (or c2m) to compare their performances to CTM, where the former models should exhibit similar performances as CTM. Finally, we conduct sparsity analysis of group distribution by varying the cutoff threshold value according to $k \in \{50, 100, 150, 200, 300, 500\}$ (Section 2.2). For comparative analysis, we apply CAMI as test data to report the log predictive distribution (Section 3.1), where a lower score entails higher generalization capability for the corresponding model.

Experimental results. While the log predictive scores for SOAP and SPREAT in Fig. 4a appear to



(a) Effect of b (b) Effect of b by collapsing (c) Effect of k with $b = 200$

Figure 4: Log predictive distribution on CAMI data.



(a) SOAP (#groups: ~ 19) (b) SPREAT (#groups: ~ 13) (c) CTM (#groups: ~ 12) (d) SOAP+c2m (#groups: ~ 12) (e) SPREAT+c2m (#groups: ~ 12)

Figure 5: Visualizing 50 randomly picked groups for each model, trained using $b = 200$, where #groups corresponds to the average number of correlated groups. The circles represent groups, and their sizes reflect the correlation strength with other groups. Far distant groups are less correlated with the near groups.

be flat across various group sizes, the CTM model projects a more realistic view where its performances are seen to be gaining by including more groups. This phenomenon is expected for SOAP and SPREAT as both absorb supplementary pathways, thereby, enforcing to learn more pathways from $\mathbf{M}$ which has an average of ~ 500 pathways that is more than double the average number of pathways in BioCyc v20.5 T2 & 3 (~ 195 pathways). By excluding $\mathbf{M}$ ("c2m" in Section 2.2) in the SOAP and SPREAT training, the log predictive distribution of these models exhibit similar performance as CTM (Fig. 4b), asserting our previous discussion. From Figs 4a and 4b, it is evident that $b = 200$ represents an optimum group size. To find an optimum k value, we fixed $b = 200$ and retrained all models. From Fig. 4c, the performances for SOAP and SPREAT are seen to decline (< -0.6) when $k > 100$.

Results from this experiment suggest that the settings $b \in \mathbb{Z}_{[150,300]}$ and $k \in \mathbb{Z}_{[50,100]}$ are optimum for discovering pathway groups in $\mathcal{P}$.

5.2 Groups Visualization

Experimental setup. Recall that groups constitute overlapping pathways. In this experiment, we visually explore the recovered groups from the three correlated models trained on BioCyc (v20.5 T2 &3) data using configurations discussed in Section 4.2. We investigate group correlations, reflected in Σ, for SOAP, SPREAT, CTM, SOAP+c2m, and SPREAT+c2m models, to analyze the influence of dual sparseness (Section 2.2) and background pathways on Σ.

Experimental results. Fig. 5 demonstrates 50 randomly picked groups and their correlations as represented by Σ for all models. The width of edges indicates the strength of correlations. Essentially for every group in these models, there are approximately 12 to 19 closely related groups. This indicates that metabolic pathways are distributed over multiple groups, therefore, forming overlapping pathways. With regard to $\mathbf{M}$, as explained in Section 5.1, background pathways in $\mathbf{M}$ consist of ~ 500 pathways for a single organismal or multi organismal genomes in comparison to BioCyc (v20.5 T2 &3) data that has an average of ~ 195 pathways. These additive pathways have influenced the construction of group correlation for both SOAP and SPREAT. Pathway groups in SOAP consist of more associated groups (~ 19 groups) than the remaining models. This has an important implication in the pathway prediction task, discussed in Section 5.3. Sparse models share a similar group structure as CTM (also they have similar log predictive scores in Section 5.1), therefore, they may exhibit similar performance in the downstream

pathway prediction task.

Results from this experiment show that SOAP and SPREAT are better contenders than CTM. Specifically, both models incorporate supplementary pathways and apply dual sparseness to reduce both the group size and the statistically irrelevant pathways.

5.3 Metabolic Pathway Prediction

Experimental setup. Pathway groups obtained from correlated models are used for the pathway prediction task. We consider five models: CTM, two models with background pathways (SOAP and SPREAT), and two collapsed models (SOAP+c2m and SPREAT+c2m). After obtaining groups, we train reMap and leADS using the configuration discussed in Section 4.2. The results are reported on golden T1 data using four evaluation metrics: *Hamming loss*, *average precision*, *average recall*, and *average F1 score*. For comparative analysis, four pathway prediction algorithms are used: i)- MinPath v1.2 (Ye and Doak, 2009), ii)- PathoLogic v21 (Karp et al., 2016), iii)- mlLGPR (elastic net with enzymatic reaction and pathway evidence features) (M. A. Basher et al., 2020), and iv)- triUMPF (M. A. Basher et al., 2021a).

Experimental results. Table 2 shows that leADS trained using groups from SOAP results in competitive performance against the other methods in terms of average F1 score with optimal performance on EcoCyc (0.8336). However, it seems to be underperforming on AraCyc, YeastCyc, and LeishCyc, yielding average F1 scores of 0.4764, 0.4914, and 0.4144, respectively. This is attributed to background pathways in $\mathbf{M}$ that have more pathways and are sometimes irrelevant (see Section 5.1), hence, impacting the training process. Interestingly, SPREAT's performances are shown to be inferior to SOAP. As alluded in Section 5.2, the average number of correlated groups for SOAP is significantly larger than SPREAT (Section 5.2), enabling leADS to revisit a true positive pathway for an organism multiple times across groups in SOAP to signal its presence in contrast to groups from SPREAT. With respect to the sensitivity score, correlated models, in general, resulted in higher scores than triUMPF, therefore, attesting the novelty of modeling groups to improve predictions.

Results from this experiment demonstrate that the group-based approach, in particular SOAP, improves the downstream metabolic pathway prediction outcomes. We suggest applying SOAP for pathway predictions using the default configurations discussed in Section 4.2.

6 CONCLUSIONS

In this paper, we presented two novel statistical hierarchical mixture models, SOAP and SPREAT, to uncover correlated pathway groups given pathway abundance data. The work is motivated by the problem of missing pathways, which is very common in genomic datasets. We empirically evaluated correlated models for the pathway prediction task using golden T1 data and compared results to other prediction methods including PathoLogic, MinPath, mlLGPR, and triUMPF. Overall, correlated models showed promising results in boosting prediction performance over ML-based algorithms, such as triUMPF. There are several directions for future study. Foremost, we intend to build a model that combines both graph-based solution (M. A. Basher et al., 2021a) and group-based strategy to improve metabolic pathway predictions with emphasis on multi-organismal genomes. We also investigate the sparseness induction in the covariance matrix for better interpretability (Fan et al., 2016). This and other possibilities we leave for future work.

ACKNOWLEDGEMENTS

This work was performed under the auspices of Genome Canada, Genome British Columbia, the Natural Sciences and Engineering Research Council (NSERC) of Canada, and Compute/Calcul Canada). ARMAB was supported by a UBC four-year doctoral fellowship (4YF) administered through the UBC Graduate Program in Bioinformatics.

REFERENCES

Aguiar-Pulido, V., Huang, W., Suarez-Ulloa, V., Cickovski, T., Mathee, K., and Narasimhan, G. (2016). Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis: supplementary issue: bioinformatics methods and applications for big metagenomics data. *Evolutionary Bioinformatics*, 12:EBO–S36436.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of machine learning research*, 9(Sep):1981–2014.

Baranwal, M., Magner, A., Elvati, P., Saldinger, J., Violi, A., and Hero, A. O. (2020). A deep learning architecture for metabolic pathway prediction. *Bioinformatics*, 36(8):2547–2553.

Bashiardes, S., Zilberman-Schapira, G., and Elinav, E. (2016). Use of metatranscriptomics in microbiome research. *Bioinformatics and biology insights*, 10:BBI–S34610.

Table 2: Predictive performance of each comparing algorithm on 6 benchmark datasets. For each performance metric, '↓' indicates the smaller score is better while '↑' indicates the higher score is better. Bold text suggests the best performance in each column.

Methods	Hamming Loss ↓					
	EcoCyc	HumanCyc	AraCyc	YeastCyc	LeishCyc	TrypanoCyc
PathoLogic	0.0610	0.0633	0.1188	**0.0424**	**0.0368**	**0.0424**
MinPath	0.2257	0.2530	0.3266	0.2482	0.1615	0.2561
mlLGPR	0.0804	0.0633	**0.1069**	0.0550	0.0380	0.0590
triUMPF	0.0435	0.0954	0.1560	0.0649	0.0443	0.0776
SOAP	**0.0392**	**0.0400**	0.1714	0.0934	0.0772	0.0479
SPREAT	0.0519	0.0827	0.1489	0.0748	0.0629	0.0503
CTM	0.0558	0.0835	0.1425	0.0804	0.0622	0.0503
SOAP+c2m	0.0590	0.0780	0.1457	0.0772	0.0614	0.0534
SPREAT+c2m	0.0542	0.0796	0.1520	0.0772	0.0598	0.0558

Methods	Average Precision Score ↑					
	EcoCyc	HumanCyc	AraCyc	YeastCyc	LeishCyc	TrypanoCyc
PathoLogic	0.7230	0.6695	0.7011	0.7194	**0.4803**	0.5480
MinPath	0.3490	0.3004	0.3806	0.2675	0.1758	0.2129
mlLGPR	0.6187	0.6686	0.7372	0.6480	0.4731	0.5455
triUMPF	0.8662	0.6080	0.7377	**0.7273**	0.4161	0.4561
SOAP	0.8611	**0.7871**	0.6215	0.4851	0.2805	0.5985
SPREAT	**0.9400**	0.6750	0.8350	0.6000	0.3200	**0.6200**
CTM	0.9150	0.6700	**0.8750**	0.5650	0.3250	**0.6200**
SOAP+c2m	0.8950	0.7050	0.8550	0.5850	0.3300	0.6000
SPREAT+c2m	0.9250	0.6950	0.8150	0.5850	0.3400	0.5850

Methods	Average Recall Score ↑					
	EcoCyc	HumanCyc	AraCyc	YeastCyc	LeishCyc	TrypanoCyc
PathoLogic	0.8078	0.8423	0.7176	0.8734	0.8391	0.7829
MinPath	**0.9902**	**0.9713**	**0.9843**	**1.0000**	**1.0000**	**1.0000**
mlLGPR	0.8827	0.8459	0.7314	0.8603	0.9080	0.8914
triUMPF	0.7590	0.3835	0.3529	0.3319	0.7126	0.6229
SOAP	0.8078	0.8746	0.3863	0.4978	0.7931	0.9371
SPREAT	0.6124	0.4839	0.3275	0.5240	0.7356	0.7086
CTM	0.5961	0.4803	0.3431	0.4934	0.7471	0.7086
SOAP+c2m	0.5831	0.5054	0.3353	0.5109	0.7586	0.6857
SPREAT+c2m	0.6026	0.4982	0.3196	0.5109	0.7816	0.6686

Methods	Average F1 Score ↑					
	EcoCyc	HumanCyc	AraCyc	YeastCyc	LeishCyc	TrypanoCyc
PathoLogic	0.7631	0.7460	0.7093	**0.7890**	0.6109	0.6447
MinPath	0.5161	0.4589	0.5489	0.4221	0.2990	0.3511
mlLGPR	0.7275	0.7468	**0.7343**	0.7392	**0.6220**	0.6768
triUMPF	0.8090	0.4703	0.4775	0.4735	0.5254	0.5266
SOAP	**0.8336**	**0.8285**	0.4764	0.4914	0.4144	**0.7305**
SPREAT	0.7416	0.5637	0.4704	0.5594	0.4460	0.6613
CTM	0.7219	0.5595	0.4930	0.5268	0.4530	0.6613
SOAP+c2m	0.7061	0.5887	0.4817	0.5455	0.4599	0.6400
SPREAT+c2m	0.7298	0.5804	0.4592	0.5455	0.4739	0.6240

Bien, J. and Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820.

Blei, D. and Lafferty, J. (2006). Correlated topic models. *Advances in neural information processing systems*, 18:147.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Caspi, R., Billington, R., Foerster, H., and et al. (2016). Biocyc: Online resource for genome and metabolic pathway analysis. *The FASEB Journal*, 30(1

Supplement):lb192–lb192.

Caspi, R., Billington, R., Keseler, I. M., Kothari, A., Krummenacker, M., Midford, P. E., Ong, W. K., Paley, S., Subhraveti, P., and Karp, P. D. (2019). The metacyc database of metabolic pathways and enzymes-a 2019 update. *Nucleic acids research*.

Dale, J. M., Popescu, L., and Karp, P. D. (2010). Machine learning methods for metabolic pathway prediction. *BMC bioinformatics*, 11(1):1.

Fan, J., Liao, Y., and Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19(1):C1–C32.

Hanson, N. W., Konwar, K. M., Hawley, A. K., and et al. (2014). Metabolic pathways for the whole community. *BMC genomics*, 15(1):1.

Hassa, J., Maus, I., Off, S., Pühler, A., Scherer, P., Klocke, M., and Schlüter, A. (2018). Metagenome, metatranscriptome, and metaproteome approaches unraveled compositions and functional relationships of microbial communities residing in biogas plants. *Applied microbiology and biotechnology*, 102(12):5045–5063.

He, J., Hu, Z., Berg-Kirkpatrick, T., and et al. (2017). Efficient correlated topic modeling with topic embedding. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 225–233. ACM.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. W. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347.

Jiao, D., Ye, Y., and Tang, H. (2013). Probabilistic inference of biochemical reactions in microbial communities from metagenomic sequences. *PLoS Comput Biol*, 9(3):e1002981.

Kanehisa, M., Furumichi, M., Tanabe, M., and et al. (2017). Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361.

Karp, P. D., Latendresse, M., Paley, S. M., and et al. (2016). Pathway tools version 19.0 update: software for pathway/genome informatics and systems biology. *Briefings in bioinformatics*, 17(5):877–890.

Lin, T., Tian, W., Mei, Q., and Cheng, H. (2014). The dual-sparse topic model: mining focused topics and focused terms in short text. In *Proceedings of the 23rd international conference on World wide web*, pages 539–550. ACM.

Loh, P.-R., Baym, M., and Berger, B. (2012). Compressive genomics. *Nature biotechnology*, 30(7):627.

M. A. Basher, A. R. and Hallam, S. J. (2021a). Relabeling metabolic pathway data with groups to improve prediction outcomes. *BioRxiv*.

M. A. Basher, A. R. and Hallam, S. J. (2021b). https://github.com/hallamlab/leADS.

M. A. Basher, A. R. and Hallam, S. J. (2021c). https://github.com/hallamlab/reMap.

M. A. Basher, A. R., McLaughlin, R. J., and Hallam, S. J. (2020). Metabolic pathway inference using multi-label classification with rich pathway features. *PLOS Computational Biology*, 16(10):1–22.

M. A. Basher, A. R., McLaughlin, R. J., and Hallam, S. J. (2021a). Metabolic pathway prediction using non-negative matrix factorization with improved precision. *Journal of Computational Biology*.

M. A. Basher, A. R., McLaughlin, R. J., and Hallam, S. J. (2021b). Metabolic pathway prediction using non-negative matrix factorization with improved precision. In *Computational Advances in Bio and Medical Sciences*, pages 33–44, Cham. Springer International Publishing.

M. A. Basher, A. R., Nallan, A. N., McLaughlin, R. J., and et al. (2021c). leads: improved metabolic pathway inference based on active dataset subsampling. *bioRxiv*.

Mascher, M., Schreiber, M., Scholz, U., Graner, A., Reif, J. C., and Stein, N. (2019). Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. *Nature genetics*, 51(7):1076–1081.

Qi, Q., Li, J., and Cheng, J. (2014). Reconstruction of metabolic pathways by combining probabilistic graphical model-based and knowledge-based methods. In *BMC proceedings*, volume 8, pages 1–10. Springer.

Sczyrba, A., Hofmann, P., Belmann, P., and et al. (2017). Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nature methods*, 14(11):1063.

Shafiei, M., Dunn, K. A., Chipman, H., Gu, H., and Bielawski, J. P. (2014). Biomenet: A bayesian model for inference of metabolic divergence among microbial communities. *PLoS Comput Biol*, 10(11):e1003918.

Stewart, F. J., Sharma, A. K., Bryant, J. A., and et al. (2011). Community transcriptomics reveals universal patterns of protein sequence conservation in natural microbial communities. *Genome biology*, 12(3):R26.

Tabei, Y., Yamanishi, Y., and Kotera, M. (2016). Simultaneous prediction of enzyme orthologs from chemical transformation patterns for de novo metabolic pathway reconstruction. *Bioinformatics*, 32(12):i278–i287.

Wang, W.-L., Xu, S.-Y., Ren, Z.-G., Tao, L., Jiang, J.-W., and Zheng, S.-S. (2015). Application of metagenomics in the human gut microbiome. *World journal of gastroenterology: WJG*, 21(3):803.

Yamanishi, Y., Tabei, Y., and Kotera, M. (2015). Metabolome-scale de novo pathway reconstruction using regioisomer-sensitive graph alignments. *Bioinformatics*, 31(12):i161–i170.

Ye, Y. and Doak, T. G. (2009). A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput Biol*, 5(8):e1000465.

Zhao, Y., Chen, M.-H., Pei, B., Rowe, D., Shin, D.-G., Xie, W., Yu, F., and Kuo, L. (2012). A bayesian approach to pathway analysis by integrating gene–gene functional directions and microarray data. *Statistics in biosciences*, 4(1):105–131.