Modeling Metabolic Pathways as Bags (with Augmentation)

Abdur Rahman M. A. Basher[1] and Steven J. Hallam [1,2,3,4,5*]

[1] Graduate Program in Bioinformatics, University of British Columbia, Genome Sciences Centre, 100-570 West 7th Avenue, Vancouver, British Columbia V5Z 4S6, Canada
[2] Department of Microbiology & Immunology, University of British Columbia, 2552-2350 Health Sciences Mall, Vancouver, British Columbia V6T 1Z3, Canada.
[3] Genome Science and Technology Program, University of British Columbia, 2329 West Mall, Vancouver, BC V6T 1Z4, Canada
[4] Life Sciences Institute, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z3
[5] ECOSCOPE Training Program, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z3

Abstract

Given unprecedented amounts of genomic data, generated by NGS technologies, many computation tools to discern knowledge became widely popular. In response, we propose the CBT (correlated bag pathway) package comprising of three hierarchical mixture models: SOAP (sparse correlated bag pathway), SPREAT (distributed sparse correlated bag pathway), and CTM (correlated topic model) to characterize pathways. The generic idea is to incorporate pathway abundance to encode each genomic data as mixture distributions of bags, and each bag, in turn, is a mixture of pathways. Empirical studies with regard to the quality of discovered pre-selected bags and their correlations show promising results, hence, providing insights to adapting bags for the pathway prediction problem.
Availability and implementation: The software package is published on github.com/cbt
Contact: shallam@mail.ubc.ca

1 Introduction

With the rapid advancements in the next-generation sequencing platforms, over the past decade, there has been an unprecedented increase in generating petabytes of genomic dataset. Consequently, several computational tools were developed to detect metabolic functions and their interactions (e.g., enzyme-enzyme and pathway-pathway) [2]. Since DNA sequence data is comprised of enzyme abundance information that can be used to infer pathways [4], it is possible to discern functional relationships of organisms. Towards achieving this objective, we developed two mixed-membership hierarchical Bayesian models, SOAP and SPREAT, aiming to capture higher-level representation of genomic data called "bags" or "concepts" where a bag is a set of expected correlated pathways. These two models extend CTM [6] by leveraging dual sparseness and supplementary (or background) pathways in modeling bag distribution to address the missing pathway information (due to noise propagated from upstream preprocessing in a bioinformatics pipeline) in order to deliver more accurate relations of organisms.

Let us start by giving some background and notations, then state our research problem that aims to recover the distribution of correlated pathways.

Definition 1.1. Pathway Collection. Let $\mathcal{P} = \{\mathbf{y}^{(i)} : 1 < i \leqslant n\}$ be a collection of n examples, 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 t known metabolic pathways, as defined from a trusted source, such as MetaCyc [11], and $\mathcal{Y}_i \subseteq \mathcal{Y}$ corresponds to a subset of true pathways associated with the sample i. ■

Recovering latent distributions of $\mathcal{P}$ mirrors the concept modeling paradigm, which was introduced to reconstruct the thematic structure, called "topics", from a corpus [8]. In genomic studies, the concept modeling has been subject to a wide range of applications, such as extracting latent microbial communities as in BioMiCo [29] and MetaTopics [33], binning metagenomic reads (TM-MCluster) [34], and inferring metabolic interactions from a microbial community as in BiomeNet [30]. A detailed overview of concept modeling in biological data is outlined in [23].

Definition 1.2. Concept Modeling. Given a collection of n samples, a concept distribution for i-th example is a multinomial distribution vector, denoted by $\eta^{(i)}$ of size b concepts, i.e., $\{p(\Phi_a|\eta^{(i)})\}_{a=1}^{a=b}$, where Φ_j in a multinomial feature distribution over the concept j, i.e., $\{p(y_k|\Phi_j)\}_{k=1}^{k=t}$. The overall goal of concept modeling is to recover the b salient concepts of each example. ■

In this report, the term concept is referred to as "bag" which is coined from the multi-graph classification (MGC) technique [32], where each bag is comprised of several correlated pathways. For brevity purposes, the following terms: *concept*, *topic*, or *bag*, are used interchangeably. Also, *features* correspond to *pathways*.

The classical studies in concept modeling attempt to discover concepts from a collection of examples that are composed of features, as in the case of latent Dirichlet allocation (LDA) [8]. This model assumes an example is a mixture of concepts derived from a Dirichlet prior, $\eta^{(i)} \sim \mathrm{Dir}(\alpha)$, and each concept j, in turn, is a mixture of features, sampled from another Dirichlet prior, i.e., $\Phi_j \sim \mathrm{Dir}(\beta)$. However, the assumptions imposed on the Dirichlet prior to the concept distributions restrains the capability of LDA to capture possible dependencies among concepts. Naturally, we expect that concepts are not independent of each other. For example, in genomic samples, we expect to exhibit a collection of observed pathways to be grouped under the metabolic nitrogen network while a subset of the same pathways may be related to the carbon cycle [12, 15].

For this, Blei and Lafferty [6] proposed an extension of LDA, referred to as correlated topic model (CTM), by incorporating logistic-normal prior, which models pairwise concept correlations with the Gaussian covariance matrix. This model has been further examined in [16] by employing continuous distributed representations for both latent concepts and examples, thus, enabling to capture concept correlations using simple Euclidean distance metric. This embedding based correlation has resulted in downscaling the computational burden required to perform inference. We take advantage of the inherent thematic structure of examples and model the concept dependencies to extract the concept distributions of examples.

Definition 1.3. Concept Correlation. Given $\mathcal{P}$, the pairwise concept-correlation is defined by a Gaussian covariance matrix, denoted by Σ. Each entry $s_{i,j}$ in Σ characterizes the i-th bag association with the bag j, where a larger score indicates both concepts are highly correlated. ■

However, as highlighted in [14], by manual inspection of the Hawaii Ocean Time-Series (HOTS) samples, the authors identified a set of pathways that are not curated in $\mathcal{P}$ for HOTS samples, such as pathway variants related to tricarboxylic acid cycle (TCA). Hypothetically, accommodating those pathways back to P may improve the precision of recovering pathways. Unfortunately, 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 a false positive in the subsequent stage of experimental studies. A good compromise would be to record those missing pathways in a separate list while keeping the original pathway collection intact for further investigation.

To this end, we store a potential set of missing pathways in a matrix $\mathbf{M} \in \mathbb{Z}_{\geq 0}^{n \times t}$, where each entry is an integer value indicating the abundance of the pathway associated with a specific sample. The matrix $\mathbf{M}$ is referred to as a "background" or "supplementary" set, analogous to the previous studies in [35, 18]. With the above definitions, we are now able to constitute the problem discussed in this section.

Problem Statement 1. *Given both $\mathcal{P}$ and $\mathbf{M}$, the goal is to efficiently recover the concept distribution η of samples.*

2 Correlated Models

In this section, we provide an overview of the correlated topic model. Then, we present two bag pathway models that incorporate background pathways while also enforcing sparsity in modeling bag proportions.

2.1 The Correlated Topic Model

The correlated topic model (CTM) is a probabilistic graphical model that extends the generative story of LDA [8] to incorporate correlation among concepts. Fig. 1 (a) shows the Bayesian graphical model for CTM using plate notation. Like latent Dirichlet allocation [8], the CTM is comprised of a hierarchical Bayesian mixture model, where features (words as described in the original paper) are mixed to constitute concepts. And, the concepts, in contrast to LDA, are assumed to be correlated to each other, as reflected by a Gaussian covariance matrix.

Formally, let n be the total number of a collection, where each example i consists of feature indices as $\mathbf{y}^{(i)}$. 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 concept $a \in \{1, \ldots, b\}$. 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 concept distributions $\theta^{(i)} = \mathrm{softmax}(\eta^{(i)})$, corresponding

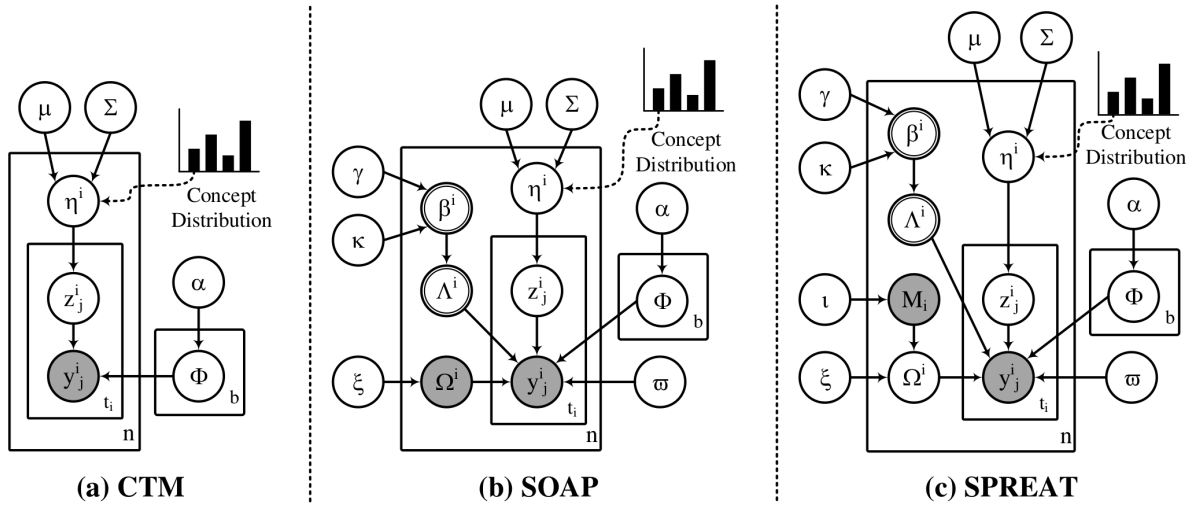


Figure 1: **Graphical representation of the correlated concept models.** The boxes are "plates" representing replicates. The outer plate represents instances, while the inner plate represents the repeated choice of features within an example. The logistic normal distribution, used to model the latent concept proportions of an example, can represent capture correlations among concepts that are impossible to capture using a single Dirichlet. The observed data for each example $\mathbf{x}^{(i)}$ are a set of annotated features $\mathbf{y}^{(i)}$ and a set of hypothetical features $\mathbf{M}_i$ while per-example concept proportions $\eta^{(i)}$, per-example concept selection parameters $\Lambda^{(i)}$, per-example hypothetical feature distributions $\Omega^{(i)}$, per-feature concept assignment $z_j^{(i)}$, and per-concept distribution over features Φ_a, and per-example beta distribution $\beta^{(i)}$ are hidden variables. The remaining hyperparameters should be provided as inputs.

1. For each concept $a \in \{1, \ldots, b\}$:
   (a) Sample a distribution over features $\Phi_a \sim \text{Dir}(.|\alpha)$;
2. For each example $i \in \{1, \ldots, n\}$:
   (a) Draw the example concept weight $\eta^{(i)} \sim \mathcal{N}(.|\mu, \Sigma)$;
   (b) Draw concept proportions $\theta^{(i)} = \text{softmax}(\eta^{(i)})$;
   (c) For each feature $j \in \{1, \ldots, t^{(i)}\}$:
      i. Sample a concept assignment $z_j^{(i)} \sim \text{Mult}(.|\theta^{(i)})$;
      ii. Sample a feature $y_j^{(i)} \sim \text{Mult}(.|\Phi_{z_j^{(i)}})$;

Algorithm 1: The generative process for CTM given a collection

the logistic-normal distribution, from which a concept 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, as indicated by it's concept assignment, i.e., $y_j^{(i)} \sim \Phi_{z_j^{(i)}}$. The CTM's generative process is outlined in Algorithm 1, which can be observed that the process is identical to LDA except the concept distributions is sampled from the logistic normal rather than a Dirichlet prior.

2.2 The Correlated Bag-Pathway Model

The correlated bag pathway is an extension to CTM and comes in two flavors, as depicted in Fig. 1 (b) and (c): i)- sparse correlated bag pathway (SOAP) and ii)- distributed sparse correlated bag pathway (SPREAT). Both models incorporate dual sparseness and supplementary pathways in modeling bag proportions. These important properties are not adopted in CTM. In contrast to SOAP, SPREAT is equipped to reconstruct the latent distribution of supplementary pathways. Let us formally explain both models in detail. For the purpose of coherence, we use features to define pathways as before.

Analogous to CTM, given n number of examples and a matrix encoding the missing features $\mathbf{M}$, the generative process for SOAP and SPREAT can be described as follows. First, we draw a multinomial feature distribution Φ_a from asymmetric Dirichlet prior $\alpha \in \mathbb{R}_{>0}$ for each concept $a \in \{1, ..., b\}$, 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 features, is inaccessible. Similar

1. For each concept $a \in \{1, ..., b\}$:
   (a) Sample a distribution over features $\Phi_a \sim \text{Dir}(.|\alpha)$;
2. For each example $i \in \{1, ..., n\}$:
   (a) Draw the example concept weight $\eta^{(i)} \sim \mathcal{N}(.|\mu, \Sigma)$;
   (b) Draw concept proportions $\theta^{(i)} = \text{softmax}(\eta^{(i)})$;
   (c) Draw beta distribution $\beta^{(i)} \sim \text{Beta}(.|\gamma, \kappa)$;
   (d) Draw a sparsity indicator vector $\Lambda^{(i)} \sim \text{Bernoulli}(.|\beta^{(i)})$;
   if `SPREAT`:
      i. Sample a vector $\mathbf{M}_i \sim \text{Prior}(.|\iota)$;
      ii. Sample background distribution $\Omega^{(i)}|\mathbf{M}_i \sim \text{Dir}(.|\xi)$;
   else:
      i. Draw background feature proportions $\Omega^{(i)} \sim \text{Dir}(.|\xi)$;
   (e) For each feature $j \in \{1, \ldots, t^{(i)}\}$:
      i. Sample a concept assignment $z_j^{(i)} \sim \text{Mult}(.|\Lambda^{(i)} \odot \theta^{(i)})$;
      ii. Sample a feature $y_j^{(i)} \sim \text{Mult}(.|(1 - \Omega^{(i)}_{z_j^{(i)}}) \odot \Phi_{z_j^{(i)}})$;

Algorithm 2: The generative process for SOAP and SPREAT

to CTM, for each example i, a concept 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 concept, it is reasonable to expect that each example is usually explained with a handful set of a mixed proportion of concepts, which may include interactions among them. Besides, a concept should cover only a few focused features, instead of absorbing all features. Thus, we borrow the idea from [22, 1, 26, 5, 16] to enforce dual sparsity to retain those relevant focused concepts and features by: i)- introducing an auxiliary Bernoulli variable $\Lambda^{(i)}$ of size b to determine whether a concept is selected for an example i or ignored, and ii)- applying a cutoff threshold to keep only the top $k \ll t$ features for each concept. 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 concepts while minimizing the negative correlation among concepts on Σ.

Next, a concept 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 feature $y_j^{(i)}$ in example i is generated from a weighted distribution $\Omega^{(i)}_{z_j^{(i)}} \odot \Phi_{z_j^{(i)}}$, as indicated by it's concept assignment, 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 feature 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 example's probability of using feature $y_j \in \mathbf{M}_i$. Here, the background feature is assumed to be drawn from a sparse binary vector prior $\iota \in \mathbb{R}_{>0}$ that is included for completeness because each example's feature $\mathbf{M}_i$ is already observed.

By 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., features) permit to contribute to other structures with different degrees of granularity (e.g., examples). The generative process of the proposed SOAP and SPREAT models is summarized in Algorithm 2. Note that by setting all entries in Ω, Λ, and ϖ to 1, SOAP and SPREAT are reduced to CTM ("collapse2ctm" or c2m), which is an additional benefit to the former models.

3 Evidence Lower Bound (ELBO) for SPREAT

Here, we discuss inference for SPREAT model. Similar expression is straightforward to derive for SOAP. Given $\mathcal{P}$, the goal of inference is to compute the posterior distribution of the per-example concept proportions $\eta^{(i)}$, the per-example concept selection parameters $\Lambda^{(i)}$ and the associated beta distributions $\beta^{(i)}$, the per-example background feature distributions $\Omega^{(i)}$, the per-feature concept assignment $z_j^{(i)}$, and the per-concept distribution over features Φ_a.

Looking at the topology of the Bayesian network, we can specify the complete-data likelihood, i.e., the joint distribution of all observed and latent variables given the hyperparameters and sparse

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

Table 1: **Correspondence between variational and original parameters.**

supplementary feature matrix using the model's independence assumptions:

$$\begin{aligned} p(z, y, \eta, \Phi, \Lambda, \beta, \Omega | \mathbf{M}, \gamma, \kappa, \alpha, \iota, \xi, \beta) = & \left[\prod_{a=1}^{b} p(\Phi_a | \alpha) \right] \left[\prod_{i=1}^{n} p(\eta | \mu, \Sigma) p(\Lambda^{(i)} | \beta^i) p(\beta^i | \gamma, \kappa) \right. \\ & \times p(\Omega^{(i)} | \mathbf{M}^{(i)}, \xi) \left[\prod_{j=1}^{t_i} p(y_j^{(i)} | z_j^{(i)}, \Omega_j^{(i)}, \Lambda^{(i)}, \Phi, \varpi) \right. \\ & \left. \left. \times p(z_j^{(i)} | \eta) \right] \right] \end{aligned} \tag{3.1}$$

By denoting all the parameters as Θ and variables as $\mathbf{V}$ while omitting the hyperparameters, we obtain the following posterior expression:

$$p(\Theta, \mathbf{V} | \mathbf{Y}, \mathbf{M}) = \frac{p(\mathbf{Y}, \mathbf{M}, \Theta, \mathbf{V})}{p(\mathbf{Y}, \mathbf{M})} \tag{3.2}$$

Unfortunately, the exact posterior distribution of the latent variables is computationally intractable. The numerator is easy to compute for any configuration of the hidden variables and parameters. The problem is the denominator, which is the marginal probability of the data:

$$p(\mathbf{Y}, \mathbf{M}) = \int_{\Theta} \int_{\mathbf{V}} p(\mathbf{Y}, \mathbf{M}, \Theta, \mathbf{V}) \tag{3.3}$$

Computing the marginal requires a complicated integral over n examples of $|\Theta|$ parameters and another integral over the $|\mathbf{V}|^n$ configurations multiplied by the size of each variable in $\mathbf{V}$. As such, we appeal to variational inference, which has been extensively employed in many complex probabilistic models. Examples include latent Dirichlet allocation (LDA) [8], sparse LDA [26], supervised topic models [24], mixed membership stochastic blockmodels [1], nested hierarchical Dirichlet process [27], and many others. The main intuition behind variational methods is to first posit a family of distributions over the hidden parameters and variables that are indexed by a set of free parameters, and then fitting the parameters to find the member of the family that is closest to the true posterior of interest in Eq. 3.2. The closeness is commonly measured using Kullback–Leibler (KL) divergence [21]. The resulting variational distribution is simpler than the true posterior so that the solution can be approximated.

However, directly minimizing the KL divergence is infeasible due to the same reason that the posterior is difficult to compute, but, we can optimize an objective function that is equal to the negative KL divergence up to a constant. This is known as the evidence lower bound (ELBO), a lower bound on the logarithm of the marginal probability in Eq. 3.3, i.e., $\log p(\mathbf{Y}, \mathbf{M})$. This ELBO can be defined using Jensen's inequality on a variational distribution over the hidden variables $q(\Theta, \mathbf{V})$ as:

$$\begin{aligned} \log p(\mathbf{Y}, \mathbf{M}) &= \log \int_{\Theta} \int_{V} p(\mathbf{Y}, \mathbf{M}, \Theta, \mathbf{V}) \\ &= \log \int_{\Theta} \int_{\mathbf{V}} p(\mathbf{Y}, \mathbf{M}, \Theta, V) \frac{q(\Theta, \mathbf{V})}{q(\Theta, \mathbf{V})} \\ &= \log(\mathbb{E}_q[\frac{p(\mathbf{Y}, \mathbf{M}, \Theta, \mathbf{V})}{q(\Theta, \mathbf{V})}]) \\ &\geq \mathbb{E}_q[\log p(\mathbf{Y}, \mathbf{M}, \Theta, \mathbf{V})] + \mathbb{H}(q) \\ &\triangleq \mathcal{L}(q) \end{aligned} \tag{3.4}$$

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 that is likely under the model, keeping both the priors and data in mind through the joint distribution. 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" [7]. Both of these terms depend on $q(\Theta, \mathbf{V})$, the variational distribution of the hidden variables.

The simplest variational family of distributions is the mean-field family where each hidden variable/parameter is fully-factorized and governed by its own parameter. This allows us to tractably

optimize the parameters to find a local minimum of the KL divergence. For SPREAT, the mean-field variational distribution is expressed as:

$$q(\eta, \Lambda, z, d, \beta, \Phi, \Omega) = \prod_{a=1}^{b} q(\Phi_a|\phi_a)\Bigg[\prod_{i=1}^{n} q(\eta^{(i)}|\nu, \zeta^2)q(\Lambda^{(i)}|\lambda^{(i)})q(\Omega^{(i)}|\omega^{(i)}) \prod_{j=1}^{j=t_i} q(z_j^{(i)}|\varsigma_j^{(i)})\Bigg] \tag{3.5}$$

where $\phi, \nu, \zeta^2, \lambda, \omega$ and ς are variational free parameters. Table 1 shows the correspondence between variational and the original parameters.

Taking together, the first term in Eq. 3.4, $\mathbb{E}_q[\log p(\mathbf{Y}, \mathbf{M}, \Theta, V)]$, can be decomposed into:

$$\begin{aligned}\mathbb{E}_q[\log p(\mathbf{Y}, \mathbf{M}, \Theta, V)] = &\sum_{a=1}^{a=b} \mathbb{E}_q[\log p(\Phi_a|\alpha)] + \sum_{i=1}^{i=n}\Bigg(\mathbb{E}_q[\log p(\eta|\mu, \Sigma)] \\ &+ \mathbb{E}_q[\log p(\Lambda^{(i)}|\beta^i)] + \mathbb{E}_q[\log p(\beta^i|\gamma, \kappa)] \\ &+ \mathbb{E}_q[\log p(\Omega^{(i)}|\mathbf{M}^{(i)}, \xi)] \\ &+ \sum_{j=1}^{j=t_i}\Big(\mathbb{E}_q[\log p(y_j^{(i)}|z_j^{(i)}, \Omega_j^{(i)}, \Lambda^{(i)}, \Phi, \varpi)] + \mathbb{E}_q[p(z_j^{(i)}|\eta)]\Big)\Bigg)\end{aligned} \tag{3.6}$$

And, the second term $\mathbb{H}(q)$ in Eq. 3.4 can be defined as:

$$\begin{aligned}\mathbb{H}(q) = -\sum_{a=1}^{a=b} \mathbb{E}_q[\log q(\Phi_a|\phi_a)] - \sum_{i=1}^{i=n}\Bigg(&\mathbb{E}_q[\log q(\eta^{(i)}|\nu, \zeta^2)] + \mathbb{E}_q[\log q(\Lambda^{(i)}|\lambda^{(i)})] \\ &+ \mathbb{E}_q[\log q(\Omega^{(i)}|\omega^{(i)})] + \sum_{j=1}^{j=t_i} \mathbb{E}_q[\log q(z_j^{(i)}|\varsigma_j^{(i)})]\Bigg)\end{aligned} \tag{3.7}$$

3.1 Variational Lower Bound

Given Eq. 3.6, we derive expressions for each term:

1. For the concept distribution over features, which are Dirichlet-distributed,

$$\begin{aligned}\mathbb{E}_q[\log p(\Phi_a|\alpha)] &= \mathbb{E}_q[\log \mathrm{Dir}(\Phi_a|\alpha)] \\ &= \mathbb{E}_q\Bigg[\log\Bigg(\frac{\Gamma(\sum_{j=1}^{j=t}\alpha_j)}{\prod_{j=1}^{j=t}\Gamma(\alpha_j)}\prod_{j=1}^{j=t}\Phi_{a,j}^{\alpha_j-1}\Bigg)\Bigg] \\ &= \mathbb{E}_q\Bigg[\log\Bigg(\frac{\Gamma(\sum_{j=1}^{j=t}\alpha_j)}{\prod_{j=1}^{j=t}\Gamma(\alpha_j)}\Bigg) + \sum_{j=1}^{j=t}\log\Phi_{a,j}^{\alpha_j-1}\Bigg] \\ &= \log\Gamma\Big(\sum_{j=1}^{j=t}\alpha_j\Big) - \sum_{j=1}^{j=t}\log\Gamma(\alpha_j) + \sum_{j=1}^{j=t}(\alpha_j - 1)\mathbb{E}_q[\log\Phi_{a,j}]\end{aligned} \tag{3.8}$$

2. For the concepts probabilities for each example, which are Gaussian distributed,

$$\begin{aligned}\mathbb{E}_q[\log p(\eta|\mu, \Sigma)] &= \mathbb{E}_q\Big[\log\Big(\mathcal{N}(\eta|\mu, \Sigma)\Big)\Big] \\ &= \mathbb{E}_q\Big[\Big(\frac{1}{2}\log|\Sigma^{-1}| - \frac{b}{2}\log 2\pi - \frac{1}{2}(\eta-\mu)^\top\Sigma^{-1}(\eta-\mu)\Big)\Big] \\ &= \frac{1}{2}\log|\Sigma^{-1}| - \frac{b}{2}\log 2\pi \\ &\quad - \frac{1}{2}\Big(\mathrm{tr}(\mathrm{diag}(\zeta^2)\Sigma^{-1}) + (\nu-\mu)^\top\Sigma^{-1}(\nu-\mu)\Big)\end{aligned} \tag{3.9}$$

3. For the focused concept distributions for each example, which are Bernoulli distributed,

$$\begin{aligned}\mathbb{E}_q[\log p(\Lambda^{(i)}|\beta^{(i)})] &= \mathbb{E}_q\Big[\log\mathrm{Bernoulli}(\Lambda^{(i)}|\beta^{(i)})\Big] \\ &= \mathbb{E}_q\Bigg[\log\Bigg(\prod_{a=1}^{a=b}\beta_a^{(i),\Lambda_a^{(i)}}(1-\beta_a)^{(i),1-\Lambda_a^{(i)}}\Bigg)\Bigg] \\ &= \sum_{a=1}^{a=b}\Big(\lambda_a^{(i)}\log\beta_a^{(i)} + (1-\lambda_a^{(i)})\log(1-\beta_a^{(i)})\Big)\end{aligned} \tag{3.10}$$

4. For selecting a set of focused concepts for each example, which are beta distributed,

$$\begin{aligned}\mathbb{E}_q[\log p(\beta^{(i)}|\gamma,\kappa))] &= \mathbb{E}_q\Big[\log \mathrm{Beta}(\beta^{(i)}|\gamma,\kappa)\Big] \\ &= \sum_{a=1}^{a=b}\Big((\gamma-1)\log(\beta_a^{(i)}) + (\kappa-1)\log(1-\beta_a^{(i)}) - \log(B(\gamma,\kappa)\Big)\end{aligned} \tag{3.11}$$

5. For the hypothetical feature distributions for each example, which are Dirichlet distributed,

$$\begin{aligned}\mathbb{E}_q[\log p(\Omega_i|\mathbf{M}^{(i)},\xi)] &= \mathbb{E}_q\Big[\log\Big(\frac{\Gamma(\sum_{j=1}^{j=t}\xi_j+\mathbf{M}_j^{(i)})}{\prod_{j=1}^{j=t}\Gamma(\xi_j+\mathbf{M}_j^{(i)})}\prod_{j=1}^{j=t}\Omega_j^{(i),\xi_j+\mathbf{M}_j^{(i)}-1}\Big)\Big] \\ &= \log\Gamma\Big(\sum_{j=1}^{j=t}\xi_j+\mathbf{M}_j^{(i)}\Big) - \sum_{j=1}^{j=t}\log\Gamma(\xi_j+\mathbf{M}_j^{(i)}) \\ &\quad + \sum_{j=1}^{j=t}(\xi_j+\mathbf{M}_j^{(i)}-1)\mathbb{E}_q[\log\Omega_j^{(i)}]\end{aligned} \tag{3.12}$$

6. For the feature assignments from both concept-feature and hypothetical feature distributions,

$$\begin{aligned}\mathbb{E}_q[\log p(y_j^{(i)}|z_j^{(i)},\Omega_j^{(i)},\Lambda^{(i)},\Phi,\varpi)] &= \mathbb{E}_q\Big[\log\Big(\prod_{c=1}^{c=t}\prod_{a=1}^{a=b}\varpi\Phi_{a,c}^{\mathbb{I}(a=z_{a,j}^{(i)}\wedge\Lambda_a^{(i)},c=y_j^{(i)}),(1-\Omega_c^{(i)}))}\Big)\Big] \\ &= \log\varpi + \sum_{c=1}^{c=t}\sum_{a=1}^{a=b}\Big(y_{j,c}^{(i)}\varsigma_{a,j}^{(i)}\lambda_a^{(i)}\mathbb{E}_q[(1-\Omega_c^{(i)})]\mathbb{E}_q[\log\Phi_{a,j}]\Big)\end{aligned} \tag{3.13}$$

7. For the concept assignments over features, the expectation of the log probability of the latent concepts is given by:

$$\begin{aligned}\mathbb{E}_q[\log p(z_j^{(i)}|\eta)] &= \mathbb{E}_q\Big[\log\Big(\frac{\exp(\eta^\top(\mathrm{diag}(z_j^{(i)})))}{\sum_{k=1}^{k=b}\exp(\eta_k)}\Big)\Big] \\ &= \mathbb{E}_q\Big[\eta^\top(\mathrm{diag}(z_j^{(i)}))\Big] - \mathbb{E}_q\Big[\log\Big(\sum_{k=1}^{k=b}\exp(\eta_k)\Big)\Big] \\ &= \sum_{a=1}^{a=b}\nu_a\varsigma_{a,j}^{(i)} - \mathbb{E}_q\Big[\log\Big(\sum_{k=1}^{k=b}\exp(\eta_k)\Big)\Big]\end{aligned} \tag{3.14}$$

The second term is hard to compute, hence, we appeal to the solution suggested by [6] in order to pertain the tightest lower bound on $-\mathbb{E}_q\Big[\log\Big(\sum_{k=1}^{k=b}\exp(\eta_k)\Big)\Big]$ using a first-order Taylor expansion. Because the function $-\log$ is convex, a first-order Taylor expansion about the point ϱ, a variational parameter, produces the following inequality:

$$\begin{aligned}-\mathbb{E}_q\Big[\log\Big(\sum_{k=1}^{k=b}\exp(\eta_k)\Big)\Big] &\geq -\log\varrho - \frac{\Big(\sum_{k=1}^{k=b}\mathbb{E}_q[\exp(\eta_k)]\Big)-\varrho}{\varrho} \\ &= 1-\log\varrho - \Big(\sum_{k=1}^{k=b}\mathbb{E}_q[\exp(\eta_k)]\Big)\varrho^{-1}\end{aligned} \tag{3.15}$$

Plugging back the results into Eq. 3.14, we obtain:

$$\mathbb{E}_q[\log p(z_j^{(i)}|\eta)] \approx 1-\log\varrho + \sum_{a=1}^{a=b}\nu_a\varsigma_{a,j}^{(i)} - \Big(\sum_{k=1}^{k=b}\mathbb{E}_q[\exp(\eta_k)]\Big)\varrho^{-1} \tag{3.16}$$

Now, for the entropy $\mathbb{H}(q)$ in Eq. 3.7, we decompose their expectations as:

1. For the concept-feature distributions, which are Dirichlet distributed,

$$\begin{aligned}\mathbb{E}_q[\log q(\Phi_a|\phi_a)] &= \mathbb{E}_q[\log\mathrm{Dir}(\Phi_a|\phi_a)] \\ &= \mathbb{E}_q\Big[\log\Big(\frac{\Gamma(\sum_{j=1}^{j=t}\phi_{a,j})}{\prod_{j=1}^{j=t}\Gamma(\phi_{a,j})}\prod_{j=1}^{j=t}\Phi_{a,j}^{\phi_{a,j}-1}\Big)\Big] \\ &= \log\Gamma\Big(\sum_{j=1}^{j=t}\phi_{a,j}\Big) - \sum_{j=1}^{j=t}\log\Gamma(\phi_{a,j}) + \sum_{j=1}^{j=t}(\phi_{a,j}-1)\mathbb{E}_q[\log\Phi_{a,j}]\end{aligned} \tag{3.17}$$

2. For the concept distributions, which are Gaussian distributed,

$$\begin{aligned}\mathbb{E}_q[\log q(\eta^{(i)}|\nu,\zeta^2)] =& \mathbb{E}_q\Big[\log\prod_{a=1}^{a=b}\mathcal{N}(\eta_a^{(i)}|\nu_a,\zeta_a^2)\Big]\\ =& -\sum_{a=1}^{a=b}\frac{1}{2}\Big(\log\zeta_a^2+\log(2\pi)+1\Big)\end{aligned} \tag{3.18}$$

3. For the concept choice parameter, which are Bernoulli distributed,

$$\begin{aligned}\mathbb{E}_q[\log q(\Lambda^{(i)}|\lambda^{(i)})] &= \mathbb{E}_q[\log\prod_{a=1}^{a=b}\text{Bernoulli}(\Lambda_a^{(i)}|\lambda_a^{(i)})]\\ &= \sum_{a=1}^{a=b}\Big(\lambda_a^{(i)}\log\lambda_a^{(i)}+(1-\lambda_a^{(i)})\log(1-\lambda_a^{(i)})\Big)\end{aligned} \tag{3.19}$$

4. For the supplementary feature distributions over examples, which are Dirichlet distributed,

$$\begin{aligned}\mathbb{E}_q[\log q(\Omega^{(i)}|\omega^{(i)})] &= \mathbb{E}_q[\log\text{Dir}(\Omega^{(i)}|\omega^{(i)})]\\ &= \mathbb{E}_q\Big[\log\Big(\frac{\Gamma(\sum_{j=1}^{j=t}\omega_j^{(i)})}{\prod_{j=1}^{j=t}\Gamma(\omega_j^{(i)})}\prod_{j=1}^{j=t}\Omega_j^{(i),\omega_j^{(i)}-1}\Big)\Big]\\ &= \log\Gamma\Big(\sum_{j=1}^{j=t}\omega_j^{(i)}\Big)-\sum_{j=1}^{j=t}\log\Gamma(\omega_j^{(i)})+\sum_{j=1}^{j=t}(\omega_j^{(i)}-1)\mathbb{E}_q[\log\Omega_j^{(i)}]\end{aligned} \tag{3.20}$$

5. For the feature assignments over examples, which are multinomially distributed,

$$\mathbb{E}_q[\log q(z_j^{(i)}|\varsigma_j^{(i)})] = \mathbb{E}_q\Big[\log\prod_{a=1}^{a=b}(\varsigma_{a,j}^{(i)})^{z_{a,j}^{(i)}}\Big] = \sum_{a=1}^{a=b}\varsigma_{a,j}^{(i)}\log\varsigma_{a,j}^{(i)} \tag{3.21}$$

where the exceptions of all the above equations can be derived using:

$$\begin{aligned}\mathbb{E}_q[\log\Phi_{a,j}] &= \Big(\Psi(\phi_{a,j})-\Psi(\sum_{k=1}^{k=t}\phi_{a,k})\Big)\\ \mathbb{E}_q[\log\Omega_j^{(i)}] &= \Big(\Psi(\omega_j^{(i)})-\Psi(\sum_{k=1}^{k=t}\omega_k^{(i)})\Big)\\ \mathbb{E}_q[(1-\Omega_c^{(i)})] &= \frac{1-\omega_c^{(i)}}{\sum_{k=1}^{k=t}(1-\omega_k^{(i)})}\\ \mathbb{E}_q[\exp(\eta_k)] &= \exp(\nu_a+\frac{1}{2}\zeta_a^2)\\ B(\gamma,\kappa) &= \frac{\Gamma(\gamma)\Gamma(\kappa)}{\Gamma(\gamma+\kappa)}\end{aligned}$$

Not that Γ denotes the Gamma function while Ψ is the logarithmic derivative of the Gamma function.

3.2 Merging All the Expectations of the ELBO Terms

Now, by joining all the terms in Section 3.1, the full ELBO can be defined as:

$$
\begin{aligned}
\mathcal{L}(q) = & \sum_{a=1}^{a=b}\Bigg(\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)\Big(\Psi(\phi_{a,j}) - \Psi(\sum_{k=1}^{k=t}\phi_{a,k})\Big)\Bigg) \\
& + \sum_{i=1}^{i=n}\Bigg(\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)\Bigg) \\
& + \sum_{i=1}^{i=n}\sum_{a=1}^{a=b}\Bigg(\lambda_a^{(i)}\log(\beta_a^{(i)}) + (1-\lambda_a^{(i)})\log(1-\beta_a^{(i)})\Bigg) \\
& + \sum_{i=1}^{i=n}\sum_{a=1}^{a=b}\Bigg((\gamma-1)\log(\beta_a^{(i)}) + (\kappa-1)\log(1-\beta_a^{(i)}) - \log(B(\gamma,\kappa)\Bigg) \\
& + \sum_{i=1}^{i=n}\Bigg(\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)\Big(\Psi(\omega_j^{(i)}) - \Psi(\sum_{k=1}^{k=t}\omega_k^{(i)})\Big)\Bigg) \\
& + \sum_{i=1}^{i=n}\sum_{j=1}^{j=t_i}\Bigg(\log\varpi + \sum_{c=1}^{c=t}\sum_{a=1}^{a=b}\Big(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)})}\Big(\Psi(\phi_{a,j}) - \Psi(\sum_{k=1}^{k=t}\phi_{a,k})\Big)\Big)\Bigg) \\
& + \sum_{i=1}^{i=n}\sum_{j=1}^{j=t_i}\Bigg(1 - \log\varrho + \sum_{a=1}^{a=b}\nu_a\varsigma_{a,j}^{(i)} - \Big(\sum_{k=1}^{k=b}\exp(\nu_k+\frac{1}{2}\zeta_k^2)\Big)\varrho^{-1}\Bigg) \\
& - \sum_{a=1}^{a=b}\Bigg(\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)\Big(\Psi(\phi_{a,j}) - \Psi(\sum_{k=1}^{k=t}\phi_{a,k})\Big)\Bigg) \\
& + \sum_{i=1}^{i=n}\sum_{a=1}^{a=b}\Bigg(\frac{1}{2}\Big(\log\zeta_a^2 + \log(2\pi) + 1\Big)\Bigg) \\
& - \sum_{i=1}^{i=n}\sum_{a=1}^{a=b}\Bigg(\lambda_a^{(i)}\log\lambda_a^{(i)} + (1-\lambda_a^{(i)})\log(1-\lambda_a^{(i)})\Bigg) \\
& - \sum_{i=1}^{i=n}\Bigg(\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)\Big(\Psi(\omega_j^{(i)}) - \Psi(\sum_{k=1}^{k=t}\omega_k^{(i)})\Big)\Bigg) \\
& - \sum_{i=1}^{i=n}\sum_{j=1}^{j=t_i}\sum_{a=1}^{a=b}\varsigma_{a,j}^{(i)}\log\varsigma_{a,j}^{(i)}
\end{aligned}
\tag{3.22}
$$

4 Optimizing the ELBO Terms

In this section, we maximize the bound in Eqs 3.6 & 3.7 with respect to each variational parameters using coordinate ascent updates, which optimizes each variational parameter while holding the remaining variables fixed. Practically, a more convenient way is to apply the mini-batch gradient approach, as outlined in [17] that alternates between subsampling a batch of examples and updating each variational parameter, after being scaled by a learning rate. This structure of learning assists us to approximate the posterior with massive examples, making the complete problem computationally scalable.

1. **Optimizing w.r.t. ς.** Gathering only the terms in the bound that contain ς, we obtain:

$$
\begin{aligned}
\mathcal{L}(q)_{[\varsigma]} = & \sum_{i=1}^{i=n}\sum_{j=1}^{j=t_i}\sum_{c=1}^{c=t}\sum_{a=1}^{a=b}y_{j,c}^{(i)}\varsigma_{a,j}^{(i)}\lambda_a^{(i)}\frac{1-\omega_c^{(i)}}{\sum_{k=1}^{k=t}(1-\omega_k^{(i)})}\Big(\Psi(\phi_{a,j}) - \Psi(\sum_{k=1}^{k=t}\phi_{a,k})\Big) \\
& + \sum_{i=1}^{i=n}\sum_{j=1}^{j=t_i}\sum_{a=1}^{a=b}\nu_a\varsigma_{a,j}^{(i)} - \sum_{i=1}^{i=n}\sum_{j=1}^{j=t_i}\sum_{a=1}^{a=b}\varsigma_{a,j}^{(i)}\log\varsigma_{a,j}^{(i)}
\end{aligned}
\tag{4.1}
$$

Taking derivatives w.r.t. $\varsigma_{a,j}^{(i)}$, we obtain:

$$\begin{aligned}\frac{\partial \mathcal{L}(q)_{[\varsigma]}}{\partial \varsigma_{a,j}^{(i)}} =& \sum_{c=1}^{c=t} y_{j,c}^{(i)} \lambda_a^{(i)} \frac{1-\omega_c^{(i)}}{\sum_{k=1}^{k=t}(1-\omega_k^{(i)})} \Big(\Psi(\phi_{a,j}) - \Psi(\sum_{k=1}^{k=t}\phi_{a,k})\Big) \\ &+ \nu_a - \log \varsigma_{a,j}^{(i)} - 1\end{aligned} \tag{4.2}$$

The analytical expression of the variational concept assignment $q(\varsigma)$ for each feature y_j and concept a is not amenable due to the non-conjugacy of logistic-normal with latent variables. Instead, we approximate the solution according to:

$$\varsigma_{a,j}^{(i)} \propto \exp\left(\sum_{c=1}^{c=t} y_{j,c}^{(i)} \lambda_a^{(i)} \frac{1-\omega_c^{(i)}}{\sum_{k=1}^{k=t}(1-\omega_k^{(i)})} \Big(\Psi(\phi_{a,j}) - \Psi(\sum_{k=1}^{k=t}\phi_{a,k})\Big) + \nu_a - 1\right) \tag{4.3}$$

where $\Psi(.)$ is the digamma function. Observe how the variational parameter $\omega_*^{(i)}$ serves as the smoothing term in selecting concepts for each feature, either from $\mathbf{M}_i$ or from $\mathcal{P}$, when $\omega_c^{(i)} > 0$. However, if $\omega_c^{(i)} = 0$, then $\varsigma_{a,j}^{(i)}$ is updated based on $\phi_{a,j}$.

2. **Optimizing w.r.t. ν.** Collecting only the terms in the bound that contain ν gives,

$$\begin{aligned}\mathcal{L}(q)_{[\nu]} =& \sum_{i=1}^{i=n}\Bigg(-\frac{1}{2}(\nu-\mu)^\top \Sigma^{-1}(\nu-\mu) + \sum_{j=1}^{j=t_i}\sum_{a=1}^{a=b}\nu_a \varsigma_{a,j}^{(i)} \\ &- \sum_{j=1}^{j=t_i}\Big(\sum_{k=1}^{k=b}\exp(\nu_k + \frac{1}{2}\zeta_k^2)\Big)\varrho^{-1}\Bigg)\end{aligned} \tag{4.4}$$

Taking derivatives w.r.t. ν_a for each concept a, we obtain:

$$\frac{\partial \mathcal{L}(q)_{[\nu]}}{\partial \nu_a} = -\Sigma^{-1}(\nu-\mu) + \sum_{j=1}^{j=t_i}\varsigma_{a,j}^{(i)} - \Big(\exp(\nu_a + \frac{1}{2}\zeta_a^2)\Big) t_i \varrho^{-1} \tag{4.5}$$

where ϱ is another variational parameter, as in CTM [6]. However, the above equation in hard to optimize, instead, we use a conjugate gradient algorithm.

3. **Optimizing w.r.t. ζ^2.** By symmetry, we gather all the terms that has ζ^2 from Eq. 3.22:

$$\begin{aligned}\mathcal{L}(q)_{[\zeta^2]} =& -\frac{1}{2}\sum_{i=1}^{i=n}\mathrm{tr}\Big(\mathrm{diag}(\zeta^2)\Sigma^{-1}\Big) - \sum_{i=1}^{i=n}\sum_{j=1}^{j=t_i}\Big(\sum_{k=1}^{k=b}\exp\Big(\nu_k + \frac{1}{2}\zeta_k^2\Big)\Big)\varrho^{-1} \\ &+ \frac{1}{2}\sum_{i=1}^{i=n}\sum_{a=1}^{a=b}\log\zeta_a^2\end{aligned} \tag{4.6}$$

Taking derivatives w.r.t. ζ_a^2 for each concept a, we obtain:

$$\frac{\partial \mathcal{L}(q)_{[\zeta^2]}}{\partial \zeta_a^2} = -\frac{1}{2}\left(\Sigma_{a,a}^{-1} + t_i\varrho^{-1}\exp\Big(\nu_a + \frac{1}{2}\zeta_a^2\Big) - \frac{1}{\zeta_a^2}\right) \tag{4.7}$$

Again, there is no analytic solution. We use Newton's method for each coordinate, constrained such that $\zeta_a \in \mathbb{R}_{>0}$.

4. **Optimizing w.r.t. ϱ.** Extracting the terms involving ϱ in the bound gives,

$$\mathcal{L}(q)_{[\varrho]} = -\sum_{i=1}^{i=n}\sum_{j=1}^{j=t_i}\log\varrho - \sum_{i=1}^{i=n}\sum_{j=1}^{j=t_i}\Big(\sum_{k=1}^{k=b}\exp(\nu_k + \frac{1}{2}\zeta_k^2)\Big)\varrho^{-1} \tag{4.8}$$

Taking derivatives w.r.t. ϱ, we obtain:

$$\frac{\partial \mathcal{L}(q)_{[\varrho]}}{\partial \varrho} = -t_i n\varrho^{-1} + t_i n\Big(\sum_{k=1}^{k=b}\exp(\nu_k + \frac{1}{2}\zeta_k^2)\Big)\varrho^{-2} \tag{4.9}$$

Equating the above formula to zero to obtain a maximum, we get:

$$\varrho = \sum_{k=1}^{k=b}\exp(\nu_k + \frac{1}{2}\zeta_k^2) \tag{4.10}$$

5. **Optimizing w.r.t. ω.** Isolating only the terms in the bound that contain variational background feature distributions $q(\omega)$, we obtain:

$$\begin{aligned}\mathcal{L}(q)_{[\omega]} =& \sum_{i=1}^{i=n}\sum_{j=1}^{j=t}\Psi(\omega_j^{(i)})(\xi_j + \mathbf{M}_j^{(i)} - \omega_j^{(i)}) - \sum_{i=1}^{i=n}\sum_{j=1}^{j=t}\Psi(\sum_{k=1}^{k=t}\omega_k^{(i)})(\xi_j + \mathbf{M}_j^{(i)} - \omega_j^{(i)}) \\ &+ \sum_{i=1}^{i=n}\sum_{j=1}^{j=t_i}\sum_{c=1}^{c=t}\sum_{a=1}^{a=b}\Big(y_{j,c}^{(i)}\varsigma_{a,j}^{(i)}\lambda_a^{(i)}\frac{1-\omega_c^{(i)}}{\sum_{k=1}^{k=t}(1-\omega_k^{(i)})}\Big(\Psi(\phi_{a,j}) - \Psi(\sum_{k=1}^{k=t}\phi_{a,k})\Big)\Big) \\ &+ \sum_{i=1}^{i=n}\sum_{j=1}^{j=t}\log\Gamma(\omega_j^{(i)}) - \sum_{i=1}^{i=n}\log\Gamma\Big(\sum_{j=1}^{j=t}\omega_j^{(i)}\Big)\end{aligned} \tag{4.11}$$

Taking derivatives w.r.t. $\omega_c^{(i)}$ gives

$$\begin{aligned}\frac{\partial\mathcal{L}(q)_{[\omega]}}{\partial\omega_c^{(i)}} =& \Big(\Psi'(\omega_c^{(i)}) - \Psi'(\sum_{k=1}^{k=t}\omega_k^{(i)})\Big)(\xi_c + \mathbf{M}_c^{(i)} - \omega_c^{(i)}) \\ &- \Big(\frac{1-\omega_c^{(i)} - \sum_{k=1}^{k=t}(1-\omega_k^{(i)})}{(\sum_{k=1}^{k=t}(1-\omega_k^{(i)}))^2}\Big)\sum_{j=1}^{j=t_i}\sum_{a=1}^{a=b}y_{j,c}^{(i)}\varsigma_{a,j}^{(i)}\lambda_a^{(i)}\Big(\Psi(\phi_{a,j}) - \Psi(\sum_{k=1}^{k=t}\phi_{a,k})\Big)\end{aligned} \tag{4.12}$$

Setting it's derivatives to zero does not lead to a closed-form solution, instead, we approximate $\omega_c^{(i)}$ for each sample i according to:

$$\begin{aligned}\omega_c^{(i)} \propto& \xi_c + \mathbf{M}_c^{(i)} - \Big(\frac{1-\omega_c^{(i)} - \sum_{k=1}^{k=t}(1-\omega_k^{(i)})}{(\sum_{k=1}^{k=t}(1-\omega_k^{(i)}))^2}\Big)\sum_{j=1}^{j=t_i}\sum_{a=1}^{a=b}y_{j,c}^{(i)}\varsigma_{a,j}^{(i)}\lambda_a^{(i)} \\ &\times \Big(\Psi(\phi_{a,j}) - \Psi(\sum_{k=1}^{k=t}\phi_{a,k})\Big)\end{aligned} \tag{4.13}$$

6. **Optimizing w.r.t. λ.** Collecting the terms that contain λ, we obtain:

$$\begin{aligned}\mathcal{L}(q)_{[\lambda]} =& \sum_{i=1}^{i=n}\sum_{a=1}^{a=b}\lambda_a^{(i)}(\log(\beta_a^{(i)}) - \log\lambda_a^{(i)}) \\ &+ \sum_{i=1}^{i=n}\sum_{a=1}^{a=b}(1-\lambda_a^{(i)})\Big(\log(1-\beta_a^{(i)}) - \log(1-\lambda_a^{(i)})\Big)\end{aligned} \tag{4.14}$$

Taking derivatives w.r.t. $\lambda_a^{(i)}$, we obtain:

$$\frac{\partial\mathcal{L}(q)_{[\lambda]}}{\partial\lambda_a^{(i)}} = \log(1-\lambda_a^{(i)}) - \log\lambda_a^{(i)} + \log(\beta_a^{(i)}) - \log(1-\beta_a^{(i)}) \tag{4.15}$$

Equating the above formula to zero to obtain a maximum, we get the canonical parameterisation of the Bernoulli distribution:

$$\theta = \log\Big(\frac{\lambda_a^{(i)}}{1-\lambda_a^{(i)}}\Big) = \log(\beta_a^{(i)}) - \log(1-\beta_a^{(i)}) \tag{4.16}$$

Therefore, we get the following updates:

$$\lambda_a^{(i)} = \frac{1}{1+\exp^{-\theta}} \tag{4.17}$$

7. **Optimizing w.r.t. ϕ.** Finally, the optimal solution of the variational concept feature distribution $q(\Phi_a|\phi_a)$ for each concept a is obtained by isolating terms involved in the bound Eq. 3.4:

$$\begin{aligned}\mathcal{L}(q)_{[\phi]} =& \sum_{a=1}^{a=b}\sum_{c=1}^{c=t}\Psi(\phi_{a,c})\Big(\alpha_c - \phi_{a,c} + \sum_{i=1}^{i=n}\sum_{j=1}^{j=t_i}y_{j,c}^{(i)}\varsigma_{a,j}^{(i)}\lambda_a^{(i)}\frac{1-\omega_c^{(i)}}{\sum_{k=1}^{k=t}(1-\omega_k^{(i)})}\Big) \\ &- \sum_{a=1}^{a=b}\sum_{j=1}^{j=t}\Psi(\sum_{k=1}^{k=t}\phi_{a,k})\Big(\alpha_c - \phi_{a,c} + \sum_{i=1}^{i=n}\sum_{j=1}^{j=t_i}y_{j,c}^{(i)}\varsigma_{a,j}^{(i)}\lambda_a^{(i)}\frac{1-\omega_c^{(i)}}{\sum_{k=1}^{k=t}(1-\omega_k^{(i)})}\Big) \\ &- \sum_{a=1}^{a=b}\log\Gamma\Big(\sum_{j=1}^{j=t}\phi_{a,j}\Big) + \sum_{a=1}^{a=b}\sum_{j=1}^{j=t}\log\Gamma(\phi_{a,j})\end{aligned} \tag{4.18}$$

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
  $s = s + 1$;
  Sample a minibatch randomly $\mathcal{B} \subset \mathcal{P}$;
  for $i \in \mathcal{B}$ **do**
    repeat
      Update $\varsigma^{(i)}$ with Eq. 4.3;
      Update $\nu^{(i)}$ with Eq. 4.5 using conjugate gradient algorithm;
      Update $\zeta^{2,(i)}$ with Eq. 4.7 using Newton's method;
      Update $\varrho^{(i)}$ with Eq. 4.10;
      Update $\omega^{(i)}$ with Eq. 4.13;
      Update $\lambda^{(i)}$ with Eq. 4.17;
    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. 4.20;
    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: Stochastic variational inference for SPREAT

After taking derivatives w.r.t. $\phi_{a,c}$, we obtain:

$$\begin{aligned}\frac{\partial \mathcal{L}(q)_{[\phi]}}{\partial \phi_{a,c}} =& \Psi'(\phi_{a,c})\Big(\alpha_c - \phi_{a,c} + \sum_{i=1}^{i=n}\sum_{j=1}^{j=t_i} y_{j,c}^{(i)}\varsigma_{a,j}^{(i)}\lambda_a^{(i)}\frac{1-\omega_c^{(i)}}{\sum_{k=1}^{k=t}(1-\omega_k^{(i)})}\Big) \\ & - \Psi'(\sum_{k=1}^{k=t}\phi_{a,k})\Big(\alpha_c - \phi_{a,c} + \sum_{i=1}^{i=n}\sum_{j=1}^{j=t_i} y_{j,c}^{(i)}\varsigma_{a,j}^{(i)}\lambda_a^{(i)}\frac{1-\omega_c^{(i)}}{\sum_{k=1}^{k=t}(1-\omega_k^{(i)})}\Big)\end{aligned} \tag{4.19}$$

Equating the above formula to zero to obtain a maximum, we get:

$$\phi_{a,c} = \alpha_c + \sum_{i=1}^{i=n}\sum_{j=1}^{j=t_i} y_{j,c}^{(i)}\varsigma_{a,j}^{(i)}\lambda_a^{(i)}\frac{1-\omega_c^{(i)}}{\sum_{k=1}^{k=t}(1-\omega_k^{(i)})} \tag{4.20}$$

The variational inference algorithm samples a mini-batch from a collection, and use it to compute the local latent parameters in Eqs 4.3, 4.5, 4.7, 4.10, 4.13, and 4.17 until the evidence lower bound (in Eq. 3.4) converges. Then, the global variational parameter ϕ is updated using the posteriors (β, Λ, η, z, Ω) collected from the previous step in Eq. 4.20, 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. We summarize our variational inference in Algorithm 3.

5 Posterior Predictive Distribution for SPREAT

We apply posterior predictive distribution to evaluate model's fitness, which estimates the distribution of unobserved values given the observed values and parameters trained on a held-out training set [17]. This metric is useful in evaluating models as it avoids comparing bounds of those models. The predictive distribution is formulated as follows:

$$
\begin{aligned}
p(\tilde{\mathbf{Y}}|\mathbf{Y}_{obs},\mathbf{M}) = \int & p(\tilde{\mathbf{Y}}|\Theta,\mathbf{M})p(\Theta|\mathbf{Y}_{obs},\mathbf{M})d\Theta \\
&+ \sum_{i=1}^{i=n}\left(\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)\right) \\
&+ \sum_{i=1}^{i=n}\sum_{a=1}^{a=b}\left(\lambda_a^{(i)}\log(\beta_a^{(i)}) + (1-\lambda_a^{(i)})\log(1-\beta_a^{(i)})\right) \\
&+ \sum_{i=1}^{i=n}\sum_{a=1}^{a=b}\left((\gamma-1)\log(\beta_a^{(i)}) + (\kappa-1)\log(1-\beta_a^{(i)}) - \log(B(\gamma,\kappa)\right) \\
&+ \sum_{i=1}^{i=n}\left(\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)\Big(\Psi(\omega_j^{(i)}) - \Psi(\sum_{k=1}^{k=t}\omega_k^{(i)})\Big)\right) \\
&+ \sum_{i=1}^{i=n}\sum_{j=1}^{j=t_i}\left(\log\varpi + \sum_{c=1}^{c=t}\sum_{a=1}^{a=b}\left(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)})}\Big(\Psi(\phi_{a,j}) - \Psi(\sum_{k=1}^{k=t}\phi_{a,k})\Big)\right)\right) \\
&+ \sum_{i=1}^{i=n}\sum_{j=1}^{j=t_i}\left(1 - \log\varrho + \sum_{a=1}^{a=b}\nu_a\varsigma_{a,j}^{(i)} - \Big(\sum_{k=1}^{k=b}\exp(\nu_k + \frac{1}{2}\zeta_k^2)\Big)\varrho^{-1}\right) \\
&- \sum_{a=1}^{a=b}\left(\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)\Big(\Psi(\phi_{a,j}) - \Psi(\sum_{k=1}^{k=t}\phi_{a,k})\Big)\right) \\
&+ \sum_{i=1}^{i=n}\sum_{a=1}^{a=b}\left(\frac{1}{2}\Big(\log\zeta_a^2 + \log(2\pi) + 1\Big)\right) \\
&- \sum_{i=1}^{i=n}\sum_{a=1}^{a=b}\left(\lambda_a^{(i)}\log\lambda_a^{(i)} + (1-\lambda_a^{(i)})\log(1-\lambda_a^{(i)})\right) \\
&- \sum_{i=1}^{i=n}\left(\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)\Big(\Psi(\omega_j^{(i)}) - \Psi(\sum_{k=1}^{k=t}\omega_k^{(i)})\Big)\right) \\
&- \sum_{i=1}^{i=n}\sum_{j=1}^{j=t_i}\sum_{a=1}^{a=b}\varsigma_{a,j}^{(i)}\log\varsigma_{a,j}^{(i)}
\end{aligned}
\tag{5.1}
$$

6 Experimental Setup

In this section, we describe the experimental settings and outline the materials used to evaluate the performance of CBT. The CBT was written in Python v3 and depends on third party libraries (e.g. Numpy [31]). Unless otherwise specified all tests were conducted on a Linux server using 10 cores of Intel Xeon CPU E5-2650.

6.1 Description of Datasets

We evaluated correlated models using a subset of datasets used in triUMPF [3]: i)- Critical Assessment of Metagenome Interpretation (CAMI) initiative low complexity dataset consisting of 40 genomes [28] that can be obtained from edwards.sdsu.edu/research/cami-challenge-datasets/ and ii)- BioCyc (v20.5 T2 & 3) [10], which consists of 9255 PGDBs (Pathway/Genome Databases) with 1463 distinct pathway labels and is constructed using the Pathway Tools v21 [20].

6.2 Parameter Settings

Unless otherwise mentioned we used the following default settings to train correlated models on BioCyc (v20.5 T2 &3) collection [10]: the pathway distribution over concepts Φ were initialized using

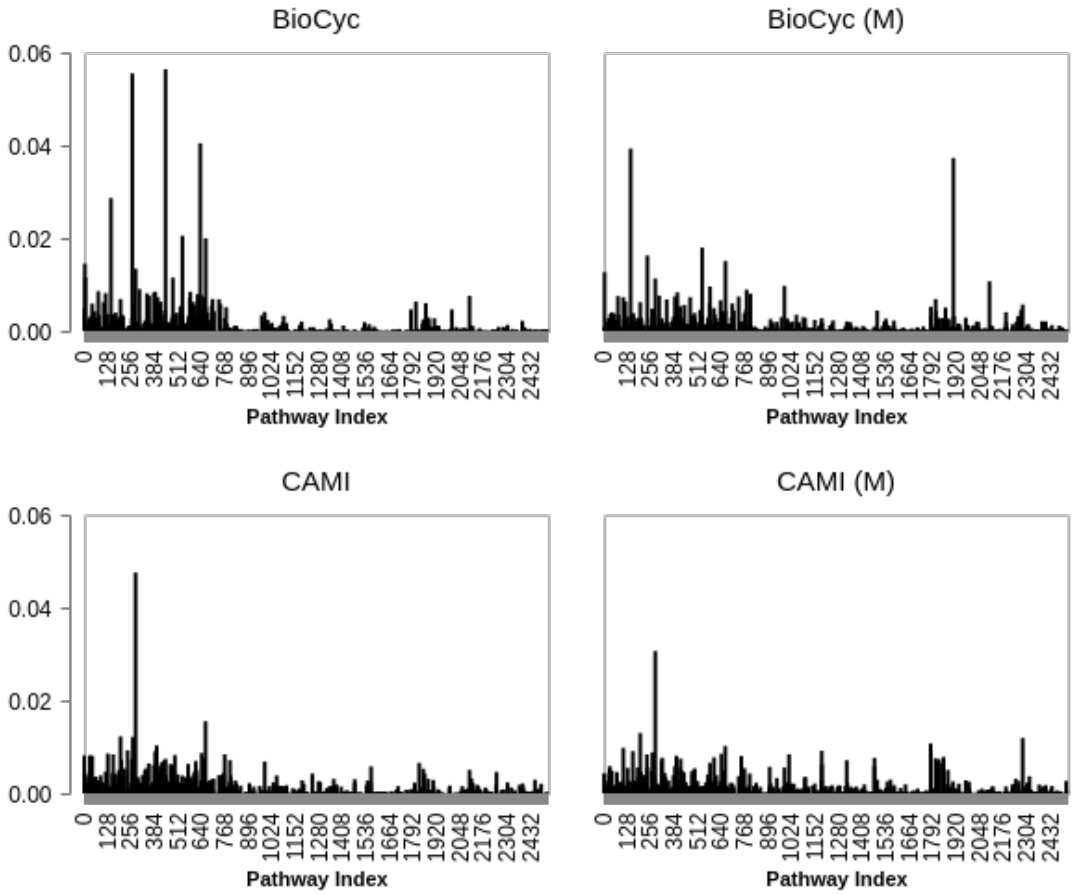


Figure 2: **Illustration of pathway frequency (averaged on all examples) in BioCyc (v20.5 T2 &3) and CAMI data, and their background pathways, indicated by M.**

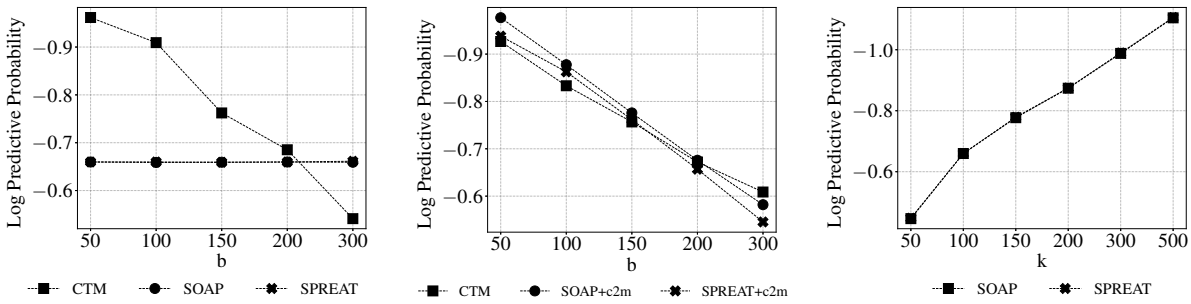
gamma distribution (with shape and scale parameters were fixed to 100 and 1/100, respectively), the forgetting rate to $g = 0.9$, the delay rate to $l = 1$, the batch size to 100, the number of epochs to 3, the number of concepts $b = 200$, top k pathways to 100 (only for SOAP and SPREAT), the Dirichlet hyperparameters α and ξ to 0.0001, and the beta hyperparameters γ and κ to 2 and 3, respectively. The supplementary pathways $\mathbf{M}$ for BioCyc, CAMI, and golden T1 datasets were obtained using mlLGPR [4]. We train mlLGPR (elastic-net) using enzymatic reaction and pathway evidence features where hyperparameters were fixed to their default values. A schematic view of pathway frequency across datasets for BioCyc T2 & 3 and CAMI, along with their augmented pathways is depicted in Fig. 2.

7 Experimental Results and Discussion

We conducted two experimental studies: parameter sensitivity and visualization for correlated models.

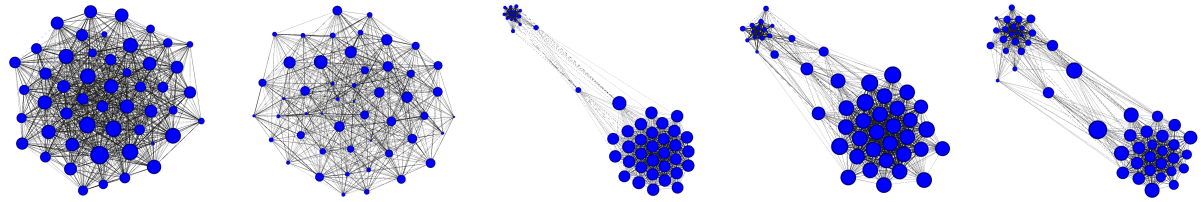
7.1 Sensitivity Analysis of Correlated Models

Experimental setup. A fundamental challenge for the reMap pipeline is to acquire a good distribution of bags and pathways from correlated models for the purpose of relabeling. Following the common practice, here we examined various hyperparameters associated with correlated models. First, we compared the sensitivity of SOAP and SPREAT against CTM by incorporating the background pathways $\mathbf{M}$ while varying the number of bags according to $b \in \{50, 100, 150, 200, 300\}$. Next, we examined the c2m option for SOAP and SPREAT to show that these two models exhibit similar performances as CTM. Finally, we conducted sparsity analysis of bag distribution by varying the cutoff threshold value according to $k \in \{50, 100, 150, 200, 300, 500\}$. For the comparative analysis, we used CAMI as a test data to report the log predictive distribution (Section 5), where a lower score entails higher generalization capability for the associated models.



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

Figure 3: **Log predictive distribution on CAMI data.**



(a) SOAP (#bags: $\sim$ 74; #pathways: $\sim$ 138) (b) SPREAT (#bags: $\sim$ 55; #pathways: $\sim$ 141) (c) CTM (#bags: $\sim$ 50; #pathways: $\sim$ 69) (d) SOAP+c2m (#bags: $\sim$ 52; #pathways: $\sim$ 8) (e) SPREAT+c2m (#bags: $\sim$ 51; #pathways: $\sim$ 8)

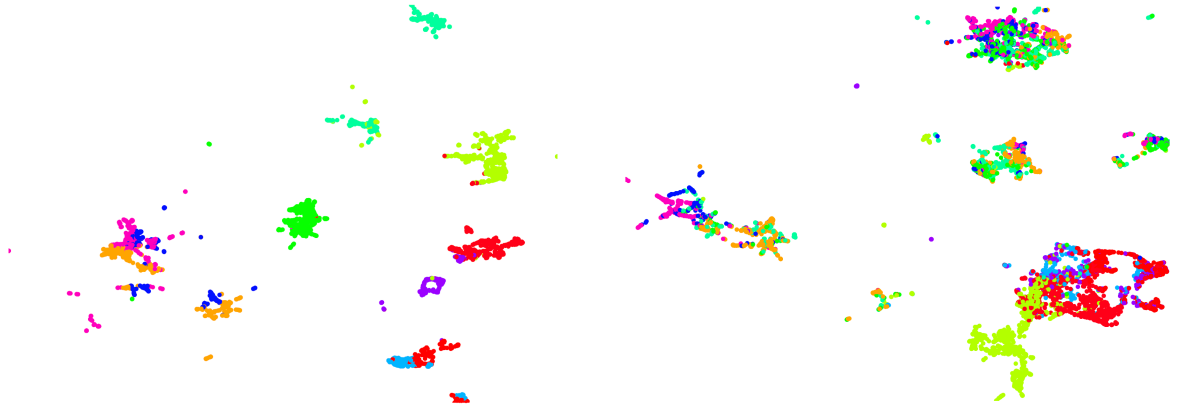
Figure 4: **Visualizing** 50 **randomly picked bags for each model, trained with** $b = 200$**.** The first term within the bracket, i.e., #bags, corresponds to the average number of correlated bags while the second term, i.e., #pathways, represents the average number of pathway size per bag. The circles represent bags, and their sizes reflect the correlation strength with other bags. Two clusters of bags can be seen for the last three models indicating the two clusters contain distinct pathways.

Experimental results. While the log predictive scores for SOAP and SPREAT in Fig. 3a appears to be horizontal across bag size, the CTM model projects a more realistic view where it's performances are seen to be gaining by including more bags. For the former models, this phenomena is not a consequence of design flaw, instead, it is expected due to the effects of supplementary pathways. That is, both models are encouraged to learn more pathways from $\mathbf{M}$ because the average pathway size for an example in $\mathbf{M}$ is $\sim$ 500 whereas in BioCyc T2 & 3 is $\sim$ 195 while only retaining 100 pathways for each bag. By excluding $\mathbf{M}$ (through enabling c2m option), we observe that the log predictive distribution of SOAP and SPREAT are similar with that of CTM, as shown in Fig. 3b, which supports our previous discussion. We found that $b = 200$ gives a good set of overlapping pathways while having on average $\sim$ 15 distinct pathways for each bag from 2526 pathways. By fixing $b = 200$, we search for an optimum k value. As illustrated in Fig. 3c, both SOAP and SPREAT deteriorate their performances (< -0.6) when $k > 100$. Taken together, we suggest the settings $b \in \mathbb{Z}_{[150,300]}$ and $k \in \mathbb{Z}_{[50,100]}$ to recover good bag and pathway distributions.

7.2 Bag Visualization

Experimental setup. Here, we visualized the discovered bags using models in Section 7.1 with the goal to assess the quality of bags. First, we examined the influence of augmented pathways on bag correlation patterns, i.e., Σ, in SOAP and SPREAT and contrast the outputs with CTM, SOAP+c2m, and SPREAT+c2m. Then, we performed an in-depth comparison between the sparse models (SOAP+c2m and SPREAT+c2m) with CTM to ensure that our modeling assumptions are aligned with the observed data, where a bag containing more focused and fewer pathways is preferred. For all experiments here, we applied the settings described in Section 6.2.

Experimental results. From Fig. 4, we notice two core findings. First, in contrast to CTM, SOAP+c2m, SPREAT, and SPREAT+c2m, bags in SOAP are densely connected ($\sim$ 74 bags) where the width of edges indicates the strength of correlations. Second, leveraging background pathways in SOAP and SPREAT resulted in gathering more pathways for each bag. These observations demonstrate the influence of $\mathbf{M}$ (obtained from mlLGPR) to pathway distribution over bags. As



(a) BioCyc (v20.5 T2 & 3)

(b) BioCyc (v20.5 T2 & 3) (M)

Figure 5: **2D UMAP projections of BioCyc (v20.5 T2 & 3) pathways and the corresponding background pathways.** Fig. 5a serves as a basis for color-coding where examples of one color in BioCyc are clustered together while the same examples are seen to be spread across the augmented BioCyc pathways (**M**) in Fig. 5b. Better viewed in color.

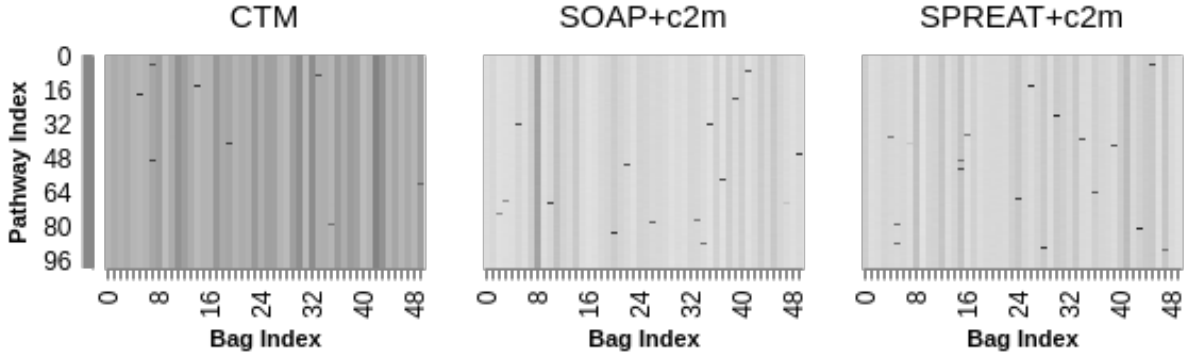


Figure 6: **Heatmap representing bag distribution of CTM, SOAP+c2m, SPREAT+c2m for randomly picked** 50 **bags with their associated** 100 **pathways.** The entries is color-coded on a gradient scale ranging from light-gray to dark-gray, where higher intensity entails higher probability.

an example, Fig. 5 shows samples from BioCyc T2 & 3 pathways and corresponding background pathways after projecting them onto 2D space using UMAP [25]. The colors encode samples from BioCyc T2 & 3 pathways that are clustered using the K-means algorithm [19] with 10 groups. While examples of the same color in BioCyc T2 & 3 pathways form a clear distinct group, the same examples are seen to be intermixed for **M**, possibly comprising of many false-positive pathways as depicted in Fig. 2 where many pathways (represented as column bars) are differentially distributed between BioCyc T2 & 3 pathways and **M**. For the collapsed models, they are observed to share similar behaviors as CTM (Fig. 3b). However, their bag distribution consists of fewer pathways than CTM (Fig. 4). For example, Fig. 6 shows 50 randomly selected bags with associated 100 pathways, where CTM is shown to encapsulate more pathways per bag (encoded by gradient darker colors) while SOAP+c2m and SPREAT+c2m exhibit a sparse distribution.

The results from these experiments link with our previous remarks, entailing that SOAP and SPREAT are equipped to reduce irrelevant pathways by applying dual sparseness. In particular, SPREAT is observed to generate fewer correlations than SOAP. With regard to incorporating supplementary pathways, SOAP and SPREAT are both sensitive to false-positive pathways, therefore, including accurate augmented pathways may recover better pathway distribution over bags.

8 Conclusion

In this paper, we have presented two novel statistical hierarchical mixture models, SOAP and SPREAT, to uncover correlated latent concepts or bags given a collection of pathway data. The

work is motivated by the problem of missing pathways, which is very common in a genomic dataset. Bag based approaches with augmented pathways (provided being unevenly distributed across examples as possible), along with applying dual sparsity in inference, are an alternative way to represent bags for examples and to model pathways over bags.

There are several directions for future study. Foremost, we intend to build a model that takes bags as an alternative solution to improve metabolic pathway prediction. We also investigate the sparseness induction in the covariance matrix for better interpretability while, at the same time, reducing the number of free parameters to estimate [13]. This and other possibilities we leave for future work.

References

[1] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of machine learning research*, 9(Sep):1981–2014, 2008.

[2] Abdur Rahman MA Basher and Steven J Hallam. Leveraging heterogeneous network embedding for metabolic pathway prediction. *bioRxiv*, feb 2020.

[3] Abdur Rahman MA Basher, Ryan J McLaughlin, and Steven J Hallam. Incorporating triple nmf with community detection to metabolic pathway inference. *bioRxiv*, 2020.

[4] Abdur Rahman MA Basher, Ryan J McLaughlin, and Steven J Hallam. Metabolic pathway inference using multi-label classification with rich pathway features. *bioRxiv*, February 2020.

[5] Jacob Bien and Robert J Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, 2011.

[6] David Blei and John Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.

[7] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *arXiv preprint arXiv:1601.00670*, 2016.

[8] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[9] Ron Caspi, Richard Billington, Luciana Ferrer, Hartmut Foerster, Carol A. Fulcher, Ingrid M. Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A. Mueller, Quang Ong, Suzanne Paley, Pallavi Subhraveti, Daniel S. Weaver, and Peter D. Karp. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Research*, 44(D1):D471–D480, 2016.

[10] Ron Caspi, Richard Billington, Hartmut Foerster, Carol A Fulcher, Ingrid Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A Mueller, Quang Ong, et al. Biocyc: Online resource for genome and metabolic pathway analysis. *The FASEB Journal*, 30(1 Supplement):lb192–lb192, 2016.

[11] Ron Caspi, Richard Billington, Ingrid M Keseler, Anamika Kothari, Markus Krummenacker, Peter E Midford, Wai Kit Ong, Suzanne Paley, Pallavi Subhraveti, and Peter D Karp. The metacyc database of metabolic pathways and enzymes-a 2019 update. *Nucleic acids research*, 2019.

[12] Paul Falkowski, RJ Scholes, EEA Boyle, Josep Canadell, D Canfield, J Elser, Nicolas Gruber, Kathy Hibbard, Peter Högberg, S Linder, et al. The global carbon cycle: a test of our knowledge of earth as a system. *science*, 290(5490):291–296, 2000.

[13] Jianqing Fan, Yuan Liao, and Han Liu. An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19(1):C1–C32, 2016.

[14] Niels W Hanson, Kishori M Konwar, Alyse K Hawley, Tomer Altman, Peter D Karp, and Steven J Hallam. Metabolic pathways for the whole community. *BMC genomics*, 15(1):1, 2014.

[15] Alyse K Hawley, Heather M Brewer, Angela D Norbeck, Ljiljana Paša-Tolić, and Steven J Hallam. Metaproteomics reveals differential modes of metabolic coupling among ubiquitous oxygen minimum zone microbes. *Proceedings of the National Academy of Sciences*, 111(31):11395–11400, 2014.

[16] Junxian He, Zhiting Hu, Taylor Berg-Kirkpatrick, Ying Huang, and Eric P Xing. Efficient correlated topic modeling with topic embedding. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 225–233. ACM, 2017.

[17] Matthew D Hoffman, David M Blei, Chong Wang, and John William Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[18] Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pp. 80–88. acm, 2010.

[19] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7):881–892, 2002.

[20] Peter D Karp, Mario Latendresse, Suzanne M Paley, Markus Krummenacker, Quang D Ong, Richard Billington, Anamika Kothari, Daniel Weaver, Thomas Lee, Pallavi Subhraveti, et al. Pathway tools version 19.0 update: software for pathway/genome informatics and systems biology. *Briefings in bioinformatics*, 17(5):877–890, 2016.

[21] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[22] Tianyi Lin, Wentao Tian, Qiaozhu Mei, and Hong Cheng. 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*, pp. 539–550. ACM, 2014.

[23] Lin Liu, Lin Tang, Wen Dong, Shaowen Yao, and Wei Zhou. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1):1608, 2016.

[24] Jon D Mcauliffe and David M Blei. Supervised topic models. In *Advances in neural information processing systems*, pp. 121–128, 2008.

[25] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

[26] David Mimno, Matt Hoffman, and David Blei. Sparse stochastic inference for latent dirichlet allocation. *arXiv preprint arXiv:1206.6425*, 2012.

[27] John Paisley, Chong Wang, David M Blei, and Michael I Jordan. Nested hierarchical dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):256–270, 2015.

[28] Alexander Sczyrba, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, Stephan Majda, Jessika Fiedler, Eik Dahms, et al. Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nature methods*, 14(11):1063, 2017.

[29] Mahdi Shafiei, Katherine A Dunn, Eva Boon, Shelley M MacDonald, David A Walsh, Hong Gu, and Joseph P Bielawski. Biomico: a supervised bayesian model for inference of microbial community structure. *Microbiome*, 3(1):1, 2015.

[30] Mahdi Shafiei, Katherine A Dunn, Hugh Chipman, Hong Gu, and Joseph P Bielawski. Biomenet: A bayesian model for inference of metabolic divergence among microbial communities. *PLoS Comput Biol*, 10(11):e1003918, 2014.

[31] Stéfan van der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30, 2011.

[32] Jia Wu, Xingquan Zhu, Chengqi Zhang, and S Yu Philip. Bag constrained structure pattern mining for multi-graph classification. *Ieee transactions on knowledge and data engineering*, 26(10):2382–2396, 2014.

[33] Jifang Yan, Guohui Chuai, Tao Qi, Fangyang Shao, Chi Zhou, Chenyu Zhu, Jing Yang, Yifei Yu, Cong Shi, Ning Kang, et al. Metatopics: an integration tool to analyze microbial community profile by topic model. *BMC genomics*, 18(1):962, 2017.

[34] Ruichang Zhang, Zhanzhan Cheng, Jihong Guan, and Shuigeng Zhou. Exploiting topic modeling to boost metagenomic reads binning. In *BMC bioinformatics*, volume 16, p. S2. BioMed Central, 2015.

[35] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *European conference on information retrieval*, pp. 338–349. Springer, 2011.