Discovery of Host-Microbiome Interactions Using Multi-Modal, Sparse, Time-Aware, Bayesian Network-Structured Neural Topic Models

Anonymous Authors[1]

Abstract

Although host-bacterial interactions have been extensively characterized for some pathogens, much less is known about how commensal bacteria in the microbiome interact with us at the molecular level. We introduce a new method, based on Neural Topic Models (NTMs), to jointly learn host and microbiome topics and a directed interaction network between them, from datasets with paired host and microbiome sequencing data. Our methodological contributions are: (a) modeling interactions between topics of the same or different data modalities, using a Directed Acyclic Graph (DAG) to capture dependencies, (b) directly inferring model structure, including DAG edges and temporal dependencies, using Bayesian Variable Selection (BVS)-style priors, and (c) an efficient end-to-end Variational Inference (VI) approach that leverages relaxations of discrete distributions. We apply our method to a new time-series dataset of paired fecal metagenomics and host blood transcriptomics measurements. We demonstrate that learning the DAG structure significantly improves topic quality, as assessed on known biological relationships. Further, we find new, biologically plausible host-microbe interactions, including microbiome effects on T-cell and dendritic cell activity, and on tuberculosis (TB) clearance.

1. Introduction and Prior Work

The human microbiome has been implicated in a variety of diseases (Sorbara & Pamer, 2022), often through epidemiological studies employing high-throughput sequencing. Despite these intriguing connections, relatively little is known about the underlying molecular cross-talk between the host and microbiome. Challenges to inferring host-microbiome interactions from high-throughput data include large inter-individual variability of the microbiome, limited perturbations or temporal information, and multi-modality, high-dimensionality, and compositionality of data. In addition, interpretability of model outputs is important.

Due to their interpretability and ability to handle counts data, topic models have been a popular choice for analyzing biological sequencing data. NTMs have recently been applied to single-cell human data (Lynch et al., 2022), and standard topic models have been applied to microbiome data (e.g., (Sankaran & Holmes, 2019)). Various methods to introduce sparsity in the number of topics have been used, including Dirichlet Processes (Teh et al., 2006) and related ideas. Our method differs in that we use BVS to introduce structural zeros, which has been shown to reduce bias as compared to shrinkage-based methods (Tadesse et al., 2005).

Different approaches have also been used to introduce dependencies in topic proportions, including the general Correlated Topic Model (Blei & Lafferty, 2005) and special predefined structures (e.g., (Li & McCallum, 2006)). Our method differs in that our model learns a DAG or Bayesian Network (BN) *de novo* to encode dependencies, which has the advantage of capturing conditional independence relationships and providing directed edges. To enable efficient and end-to-end inference, we leverage a recent VI-based approach that uses relaxed permutations to represent the DAG (Charpentier et al., 2022). BNs have previously been applied to microbiome data (Ruiz-Perez et al., 2021); however, they were not part of an end-to-end model including noise in the data, and did not integrate clustering/topics. There is also considerable prior work on time-dependence in topic models starting with (Blei & Lafferty, 2006). Because our primary application is to short time-series with fairly long sampling intervals, we use a relatively simple formulation; our main contribution in this regard is using the BVS approach to infer which topics are time-dependent.

2. Method Description

2.1. Model

Assume we have data from S subjects, each profiled at T time-points, and with measurements at each time-point of

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by *the 2023 ICML Workshop on Computational Biology*. Do not distribute.

M different modalities (i.e., host blood transcriptomics and fecal microbiome metagenomics). We model each multimodal biological specimen set as M bags of words with W_m words each, in which each word occurrence corresponds to a sequencing read. Let r_{mstl} denote the l^{th} read from modality m for subject s at time-point t, where there are L_{mst} sequencing reads. We model reads as being generated by a NTM with a DAG structure capturing dependencies among topics. Sparsity in both the number of topics and the number of edges in the DAG are modeled using BVS. Figure 1a shows the complete model using plate notation.

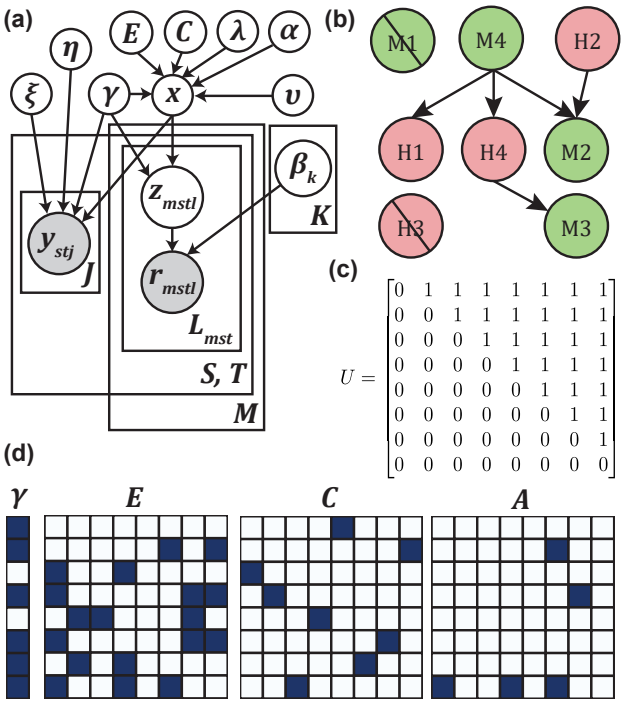

Figure 1. (a) Plate representation of our model. (b) Example topic BN with 4 topics for host (pink) and green (microbes); slashes indicate topic is not selected. (c) Mask matrix. (d) Selector γ, edge $\mathbf{E}$, permutation $\mathbf{C}$, and adjacency matrix $\mathbf{A}$ for the example BN.

To facilitate efficient variational inference, as described below, we decomposed the sparse DAG adjacency matrix as follows (Figures 1b-d illustrate an example). Let K_m denote the maximum number of topics for modality m and let $K = \sum_m K_m$. Let $\mathbf{E}$ denote a $K \times K$ edge matrix, $\mathbf{C}$ a $K \times K$ permutation matrix, and $\boldsymbol{\gamma}$ a K-dimensional vector of indicator variables sampled as:

$$\begin{aligned} E_{kk'} &\sim \text{Bernoulli}(\rho_e) \\ \mathbf{C} &\sim \text{Permutation}_{K\times K} \\ \gamma_k &\sim \text{Bernoulli}(\rho_{\gamma m}) \text{ for } k \in \text{modality } m \end{aligned}$$

Here, ρ_e and $\rho_{\gamma m}$ are hyperparameters that influence the sparsity of edges in the DAG and number of topics, respectively (we set $\rho_e = 0.05$ and $\rho_{\gamma m} = 1/K_m$). Note that we set $\gamma_k = 1$ for the first topic in each modality to ensure one topic per modality is always present . The adjacency matrix for the BN is then given by $\mathbf{A} = (\mathbf{C}^T(\mathbf{E} \odot U)\mathbf{C}) \odot \boldsymbol{\gamma}\boldsymbol{\gamma}^T$, where U is an upper-triangular mask matrix.

The sampling scheme for words (reads) is then:

$$\begin{aligned} \alpha_{tk} &\sim \text{Bernoulli}(\rho_\alpha) \\ x_{stk} &\sim \text{Normal}(\textstyle\sum_{k'} A_{k'k}\lambda_{k'k}x_{stk'} + \alpha_{tk}\nu_{tk}, 1) \\ \theta_{mstk} &= \frac{\gamma_k e^{x_{stk}}}{\sum_{k'\in\text{modality}(m)} \gamma_{k'} e^{x_{stk'}}} \\ z_{mstl} &\sim \text{Categorical}(\theta_{mst}) \\ r_{mstl} &\sim \text{Categorical}(\beta_{z_{mstl}}) \end{aligned}$$

The variables α_{tk} select whether there is time-dependence for each topic-time-point pair. We model edge and temporal weights, $\lambda_{k'k}$ and ν_{tk}, respectively, as parameters that are optimized during inference as in (Charpentier et al., 2022), rather than random variables. Word distributions β_{mk} are similarly modeled as W_m parameters per topic that are optimized during inference, as in (Miao et al., 2017).

We also include a supervised component of the model, that predicts J binary features, y_{stj}, from the document distributions, i.e., $y_{stj} \sim \text{Bernoulli}(1/(1 + e^{-\sum_k \xi_{kj}\eta_{kj}x_{stk}}))$, where ξ_{kj} is a Bernoulli-distributed selector random variable, and η_{kj} are parameters optimized during inference.

2.2. Inference

To achieve end-to-end inference, we used VI with relaxations of Bernoulli-distributed variables and of the permutation matrix. We implemented inference in Pytorch and used Adam with default parameters for optimization. The topic selection variable, z, was marginalized out for inference. For Bernoulli-distributed variables, we used a Gumbel Softmax approach with common parameters across the dataset for their variational approximations (Jang et al., 2016). The permutation matrix $\mathbf{C}$ was handled using the approach described in (Charpentier et al., 2022) for differentiable DAG sampling, which assumes no prior on $\mathbf{C}$ and approximates the ELBO using samples that are generated with i.i.d. Gumbel noise and then passed through the SoftSort function.

Approximations for x_{stk} were handled using a variational auto-encoder approach (Kingma & Welling, 2013). Let $\widehat{D}_{st} = (\widehat{D}_{1st}, \ldots, \widehat{D}_{mst})$ denote the vector concatenating relative abundance vectors for all modalities for subject s at time-point t. An inference network conditioned on $\widehat{D}_{st}$ was then used to generate the variational parameters, i.e., $\tilde{x}_{stk} \sim \text{Normal}(\text{MLP}_\mu(\widehat{D}_{st}), \text{MLP}_{\sigma^2}(\widehat{D}_{st}))$. For all experiments discussed here, we used a 3-layer fully connected MLP with 100 nodes in each layer and SoftPlus activations. Outputs were then fed through final linear layers to generate μ or

Table 1. Numbers of topics and percent significantly enriched (E) for Family level taxa for microbe (M) or GO categories for host (H) topics. S = topic sparsity, B = BN, T = time-dependence. S-B-T- is a standard NTM, in which host and microbe topics are independent. The maximum number of topics was set to 100 in all cases.

	#MT	#HT	%ME	%HE
S-B-T-	100 ± 0	100 ± 0	31 ± 2.8	12 ± 0.6
S-B+T-	100 ± 0	100 ± 0	32 ± 2.1	12 ± 1.2
S-B+T+	100 ± 0	100 ± 0	36 ± 1.9	12 ± 0.7
S+B-T-	50 ± 1.0	48 ± 5.0	36 ± 4.7	20 ± 2.6
S+B+T-	36 ± 2.2	29 ± 2.1	33 ± 2.3	24 ± 2.6
S+B+T+	37 ± 2.3	27 ± 1.6	38 ± 5.7	29 ± 4.0

log σ. We used a Stochastic Gradient Variational Bayes estimator (Kingma & Welling, 2013) for the KL term.

3. Results

3.1. Longitudinal Tuberculosis (TB) Cohort

We applied our method to a time-series dataset we recently collected that measures host transcriptomics and gut metagenomics from 98 paired blood and stool samples in a cohort of 24 participants undergoing antibiotic treatment for drug-resistant TB. The dataset captures naturally occurring and introduced perturbations to both the microbiome and host, and is thus a potentially rich source of information for inferring interactions. Specimens were collected pre-treatment, then at 2 weeks, 2 months, 6 months, and 2 years (post-treatment); not all time-points were collected for all participants. Specimen processing, sequencing, and bioinformatics were as described previously (Olendzki et al., 2022; Wipperman et al., 2021). Datasets were filtered to remove features with low abundance or coefficient of variation across samples with the following settings: $> 90\%$ with less than 200 counts for transcriptomics or 50 counts for metagenomics, or $CV < 0.5$. Transcriptomic data was further filtered to remove highly abundant ubiquitous transcripts, including ribosomal and RBC-specific genes. After filtering, 1,199 host genes and 103 taxa remained for downstream analyses.

3.2. Sparsity and Topic-Topic or Time Dependencies

We first removed different capabilities from our model to understand their influence on topic learning (Table 1). For all models, we ran inference 5 times with different random seeds, for 100,000 epochs each. Looking first at topic sparsity (S+), we found that our BVS approach always eliminated at least $\approx 50\%$ topics, and thus was effective at inducing sparsity. When BN learning (B+) was enabled, models inferred even fewer topics, suggesting their ability to use BNs to efficiently capture dependencies in data when learning topics. Including temporal information (T+) in models with topic sparsity also resulted in fewer topics, but the effect was smaller. Interestingly, all models with topic sparsity learned proportionately more microbe than host topics, despite their being $> 10\times$ as many genes as microbial taxa in the dataset. This is consistent with the microbiome composition being more "personalized," and reflecting less coordinated behavior than human gene expression.

We next investigated topic quality, by computing the number of topics that were significantly enriched (after BH correction) for either Family level taxa or GO categories using a z-score cutoff of ≥ 3 (Lynch et al., 2022), for host or microbe topics, respectively. We found that BN learning, when coupled with topic sparsity, significantly improved topic quality, particularly for host topics. Including time in models with topic sparsity also improved topic quality, but more modestly.

3.3. Examples of Learned Microbe $\rightarrow$ Host Topics

We used the full model (S+B+T+), and chose the run with largest ELBO, which yielded 41 microbe topics, 26 host topics, and 124 edges in the BN (with Bayes Factor [BF] > 10 [strong evidence]). Here, we present two examples of learned microbiome-to-host dependencies, an area of particular interest in the field (Tuganbaev & Honda, 2021).

Figure 2a shows all the children of topic M9. Both taxa in M9 are Gram-negative organisms that are in the Bacteroidales Order and have abilities to degrade complex polysaccharides (McKee et al., 2021). M9 has a positive edge to M21, which contains *Bacteroides vulgatus*, another Bacteroidales that has broad carbohydrate utilization capabilities, suggesting possible cross-feeding from M9 organisms. M9 has a negative edge into M13 that contains mostly Gram-positive organisms, many of which have pathogenic properties, suggesting that M9 organisms may be active in suppressing pathobioants. M9 has a positive edge to H62, a topic with several genes involved in immune regulation, including TRAV17 and TRBV27 (T-cell receptor genes), XRRA1 (associated with proliferation of white blood cells), and C4BPA (associated with promoting inflammatory cytokines and TL-4 pathway-related genes). This finding is consistent with reports that *Prevotella copri* is associated with increased host inflammation (Tett et al., 2021). M9 has a negative edge into H48, which is significantly enriched for genes associated with differentiation of subsets of dendritic cells (DC) (Balan et al., 2019). This previously unknown relationships suggests M9 organisms, in addition to promoting inflammation, may also skew DC development.

Figure 2b shows all the children of M38. Its taxa are two *Bacteroides* species and *Phascolarctobacterium faecium*, a Gram-negative organism in Order Acidaminococcales that is also in M37 (a topic that M38 positively influences). *P. faecium* has been shown to cross-feed on succinate pro-

duced by *Bacteroides* species (Ikeyama et al., 2020), providing a biologically plausible explanation for the learned relationships. M37 contains several other organisms that are plausible cross-feeders on products from M38 organisms, including *Bacteroides caccae* and *Methanobrevibacter smithii* (the dominant archaeon in human gut, which consumes end products of bacterial fermentation to produce methane). M38 also has a positive edge to H57, a host topic that is significantly enriched for T-cell receptor cells. *B. fragilis* and other *Bacteroides* species produce capsular polysaccharide-A, which has been shown to activate CD4+ T-cells (Eribo et al., 2022), providing a plausible mechanism for the learned M38 $\rightarrow$ H57 edge.

3.4. Time-Dependent and TB-Clearance Related Topics

We found that 76% of topics showed time-dependence (BF $>$ 10). Many of the time-dependent microbe topics can be clearly tied to antibiotic treatment (Nel Van Zyl et al., 2022). For example, a number of microbe topics decrease proportionally during the entire antibiotic treatment period and then recover post-antibiotics at 2 years, such as M19, which contains multiple *Roseburia* species and other fastidious anaerobes that are known to be highly affected by antibiotics. In other cases, microbe topics capture blooms of organisms seen in dysbiotic guts, such as M7 that contains 3 *Klebsiella* species and M29 that is dominated by *E. coli*.

We also found time-dependent host topics. Examples of host topics that showed relative decreases during treatment include M53 (highly significantly enriched, with 52/83 genes in the immunoglobulin complex category) and M44 and M56 (containing myeloid marker genes). In contrast, topics that showed increases during treatment included M52 (containing T-cells genes, including PTGDR2 that is preferentially expressed in CD4+ cells). These results suggest time-dependent skewing of immune populations, in both lymphocyte and myeloid compartments during treatment.

When we included TB clearance measured by sputum load as the supervised variable in our model, a negative M35 $\rightarrow$ TB clearance edge was found (BF $>$ 100). M35 contains *Blautia wexlerae*, *Ruminococcus torques*, *Clostridium sp. AM22-11AC*, *Bifidobacterium longum* and *Anaerostipes hadrus*. Interestingly, many of these organisms have bile acid transformation capabilities and have been negatively associated with obesity (Aron-Wisnewsky et al., 2021). One intriguing possibility is that alterations in host fat metabolism induced by these organisms influence the course of TB infection; however, targeted studies would certainly be necessary to investigate this hypothesis.

3.5. Conclusion and Future Work

We introduced a model and accompanying VI inference algorithm that simultaneously encodes sparsity using BVS priors, captures topic-to-topic dependency using Bayesian Networks, and additionally allows for time-dependence and supervised variables. We applied our method to a time-series dataset with paired gut metagenomics and host gene expression measurements, and demonstrated that our method finds sparse and biologically interpretable topics, and a topic-topic network with putative new connections between the microbiome, the host immune system, and TB infection. One limitation of our results is that we analyzed only one dataset; this was in part because joint host-microbe data is currently limited. In future work, we plan to apply our model to more datasets, including larger ones we are generating. Another limitation is that we only evaluated two measures of topic quality (GO categories and bacterial taxonomy). We plan to investigate additional measures, such as gene and microbiome signatures. Although in this work we applied our model to two data types, our formulation is general, and in future work we plan to incorporate other modalities such as metabolomics. Also, we can extend our model to non-Multinomial distributions (e.g., Log-normal or Negative Binomial) to capture noise characteristics of additional data modalities. Our BN, although currently Gaussian, could be extended to capture nonlinear relationships; inference would be straight-forward using our VI framework with SVGB estimators.

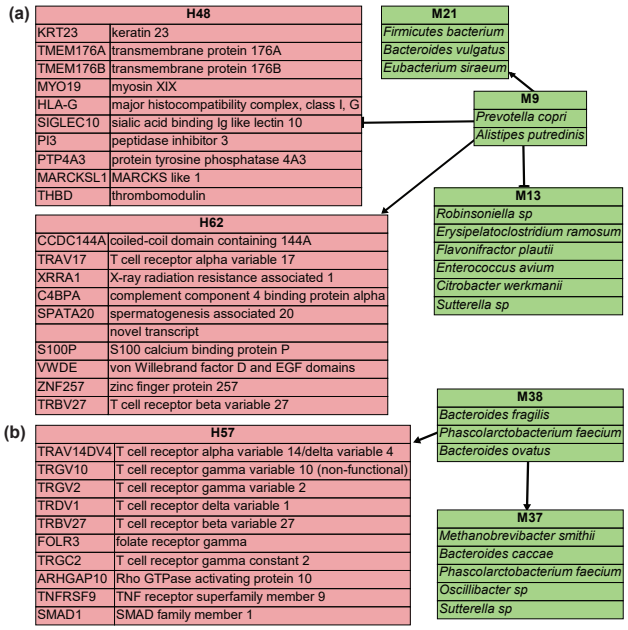

Figure 2. Example topic sub-networks, showing all children of microbe topics (a) M9 and (b) M38. Arrow thickness denotes the Bayes Factor (BF) for edge presence ($>$ 10 for all edges shown). Microbe topics are green and host topics are pink. Top 10 words (taxa or genes) with z-scores $\geq$ 2 are shown.

References

Aron-Wisnewsky, J., Warmbrunn, M. V., Nieuwdorp, M., and Clément, K. Metabolism and Metabolic Disorders and the Microbiome: The Intestinal Microbiota Associated With Obesity, Lipid Metabolism, and Metabolic Health-Pathophysiology and Therapeutic Strategies. *Gastroenterology*, 160(2):573–599, 2021.

Balan, S., Saxena, M., and Bhardwaj, N. Dendritic cell subsets and locations. *Int Rev Cell Mol Biol*, 348:1–68, 2019.

Blei, D. M. and Lafferty, J. D. Correlated Topic Models. In *NIPS*, 2005.

Blei, D. M. and Lafferty, J. D. Dynamic topic models. *Proceedings of the 23rd international conference on Machine learning*, 2006.

Charpentier, B., Kibler, S., and Günnemann, S. Differentiable DAG Sampling. *ArXiv*, abs/2203.08509, 2022.

Eribo, O. A., du Plessis, N., and Chegou, N. N. The Intestinal Commensal, Bacteroides fragilis, Modulates Host Responses to Viral Infection and Therapy: Lessons for Exploration during Mycobacterium tuberculosis Infection. *Infect Immun*, 90(1):e0032121, 2022.

Ikeyama, N., Murakami, T., Toyoda, A., and et al. Microbial interaction between the succinate-utilizing bacterium Phascolarctobacterium faecium and the gut commensal Bacteroides thetaiotaomicron. *Microbiologyopen*, 9(10): e1111, 2020.

Jang, E., Gu, S. S., and Poole, B. Categorical Reparameterization with Gumbel-Softmax. *ArXiv*, abs/1611.01144, 2016.

Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. *CoRR*, abs/1312.6114, 2013.

Li, W. and McCallum, A. Pachinko allocation: DAG-structured mixture models of topic correlations. *Proceedings of the 23rd international conference on Machine learning*, 2006.

Lynch, A. W., Theodoris, C. V., Long, H. W., and et al. MIRA: joint regulatory modeling of multimodal expression and chromatin accessibility in single cells. *Nat Methods*, 19(9):1097–1108, 2022.

McKee, L. S., La Rosa, S. L., Westereng, B., and et al. Polysaccharide degradation by the Bacteroidetes: mechanisms and nomenclature. *Environ Microbiol Rep*, 13(5): 559–581, 2021.

Miao, Y., Grefenstette, E., and Blunsom, P. Discovering Discrete Latent Topics with Neural Variational Inference. In *International Conference on Machine Learning*, 2017.

Nel Van Zyl, K., Matukane, S. R., Hamman, B. L., and et al. Effect of antibiotics on the human microbiome: a systematic review. *Int J Antimicrob Agents*, 59(2):106502, 2022.

Olendzki, B., Bucci, V., Cawley, C., and et al. Dietary manipulation of the gut microbiome in inflammatory bowel disease patients: Pilot study. *Gut Microbes*, 14(1): 2046244, 2022.

Ruiz-Perez, D., Lugo-Martinez, J., Bourguignon, N., and et al. Dynamic Bayesian Networks for Integrating Multiomics Time Series Microbiome Data. *mSystems*, 6(2), 2021.

Sankaran, K. and Holmes, S. P. Latent variable modeling for the microbiome. *Biostatistics*, 20(4):599–614, 2019.

Sorbara, M. T. and Pamer, E. G. Microbiome-based therapeutics. *Nat Rev Microbiol*, 20(6):365–380, 2022.

Tadesse, M. G., Sha, N., and Vannucci, M. Bayesian Variable Selection in Clustering High-Dimensional Data. *Journal of the American Statistical Association*, 100:602 – 617, 2005.

Teh, Y. W., Jordan, M. I., Beal, M. J., and et al. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101:1566 – 1581, 2006.

Tett, A., Pasolli, E., Masetti, G., and et al. Prevotella diversity, niches and interactions with the human host. *Nat Rev Microbiol*, 19(9):585–599, 2021.

Tuganbaev, T. and Honda, K. Non-zero-sum microbiome immune system interactions. *Eur J Immunol*, 51(9):2120–2136, 2021.

Wipperman, M. F., Bhattarai, S. K., Vorkas, C. K., and et al. Gastrointestinal microbiota composition predicts peripheral inflammatory state during treatment of human tuberculosis. *Nat Commun*, 12(1):1141, 2021.