HM-GIM: A Probabilistic Neural Model for Discovering Heterogeneous Microbiome or Human Gene Groupings and Their Interactions

Jiening Zhu [1] **Christine Tataru** [1] **Isin Y Comba** [2] **Shakti K. Bhattarai** [3] **Vanni Bucci** [3] **Georg K. Gerber** [1]

Abstract

Coordinated behavior among groups of biological units, such as co-expressed genes, is common in biological systems including the human microbiome, which is important in a variety of physiologic and pathological processes. While many methods infer such groupings, principled identification of interactions between groups remains under-explored. We present Host-Microbe Groups Interaction Model (HM-GIM), a generative Bayesian deep learning approach for uncovering group-level interactions in host-microbiome data. HM-GIM jointly infers groups of host genes or microbes, along with latent factors that induce a sparse, undirected dependency structure among them. Key innovations include: (1) modeling undirected and multi-way interactions, (2) avoiding distortions from simplex-based models, and (3) enabling flexible count-based error models. We demonstrate on paired human gene expression and microbiome data from a longitudinal cohort of patients with tuberculosis (TB) that HM-GIM outperforms existing methods in finding biologically meaningful groupings, and provide a case study identifying host-microbe interactions involved in innate and adaptive host immunity.

1. Introduction

The human gut microbiome influences diverse host physiology, including metabolism and immune function (Wu & Wu, 2012; Zhao et al., 2023) as well as pathological processes in a variety of human diseases. Advances in high-throughput sequencing allow joint profiling of microbial composition via metagenomics and host responses via transcriptomics (Zhou et al., 2015; Wang et al., 2009), offering opportunities to study host-microbiome interactions at scale. However, extracting meaningful insights from these data remains challenging in part due to their high-dimensionality, i.e., tens of thousands of genes and hundreds of taxa. A popular approach to address this challenge are topic models, which have been applied to host or microbiome data individually to find overlapping groups (topics) of biological units (genes or taxa) (Lafferty & Blei, 2005). Standard topic models assume independence of topics and thus cannot capture interactions between them, and additionally assume data is multinomially distributed, which does not accurately model over-dispersion observed with sequencing-based data. Alternatives such as nonnegative matrix factorization (NMF) or other latent factor models are also widely used to find groupings in high-dimensional data, and allow for more flexible error models, but do not directly model dependencies between groups. A recent neural topic model approach with Bayesian Networks (Gerber et al., 2023) allows for dependencies between topics, but presents challenges with interpreting interactions due to restriction to the simplex.

To address these limitations, we introduce HM-GIM, a generative Bayesian deep learning model that jointly infers groups of host genes or microbes and undirected interactions among the groups from paired host and microbiome counts-based sequencing data. We assume multi-way undirected interactions because in practice, inferring edge directionality may be unreliable with limited data, and multi-way interactions may more accurately describe certain biological relationships. HM-GIM assumes that the data arises from by groups of either genes or microbes, with specimen-specific usages for each group. Latent factors are used to model relationships among usages of the groups, e.g., a common factor capturing the tendency of several groups of microbes or human genes to positively and/or negatively covary in a specimen. These latent factors induce sparse, undirected edges between the nodes (groups), providing a compact and interpretable representation of the interaction structure. The observed counts-based data is then generated with a flexible noise model that allows for over-dispersion.

The remainder of this manuscript is organized as follows. We first introduce the generative process and inference al-

[1] Brigham and Women's Hospital, Boston, MA, USA [2] Mayo Clinic, Rochester, MN, USA [3] UMass Chan Medical School, Worcester, MA, USA.. Correspondence to: Georg K. Gerber <ggerber@bwh.harvard.edu>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

gorithm for HM-GIM. Next, we apply HM-GIM to a rich longitudinal dataset of 98 paired host transcriptomic and gut metagenomic profiles from 24 participants undergoing antibiotic treatment for multi-drug-resistant tuberculosis (MDR-TB). Samples were collected at up to five time points: pre-treatment, and at 2 weeks, 2 months, 6 months, and 2 years post-treatment initiation. We demonstrate HM-GIM's superior benchmarking performance on this dataset in finding biologically relevant groupings. Finally, we provide a case-study, showing that our method uncovers biologically interpretable host-microbiome interactions on this dataset.

2. Model

2.1. Generative process

Let y_{stmw} denote the observed sequencing counts for analysis unit w (gene or microbe) in modality m ($m = 0$ for host, $m = 1$ for microbes) for participant s at time-point t ($t \in \{T_1, T_2, ..., T_{N_t}\}$). We assume that the observed count data is generated by up to K_m groups —either gene or microbe groups—selected probabilistically to induce sparsity, and shaped by the N_t time covariates and up to N_f latent factors with sparse loadings that capture interactions among groups. Let $\mathcal{B}_{mkw}$ denote the frequency of occurrence of gene or microbe w in modality m in group k; these two matrices ($m = 0$ or 1) of parameters are non-probabilistic and learned during inference.

The overall generative process for HM-GIM (Figure 1) is then specified as:

1. Sample binary indicator variables that select whether group k in modality m is active: $\gamma_{mk} \sim \text{Bernoulli}\left(\rho_{\gamma_m}\right)$.

2. Sample binary indicator variables that select whether factor i influences group k in modality m: $\eta_{mki} \mid \gamma_{mk} \sim (1 - \gamma_{mk}) \cdot \delta_0 + \gamma_{mk} \cdot \text{Bernoulli}\left(\rho_\eta\right)$. Note that only active groups are influenced by factors.

3. Sample binary indicator variables that select whether group k in modality m has a time-dependent effect at time-point t: $\alpha_{tmk} \mid \gamma_{mk} \sim (1 - \gamma_{mk}) \cdot \delta_0 + \gamma_{mk} \cdot \text{Bernoulli}\left(\rho_\alpha\right)$.

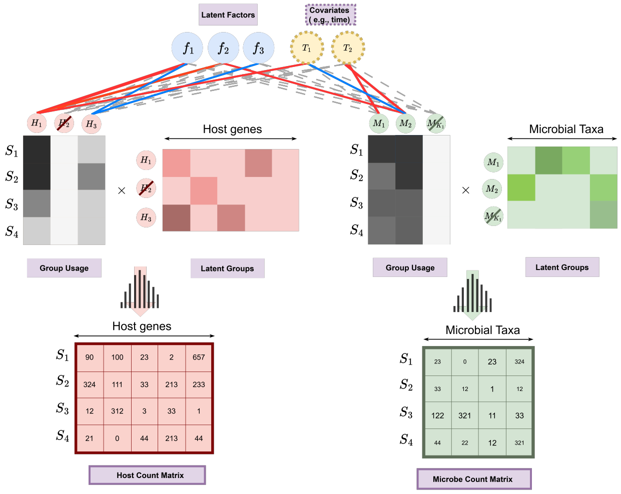
4. Sample latent factor i for participant s at time-point t: $f_{sti} \sim \text{Normal}(0, 1)$.

5. Sample the log of the latent group usage for participant s at time-point t of group k in modality m: $x_{stmk} \sim \text{Normal}(\lambda_{stmk}, 1)$. Here,

$$\lambda_{stmk} = b_{mk} + \sum_i \eta_{mki} \omega_{mki} f_{sti} + \alpha_{tmk} \phi_{tmk}$$

is the the mean of the log of latent usage for participant s at time-point t of group k in modality m. The other terms, b_{mk}, $\omega_{mki} \in \mathbb{R}$ and $\phi_{tmk} \in \mathbb{R}$ are non-probabilistic weights that are learned during inference, specifying offsets, factor loadings and covariate effects respectively.

6. Sample observed sequencing counts for analysis unit w (gene or microbe) in modality m for participant s at time-point t: $y_{stmw} \sim \text{NegativeBinomial}\left(\mu_{stm}, \epsilon_{mw}\right)$, where ϵ_{mw} is the dispersion parameter while $\mu_{stm} = \psi_{stm} (\sum_k \theta_{stmk} \mathcal{B}_{mkw})^2$ is the mean parameter. Here, $\theta_{stmk} = \gamma_{mk} e^{x_{stmk}}$ is the usage of group k. The scaling factor $\psi_{stm} = \frac{C_{stm}}{\text{median}_{s,t} C_{stm}}$ accounts for sequencing depth, where $C_{stm} = \sum_w y_{stmw}$ is the total observed count per specimen.



S_1	90	100	23	2	657
S_2	324	111	33	213	233
S_3	12	312	3	33	1
S_4	21	0	44	213	44

S_1	23	0	23	324
S_2	33	12	1	12
S_3	122	321	11	33
S_4	44	22	12	321

Figure 1. Overview of HM-GIM: our Bayesian generative model learns latent factors (blue) connecting latent groups (pink = host genes, green = microbial taxa) through factor loadings, capturing dependencies within and across modalities. Probabilistic binary indicators induce sparsity in the number of active groups for each modality (shown as slashes) and factor loadings on each group-factor pair (gray dashed lines denote non-selected loadings; solid blue and red lines denote positive or negative loadings respectively). Group usage for each subject $S_\cdot$ is generated from the the latent factors and covariates (dotted frames). Observed count-based data (thick frames) is then generated from Negative Binomial distributions parameterized by the product of group usage and frequencies of host genes or microbial taxa in groups.

2.2. Hyperparameters for Prior Distributions and Noise Model

We selected hyperparameters governing sparsity, variance, and dispersion based on a combination of domain knowledge and empirical performance. For group activation, we set the sparsity prior $\rho_{\gamma_m} = 1/K_m$, modeling the prior expectation of only one activated group. Similarly, factor loading sparsity was set to be $\rho_\eta = 1/(\sum_m K_m * N_f)$. The prior for temporal effects was set to $\rho_\alpha = 0.05$, modeling the prior expectation of a 5% probability of any given group having an effect from a time-point. For the Negative Binomial Noise model, we assumed a fixed dispersion $\epsilon_{mw} = 0.03$, reflecting a relative standard deviation of $\sim$20% at moderate count levels ($\sim$200).

2.3. Inference

The posterior distribution was approximated using stochastic amortized variational inference. A three-layer neural network with SoftPlus activations that takes concatenated log normalized microbial and host count data as inputs was used as the encoder. Discrete variables (e.g., group selectors and factor loading selectors) were approximated using Concrete (Gumbel-Softmax) relaxations. We exploited conjugacy of latent factors and marginalized them out to improve inference efficiency. The inference algorithm was implemented in PyTorch 2.6 using the Adam optimizer with default parameters. More details about the inference algorithm is included in Appendix A.

3. Results

3.1. Data Preprocessing

To ensure data quality, we filtered out low-abundance and invariant features based on relative abundance and coefficient of variation thresholds. Features were removed if they occurred in over 90% of samples with low abundance or had coefficient of variation (CV) < 0.5. Abundance thresholds were selected using the elbow points of retention curves. We also excluded highly abundant, non-informative transcripts (e.g., ribosomal and RBC genes). After filtering, 1729 host genes and 109 microbial taxa were retained for analysis.

3.2. Learning Biologically Relevant Groupings

We compared HM-GIM's ability to learn biologically relevant groupings against five popular methods in the microbiome field: tensor decomposition (TD) (Hore et al., 2016), topic models or latent Dirichlet allocation (LDA) (Blei et al., 2003), neural topic models with Bayesian network interactions (BN) (Gerber et al., 2023), multi-omics factor analysis (MOFA) (Argelaguet et al., 2018) , and non-negative matrix factorization (NMF) (Lee & Seung, 1999).

The number of topics or groups is a critical parameter in this analysis. We followed standard best practices for selecting this hyperparameter when not determined intrinsically by the method. Specifically, we used default or built-in selection strategies for TD, BN, and MOFA. For LDA, we applied 5-fold cross-validation to choose the topic number that minimized perplexity (Gan & Qi, 2021). For NMF, we selected the optimal rank based on the stability point before the largest drop in the cophenetic correlation coefficient (Brunet et al., 2004).

The biological relevance of host or microbial groups was assessed via functional enrichment analyses (hypergeometric test) using Gene Ontology (GO) annotations (release version 2024.1) (Consortium, 2021) for human genes, and KEGG pathways (Kanehisa et al., 2021) for microbial genes. GO annotations were accessed through the Human MSigDB Collections. Microbial genomes were annotated using Prokka (Seemann, 2014) and the presence of KEGG pathways was determined using MinPath (Ye & Doak, 2009). To focus on biologically meaningful categories, GO terms or KEGG pathways with fewer than 5 or more than 50 genes or taxa were filtered out. Multiple hypothesis testing corrections were performed using the Benjamini-Hochberg procedure to control the false discovery rate (FDR), and an adjusted p-value cutoff of 0.05 was considered significant for both GO and KEGG enrichment results.

To quantify performance variability and assess the robustness of each method, we ran all algorithms using 10 independent random seeds. Summary statistics and confidence intervals reported in the results are based on the distribution across these 10 runs (Figure 2). HM-GIM consistently outperformed the other methods on both tasks, including the more challenging task for microbes (due to relatively poor annotation of microbial genes), highlighting the superior ability of our method to uncover biologically relevant groupings compared to the state-of-the-art.

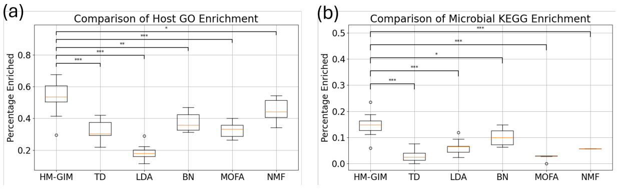
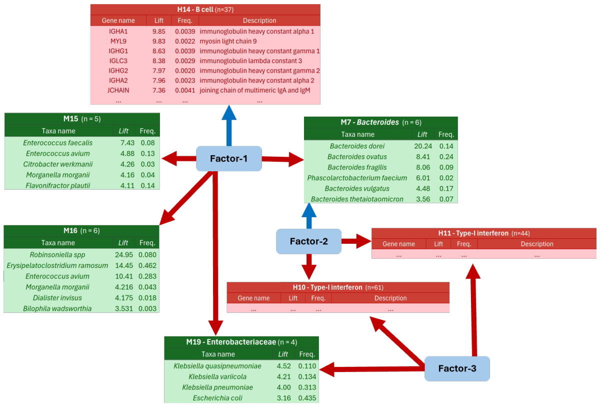


Figure 2. Ability of methods to learn biologically relevant groupings. The quality of groupings was assessed via functional enrichment analyses using (a) GO terms for human genes, and (b) KEGG pathways for microbes. Ten random seeds were run for each method. Statistical significance of pairwise differences was assessed using Mann–Whitney U tests. Stars denote statistical significance levels after Benjamini-Hochberg procedure: $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***). TD = tensor decomposition, LDA = topic models/latent Dirichlet allocation, BN = neural topic models with Bayesian network interactions, MOFA = multi-omics factor analysis, NMF = non-negative matrix factorization.

3.3. Case Study

HM-GIM found 19 microbial groups (21% significantly enriched for $\geq$ 1 KEGG pathway) and 28 host groups (39% significantly enriched for $\geq$ 1 GO category). Seven latent factors were identified, providing 68 factor-group interactions. Groups or interactions were reported here if the Bayes Factor was $>$ 100 (indicating decisive evidence (Kass & Raftery, 1995)) for the relevant binary selector. As a case study to focus on interesting host-microbe interactions, we analyzed the top 10% of enriched host pathways (ranked by enrichment p-value), which corresponded to 3 host topics linked to 4 microbial groups (Figure 3). Our case study identified interesting and plausible host-microbiome interactions involving both adaptive and innate immunity.



H14 - B cell (n=37)			
Gene name	Lift	Freq.	Description
IGHA1	9.85	0.0039	immunoglobulin heavy constant alpha 1
MYL9	9.83	0.0022	myosin light chain 9
IGHG1	8.63	0.0039	immunoglobulin heavy constant gamma 1
IGLC3	8.38	0.0029	immunoglobulin lambda constant 3
IGHG2	7.97	0.0020	immunoglobulin heavy constant gamma 2
IGHA2	7.96	0.0023	immunoglobulin heavy constant alpha 2
JCHAIN	7.36	0.0041	joining chain of multimeric IgA and IgM
...	...	...	...

M15 (n = 5)		
Taxa name	*Lift*	Freq.
Enterococcus faecalis	7.43	0.08
Enterococcus avium	4.88	0.13
Citrobacter werkmanii	4.26	0.03
Morganella morganii	4.16	0.04
Flavonifractor plautii	4.11	0.14

M7 - *Bacteroides* (n = 6)		
Taxa name	*Lift*	Freq.
Bacteroides dorei	20.24	0.14
Bacteroides ovatus	8.41	0.24
Bacteroides fragilis	8.06	0.09
Phascolarctobacterium faecium	6.01	0.02
Bacteroides vulgatus	4.48	0.17
Bacteroides thetaiotaomicron	3.56	0.07

M16 (n = 6)		
Taxa name	*Lift*	Freq.
Robinsoniella spp	24.95	0.080
Erysipelatoclostridium ramosum	14.45	0.462
Enterococcus avium	10.41	0.283
Morganella morganii	4.216	0.043
Dialister invisus	4.175	0.018
Bilophila wadsworthia	3.531	0.003

H11 - Type-I interferon (n=44)			
Gene name	Lift	Freq.	Description
...	...	...	...

H10 - Type-I interferon (n=61)			
Gene name	Lift	Freq.	Description
...	...	...	...

M19 - Enterobacteriaceae (n = 4)		
Taxa name	*Lift*	Freq.
Klebsiella quasipneumoniae	4.52	0.110
Klebsiella variicola	4.21	0.134
Klebsiella pneumoniae	4.00	0.313
Escherichia coli	3.16	0.435

Figure 3. Case study example of host gene (pink) and microbial taxa (green) groups linked through factors (light blue). Dark blue and red lines denote positive or negative factor loadings respectively. Tables list microbes with lift>3. Some example host genes for H14 are shown; full host gene lists are omitted due to space constraints. Freq. = Frequency in group.

Starting with adaptive immunity, H14, a host group, was highly enriched for GO categories related to B-cell mediated responses with key genes including immunoglobulin (Ig) light and heavy chains. Interestingly, H14 also included MHC class II genes regions, which are involved in T-cell mediated antigen presentation that facilitate class-switching (Ng et al., 2022), and the JCHAIN gene that is essential for sIgA polymeric antibody secretion at mucosal surfaces, which plays a central role in maintaining appropriate interactions with the microbiome in homeostasis (Benckert et al., 2011; Okai et al., 2016). We note that although our study used blood transcriptomics, systemic IgA profiles are known to closely mirror mucosal sIgA likely due to B-cell and T-cell trafficking (Iversen et al., 2017). H14 was linked to four microbe topics, M7, M15, M16, M19 via Factor-1, with the signs of factor loadings indicating negative covariation between the host group and the microbe groups. Three of the microbe groups (M15, M16, and M19) were dominated by functionally/phylogenetically distinct microbes that included mostly pathobionts. For example, M19 was predominantly composed of *Enterobacteriaceae* members, which are known to overgrow in settings of inflammation or other insults to gut homeostasis. M7, on the other hand, was composed predominantly of *Bacteroides* species, which generally function as commensals, although some species such as *B. fragilis* can be opportunistic pathogens. Interestingly, *Phascolarctobacterium faecium* (Anthamatten et al., 2024), has been shown to cross-feed on succinate produced by *Bacteroides*, providing a plausible explanation for its membership in this group. Taken together, the interaction structure HM-GIM uncovered is highly plausible biologically: increased levels of secreted Ig lead to lower levels of distinct groups of microbes. Some identified interactions are already known, such as strong IgA responses to *Enterobacteriaceae* (Conrey et al., 2023) and *Enterococcus* overgrowth and intestinal crypt invasion (Berbers et al., 2025) in IgA deficiency (Berbers et al., 2025). However, the interactions with microbes in M16 and M7 and Ig are not well characterized and present discovery opportunities.

Regarding innate immunity, HM-GIM identified two host groups (H10 and H11), highly enriched for type I interferon-stimulated genes (e.g., IFI44L, OAS1, MX1). These host groups were linked to microbial groups M7 and M19 via Factor-2 and Factor-3, with the factor loadings indicating positive covariation between M19 and the host groups and negative covariation between M7 and the host groups. Type I interferons are known to be stimulated by lipopolysaccharide (LPS) via TLR-4 signaling (Stefan et al., 2020). LPS is produced by gram-negative bacteria, including the *Enterobacteriaceae*, which dominate M19, providing a biologically plausible relationship for this interaction. The interaction in the opposite direction between the *Bacteroides* dominated group, M7, and interferon genes has not previously been described and suggests an interesting hypothesis that known T-cell mediated anti-inflammatory properties of the *Bacteroides* may also mitigate interferon responses.

4. Discussion

Our model, HM-GIM, finds biologically relevant groupings of host genes and microbes from high-throughput host-microbiome data and outperforms competing methods in terms of interpretability. Moreover, in our case study, we demonstrated that HM-GIM can uncover complex but interpretable and biologically interesting relationships among groups. To further enhance the utility and robustness of HM-GIM, we will pursue several key directions in future work. First, we aim to improve the quality of inferred groups by increasing both identifiability and interpretability. Possible strategies include stronger sparsity priors or those discouraging topic overlap; constraints inspired by anchor-word methods from identifiable topic modeling frameworks; and leveraging embeddings from large language models trained on biological data or literature. Second, we plan to refine our noise model to include more complicated dispersion relationships and modality-specific noise. Finally, we plan to evaluate our model on additional datasets and perform biological validation of key findings.

Impact Statement

Our factor-based group interaction modeling framework uncovers biologically relevant groupings and host-microbe interaction patterns, revealing potential interactions between the host and microbiome at the molecular level. More broadly, our approach offers a robust tool for analyzing multi-omic data in a scalable and interpretable manner.

References

Anthamatten, L., von Bieberstein, P. R., Menzi, C., Zünd, J. N., Lacroix, C., de Wouters, T., and Leventhal, G. E. Stratification of human gut microbiomes by succinotype is associated with inflammatory bowel disease status. *Microbiome*, 12(1), September 2024. ISSN 2049-2618. doi: 10.1186/s40168-024-01897-8. URL http://dx.doi.org/10.1186/s40168-024-01897-8.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6), June 2018. ISSN 1744-4292. doi: 10.15252/msb.20178124. URL http://dx.doi.org/10.15252/msb.20178124.

Benckert, J., Schmolka, N., Kreschel, C., Zoller, M. J., Sturm, A., Wiedenmann, B., and Wardemann, H. The majority of intestinal iga+ and igg+ plasmablasts in the human gut are antigen-specific. *Journal of Clinical Investigation*, 121(5):1946–1955, May 2011. ISSN 0021-9738. doi: 10.1172/jci44447. URL http://dx.doi.org/10.1172/JCI44447.

Berbers, R.-M., Paganelli, F. L., van Montfrans, J. M., Ellerbroek, P. M., Viveen, M. C., Rogers, M. R. C., Salomons, M., Schuurmans, J., van Stigt Thans, M., Vanmaris, R. M. M., Brosens, L. A. A., van der Wal, M. M., Dalm, V. A. S. H., van Hagen, P. M., van de Ven, A. A. J. M., Uh, H.-W., van Wijk, F., Willems, R. J. L., and Leavis, H. L. Gut microbial dysbiosis, iga, and enterococcus in common variable immunodeficiency with immune dysregulation. *Microbiome*, 13(1), January 2025. ISSN 2049-2618. doi: 10.1186/s40168-024-01982-y. URL http://dx.doi.org/10.1186/s40168-024-01982-y.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3: 993–1022, 2003. URL https://www.jmlr.org/papers/volume3/blei03a.

Brunet, J.-P., Tamayo, P., Golub, T. R., and Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, March 2004. ISSN 1091-6490. doi: 10.1073/pnas.0308531101. URL http://dx.doi.org/10.1073/pnas.0308531101.

Conrey, P. E., Denu, L., O'Boyle, K. C., Rozich, I., Green, J., Maslanka, J., Lubin, J.-B., Duranova, T., Haltzman, B. L., Gianchetti, L., Oldridge, D. A., De Luna, N., Vella, L. A., Allman, D., Spergel, J. M., Tanes, C., Bittinger, K., Henrickson, S. E., and Silverman, M. A. Iga deficiency destabilizes homeostasis toward intestinal microbes and increases systemic immune dysregulation. *Science Immunology*, 8(83), May 2023. ISSN 2470-9468. doi: 10.1126/sciimmunol.ade2335. URL http://dx.doi.org/10.1126/sciimmunol.ade2335.

Consortium, G. O. The gene ontology resource: enriching a gold mine. *Nucleic Acids Research*, 49(D1):D325–D334, 2021. doi: 10.1093/nar/gkaa1113. URL https://academic.oup.com/nar/article/49/D1/D325/6027811.

Gan, J. and Qi, Y. Selection of the optimal number of topics for lda topic model—taking patent policy analysis as an example. *Entropy*, 23(10):1301, October 2021. ISSN 1099-4300. doi: 10.3390/e23101301. URL http://dx.doi.org/10.3390/e23101301.

Gerber, G. K., Bhattarai, S. K., Du, M., Glickman, M. S., and Bucci, V. Discovery of host-microbiome interactions using multi-modal, sparse, time-aware, bayesian network-structured neural topic models. In *Proceedings of the Workshop on Computational Biology at ICML 2023*, 2023. URL https://icml-compbio.github.io/2023/papers/WCBICML2023_paper57.pdf.

Hore, V., Viñuela, A., Buil, A., Knight, J., McCarthy, M. I., Small, K., and Marchini, J. Tensor decomposition for multiple-tissue gene expression experiments. *Nature Genetics*, 48(9):1094–1100, August 2016. ISSN 1546-1718. doi: 10.1038/ng.3624. URL http://dx.doi.org/10.1038/ng.3624.

Iversen, R., Snir, O., Stensland, M., Kroll, J. E., Steinsbø, , Korponay-Szabó, I. R., Lundin, K. E., de Souza, G. A., and Sollid, L. M. Strong clonal relatedness between serum and gut iga despite different plasma cell origins. *Cell Reports*, 20(10):2357–2367, September 2017. ISSN 2211-1247. doi: 10.1016/j.celrep.2017.

08.036. URL http://dx.doi.org/10.1016/j.celrep.2017.08.036.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. Kegg: integrating viruses and cellular organisms. *Nucleic Acids Research*, 49(D1):D545–D551, 2021. doi: 10.1093/nar/gkaa970. URL https://doi.org/10.1093/nar/gkaa970.

Kass, R. E. and Raftery, A. E. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, June 1995. ISSN 1537-274X. doi: 10.1080/01621459.1995.10476572. URL http://dx.doi.org/10.1080/01621459.1995.10476572.

Lafferty, J. and Blei, D. Correlated topic models. In *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005. URL https://proceedings.neurips.cc/paper_files/paper/2005/hash/9e82757e9a1c12cb710ad680db11f6f1-Abstract.html.

Lee, D. D. and Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. doi: 10.1038/44565.

Ng, K. W., Hobbs, A., Wichmann, C., Victora, G. D., and Donaldson, G. P. *B cell responses to the gut microbiota*, pp. 95–131. Elsevier, 2022. ISBN 9780323989459. doi: 10.1016/bs.ai.2022.08.003. URL http://dx.doi.org/10.1016/bs.ai.2022.08.003.

Okai, S., Usui, F., Yokota, S., Hori-i, Y., Hasegawa, M., Nakamura, T., Kurosawa, M., Okada, S., Yamamoto, K., Nishiyama, E., Mori, H., Yamada, T., Kurokawa, K., Matsumoto, S., Nanno, M., Naito, T., Watanabe, Y., Kato, T., Miyauchi, E., Ohno, H., and Shinkura, R. High-affinity monoclonal iga regulates gut microbiota and prevents colitis in mice. *Nature Microbiology*, 1(9), July 2016. ISSN 2058-5276. doi: 10.1038/nmicrobiol.2016.103. URL http://dx.doi.org/10.1038/nmicrobiol.2016.103.

Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, 2014. doi: 10.1093/bioinformatics/btu153. URL https://doi.org/10.1093/bioinformatics/btu153.

Stefan, K. L., Kim, M. V., Iwasaki, A., and Kasper, D. L. Commensal microbiota modulation of natural resistance to virus infection. *Cell*, 183(5):1312–1324.e10, November 2020. ISSN 0092-8674. doi: 10.1016/j.cell.2020.10.047. URL http://dx.doi.org/10.1016/j.cell.2020.10.047.

Wang, Z., Gerstein, M., and Snyder, M. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, January 2009. ISSN 1471-0064. doi: 10.1038/nrg2484. URL http://dx.doi.org/10.1038/nrg2484.

Wu, H.-J. and Wu, E. The role of gut microbiota in immune homeostasis and autoimmunity. *Gut Microbes*, 3(1):4–14, January 2012. ISSN 1949-0984. doi: 10.4161/gmic.19320. URL http://dx.doi.org/10.4161/gmic.19320.

Ye, Y. and Doak, T. G. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Computational Biology*, 5(8):e1000465, 2009. doi: 10.1371/journal.pcbi.1000465. URL https://doi.org/10.1371/journal.pcbi.1000465.

Zhao, M., Chu, J., Feng, S., Guo, C., Xue, B., He, K., and Li, L. Immunological mechanisms of inflammatory diseases caused by gut microbiota dysbiosis: A review. *Biomedicine amp; Pharmacotherapy*, 164:114985, August 2023. ISSN 0753-3322. doi: 10.1016/j.biopha.2023.114985. URL http://dx.doi.org/10.1016/j.biopha.2023.114985.

Zhou, J., He, Z., Yang, Y., Deng, Y., Tringe, S. G., and Alvarez-Cohen, L. High-throughput metagenomic technologies for complex microbial community analysis: Open and closed formats. *mBio*, 6(1), February 2015. ISSN 2150-7511. doi: 10.1128/mbio.02288-14. URL http://dx.doi.org/10.1128/mBio.02288-14.

A. Inference details

We employed stochastic amortized variational inference to approximate the posterior over latent variables in our model, which includes both continuous and discrete variables.

A.1. Amortized Variational Network For Continuous Latent Variables

We use an amortized inference network to parameterize the variational posterior $q(\mathbf{x}_{st};\, \boldsymbol{\xi}, \mathbf{y}_{st}) = \mathcal{N}\left(\mathbf{x}_{st}; NN_{\mu}(\mathbf{y}_{st};\, \boldsymbol{\xi}), NN_{\sigma^2}(\mathbf{y}_{st};\, \boldsymbol{\xi})\right)$ for the log of the latent group usage variable for subject s at time-point t. Here, $\boldsymbol{\xi}$ are parameters for the inference network.

The inference network is a three-layer fully connected neural network with SoftPlus activations and hidden layer dimensions of 100, with the input being the the concatenated log-transformed microbial and host gene expression counts $\mathbf{y}_{st}$, and the outputs of dimension $2K$ where K is the total number of groups. In the case of NN_{μ}, the outputs are real numbers; for NN_{σ^2}, the outputs are positive real numbers (enforced with an additional SoftPlus activation in the output layer).

Posterior samples of $\mathbf{x}_{st}$ are drawn using the reparameterization trick as:

$$\epsilon \sim Normal(0, \mathbf{I})$$

$$\mathbf{x}_{st} = NN_{\mu}(\mathbf{y}_{st};\, \xi) + \epsilon\sqrt{NN_{\sigma^2}(\mathbf{y}_{st};\, \xi)}$$

A.2. Concrete Distributions for Discrete Variables

We approximate binary latent variables for group activation (γ_{mk}), factor loading mask (η_{mki}), and time-specificity (α_{tmk}) using the Binary Concrete distribution:

$$\sigma\left(\frac{L + \log u - \log(1-u)}{\tau}\right), \quad u \sim \text{Uniform}(0,1)$$

Here, L are logit parameters and τ is the temperature. We anneal τ from 1.0 to 0.05 over training.

A.3. Marginalization of factors

To improve inference efficiency, we analytically marginalize over the latent factor vector $\vec{f}_{st\cdot}$ when computing the posterior distribution.

Recall that the latent log group usages $\vec{x}_{st\cdot\cdot} \in \mathbb{R}^{N_t}$ are modeled as linear combinations of factor weights:

$$\epsilon_{stmk} \sim \text{Normal}(0,1)$$

$$x_{stmk} = b_{mk} + \sum_{i} \eta_{mki}\omega_{mki}f_{sti} + \alpha_{tmk}\phi_{tmk} + \epsilon_{stmk}$$

This can be rewritten in a vector form as:

$$\vec{x}_{st\cdot\cdot} = (\vec{b}_{\cdot\cdot} + \vec{\alpha}_{t\cdot\cdot}\vec{\phi}_{t\cdot\cdot}) + W\vec{f}_{st\cdot} + \vec{\epsilon}$$

where $W \in \mathbb{R}^{N_t \times N_f}$ and $\vec{f}_{st\cdot} \sim Normal(0, I_{N_f})$. Under this formulation, $\vec{x}_{st\cdot\cdot}$ is a linear transformation of Gaussian variables. Hence, we can marginalize out the latent factor vector $\vec{f}_{st\cdot}$ analytically. The marginal distribution of $\vec{x}_{st\cdot\cdot}$ becomes:

$$\vec{x}_{st\cdot\cdot} \sim Normal(\vec{b}_{\cdot\cdot} + \vec{\alpha}_{t\cdot\cdot}\vec{\phi}_{t\cdot\cdot}, WW^T + I)$$

A.4. Overall ELBO

The overall evidence lower bound (ELBO) optimized during training consists of a data likelihood term and multiple KL terms:

$$\begin{aligned} \mathcal{L} = \ & \mathbb{E}_q[\log p(\boldsymbol{Y} \mid \boldsymbol{z}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\alpha})] \\ & -\text{KL}(q(\boldsymbol{z}) \| p(\boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\alpha})) \\ & -\text{KL}(q(\boldsymbol{\gamma}) \| p(\boldsymbol{\gamma})) - \text{KL}(q(\boldsymbol{\alpha}) \| p(\boldsymbol{\alpha})) \end{aligned}$$

Terms involving the Concrete distribution are estimated using the reparameterization trick and assuming Bernoulli distributions when computing the KL divergence. All expectations are estimated by taking one sample, and the ELBO is maximized using the Adam optimizer with other non-probabilistic model parameters, $\boldsymbol{\omega}$, $\boldsymbol{\phi}$ and $\boldsymbol{b}$, as well as the inference network parameters $\boldsymbol{\xi}$.