Learning accurate representations of microbe-metabolite interactions

James T. Morton,[1,2] Alexander A. Aksenov,[3,4] Louis Felix Nothias,[3,4] James R. Foulds,[5] Robert A. Quinn,[6] Michelle H. Badri,[7] Tami L. Swenson,[8] Marc W. Van Goethem,[8] Trent R. Northen,[8,9] Yoshiki Vazquez-Baeza,[1] Mingxun Wang,[3,4] Aaron Watters,[10] Se Jin Song,[1,11] Richard Bonneau,[7,10,12,13] Pieter C. Dorrestein,[3,4] and Rob Knight[1,2,14,11]

[1]*Department of Pediatrics, University of California, San Diego, La Jolla, CA, USA*

[2]*Department of Computer Science & Engineering, University of California, San Diego, La Jolla, CA, USA*

[3]*Collaborative Mass Spectrometry Innovation Center, University of California San Diego, La Jolla, CA, USA*

[4]*Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA*

[5]*Department of Information Systems, University of Maryland Baltimore County, Baltimore, MD, USA*

[6]*Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA*

[7]*Department of Biology, New York University, New York, 10012 NY, USA*

[8]*Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Rd, Berkeley, CA, 94720, USA*

[9]*DOE Joint Genome Institute, 2800 Mitchell Dr., Walnut Creek, CA, 94598, USA*

[10]*Flatiron Institute, Simons Foundation, New York, 10010 NY, USA*

[11]*Center for Microbiome Innovation, University of California, San Diego, La Jolla, CA, USA*

[12]*Computer Science Department, Courant Institute, New York, 10012 NY, USA*

[13]*Center For Data Science, NYU, New York, NY 10008, USA*

[14]*Department of Bioengineering University of California, San Diego, La Jolla, CA, USA*

I. ABSTRACT

Integrating multi-omics datasets is critical for microbiome research, but multiple statistical challenges can confound traditional correlation techniques. We solve this problem by using neural networks to estimate the conditional probability that each molecule is present given the presence of each specific microbe. We show with known environmental (desert biological soil crust wetting) and clinical (cystic fibrosis lung) examples, our ability to recover microbe-metabolite relationships, and demonstrate how the method can discover relationships between microbially-produced metabolites and inflammatory bowel disease.

II. INTRODUCTION

Knowledge gained by integrating complementary "-omics" data with a multi-omics approach will lead to improved diagnostics, automated drug discovery, and optimized culturing conditions for uncharacterized microbes [1]. However, because conventional correlation techniques have unacceptably high false discovery rates, finding meaningful relationships between genes within complex microbiomes and their products in the metabolome is challenging.

Although there has been a widespread effort to develop multi-omics approaches, several conceptual challenges limit techniques that integrate disparate "omics" data in general, including linking the microbial sequencing and untargeted mass spectrometry. Therefore, new approaches are needed that can handle disparate data types [2]. Relative abundances of thousands of microbes and metabolites can be measured using sequencing and mass spectrometry, which can result in the generation of very high dimensional microbiome and metabolomics datasets. Quantifying microbe-metabolite interactions requires estimating a distribution across all possible microbe-metabolite interactions.

Techniques such as Canonical Correspondence Analysis (CCA) and Partial Least Squares (PLS) approximate this joint distribution using a low dimensional representations [3–5]. Network models have been shown to improve classification accuracy using multiple datasets [6]. Factor models have been proposed to incorporate multiple datasets for biomarker analysis [7]. Despite of the wide application of these methods, they are notoriously difficult to interpret [8–10] and it remains unclear whether these models can obtain individual microbe-metabolite interactions.

Pearson and Spearman correlations assume independence between interactions, simplifying the estimation procedure by reducing it to a combination of independent two dimensional problems. However, many studies have shown that these methods are not statistically valid for compositional data, a fact first recognized by Pearson in 1895 and followed up in numerous studies [11–15]. This problem is further complicated because both microbiome [15] and mass spectrometry [16–19] datasets are also compositional, meaning that the absolute abundances are not measured, which can confound statistical inference. For example, in untargeted mass spectrometry experiments, the set of molecules detected and their relative abundance vary depending on the extraction protocol and analytic methods used, which leads to only a partial snapshot of the metabolome. Moreover, measuring the total mass of molecules extracted is often not performed in large scale metabolomics efforts, due to the highly laborious nature of that step.

To understand how issues associated with compositional data impact inference on microbe-metabolite interactions, consider the example in Figure S1. There are two microbes and two metabolites in Figure S1a. All are increasing exponentially at different rates and are highly correlated with each other. If proportions are estimated from the absolute abundances via sampling, the information about the total microbe population size and the total metabolite abundance is lost, and the correlations between the microbes and the metabolites disappear. False positives can also appear as shown in Figure S1b, microbe and metabolite interactions that have no apparent correlation structure may appear to be correlated when investigating the proportions. These issues alone can give rise to overwhelming false positives and false negatives, making Pearson and Spearman in some scenarios comparable to random coin flips. Experimental validation currently takes large laboratories multiple years to perform [20], often requiring time-consuming manual examinations of erroneous correlations.

There are other compositional techniques such as SparCC[11] and proportionality[21] that are scale-invariant when analyzing a single dataset, but lose scale-invariance when analyzing multiomics datasets. This was shown in the context of identifying microbe-fungal interactions [22], which provided motivation to extend SPIEC-EASI [12] to handle multiomics datasets. We show that this approach does not work for microbe-metabolite interactions because of differences of measurement units between sequencing and mass spectrometry measurements (Supplementary materials). An alternative approach is to consider co-occurrence

probabilities instead of correlations. Here, co-occurrence probabilities refer to the conditional probability of observing a metabolite given that a microbe was observed, thereby allowing us to identify the most likely microbe-metabolite interactions. To do this, we propose "mmvec", (microbe-metabolite vectors), to learn these co-occurrence probabilities between microbes and metabolites. Due to its scale-robustness properties, the microbial-metabolite relationships learned by mmvec are consistent between the absolute and relative abundances. The microbe-metabolite interactions can be ranked [23] and visualized through standard dimensionality reduction interfaces, enabling interpretable findings. The computations behind mmvec can take advantage of modern GPU architectures using Tensorflow [24], enabling scalable inference on large multiomics datasets. Furthermore, we provide evidence in two benchmarks and four case studies that mmvec outperforms existing statistical methods.

III. RESULTS AND DISCUSSION

We performed benchmarks comparing mmvec to Pearson, Spearman, SPIEC-EASI, SparCC and Proportionality [21] using a cystic fibrosis biofilm simulation. We then show that mmvec can resolve contradictory cyanobacteria-metabolite relationships in a desert soil biocrust wetting study. We also demonstrate recovery of known associations of *P. aeruginosa*-produced metabolites observed in cystic fibrosis [25]. Finally, we explore the relationships of microbiota and metabolic changes in mice fed a high fat diet [26] and inflammatory bowel disease [27], showing how this approach can be used to determine microbial origin of novel molecules even in extremely complex real-life biological systems with limited knowledge of existing associations.

A. Simulation benchmarks

To compare mmvec performance to Pearson, Spearman, Proportionality, SparCC and SPIEC-EASI correlations, we used data from existing studies in which the relationships between microbes and metabolites were the central focus of investigation. One such study simulated spatial-temporal dynamics in a microbial biofilm [25]. The original study tested the hypothesis that the cystic fibrosis (CF) microbiome community within human lungs can

be manipulated by altering its chemical environment. Changes in pH and oxygen saturation suppress the principal pathogen, *P. aeruginosa*, without using antibiotics, by promoting the growth of a community of fermenters that out-compete the pathogen. The simplicity of this system allowed the high-level ecological patterns to be modelled. In the original simulations, the interactions between two microbes (fermenters denoted by θ_f and *P. aeruginosa* denoted by θ_p) and multiple molecules were modeled using Monad kinetics and diffusion processes[25] (Figure 2a).

We simulated the measurement process for microbial DNA sequencing and untargeted mass spectrometry for metabolites as discussed in the Online Methods, providing ground truth information on their interactions. The model simulates interactions between *P. aeruginosa* and the fermenters, and their interactions with the environment. It also simulates known interactions between microbes and molecules, such as sugar consumption by fermenters and ammonia production by the pathogen. For example, the fermenters are positively associated with sugars and ammonium concentration, and negatively associated with inhibitor concentration; *P. aeruginosa* is positively associated with amino acids and pH.

Therefore, we can test whether the top K metabolites associated with each microbe by each tool includes the correct microbe-metabolite interactions. Figure 2c shows specificity and sensitivity for each tools as a function of K. In these simulations, random chance outperformed all of the tools except for mmvec and SPIEC-EASI, with mmvec performing the best. As shown in Figure 2d and Figure S2, mmvec is the only method robust to scale deviations amongst the methods tested. This is critical for maintaining consistency between absolute and relative abundances, which can otherwise lead to inflated false positives and false negatives [14].

B. Soil biocrust wetting event case study

Many studies produce inconsistent results that can be resolved with improved data analysis, especially in environmental and clinical settings. To test whether mmvec can resolve unexplained discrepancies in microbe-metabolite interactions across studies, we applied it to a study of biocrust wetting [28]. In this study, laboratory-based exometabolite patterns observed with bacterial isolates were reproduced in the environment. Specifically, in this work authors identified metabolites that were consumed and released by multiple biocrust

isolates including *Microcoleus vaginatus* and two *Bacillus* strains [29], and compared these patterns with closely-related environmental taxa and metabolites observed in situ [28].

While almost 70% of the examined microbe-metabolite relationships following the wetting event were validated [28], some contradicted the microbe-metabolite relationships observed in cultures [29]. These contradictions stemmed from Spearman correlations between *M. vaginatus* abundances and the observed metabolite abundances, but were resolved by mmvec (Figure 3a).

All metabolites released from the *M. vaginatus* isolate have higher conditional probabilities than the average metabolite following biocrust wetting, and are among the top 80 co-occurring metabolites with *M. vaginatus* (of 485 molecules total). This result supports the original finding that *M. vaginatus* actually releases these molecules after the wetting event. In contrast, Spearman labels 7 of 13 of these molecules with a negative correlation, indicating that these molecules were consumed by *M. vaginatus* rather than released, as originally stated in [28]. When the annotation detection rates amongst different statistical methodologies, mmvec has a substantially higher true positive rate as shown in Figure 3b.

The conflicting results between mmvec and Spearman could be explained by the growing microbial biomass and shift in available resources after wetting (Figure 3 c, d). Total biomass is expected to increase, because *M. vaginatus* releases metabolites that enable the growth of many other microbes. Because DNA sequencing can only measure proportions, the growth in other microbes could cause the proportions of *M. vaginatus* to decrease, leading to a misleading anti-correlation with 4-guanidinobutanoate (Figure 3d). However, it is not possible to infer whether *M. vaginatus* is decreasing in abundance [23] or 4-guanidinobutanoate is increasing in abundance.

The change in the total biomass and the total available resources could explain the contradiction between the Spearman correlations and the isolate results. *M. vaginatus* likely grows at a slower rate relative to other microbes that benefit from the metabolite release. Because mmvec does not rely on knowledge of the total biomass or normalize to relative abundance, these contradictions are avoided.

C. Cystic Fibrosis case study

To further validate if mmvec can detect known microbe-metabolite interactions in a biological setting, we re-analyzed a study on lung mucus microbiome of patients with cystic fibrosis [25, 30]. Cystic fibrosis has been shown to be dominated by two major groups of microbes, anaerobes and pathogens that occupy unique niches, and their interactions are defined by the environment. Anaerobes dominate in low oxygen and low pH environments, while pathogens, in particular *P. aeruginosa*, dominate in the opposite conditions [25]. Mmvec clearly separates anaerobes and pathogens (Figure 4a), with known anaerobic microbes (*Veillonella*, *Fusobacterium*, *Prevotella* and *Streptococcus*) on the left, and notable pathogens, such as *P. aeruginosa*, on the right.

P. aeruginosa is known to produce small-molecule virulence factors [31]. In the original study, based on annotations from GNPS[32], the bacterium was found to produce six molecules: 4-hydroxy-2-heptylquinoline (HHQ), Pyocyanin (PYO), Phenazine-1-carboxylic acid (PCA), 2-nonyl-4-hydroxy-quinoline (NHQ), 2-heptyl-3,4-dihydroxyquinoline (PQS, Pseudomonas quinolone signal) and Pyochelin [25]. As shown in Figure 4a, mmvec identifies these molecules with a high co-occurrence probability with *P. aeruginosa*. Mmvec also identifies a cluster of rhamnolipids likely produced by *P. aeruginosa*. Rhamnolipids are well characterized and are an important virulence factors for *P. aeruginosa*, contributing to biofilm development, motility on surfaces and antagonistic interactions with host inflammatory cells [33, 34]. These rhamnolipids were not identified in the original study [25]. The annotations for these compounds have been estiblished using GNPS [32].

There is a negative correlation between the first principal component learned from mmvec and the metabolites log-fold change across the oxygen gradient (Figure 4b) (Pearson r=-0.59, p-value 1.8×10^{-44}), which is consistent with the findings in the original work. No such correlation between the oxygen gradient and the first microbial principal component was found by Pearson (r=0.01, p=0.89). There exist two notable microbes on opposing ends of the first microbial principal component: *P. aeruginosa*, a known pathogen, and *Streptococcus*, a known anaerobe. The top 100 metabolites that are specific to *P. aeruginosa* and *Streptococcus* are shown to have drastically different profiles in samples where *P. aeruginosa* and *Streptococcus* were the most abundant species (Figure 4d,e) (logratio t-test=6.51, p=4.4×10^{-8}). This provides evidence that in the context of this study, the

metabolomic profiles can be largely influenced by the most abundant microbes, a notion that has important implications for understanding CF etiology. To further support this, the learned metabolite conditional probabilities for *P. aeruginosa* can be used to predict the metabolite proportions in the 41 samples where *P. aeruginosa* is the most abundant taxa. The predicted *P. aeruginosa* metabolite profiles alone can explain 10% of the metabolite variation in these samples (r=0.319, p=1.18×10^{-11}).

Of 14 quinolone molecules known to be produced by *P. aeruginosa*, Pearson correlation detected 9 with $p < 0.05$ without FDR correction, and only 5 with FDR correction. For example, Pyocyanin, does not appear related to *P. aeruginosa* by the raw proportions (r=0.158, FDR-corrected pvalue=0.089, rank=96), but is ranked 34th most associated with *P. aeruginosa* by mmvec (Figure S3c), consistent with culturing experiments that demonstrate that *P. aeruginosa* produces this molecule [35]. 18 rhamnolipids are among the top 25 metabolites most associated with *P. aeruginosa* by mmvec, and have higher ranks with mmvec than with Pearson correlation (Figure S3b).

D. Effects of high fat diet in murine model case study

We then tested whether mmvec could determine the microbial origin of specific molecules in a complex biological system. We recently discovered a new kind of bile acid, where cholate is conjugated to amino acids other than glycine and taurine [36]. These molecules increased in abundance with high-fat diet in humans. We determined that these molecules are microbially-made since they were present in specific pathogen free, but not in germ free mice. We therefore set out to identify candidate producers. We were able to confirm that one of these bile acids, cholate phenylalanine amidate, was associated with high-fat diet in well-controlled study that investigated the development of non-alcoholic fatty liver disease (NAFLD), cirrhosis, and hepatocarcinoma (HCC) in a mouse model [26]. When re-analyzing these datasets for differential abundances via multinomial regression, the strong association of the novel bile acid with HFD became immediately apparent. The use of mmvec showed distinct associated groups of microbes and HFD (Figure 5a) and a clear stratification of the mass spectrometry data according to diet (Figure 5b). Several *Clostridium spp.* correlated with the cholate phenylalanine conjugate. Indeed, we showed that *Clostridium spp.* were found to produce this bile acid [36]. This result demonstrates mmvec's ability to streamline

the discovery of microbes that produce specific molecules of interest *in vivo*.

E. Microbe-metabolite interactions in Inflammatory Bowel Disease

Finally, microbe-metabolite interactions were investigated for samples of IBD patients generated under the integrative Human Microbiome Project [27]. The role of the microbiome in IBD is acknowledged, but still poorly understood. The original study uncovered shifts in metabolomic and microbial profiles associated with the IBD. In particular, levels of carnitines and bile acids were shown to be affected [27]. Using mmvec we confirmed the core findings in the previous study, such as the co-occurrence between *R. hominis* and multiple carnitines, including previously noted C20, which have anti-inflammatory properties (Figure 6a) [27]. We also found high correlation of *Klebsiella spp.* with IBD status and that it co-occurs with high probability with several bile acids (Figure 6b). Although *Klebsiella* itself does not produce these compounds, some pathogens (including *Klebsiella*) are known to be resistant to bile acids [37]. Excessive production of some bile acids and bile acid malabsorption can lead to overabundance of bile acids, which is a hallmark of IBD [38], although the exact mechanisms remain unknown. The ability of *Klebsiella* to thrive in concentrated bile acid environments is consistent with the high co-occurrence probabilities shown in Figure 6b. We also noted that three *Klebsiella* species are the top drivers of the IBD- associated molecules (Figure 6c). It is important to delineate different reasons for co-occurrence. Unlike *Klebsiella*, *Clostridium* species are known for bile acid manipulation, including production of bile that can germinate *Clostridium difficile* spores or that have anti-microbial properties [39, 40].

Therefore, it is possible that in case of *Clostridia*, the existing co-occurrences (Figure 6b) are due to actual biosynthesis of the metabolites by the microbial species indicated rather than ability to withstand them.

In addition to recapitulating reported findings, mmvec also yielded previously undetected relationships. The major microbe that was found to be associated with healthy patients is *Propionibacteriaceae*, which was not detected in Price et al 2019 (Figure 6cd). This relationship is corroborated by other published studies. In one study, it has been shown that some members of the *Propionibacterium* genus produce 1.4-Dihydroxy-2-naphthoic acid (DHNA), a growth stimulator for bacteria such as *Bifidobacterium* that are thought to reduce the

symptoms of IBD [41]. Also, in a survey of *in vivo* vs. *in vitro* bacterial activity, *Propionibacterium freudenreichii* was shown to play an immunomodulatory role in the context of an ulcerative colitis mice model [42]. In another study it was shown that *Propionibacterium freudenreichii* is a viable core component in an anti-inflammatory probiotic fermented dairy product [43]. The members of this family have been considered beneficial for intestinal immunoregulation; *Propionibacteriaceae* have been observed to be enriched in human breast milk and have been shown to restore Th17 differentiation [44]. Thus, it appears that the existing knowledge supports the statistically-inferred interaction uncovered by mmvec, but not identified in the published analysis of the dataset

IV. CONCLUSION

In both simulation benchmarks and annotated dataset, mmvec shows promise for inferring microbe-metabolite interactions from multiomics datasets. Our benchmarks suggest that mmvec outperforms all existing tools that aim to infer interactions between paired microbe-metabolite abundance datasets, both in simulations and in experimental data. In the biocrust wetting experiment, mmvec resolved conflicting findings between the *in vitro* validated *M. vaginatus* released metabolites and the sequencing/mass spectrometry analysis of environmental samples. In the cystic fibrosis study, mmvec can reliably identify all of the experimentally determined *P. aeroginosa*-produced molecules of interest. We show in the example of bile acid production that mmvec enables exploratory analysis in complex biological systems and streamlined discovery of the microbial origin of specific metabolites. Finally, mmvec was able to identify the strongest microbial contributions to the metabolite abundances in the IBD study, where one of those microbes was missed in the original study.

In light of these findings, the current methodology still has limitations. It remains unclear how to access statistical significance of an interaction using co-occurrence probabilities. Similarly, confidence intervals for the strength of each microbe-metabolite interaction can not yet be calculated. Furthermore, more theoretical work will be required to handle continuous-valued inputs.

The concepts outlined here should generalize beyond microbe-metabolite interactions to handle other paired multi-omic data types, provided that the input dataset is made up of counts (as in metagenomics, transcriptomics, etc.). With the exponential growth of multi-

omics datasets, there is much potential to use these methods to reveal microbial metabolism, including for microbes that are not cultivable in the laboratory. Approaches utilizing co-occurrence probabilities have the potential to enable more targeted experimental assays, accelerating the discovery of microbe-metabolite interactions, paving the way towards new ecosystems engineering approaches in clinical, environmental and industrial applications.

V. ACKNOWLEDGEMENTS

We would like to thank Vera Pawlowsky, Juan Jose Egozcue and Susan Holmes for their insights behind the geometry of this neural network model. T.L.S., M.W.V.G and T.R.N greatly acknowledge funding from the Office Science Early Career Research Program, Office of Biological and Environmental Research, of the U.S. Department of Energy under contract number DE-AC02-05CH11231 to Lawrence Berkeley National Laboratory. This was in part supported by P41GM103484 Center for Computational Mass Spectrometry, Instrument support through NIH S10RR029121 and R03 CA211211 on reuse of metabolomics data. Y. V. B. is funded by the Janssen Human Microbiome Institute through a collaboration with the Center for Microbiome Innovation. J.T.M. was funded by NSF grant GRFP DGE-1144086

VI. AUTHOR CONTRIBUTIONS

J.T.M wrote the mmvec algorithm, conducted the benchmarks and ran all of the analyses. A.A. and L.F.N. preprocessed and annotated the metabolomics data. A.A. provided insights in the high fat diet study. J.F. provided insights behind word2vec and topic modeling. M.H.B. benchmarked SPIEC-EASI. R.A.Q. provided insights behind the cystic fibrosis study and simulations. Y.V.B. provided insights behind the interpretation of the IBD analysis. M.W. developed the GNPS workflow for mmvec. A.W developed the network visualizations. T.S. M.V.G and T.N. provided insights behind the biocrust soils experiment. All authors were involved with writing the manuscript.

VII. COMPETING INTERESTS

None of the authors have any competing interests.

VIII. REFERENCES

[1] Janet K Jansson and Erin S Baker. A multi-omic future for microbiome studies. *Nat Microbiol*, 1(16049):645, 2016.

[2] Rob Knight, Alison Vrbanac, Bryn C Taylor, Alexander Aksenov, Chris Callewaert, Justine Debelius, Antonio Gonzalez, Tomasz Kosciolek, Laura-Isobel McCall, Daniel McDonald, et al. Best practices for analysing microbiomes. *Nature Reviews Microbiology*, page 1, 2018.

[3] Chen Meng, Oana A Zeleznik, Gerhard G Thallinger, Bernhard Kuster, Amin M Gholami, and Aedín C Culhane. Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinform.*, 17(4):628–641, July 2016.

[4] Gwénaëlle Le Gall, Samah O Noor, Karyn Ridgway, Louise Scovell, Crawford Jamieson, Ian T Johnson, Ian J Colquhoun, E Kate Kemsley, and Arjan Narbad. Metabolomics of fecal extracts

detects altered metabolic activity of gut microbiota in ulcerative colitis and irritable bowel syndrome. *Journal of proteome research*, 10(9):4208–4218, 2011.

[5] Florian Rohart, Benoit Gautier, Amrit Singh, and Kim-Anh Le Cao. mixomics: An r package for 'omics feature selection and multiple data integration. *PLoS computational biology*, 13(11):e1005752, 2017.

[6] Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333, 2014.

[7] Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, 14(6):e8124, 2018.

[8] Cajo JF Ter Braak and Piet FM Verdonschot. Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquatic sciences*, 57(3):255–289, 1995.

[9] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

[10] Antoine Bodein, Olivier Chapleur, Arnaud Droit, and Kim-Anh Lê Cao. A generic multivariate framework for the integration of microbiome longitudinal studies with other data types. *bioRxiv*, page 585802, 2019.

[11] Jonathan Friedman and Eric J Alm. Inferring correlation networks from genomic survey data. *PLoS computational biology*, 8(9):e1002687, 2012.

[12] Zachary D Kurtz, Christian L Müller, Emily R Miraldi, Dan R Littman, Martin J Blaser, and Richard A Bonneau. Sparse and compositionally robust inference of microbial ecological networks. *PLoS computational biology*, 11(5):e1004226, 2015.

[13] Sophie Weiss, Will Van Treuren, Catherine Lozupone, Karoline Faust, Jonathan Friedman, Ye Deng, Li Charlie Xia, Zhenjiang Zech Xu, Luke Ursell, Eric J Alm, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME journal*, 10(7):1669, 2016.

[14] Doris Vandeputte, Gunter Kathagen, Kevin D'hoe, Sara Vieira-Silva, Mireia Valles-Colomer, João Sabino, Jun Wang, Raul Y Tito, Lindsey De Commer, Youssef Darzi, Séverine Vermeire,

Gwen Falony, and Jeroen Raes. Quantitative microbiome profiling links gut community variation to microbial load. *Nature*, 551(7681):507–511, November 2017.

[15] Gregory B Gloor, Jean M Macklaim, Vera Pawlowsky-Glahn, and Juan J Egozcue. Microbiome datasets are compositional: And this is not optional. *Front. Microbiol.*, 8, 2017.

[16] Keqi Tang, Jason S Page, and Richard D Smith. Charge competition and the linear dynamic range of detection in electrospray ionization mass spectrometry. *J. Am. Soc. Mass Spectrom.*, 15(10):1416–1423, 2004.

[17] Richard King, Ryan Bonfiglio, Carmen Fernandez-Metzler, Cynthia Miller-Stein, and Timothy Olah. Mechanistic investigation of ionization suppression in electrospray ionization. *J. Am. Soc. Mass Spectrom.*, 11(11):942–950, 2000.

[18] B K Matuszewski, M L Constanzer, and C M Chavez-Eng. Strategies for the assessment of matrix effect in quantitative bioanalytical methods based on HPLCMS/MS. *Anal. Chem.*, 75(13):3019–3030, 2003.

[19] Alžběta Kalivodová, Karel Hron, Peter Filzmoser, Lukáš Najdekr, Hana Janečková, and Tomáš Adam. Pls-da for compositional data with application to metabolomics. *Journal of Chemometrics*, 29(1):21–28, 2015.

[20] J. K. Jansson and E. S. Baker. A multi-omic future for microbiome studies. *Nat Microbiol*, 1:16049, 04 2016.

[21] David Lovell, Vera Pawlowsky-Glahn, Juan José Egozcue, Samuel Marguerat, and Jürg Bähler. Proportionality: a valid alternative to correlation for relative data. *PLoS computational biology*, 11(3):e1004075, 2015.

[22] Laura Tipton, Christian L Müller, Zachary D Kurtz, Laurence Huang, Eric Kleerup, Alison Morris, Richard Bonneau, and Elodie Ghedin. Fungi stabilize connectivity in the lung and skin microbial ecosystems. *Microbiome*, 6(1):12, 2018.

[23] James T Morton, Clarisse Marotz, Alex Washburne, Justin Silverman, Livia S Zaramela, Anna Edlund, Karsten Zengler, and Rob Knight. Establishing microbial composition measurement standards with reference frames. *Nature Communications*, 10(1):2719, 2019.

[24] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[25] Robert A Quinn, William Comstock, Tianyu Zhang, James T Morton, Ricardo da Silva, Alda Tran, Alexander Aksenov, Louis-Felix Nothias, Daniel Wangpraseurt, Alexey V Melnik, Gail Ackermann, Douglas Conrad, Isaac Klapper, Rob Knight, and Pieter C Dorrestein. Niche partitioning of a pathogenic microbiome driven by chemical gradients. *Sci Adv*, 4(9):eaau1908, September 2018.

[26] Shabnam Shalapour, Xue-Jia Lin, Ingmar N Bastian, John Brain, Alastair D Burt, Alexander A Aksenov, Alison F Vrbanac, Weihua Li, Andres Perkins, Takaji Matsutani, et al. Inflammation-induced iga+ cells dismantle anti-liver cancer immunity. *Nature*, 551(7680):340, 2017.

[27] J. Lloyd-Price, C. Arze, A. N. Ananthakrishnan, M. Schirmer, J. Avila-Pacheco, T. W. Poon, E. Andrews, N. J. Ajami, K. S. Bonham, C. J. Brislawn, D. Casero, H. Courtney, A. Gonzalez, T. G. Graeber, A. B. Hall, K. Lake, C. J. Landers, H. Mallick, D. R. Plichta, M. Prasad, G. Rahnavard, J. Sauk, D. Shungin, Y. Vazquez-Baeza, R. A. White, J. Braun, L. A. Denson, J. K. Jansson, R. Knight, S. Kugathasan, D. P. B. McGovern, J. F. Petrosino, T. S. Stappenbeck, H. S. Winter, C. B. Clish, E. A. Franzosa, H. Vlamakis, R. J. Xavier, C. Huttenhower, J. Bishai, K. Bullock, A. Deik, C. Dennis, J. L. Kaplan, H. Khalili, L. J. McIver, C. J. Moran, L. Nguyen, K. A. Pierce, R. Schwager, A. Sirota-Madi, B. W. Stevens, W. Tan, J. J. Ten Hoeve, G. Weingart, R. G. Wilson, and V. Yajnik. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758):655–662, May 2019.

[28] Tami L Swenson, Ulas Karaoz, Joel M Swenson, Benjamin P Bowen, and Trent R Northen. Linking soil biology and chemistry in biological soil crust using isolate exometabolomics. *Nature communications*, 9(1):19, 2018.

[29] Richard Baran, Eoin L Brodie, Jazmine Mayberry-Lewis, Eric Hummel, Ulisses Nunes Da Rocha, Romy Chakraborty, Benjamin P Bowen, Ulas Karaoz, Hinsby Cadillo-Quiroz, Ferran Garcia-Pichel, et al. Exometabolite niche partitioning among sympatric soil bacteria. *Nature communications*, 6:8289, 2015.

[30] Robert A Quinn, Katrine Whiteson, Yan-Wei Lim, Peter Salamon, Barbara Bailey, Simone Mienardi, Savannah E Sanchez, Don Blake, Doug Conrad, and Forest Rohwer. A winogradsky-based culture system shows an association between microbial fermentation and cystic fibrosis exacerbation. *ISME J.*, 9(4):1024–1038, March 2015.

[31] Wilna J Moree, Vanessa V Phelan, Cheng-Hsuan Wu, Nuno Bandeira, Dale S Cornett, Brendan M Duggan, and Pieter C Dorrestein. Interkingdom metabolic transformations captured by microbial imaging mass spectrometry. *Proceedings of the National Academy of Sciences*, 109(34):13811–13816, 2012.

[32] Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, Jeramie Watrous, Clifford A Kapono, Tal Luzzatto-Knaan, Carla Porto, Amina Bouslimani, Alexey V Melnik, Michael J Meehan, Wei-Ting Liu, Max Crüsemann, Paul D Boudreau, Eduardo Esquenazi, Mario Sandoval-Calderón, Roland D Kersten, Laura A Pace, Robert A Quinn, Katherine R Duncan, Cheng-Chih Hsu, Dimitrios J Floros, Ronnie G Gavilan, Karin Kleigrewe, Trent Northen, Rachel J Dutton, Delphine Parrot, Erin E Carlson, Bertrand Aigle, Charlotte F Michelsen, Lars Jelsbak, Christian Sohlenkamp, Pavel Pevzner, Anna Edlund, Jeffrey McLean, Jörn Piel, Brian T Murphy, Lena Gerwick, Chih-Chuang Liaw, Yu-Liang Yang, Hans-Ulrich Humpf, Maria Maansson, Robert A Keyzers, Amy C Sims, Andrew R Johnson, Ashley M Sidebottom, Brian E Sedio, Andreas Klitgaard, Charles B Larson, Cristopher A Boya P, Daniel Torres-Mendoza, David J Gonzalez, Denise B Silva, Lucas M Marques, Daniel P Demarque, Egle Pociute, Ellis C O'Neill, Enora Briand, Eric J N Helfrich, Eve A Granatosky, Evgenia Glukhov, Florian Ryffel, Hailey Houson, Hosein Mohimani, Jenan J Kharbush, Yi Zeng, Julia A Vorholt, Kenji L Kurita, Pep Charusanti, Kerry L McPhail, Kristian Fog Nielsen, Lisa Vuong, Maryam Elfeki, Matthew F Traxler, Niclas Engene, Nobuhiro Koyama, Oliver B Vining, Ralph Baric, Ricardo R Silva, Samantha J Mascuch, Sophie Tomasi, Stefan Jenkins, Venkat Macherla, Thomas Hoffman, Vinayak Agarwal, Philip G Williams, Jingqui Dai, Ram Neupane, Joshua Gurr, Andrés M C Rodríguez, Anne Lamsa, Chen Zhang, Kathleen Dorrestein, Brendan M Duggan, Jehad Almaliti, Pierre-Marie Allard, Prasad Phapale, Louis-Felix Nothias, Theodore Alexandrov, Marc Litaudon, Jean-Luc Wolfender, Jennifer E Kyle, Thomas O Metz, Tyler Peryea, Dac-Trung Nguyen, Danielle VanLeer, Paul Shinn, Ajit Jadhav, Rolf Müller, Katrina M Waters, Wenyuan Shi, Xueting Liu, Lixin Zhang, Rob Knight, Paul R Jensen, Bernhard O Palsson, Kit Pogliano, Roger G Linington, Marcelino Gutiérrez, Norberto P Lopes, William H Gerwick, Bradley S Moore, Pieter C Dorrestein, and Nuno Bandeira. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat. Biotechnol.*, 34(8):828–837, August 2016.

[33] Raina Margaret Maier and G Soberon-Chavez. Pseudomonas aeruginosa rhamnolipids: biosynthesis and potential applications. *Applied Microbiology and Biotechnology*, 54(5):625–633, 2000.

[34] Thammajun L Wood, Ting Gong, Lei Zhu, James Miller, Daniel S Miller, Bei Yin, and Thomas K Wood. Rhamnolipids from pseudomonas aeruginosa disperse the biofilms of sulfate-reducing bacteria. *NPJ biofilms and microbiomes*, 4(1):22, 2018.

[35] Lucy Allen, David H Dockrell, Theresa Pattery, Daniel G Lee, Pierre Cornelis, Paul G Hellewell, and Moira KB Whyte. Pyocyanin production by pseudomonas aeruginosa induces neutrophil apoptosis and impairs neutrophil-mediated host defenses in vivo. *The Journal of Immunology*, 174(6):3643–3649, 2005.

[36] Robert A. Quinn, Alison Vrbanac, Alexey V. Melnik, Kathryn A. Patras, Mitchell Christy, Andrew T. Nelson, Alexander Aksenov, Anupriya Tripathi, Greg Humphrey, Ricardo da Silva, Robert Bussell, Taren Thron, Mingxun Wang, Fernando Vargas, Julia M. Gauglitz, Michael J. Meehan, Orit Poulsen, Brigid S. Boland, John T. Chang, William J. Sandborn, Meerana Lim, Neha Garg, Julie Lumeng, Barbara I. Kazmierczak, Ruchi Jain, Marie Egan, Kyung E. Rhee, Gabriel G. Haddad, Dionicio Siegel, Sarkis Mazmanian, Victor Nizet, Rob Knight, and Pieter C. Dorrestein. Chemical impacts of the microbiome across scales reveal novel conjugated bile acids. *bioRxiv*, 2019.

[37] Michelle K. Paczosa and Joan Mecsas. Klebsiella pneumoniae: Going on the offense with a strong defense. *Microbiology and Molecular Biology Reviews*, 80(3):629–661, 2016.

[38] Elisa Tiratterra, Placido Franco, Emanuele Porru, Konstantinos H Katsanos, Dimitrios K Christodoulou, and Giulia Roda. Role of bile acids in inflammatory bowel disease. *Annals of gastroenterology*, 31(3):266, 2018.

[39] Alan F Hofmann and Lars Eckmann. How bile acids confer gut mucosal protection against bacteria. *Proceedings of the National Academy of Sciences*, 103(12):4333–4334, 2006.

[40] Máire Begley, Cormac GM Gahan, and Colin Hill. The interaction between bacteria and bile. *FEMS microbiology reviews*, 29(4):625–651, 2005.

[41] Y. Okada, Y. Tsuzuki, J. Miyazaki, K. Matsuzaki, R. Hokari, S. Komoto, S. Kato, A. Kawaguchi, S. Nagao, K. Itoh, T. Watanabe, and S. Miura. Propionibacterium freudenreichii component 1.4-dihydroxy-2-naphthoic acid (DHNA) attenuates dextran sodium sulphate induced colitis by modulation of bacterial flora and lymphocyte homing. *Gut*, 55(5):681–688,

May 2006.

[42] B. Foligne, S. Parayre, R. Cheddani, M. H. Famelart, M. N. Madec, C. Ple, J. Breton, J. Dewulf, G. Jan, and S. M. Deutsch. Immunomodulation properties of multi-species fermented milks. *Food Microbiol.*, 53(Pt A):60–69, Feb 2016.

[43] C. Ple, J. Breton, R. Richoux, M. Nurdin, S. M. Deutsch, H. Falentin, C. Herve, V. Chuat, R. Lemee, E. Maguin, G. Jan, M. Van de Guchte, and B. Foligne. Combining selected immunomodulatory Propionibacterium freudenreichii and Lactobacillus delbrueckii strains: Reverse engineering development of an anti-inflammatory cheese. *Mol Nutr Food Res*, 60(4):935–948, Apr 2016.

[44] Natacha Colliou, Yong Ge, Bikash Sahay, Minghao Gong, Mojgan Zadeh, Jennifer L Owen, Josef Neu, William G Farmerie, Francis Alonzo, Ken Liu, et al. Commensal propionibacterium strain uf1 mitigates intestinal inflammation via th17 cell regulation. *The Journal of clinical investigation*, 127(11):3970–3986, 2017.

[45] Nasser M Nasrabadi. Pattern recognition and machine learning. *Journal of electronic imaging*, 16(4):049901, 2007.

[46] Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado. *Modeling and Analysis of Compositional Data*. John Wiley & Sons, February 2015.

[47] Mikolov, Tomas and Sutskever, Ilya and Chen, Kai and Corrado, Greg S and Dean, Jeff. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[48] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.

[49] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[51] Kris Sankaran and Susan P Holmes. Latent variable modeling for the microbiome. *arXiv preprint arXiv:1706.04969*, 2017.

[52] John Aitchison and Michael Greenacre. Biplots of compositional data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(4):375–392, 2002.

[53] John Aitchison, KW Ng, et al. Conditional compositional biplots: theory and application. 2005.

[54] JA Martín-Fernández, V Pawlowsky-Glahn, JJ Egozcue, and R Tolosona-Delgado. Advances in principal balances for compositional data. *Mathematical Geosciences*, 50(3):273–298, 2018.

[55] Michael A Skinnider, Jordan W Squair, and Leonard J Foster. Evaluating measures of association for single-cell transcriptomics. *Nature methods*, 16(5):381, 2019.

[56] Han Liu, Kathryn Roeder, and Larry Wasserman. Stability approach to regularization selection (stars) for high dimensional graphical models. In *Advances in neural information processing systems*, pages 1432–1440, 2010.

[57] Evan Bolyen, Jai Ram Rideout, Matthew R Dillon, Nicholas A Bokulich, Christian Abnet, Gabriel A Al-Ghalith, Harriet Alexander, Eric J Alm, Manimozhiyan Arumugam, Francesco Asnicar, et al. Qiime 2: Reproducible, interactive, scalable, and extensible microbiome data science. Technical report, PeerJ Preprints, 2018.

[58] Yoshiki Vázquez-Baeza, Meg Pirrung, Antonio Gonzalez, and Rob Knight. Emperor: a tool for visualizing high-throughput microbial community data. *Gigascience*, 2(1):16, 2013.

[59] Oliver Fiehn, Don Robertson, Jules Griffin, Mariet van der Werf, Basil Nikolau, Norman Morrison, Lloyd W Sumner, Roy Goodacre, Nigel W Hardy, Chris Taylor, et al. The metabolomics standards initiative (msi). *Metabolomics*, 3(3):175–178, 2007.

IX. METHODS

A. Mmvec neural network architecture

The development of our proposed neural network was inspired by applications in natural language processing. The underlying model can also be referred to as a bi-loglinear multinomial regression. Our mmvec model posits an assumed generative process for the data, which leads to an inference algorithm to recover the model's parameters from multi-omics data. The model's assumed generative model for metabolite ν, microbe μ and sample k given as follows.

First generate microbe vector $\boldsymbol{u}_\mu$ for microbe $\mu \in \{1, ...N\}$ and metabolite vectors $\boldsymbol{v}_{\boldsymbol{\nu}}$ for

metabolite $\nu \in \{1, ...M\}$,

$$\boldsymbol{u}_\mu \sim \mathcal{N}(\boldsymbol{0}, \sigma_u I) \qquad \boldsymbol{v_\nu} \sim \mathcal{N}(\boldsymbol{0}, \sigma_v I) \ ,$$

These vectors are length p, corresponding to the number of latent vectors dimensions. Each of these vectors are drawn from a normal prior centered around zero and a diagonal covariance matrix with variances σ_u and σ_v, namely to serve regularization purposes and avoid overfitting. For a given microbial sample x_k, the models generative process draws a single microbe from a single draw from the categorical distribution

$$\mu \sim Categorical(\boldsymbol{x_k}) \ .$$

That microbe μ can be used to index U in order to generate conditional probabilities $\boldsymbol{q_\mu}$

$$p(\nu|\mu) = \frac{\exp(\boldsymbol{v_\nu} \cdot \boldsymbol{u}_\mu + \nu_{\nu 0} + u_{\mu 0})}{\sum_j \exp(\boldsymbol{v_j} \cdot \boldsymbol{u}_\mu + \nu_{j0} + u_{\mu 0})} \ ,$$

$$\boldsymbol{q_\mu} = [p(\nu_1|\mu), \ldots p(\nu_M|\mu)]$$

Here, $\nu_j 0 + u_{\mu 0}$ are row and column biases, which are required to accurately estimate the conditional probabilities. The above transformation is the softmax transform [45] to compute probabilities from real-valued quantities. This transformation is also known as the inverse clr transform [46], which enforces scale invariance as shown in the simulations. In the mmvec model's generative process, these conditional probabilities generate the metabolite abundances $\boldsymbol{y_k}$ for a given sample k through a multinomial distribution.

$$\boldsymbol{y_k} \sim Multinomial(n, \boldsymbol{q_\mu}) \ ,$$

where n is the total metabolite abundances across sample k. It is important to note that metabolite abundances themselves are not counts, but rather a continuous representation of molecule counts. We make the simplifying assumption that these continuous valued abundances can be approximated by Multinomial count models.

This model bears resemblance to how word2vec estimates word probabilities conditioned on a single particular word [47]. There are a couple of majors differences to be considered. First, in the original application of word2vec, a skipgram was proposed. Skipgrams [47] have been designed to account for the sequential nature of text. There is no such sequential

nature with microbiome or metabolite samples, the only ordering information that is known is the sample membership. As a result, the skipgrams can be replaced using multinomial sampling, where a single microbe is randomly sampled from a microbiome sample at each gradient descent step.

Second, in the original word2vec application a single input/output word pair were evaluated at each gradient descent step, which is required to incorporate the contextual information of words within sentences. In the application of multiomics, this is unnecessasrily complicated, since there is no such contextual with regards to microbes and metabolites. Instead, all of the metabolite abundances can be simultaneously evaluated for each gradient descent step, ultimately speeding up computations. Specifically, these metabolite abundances are simultaneously considered in order to estimate the conditional probabilities q_k for the given microbial count u_{jk}. From these conditional probabilities, the metabolite abundances y_k are generated from a Multinomial distribution. This process is repeated across all of the microbial reads. To show that $p(\nu|\mu)$ truly approximates the probability of observing a metabolite given a microbe, we first need to make the simplifying assumption that the conditional distribution of a metabolite given the presence of a single microbe also follows a multinomial distribution as follows

$$p(Y = y|X_\mu = 1) = Multinomial(y|q_\mu)$$

Where y is the vector of observed metabolites, Y is the random variable modeling metabolite abundances, X is a random variable modeling microbe abundances, x is a vector of observed microbes and μ is a single microbe. Given these modeling assumptions, we can parameterize the conditional Multinomial distributions with embedding vectors as described above. This estimation procedure can be reformulated as a matrix factorization, where the conditional probability matrix is decomposed into two weight matrices $\boldsymbol{U}$ and $\boldsymbol{V}$, which are comprised of microbe-metabolite vectors as follows

$$\boldsymbol{U} = [\mathbf{0}, \boldsymbol{u_0}, \boldsymbol{u_1}, ..., \boldsymbol{u_N}]^T \qquad\qquad \boldsymbol{V} = [\boldsymbol{v_0}, \mathbf{0}, \boldsymbol{v_1}, ..., \boldsymbol{v_M}] \ .$$

Here $\boldsymbol{U} \in \mathbb{R}^{N \times p}$ and $\boldsymbol{V} \in \mathbb{R}^{(M-1) \times p}$ represents the corresponding embeddings for N microbes and M metabolites. The number dimensions p for both $\boldsymbol{U}$ and $\boldsymbol{V}$ as well as the priors are specified by the user, but can also be evaluated during cross-validation. The biases $\boldsymbol{u_0}$ and

$\boldsymbol{v_0}$ are critical for estimating accurate co-occurrence probabilities, as suggested by similar methodologies used in recommender systems [48]. The $\boldsymbol{U}$ and $\boldsymbol{V}$ matrices are estimated through maximum a posteriori (MAP) estimation using ADAM [49] with the following log-posterior

$$\begin{aligned}
\mathcal{L} &= \mathcal{L}_Y + \mathcal{L}_U + \mathcal{L}_V \\
\mathcal{L}_U &= \sum_{\mu} \sum_{\rho=1}^{p} \mathcal{N}(U_{\mu,\rho}|0, \sigma_u) \\
\mathcal{L}_V &= \sum_{\nu} \sum_{\rho=1}^{p} \mathcal{N}(V_{\nu,\rho}|0, \sigma_v) \\
\mathcal{L}_Y &= \sum_{k} \sum_{r \in x_k} Multinomial(\boldsymbol{y_k}|\boldsymbol{q_\mu}) \ .
\end{aligned}$$

Within a single iteration of stochastic gradient descent a single microbial sequence i is randomly drawn and compared to a complete set of metabolite abundances y_i for that given sample. If there are a total of R microbial reads across all of the microbial samples, there will be R iterations for a complete epoch over the microbial dataset. This means that the running time of this training process is $O(RM)$ for a single epoch. Cross validation can be performed by holding out samples measuring the predictive power by looking at the sum of squares errors. Predictions can be made as follows

$$SSE = \sum_{k,i} (y_k - m_k \cdot softmax(\boldsymbol{V}\boldsymbol{U}_{u_{ik},\cdot}))^2 \ .$$

Where the predictive metabolite abundances are compared to the holdout abundances y_k across all microbial reads i in the holdout samples k. m_k denotes the total metabolite abundances in sample k

B. Microbe-metabolite vectors in simplicial coordinates

Here, we will provide some insights behind the underlying geometry behind this neural network. Doing so will provide intuition behind the algebraic operations commonly applied in the context of word2vec, suggesting the possibility of performing similar tasks in the context of microbe-metabolite interactions. Furthermore, this will motivate the use of the

Aitchison distance to quantify microbe-microbe and metabolite-metabolite interactions. Finally we will make a connection to topic modeling, providing another means to potentially interpret the latent dimensions in the model. The connection between the softmax and the inverse clr transform suggests that the inputs to this transform can be represented in clr coordinates. The softmax function and its corresponding inverse, the clr transform, is given as follows

$$softmax(x) = \left[\frac{e^{x_1}}{\sum_i e^{x_i}}, \ldots, \frac{e^{x_1}}{\sum_i e^{x_i}}\right]$$

$$clr(z) = \left[\log \frac{z_1}{g(z)}, \ldots, \log \frac{z_D}{g(z)}\right]$$

Since biases are incorporated into the mmvec model, by construction $Q = UV^T$ is both row centered and column centered, meaning that the sum of rows are zero and the sum of the columns are zero. Given this the following holds

Theorem: If $\boldsymbol{Q} = \boldsymbol{UV}$ and $\mathbf{1_N Q} = \mathbf{0}$ and $\boldsymbol{Q}\mathbf{1_M} = \mathbf{0}$ then $\boldsymbol{U}\mathbf{1_p} = \mathbf{0}$ and $\boldsymbol{V}\mathbf{1_p} = \mathbf{0}$

Suppose that there exists another solution $\boldsymbol{Q} = \boldsymbol{UV^{*T}}$ where $\boldsymbol{V} = \boldsymbol{V} - \mathbf{1_M}\boldsymbol{\lambda_v^T}$ and $\boldsymbol{\lambda_v} \in \mathbb{R}^p$. Then

$$\boldsymbol{Q} = \boldsymbol{U}(\boldsymbol{V} - \mathbf{1_M}\boldsymbol{\lambda_v^T}).$$

Given that the rows of $\boldsymbol{Q}$ sum to 0, then

$$\boldsymbol{U}(\boldsymbol{V} - \mathbf{1_M}\boldsymbol{\lambda_v^T})^T\mathbf{1_M} = 0$$

$$\boldsymbol{U}\boldsymbol{\lambda_v}M = 0.$$

This means that only the trivial solution $\boldsymbol{\lambda_v} = 0$ exists, therefore the rows of $\boldsymbol{V}$ do sum to 0.

Using the same reasoning above, suppose that there exists another solution $\boldsymbol{Q} = \boldsymbol{U^*V^T}$ where $\boldsymbol{U}^* = \boldsymbol{U} - \mathbf{1_N}\boldsymbol{\lambda_u}^T$ and $\boldsymbol{\lambda_u} \in \mathbb{R}^p$. Then

$$\boldsymbol{Q} = (\boldsymbol{U} - \mathbf{1_N}\boldsymbol{\lambda_u}^T)\boldsymbol{V}^T.$$

Given that the columns of $\boldsymbol{Q}$ sum to 0, then

$$\mathbf{1_N}^T(\boldsymbol{U} - \mathbf{1_N}\boldsymbol{\lambda_u}^T)\boldsymbol{V}^T = 0$$

$$N\boldsymbol{\lambda_u}^T\boldsymbol{V} = 0.$$

This means that only the trivial solution $\boldsymbol{\lambda_u} = 0$ exists, therefore the rows of $\boldsymbol{U}$ do sum to 0.

Therefore the rows of both $\boldsymbol{U}$ and $\boldsymbol{V}$ must sum to zero if $\boldsymbol{U}$ and $\boldsymbol{V}$ are non-trivial.

As noted in previous compositional data analysis work, the sum of the components within a vector in clr coordinates is zero. Given that the row vectors within U and V both sum to zero, that suggests that each of these vectors are also in clr coordinates. This means the following properties are satisfied

Topic proportions

Since the U and V row vectors are in clr coordinates, that implies that these row vectors can be directly converted to p-dimensional proportions, yielding a similar interpretation to topics used in models such as LDA [50, 51].

Linearity

Vectors in clr coordinates are known to satisfy linearity, namely

$$clr(\alpha x + y) = \alpha clr(x) + clr(y)$$

for $\alpha \in \mathbb{R}$, $x \in S^p$ and $y \in S^p$. This linearity property was leveraged in word2vec models to perform analogy reasoning. Since both microbes and metabolites are in clr coordinates, it should be possible to categorize microbe-microbe and metabolite-metabolite interactions.

Isometry

The clr transform is distance preserving, meaning that the Aitchison distance on proportions is equivalent to the Euclidean distance on clr vectors. This provides motivation for using Euclidean distances to compute microbe-microbe and metabolite-metabolite similarities.

C. Visualization through biplots

Visualization techniques from compositional data analysis can aid with interpretation [52, 53]. $\boldsymbol{U}$ and $\boldsymbol{V}$ can be visualized as factors within a biplot to visualize the microbe-metabolite embeddings on a single plot. The first two latent dimensions of $\boldsymbol{U}$ represent

microbial coordinates on a 2D scatter plot and the first two latent dimensions of $\boldsymbol{V}$ represent metabolite coordinates on a 2D scatter plot. Typically the coordinate from the $\boldsymbol{V}$ matrix are plotted as arrows from the origin in order to identify features that explain the variance in $\boldsymbol{U}$. However, in our case studies, there are typically many more metabolites than microbes - so we opt to visualize the metabolites as points and microbes as arrows for a simpler visualization

As suggested by the above theorem, the distance between points approximates the Aitchison distance between metabolites, and the distance between arrow tips approximates the Aitchison distance between microbes. As suggested in [54], the Aitchison distance is also equivalent to the variance of the log ratios, suggesting that microbe-microbe and metabolite-metabolite distances could also be interpreted as a measure of proportionality [21]

D. Benchmarks

The simulated data was based on a cystic fibrosis biofilm model derived in Quinn et al [25] shown in Figure S12 in the paper. The biofilm model was built to explain how fermenters and *P. aeruginosa* responded to different concentrations of sugars, amino acids, pH, oxgygen and antibiotics across the Winogradsky column. These models solved for differential equations integrating Monad kinetics and diffusion processes and was run in Matlab using the code provided at https://github.com/zhangzhongxun/WinCF_model_Code

From this simulation, we only focus 2 microbes and 5 compounds. The two microbes are P. aeroginosa (θ_p) and fermenters (θ_f). The five compounds (SG), acids (F), ammonium (P), amino acids (SA) and inhibition molecules (I). In order to simulate a high dimensional dataset, each microbial taxon was split into 50 different subtaxa and each compound was split into 50 molecular subclasses. The partitioning procedure is given as follows

$$\begin{aligned} \boldsymbol{p_i} &\sim \mathcal{N}(\boldsymbol{0}, \sigma_o \boldsymbol{I}) & \boldsymbol{q_i} &\sim \mathcal{N}(\boldsymbol{0}, \sigma_c \boldsymbol{I}) \\ \boldsymbol{o_{ij}} &= \kappa_{ij} ilr^{-1}(\boldsymbol{p_i}) & \boldsymbol{c_{ik}} &= \eta_{ik} ilr^{-1}(\boldsymbol{q_i}) \ , \end{aligned}$$

where $\boldsymbol{p_i}$ is a vector proportions representing how the subtaxa corresponding to j will be distributed in sample i. κ_{ij} represents the absolute abundance of taxon j in sample i. $\boldsymbol{o_{ij}}$ represents a vector of the absolute abundances for all of the subtaxa corresponding to taxon

j. These are the absolute abundances that are used for comparison in Figure 2.

Here we use the ilr^{-1} transform to generate proportions from a multivariate normal distribution. Here the multivariate normal distribution is centered around zero, and the covariance matrix $\sigma_o \boldsymbol{I}$ has only a constant diagonal structure with a tunable parameter σ_o specifying the variability of the partitioning procedure. Larger values of σ_o will cause the allocations of the microbes to be increasingly uneven.

The partitioning procedure is identical for the metabolites. $\boldsymbol{q_i}$ is a vector proportions representing how the subcompounds corresponding to k will be distributed in sample i. η_{ik} represents the absolute abundance of compound k in sample i. $\boldsymbol{c_{ik}}$ represents a vector of the absolute abundances for all of the subtaxa corresponding to compound k. The multivariate normal distribution used to generate the proportions is centered around zero. The covariance matrix $\sigma_c \boldsymbol{I}$ has only a constant diagonal structure with a tunable parameter σ_c specifying the variability of the partitioning procedure. Larger values of σ_c will cause the allocations of the metabolites to be increasingly uneven.

Once the subtaxa and subcompounds absolute abundances have been simulated, the microbial relative counts and metabolite abundances are simulated. The sampling procedure is performed as follows

$$\begin{aligned} \zeta_i &\sim \mathcal{LN}(n, \tau_o) & \omega_i &\sim \mathcal{LN}(m, \tau_c) \\ x_i &\sim \mathcal{PLN}(\zeta_i \, C(\boldsymbol{o_i}), \epsilon_o) & y_i &\sim \mathcal{LN}(\omega_i \, C(\boldsymbol{c_i}), \epsilon_c) \; . \end{aligned}$$

The total sequencing depths and total intensities for sample i are draw from Lognormal distributions with means parameterized by n and m and overdispersion parameters τ_o and τ_c. We chose to use the lognormal distribution for three reasons. First, the lognormal distribution models overdispersion. Second, the lognormal distribution has a simpler interpretation than other overdispersed distributions such as the negative binomial, since the parameters can be directly interpreted as a normal distribution and consequentially has a compositional interpretation due to its connection to the ilr transform. Finally, the lognormal distribution commonly used for modeling in the the ecological literature in the context of studying species populations in Niche theory and Neutral theory, leading to a natural biological interpretation.

Once the total sequencing depth and the total intensities are sampled, the microbial

sequencing counts and metabolite abundances are then sampled. A Poisson lognormal distribution is used to generate the microbial counts from the microbial proportions $C(\boldsymbol{o_i})$ scaled by the sequencing depth ζ_i. The counts are sampled with error ϵ_o. A Lognormal distribution is used to generate the metabolite abundances from metabolite proportions $C(\boldsymbol{c_i})$ scaled by the total intensity ω_i. The abundances are sampled with error ϵ_c. All of the code used to generate the benchmarks can be found at https://github.com/knightlab-analyses/multiomic-cooccurrences

E. Data Analysis

Due to the overwhelming sparsity in microbiome datasets, some filtering is required in order to infer microbe-metabolite interactions. We chose to filter out microbes that appear in less than 10 samples, since these microbes don't have enough information to infer which metabolites are co-occurring with them. In other words the mmvec model has too many degrees of freedom to perform inference on these microbes. For the cystic fibrosis study, there were 172 samples and after filtering there were 138 unique microbial taxa and 462 metabolite features. For the biocrust soils study, there were 19 samples and after filtering there were 466 unique microbial taxa and 85 metabolite features. For the murine high fat diet study, there were 434 samples and after filtering there were 902 microbes and 11978 metabolites. For the IBD dataset, there were 13920 features in the c18 LCMS dataset, 26966 features in the c8 LCMS dataset and 562 taxa. Cross validation was performed across all studies to evaluate overfitting. In the desert biocrust soils experiment, 1 sample out of 19 samples was randomly chosen to be left out for cross-validation. In all of the other studies, 10 samples were randomly chosen to be left out for cross-validation. All of the analyses can be found under https://github.com/knightlab-analyses/multiomic-cooccurences.

F. Data availability

The cystic fibrosis sequencing and metadata data can be found under http://qiita.microbio.me; study id: 10863. The corresponding GNPS analysis can be accessed at http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=34d825dbf4e9466e81d809faf814995b.

The biocrust soils data was retrieved from the supplemental section in Swenson et al [28]. The High fat diet murine model case study 16S rRNA data can be found under http://qiita.microbio.me; study id: 10856. The High fat diet murine model case study are publicly available at https://massive.ucsd.edu/ at MassIVE ID MSV000080918. The GNPS analysis for this study can be accessed at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=977d85bba47b4e96bf69872b961b8edd The IBD data used can be found under https://ibdmdb.org/

G. Software availability

The software implementing the mmvec algorithm can be found under https://github.com/biocore/mmvec Differential abundance analyses in the high fat diet study was performed using L2-regularized multinomial regression using software available at https://github.com/biocore/songbird The software used to build the multiomics network can be found at https://github.com/mortonjt/multiomics_network

X. SUPPLEMENTAL MATERIALS

A. Challenges of analyzing multiple compositional datasets

One of the challenges involved with inferring microbe-metabolite interactions is resolving the differences between the absolute abundances in the original environment, and the measured relative abundances from sequencing and mass spectrometry. In order to guarantee consistency between absolute and relative abundances, scale invariance must be maintained [23], otherwise overwhelming false positive or false negative rates can occur (Figure S1).

As shown in Figure 2d, most tools are not scale invariant. The reason for the contradiction is further clarified in Figure S2, from the proportions, it looks like most of the microbes are decreasing when in fact they are merely increasing with a slower growth rate compared to the fastest growing microbe. The inability to determine which microbes are actually increasing or decreasing caused Pearson and Spearman to misannotate the vast majority of

the interactions, with the except of the interactions of the fastest growing microbe in this scenario.

It may appear the benchmark in Figure 2d that the proportionality metric rho is scale invariant in the context of multiomics analysis. However, another benchmark in Figure S2 reveals that rho is not scale invariant. The reason why scale-invariance breaks for phi, rho and SparCC is because the microbe and metabolite datasets have differing absolute sums. When analyzing a single dataset all three of these metrics rely on the following quantity to hold

$$V\left(\log \frac{x_i}{x_j}\right) = V\left(\log \frac{Np_{x_i}}{Np_{x_j}}\right) = V\left(\log \frac{p_{x_i}}{p_{x_j}}\right),$$

where x_i correspond to random variable quantifying absolute abundances of microbe i, N corresponds to a random variable quantifying the total population size of the microbes, and p_{x_i} correspond to the proportions of microbe i. Due to the log-ratio, the dependence on the total population size N drops out, negating the need to quantify total microbial load. This is critical for microbiome sequencing applications, since quantifying total microbial load can be challenging [14, 23]. Furthermore, methods that satisfy scale invariance have shown shown to be superior to other tools in the context of co-occurrence analysis [55].

However, scale invariance is much harder to enforce when analyzing co-occurrence relationships across multiple datasets. When evaluating the variance of the log ratios across multiple datasets, the scale-invariance relationship is not immediately satisfied

$$V\left(\log \frac{x_i}{y_j}\right) = V\left(\log \frac{Np_{x_i}}{Mp_{y_j}}\right) \neq V\left(\log \frac{p_{x_i}}{p_{y_j}}\right).$$

Here, y_j refers to the absolute abundances of the metabolite j, M refers to the total number of metabolites in the original environment and p_{y_j} is the proportion of metabolite j.

This was recognized in Tipton et al [22] and additional modifications were added to SPIEC-EASI. These modifications explain the superior performance of SPIEC-EASI in the benchmarks. However, there are two major impediments to the application of SPIEC-EASI, namely zeros and StARs [56] regularization. SPIEC-EASI still relies on using pseudocounts, adding bias into the resulting inference. Furthermore, STaRs has been shown to enhance the interpretation of the SPIEC-EASI results, but STaRs is not a scale-invariant procedure. Due to this alone, the absolute and relative estimates will not match as shown in Figure 2

and Figure S2. This may not be a problem when analyzing multiomics datasets with similar scales, such as 16S and ITS sequencing datasets. However, these problems will become exacerbated when analyzing datasets with drastically different scales. Sequencing counts are usually below 100k reads per samples, where as MS intensities are up to 10e9 intensity units.

In light of the challenges discussed above, there are some scenarios where standard statistical methods will be consistent with the biological reality. As discussed in [23] the differences between absolute and relative abundances is essentially a constant factor attributed to the changes in the total biomass. If the total biomass is constant, then traditional statistical methods will work fine. In the case of the cystic fibrosis dataset [25], microbial communities were grown in fixed size Winogradsky columns. As a result, the total size of the community could be constrained due to the limited resources and space. This could explain the consistency between Pearson and mmvec in this particular study (Figure S3). On the other hand, in the biocrust soils study, the drastic differences between Spearman and mmvec could be explained by the rapid increase in biomass following the wetting event.

B. Software workflows

To facilitate utilization of the mmvec tool, we have developed two different user facing interfaces. First, we have developed a qiime2 plugin [57], where mmvec can be run using a simple command line interface. This interface is complemented using [24], where users can monitor convergence rates for their models in real-time and evaluate how different parameters will affect their model fit (Figure S4). Second, we have integrated mmvec into the Global Natural Product Social Molecular Networking (GNPS) platform that can be accessed by the public. The online interface through GNPS resolves several usability issues. First, GNPS facilitates import of metabolomics data into qiime2 by pre-processing, importing, and sample renaming, This is performed as part of the standard metabolomics analysis at GNPS (e.g. molecular networking and feature-based molecular networking). Second, since it is possible to both download and re-use outputs of workflows run at GNPS directly, it is straightforward to select the GNPS qza and molecule annotations needed for mmvec. The user will need to upload the accompanying feature and taxonomy data for qiime2 and the analysis will be begin. Once the workflow completes, the biplots can be viewed directly in

the browser and other outputs (e.g. ranks) are available for download (Figure S5).

The mmvec implementation is written using Tensorflow and can leverage GPUs for computation. The number of gradient descent iterations is specified by the user and model fit diagnostics can be monitored in real time using Tensorboard. The runtime of mmvec across 16 cores can take multiple days until a model convergence reaches convergence. With GPUs, the running time is reduced to a few hours. Using a Telsa GPU, the model can reach convergence within 4 hours on the IBD dataset comprised of 562 microbial taxa, 26,966 metabolite features and 400 samples. However, there is a trade-off of accuracy and running time. More accurate models require smaller learning rates and may take longer to run.

XI. FIGURE LEGENDS

Figure legends are below

Figure 1: Input data types and mmvec neural network architecture. (a) The neural network architecture where the input layer represents one-hot encodings of N microbes and the output layer represents the proportions of M metabolites. U corresponds to microbial vectors and V corresponds to metabolite vectors. (b) The pipeline for training mmvec. The objective behind mmvec is to predict metabolite abundances (y) given a single input microbe sequence (x), also known as a one-hot encoding. This training procedure will estimate conditional probabilities of observing a metabolite given the input microbe sequence. Cross-validation can be performed on hold-out samples to access overfitting.

Figure 2: Simulation benchmarks. (a) Absolute abundances of microbes and metabolites simulated from differential equations derived in [25] for a specific spatial point. (b) Proportions of the abundances shown in (a). (c) F1 score, precision and recall curves comparing mmvec to Pearson, Spearman, SparCC, SPIEC-EASI, and proportionality metrics phi and rho across the top 100 metabolites for each microbe. (d) comparisons of coefficients learned from absolute abundances and relative abundances all of the benchmarked methods.

Figure 3: *M. vaginatus* released metabolites after the biocrust wetting event. (a) Comparison of *M. vaginatus* metabolite interactions estimated from Spearman and mmvec from (n=19 samples). All of the experimentally validated *M. vaginatus* released metabolites are labeled. All metabolites with contradicting findings between the wetting experiment and the *in vitro* experimental results are highlighted in red. Points are resized according to the 10 log(p-value) obtained from Spearman correlation. Dashlines mark the cutoff for a Spearman correlation of zero, and the conditional log probabilities of zero. Here a zero log conditional probability represents the conditional probability of the average metabolite because all probabilities here are mean centered. (b) Benchmarks comparing the detection rate of the experimentally validated molecules across different statistical methodologies. (c) *M. vaginatus* proportions and (d) 4-guanidinobutanoate proportions following a wetting event.

Figure 4: Investigation of *P.aeruginosa*-associated molecules. (a) Biplot drawn from the mmvec conditional probabilities estimated for the cystic fibrosis dataset [25]. Arrows represent microbes and dots represent metabolites. The x and y axes represent principal components from the SVD of the microbe-metabolite conditional probabilities estimated from mmvec (n=138 samples). Distances between points quantify co-occurrence strength between metabolites, with small distances indicating metabolites that have a high probability of co-occurring with high probability. Distances between arrow tips quantify co-occurrence strength between microbes. The directionality of the arrows can be used to pinpoint which microbes can explain the metabolite co-occurrence patterns. Arrows highlighted in green correspond to putative cystic fibrosis pathogens and yellow arrows highlight known anaerobes. Only known molecules produced by *P. aeruginosa* are labeled. (b) Scatter plot of molecules with respect to the oxygen gradient differential and the first principal component learned from mmvec (n=442 molecules) with linear regression model and 95% confidence interval for regression estimate. (c) The first principal component vs the number of samples where the taxa was the most abundant taxa in that sample . (d) Heatmap of *P. aeruginosa* and *Streptococcus* abundances in samples where they are the most abundant species. (e) Heatmap of the top 100 molecules that co-occur with *P. aeruginosa* and *Streptococcus*.

Figure 5: Microbe/metabolite co-occurrences across study of HCC progression in the context of innate immunity in a mouse model [26]. (a) Visualization of microbial co-occurrence patterns, where distances between points approximates the Aitchison distance between microbes, which quantities microbial occurrences. Small distances are indicative of microbes with high probability of co-occurring together. Microbes are colored according to their association with HFD, which was estimated using differential abundance analysis via multinomial regression. (b) Emperor [58] biplot of microbe-metabolite interactions, with metabolites colored according to their association with HFD. HFD association was estimated through differential abundance analysis via multinomial regression. Distances between points approximate Aitchison distances between metabolites and distances between arrow tips approximate Aitchison distances between microbes. Several *Clostridium spp.* appear to co-occur with the new bile acid molecule cholate phenylalanine amidate, also referred to as Phe conjugated cholic acid.

Figure 6: Microbe-metabolite interactions of the human microbiome in association with IBD samples [27]. (a) Heatmap visualization of the inferred conditional probabilities for various bile acids given the presence of *Klebsiella*, *Roseburia* and *Clostridium bolteae*. (b) Heatmap visualization of the inferred conditional probabilities for the carnitines given the presence of *Klebsiella*, *Roseburia*, and *Clostridium bolteae*. (c) Multiomics biplot of the microbe-metabolite interactions learned from metagenomics profiles and C18 negative ion mode LC-MS. Microbes (arrows) and metabolites (spheres) are colored according to their differentials estimated from multinomial regression. *Klebsiella spp.* appears to be strongly associated with IBD, while *Propionibacterium spp.* has strong negative association. (d) Network of the top 300 edges where only the edges that contain *Klebsiella* and *Propionibacteriaceae* are visualized.

Figure S1: Description of the compositionality issue. (a) An illustration of how false negatives can occur - in the absolute abundance data, there is a strong Pearson correlation between the microbes and the metabolites. These correlations disappear when considering the corresponding proportions. (b) An illustration of how false positives can occur - in the absolute abundance data, there is no correlation between the dark green molecule and the dark blue microbe. However, the proportions of the same dataset show that there is a very strong correlation between the dark blue and the dark green molecule.

Figure S2: Illustration of how excessive misannotation rates can occur. (a) Absolute abundances and relative abundances of microbes/metabolites observed in an environment over time, with each microbe/metabolite colored according to its rate of increase / decrease. (b) A scale-invariance comparison of statistical methodologies. Points are colored by the corresponding microbes in the interactions; triangle markers represent increasing metabolites and decreasing metabolites. Mmvec is the only method that remains consistent between the absolute and relative abundances.

Figure S3: Comparison of Pearson and mmvec on Cystic Fibrosis study. (a) Estimates of *P. aeruginosa* associated molecules between Pearson and the conditional probabilities calculated from the mmvec applied to the cystic fibrosis study dataset. The annotations correspond to level 2 or 3 of the metabolomics standards initiative [59] and may correspond to different isomeric species (n=462 molecules). (b) Ranks of Pearson coefficients and conditional probabilities from the mmvec for the Rhamnolipids (n=462 molecules). (c) Pyochelin proportions vs *P. aeruginosa* proportions.

Figure S4 : Negative log likelihood and prediction accuracy of mmvec. Tensorboard visualization of training error and cross-validation error of mmvec on the IBG dataset. Five different runs with differing initialization conditions are shown.

936 Figure S5: GNPS [32] job output. An example of job on the GNPS website with the job description and the downloadable output files from mmvec.

937