ARTICLES
https://doi.org/10.1038/s41592-019-0616-3
nature methods

Learning 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[10,11], Mingxun Wang[3,4], Nicholas A. Bokulich[12,13], Aaron Watters[14], Se Jin Song[1,11], Richard Bonneau[7,14,15,16], Pieter C. Dorrestein[3,4] and Rob Knight[1,2,11,17]*

Integrating multiomics datasets is critical for microbiome research; however, inferring interactions across omics datasets has multiple statistical challenges. We solve this problem by using neural networks (https://github.com/biocore/mmvec) to estimate the conditional probability that each molecule is present given the presence of a specific microorganism. We show with known environmental (desert soil biocrust 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.

Knowledge gained by integrating complementary omics data will lead to improved detection of microbial products and optimized culturing conditions for uncharacterized microorganisms[1]. Previous work has been able to predict metabolite abundance profiles from microbe abundance profiles[2,3]. 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 multiomics approaches, several conceptual challenges limit techniques that integrate disparate omics data in general, for example, linking microbial sequencing and untargeted mass spectrometry. Therefore, new approaches are needed to handle disparate data types[4]. Relative abundances of thousands of microbes and metabolites can be measured using sequencing technology and mass spectrometry, respectively, resulting in the generation of high-dimensional microbiome and metabolomics datasets. Quantifying microbe–metabolite interactions from these abundances 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 low-dimensional representations[5–7]. Network models have been shown to improve classification accuracy using multiple datasets[8]. Factor models have been proposed to incorporate multiple datasets for biomarker analysis[9]. Despite the wide application of these methods, they are notoriously difficult to interpret[10–12] and it remains unclear whether these models can obtain individual microbe–metabolite interactions.

Pearson's and Spearman's 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 the simplifications in these methods are not statistically valid for compositional data, a fact first recognized by Pearson in 1895 and followed up in numerous studies[13–17]. This problem is further complicated because both microbiome[17] and mass spectrometry[18–21] 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 depend on the extraction protocol and analytic methods used, which leads to a partial snapshot of the metabolome. Moreover, measuring the total mass of molecules extracted is often not performed in large-scale metabolomics efforts, owing to the highly laborious nature of that step.

To understand how issues associated with compositional data impact inference on microbe–metabolite interactions, we illustrate one example in Supplementary Fig. 1. These issues alone can give rise to overwhelming false positives and false negatives, making Pearson's and Spearman's in some scenarios comparable to random coin flips. Experimental validations currently take large laboratories multiple years to perform[22], often requiring time-consuming manual examinations of erroneous correlations.

[1]Department of Pediatrics, University of California, San Diego, La Jolla, CA, USA. [2]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA. [3]Collaborative Mass Spectrometry Innovaftion 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, NY, USA. [8]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [9]DOE Joint Genome Institute, Walnut Creek, CA, USA. [10]Jacobs School of Engineering, University of California, San Diego, La Jolla, CA, USA. [11]Center for Microbiome Innovation, University of California, San Diego, La Jolla, CA, USA. [12]The Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, USA. [13]Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ, USA. [14]Flatiron Institute, Simons Foundation, New York, NY, USA. [15]Computer Science Department, Courant Institute, New York, NY, USA. [16]Center For Data Science, New York University, New York, NY, USA. [17]Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA. *e-mail: rknight@ucsd.edu

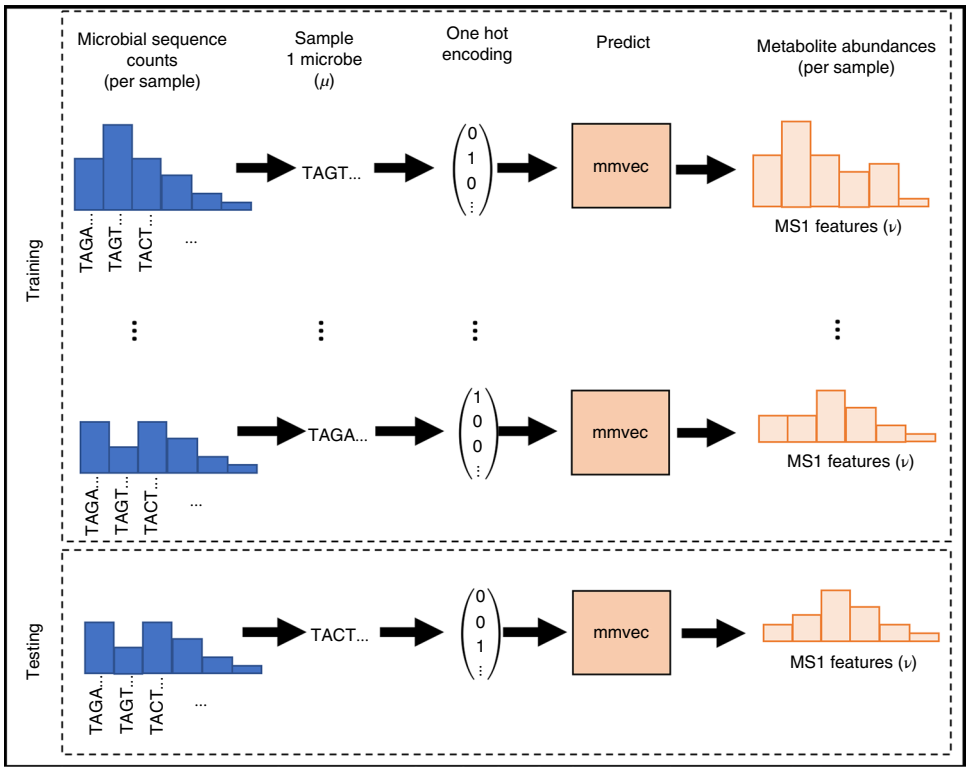

Fig. 1 | Input data types and mmvec neural network architecture. The pipeline for training mmvec. The objective behind mmvec is to predict metabolite abundances (y) given a single input microbe sequence (x), also known as a one-hot encoding. This training procedure will estimate conditional probabilities of observing a metabolite given the input microbe sequence. Cross validation can be performed on hold-out samples to assess overfitting.

There are other compositional techniques such as SparCC[13] and proportionality[23] 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[24], which provided motivation to extend SPIEC-EASI[14] to handle multiomics datasets. We show that this approach does not work for microbe–metabolite interactions because of differences in measurement units between sequencing and mass spectrometry measurements (Supplementary Note). 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), a neural network that predicts an entire metabolite abundance profile from a single microbe sequence (Fig. 1). Through iterative training, mmvec can learn the co-occurrence probabilities between microbes and metabolites. The microbe–metabolite interactions can be ranked[25] and visualized through standard dimensionality reduction interfaces, enabling interpretable findings. The computations behind mmvec take advantage of modern graphics processing unit (GPU) architectures using Tensorflow[26], enabling scalable inference on large multiomics datasets. Furthermore, we provide evidence in two benchmarks and four case studies that mmvec outperforms existing statistical methods.

Results

We performed benchmarks comparing mmvec to Pearson's, Spearman's, SPIEC-EASI[14], SparCC[13] and proportionality[23] using datasets from a simulated cystic fibrosis biofilm. We then show that mmvec can resolve contradictory cyanobacteria–metabolite relationships in a study of desert soil biocrust wetting. We also demonstrate recovery of known associations of metabolites produced by *Pseudomonas aeruginosa* that are observed in cystic fibrosis[27]. Finally, we explore the relationship between microbiota and metabolic changes in mice fed a high-fat diet (HFD)[28] and during inflammatory bowel disease[29], showing how this approach can be used to determine the microbial origin of molecules even in extremely complex real-life biological systems with limited knowledge of existing associations.

Simulation benchmarks. To compare the performance of mmvec to Pearson's, Spearman's, 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[27]. The original study tested the hypothesis that the cystic fibrosis 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 outcompete the pathogen. The simplicity of this system allowed high-level ecological patterns to be modeled. 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[27] (Fig. 2a).

We simulated the measurement process for microbial DNA sequencing and untargeted mass spectrometry for metabolites (Methods), providing ground truth information on their interactions. The model simulates interactions between *P. aeruginosa* and the fermenters, as well as 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.

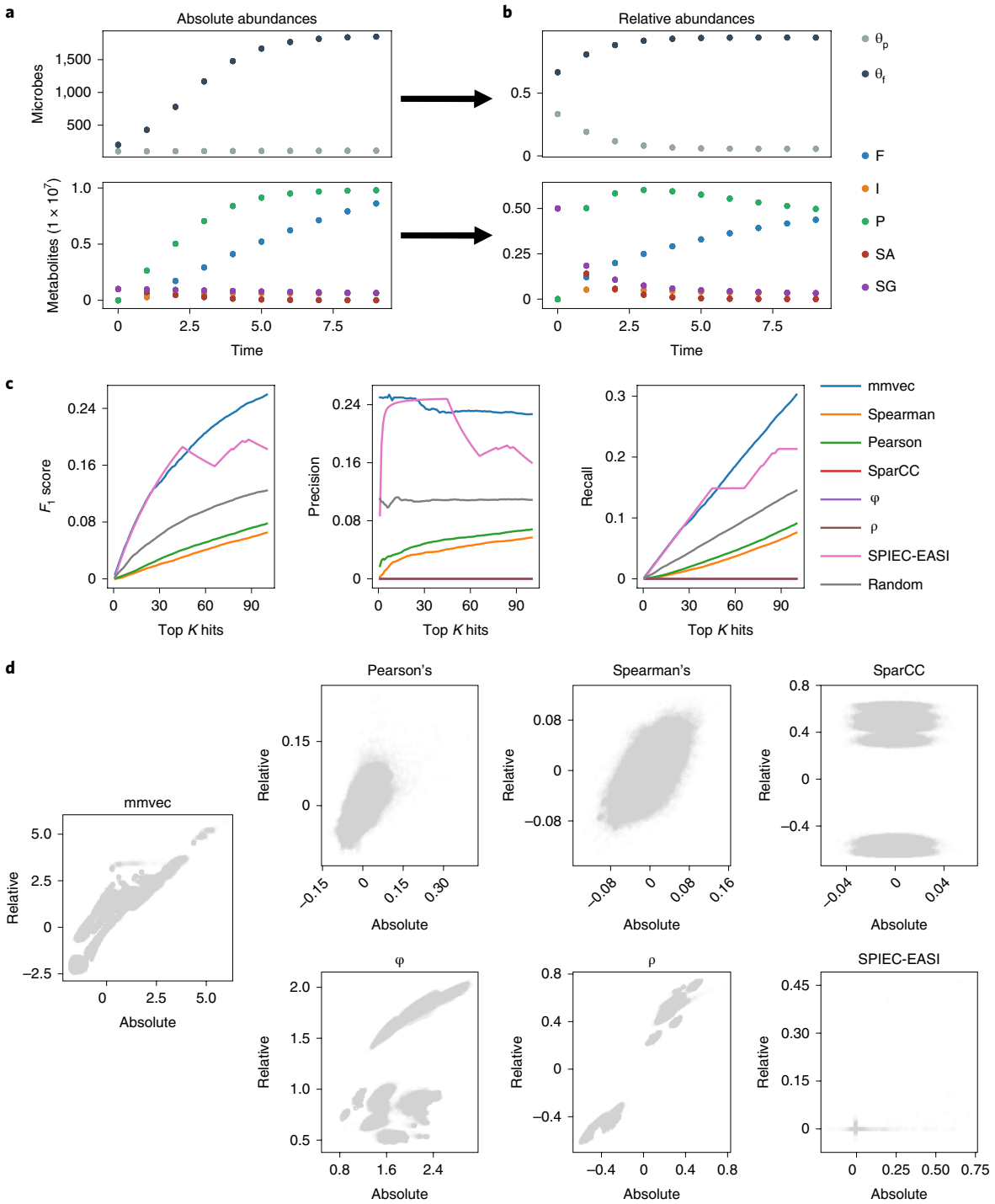

Fig. 2 | Simulation benchmarks. a, Absolute abundances of microbes and metabolites simulated from differential equations derived in ref. [27] for a specific spatial point. Fermenters are denoted by θ_f and *P. aeruginosa* are denoted by θ_p. Here five types of metabolites were simulated, namely sugars (SG), inhibitors (I), acids (F), ammonium (P) and amino acids (SA). **b**, Proportions of the abundances shown in **a**. **c**, F_1 score, precision and recall curves comparing mmvec to Pearson's, Spearman's, SparCC, SPIEC-EASI and proportionality metrics φ and ρ across the top 100 metabolites for each microbe. **d**, Comparisons of coefficients learned from absolute abundances and relative abundances from all of the benchmarked methods.

Therefore, we can test whether the top K metabolites associated with each microbe include the correct microbe–metabolite interactions. Figure 2c shows specificity and sensitivity for each tool as a function of K, in which random chance outperformed all of the tools except for mmvec and SPIEC-EASI, with mmvec performing the best. As shown in Fig. 2d and Supplementary Fig. 2, mmvec is the only method robust to scale deviations. This is critical for maintaining consistency between absolute and relative abundances, which can otherwise lead to inflated false positives and false negatives[16].

Soil biocrust wetting event. Improved data analysis can help to resolve previously inconsistent experimental results, especially in environmental and clinical settings. To test whether mmvec can resolve unexplained discrepancies in microbe–metabolite interactions

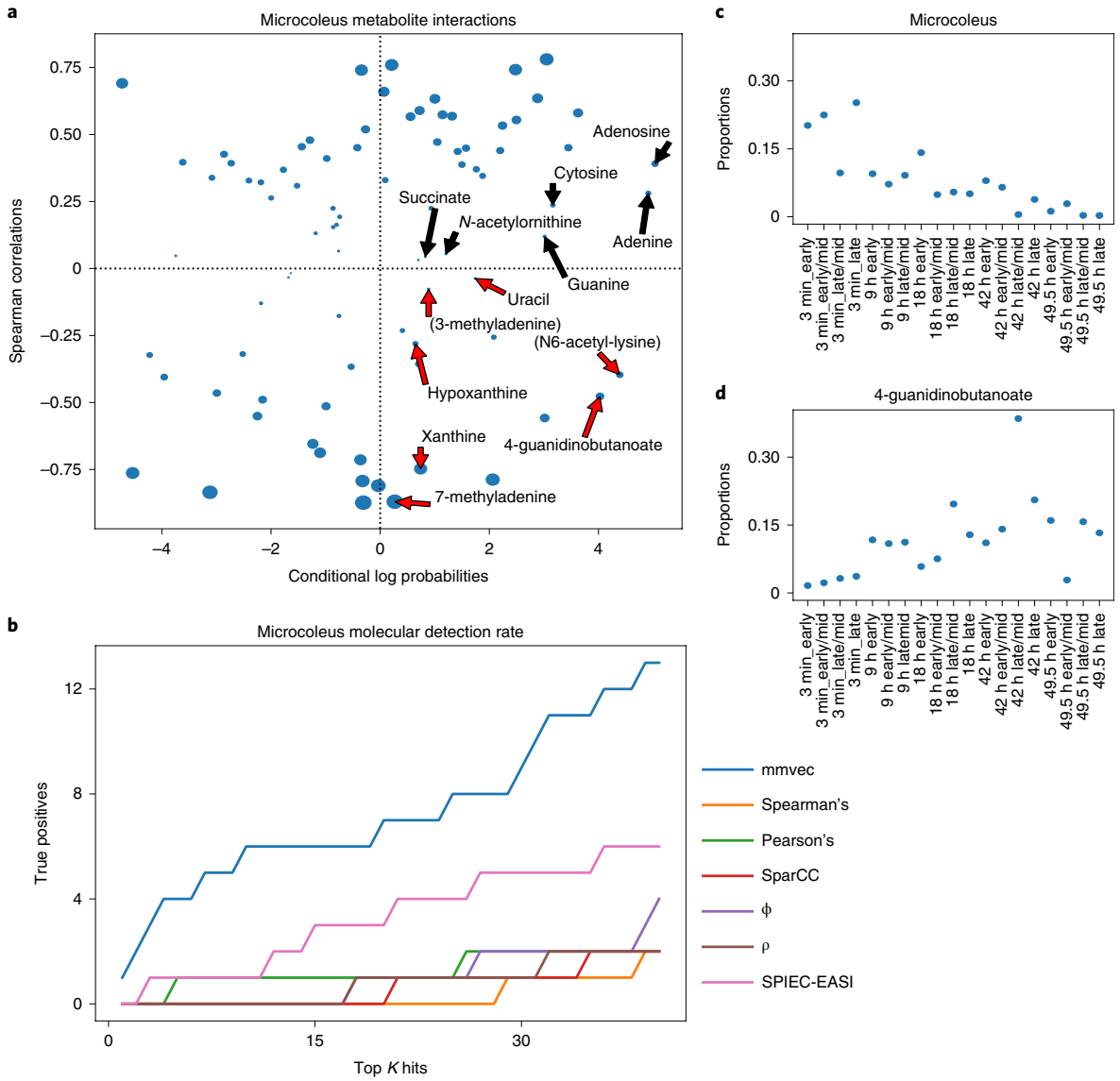

Fig. 3 | Metabolites released by *M. vaginatus* after the biocrust wetting event. a, Comparison of *M. vaginatus*–metabolite interactions estimated from Spearman's and mmvec from ($n = 19$ samples). All of the experimentally validated metabolites released by *M. vaginatus* are labeled. All metabolites with contradictory findings between the wetting experiment and the in vitro experimental results are highlighted in red. Points are resized according to the $-10\log P$ obtained from Spearman's correlation. Dashed lines mark the cutoff for a Spearman's 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 are mean centered. **b**, Benchmark comparisons of the detection rate of the experimentally validated molecules across different statistical methodologies. **c**,**d**, *M. vaginatus* (**c**) and 4-guanidinobutanoate (**d**) proportions after a wetting event.

across studies, we applied it to a study of biocrust wetting[30]. In this work, the authors identified metabolites that were consumed and released by multiple biocrust isolates including *Microcoleus vaginatus* and two *Bacillus* strains[31], and compared these patterns with closely related environmental taxa and metabolites observed in situ[30].

While almost 70% of the examined microbe–metabolite relationships following the wetting event were validated[30], some contradicted microbe–metabolite relationships observed in cultures[31]. These contradictions stemmed from Spearman's correlations between *M. vaginatus* abundances and the observed metabolite abundances, but were resolved by mmvec (Fig. 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 40 co-occurring metabolites with *M. vaginatus* (of 85 molecules total). This result supports the original finding that *M. vaginatus* actually releases these molecules after the wetting event. By contrast, Spearman's labels 7 of 13 of these molecules with a negative correlation (Fig. 3a), indicating that these molecules were consumed by *M. vaginatus* rather than released, as originally stated[30]. When the annotation detection rates differs among statistical methodologies, mmvec has a substantially higher true-positive rate as shown in Fig. 3b.

The conflicting results between mmvec and Spearman's could be explained by the growing microbial biomass and shift in available resources after wetting (Fig. 3c,d). Total biomass is expected to increase because *M. vaginatus* releases metabolites that enable the growth of many other microbes. DNA sequencing can only measure proportions, the growth in other microbes could thus cause the proportions of *M. vaginatus* to decrease, leading to a misleading anticorrelation with 4-guanidinobutanoate (Fig. 3d).

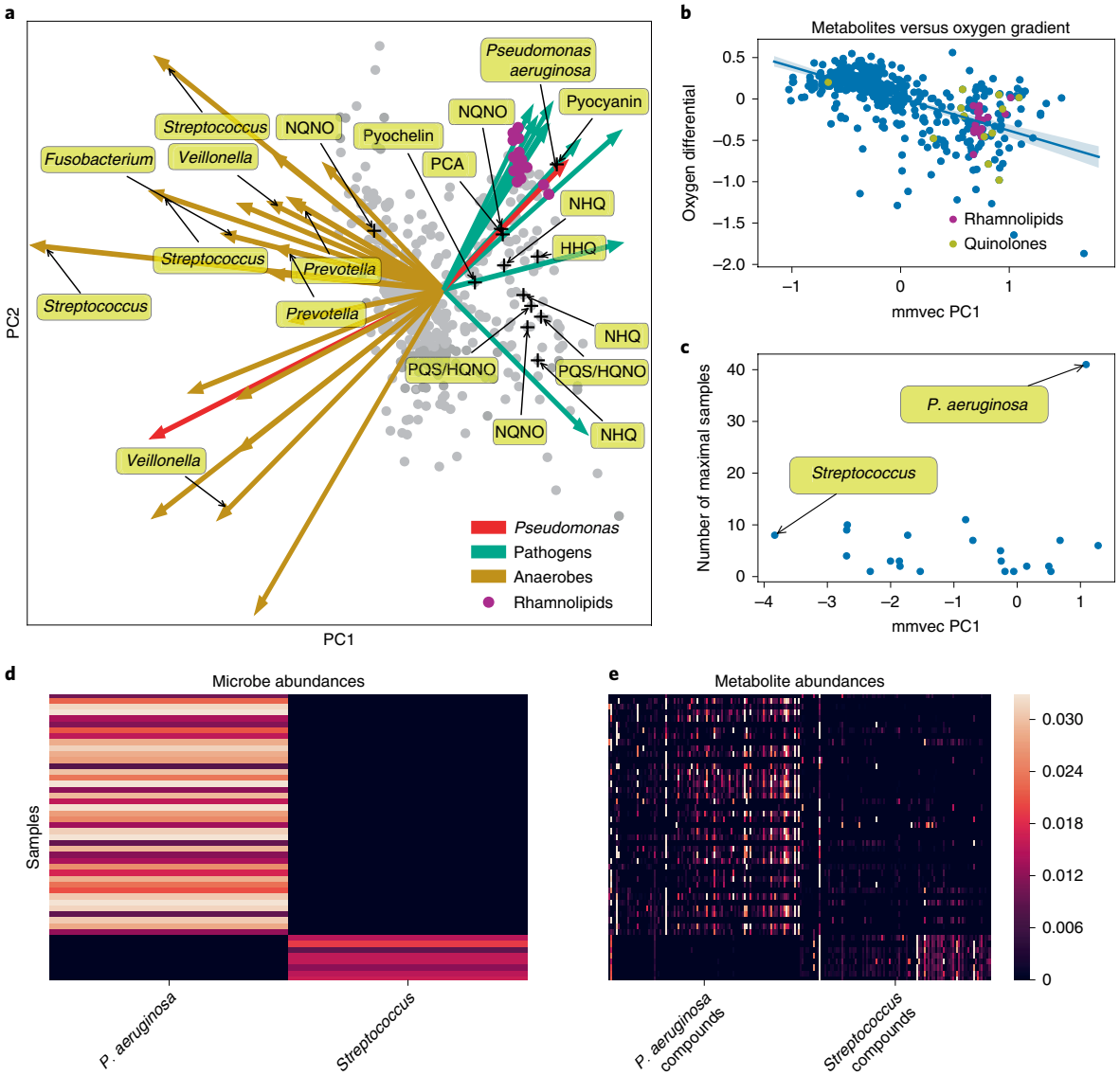

Fig. 4 | Investigation of *P. aeruginosa*-associated molecules. a, Biplot drawn from the mmvec conditional probabilities estimated for the cystic fibrosis dataset[27]. Arrows represent microbes and dots represent metabolites. The *x* and *y* axes represent principal components (PCs) from the singular value decomposition (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. 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 a linear regression model and 95% confidence intervals for the regression estimate. **c**, The first principal component versus the number of samples for which the taxa was the most abundant taxa in that sample. **d**, Heat map of *P. aeruginosa* and *Streptococcus* abundances from samples in which they were the most abundant species. **e**, Heat map of the top 100 molecules that co-occur with *P. aeruginosa* and *Streptococcus*.

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. However, it is not possible to infer whether *M. vaginatus* is decreasing in abundance[25] or 4-guanidinobutanoate is increasing in abundance.

Cystic fibrosis. To further validate whether mmvec can detect known microbe–metabolite interactions, analyzed a study on the lung mucus microbiome of patients with cystic fibrosis[27,32]. 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[27]. mmvec clearly separates anaerobes and pathogens (Fig. 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[33]. In the original study, on the basis of annotations from the Global Natural Product Social Molecular Networking

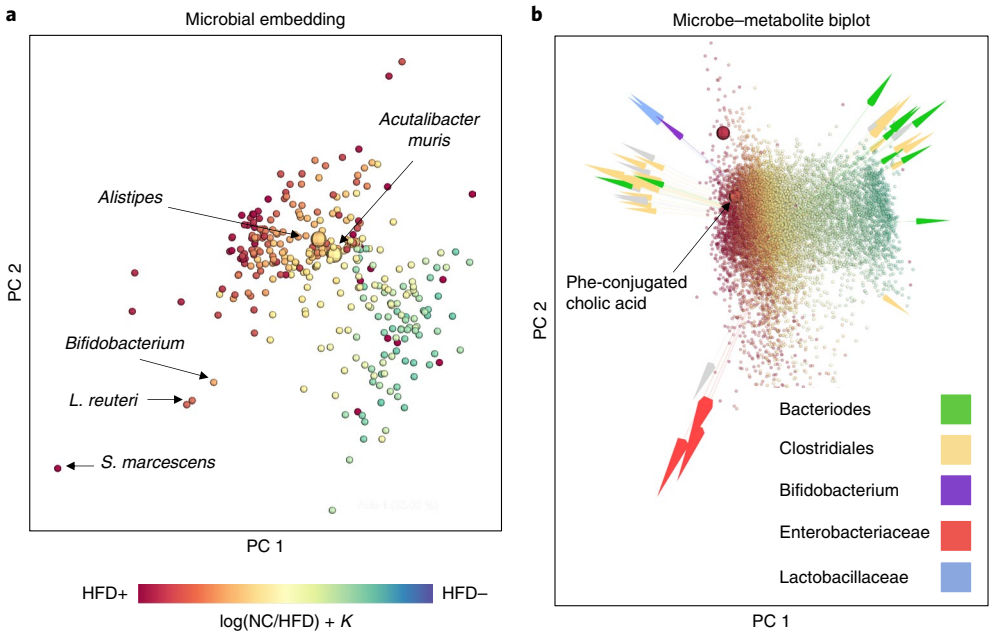

Fig. 5 | Microbe–metabolite co-occurrences across study of hepatocarcinoma progression in the context of innate immunity in a mouse model. **a**, Visualization of microbial co-occurrence patterns, where distances between points approximate the Aitchison distance between microbes, which quantifies microbial occurrences. Small distances are indicative of microbes with a high probability of co-occurring. Microbes are colored according to their association with HFD, which was estimated using differential abundance analysis via multinomial regression. **b**, Emperor[47] 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.

platform (GNPS)[34], the bacterium was found to produce six molecules: 4-hydroxy-2-heptylquinoline, pyocyanin, phenazine-1-carboxylic acid, 2-nonyl-4-hydroxy-quinoline, 2-heptyl-3,4-dihydroxyquinoline (*Pseudomonas* quinolone signal) and pyochelin[27]. As shown in Fig. 4a, mmvec identifies these molecules with a high probability of co-occurrence 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[35,36]. These rhamnolipids were not identified in the original study[27]. The annotations for these compounds have been established using the GNPS[34].

There is a negative correlation between the first principal component learned from mmvec and the log fold change of the metabolites across the oxygen gradient (Fig. 4b) (Pearson's $r=-0.59$, $P=1.8\times10^{-44}$, $n=442$ molecules), 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's ($r=0.11$, $P=0.16$, $n=138$ microbes). 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 (Fig. 4d,e) (log ratio t test$=6.51$, $P=4.4\times10^{-8}$, $n=49$ samples). 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 cystic fibrosis 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 ($r=0.319$, $P=1.18\times10^{-11}$, $n=442$ molecules).

Of 14 quinolone molecules known to be produced by *P. aeruginosa*, Pearson's correlation detected nine with $P<0.05$ without false-discovery rate (FDR) correction, and only five with FDR correction. For example, pyocyanin, does not appear related to be related to *P. aeruginosa* by the raw proportions ($r=0.158$, FDR-corrected $P=0.089$, rank$=96$, $n=172$ samples), but is ranked 34th most associated with *P. aeruginosa* by mmvec (Supplementary Fig. 3c), consistent with culturing experiments that demonstrate that *P. aeruginosa* produces this molecule[37]. Eighteen rhamnolipids are among the top 25 metabolites most associated with *P. aeruginosa* by mmvec, and have higher ranks with mmvec than with Pearson's correlation (Supplementary Fig. 3b).

Effects of a high-fat diet in murine model. 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[38]. These molecules increased in abundance with HFD in humans. We determined that these molecules are microbially made as they were present in specific-pathogen-free mice, 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 HFD in a well-controlled study that investigated the development of non-alcoholic fatty liver disease, cirrhosis and hepatocarcinoma in a mouse model[28]. When reanalyzing these datasets for differential abundances via multinomial regression, the strong association of bile acid with HFD became

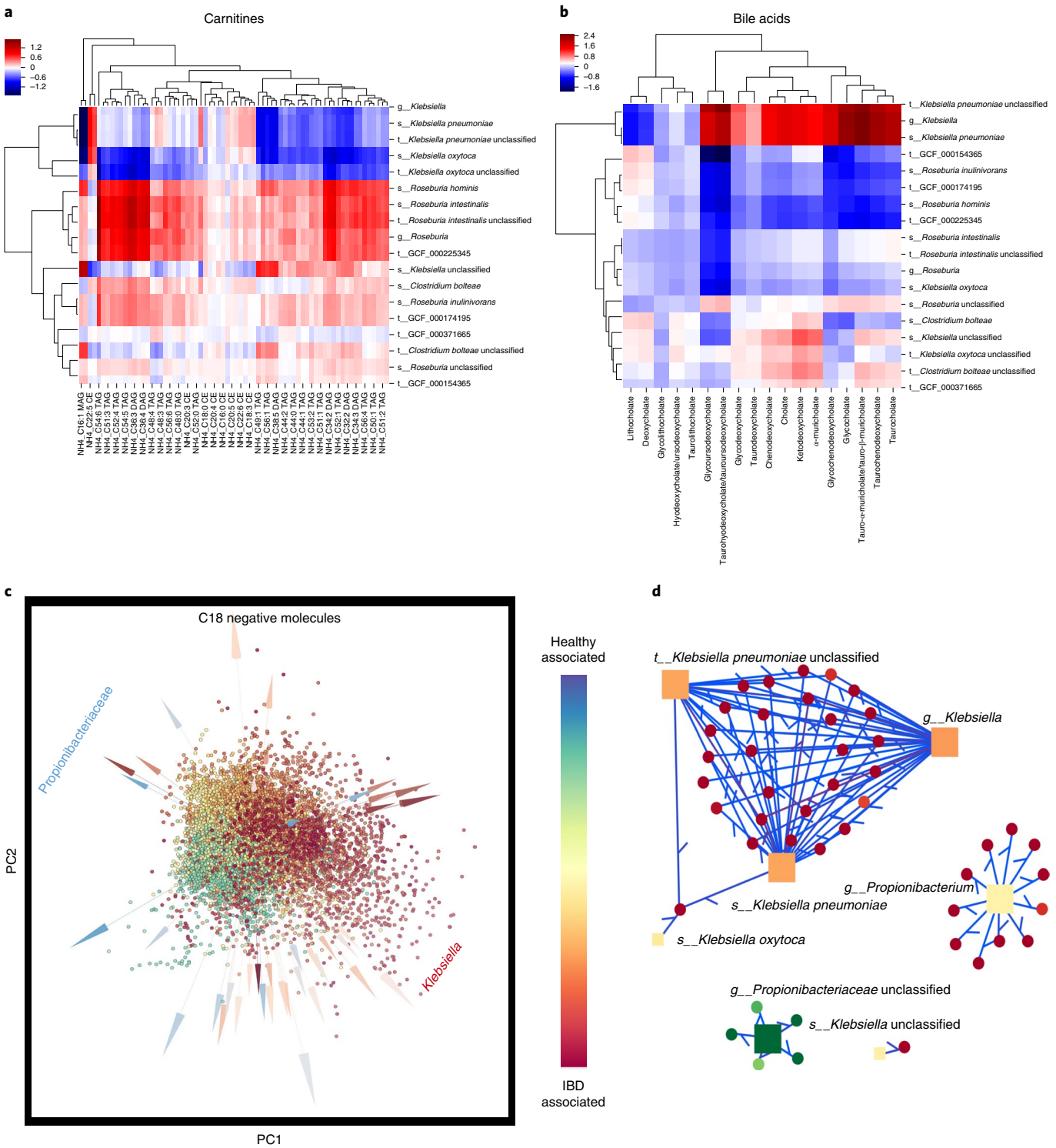

Fig. 6 | Microbe–metabolite interactions of the human microbiome in association with IBD samples. a, Heat map of the inferred conditional probabilities for various bile acids given the presence of *Klebsiella*, *Roseburia* and *Clostridium bolteae*. **b**, Heat map of the inferred conditional probabilities 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*. appear to be strongly associated with IBD, while *Propionibacterium spp*. have a strong negative association. **d**, Network of the top 300 edges where only the edges that contain *Klebsiella* and *Propionibacteriaceae* are visualized.

immediately apparent. The use of mmvec showed distinct groups of microbes that associated with HFD (Fig. 5a) and a clear stratification of the mass spectrometry data according to diet (Fig. 5b). Several *Clostridium spp*. correlated with the cholate phenylalanine conjugate. Indeed, we showed that *Clostridium spp*. were found to produce this bile acid[38]. This result demonstrates the ability of mmvec to streamline the discovery of microbes that produce specific molecules of interest.

Microbe–metabolite interactions in inflammatory bowel disease. Finally, microbe–metabolite interactions were investigated for samples from patients with inflammatory bowel disease (IBD) generated under the integrative Human Microbiome Project[29]. 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 IBD. In particular, levels of carnitines and bile acids were shown to be affected[29]. Using mmvec, we confirmed the core findings in the previous study, such as the co-occurrence between *Roseburia hominis* and multiple carnitines (including the previously noted C20), which have anti-inflammatory properties[29] (Fig. 6a). We also found high correlation between *Klebsiella spp*. and IBD status and that *Klebsiella* co-occurs with high probability with several bile acids (Fig. 6b). Although *Klebsiella* itself does not produce these compounds, some pathogens (including *Klebsiella*) are known to be resistant to bile acids[39]. Excessive production of bile acids and bile acid malabsorption can lead to overabundance of bile acids, which is a hallmark of IBD[40], 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 Fig. 6b. We also noted that three *Klebsiella* species are the top drivers of the IBD-associated molecules (Fig. 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 acid that can germinate *Clostridium difficile* spores or that has antimicrobial properties[41,42].

Therefore, it is possible that in the case of *Clostridia*, the existing co-occurrences (Fig. 6b) are due to actual biosynthesis of the metabolites by the microbial species indicated rather than an 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 Lloyd-Price et al.[29] (Fig. 6c,d). This relationship is corroborated by other published studies[43–46]. 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[43]. Also, in a survey of in vivo versus in vitro bacterial activity, *Probionibacterium freudenreichii* was shown to play an immunomodulatory role in the context of an ulcerative colitis mouse model[44]. In another study it was shown that *Propionibacterium freudenreichii* is a viable core component in an anti-inflammatory probiotic fermented dairy product[45]. 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[46]. Thus, it appears that the existing knowledge supports the statistically inferred interaction uncovered by mmvec, but not identified in the original analysis.

Discussion

In both simulation benchmarks and annotated datasets, mmvec presents improved performance for inferring microbe–metabolite interactions from multiomics datasets. Our results suggest that mmvec outperforms all existing tools that aim to infer interactions between paired microbe–metabolite abundance datasets, both in simulations and experimental data. In the biocrust wetting experiment, mmvec resolved conflicting findings between the in vitro-validated metabolites that are released *M. vaginatus* and the sequencing and mass spectrometry analysis of environmental samples. In the cystic fibrosis study, mmvec can reliably identify all of the experimentally determined molecules of interest produced by *P. aeruginosa*. 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 cannot yet be calculated. Furthermore, theoretical work will be required to handle inputs with continuous values.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information, details of author contributions and competing interests, and statements of code and data availability are available at https://doi.org/10.1038/s41592-019-0616-3.

Received: 4 April 2019; Accepted: 19 September 2019;
Published online: 4 November 2019

References

1. Jansson, J. K. & Baker, E. S. A multi-omic future for microbiome studies. *Nat. Microbiol.* **1**, 645 (2016).
2. Noecker, C. et al. Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. *MSystems* **1**, e00013–e00015 (2016).
3. Mallick, H. et al. Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. *Nat. Commun.* **10**, 3136 (2019).
4. Knight, R. et al. Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* **16**, 410–422 (2018).
5. Meng, C. et al. Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinform.* **17**, 628–641 (2016).
6. Gall, G. Le et al. Metabolomics of fecal extracts detects altered metabolic activity of gut microbiota in ulcerative colitis and irritable bowel syndrome. *J. Proteome Res.* **10**, 4208–4218 (2011).
7. Rohart, F. et al. mixomics: an r package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* **13**, e1005752 (2017).
8. Wang, B. et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11**, 333 (2014).
9. Argelaguet, R. et al. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Sys. Biol.* **14**, e8124 (2018).
10. Ter Braak, C. J. F. & Verdonschot, P. F. M. Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquat. Sci.* **57**, 255–289 (1995).
11. Witten, D. M., Tibshirani, R. & Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–534 (2009).
12. Bodein, A., Chapleur, O., Droit, A. & Lê Cao K. A. A generic multivariate framework for the integration of microbiome longitudinal studies with other data types. Preprint at *bioRxiv* https://doi.org/10.1101/585802 (2019).
13. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, e1002687 (2012).
14. Kurtz, Z. D. et al. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11**, e1004226 (2015).
15. Weiss, S. et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* **10**, 1669–1681 (2016).
16. Vandeputte, D. et al. Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**, 507–511 (2017).
17. . & Gloor, G. B. et al. Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* **8**, 2224 (2017).
18. Tang, K., Page, J. S. & Smith, R. D. Charge competition and the linear dynamic range of detection in electrospray ionization mass spectrometry. *J. Am. Soc. Mass Spectrom.* **15**, 1416–1423 (2004).
19. King, R., Bonfiglio, R., Fernandez-Metzler, C., Miller-Stein, C. & Olah, T. Mechanistic investigation of ionization suppression in electrospray ionization. *J. Am. Soc. Mass Spectrom.* **11**, 942–950 (2000).
20. Matuszewski, B. K., Constanzer, M. L. & Chavez-Eng, C. M. Strategies for the assessment of matrix effect in quantitative bioanalytical methods based on HPLC–MS/MS. *Anal. Chem.* **75**, 3019–3030 (2003).
21. Kalivodová, A. et al. Pls-da for compositional data with application to metabolomics. *J. Chemom.* **29**, 21–28 (2015).
22. Jansson, J. K. & Baker, E. S. A multi-omic future for microbiome studies. *Nat. Microbiol.* **1**, 16049 (2016).

23. Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S. & Bähler, J. Proportionality: a valid alternative to correlation for relative data. *PLoS Comput. Biol.* **11**, e1004075 (2015).
24. Tipton, L. et al. Fungi stabilize connectivity in the lung and skin microbial ecosystems. *Microbiome* **6**, 12 (2018).
25. Morton, J. T. et al. Establishing microbial composition measurement standards with reference frames. *Nat. Commun.* **10**, 2719 (2019).
26. Abadi, M. et al. Tensorflow: A system for large-scale machine learning. In *Proc 12th Symposium on Operating Systems Design and Implementation* 265–283 (USENIX Association, 2016).
27. Quinn, R. A. et al. Niche partitioning of a pathogenic microbiome driven by chemical gradients. *Sci. Adv.* **4**, eaau1908 (2018).
28. Shalapour, S. et al. Inflammation-induced IgA$^+$ cells dismantle anti-liver cancer immunity. *Nature* **551**, 340–345 (2017).
29. Lloyd-Price, J. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
30. Swenson, T. L., Karaoz, U., Swenson, J. M., Bowen, B. P. & Northen, T. R. Linking soil biology and chemistry in biological soil crust using isolate exometabolomics. *Nat. Commun.* **9**, 19 (2018).
31. Baran, R. et al. Exometabolite niche partitioning among sympatric soil bacteria. *Nat. Commun.* **6**, 8289 (2015).
32. Quinn, R. A. et al. A Winogradsky-based culture system shows an association between microbial fermentation and cystic fibrosis exacerbation. *ISME J.* **9**, 1024–1038 (2015).
33. Moree, W. J. et al. Interkingdom metabolic transformations captured by microbial imaging mass spectrometry. *Proc. Natl Acad. Sci. USA* **109**, 13811–13816 (2012).
34. Wang, M. et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
35. Maier, R. M. & Soberon-Chavez, G. *Pseudomonas aeruginosa* rhamnolipids: biosynthesis and potential applications. *Appl. Microbiol. Biotechnol.* **54**, 625–633 (2000).
36. Wood, T. L. et al. Rhamnolipids from *Pseudomonas aeruginosa* disperse the biofilms of sulfate-reducing bacteria. *NPJ Biofilms Microbiomes* **4**, 22 (2018).
37. Allen, L. et al. Pyocyanin production by *Pseudomonas aeruginosa* induces neutrophil apoptosis and impairs neutrophil-mediated host defenses in vivo. *J. Immunol.* **174**, 3643–3649 (2005).
38. Quinn, R. A. et al. Chemical impacts of the microbiome across scales reveal novel conjugated bile acids. Preprint at *bioRxiv* https://doi.org/10.1101/654756 (2019).
39. Paczosa, M. K. & Mecsas, J. *Klebsiella pneumoniae*: going on the offense with a strong defense. *Microbiol. Mol. Biol. Rev.* **80**, 629–661 (2016).
40. Tiratterra, E. et al. Role of bile acids in inflammatory bowel disease. *Ann. Gastroenterol.* **31**, 266 (2018).
41. Hofmann, A. F. & Eckmann, L. How bile acids confer gut mucosal protection against bacteria. *Proc. Natl Acad. Sci. USA* **103**, 4333–4334 (2006).
42. Begley, M., Gahan, C. G. M. & Hill, C. The interaction between bacteria and bile. *FEMS Microbiol. Rev.* **29**, 625–651 (2005).
43. Okada, Y. et al. *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**, 681–688 (2006).
44. Foligne, B. et al. Immunomodulation properties of multi-species fermented milks. *Food Microbiol.* **53**, 60–69 (2016).
45. Ple, C. et al. Combining selected immunomodulatory *Propionibacterium freudenreichii* and *Lactobacillus delbrueckii* strains: reverse engineering development of an anti-inflammatory cheese. *Mol. Nutr. Food Res.* **60**, 935–948 (2016).
46. Colliou, N. et al. Commensal *Propionibacterium* strain uf1 mitigates intestinal inflammation via th17 cell regulation. *J. Clin. Invest.* **127**, 3970–3986 (2017).
47. Vázquez-Baeza, Y., Pirrung, M., Gonzalez, A. & Knight, R. Emperor: a tool for visualizing high-throughput microbial community data. *Gigascience* **2**, 16 (2013).

Methods

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 multiomics data. The model's assumed generative model for metabolite ν, microbe μ and sample k given as follows.

First generate microbe vector $\mathbf{u}_\mu$ for microbe $\mu \in \{1,\ldots N\}$ and metabolite vector $\mathbf{v}_\nu$ for metabolite $\nu \in \{1,\ldots M\}$,

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

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 $\boldsymbol{I}$ with variances σ_u and σ_v, 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

$$\mu \sim \text{Categorical}(\mathbf{x_k})$$

That microbe μ can be used to index U to generate conditional probabilities $\mathbf{q}_\mu$

$$P(\nu|\mu) = \frac{\exp(\mathbf{v}_\nu \cdot \mathbf{u}_\mu + \nu_{\nu 0} + u_{\mu 0})}{\sum_j \exp(\mathbf{v}_j \cdot \mathbf{u}_\mu + \nu_{j0} + u_{\mu 0})},$$

$$\mathbf{q}_\mu = [P(\nu_1|\mu), \ldots P(\nu_M|\mu)]$$

Here $\nu_{j0} + u_{\mu 0}$ are row and column biases, which are required to accurately estimate the conditional probabilities. The above transformation is the softmax transform[48] to compute probabilities from real-valued quantities. This transformation is also known as the inverse clr transform[49], which enforces scale invariance as shown in the simulations. In the generative process of mmvec model, these conditional probabilities generate the metabolite abundances $\mathbf{y}_k$ for a given sample k through a multinomial distribution.

$$\mathbf{y}_k \sim \text{Multinomial}(n, \mathbf{q_\mu}),$$

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[50]. There are a couple of majors differences to be considered. First, in the original application of word2vec, a skipgram was proposed. Skipgrams[50] 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 was 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 unnecessasarily complicated, as there is no such context 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 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:

$$P(Y = y|X_\mu = 1) = \text{Multinomial}(y|q_\mu)$$

where y is the vector of observed metabolites, Y is the random variable modeling metabolite abundances, X is a random variable modeling microbe abundances, $\mathbf{x}$ 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:

$$\mathbf{U} = [\mathbf{0}, \mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_N]^T \qquad \mathbf{V} = [\mathbf{v}_0, \mathbf{0}, \mathbf{v}_1, \ldots, \mathbf{v}_M]$$

Here $\boldsymbol{U} \in R^{N \times p}$ and $\boldsymbol{V} \in R^{(M-1) \times p}$ 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 $\mathbf{u}_0$ and $\mathbf{v}_0$ are critical for estimating accurate co-occurrence probabilities, as suggested by similar methodologies used in recommender systems[51]. The $\boldsymbol{U}$ and $\boldsymbol{V}$ matrices are estimated through maximum a posteriori estimation using ADAM[52] with the following log-posterior

$$\mathcal{L} = \mathcal{L}_Y + \mathcal{L}_U + \mathcal{L}_V$$
$$\mathcal{L}_U = \sum_\mu \sum_{\rho=1}^p \mathcal{N}(U_{\mu,\rho}|0, \sigma_u)$$
$$\mathcal{L}_V = \sum_\nu \sum_{\rho=1}^p \mathcal{N}(V_{\nu,\rho}|0, \sigma_v)$$
$$\mathcal{L}_Y = \sum_k \sum_{r \in x_k} \text{Multinomial}(\mathbf{y}_k|\mathbf{q}_\mu)$$

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 error (SSE). Predictions can be made as follows

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

where the predictive metabolite abundances are compared to the hold-out abundances y_k across all microbial reads i in the hold-out samples k. m_k denotes the total metabolite abundances in sample k.

Microbe–metabolite vectors in simplicial coordinates. Here we will provide some insights behind the underlying geometry behind this neural network. Doing so will support the 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, are given as follows

$$\text{softmax}(\mathbf{x}) = \left[\frac{\mathbf{e}^{\mathbf{x}_1}}{\sum_i \mathbf{e}^{\mathbf{x}_i}}, \ldots, \frac{\mathbf{e}^{\mathbf{x}_D}}{\sum_i \mathbf{e}^{\mathbf{x}_i}}\right]$$

$$\text{clr}(\mathbf{z}) = \left[\log \frac{\mathbf{z}_1}{\mathbf{g}(\mathbf{z})}, \ldots, \log \frac{\mathbf{z}_D}{\mathbf{g}(\mathbf{z})}\right]$$

As biases are incorporated into the mmvec model, by construction, $\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{V}^T$ is both row centered and column centered, meaning that the sum of rows is zero and the sum of the columns is zero. Given this the following holds:

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

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

$$\mathbf{Q} = \mathbf{U}(\mathbf{V} - \mathbf{1}_M \boldsymbol{\lambda}_\mathbf{v}^T)$$

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

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

$$\mathbf{U}\lambda_\mathbf{v} M = 0$$

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

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

$$\mathbf{Q} = (\mathbf{U} - \mathbf{1}_N \boldsymbol{\lambda}_\mathbf{u}^T)\mathbf{V}^T$$

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

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

$$N \boldsymbol{\lambda}_\mathbf{u}^T \mathbf{V} = 0$$

This means that only the trivial solution $\boldsymbol{\lambda}_\mathbf{u} = 0$ 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 $\boldsymbol{U}$ and $\boldsymbol{V}$ 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. As the $\boldsymbol{U}$ and $\boldsymbol{V}$ 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 latent dirichlet allocation[53,54].

Linearity. Vectors in clr coordinates are known to satisfy linearity, namely

$$\mathrm{clr}(\alpha\mathbf{x} + \mathbf{y}) = \alpha\mathrm{clr}(\mathbf{x}) + \mathrm{clr}(\mathbf{y})$$

for $\alpha \in R$, $x \in S^p$ and $y \in S^p$. This linearity property was leveraged in word2vec models to perform analogy reasoning. As 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.

Visualization through biplots. Visualization techniques from compositional data analysis can aid with interpretation[55,56]. $\boldsymbol{U}$ 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 two-dimensional scatter plot and the first two latent dimensions of $\boldsymbol{V}$ represent metabolite coordinates on a two-dimensional scatter plot. Typically the coordinates from the $\boldsymbol{V}$ matrix are plotted as arrows from the origin to identify features that explain the variance in $\mathbf{U}$. 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 ref. [57], 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[23].

Benchmarks. The simulated data were based on a cystic fibrosis biofilm model derived by Quinn et al. (shown in Figure S12 of ref. [27]). The biofilm model was built to explain how fermenters and *P. aeruginosa* responded to different concentrations of sugars, amino acids, pH, oxygen and antibiotics across the Winogradsky column. These models solved for differential equations integrating Monad kinetics and diffusion processes and were run in Matlab using the code provided at https://github.com/zhangzhongxun/WinCF_model_Code.

From this simulation, we only focus two microbes and five compounds. The two microbes are *P. aeruginosa* (Θ_p) and fermenters (Θ_f). The model also includes the five compounds (SG), acids (F), ammonium (P), amino acids (SA) and inhibition molecules (I). 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:

$$\mathbf{p}_i \sim \mathcal{N}(0, \sigma_o \mathbf{I}) \quad \mathbf{q}_i \sim \mathcal{N}(0, \sigma_c \mathbf{I})$$

$$\mathbf{o}_{ij} = \kappa_{ij}\mathrm{ilr}^{-1}(\mathbf{p}_i) \quad \mathbf{c}_{ik} = \eta_{ik}\mathrm{ilr}^{-1}(\mathbf{q}_i)$$

where $_{pi}$ is a vector of 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. $\mathbf{o}_{ij}$ 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 Fig. 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. $\mathbf{q}_i$ is a vector of 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. $\mathbf{c}_{ik}$ 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 absolute abundances of subtaxa and subcompounds have been simulated, the microbial relative counts and metabolite abundances are simulated. The sampling procedure is performed as follows:

$$\zeta_i \sim \mathcal{LN}(n, \tau_o) \quad \omega_i \sim \mathcal{LN}(m, \tau_c)$$

$$x_i \sim \mathcal{PLN}(\zeta_i\, C(\mathbf{o}_i), \varepsilon_o) \quad y_i \sim \mathcal{LN}(\omega_i\, C(\mathbf{c}_i), \varepsilon_c)$$

The total sequencing depths and total intensities for sample i are drawn 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, as the parameters can be directly interpreted as a normal distribution, and consequently has a compositional interpretation owing to its connection to the ilr transform. Finally, the lognormal distribution is 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 $C(\mathbf{o}_i)$ 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 $C(\mathbf{c}_i)$ 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.

Software workflows. To facilitate utilization of the mmvec tool, we have developed two different user interfaces. First, we have developed a qiime2 plugin[58], where mmvec can be run using a simple command line interface. This interface is complemented using ref. [26], where users can monitor convergence rates for their models in real time and evaluate how different parameters will affect their model fit (Supplementary Fig. 4). Second, we have integrated mmvec into the GNPS platform that can be accessed by the public. The online interface through the GNPS resolves several usability issues. First, the GNPS facilitates import of metabolomics data into qiime2 by preprocessing, importing and sample renaming, This is performed as part of the standard metabolomics analysis on the GNPS (for example, molecular networking and feature-based molecular networking). Second, as it is possible to both download and reuse outputs of workflows run on the 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 begin. Once the workflow completes, the biplots can be viewed directly in the browser and other outputs (for example, ranks) are available for download (Supplementary Fig. 5).

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 Tesla GPU, the model can reach convergence within 4 h on the IBD dataset that comprises 562 microbial taxa, 26,966 metabolite features and 400 samples. However, there is a trade-off between accuracy and running time. More accurate models require smaller learning rates and may take longer to run.

Data analysis. Owing to the overwhelming sparsity in microbiome datasets, some filtering is required to infer microbe–metabolite interactions. We chose to filter out microbes that appeared in less than ten samples, as 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 HFD study, there were 434 samples and after filtering there were 902 microbes and 11,978 metabolites. For the IBD dataset, there were 13,920 features in the c18 LC–MS dataset, 26,966 features in the c8 LC–MS dataset and 562 taxa. Cross validation was performed across all studies to evaluate overfitting. In the desert biocrust soils experiment, one sample of 19 was randomly chosen to be left out for cross validation. In all of the other studies, ten 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.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The cystic fibrosis sequencing and metadata data can be found at https://qiita.microbio.me/ under study ID 10863. The corresponding GNPS analysis can be

accessed at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=34d825dbf4e9466e81d809faf814995b. The biocrust soil data were retrieved from the supplemental section in Swenson et al.[30]. The HFD murine model case study 16S rRNA data can be found at https://qiita.microbio.me/ under study ID 10856. The HFD murine model case study data are publicly available at https://massive.ucsd.edu/ under 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/.

Code availability

The software implementing the mmvec algorithm can be found under https://github.com/biocore/mmvec. Differential abundance analyses in the HFD study were 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. Biplots were generated using Emperor[47].

References

48. Nasrabadi, N. M. Pattern recognition and machine learning. *J. Electron. Imaging* **16**, 049901 (2007).
49. Pawlowsky-Glahn, V., Egozcue, J. J. & Tolosana-Delgado, R. *Modeling and Analysis of Compositional Data*. (John Wiley & Sons, 2015).
50. Tomas, M., Ilya, S., Kai, C., Corrado, G. S. & Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26* (eds Burges, C. J. C. et al.) 3111–3119 (NIPS, 2013).
51. Koren, Y., Bell, R. & Volinsky, C. Matrix factorization techniques for recommender systems. *Computer* **42**, 30–37 (2009).
52. D. P. Kingma and J. Ba. Adam: a method for stochastic optimization. Preprint at *arXiv* https://arxiv.org/abs/1412.6980 (2014).
53. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
54. Sankaran, K. & Holmes, S. P. Latent variable modeling for the microbiome. *Biostatistics* **20**, 599–614 (2019).
55. Aitchison, J. & Greenacre, M. Biplots of compositional data. *J. R. Stat. Soc. Ser. C Appl. Stat.* **51**, 375–392 (2002).
56. Aitchison, J. & Ng, K. W. Conditional compositional biplots: theory and application. *DUGiDocs* https://dugi-doc.udg.edu/handle/10256/657 (2005).
57. Martín-Fernández, J. A., Pawlowsky-Glahn, V., Egozcue, J. J. & Tolosona-Delgado, R. Advances in principal balances for compositional data. *Math. Geosci.* **50**, 273–298 (2018).
58. Bolyen, E. et al. Reproducible, interactive, scalable, and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).

Acknowledgements

We would like to thank V. Pawlowsky, J. J. Egozcue and S. Holmes for their insights on the geometry of this neural network model. In addition, we would also like to thank N. Bokulich for feedback and contributions on the mmvec software package. T.L.S., M.W.V.G. and T.R.N. acknowledge funding from the Office of 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 study was in part supported by grant P41GM103484 for the Center for Computational Mass Spectrometry and instrument support through National Institutes of Health grants 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 National Science Foundation grant GRFP DGE-1144086. R.K. and S.J.S. have been funded by Janssen under grant number 20175015 and the Alfred P. Sloan Foundation under grant number G-2017-9838.

Author contributions

J.T.M. wrote the mmvec algorithm, conducted the benchmarks and ran all of the analyses. A.A.A. and L.F.N. preprocessed and annotated the metabolomics data. A.A.A. provided insights in the HFD study. J.R.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. N.A.B developed the heat map visualizations. A.W. developed the network visualizations. T.L.S., M.W.V.G and T.N. provided insights into the biocrust soils experiment. R.B. provided insights behind the simulation benchmarks. S.J.S provided ecological insights. P.C.D provided insights behind metabolomics. All authors were involved with writing the manuscript.

Competing interests

Mingxun Wang is the founder of Ometa Labs LLC. The remaining authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0616-3.

Correspondence and requests for materials should be addressed to R.K.

Peer review information Lei Tang was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

nature research

Corresponding author(s): NMETH-A38236

Last updated by author(s): Sep 10, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Only published data is used. No software was used for data collection
Data analysis	All data analysis scripts can be found here: https://github.com/knightlab-analyses/multiomic-cooccurences

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on 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
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

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was chosen based on the published datasets. 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.
Data exclusions	Taxa that appeared in less than 10 samples for each study were removed, since there are fewer samples than degrees of freedom in the model to infer these microbes co-occurrence patterns. This exclusion criteria was not pre-established.
Replication	Extensive software unittests have been developed to ensure that the algorithm is been reproducible. Tutorials are also available to show case this.
Randomization	The experimental designs were pre-established in previous studies, so this is not applicable.
Blinding	The experimental designs were pre-established in previous studies, so this is not applicable

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging