Variational Autoencoders for Metabolomics: Data Imputation, Deconfounding, and Correlation Discovery

Huang Lin[1,*], Lijun Zhang[2], Alexander Aksenov[3], Alan Jarmusch[4], Iris Lee[5], and James T. Morton[6]

[1]Department of Epidemiology and Biostatistics, University of Maryland, College Park, MD, 20742, USA
[2]Tiposi, 1900 McCarthy Blvd Ste 106, Milpitas, CA, 95035, USA
[3]University of Connecticut, Storrs, CT, 06269, USA
[4]National Institute of Environmental Health Sciences, Durham, NC, 27709, USA
[5]Division of Rheumatology, Department of Medicine, Washington University in St. Louis, St. Louis, MO, USA
[6]Gutz Analytics L.L.C., Boulder, CO, 80304, USA
[*]hlin1239@umd.edu

ABSTRACT

Untargeted metabolomics faces significant statistical challenges, including compositionality, missing values, and confounding effects that can lead to spurious correlations and impede accurate metabolite relationship analysis. Here we present MetVAE, a variational autoencoder-based model that addresses these challenges. In simulations, MetVAE demonstrates superior performance over competing methods in controlling false discovery rates while maintaining high true positive rates. Applied to inflammatory bowel disease data, MetVAE identifies three metabolite clusters with disease-specific correlation patterns. Analysis of hepatocellular carcinoma data shows MetVAE can complement structural similarity measures by uncovering functional relationships between metabolites. Using a large-scale dataset of over 20,000 samples, we also demonstrate MetVAE's scalability and effectiveness in reducing technical variation while preserving biological signals. Our results establish MetVAE as a powerful and versatile tool for advancing metabolomics research, enabling deeper exploration of disease mechanisms and metabolic interactions.

Introduction

Metabolomics stands as a critical discipline within the 'omics' sciences, utilizing state-of-the-art analytical instruments and advanced pattern recognition techniques to reveal metabolic alterations in subjects, arising from disease states or in response to medical interventions or environmental changes[1]. The field is divided into two core strategies: targeted and untargeted metabolomics, employing technologies, such as nuclear magnetic resonance (NMR), gas chromatography-mass spectrometry (GC-MS), and liquid chromatography-mass spectrometry (LC-MS), for data acquisition[2–4]. Targeted metabolomics, which is hypothesis-driven, typically focuses on validating known processes. In contrast, untargeted metabolomics offers a broad and exhaustive analysis, aiming to measure as many metabolite as possible present in a sample, including those not previously identified[5]. This comprehensive approach seeks to provide extensive metabolic insights, facilitating the discovery of functional compounds and biomarkers[6,7].

The identification of metabolites and the biological interpretation of the vast and complex datasets generated through untargeted metabolomics pose significant challenges. An effective strategy to surmount these challenges involves the analysis of metabolic interactions. Metabolites are typically not identified as isolated units but rather as constituents of broader groups within identical chemical classes[8]. The examination of metabolite interactions has been instrumental in guiding the annotation of unknown metabolites[9], pinpointing potential biomarkers for disease diagnosis[10], and elucidating the underlying mechanisms of disease progression[11].

The initial step in characterizing metabolite interactions typically involves identifying pairwise relationships between metabolites. However, this task presents significant challenges due to the unique characteristics of untargeted

metabolomics data. Untargeted metabolomics, with its broad and exploratory nature, encompasses a vast spectrum of both known and unknown compounds, aiming for the relative quantification of thousands of metabolites in biological samples. This approach to quantification largely stems from the absence of standards and the impracticality of establishing calibration curves for hundreds to thousands of metabolites, rendering the data from untargeted metabolomics compositional in nature[12]. The Supplementary Methods section provides an in-depth discussion of the compositionality inherent in untargeted metabolomics, examining it through the lens of calibration curves. Notably, this issue of compositionality is not exclusive to metabolomics but is also prevalent in other high-throughput sequencing data, such as microbiome data[13–18]. The literature strongly cautions against applying Pearson or Spearman correlation coefficients directly to compositional data[19–23]. For instance, these coefficients produce spurious results when working with relative data due to their shared denominator[24]. To illustrate, we simulated true abundance data for 20 metabolites with 100 samples, using a log-normal distribution. Sample-specific and feature-specific biases were added to mimic the relative nature of untargeted metabolomics data. Out of all pairwise relationships, we designed only four pairs to be truly correlated (Extended Data Fig. 1a; details in Supplementary Methods). Applying Pearson correlation to this simulated data generated numerous spurious correlations caused by the biases, leading to significantly inflated false positives (Extended Data Fig. 1b). To address this issue, several compositionally robust methods have been developed. These include co-occurrence based methods like Sparse Correlations for Compositional data (SparCC)[19], variance of log-ratios (VLRs)[25], proportionality[20], and Sparse Estimation of Correlations among Microbiomes (SECOM)[23], alongside conditional dependence approaches such as Sparse Inverse Covariance Estimation for Ecological Association Inference (SPIEC-EASI)[21] and Semi-Parametric Rank-based approach for Inference in Graphical model (SPRING)[26]. These methods have proven effective in reducing false positives and accurately uncovering the true underlying relationships (results from VLR, SparCC, SECOM, and SPIEC-EASI are presented in Extended Data Fig. 1c, d, e, f, respectively). Note that for hypothesis testing concerning VLR, we have developed an innovative permutation testing framework that utilizes what we refer to as normalized VLR values. Further elaboration on this technique is provided in the Supplementary Methods.

Moreover, untargeted metabolomics data are frequently plagued by missing values, which are broadly classified into three categories: missing not at random (MNAR), missing at random (MAR), and missing completely at random (MCAR)[27,28]. While some missing values arise from random errors (MCAR) or suboptimal data preprocessing (MAR)[29], a significant portion of missing values, known as censored values, occur when measurements fall below the limits of quantification (LOQ) and are thus considered MNAR[30]. Most compositionally robust methods involve taking logarithms or log ratios of the data, necessitating the addition of a pseudo count to circumvent numerical issues. However, while adding a pseudo count is a straightforward adjustment, it can significantly alter the distribution of the original data, particularly when missing values are MNAR. For example, introducing a pseudo count in the calculation of VLR can obscure the relationship between truly correlated metabolite pairs and artificially create correlations among uncorrelated pairs, given that the data effectively originate from a mixture of distributions (Extended Data Fig. 2a). Similarly, SparCC encounters challenges with both a loss of true positives and an increase in false positives (Extended Data Fig. 2b). Conversely, methods based on complete data like SECOM and probabilistic graphical model like SPIEC-EASI experience a reduction in true positive rates (Extended Data Fig. 2c and d, respectively).

Lastly, the accurate quantification of pairwise relationships is often compromised by the presence of unwanted confounders. A prime example of such confounders are batch effects—systematic technical discrepancies arising during the sample processing phase, which bear no relation to biological variations. To date, there are limited methods specifically designed to quantify pairwise relationships that can concurrently address these confounders. To evaluate the impact of confounders on the outcomes of pairwise relationship quantification, we introduced one continuous and one binary confounder into the illustrative dataset (with details provided in the Supplementary Methods). Note that in the absence of confounders, we still have four pairs that are truly correlated as shown in Extended Data Fig. 1a. Upon applying the previously mentioned compositionally robust methods to the data now confounded, it was unsurprising to find that none were capable of accurately identifying true positives, and all demonstrated a significant increase in false positives, with the Pearson correlation coefficient performing the worst (results from Pearson correlation coefficient, SparCC, SECOM, and SPIEC-EASI are depicted in Extended Data Fig. 3a, b, c, d, respectively).

Given the challenges associated with current methods for quantifying pairwise relationships, this study introduces a Variational Autoencoder (VAE) based model named Metabolomics Variational Autoencoder (MetVAE). The MetVAE model employs a multifaceted approach to address prevalent issues in metabolomics data analysis: (1) compositionality

is addressed by applying a centered log-ratio (CLR) transformation to the input data; (2) the generative architecture of the VAE is harnessed to capture latent structure and enable robust multiple imputation of missing values; and (3) known confounders are explicitly incorporated into the framework to eliminate their effects in computing pairwise correlations. MetVAE's efficacy is assessed through a series of simulation studies, complemented by applications to real-world datasets, namely a inflammatory bowel disease (IBD) data, a hepatocellular carcinoma (HCC) mouse experiment data, and the dataset from the Reanalysis of Data User (ReDU) interface. Detailed methodologies supporting these analyses are elaborated in the Methods section.

Results

Simulation benchmark

In this section, we utilized simulated data to benchmark the MetVAE model against several established methodologies, including the standard Pearson correlation coefficient, SECOM[23], SparCC[19], SPIEC-EASI[21], and SPRING[26]. As described in the Methods section, MetVAE employs a regularization method to obtain a sparse correlation matrix. For competing methods that have an embedded approach to achieve sparsity, we utilized their default methods. Specifically, for SPIEC-EASI and SPRING, we selected the neighborhood selection strategy proposed by Meinshausen and Bühlmann (the MB method)[31], as recommended for its enhanced performance over the graphical lasso algorithm[32] in numerical simulations[26]. For SECOM (Thresholding), a hard thresholding method was implemented with parameters set to default values. For competing methods without a specific approach for obtaining sparsity but where P-values were derived, i.e., Pearson, SparCC, and SECOM (Filtering), we derived a sparse correlation matrix by employing adjusted P-value filtering. This means that correlation coefficients were retained only if their corresponding adjusted P-values fell below a predetermined threshold (set to 0.05 in our simulations). The adjusted P-values were derived by applying the Benjamini–Hochberg (BH) procedure[33] to the raw P-values. It is important to note that the BH procedure assumes that the null hypotheses are either independent or positively dependent. However, this assumption may not hold in our scenario, as pairwise correlations can exhibit complex interdependencies. Properly addressing the P-values given these inter-relationships is beyond the scope of this paper. Therefore, the application of the BH procedure in our analysis was primarily to "somewhat" control the false discovery rate (FDR) for illustrative purposes, rather than to assert definitive statistical control. We evaluated the performance of various methods based on their ability to accurately identify true nonzero correlations, which is quantified by the true positive rate (TPR), also known as the statistical power. Additionally, we assessed the proportion of erroneously identified correlations relative to all correlations identified, measured by the false discovery rate (FDR). For comprehensive details on the simulation settings, we direct readers to the Supplementary Methods.

Data without confounders. Fig. 1 provides a summary of the results using various methods to estimate the correlation matrix in scenarios without confounders. In comparison to compositionally robust methods, standard Pearson correlation coefficients demonstrated significantly higher FDR when missing values were minimal. For example, in a basic low-dimensional scenario with $N = 100, D = 50$, and no missing values, the Pearson correlation coefficients displayed an FDR nearly equal to 1, indicating that nearly all its identified correlations were likely spurious. As the proportion of missing values increased, the calculation of Pearson coefficients, which relied solely on complete data, experienced a marked decrease in TPR and FDR. Aside from the standard Pearson correlation coefficients, the conditional dependence-based methods, SPIEC-EASI and SPRING, exhibited the weakest performance in controlling false positives, as reflected in their high FDR. In the low-dimensional scenario of $N = 100, D = 50$, SPRING showed an average FDR of approximately 85%, with the FDR escalating to almost 100% in both high-dimensional cases. Concurrently, SPRING experienced a significant loss in TPR not only in the low-dimensional scenario ($N = 100, D = 50$) but also in the high-dimensional cases ($N = 100, D = 200$), where the TPR was only around 50%. SPIEC-EASI recorded an average FDR of around 40% in the low-dimensional scenario and saw this rate escalate to above 75% in high-dimensional cases. On the other hand, SPIEC-EASI consistently exhibited a larger TPR compared to SPRING across most simulation cases. Among the co-occurrence based methods, SparCC exhibited the highest FDR, averaging about 50% in the low-dimensional scenario. Its TPR was comparable to that of SECOM (Thresholding) and SECOM (Filtering). However, SparCC utilizes a permutation-based algorithm to obtain P-values, which is very computationally intensive. In practical terms, processing one simulation dataset with $N = 100, D = 200$ using a single CPU took approximately 6 hours, making SparCC impractical for bench-marking in high-dimensional scenarios. Consequently, SparCC was not considered

for further analysis in high-dimensional conditions due to these computational constraints. SECOM (Thresholding) achieved the lowest FDR among all the methods tested, though this came at the cost of a TPR that was about 5% lower than that observed with SECOM (Filtering). This trade-off between FDR and TPR is consistent with findings reported in the original publication of SECOM[23]. MetVAE demonstrated a comparable FDR to SECOM (Thresholding) while remaining significantly lower than all other competing methods. Notably, MetVAE also consistently maintained a sufficiently large TPR across all test scenarios. Its TPR was nearly always above 90%, more than double that of both SECOM methods when the level of missing data escalated to 30%. This marked disparity arises because SECOM primarily relies on complete datasets to effectively control false positives, thereby limiting its performance in scenarios with substantial missing data.

Data with confounders. We also evaluated how various methods performed in identifying correlations when confounders were present in the data. The results are summarized in Fig. 2. As expected, since none of the competing methods examined were specifically designed to handle confounders, they experienced significantly inflated FDR and reduced TPR across all simulation scenarios. On the contrary, since MetVAE explicitly incorporates confounders in its generative model, it effectively managed to maintain its FDR at levels similar to those observed in scenarios without confounders, which are significantly lower than all competing methods listed. Additionally, MetVAE was also able to sustain the highest TPR, particularly when the proportion of missing data was minimal. This highlights its exceptional capacity to detect genuine correlations accurately even in the presence of confounders.

MetVAE identifies differential metabolic correlations

We first demonstrate that MetVAE can identify biologically meaningful differential correlation metabolites using untargeted metabolomics data from the Inflammatory Bowel Disease Multi'omics Database (IBDMDB)[34]. Inflammatory bowel disease (IBD), encompassing both Crohn's disease (CD) and ulcerative colitis (UC), affects over one million Americans[35]. The precise etiology of IBD remains unknown, but it is believed to result from a complex interplay between genetic predisposition and an abnormal host immune response to environmental factors. Microbial dysbiosis is a common feature in IBD. Metabolomics offers a unique opportunity to characterize disease-associated metabolic alterations and is particularly valuable for quantifying gut microbial metabolism[36]. The metabolomics data in the IBDMDB were generated using four LC-MS methods on fecal homogenates. Detailed information on study design, sample collection, and processing is available on the IBDMDB website (`https://ibdmdb.org/`). For this re-analysis, we focused on baseline visits, resulting in a total of 70 subjects: 31 with CD, 20 with UC, and 19 non-IBD controls.

Overall, the highest number of correlated metabolites was observed in CD samples, followed by UC samples, with only a small number of pairwise correlations identified in non-IBD controls (Extended Data Fig. 5). Three distinct clusters of metabolites with unique correlation patterns across disease states were identified (Supplementary Fig. 1). The first cluster, consisting of 37 annotated metabolites (purple box, Fig. 3), and the second cluster, comprising 22 annotated and 1 unannotated metabolite (orange box, Fig. 3), exhibited strong positive correlations specifically within CD patients, but not in UC or non-IBD controls. Notably, the majority of metabolites in Cluster 1 (purple box, Fig. 3) are triglycerides (TGs) and Lysophosphatidylethanolamine (LysoPEs). Triglycerides have been observed to be elevated in CD patients compared to both UC patients and non-IBD controls[37]. Cluster 2 (orange box, Fig. 3) primarily consists of phosphatidylcholines (PCs), which are commonly supplemented to manage inflammation in both CD and UC due to their role in maintaining gut barrier integrity and have been found to be elevated in stool metabolomics analyses of patients with these conditions[38,39].

Cluster 3 comprises 13 metabolites, all of which are cholesteryl esters (CEs)—cholesterol molecules esterified with fatty acids. These molecules are critical for cholesterol transport, and disruptions in CE metabolism have been linked to inflammation[40]. CEs serve as precursors to cholesterol in the gut, and their metabolism is notably altered in IBD. Specifically, IBD is associated with changes in cholesterol synthesis, intracellular cholesterol trafficking, and cholesterol absorption[41–43]. Dyslipidemia in IBD is well-documented[44–46] and associated with high disease activity[37]. Interestingly, the CEs in this study exhibit strong positive correlations in both CD and UC patients, but not in non-IBD controls (green box, Fig. 3).

Together, these three clusters of metabolites suggest a disruption in lipid metabolism. The increased density of positively correlated metabolites in CD patients may reflect the heightened severity of inflammatory processes in CD compared to UC. Details of these metabolites are provided in Supplementary Table 1.

We also explored the structural similarity among these metabolites by computing their pairwise Tanimoto scores[47]. The Tanimoto score ranges from 0 to 1, with higher scores indicating greater structural similarity. The metabolites identified exhibited high Tanimoto scores within clusters but not between clusters (Extended Data Fig. 5), highlighting the high degree of molecular fingerprint similarity within each cluster. Notably, Cluster 2 contains one unannotated metabolite. Given its high structural similarity to other PCs in the cluster, it is likely another type of PC. This observation underscores MetVAE's potential as a tool for chemical annotation, addressing a critical need in untargeted metabolomics for accurate identification of unknown metabolites. To gain further functional insights into these metabolites, we mapped their HMDB IDs to Kyoto Encyclopedia of Genes and Genomes (KEGG) IDs. While many TGs and CEs are annotated in KEGG, only one PC has a KEGG ID, and none of the LysoPEs are mapped to KEGG. This highlights the limited understanding and annotation of biological pathways involving phosphatidylcholine and lysophosphatidylethanolamine species. As a comparison to the MetVAE results, we also computed Pearson and Spearman correlations on CLR-transformed data, retaining only correlations with a p-value less than 0.005. In general, distinct clusters of metabolites were not identified. For instance, the first cluster of metabolites identified by MetVAE did not exhibit differential correlations across disease states when analyzed using either Pearson (Extended Data Fig. 6) or Spearman correlations (Extended Data Fig. 7). Additionally, the differential correlation patterns observed in the second and third clusters were notably weakened.

MetVAE enhances biological insights beyond structural similarities

We further demonstrate that MetVAE complements structural similarity measures by providing additional biological insights using untargeted metabolomics data from a hepatocellular carcinoma (HCC) mouse experiment[48]. Among the 2084 metabolites present in the dataset, only 168 metabolites (around 8%) were annotated, highlighting a significant gap in feature annotation within untargeted metabolomics datasets. In the original study, mice were assigned to one of two diet plans: normal chow (NC) or a high-fat diet (HFD). HFD was identified as a significant risk factor for developing HCC (Supplementary Fig. 2). As reported, HFD feeding induced microbial dysbiosis and altered fecal metabolites. To account for this, we adjusted for diet patterns as a confounder in the MetVAE model.

We constructed a microbial correlation network using MetVAE and compared it to a structural similarity network based on cosine similarity (> 0.9) calculated from MS2 features. Notably, MetVAE revealed connections that bridged otherwise disjointed clusters identified by cosine similarity. For instance, in Fig. 4, metabolites colored in red, comprising small peptides (di- and tri-) and bile acids, formed four disjoint clusters (subgraphs 2, 13, 59, and 338) with no interactions based on cosine similarity. MetVAE, however, identified several highly positive correlations (>0.9) between these clusters. The positive correlations among bile acids across the red clusters suggest coordinated shifts in bile acid profiles, while the connections between small peptides and bile acids point to increased protein catabolism during disease progression, altered gut microbiota activity influencing both bile acid transformation and protein metabolism, and possible inflammatory responses driving peptide production.

Similarly, the blue cluster in Fig. 4 (subgraph 5), which predominantly contains phosphocholines (lipids), was disconnected from the red clusters (small peptides and bile acids) based on cosine similarity. However, MetVAE uncovered several strong negative correlations (<-0.7) between molecules in the blue and red clusters. These negative correlations suggest altered lipid metabolism during nonalcoholic fatty liver disease (NAFLD) progression, potential compensatory mechanisms wherein increased bile acid production is associated with decreased phosphocholine levels, and feedback loops linking lipid metabolism and bile acid synthesis.

Using cosine-based networking, the small peptides could not be clustered into the same graph. However, by applying MetVAE, we uncovered that these molecular networks are, in fact, highly co-occurring. This observation aligns with biological intuition, highlighting the potential of MetVAE in enriching functional annotations and deepening our understanding of metabolic networks.

Scalability of MetVAE

Lastly, we demonstrate the scalability of MetVAE using a combined dataset from the Reanalysis of Data User (ReDU) interface[49]. ReDU provides access to uniformly formatted public MS/MS data in the Global Natural Product Social Molecular Networking Platform (GNPS) via formatted metadata (`https://gnps.ucsd.edu/`). We fitted MetVAE on a subset of RedU consisting of 20,330 samples and 3,929 features. The resulting sparse correlation coefficient matrix identified 10,376 pairs of metabolites with correlation coefficients greater than 0.1, involving 1,249 unique metabolites. This analysis demonstrates that MetVAE is highly scalable, and it opens opportunities for transfer learning.

Users can pre-train the model on general databases, such as ReDU, and later fine-tune the parameters using tailored datasets. Our simulation results showed that transfer learning with MetVAE significantly reduced the convergence time of the algorithm (data not shown). Using ReDU data, we also demonstrated MetVAE's effectiveness in reducing technical variation, visualized by applying UMAP[50] to both raw and reconstructed data (Fig. 5). ReDU integrates datasets from various types of mass spectrometers, leading to significant technical variation. In Fig. 5a, raw data clearly show separation by mass spectrometer type, indicating high technical variability. However, after applying MetVAE and specifying mass spectrometer type in the model, the reconstructed data display reduced technical variation (Fig. 5b), as quantified by normalized between-group sum of squares (SSB), which decreased from 0.44 in the raw data to 0.33 in the reconstructed data. In contrast, biological variation, such as health status (Fig. 5), was well-preserved and even enhanced, with SSB increasing from 0.04 to 0.08 in the reconstructed data.

Discussion

Metabolomics represents the downstream output of the genome while also serving as an upstream input from the environment, positioning it at the foundation of the "omics" cascade. While genomics and proteomics may indicate potential events, metabolomics reveals what actually occurred. Since its inception, metabolomics research has rapidly expanded, finding applications in early disease diagnosis, personalized treatments, environmental health, and more. However, the analysis of metabolomics data, especially untargeted metabolomics, faces several statistical challenges, such as the inherent compositional nature of the data, high rates of missing values, and confounding effects like batch effects in large-scale studies. If these challenges are not properly addressed, standard methods can suffer from high FDR and reduced TPR. To address these issues, we developed MetVAE, based on the VAE framework, which imputes missing values while preserving compositional structure and accounting for confounding effects. Simulation studies demonstrated that MetVAE outperforms competing methods in terms of FDR control and maintaining a high TPR.

In addition to simulations, we also applied MetVAE to three real datasets to further illustrate its effectiveness. First, using the IBDMDB dataset, we showed that MetVAE can identify differential correlation patterns that standard methods fail to detect, offering valuable insights into the metabolic dysregulation associated with diseases. MetVAE uncovered three distinct clusters of metabolites, each exhibiting unique correlation patterns across disease states. The first metabolite cluster, comprising TGs and LysoPEs, demonstrated strong positive correlations specifically within CD patients. These metabolites are intricately involved in various metabolic pathways (Supplementary Table 2), including fat digestion and absorption, lipid metabolism, and insulin resistance[51]. The dysregulation of these pathways in CD patients suggests a potential link between altered lipid metabolism and the inflammatory processes characteristic of the disease. For instance, the accumulation of visceral fat, which is associated with increased risk for postoperative recurrence in CD, highlights the importance of fat metabolism in disease management[52]. Furthermore, the association of TGs and LysoPEs with insulin resistance may indicate a metabolic syndrome-like state in CD patients, which could complicate their clinical management and increase the risk of comorbidities such as cardiovascular diseases[51]. The second cluster, predominantly composed of PCs, also exhibited strong positive correlations specifically within CD patients. Enrichment analysis revealed several statistically significant pathways, with the top five including basal cell carcinoma, cholesterol metabolism, choline metabolism in cancer, cortisol synthesis and secretion, and cushing syndrome. PCs play a critical role in maintaining gut integrity and modulating inflammatory responses, which are pivotal in the pathogenesis of CD[53,54]. The association of PCs with cholesterol metabolism suggests a potential dysregulation of lipid homeostasis in CD, which may contribute to the inflammatory processes and complications associated with the disease[55]. Additionally, the correlation with pathways related to cortisol synthesis and stress responses indicates a potential link between psychosocial factors and CD, which could inform patient management strategies that address these factors alongside traditional medical treatments[56]. Lastly, the third cluster consists of CEs, which exhibit strong positive correlations in both CD and UC patients, but not in non-IBD controls. Enrichment analysis revealed several statistically significant pathways, with the top five including cholesterol metabolism, fat digestion and absorption, lipid and atherosclerosis, efferocytosis, and ovarian steroidogenesis. Biologically, the involvement of CEs in pathways such as cholesterol metabolism and fat digestion suggests a dysregulation of lipid homeostasis in IBD patients, which may contribute to the chronic inflammation characteristic of these diseases[57]. The association with efferocytosis, a process critical for the resolution of inflammation, indicates that impaired clearance of apoptotic cells may exacerbate the inflammatory milieu in IBD[58,59]. Furthermore, the correlation with pathways related to atherosclerosis underscores the increased cardiovascular risk associated with IBD, suggesting that monitoring lipid profiles and implementing

cardiovascular risk reduction strategies may be essential components of comprehensive care for these patients[60,61]. Overall, these insights underscore the importance of metabolic profiling in understanding IBD, which could pave the way for novel individualized therapeutic interventions.

Additionally, we demonstrated that MetVAE can enhance biological insights by providing functional annotations to structural similarities, using data from the HCC mouse experiment. We identified several network clusters formed by cosine similarities, including dipeptides, bile acids, and phosphocholines. While these clusters were originally disjointed, MetVAE revealed high correlations bridging them, offering new insights into their biochemical relationships. The interplay between dipeptides, bile acids, and phosphocholines reflects their roles in lipid metabolism and cellular homeostasis. For instance, specific dipeptides, such as leucine-leucine (Leu-Leu), positively correlate with bile acids, potentially through pathways that regulate bile acid synthesis and secretion. Bile acids, synthesized from cholesterol in the liver, play a key role in lipid digestion and absorption. The presence of dipeptides can upregulate genes involved in bile acid production, such as those activated by the farnesoid X receptor (FXR)[62]. This interaction supports bile acid homeostasis and enhances lipid metabolism, as dipeptides may stimulate bile acid secretion, facilitating lipid absorption[63]. In contrast, the relationship between dipeptides and phosphocholines is characterized by a negative correlation. Phosphocholines, essential components of cell membranes, contribute to membrane integrity and signaling. However, high bile acid levels can disrupt membrane phospholipid composition, reducing phosphocholine availability[63]. Excess bile acids solubilize membrane phospholipids, impairing cellular function. Elevated dipeptides, particularly in metabolic disorders, may exacerbate this disruption. Increased dipeptide concentrations can drive lipogenesis, potentially competing with phosphocholines for membrane incorporation[64]. Metabolic dysregulation, such as in obesity or non-alcoholic fatty liver disease (NAFLD), may further complicate these interactions. Excess bile acids and altered lipid profiles can deplete phosphocholine precursors, reinforcing the negative correlation[62]. This underscores the intricate balance required for maintaining metabolic health and highlights the role of dipeptides and bile acids in regulating lipid metabolism.

Lastly, an extremely large and high-dimensional ReDU dataset, containing over 40,000 samples and 15,000 features, was used to demonstrate the scalability of MetVAE and its effectiveness in mitigating confounding effects.

Metabolomics data has enormous potential for revealing insights into alterations in metabolism, but there are multiple challenges with handling technical variability and confounding. We have provided a comprehensive methodology that can handle many of these technical challenges while enabling the estimation of metabolic interactions.

Acknowledgements

This research of JTM was supported through funding by Department of Energy grant DE-SC0024320. This research was supported, in part, by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (ZIC ES103363).

Author contributions

JM and HL contributed equally to the theory and methodology described in this paper. All numerical works and computations were conducted by HL, who developed MetVAE pipeline in Python that is freely and publicly available.

Competing interests

Authors declare no conflicts of interest.

Tables

Table 1. Demographic and Clinical Features Adjusted in the MetVAE Model for IBD Data

	Non-IBD (N=19)	CD (N=31)	UC (N=20)	Overall (N=70)	P-value
Antibiotic usage (%)					
No	19 (100)	24 (77.4)	18 (90.0)	61 (87.1)	0.071
Yes	0 (0)	7 (22.6)	2 (10.0)	9 (12.9)	
Gender (%)					
Female	9 (47.4)	17 (54.8)	15 (75.0)	41 (58.6)	0.185
Male	10 (52.6)	14 (45.2)	5 (25.0)	29 (41.4)	
Age					
Mean (SD)	30.7 (20.8)	30.4 (18.8)	30.2 (17.3)	30.4 (18.7)	0.948
Median [Min, Max]	28.0 [6.00, 69.0]	24.0 [8.00, 76.0]	29.0 [7.00, 76.0]	27.0 [6.00, 76.0]	

For numeric variables, p-values were obtained using a standard two-sample t-test (two-sided). For categorical variables, p-values were obtained using a chi-squared test of independence.

Figures

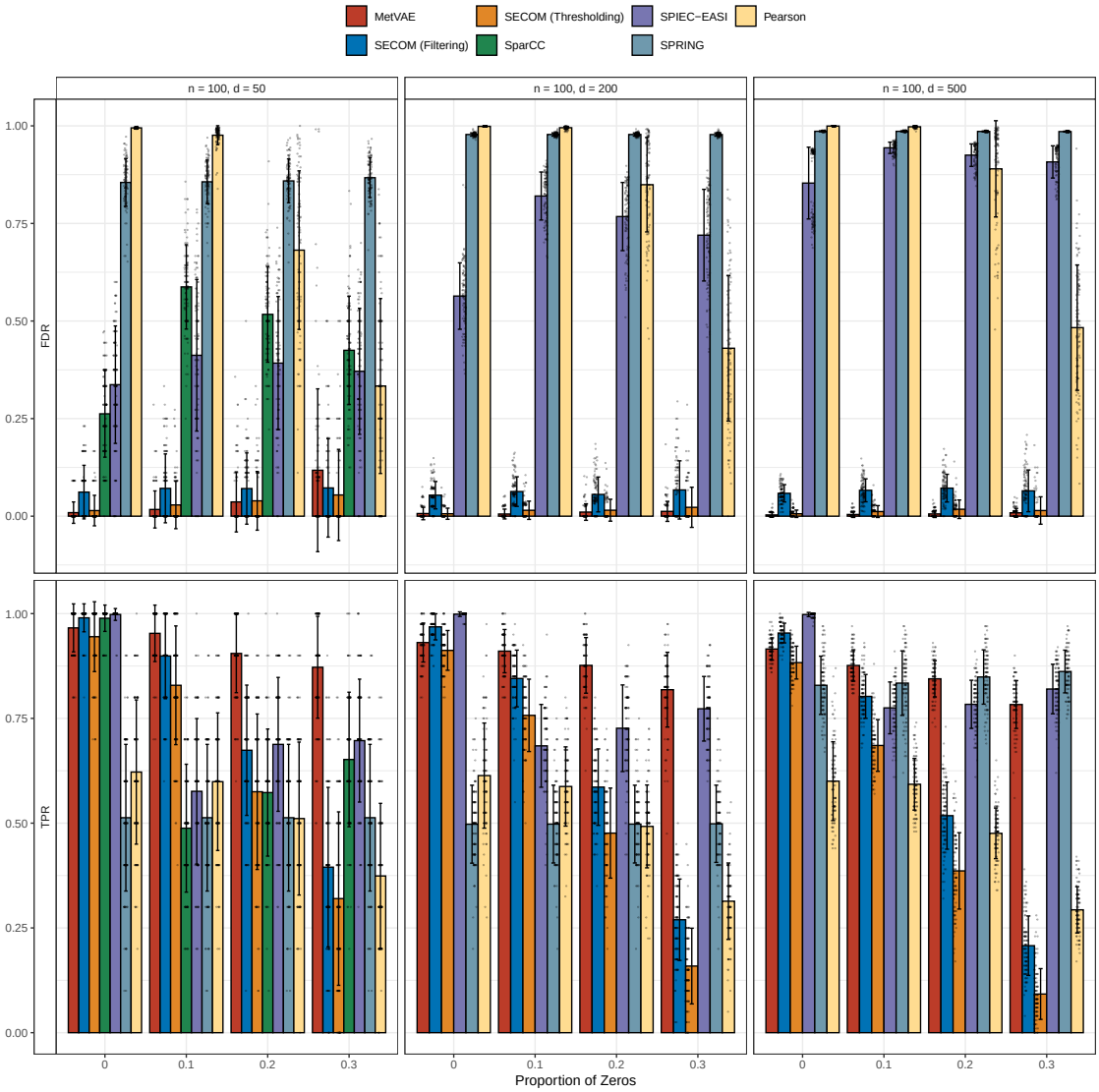


Figure 1. Comparison of methods for detecting correlations in the absence of confounders. True absolute abundance data were generated from log-normal distributions, with observed abundances derived by adding sample-specific sampling fractions and taxon-specific sequencing efficiencies to the absolute abundances. The columns across the figure represent different simulation settings defined by combinations of sample size (N) and number of features (D). The x-axis indicates the proportion of missing data, ranging from 0 to 0.3, while the y-axis depicts both the false discovery rate (FDR) and the true positive rate (TPR), which are illustrated in the upper and lower panels of the figure, respectively. The results are summarized as the mean values of FDR and TPR, accompanied by standard deviations represented as error bars, derived from 100 simulation iterations for each specific N/D configuration. Data points are overlaid on the bar charts as dots with jittering effects. The color scheme and labels identifying each method are displayed at the top of the graph.

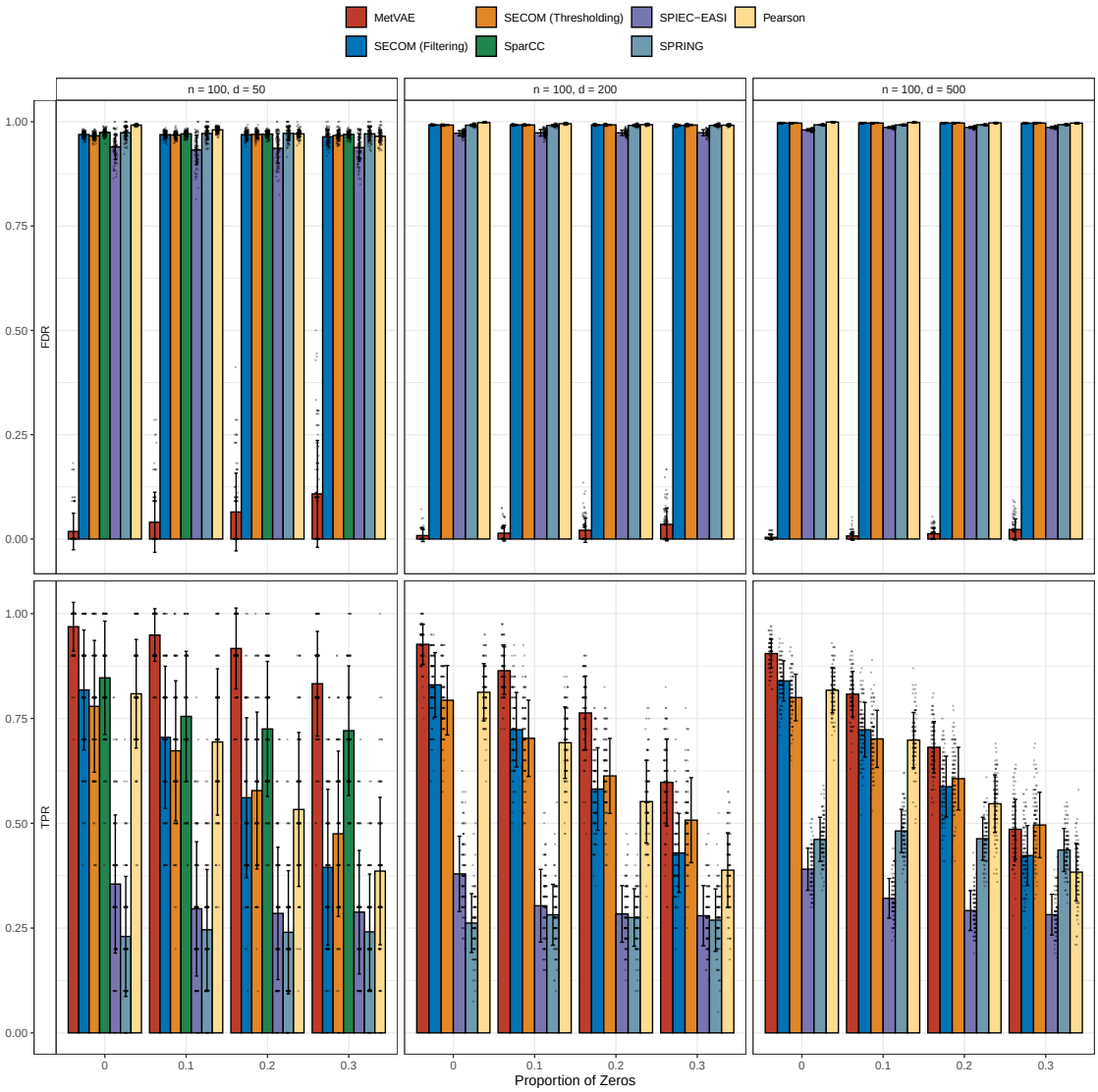


Figure 2. Comparison of methods for detecting correlations in the presence of confounders. True absolute abundance data were generated from log-normal distributions, incorporating effect sizes of one continuous and one binary confounder into the abundance profiles of differentially abundant metabolites. The observed abundances were derived by adding sample-specific sampling fractions and taxon-specific sequencing efficiencies to the absolute abundances. The columns across the figure represent different simulation settings defined by combinations of sample size (N) and number of features (D). The x-axis indicates the proportion of missing data, ranging from 0 to 0.3, while the y-axis depicts both the false discovery rate (FDR) and the true positive rate (TPR), which are illustrated in the upper and lower panels of the figure, respectively. The results are summarized as the mean values of FDR and TPR, accompanied by standard deviations represented as error bars, derived from 100 simulation iterations for each specific N/D configuration. Data points are overlaid on the bar charts as dots with jittering effects. The color scheme and labels identifying each method are displayed at the top of the graph.

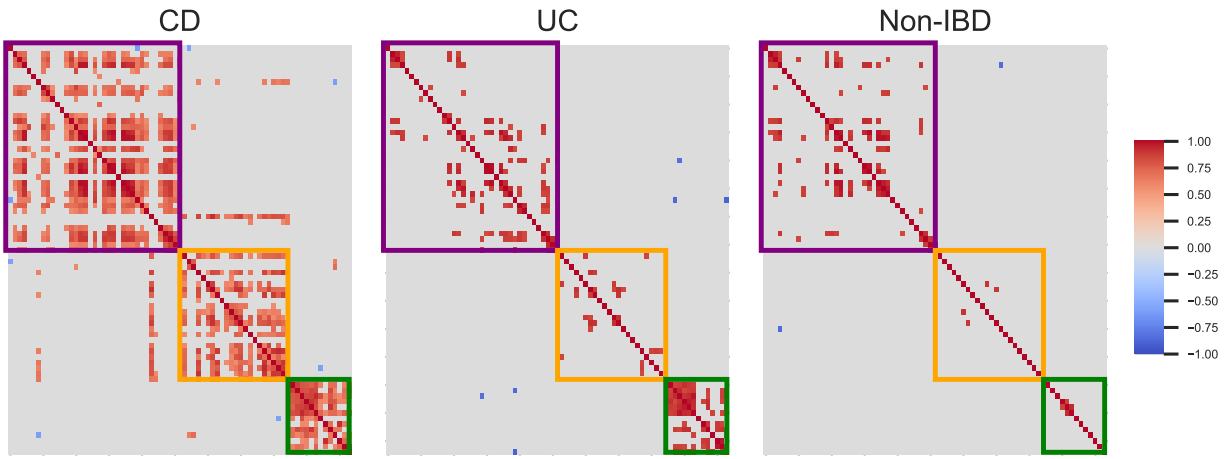


Figure 3. Heterogeneous metabolite correlations across disease states identified by MetVAE. Heatmaps displaying the correlation matrices of metabolites identified by MetVAE for CD (first column), UC (second column), and non-IBD controls (third column), with correlation values ranging from -1 to 1 (blue indicates negative correlations, and red indicates positive correlations). Three clusters of metabolites are highlighted: (1) Purple Box and (2) Orange Box: Metabolites in these clusters display strong positive correlations specifically within CD patients, with minimal correlations observed in UC or non-IBD controls. (3) Green Box: Metabolites in this cluster show strong positive correlations in both CD and UC patients, but these correlations are absent in non-IBD controls.

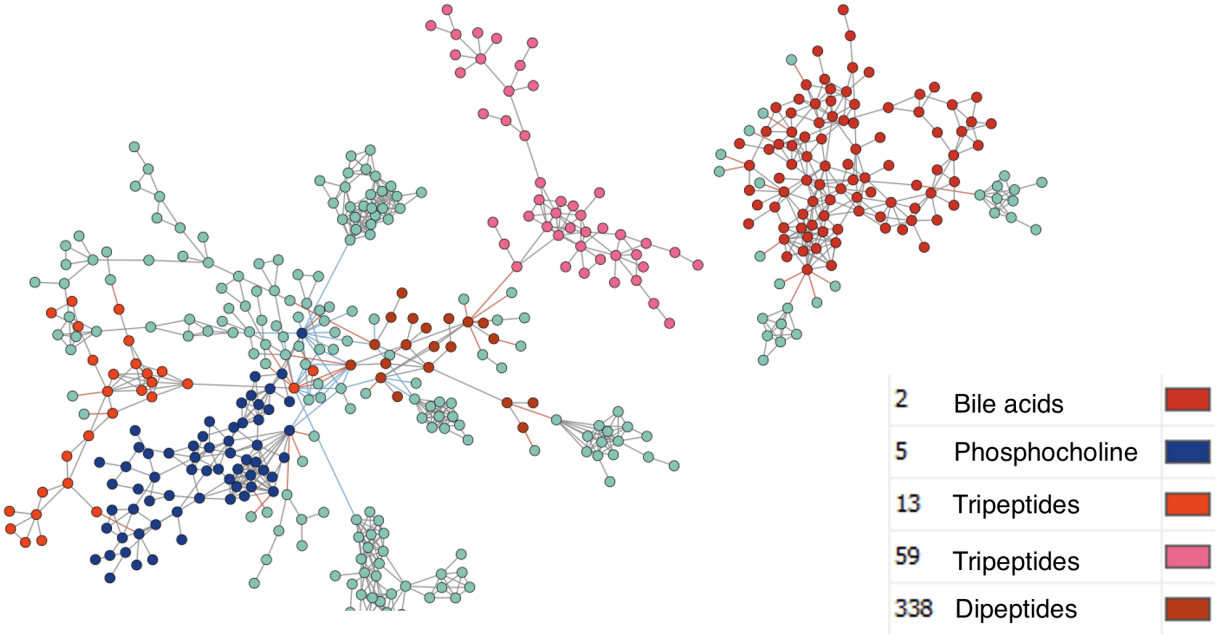


Figure 4. A molecular network focused on five subgraphs. Gray edges represent cosine similarity between molecules greater than 0.7. Red edges represent pairs of molecules with correlation greater than 0.9. Blue edges represent pairs of molecules with correlation less than -0.7.

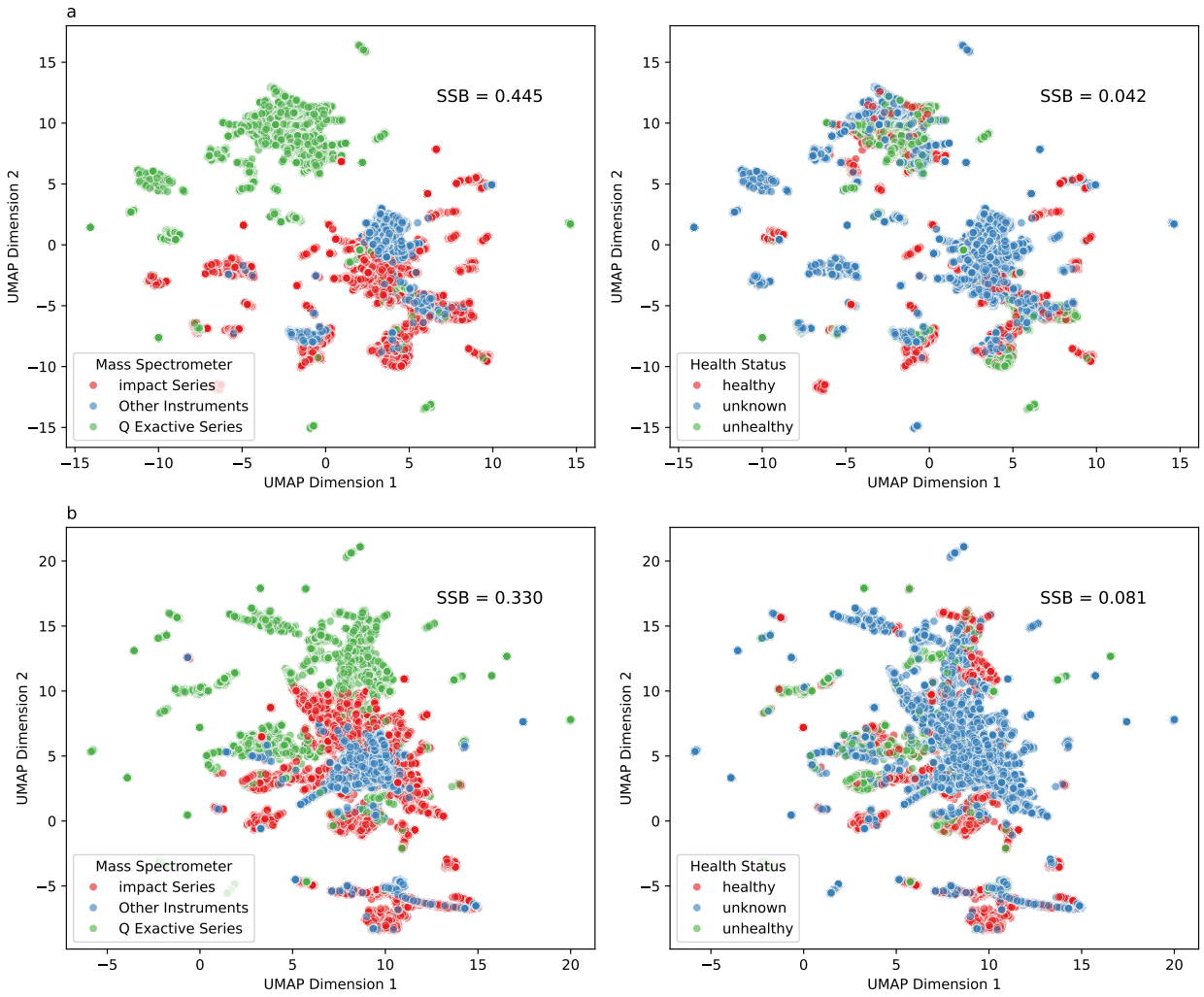


Figure 5. UMAP visualization of raw and imputed data reveals clustering patterns by mass spectrometer type and health status. **(a)** UMAP using raw data, with samples colored by mass spectrometer type (left) and health status (right). **(b)** UMAP using imputed data by MetVAE, with samples similarly colored. The x-axis represents UMAP Dimension 1 and the y-axis represents UMAP Dimension 2 in both subfigures. The normalized between-group sum of squares (SSB), which is defined as the SSB to total sum of squares (SST) to total sum of squares (SST), is displayed in each panel.

References

1. Beger, R. D. A review of applications of metabolomics in cancer. *Metabolites* **3**, 552–574 (2013).
2. Smolinska, A., Blanchet, L., Buydens, L. M. & Wijmenga, S. S. Nmr and pattern recognition methods in metabolomics: from data acquisition to biomarker discovery: a review. *Anal. chimica acta* **750**, 82–97 (2012).
3. Koek, M. M., Jellema, R. H., van der Greef, J., Tas, A. C. & Hankemeier, T. Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives. *Metabolomics* **7**, 307–328 (2011).
4. Fenaille, F., Saint-Hilaire, P. B., Rousseau, K. & Junot, C. Data acquisition workflows in liquid chromatography coupled to high resolution mass spectrometry-based metabolomics: Where do we stand? *J. Chromatogr. a* **1526**, 1–12 (2017).
5. Schrimpe-Rutledge, A. C., Codreanu, S. G., Sherrod, S. D. & McLean, J. A. Untargeted metabolomics strategies—challenges and emerging directions. *J. Am. Soc. for Mass Spectrom.* **27**, 1897–1905 (2016).
6. Gowda, G. N. *et al.* Metabolomics-based methods for early disease diagnostics. *Expert. review molecular diagnostics* **8**, 617–633 (2008).
7. Yin, P. & Xu, G. Current state-of-the-art of nontargeted metabolomics based on liquid chromatography–mass spectrometry with special emphasis in clinical applications. *J. Chromatogr. A* **1374**, 1–13 (2014).
8. Amara, A. *et al.* Networks and graphs discovery in metabolomics data analysis and interpretation. *Front. Mol. Biosci.* **9**, 841373 (2022).
9. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat. biotechnology* **34**, 828–837 (2016).
10. Liu, W. *et al.* Metabolic interactions and differences between coronary heart disease and diabetes mellitus: a pilot study on biomarker determination and pathogenesis. *J. Proteome Res.* **20**, 2364–2373 (2021).
11. Faubert, B., Solmonson, A. & DeBerardinis, R. J. Metabolic reprogramming and cancer progression. *Science* **368**, eaaw5473 (2020).
12. Kalivodová, A. *et al.* Pls-da for compositional data with application to metabolomics. *J. Chemom.* **29**, 21–28 (2015).
13. Lin, H. & Peddada, S. D. Analysis of compositions of microbiomes with bias correction. *Nat. communications* **11**, 3514 (2020).
14. Morton, J. T. *et al.* Establishing microbial composition measurement standards with reference frames. *Nat. communications* **10**, 2719 (2019).
15. Mandal, S. *et al.* Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. ecology health disease* **26**, 27663 (2015).
16. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: and this is not optional. *Front. microbiology* **8**, 294209 (2017).
17. Morton, J. T. *et al.* Balance trees reveal microbial niche differentiation. *MSystems* **2**, e00162–16 (2017).
18. Lin, H. & Peddada, S. D. Multigroup analysis of compositions of microbiomes with covariate adjustments and repeated measures. *Nat. Methods* **21**, 83–91 (2024).
19. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLOS computational biology* **8**, e1002687 (2012).
20. Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S. & Bähler, J. Proportionality: a valid alternative to correlation for relative data. *PLoS computational biology* **11**, e1004075 (2015).
21. Kurtz, Z. D. *et al.* Sparse and compositionally robust inference of microbial ecological networks. *PLoS computational biology* **11**, e1004226 (2015).
22. Morton, J. T. *et al.* Learning representations of microbe–metabolite interactions. *Nat. methods* **16**, 1306–1314 (2019).

23. Lin, H., Eggesbø, M. & Peddada, S. D. Linear and nonlinear correlation estimators unveil undescribed taxa interactions in microbiome data. *Nat. communications* **13**, 4946 (2022).

24. Pearson, K. Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia. *Philos. Transactions Royal Soc. London. Ser. A, containing papers a mathematical or physical character* 253–318 (1896).

25. Aitchison, J. The statistical analysis of compositional data. *J. Royal Stat. Soc. Ser. B (Methodological)* **44**, 139–160 (1982).

26. Yoon, G., Gaynanova, I. & Müller, C. L. Microbial networks in spring-semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data. *Front. genetics* **10**, 449195 (2019).

27. Gelman, A. & Hill, J. *Data analysis using regression and multilevel/hierarchical models* (Cambridge university press, 2006).

28. Little, R. J. & Rubin, D. B. *Statistical analysis with missing data*, vol. 793 (John Wiley & Sons, 2019).

29. Wei, R. *et al.* Missing value imputation approach for mass spectrometry-based metabolomics data. *Sci. reports* **8**, 663 (2018).

30. Karpievitch, Y. V., Dabney, A. R. & Smith, R. D. Normalization and missing value imputation for label-free lc-ms analysis. *BMC bioinformatics* **13**, 1–9 (2012).

31. Meinshausen, N. & Bühlmann, P. High-dimensional graphs and variable selection with the Lasso. *The Annals Stat.* **34**, 1436 – 1462, DOI: 10.1214/009053606000000281 (2006).

32. Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441 (2008).

33. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).

34. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).

35. Kaplan, G. G. The global burden of ibd: from 2015 to 2025. *Nat. reviews Gastroenterol. & hepatology* **12**, 720–727 (2015).

36. Gallagher, K., Catesson, A., Griffin, J. L., Holmes, E. & Williams, H. R. Metabolomic analysis in inflammatory bowel disease: a systematic review. *J. Crohn's Colitis* **15**, 813–826 (2021).

37. Koutroumpakis, E. *et al.* Association between long-term lipid profiles and disease severity in a large cohort of patients with inflammatory bowel disease. *Dig. diseases sciences* **61**, 865–871 (2016).

38. Schirmer, M. *et al.* Linking microbial genes to plasma and stool metabolites uncovers host-microbial interactions underlying ulcerative colitis disease course. *Cell Host & Microbe* **32**, 209–226 (2024).

39. Franzosa, E. A. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. microbiology* **4**, 293–305 (2019).

40. Choi, S.-H., Sviridov, D. & Miller, Y. I. Oxidized cholesteryl esters and inflammation. *Biochimica et Biophys. Acta (BBA)-Molecular Cell Biol. Lipids* **1862**, 393–397 (2017).

41. Verma, N., Ahuja, V. & Paul, J. Profiling of abc transporters during active ulcerative colitis and in vitro effect of inflammatory modulators. *Dig. diseases sciences* **58**, 2282–2292 (2013).

42. Kwon, J., Lee, C., Heo, S., Kim, B. & Hyun, C.-K. Dss-induced colitis is associated with adipose tissue dysfunction and disrupted hepatic lipid metabolism leading to hepatosteatosis and dyslipidemia in mice. *Sci. reports* **11**, 5283 (2021).

43. Hrabovskỳ, V. *et al.* Cholesterol metabolism in active crohn's disease. *Wiener Klinische Wochenschrift* **121**, 270–275 (2009).

44. van Leuven, S. I. *et al.* Enhanced atherogenesis and altered high density lipoprotein in patients with crohn's disease. *J. lipid research* **48**, 2640–2646 (2007).

45. Fan, F. *et al.* Lipidomic profiling in inflammatory bowel disease: comparison between ulcerative colitis and crohn's disease. *Inflamm. bowel diseases* **21**, 1511–1518 (2015).

46. Agouridis, A. P., Elisaf, M. & Milionis, H. J. An overview of lipid abnormalities in patients with inflammatory bowel disease. *Annals gastroenterology: quarterly publication Hellenic Soc. Gastroenterol.* **24**, 181 (2011).

47. Maggiora, G., Vogt, M., Stumpfe, D. & Bajorath, J. Molecular similarity in medicinal chemistry: miniperspective. *J. medicinal chemistry* **57**, 3186–3204 (2014).

48. Shalapour, S. *et al.* Inflammation-induced iga+ cells dismantle anti-liver cancer immunity. *Nature* **551**, 340–345 (2017).

49. Jarmusch, A. K. *et al.* Redu: a framework to find and reanalyze public mass spectrometry data. *Nat. Methods* **17**, 901–904 (2020).

50. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

51. Scoville, E. A. *et al.* Alterations in lipid, amino acid, and energy metabolism distinguish crohn's disease from ulcerative colitis and control subjects by serum metabolomic profiling. *Metabolomics* **14**, 1–12 (2018).

52. Li, Y. *et al.* Visceral fat area is associated with a high risk for early postoperative recurrence in crohn's disease. *Color. Dis.* **17**, 225–234 (2015).

53. Bagheri, M. *et al.* Genotype-based "virtual" metabolomics in a clinical biobank identifies novel metabolite-disease associations. *Front. Genet.* **15**, 1392622 (2024).

54. Heischmann, S., Dzieciatkowska, M., Hansen, K., Leibfritz, D. & Christians, U. The immunosuppressant mycophenolic acid alters nucleotide and lipid metabolism in an intestinal cell model. *Sci. Reports* **7**, 45088 (2017).

55. Liu, H. *et al.* Untargeted serum metabolomics reveals specific metabolite abnormalities in patients with crohn's disease. *Front. Medicine* **9**, 814839 (2022).

56. Metwaly, A. *et al.* Integrated microbiota and metabolite profiles link crohn's disease to sulfur metabolism. *Nat. communications* **11**, 4322 (2020).

57. Ma, Z. *et al.* Extracellular vesicles containing mfge8 from colorectal cancer facilitate macrophage efferocytosis. *Cell Commun. Signal.* **22**, 295 (2024).

58. Dave, M. *et al.* Mesenchymal stem cells ameliorate inflammation in an experimental model of crohn's disease via the mesentery. *bioRxiv* 2023–05 (2023).

59. Wu, M.-Y. *et al.* Enhancement of efferocytosis through biased fpr2 signaling attenuates intestinal inflammation. *EMBO molecular medicine* **15**, e17815 (2023).

60. Rungoe, C. *et al.* Risk of ischaemic heart disease in patients with inflammatory bowel disease: a nationwide danish cohort study. *Gut* **62**, 689–694 (2013).

61. Zanoli, L. *et al.* Inflammation and aortic stiffness: an individual participant data meta-analysis in patients with inflammatory bowel disease. *J. Am. Hear. Assoc.* **6**, e007003 (2017).

62. Esteves, M. *et al.* Interaction of cytotoxic and cytoprotective bile acids with model membranes: Influence of the membrane composition. *Langmuir* **31**, 8901–8910 (2015).

63. Iwasaki, T. *et al.* Possible role of bent structure of methylated lithocholic acid on artificial and plasma membranes. *Membranes* **12**, 997 (2022).

64. Demiankova, M. V. *et al.* Antibacterial activity of peptide derivatives of phosphinothricin against multidrug-resistant klebsiella pneumoniae. *Molecules* **28**, 1234 (2023).

65. Lucas, J., Tucker, G., Grosse, R. B. & Norouzi, M. Don't blame the elbo! a linear vae perspective on posterior collapse. *Adv. Neural Inf. Process. Syst.* **32** (2019).

66. Morton, J. T., Silverman, J., Tikhonov, G., Lähdesmäki, H. & Bonneau, R. Scalable estimation of microbial co-occurrence networks with variational autoencoders. *bioRxiv* 2021–11 (2021).

67. Burda, Y., Grosse, R. & Salakhutdinov, R. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519* (2015).

68. Bickel, P. J. & Levina, E. Covariance regularization by thresholding. *The Annals Stat.* **36**, 2577–2604 (2008).

69. Rothman, A. J., Levina, E. & Zhu, J. Generalized thresholding of large covariance matrices. *J. Am. Stat. Assoc.* **104**, 177–186 (2009).

70. Cao, Y., Lin, W. & Li, H. Large covariance estimation for compositional data via composition-adjusted thresholding. *J. Am. Stat. Assoc.* **114**, 759–772 (2019).

71. Cui, Y., Leng, C. & Sun, D. Sparse estimation of high-dimensional correlation matrices. *Comput. Stat. & Data Analysis* **93**, 390–403 (2016).

72. Bijlsma, S. *et al.* Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. *Anal. chemistry* **78**, 567–574 (2006).

Methods

The development of the MetVAE model drew inspiration from a recent research of Lucas et al.[65], which demonstrated that a linear Variational Autoencoder (VAE) and Probabilistic Principal Component Analysis (pPCA) could converge to the same global optimum corresponding to principal component directions. This finding effectively bridges the gap between deep learning models often perceived as black-boxes and their biologically meaningful interpretations. Additionally, our work was influenced by the research conducted by Morton et al.[66]. They introduced modifications to VAE models to estimate Multinomial Logistic Normal (MLN) distributions for microbiome count data. These modifications enabled scalable estimation of microbial co-occurrence networks. Our MetVAE model introduces a novel framework for analyzing metabolic correlations: (1) Interpretability: it employs a linear VAE architecture to enhance the interpretability of neural network structures, (2) Compositionality: it applies the central log-ratio (CLR) transformation to address the compositional nature of untargeted metabolomics data, (3) Missing Data: it leverages VAE-generated reconstructions for improved imputation of missing values, and (4) Covariate/Confounder Control: it directly integrates covariates and confounders into the data generating process to mitigate their impact.

General workflow

Given a dataset of N samples (indexed by i) and D metabolites (indexed by j), we represent the log observed abundances of metabolites in sample i as a vector $\mathbf{o}_i \in \mathbb{R}^D$, where $\mathbf{o}_i = [o_{ij}]_{j=1,\ldots,D}$. Some metabolites may be unobserved due to factors like random errors, stochastic fluctuations, data preprocessing limitations, or values below the limits of quantification (LOQ). This can result in zero entries. To denote the observation status, we employ the notation $(\cdot)^{\mathrm{o}}$ for the subvector of elements in positions where the abundances are observed (non-missing), and $(\cdot)^{\mathrm{m}}$ for the subvector of elements in positions where the abundances are missing. With these definitions in place, we outline the general workflow of the MetVAE model as follows:

Missing Data Imputation
$\mathbf{y}_i^{\mathrm{o}} = \mathrm{clr}(\mathbf{o}_i^{\mathrm{o}})$ $\mathbf{y}_i = \iota\left(\mathbf{y}_i^{\mathrm{o}}\right)$
Encoder
$p(\mathbf{z}_i) = \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$ $q_{\boldsymbol{\Phi}}(\mathbf{z}_i \vert \mathbf{x}_i, \mathbf{y}_i) = \mathcal{N}\left(\mathbf{V}(\mathbf{y}_i - \mathbf{B}\mathbf{x}_i), \Omega\right)$
Decoder
$p_{\boldsymbol{\Theta}}(\mathbf{y}_i \vert \mathbf{z}_i, \mathbf{x}_i) = \mathcal{N}\left(\mathbf{W}\mathbf{z}_i + \mathbf{B}\mathbf{x}_i, \sigma^2\mathbf{I}_d\right)$
Data Reconstruction
$\mathbf{y}_i^{\mathrm{m}} \sim p_{\boldsymbol{\Theta}}(\mathbf{y}_i \vert \mathbf{z}_i, \mathbf{x}_i)$ $\mathbf{y}_i = (\mathbf{y}_i^{\mathrm{o}}, \mathbf{y}_i^{\mathrm{m}})$

where

(1) $\mathrm{clr}(\cdot)$: the CLR transformation. $\mathrm{clr}(\mathbf{x}) = \mathbf{x} - \bar{x}$ if $\mathbf{x}$ is in log scale. See Supplementary Methods for details on log-ratio transformations,

(2) $\mathbf{y} \in \mathbb{R}^D$: the CLR-transformed abundances,

(3) $\iota\left(\mathbf{y}_i^{\mathrm{o}}\right)$ is an imputation function that transforms $\mathbf{y}_i^{\mathrm{o}}$ into a complete vector $\mathbf{y}_i$ such that $\left(\iota\left(\mathbf{y}_i^{\mathrm{o}}\right)\right)^{\mathrm{o}} = \mathbf{y}_i^{\mathrm{o}}$,

(4) $\mathbf{z}_i \in \mathbb{R}^K$: the latent variables, $K \leq D$,

(5) $p(\mathbf{z}_i)$: the prior distribution of the latent variables,

(6) $\mathbf{x}_i \in \mathbb{R}^P$: the covariates/confounders,

(7) $\mathbf{B} \in \mathbb{R}^{D \times P}$: the effect sizes of covariates/confounders,

(8) $q_{\boldsymbol{\Phi}}(\mathbf{z}_i|\mathbf{x}_i, \mathbf{y}_i)$: the variational distribution represented by the encoder network, which is characterized by parameters $\boldsymbol{\Phi} = (\mathbf{V}, \boldsymbol{\Omega})$,

(9) $p_{\boldsymbol{\Theta}}(\mathbf{y}_i|\mathbf{z}_i, \mathbf{x}_i)$: The posterior distribution represented by the decoder network, with parameters $\boldsymbol{\Theta} = (\mathbf{W}, \boldsymbol{\sigma}^2)$. Here, the weight matrix $\mathbf{W}$ is also known as the scaled principal components for the CLR-transformed data. See Supplementary Methods for details.

Objective function

In the standard VAE approach, we use the evidence lower bound (ELBO) as the objective function:

$$\mathscr{L}(\boldsymbol{\Theta}, \boldsymbol{\Phi}; y, x) = \frac{1}{N} \sum_{i=1}^{N} \mathscr{L}(\boldsymbol{\Theta}, \boldsymbol{\Phi}; \mathbf{y}_i, \mathbf{x}_i),$$

where

$$\begin{aligned}
\mathscr{L}(\boldsymbol{\Theta}, \boldsymbol{\Phi}; \mathbf{y}_i, \mathbf{x}_i) &= \mathbb{E}_{\mathbf{z}_i \sim q_{\boldsymbol{\Phi}}(\mathbf{z}_i|\mathbf{x}_i, \mathbf{y}_i)} \left[\ln \frac{p_{\boldsymbol{\Theta}}(\mathbf{y}_i|\mathbf{z}_i, \mathbf{x}_i) p(\mathbf{z}_i)}{q_{\boldsymbol{\Phi}}(\mathbf{z}_i|\mathbf{x}_i, \mathbf{y}_i)} \right] \\
&= \mathbb{E}_{\mathbf{z}_i \sim q_{\boldsymbol{\Phi}}(\mathbf{z}_i|\mathbf{x}_i, \mathbf{y}_i)} \left[\ln \frac{p_{\boldsymbol{\Theta}}(\mathbf{y}_i, \mathbf{z}_i|\mathbf{x}_i)}{q_{\boldsymbol{\Phi}}(\mathbf{z}_i|\mathbf{x}_i, \mathbf{y}_i)} \right] \\
&\leq \ln \mathbb{E}_{\mathbf{z}_i \sim q_{\boldsymbol{\Phi}}(\mathbf{z}_i|\mathbf{x}_i, \mathbf{y}_i)} \left[\frac{p_{\boldsymbol{\Theta}}(\mathbf{y}_i, \mathbf{z}_i|\mathbf{x}_i)}{q_{\boldsymbol{\Phi}}(\mathbf{z}_i|\mathbf{x}_i, \mathbf{y}_i)} \right] \quad \text{by Jensen's Inequality} \\
&= \ln p_{\theta}(\mathbf{y}_i \mid \mathbf{x}_i).
\end{aligned}$$

Alternatively, for a tighter bound on the log-likelihood, one can also use the Importance Weighted Autoencoder (IWAE)[67]. While sharing the same architecture as a VAE, the IWAE introduces multiple samples during the optimization process, leading to an improved tightness of the bound. The objective function is:

$$\mathscr{L}_M(\boldsymbol{\Theta}, \boldsymbol{\Phi}; y, x) = \sum_{i=1}^{N} \mathbb{E} \left[\log \frac{1}{M} \sum_{m=1}^{M} w_{mi} \right],$$

where M refers to the number of importance samples used to estimate the variational lower bound, and for $m \leq M$ and $i \leq N$,

$$w_{mi} = \frac{p_{\boldsymbol{\Theta}}(\mathbf{y}_{mi} \mid \mathbf{z}_{mi}, \mathbf{x}_i)\, p(\mathbf{z}_{mi})}{q_{\boldsymbol{\Phi}}(\mathbf{z}_{mi} \mid \mathbf{x}_i, \mathbf{y}_{mi})},$$

and $(\mathbf{z}_{1i}, \mathbf{y}_{1i}), \ldots, (\mathbf{z}_{Mi}, \mathbf{y}_{Mi})$ are M i.i.d. samples from $q_{\boldsymbol{\Phi}}(\mathbf{z}_i \mid \mathbf{y}_i, \mathbf{x}_i)\, p_{\boldsymbol{\Theta}}(\mathbf{y}_i \mid \mathbf{z}_i, \mathbf{x}_i)$.

Missing data imputation

During the training phase, at each epoch, missing values for each metabolite are imputed using randomly generated values from a censored Normal distribution. This approach diverges significantly from the method of adding pseudo-counts to the original data, which can introduce artificial correlations. A key advantage of our imputation strategy is its capacity to closely mimic the distribution of the complete underlying data, effectively reducing the occurrence of false positives. Moreover, by imputing missing values with different values in each epoch, the strategy encourages the VAE model to de-emphasize missing data, instead prioritizing the analysis of non-missing data. This approach aids the model in accurately learning the correct low-dimensional latent representations, culminating in a more accurate reconstruction of the data.

Given that the observed (relative) abundances contain zeros and that the CLR-transformation is bijectively related to the relative abundances, the observed CLR-transformed abundances exhibit missing values in the same positions as their original relative abundance counterparts. Assuming the complete CLR-transformed abundance y_{ij} adheres to a Normal distribution $\mathscr{N}(\mu_j, \sigma_j^2)$, with $\mu_j = \boldsymbol{\beta}_j^T \mathbf{x}_i$ where $\mathbf{x}_i$ includes the intercept term and any confounders. The observed

CLR-transformed abundance y_{ij} is then derived from a censored Normal distribution. Specifically, we introduce a limit of quantification (LOQ) denoted by τ, which, in practical terms, can be determined as either the minimum CLR-transformed abundance of the metabolite or a certain percentile (e.g., the 90th percentile) of its minimum. The estimation of β and σ is accomplished by maximizing the log-likelihood function using numerical methods. The detailed computational procedures are provided in the Supplementary Methods.

During the inference phase, we implement multiple imputation by utilizing the reconstructed values from the VAE decoder. Specifically, observed (non-missing) values are retained, while missing entries are imputed with their VAE-reconstructed counterparts in each iteration. Final pairwise correlation estimates are computed by averaging correlation matrices across all imputed datasets, thereby accounting for imputation uncertainty.

Model properties

We initiate our discussion with the examination of parameter identifiability within the MetVAE model, specifically $\mathbf{W}$, $\mathbf{B}$, and σ, as they pertain to the neural network architecture.

Theorem 1. *Within the linear VAE framework of the MetVAE model, the parameters* $\mathbf{W}$*,* $\mathbf{B}$*, and* σ *converge to the same global optimum as that of pPCA.*

The proof of Theorem 1 parallels the arguments presented in Lucas et al.[65] and is detailed in the Supplementary Methods. Consequently, this demonstrates that the parameters within the MetVAE model are identifiable.

Theorem 2. *Within the framework of the MetVAE model, the Variance of Log-Ratio (VLR) for each pair of metabolites, denoted as* t_{lm} *with* $l,m = 1,\ldots,M$*, can be determined by*

$$t_{lm} := \mathrm{Var}\left(\ln \frac{o_l}{o_m}\right) = \frac{1}{n}\|(\mathrm{W})_l - (\mathrm{W})_m\|_2^2,$$

where $(\mathrm{W})_j$ *represents the* j^{th} *row of matrix* W*, and* o_j *represents the random variables corresponding to the log observed abundances of metabolite* j *rather than specific observations.*

Theorem 2 illustrates that the MetVAE model, despite utilizing neural networks akin to many deep learning-based methodologies, distinguishes itself through its interpretability. The weight matrix $\mathbf{W}$ bears biological meaning; specifically, the squared row-wise Euclidean distance within $\mathbf{W}$ encapsulates the information of metabolic co-occurrence. This relationship is empirically validated using synthetic data, as depicted in Supplementary Fig. 3. The proof of Theorem 2 is provided in the Supplementary Methods.

Identifiability of the correlation

We firstly assume the following data-generating process for the log observed metabolic abundances, which is similar to the model proposed in the Analysis of Compositions of Microbiomes with Bias Correction 2 (ANCOM-BC2)[18].

Assumption 1 (Data-generating process)**.** *The log-transformed observed abundance,* o_{ij}*, can be decomposed additively into the following components::*

$$o_{ij} = s_i + c_j + a_{ij} + e_{ij},$$

where s_i *represents the sample-specific bias,* c_j *represents the feature-specific bias,* a_{ij} *is the log absolute abundance, and* e_{ij} *is the random error, with* $e_{ij} \sim_{i.i.d.} F_e$ *such that* $\mathrm{E}(e_{ij}) = 0, \mathrm{Var}(e_{ij}) = \sigma_e^2$*.*

Furthermore, it is assumed that the log absolute abundance is independent of the random error:

$$a_{ij} \perp e_{ij}.$$

Note that in the presence of covariates or confounders, the log absolute abundance, a_{ij} can be further decomposed into $a_{ij} = \mathbf{b_j}^T\mathbf{x_i} + \varepsilon_{ij}$, where $\mathbf{b_j}$ represents the effect of the covariates $\mathbf{x_i}$ on the log absolute abundance, and ε_{ij} is the residual. Consequently, the data-generating model can be readily extended as follows:

$$o_{ij} = s_i + c_j + \mathbf{b_j}^T\mathbf{x_i} + \varepsilon_{ij} + e_{ij},$$
$$\varepsilon_{ij} \perp e_{ij}.$$

To avoid redundancy, we will focus solely on the case where covariates or confounders are absent in the following discussion.

Next, we define the true covariance matrix, $\Sigma_0 = \left(\sigma^0_{lm}\right)_{D\times D}$ as follows:

$$\sigma^0_{lm} = \mathrm{Cov}(a_l, a_m),$$

where a_l, a_m are log absolute abundances for feature l and m (with $l, m \in \{1, \ldots, D\}$), respectively. We assume that Σ_0 belongs to a class of sparse covariance matrices[68].

Assumption 2 (Sparse covariance matrices)**.**

$$\mathscr{U}(q, K) = \left\{ \Sigma : \Sigma \succ 0, \max_j \sigma_{jj} \leq K, \max_l \sum_{m=1}^{D} |\sigma_{lm}|^q \leq o(D) \right\},$$

where $0 \leq q < 1$, $\Sigma \succ 0$ denotes that Σ is positive definite, and $o(\cdot)$ is the little-o notation which means "is ultimately smaller than".

This assumption delineates the covariance matrix for log absolute abundances as sparse, which is a common assumption in the estimation of high-dimensional covariance matrices[19,23,69,70].

Lastly, we make the following assumption regarding the noise-to-signal ratio.

Assumption 3 (Small noise-to-signal ratio)**.**

$$\sigma^2_e \ll \sigma^0_{jj}.$$

With all the above assumptions in place, we present the following theorem regarding the identifiability of the true correlations between the log absolute abundances.

Theorem 3. *Let the true Pearson correlation coefficient matrix, calculated based on the log absolute abundances, a_j, be denoted as $R_0 = \left(\rho^0_{lm}\right)_{D\times D}$, where:*

$$\rho^0_{lm} = \frac{\sigma^0_{lm}}{\sqrt{\sigma^0_{ll}\sigma^0_{mm}}}.$$

Additionally, define covariance matrix for the CLR-transformed abundances, y_j, and the matrix of Variance of Log-Ratios (VLRs) as $\Sigma = (\sigma_{lm})_{D\times D}$ and $T = (t_{lm})_{D\times D}$, respectively, where:

$$\sigma_{lm} = \mathrm{Cov}(y_l - y_m),$$
$$t_{lm} = \mathrm{Var}(o_l - o_m).$$

Given Assumptions 1, 2, and 3, the following estimator:

$$\rho_{lm} = \frac{\sigma_{ll} + \sigma_{mm} - t_{lm}}{2\sqrt{\sigma_{ll}\sigma_{mm}}}, \tag{1}$$

asymptotically approximates ρ^0_{lm}, i.e.,

$$\rho_{lm} \to \rho^0_{lm} \quad \text{as } D \to \infty. \tag{2}$$

Proof. Starting with the definition of VLR, we express t_{lm} as follows:

$$\begin{aligned} t_{lm} &= \mathrm{Var}(o_l - o_m) = \mathrm{Var}(c_l + a_l + e_l - c_m - a_m - e_m) = \mathrm{Var}(a_l - a_m) + \mathrm{Var}(e_l - e_m), \\ &= \sigma^0_{ll} + \sigma^0_{mm} - 2\sigma^0_{lm} + 2\sigma^2_e = \sigma^0_{ll} + \sigma^0_{mm} - 2\rho^0_{lm}\sqrt{\sigma^0_{ll}\sigma^0_{mm}} + 2\sigma^2_e. \end{aligned} \tag{3}$$

Rearranging equation (3) yields that:

$$\rho^0_{lm} = \frac{\sigma^0_{ll} + \sigma^0_{mm} + 2\sigma^2_e - t_{lm}}{2\sqrt{\sigma^0_{ll}\sigma^0_{mm}}}. \tag{4}$$

To demonstrate that ρ_{lm} asymptotically approximates ρ^0_{lm}, we begin by showing that numerator of ρ_{lm} is asymptotically indistinguishable from the numerator of ρ^0_{lm}. Based on Assumption 1 and the definition of the CLR transformation, we have:

$$\begin{aligned} o_j &= s + c_j + a_j + e_j, \\ y_j &= o_j - \frac{1}{D}\sum_{j=1}^{D} o_j = c_j - \bar{c} + a_j - \bar{a} + e_j - \bar{e}. \end{aligned}$$

Thus,

$$\begin{aligned} \mathrm{Var}(y_j) &= \mathrm{Var}(c_j - \bar{c} + a_j - \bar{a} + e_j - \bar{e}) \\ &= \underbrace{\mathrm{Var}(a_j - \bar{a})}_{①} + \underbrace{\mathrm{Var}(e_j - \bar{e})}_{②}. \end{aligned}$$

It is straightforward to verify that, for ②:

$$② = \sigma^2_e + \frac{1}{D^2}D\sigma^2_e - \frac{2}{D}\sigma^2_e = \frac{D-1}{D}\sigma^2_e \to \sigma^2_e \quad \text{as } D \to \infty. \tag{5}$$

For ①, we have:

$$① = \sigma^0_{jj} - \frac{2}{D}\sum_{j'}\sigma^0_{jj'} + \frac{1}{D^2}\sum_{j}\sigma^0_{jj} + \frac{1}{D^2}\sum_{j}\sum_{j'\neq j}\sigma^0_{jj'} = \frac{D-2}{D}\sigma^0_{jj} + \frac{1}{D^2}\sum_{j}\sigma^0_{jj} - \frac{2}{D}\sum_{j'\neq j}\sigma^0_{jj'} + \frac{1}{D^2}\sum_{j}\sum_{j'\neq j}\sigma^0_{jj'}. \tag{6}$$

Based on Assumption 2, one can show that:

$$\sum_{j'\neq j}\sigma^0_{jj'} \le \sum_{j'}\sigma^0_{jj'} \le \sum_{j'}|\sigma^0_{jj'}| \le \max_j \sum_{j'}|\sigma^0_{jj'}|^{1-q}|\sigma^0_{jj'}|^q \le \max_j \sum_{j'}(\sigma^0_{jj}\sigma^0_{j'j'})^{(1-q)/2}|\sigma^0_{jj'}|^q \le K^{1-q}o(D) = o(D).$$

Similarly, we obtain:

$$\sum_{j}\sum_{j'\neq j}\sigma^0_{jj'} \le \sum_{j}\sum_{j'}\sigma^0_{jj'} \le D\max_{j'}\sum_{j'}\sigma^0_{jj'} = o(D^2).$$

Therefore, we have:

$$① \to \sigma^0_{jj}, \quad \text{as } D \to \infty.$$

Combining ① and ②, we demonstrate that:

$$\sigma_{jj} := \mathrm{Var}(y_j) \to \sigma^0_{jj} + \sigma^2_e \quad \text{as } D \to \infty.$$

Thus, the numerator of ρ_{lm},

$$\sigma_{ll} + \sigma_{mm} - t_{lm} \to \sigma^0_{ll} + \sigma^0_{mm} + 2\sigma^2_e - t_{lm}, \quad \text{as } D \to \infty,$$

converges to the numerator of ρ^0_{lm}.

Next, we demonstration the approximation of the denominators between ρ_{lm} and ρ^0_{lm}. Based on Assumption 3, we have:

$$\sigma_{jj} \to \sigma^0_{jj} + \sigma^2_e \approx \sigma^0_{jj} \quad \text{as } D \to \infty.$$

Therefore, the denominator of ρ_{lm}:

$$2\sqrt{\sigma_{ll}\sigma_{mm}} \to 2\sqrt{(\sigma_{ll}^0+\sigma_e^2)(\sigma_{mm}^0+\sigma_e^2)} \approx 2\sqrt{\sigma_{ll}^0\sigma_{mm}^0} \quad \text{as } D\to\infty,$$

approximates the denominator of ρ_{lm}^0.

Combining all the above demonstrations, we have:

$$\rho_{lm} \approx \rho_{lm}^0, \quad \text{as } D\to\infty.$$

□

As detailed in the Supplementary Methods, VLR can be effectively calculated utilizing the CLR-transformed data. Consequently, Theorem 3 establishes a connection between the correlations among log absolute abundances and their CLR-transformed counterparts.

Remark 1. With a limited number of features, i.e., when D is not large, we suggest the following modifications. Based on equations (5) and (6), and under Assumption 2, which allows us to "ignore" all covariances, we have:

$$\sigma_{jj} = \frac{D-2}{D}\sigma_{jj}^0 + \frac{1}{D^2}\sum_j \sigma_{jj}^0 + \frac{D-1}{D}\sigma_e^2. \tag{7}$$

With a simple rearrangement of equation (7), we obtain:

$$\sigma_{jj}^0 + \sigma_e^2 = \frac{D}{D-2}\sigma_{jj} - \frac{1}{D(D-2)}\sum_j \sigma_{jj}^0 - \frac{1}{D-2}\sigma_e^2. \tag{8}$$

On the other hand, summing equation (7) across features, j, leads to:

$$\sum_j \sigma_{jj} = \frac{D-1}{D}\sum_j \sigma_{jj}^0 + (D-1)\sigma_e^2. \tag{9}$$

Substituting equation (9) into equation (8) leads to the modified estimator:

$$\sigma_{jj}^* = \frac{D}{D-2}\sigma_{jj} - \frac{1}{(D-1)(D-2)}\sum_j \sigma_{jj}, \tag{10}$$

and

$$\rho_{lm}^* = \frac{\sigma_{ll}^* + \sigma_{mm}^* - t_{lm}}{2\sqrt{\sigma_{ll}^*\sigma_{mm}^*}}. \tag{11}$$

Sparse correlation estimation

Let R_n denote the empirical correlation matrix estimated from the CLR-transformed abundances, y_j. We estimate the sparse correlation matrix using the Sparse Estimation of the Correlation matrix (SEC) algorithm[71]:

$$\hat{R} = \arg\min_R \frac{1}{2}\|R-R_n\|_F^2 + \lambda|R|_1, \text{ s.t. } R \succeq \varepsilon I, \sigma_{jj}=1, j=1,\ldots,D. \tag{12}$$

where $\|\cdot\|$ is the Frobenius norm, $R\succeq\varepsilon I$ indicates that $R-\varepsilon I$ is semipositive definite, ε is set to 10^{-5} by default, and λ is the regularization parameter for the l_1 norm.

When the distributions of CLR-transformed abundances, y_j, exhibit either exponential-type or polynomial-type tail probabilities[71], the convergence of $\hat{R}$ to R in Frobenius norm is established in the following theorem.

Theorem 4. *Denote the non-diagonal support of R as $A_0=\{l,m : l\neq m, \rho_{lm}\neq 0\}$ and its cardinality as q. If we set $\lambda = K\sqrt{\frac{\log D}{N}}$ for some constant K, then:*

$$\|\hat{R}-R\|_F^2 = O_p(q\frac{\log D}{N}).$$

The proof of this theorem largely follows the proof of Theorem 2 provided in the original SEC paper. Interested readers are encouraged to refer to that paper for detailed explanations.

For ease of computation, the objective function of the penalized likelihood in the SEC algorithm can be replaced with the general thresholding operator[69,70]:

$$\hat{R} = (\hat{\rho}_{lm})_{D\times D} \quad \text{with } \hat{\rho}_{lm} = S_\lambda(\rho_{lm}),$$

where the general thresholding functions $S_\lambda(\cdot)$ satisfies the following conditions:

1 $S_\lambda(z) = 0$ for $|z| \leq \lambda$,

2 $|S_\lambda(z) - z| \leq \lambda$ for all $z \in \mathbb{R}$.

The tuning parameter λ can be selected by cross-validation, specifically by minimizing the Frobenius norm: $\lambda = \min_\lambda \|\hat{R} - R_n\|_F$.

Simulation benchmark

In our study, we investigated three sample-to-feature (N/D) ratios to address varying dimensional contexts: a low-dimensional case with $N = 100, D = 50$, a high-dimensional scenario with $N = 100, D = 200$, and an even more extreme high-dimensional case with $N = 100, D = 500$. The proportion of missing values was explored from 0% (complete data) up to 30%, acknowledging that 20% is often the maximum threshold for discarding samples based on missing values during data pre-processing in metabolomics studies[72]. We simulated metabolic abundances using a log-Normal distribution, specifying that $0.2D$ pairs of metabolites (i.e., $0.4D$ individual metabolites) exhibit true correlations. Our analysis encompassed two scenarios: one without confounders and another with the inclusion of confounders. For the scenarios involving confounders, we incorporated one continuous and one binary confounders, affecting 10% of the metabolites ($0.1D$), rendering these metabolites differentially abundant. The rest of the metabolites remained unaffected by these confounders.

IBDMDB analysis

We stratified the data based on disease status (CD, UC, and non-IBD) and fitted the MetVAE model separately for each group, adjusting for age, gender, and antibiotic usage. During data preprocessing, we retained only annotated metabolites and, for duplicated metabolites due to different ion modes, we selected the one with the lowest pooled QC sample coefficient of variation (CV). We retained metabolites with less than 50% missingness across samples. After preprocessing, we identified 422 annotated and 17 unannotated metabolites common to all disease groups. Important demographic and clinical features, such as age, gender, and antibiotic usage, were adjusted for in the MetVAE model. Summary statistics of these covariates are provided in Table 1. We implemented a thresholding method (equation (12)) to obtain the final sparse correlation matrix for each disease group.

HCC analysis

The original study included 203 HCC mice and 235 non-HCC controls. Metabolomic data were collected using reverse-phase ultra-high-pressure liquid chromatography–mass spectrometry (UHPLC-MS), yielding 7,713 MS1 features and 2,122 MS2 features. For this re-analysis, we focused on HCC samples and metabolites with MS2 features. After excluding metabolites with excessive zeros (>50% across all samples), the dataset comprised 203 HCC samples and 2,084 metabolites.

ReDU analysis

The dataset snapshot used in this analysis consists of 40,919 samples and 15,658 features. Initially, we trimmed the dataset by requiring features to be present in a minimum of 60 samples and each sample to contain at least 60 nonzero features. This filtering resulted in a reduced dataset of 20,330 samples and 3,929 features. We ran MetVAE on this reduced dataset using an NVIDIA A100 Tensor Core GPU with 16 CPUs on the University of Maryland's Zaratan High-Performance Computing (HPC) cluster. The job completed in approximately 40 hours.

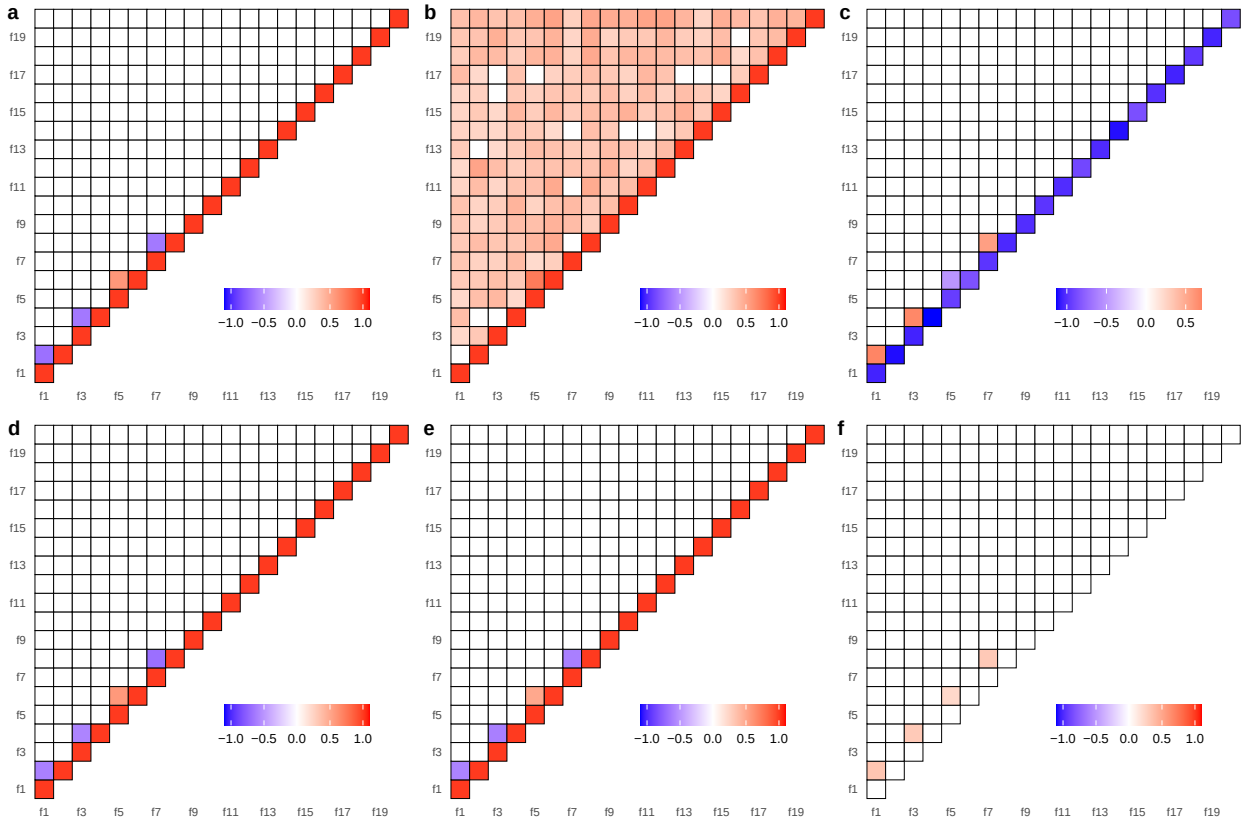
Data availability

The IBD metabolomics data were sourced from the Inflammatory Bowel Disease Multi'omics Database (IBDMDB) https://ibdmdb.org/. The Reanalysis of Data User (ReDU) data are not publicly available but may be obtained by contacting the corresponding author Dr. James Morton.

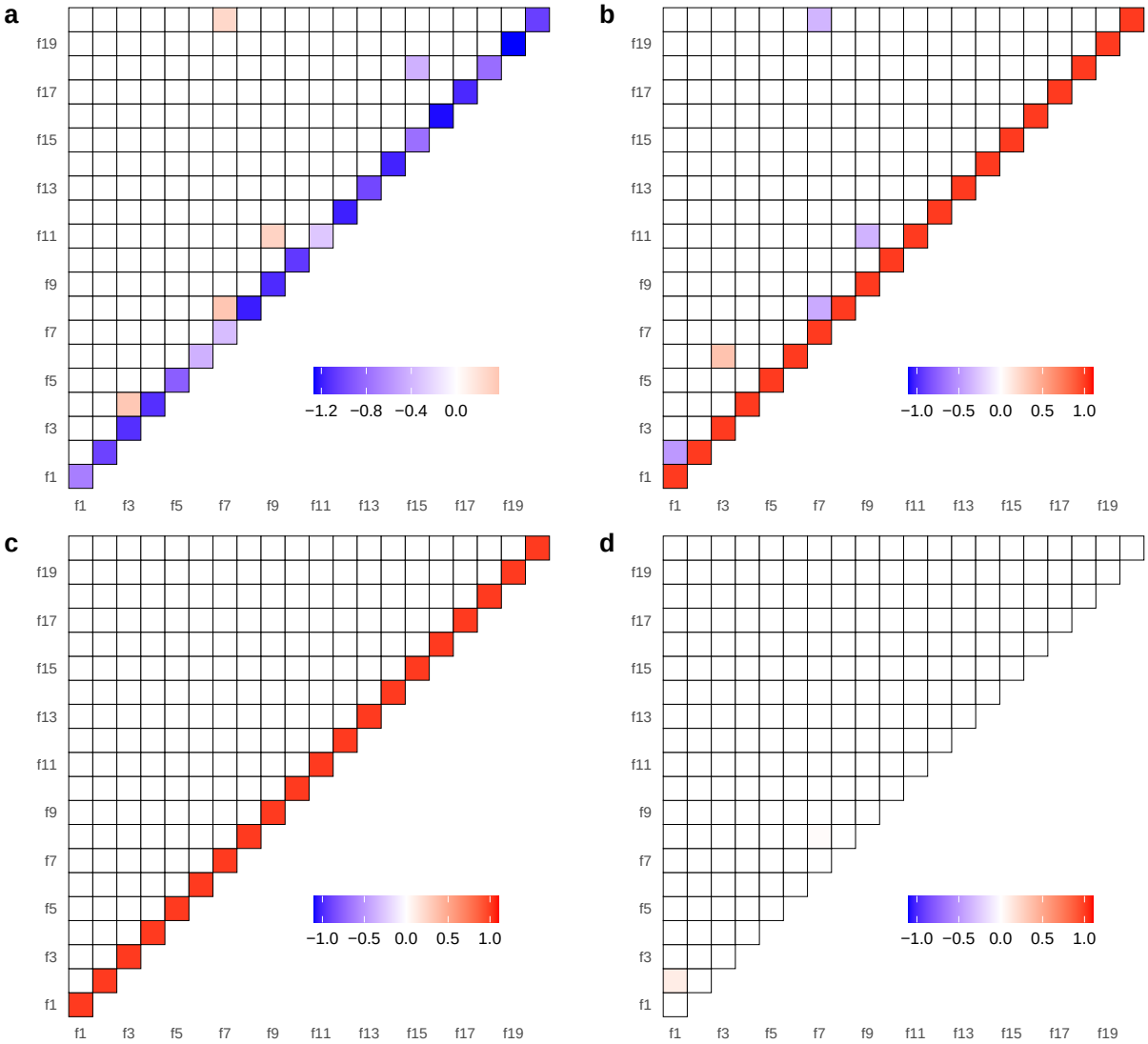
Code availability

MetVAE has been developed as a Python package and is available on PyPI at https://pypi.org/project/metvae/. The complete code used for all analyses can be accessed in the associated GitHub repository: https://github.com/FrederickHuangLin/MetVAE-Code-Archive/tree/main.

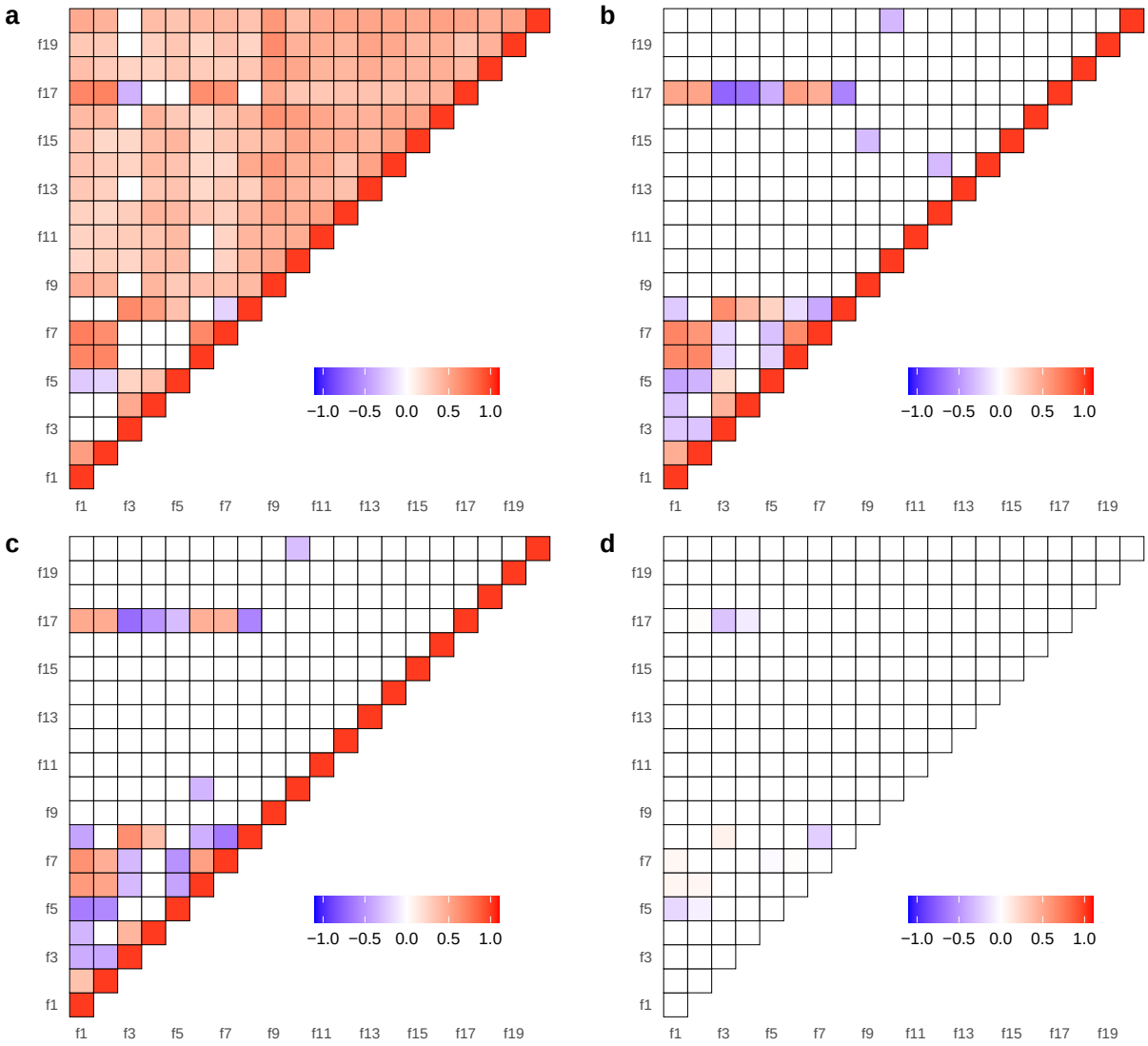
Extended data figures



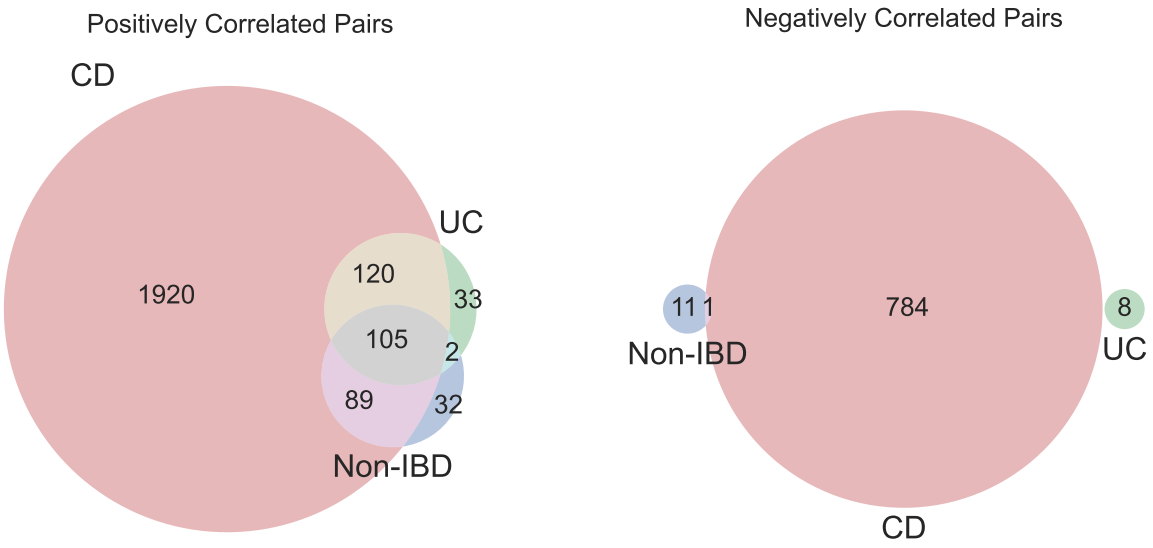
Extended Data Figure 1. The impact of compositionality on different methods for identifying true correlations. True absolute abundance data were generated from log-normal distributions, with observed abundances derived by adding sample-specific sampling fractions and taxon-specific sequencing efficiencies to the absolute abundances. **(a)** The underlying true Pearson correlation coefficients based on absolute abundances among 20 simulated features, highlighting 4 truly correlated pairs and the remainder as uncorrelated. **(b), (c), (d), (e), (f)** The identified correlated pairs using observed abundance data, provided by Pearson correlation coefficients, VLR, SECOM, SparCC, and SPIEC-EASI, respectively. For methods utilizing p-values (two-sided) to determine sparsity (Pearson correlation coefficients, VLR, and SparCC), the Benjamini-Hochberg (B-H) procedure was implemented to correct for multiple comparisons, with sparsity achieved by filtering out correlations where the adjusted p-values exceeded 0.05. Correlation coefficients, ranging from -1 to 1, are color-coded from blue to red, respectively.



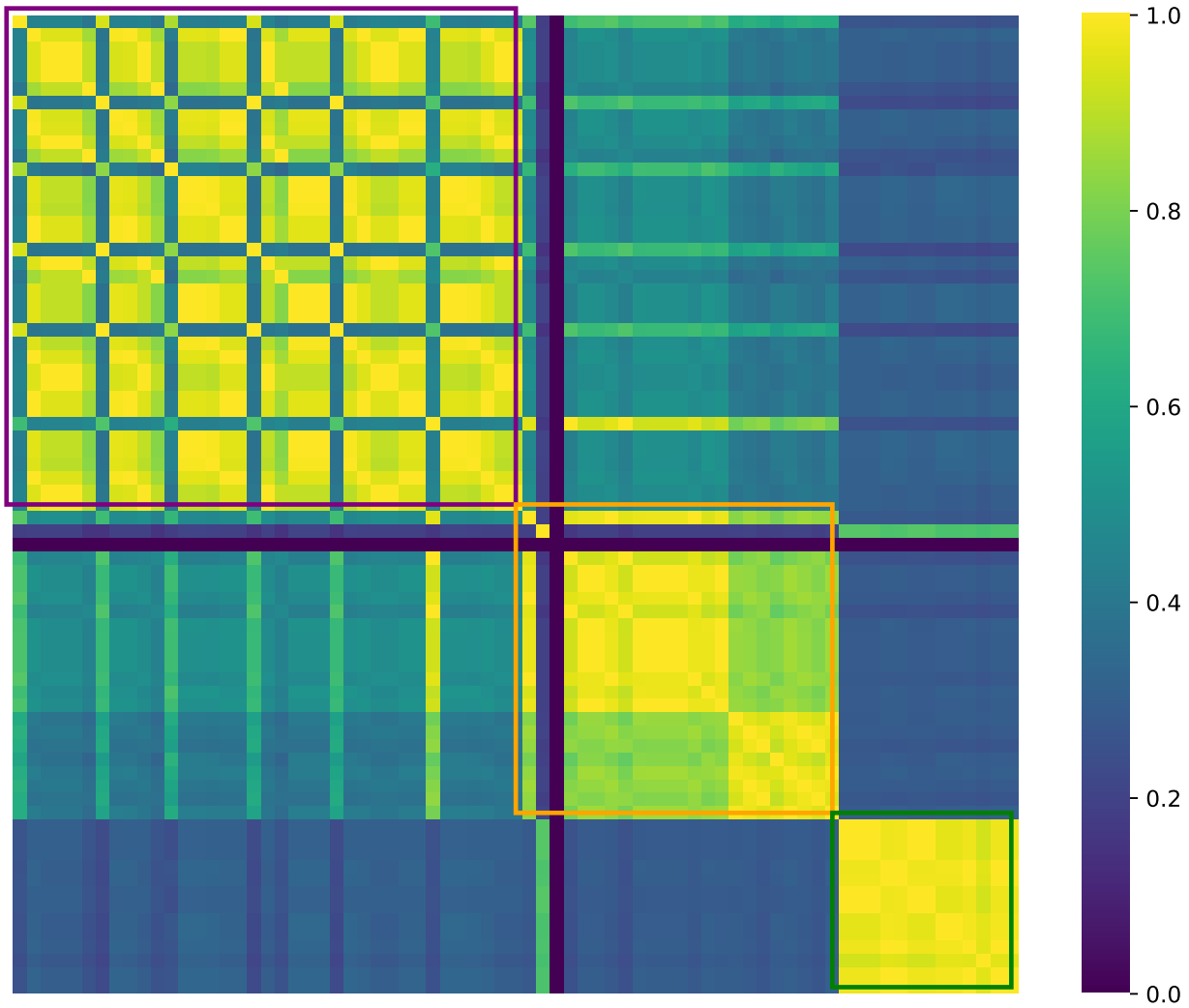
Extended Data Figure 2. The impact of missingness on different methods for identifying true correlations. True absolute abundance data were generated from log-normal distributions, with observed abundances derived by adding sample-specific sampling fractions and taxon-specific sequencing efficiencies to the absolute abundances. To simulate a not missing at random (NMAR) scenario, observed abundances below the 30th percentile were replaced with zeros. This figure builds on the setup from Extended Data Fig. 1, featuring 4 truly correlated pairs (based on complete absolute abundances) among 20 features, with the remainder uncorrelated. **(a), (b), (c), (d)** The identified correlated pairs using observed abundance data, provided by Pearson correlation coefficients, VLR, SECOM, SparCC, and SPIEC-EASI, respectively. For methods utilizing p-values (two-sided) to determine sparsity (Pearson correlation coefficients, VLR, and SparCC), the Benjamini-Hochberg (B-H) procedure was implemented to correct for multiple comparisons, with sparsity achieved by filtering out correlations where the adjusted p-values exceeded 0.05. Correlation coefficients, ranging from -1 to 1, are color-coded from blue to red, respectively.



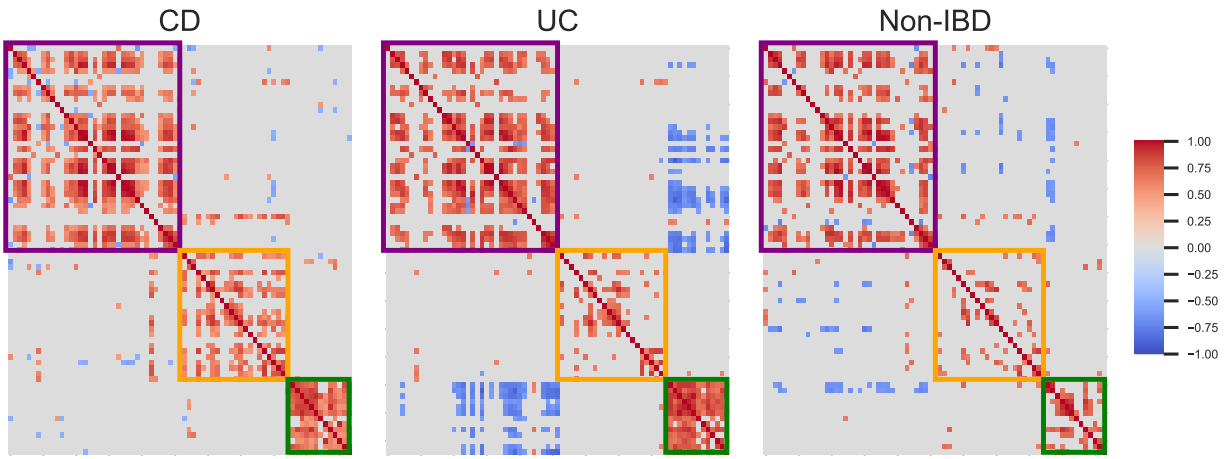
Extended Data Figure 3. The impact of confounders on different methods for identifying true correlations. True absolute abundance data were generated from log-normal distributions, with effect sizes from one continuous and one binary confounder incorporated into the abundance profiles of differentially abundant metabolites. Observed abundances were then derived by adding sample-specific sampling fractions and taxon-specific sequencing efficiencies to the absolute abundances. This figure builds on the setup from Extended Data Fig. 1, featuring 4 truly correlated pairs (determined from the unconfounded absolute abundances) among 20 features, with the remainder uncorrelated. **(a), (b), (c), (d)** The identified correlated pairs using observed abundance data, provided by Pearson correlation coefficients, VLR, SECOM, SparCC, and SPIEC-EASI, respectively. For methods utilizing p-values (two-sided) to determine sparsity (Pearson correlation coefficients, VLR, and SparCC), the Benjamini-Hochberg (B-H) procedure was implemented to correct for multiple comparisons, with sparsity achieved by filtering out correlations where the adjusted p-values exceeded 0.05. Correlation coefficients, ranging from -1 to 1, are color-coded from blue to red, respectively.



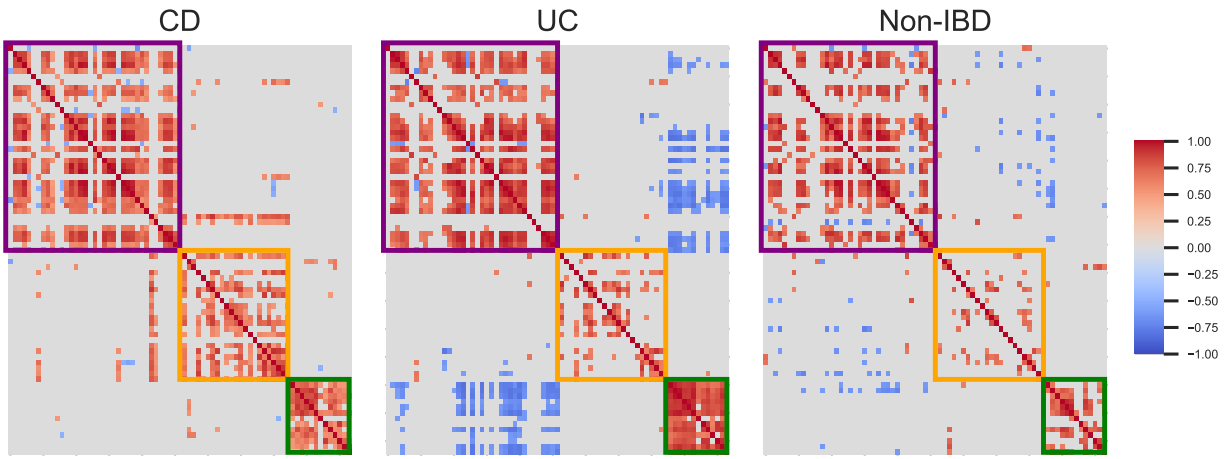
Extended Data Figure 4. Overlap of Correlated Metabolites Across Disease Statuses in IBD.Venn diagram showing the number of positively (left panel) and negatively (right panel) correlated metabolite pairs identified in CD, UC, and non-IBD controls.



Extended Data Figure 5. Tanimoto similarities of metabolites of interest across disease states. Heatmaps displaying the pairwise Tanimoto similarities of metabolites, with correlation values ranging from 0 (blue) to 1 (yellow). Three clusters of metabolites are highlighted: (1) Purple Box and (2) Orange Box: Metabolites in these clusters display strong positive correlations specifically within CD patients, with minimal correlations observed in UC or non-IBD controls. (3) Green Box: Metabolites in this cluster show strong positive correlations in both CD and UC patients, but these correlations are absent in non-IBD controls. Overall, metabolites display strong within-cluster similarities and distinct structural differences between clusters.



Extended Data Figure 6. Pearson correlations across disease states on CLR-transformed data. Heatmaps displaying the Pearson correlation matrices of metabolites for CD (left panel), UC (middle panel), and non-IBD controls (right panel), with correlation values ranging from -1 to 1 (blue indicates negative correlations, and red indicates positive correlations). The three clusters identified by MetVAE are highlighted. However, compared to the MetVAE results, the differential correlation patterns are either absent (purple box) or diminished (orange and green boxes).



Extended Data Figure 7. Spearman correlations across disease states on CLR-transformed data. Heatmaps displaying the Spearman correlation matrices of metabolites for CD (left panel), UC (middle panel), and non-IBD controls (right panel), with correlation values ranging from -1 to 1 (blue indicates negative correlations, and red indicates positive correlations). The three clusters identified by MetVAE are highlighted. However, compared to the MetVAE results, the differential correlation patterns are either absent (purple box) or diminished (orange and green boxes).