Unfolding and De-confounding: Biologically meaningful causal inference from longitudinal multi-omic networks using METALICA

Daniel Ruiz-Perez,[a] Isabella Gimon,[a] Musfiqur Sazal,[a] Kalai Mathee,[b,c] Giri Narasimhan[a,c]

Bioinformatics Research Group (BioRG), Florida International University, Miami, FL 33199, USA[a]; Florida International University, Miami, FL 33199, USA[b]; Biomolecular Sciences Institute, Florida International University, Miami, FL 33199, USA[c];

Compiled December 13, 2023

This is a draft manuscript, pre-submission

Address correspondence to Giri Narasimhan, giri@fiu.edu.

ABSTRACT

A key challenge in the analysis of microbiome data is the integration of multi-omic datasets and the discovery of interactions between microbial taxa, their expressed genes, and the metabolites they consume and/or produce. In an effort to improve the state-of-the-art in inferring biologically meaningful multi-omic interactions, we sought to address some of the most fundamental issues in causal inference from longitudinal multi-omics microbiome data sets. We developed METALICA, a suite of tools and techniques that can infer interactions between microbiome entities. METALICA introduces novel *unrolling* and *de-confounding* techniques used to uncover multi-omic entities that are believed to act as confounders for some of the relationships that may be inferred using standard causal inferencing tools. The results lend support to predictions about biological models and processes by which microbial taxa interact with each other in a microbiome. The *unrolling* process helps to identify putative intermediaries (genes and/or metabolites) to explain the interactions between microbes; the *de-confounding* process identifies putative common causes that may lead to spurious relationships to be inferred. METALICA was applied to the networks inferred by existing causal discovery and network inference algorithms applied to a multi-omics data set resulting from a longitudinal study of IBD microbiomes. The most significant unrollings and de-confoundings were manually validated using the existing literature and databases.

Importance: We have developed a suite of tools and techniques capable of inferring interactions between microbiome entities. METALICAintroduces novel techniques called unrolling and de-confounding that are employed to uncover multi-omic entities considered to be confounders for some of the relationships that may be inferred using standard causal inferencing tools. To evaluate our method, we conducted tests on the Inflammatory Bowel Disease (IBD) dataset from the iHMP longitudinal study, which we pre-processed in accordance with our previous work.

KEYWORDS: Longitudinal microbiome analysis, Multi-omic integration, Causal inference, unfolding, de-confounding.

BACKGROUND

Microbiomes are communities of microbes inhabiting an environmental niche. *Metagenomics* data sets contain sequenced reads from samples of a microbial community and are used to infer a detailed abundance profile of the microbial taxa present in that community (1, 2). More recently, additional types of biological data are being generated from microbiome studies, including but not limited to:

- *Metatranscriptomics* and *Metaproteomics*, which helps survey the expression of the totality of genes and proteins in the microbial community (3);
- *Metabolomics*, which helps profile the concentrations of the entire set of small molecules (metabolites) present in the microbiome's environmental niche (4);
- *Metaresistomics*, which helps to capture the repertoire of antibiotic resistance genes present in the microbial community (5); and
- *Host transcriptomics*, which provides information about the expression levels of the host genes (6).

Such multi-omic data sets are critical for a more in-depth and functional understanding of microbial communities. They also shed light on some of the interactions between the entities in the microbiome (7). Thus, the study of microbial communities offers a powerful approach for inferring interactions within the community (8, 9), their impact on the host environment (5), and their role in disease and health (10, 11).

A major bioinformatic challenge is the "integrative" analysis of multi-omic data sets from microbiomes (12). Most multi-omic studies focus on a separate analysis of each omic data set without building a unified model (13). There have been some attempts (14, 15, 16, 17, 18) to build tools and develop techniques to facilitate an integrative analysis (19, 20). Significant advances were recently made on analyzing multi-omic longitudinal data sets by Ruiz-Perez et al. (21). Questions related to reproducibility, flexibility, interpretation, and biological validity continue to be challenges in the area of multi-omic microbiome analysis (21, 22, 23).

Deep Learning approaches for integrating multi-omics (24, 25) have also been developed, but they are either hard to interpret or limited to predicting just one of the omic profiles. Additionally, the high computational cost of deep learning further prevents these models from being useful at providing insights into the interplay between the different omic entities. *Partial Least Squares* models have also been used to facilitate this integration (26). Their limitations depend on the underlying data generation model, and are generally prone to produce spurious results when applied to high-dimensional data sets (27).

Given that microbiomes are inherently dynamic, longitudinal multi-omic data sets are important to fully understand the complex interactions that take place within these communities (28). Many attempts have been made to analyze data from longitudinal studies (17, 18, 29); however, these approaches do not attempt to study interactions between taxa. An alternative approach involves the use of dynamical systems such as the generalized Lotka-Volterra (gLV) models (30, 31). As was noted by Ruiz-Perez et al. (21), the large set of parameters in these probabilistic models diminishes their utility for use in inference.

In previous work (32, 21), we have described sophisticated methods to model and analyze data from longitudinal microbiome studies using *Dynamic Bayesian Networks* (DBNs). Our approach involved starting from next generation sequencing data and other omics measurements. Every attempt was made to ensure that the resulting networks had biologically meaningful edges and were not a result of overfitting. However, even if an edge was directed from an entity measured at a previous time point to an entity measured at a later one, it did not guarantee that it represented a true and direct causal interaction. It could be possible for the edge to be merely the result of a statistical correlation caused by an indirect causal relationship or model overfitting.

Microbiomes are complex environments with many subtle relationships. However, causal discovery relies on noisy data from error-prone technologies, and has to contend with a host of hidden confounders that may be hard or impossible to identify, let alone be measured. The jump to infer causality is a natural next step in understanding multi-omic interactions, and the lack of research in this area is striking. Most of the causal microbiome literature focuses on the causal impact of the microbiome on health or disease, but not on the causal interactions between these microorganisms (33, 34, 35, 36). This shortcoming was addressed in our previous work (10, 11). Finally, another major challenge in building true models of biological interactions lies in developing methods to validate them and in providing confidence measures.

METHODS

Overview. In this section, we have considered three network learning methods, Dynamic Bayesian Networks (DBNs) using PALM (21), TETRAD (37, 38, 39), and Tigramite (40), and applied them to a rich, multi-omics data set. We then describe *unrolling*, a novel method to extract well-supported, biologically-relevant conjectures on entities that appear to mediate complex relationships between microbes in a microbiome. Finally, we describe *de-confounding*, another novel method to identify network edges for which there is strong support for conjecturing that they are *spurious*, i.e., **not** causal. The two methods, unrolling and de-confounding constitute the heart of the METALICA (MicrobiomE Temporal AnaLysIs using CAusality) package presented here.

In what follows, we describe the experiments that were performed. We start by describing the data sets used for the experiments and the preprocessing of the data. Next we discuss the theory behind the first of the network learning methods, i.e., DBNs, and follow it up with the constraining structures used and the procedure to create a collection of DBNs with the help of PALM. This is followed by a brief description of two well-known methods, TETRAD and Tigramite, to create causal networks for the above data set. Finally, we describe the methods of *unrolling* and *de-confounding* to evaluate and compare the causal discoveries made by all the three network learning algorithms.

Data sets. To test the three proposed methods, the Inflammatory Bowel Disease (IBD) cohort from a study that included 132 individuals across five clinical centers was used (18). During a period of one year, each subject was profiled (biopsies, blood draws, and stool samples) every two weeks on average. This yielded temporal profiles for the metagenomes, metatranscriptomes, metaproteomes, metabolomes and viromes across all subjects. Additionally, for each subject, host- and microbe-targeted human RNA sequencing was yielded from biopsies collected at initial screening colonoscopy sampled from two sites in the gut (ileum and rectum) to obtain the host transcriptomic profile. All data are fully described and available at https://ibdmdb.org.

Preprocessing the data. We used the processed version of the IBD dataset generated by our previous work (21), which provided temporally aligned and unaligned versions of metagenomics, metatranscriptomics, metabolomics, and host transcriptomics data. As explained in Ruiz-Perez et al. (21), the data were normalized and centered, the time series were smoothed, and then temporally aligned. For completeness, a summary of this process is described here. The different omics data types were processed separately. First, the taxon, metabolite, and gene abundance values were normalized to make each type separately add up to 1 for each subject, thus expressing each abundance value as a fraction of the whole metagenome, metabolome, and metatranscriptome. Then, the intensities of the metabolites and genes were scaled to match the mean of the taxa because the larger number of genes and metabolites had made their average values much smaller. Metabolites without an HMDB ID or with near-zero variance over the originally sampled time points were removed. Any sample that had less than five measured time points in any of the multi-omics measurements was also removed. The multi-omic time series were then smoothed using B-splines to deal with irregular sampling rates and missing time points. Then, temporal alignment of the time series data from individuals was performed as described in Lugo-Martinez et al. (32). This was done because they assumed that even though the underlying biological process of the different subjects may be the same, the speed at which the processes occur in each patient could be different. These temporal alignments use a linear time transformation function to “warp” one time series into a common, representative sample time series used as the “reference” (32), which was selected as follows for each omics data: All possible pairwise alignments were generated between them and the time series that resulted in the least total overall error in the alignments was selected as the reference. Abnormal and noisy samples from the resulting set of alignments were filtered out. Given an individual’s warped/aligned time series for a specific omic type (represented by a transformation), the other multi-omics data were also aligned using the same transformation. The resulting data set comprised of 51 sets of multi-omics time series, one set per subject. We also further restricted ourselves to just the Crohn’s disease patients for some analyses, which after the same filtering as described above, resulted in 11 patients.

Due to the relatively small number of time points in each time series, new datasets were generated by simply increasing the sampling frequency from each smoothed time series. Thus, a time series with a sampling rate of seven days was created. The three preprocessed omics data were then separated, resulting in sets denoted by $\mathbb{T}$, $\mathbb{G}$, and $\mathbb{M}$, representing the data involving just taxa, genes, and metabolites, respectively. They were also combined to generate different subsets and denoted in a natural way by concatenating the individual symbols. The resulting datasets were the temporally aligned and unaligned versions of the following: $\{\mathbb{T}, \mathbb{G}, \mathbb{M}, \mathbb{TG}, \mathbb{TM}, \mathbb{GM}, \mathbb{TGM}\}$.

In an effort to increase the number of biologically interpretable results and to get the most significant validations of the interactions, the attributes that were cataloged in KEGG (41) were used. This resulted in the selection of 27 bacterial species, 34 genes, and 19 metabolites, in addition to one so-called “clinical” variable (sampling time, represented by the week during which the sample was obtained). The process described above is generalizable, meaning that more omics data sets, metadata, and clinical variables can be added with relative ease.

Dynamic Bayesian Networks. DBNs are a variety of Bayesian Networks (BNs) designed to represent temporal connections between variables as their edges represent lagged dependencies. DBNs can be used to conduct time-varying probabilistic inference and causal discovery. They were developed to unify models such as Kalman filters, autoregressive–moving-average models (ARIMA), and hidden Markov models (HMMs) into a general probabilistic model and inference mechanism (42, 43), and are conceptually similar to Probabilistic Boolean Networks (PBN) (44). DBNs can model the types of relationships supported by the above methods, and can capture even more complex relationships with both discrete and continuous variables conditioned on either temporal and non-temporal variables.

This work, focuses on a version of DBNs called Two-Timeslice BN (2TBN) (45), which finds relationships between variables over adjacent time steps. Let X_i^t denote the value of variable X_i at time t. It can be calculated from the internal regressors if the values of the other variables are known at the previous time point, $t-1$. We employed a tool called PALM, which uses a multi-omics DBN model proposed by Ruiz-Perez *et al.* (21). PALM integrates different omics datasets with flexible structure constraints.

In particular, we also used their proposed *Skeleton* and *Augmented* constraints. These constraints are described below in the "Constraining structures" section. Idealized DBN construction methods require an exponential-time exhaustive search using all subsets of nodes. However, it is possible to construct DBNs more efficiently by limiting the number of "parents" for each node (i.e., bounding the number of incoming edges for each node).

Constraining structures. The above input was fed into PALM (21).The set of allowable edges was constrained by providing a *Skeleton* structure as input to the DBN construction step as described by Ruiz-Perez et al. (21). These constraints, which are provided in the form of a matrix, only allow edges between certain types of nodes, greatly reducing the complexity of searching over possible structures and preventing over-fitting. Specifically, *intra* edges (i.e., edges within same time point) from taxa nodes to gene (expression) nodes and from gene nodes to metabolites (concentration) nodes were allowed. All other interactions within the same time point (for example, direct gene to taxa) were disallowed. In addition, *inter* edges (i.e., edges between nodes from adjacent time points) were only allowed from metabolites to taxa nodes in the next time point, and *self-loops*, i.e., edges from node X_i^t to X_i^{t+1} for all types of nodes. (Note that, whenever it is obvious by the context, random variables and the nodes in the networks that represent them are not differentiated.) The restrictions in the *Skeleton* reflect the basic ways the different entities interact with each other, i.e., taxa express genes that they carry on their genomes; these, in turn, are involved in metabolic pathways for the synthesis of metabolites; subsequently the metabolites impact the growth of taxa (in the next time slice).

A less constrained framework referred to as the *Augmented* skeleton was also used to produce an alternative set of networks. Unlike the original Skeleton, the Augmented framework also allows intra edges from taxa to metabolites to account for cases where noise or other issues related to gene-profiling may limit our ability to indirectly connect taxa and the metabolites they produce. All other edges from the skeleton were retained.

Computing DBNs using PALM. DBNs were learned using PALM for all subsets of the omics datasets from Section 2.2 (i.e., $\{\mathbb{T}, \mathbb{G}, \mathbb{M}, \mathbb{TG}, \mathbb{TM}, \mathbb{GM}, \mathbb{TGM}\}$), for several different number of allowable parents ($\{3, 4, 5, 6\}$), for temporally aligned and unaligned datasets, and for the Skeleton and Augmented constraint frameworks, thus resulting in a total of $7 \times 4 \times 2 \times 2 = 112$ potential DBN networks. A total of 100 networks were learned by subsampling subjects with replacement (i.e., 100 bootstrap repetitions) for each model. The networks were then combined, averaging the regression coefficient (weight) of the edges as long as they appeared in at least 10% of the repetitions. Each edge was also labeled with the bootstrap score or support (proportion of times that edge appears). Each repetition was set to run independently on a separate processor using Matlab's Parallel Computing Toolbox.

In order to explore causal inferencing, two other well-known methods (TETRAD and Tigramite) (37, 38, 39, 40) were applied on our data sets. Note that the exact same set of nodes were used as those in the two-time-slice DBN, meaning that every microbiome quantity (taxon abundance, gene expression, metabolite concentration) is represented by two nodes, one from a "previous" time instant and one from the "current" time instant. Since all the networks were on the same set of nodes, it facilitates the comparison between all three methods. We also note that TETRAD and Tigramite do not learn based on a global score such as likelihood, but rather on conditional independence tests.

Causal Networks using the TETRAD Suite. The tsGFCI (SVAR-GFCI) (46) algorithm is implemented in the TETRAD package (37, 38, 39), for which the wrapper PyCausal (47) was used. The tsGFCI algorithm is a version of tsFCI (48) and GFCI, while tsFCI is, in turn, the evolution of FCI (49). FCI is in turn a modification of PC-stable, which was designed by modifying PC, an adaptation of the SGS algorithm (50).

Algorithm tsFCI (SVAR-FCI) is based on a modified version of the FCI algorithm. Briefly, it uses the direction of time to orient interactions and enforces repeating structures for both adjacencies and orientations based on the stationarity assumption. Since the hybrid score-based GFCI is usually more accurate in finite samples than FCI, similar modifications were made in the development of tsGFCI. In this case, a greedy initial adjacency search is used, enforcing time order and repeating structures, and scores the structures using BIC (51).

For each significance threshold $\alpha \in \{0.0001, 0.001, 0.01, 0.1\}$, different networks were learned with the PositiveCorr CI test, the FisherZScore network score, and for each combination of omics datasets and alignment. A total of $4 \times 7 \times 2 = 56$ experiments were performed with TETRAD. Each TETRAD experiment was repeated with N bootstrapping repetitions. Here, $N = 10$ was used.

Causal Networks with Tigramite. For the discussion below, the following notation is needed. Let $Pa_G(X)$ represent the parents of node X in network G. When the context is clear, G is dropped and simply denoted as $Pa(X)$. Let $Pa^p(X)$ denote the p "strongest" parents. Independence of A and B conditioned on C is denoted by $A \perp\!\!\!\perp B | C$. Tigramite (40) implements the PCMCI algorithm, which works in two stages – conditional selections followed by causal discovery.

1. **Conditional selections**: A modified version of the PC-stable algorithm (adapted for time series and with the skeleton constraints) is used to compute a set of variables that are inferred to have a causal effect on each node X. It obtains the set of parents, $Pa_G(X_i)$, estimated from the data (which may be superset of the true set) for all variables $X_i, i = 1, \ldots, n$. This is achieved as follows. For every variable, the set of parents are initialized to all allowable parents. Then conditional independence tests are applied for each edge, $(X_i^{t'}, X_j^t)$, using conditioning sets of increasing size, removing the edge as soon as a test fails. (Note that, as per our constraints, $t' = t$ or $t' = t - 1$.) In each case, the null hypothesis states that the two variables at the endpoint of the edge being considered remain dependent even when conditioned on an appropriate set of size $p \geq 0$, as stated below:

$$\mathbf{H}_0 : X_i^{t'} \not\perp\!\!\!\perp X_j^t | \boldsymbol{S}, \text{ for any } \boldsymbol{S} \subseteq Pa(X_j^t) \setminus \{X_i^{t'}\} \text{ with } |\boldsymbol{S}| = p. \tag{1}$$

   The rejection of the null hypothesis $\mathbf{H}_0$ requires a significance threshold α. All possible sets $\boldsymbol{S} \subseteq Pa(X_j^t) \setminus \{X_i^{t'}\}$ with cardinality p are considered such that $1 \leq p \leq q_{max}$.

2. **Causal discovery stage**: Next the MCI algorithm is applied, which employs a more stringent conditional independence test, for each surviving edge $X_i^{t'} \rightarrow X_j^t$, retaining it if and only if

$$X_i^{t'} \not\perp\!\!\!\perp X_j^t | Pa(X_j^t) \setminus \{X_i^{t'}\} \cup Pa^p(X_j^t). \tag{2}$$

Since Tigramite assumes that all the data points belong to a single subject, bootstrap cannot be implemented in the usual way of subsampling subjects with replacement. Instead, a different network was learned for each subject, and the resulting networks were then combined. The percentage of times that a given edge appears in all the different networks was annotated in the edge, together with the averaged cross-link strength. Different networks were learned for different significance threshold values, $\alpha \in \{0.0001, 0.001, 0.01, 0.1\}$, for each CI test available (GPDC, CMIknn, ParCorr) (40), and for each omics dataset. A total of $4 \times 3 \times 7 \times 2 = 168$ experiments were performed with Tigramite.

The following sections introduce the two causal network analysis techniques in METALICA, which will be applied to the networks learned with the methods introduced in Sections 2.6 – 2.8 using DBNs, TETRAD, and Tigramite.

Unrolling. Typical algorithms for network learning and analysis fail to elucidate the actual reasons why two entities may be causally related to each other. An important challenge in microbiome analysis is to use multi-omics data to determine whether and how two taxa may be interacting with each other. The term *unrolling* is hereby introduced as the process of determining the sequential steps by which two omic entities potentially interact with each other. This is done by learning independent networks using different subsets of omics data. For example, by learning two separate networks with the $\mathbb{T}$ and the $\mathbb{TM}$ datasets, an interaction between two microbial taxa (as suggested by the former) can be surmised to be via metabolic intermediaries (as suggested by the latter).

To make this more formal, let $G_{\mathbb{X}} = (V_{\mathbb{X}}, E_{\mathbb{X}})$ represent the network learned using dataset $\mathbb{X}$, with vertex set $V_{\mathbb{X}}$ and edge set $E_{\mathbb{X}}$. Now, an explanation by unrolling occurs if the following three conditions are true:

1. There is an edge from T_i to T_j in $G_{\mathbb{T}}$, for some $T_i, T_j \in V_{\mathbb{T}}, i \neq j$.
2. There is *no* edge from T_i to T_j in the network $G_{\mathbb{TM}}$.
3. There exists some metabolite $M_x \in V_{\mathbb{TM}}$ such that edges (T_i, M_x) and (M_x, T_j) exist in $G_{\mathbb{TM}}$.

If the above three conditions are met, the interaction between the taxa T_i and T_j is inferred to be happening through an intermediary metabolite M_x, which is “produced” by T_i and “consumed” by T_j.

This process can be replicated by unrolling the edges of the network inferred from $\mathbb{T}$ with the one inferred from $\mathbb{TG}$ to discover the genes that are likely driving the interaction between the same pair of taxa. Finally, the networks, $G_{\mathbb{TG}}$ from $\mathbb{TG}$ or $G_{\mathbb{TM}}$ from $\mathbb{TM}$ can be unrolled using the more detailed network, $G_{\mathbb{TGM}}$ to find fully unrolled chains of the form $T_i \to G_y \to M_x \to T_j$ in $G_{\mathbb{TGM}}$ with the capability to simultaneously explain the edges $T_i \to T_j$ in $G_{\mathbb{T}}$, the chain $T_i \to M_x \to T_j$ in $G_{\mathbb{TM}}$, and the chain $T_i \to G_y \to T_j$ in $G_{\mathbb{TG}}$.

This step-wise unrolling is necessary to discover relationships with strong support from the data, where the network learned from $\mathbb{T}$ was unrolled in a network learned from some subset of $\{\mathbb{TG}, \mathbb{TM}, \mathbb{TGM}\}$. The number of the networks from $\{\mathbb{TG}, \mathbb{TM}, \mathbb{TGM}\}$ that support the unrolling provide a degree of confidence for that unrolling. Furthermore, the bootstrap score for each of the edges involved in the process is reported, together with an *Overall Score* that is computed as the product of the individual bootstrap scores of the two replacement edges. This unrolling approach is explained with concrete examples in the *Discussion* Section under *Uncovering unrolled biological relationships*.

De-confounding Most current causal inference techniques rely on the *causal sufficiency* assumption, which assumes that there are no hidden confounders (for any pair of variables) in the data. Confounders are variables that are either (a) unknown, (b) known but not measured, or (c) measured but not used in the analysis, but affect both the cause and the effect of at least one predicted interaction. Predictions of interactions with hidden confounders could be incorrect. The strength of a predicted interaction may be enhanced or diminished when the hidden confounder is not used in the analysis. It is also possible that the predicted interaction may introduce spurious edges when the hidden confounder is not used in the analysis.

In general, the causal sufficiency assumption may be "too strong" and may be impossible to verify, even with the availability of richer data sets that include multi-omics data, thus making this assumption a key obstacle to performing accurate causal inference (52). Going beyond the multi-omic domain, causal sufficiency is an assumption that does not strictly hold in most observational datasets, since it is difficult or impossible to include all possible explanatory variables in a study.

A recent paper by Wang and Blei (53) attempts to perform *de-confounding*, which is the process of removing the effect of *all* confounders. They introduce the concept of "substitute confounders", which attempts to account for the effect of all hidden confounders in order to arrive at unbiased estimates of causal effects. A major limitations of their method is that the de-confounded interactions are not identified, which is important for understanding the interactions. Furthermore, there may not be a one-to-one correspondence between the substitute confounder and some real confounder, meaning that one substitute confounder may be an approximation for a combination of several hidden confounders.

In this work, a different approach for the task of de-confounding interactions is taken, inspired by the unrolling approach of Section 2.9. Independent networks are iteratively learned with different subsets of data with the hope that by adding a new omics layer it would be possible to identify some of the relevant intermediate entities and the corresponding interactions. As before, $G_{\mathbb{X}} = (V_{\mathbb{X}}, E_{\mathbb{X}})$ represent the network learned using dataset $\mathbb{X}$, with vertex set $V_{\mathbb{X}}$ and edge set $E_{\mathbb{X}}$. For example, by learning a network with the $\mathbb{T}$ and $\mathbb{TM}$ datasets, interactions can be de-confounded if the following three conditions are satisfied:

1. There is an edge (T_i, T_j) in $G_{\mathbb{T}}$, i.e., $(T_i, T_j) \in E_{\mathbb{T}}$, for some $T_i, T_j \in V_{\mathbb{T}}, i \neq j$.
2. There is *no* edge from T_i to T_j in $G_{\mathbb{TM}}$, i.e., $(T_i, T_j) \notin E_{\mathbb{TM}}, i \neq j$.
3. Edges (M_x, T_i) and (M_x, T_j) exist in $G_{\mathbb{TM}}$, i.e., $(M_x, T_i), (M_x, T_j) \in E_{\mathbb{TM}}, i \neq j$, for some metabolite $M_x \in V_{\mathbb{TM}}$.

Using this method, if the above conditions are satisfied for a pair of taxa, T_i and T_j, the direction for the directed edge $(T_i, T_j) \in E_{\mathbb{T}}$ is deduced and the inferred interaction between the two taxa is spuriously introduced by the metabolite M_x acting as a confounder. The metabolite can also be inferred to impact the abundance of both taxa, T_i and T_j. One possible scenario is that the metabolite, M_x, could be an essential metabolite for both taxa, and its presence or absence from the data could make the abundance of the taxa to appear correlated.

As with metabolites, this process can be repeated by de-confounding $G_{\mathbb{T}}$ with edges from $G_{\mathbb{TG}}$ to discover genes/proteins that could confound a presumed causal connection between the taxa. In general, the networks learned using the $\mathbb{T}, \mathbb{G}$, and/or $\mathbb{M}\}$ datasets can be de-confounded by the networks learned using one or more of the datasets from $\{\mathbb{TG}, \mathbb{TM}, \mathbb{GM}, \mathbb{TGM}\}$. Similarly, networks learned using one of $\mathbb{TG}, \mathbb{TM}$, or $\mathbb{GM}\}$ datasets can be de-confounded by the networks learned using $\mathbb{TGM}$. This could lead to chains of de-confoundings, where an interaction that led to the de-confounding a relationship is itself later de-confounded.

As before, for each de-confounding discovery, the following is reported: (a) the confounded edge, (b) the de-confounder, (c) the bootstrap score for the edges involved in the discovery, (d) the overall score of the discovery computed as the product of the individual bootstrap scores of the two replacement edges, and (e) the two data sets that were used to discover the specific de-confounding. The results of the de-confounding approach is explained with examples in the *Discussion* section.

RESULTS

A large number of networks were learned with the different data subsets, the different methods, and the parameter settings, as mentioned in Sections 2.6, 2.7, 2.8, respectively for DBN, TETRAD, and Tigramite. Unrolling and de-confounding were implemented in METALICA and applied to all the resulting networks, as described in the *Methods* section. The results from the experiments are presented below.

Resulting networks Figure 1 shows the DBNs learned from the $\mathbb{T}$, $\mathbb{TM}$, $\mathbb{TG}$, and $\mathbb{TGM}$ versions of the Crohn's disease datasets without temporal alignment. The structure of the networks learned by the other tools were similar to those shown and can be found in the Supplement. Self loops were hidden in the visualization to avoid unnecessary clutter. The remarkable information gain obtained by using additional omics data sets is readily observable in Figure 1 d), with a more complete picture of the state of the whole system, thus setting the stage for biologically-relevant interpretations. The one non-omics variable (week of sample obtained), which is generically referred to as a "clinical variable" did not have any incident edges in the $\mathbb{TG}$ network, but it did in the other networks.

Tool analysis Network validation is a challenging problem because we do not have the ground truth network, which is what these methods try to approximate. In addition to analyzing the networks, the effect of the different network parameters was also explored. The heatmap in Figure 2 shows the percentage of unrolling that is effected by METALICA on the networks learned by PyCausal (TETRAD). The columns labeled $\mathbb{TGMT}$, $\mathbb{TGT}$, and $\mathbb{TMT}$ represent the proportion of taxon to taxon interactions in the network learned with $\mathbb{T}$ that got unrolled with the networks learned with $\mathbb{TGM}$, $\mathbb{TG}$, and $\mathbb{TM}$, respectively. The alpha parameter for experiments with TETRAD is the significance threshold for the conditional independence tests.

The last column shows the average overall score of each unrolling, which is defined as the product of the individual bootstrap scores of the two replacement edges. Edge bootstrap scores represent the proportion of times an edge appears in bootstrap repetitions as described earlier.

Figure 3 shows the unrolling details output by METALICA in the experiments conducted with different methods, averaged over all parameters. All values except the last column represent the proportion of taxon to taxon interactions in the network learned with $\mathbb{T}$ that got unrolled with the networks learned with $\mathbb{TGM}$, $\mathbb{TG}$, and $\mathbb{TM}$, respectively. Tigramite networks showed the highest percentage of unrolled edges with $\mathbb{TGT}$ and $\mathbb{TMT}$ when compared with the other two methods, but fell short with $\mathbb{TGM}$, where DBNs resulted in significantly higher percentage of unrolled edges. Note that applying temporal alignments to the data sets seemed to significantly improve the percentage of edges unrolled for the DBN method, especially with $\mathbb{TGMT}$, where the percentage rose from 24.7% to 78.8%. The increase was significantly lower with the other two datasets. The impact of temporal alignments on the other methods was inconsistent, where it showed both increase and decrease in the different columns. We also note that temporal alignments were used to normalize the "rates" of the underlying biological process of the different subjects.

DISCUSSION

As shown in Figure 2, as the alpha parameter decreases, the proportion of edges unrolled by METALICA decreases substantially. The smaller the alpha, the easier it is for two variables to be dependent, resulting in networks with more edges. This also

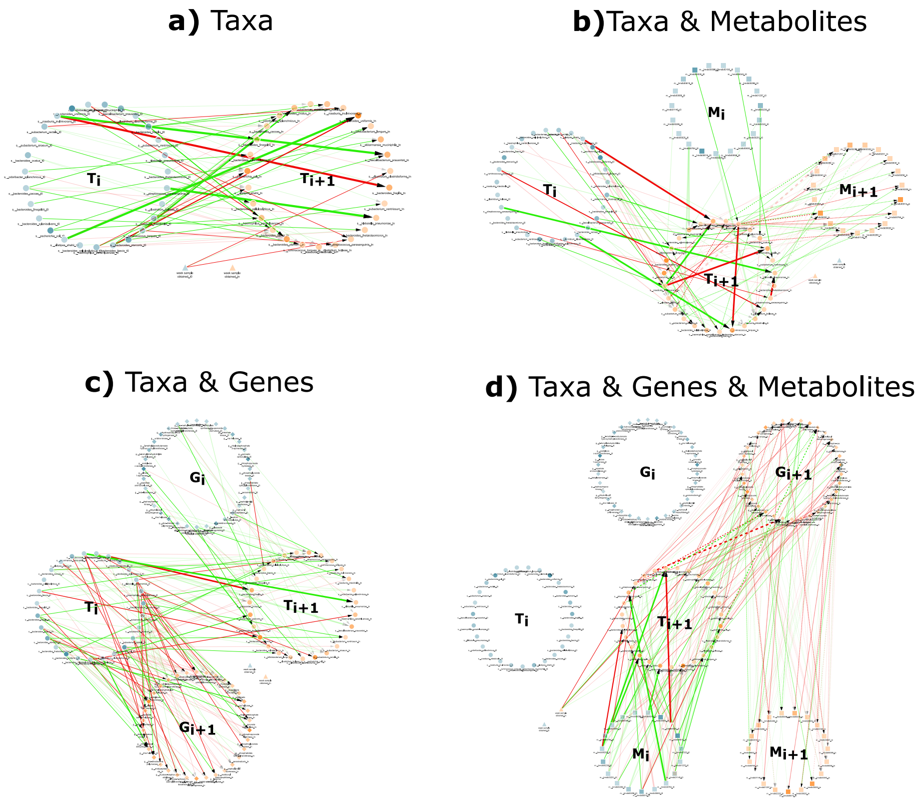
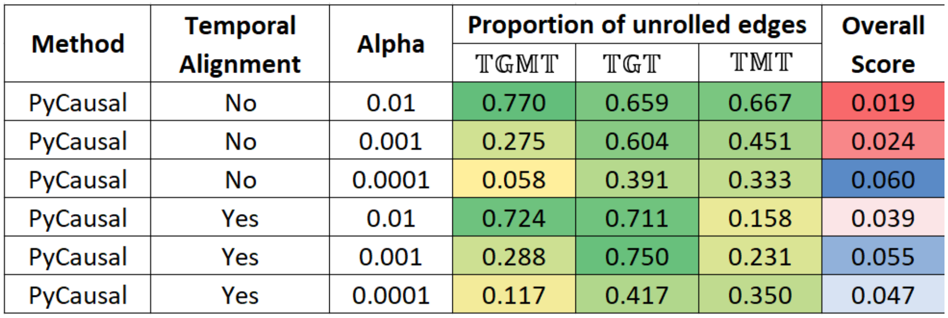
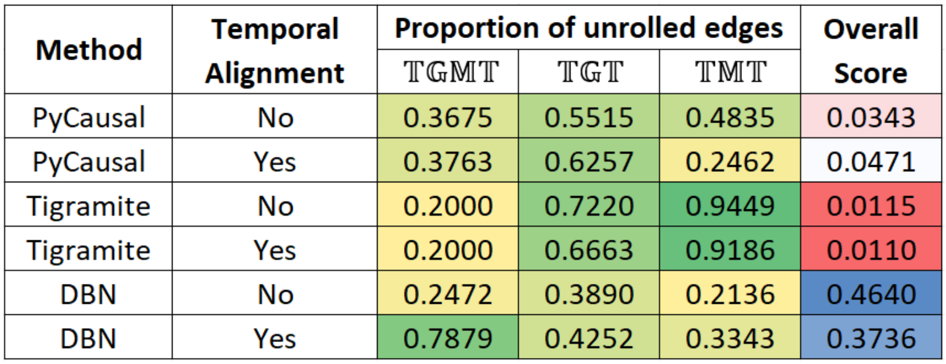


FIG 1 Samples of the ***two-time-slice DBN networks for the four different multi-omic subsets*** **produced by PALM. Self-edges are not displayed to avoid clutter. Networks were learned with a maximum number of parents of 3. The four networks show the nodes representing variables from each omics data source organized in two large circles, one representing the variables for the current time point (blue) and the other for the next time point (orange). Node shapes represent the omics data source of the variable. Taxa nodes are represented as filled circles, metabolites as filled squares, genes as filled diamonds, and clinical variables as filled triangles. Red (green) edges represent negative (positive resp.) regression coefficients. Edge width is proportional to the regression coefficient and edge opacity to the bootstrap score. Finally, node opacity is proportional to abundance. a) DBN learned with just taxa abundance ($\mathbb{T}$). The dataset included abundance of 27 bacteria and a clinical variable indicating the week the sample was obtained and resulted in a network with 95 edges. b) DBN learned with taxa and metabolites ($\mathbb{TM}$). A set of 19 metabolites were added to the previous dataset, and 164 edges were learned in this network. c) DBN learned with the taxa and genes dataset ($\mathbb{TG}$). A set of 34 genes were added to the taxa dataset, and a network with 230 edges was learned. d) DBN learned with the 27 taxa, 34 genes, and 19 metabolites ($\mathbb{TGM}$), resulting in a total of 311 edges.**



Method	Temporal Alignment	Alpha	Proportion of unrolled edges			Overall Score
			TGMT	TGT	TMT	
PyCausal	No	0.01	0.770	0.659	0.667	0.019
PyCausal	No	0.001	0.275	0.604	0.451	0.024
PyCausal	No	0.0001	0.058	0.391	0.333	0.060
PyCausal	Yes	0.01	0.724	0.711	0.158	0.039
PyCausal	Yes	0.001	0.288	0.750	0.231	0.055
PyCausal	Yes	0.0001	0.117	0.417	0.350	0.047

FIG 2 Heatmap showing the proportion of edges unrolled by METALICA **in the Crohn's disease datasets for the networks obtained from PyCausal (**TETRAD**) as the alpha parameter varies using datasets with and without temporal alignment. Last column shows the overall bootstrap score.**



Method	Temporal Alignment	Proportion of unrolled edges			Overall Score
		TGMT	TGT	TMT	
PyCausal	No	0.3675	0.5515	0.4835	0.0343
PyCausal	Yes	0.3763	0.6257	0.2462	0.0471
Tigramite	No	0.2000	0.7220	0.9449	0.0115
Tigramite	Yes	0.2000	0.6663	0.9186	0.0110
DBN	No	0.2472	0.3890	0.2136	0.4640
DBN	Yes	0.7879	0.4252	0.3343	0.3736

FIG 3 Heatmap showing percentages of edges unrolled by METALICA **in the Crohn's disease datasets for all the methods averaged over all parameter choices. The last column shows the overall bootstrap score.**

means that higher alpha values result in networks with higher average confidence on each edge, since it is also more difficult for it to be learned by chance. This is consistent with the higher percentage of unrolling for larger alpha values, indicating that the edges with higher support get unrolled more frequently, adding support for the unrolling process. Interestingly, there is a clear reversal of the pattern for the overall bootstrap score (last column) for the experiments without temporal alignment, where, contrary to our intuition, the smaller alpha values result in higher overall scores. Interestingly, temporally aligning the data set seems to fix this problem, which would support the necessity of alignment as a pre-processing step.

Also, as shown in Figure 3, the DBN/PALM method seems more stable than the other two algorithms, since the much higher average overall bootstrap score indicates that in each bootstrap, the edges learned are consistent with the ones learned in other bootstrap runs. This lower variability across the different random data subsamples used is a clear advantage of the DBN/PALM method.

The top unrollings and de-confoundings discovered by METALICA using the networks from all the methods were sorted based on the overall bootstrap score, and other factors like the number of networks they appear in, or the different network types that supported this particular finding. We discuss below some particularly interesting results from the METALICA analysis described above.

Uncovering unrolled biological relationships Here, we discuss the unrolling of specific edges from the METALICA results using the dataset containing all diseases.

First, we consider the edge *Eubacterium siraeum* → *Bacteroides thetaiotaomicron* in $G_{\mathbb{T}}$, i.e., the edge between the abundance of the two bacterial taxa, *E. siraeum* and *B. thetaiotaomicron*. It manifests itself as the unrolled path *E. siraeum* → uridine kinase → cytidine → *B. thetaiotaomicron* in $G_{\mathbb{TGM}}$, as shown in Figure 4. The following is the support for each edge in the unrolled path from the literature and the knowledge-bases. Both *E. siraeum* and *B. thetaiotaomicron* contain the gene to produce enzyme uridine kinase (54, 55). This enzyme, when present in prokaryotes and eukaryotes, phosphorylates both uridine and cytidine to their mono-phosphate forms, and vice-versa. The specific reactions that this enzyme is capable of performing are the following (56, 57, 58):

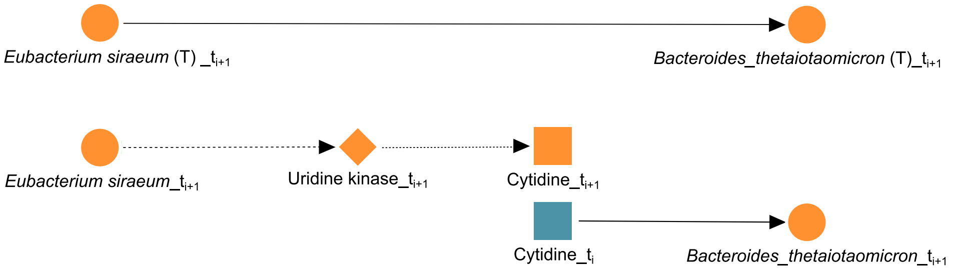

Eubacterium siraeum (T) _t_{i+1}
Bacteroides_thetaiotaomicron (T)_t_{i+1}
Eubacterium siraeum_t_{i+1}
Uridine kinase_t_{i+1}
Cytidine_t_{i+1}
Cytidine_t_i
Bacteroides_thetaiotaomicron_t_{i+1}

FIG 4 Biologically confirmed unrolling. The edge *Eubacterium siraeum* → *Bacteroides thetaiotaomicron* learned in $G_{\mathbb{T}}$ (T) is unrolled into *Eubacterium siraeum* → uridine kinase → cytidine → *Bacteroides thetaiotaomicron* in $G_{\mathbb{TGM}}$.

- ATP + Uridine ⇌ ADP + UMP, and
- ATP + Cytidine ⇌ ADP + CMP,

where ATP stands for adenosine tri-phosphate, ADP stands for adenosine di-phosphate, UMP stands for uridine mono-phosphate, and CMP stands for cytidine mono-phosphate. Since *B. thetaiotaomicron* carries the gene for uridine kinase, it has the ability to perform the forward reaction and consume it by phosphorylating cytidine to CMP. More importantly, *B. thetaiotaomicron* also has the gene for cytidine deaminase, which scavenges exogenous and endogenous cytidine for UMP synthesis (59). The reaction performed by this enzyme is cytidine + H2O ⇌ uridine + Ammonia (60, 61, 62), which validates the third and last edge (cytidine → *B. thetaiotaomicron*) in Figure 4. In addition, experimental results show that a cytidine-scavenging system confers colonization fitness to *B. thetaiotaomicron*, and therefore positively impact its abundance (63). Interestingly, uridine may be playing a role in this connection between the two taxa, since both enzymes discussed involve uridine, so both taxa can produce and consume uridine. Reinforcing this argument is the fact that the edge uridine → *B. thetaiotaomicron* is also present in the same network $G_{\mathbb{TGM}}$. Moreover, this unrolling can be important for IBD. Treatment for Crohn's disease with live *B. thetaiotaomicron* or its products displays strong efficacy in preclinical models of IBD, with multiple benefits (64). Similarly, there is precedent to treat gastrointestinal problems with *E. Siraeum* (65), and activation-induced cytidine deaminase seems to prevent colon cancer development despite persistent inflammation in the colon (66).

In summary, our unrolling methods allow us to make biological sense out of a set of related edges in the series of networks generated from the multi-omics data.

As a second example, the path: *Bacteroides stercoris* → uridine kinase → cytidine → *Bacteroides stercoris* can also be validated, which can be thought of as an unrolling of the self-loop from *Bacteroides stercoris* to itself in $G_{\mathbb{T}}$ as shown in Figure 5. The taxon, *B. stercoris*, carries the gene for both uridine kinase (67) and cytidine deaminase (68), so it

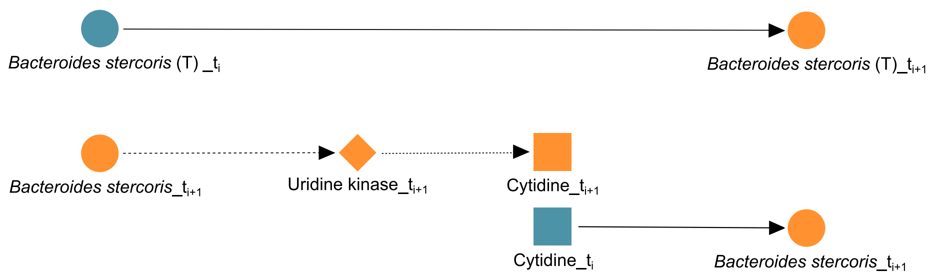

Bacteroides stercoris (T) _t_i
Bacteroides stercoris (T)_t_{i+1}
*Bacteroides stercoris*_t_{i+1}
Uridine kinase_t_{i+1}
Cytidine_t_{i+1}
Cytidine_t_i
*Bacteroides stercoris*_t_{i+1}

FIG 5 Biologically confirmed unrolling. The edge *Bacteroides stercoris* → *Bacteroides stercoris* learned in $G_{\mathbb{T}}$ (T) is unrolled into *Bacteroides stercoris* → uridine kinase → cytidine → *Bacteroides stercoris* in $G_{\mathbb{TGM}}$

can both produce and consume cytidine, and since cytidine deaminase can scavenge endogenous cytidine, this lends further support to the self-loop edge from *B. stercoris* to itself; it might be regulating itself through the cytidine or uridine internally. Interestingly, *B. stercoris* is linked to colorectal cancer (69), and its increased abundance was detected in fecal samples of Crohn's Disease (CD) patients (70). Also, an increased reactivity of Immunoglobulin G from Crohn's Disease patients toward *B. stercoris* and other species of *Bacteroides* has been shown in the serum of CD patients (71).

Two examples of "partial" validations of unrollings from our experiments are also provided. The unrolled path *Bacteroides finegoldii* → phosphatidate cytidylyltransferase → Betaine → *Eubacterium ventriosum* was discovered by our search. It first appeared as an edge *B. finegoldii* → *E. ventriosum* in $\mathbb{T}$, which then got unrolled in $\mathbb{TG}$, $\mathbb{TM}$, and $\mathbb{TGM}$. *B. finegoldii* is an anaerobic gram-negative bacteria that has been found to be generally beneficial in the gut (72). It contains the gene BN532_01044 which expresses the phosphatidate cytidylytransferase protein. This is a membrane-bound enzyme that participates in the glycerophospholipid metabolism and phosphatidylinositol signaling system. Moreover, *B. finegoldii* is known to produce the metabolite Betaine (73). Increased levels of betaine have been found to benefit IBD patients, allowing for proper digestion and assimilation of nutrients. Over the last decade, doctors have recommended betaine-rich foods as a way to help IBD patients rapidly absorb and distribute vital vitamins and minerals needed to maintain diversity in the gut (73). Additionally, recent studies have shown betaine to be correlated to the *Eubacterium* genus and to be of general importance for osmotic adaptation of most species of *Eubacterium* (74). Even though no specific study was found about the species *Eubacterium ventriosum*, the fact that betaine was found to increase the abundance of the *Eubacterium* genus lends support to the argument that *Eubacterium* members consume betaine through the conversion of Acetate (75), thus partially validating the unrolling. Moreover, while Acetate was not contemplated in the dataset, one of its precursors, Choline, was. Many strong unrollings have a link from Choline to a member of the *Eubacterium* genus in the dataset (*E. ventriosum, E. siraeum, E. rectale*), and almost every method learned the edge Betaine → *E. ventriosum* as part of specific unrollings, which could be an indication of a pathway transforming Choline to Acetate to Betaine, which may be facilitated by members of the genus, *Eubacterium*.

The path: *Bacteroides ovatus* → DNA helicase → Pyridoxine → *Bacteroides ovatus* in $\mathbb{TGM}$ can be thought of as an unrolling of a self-loop edge in $\mathbb{T}$ from *B. ovatus* to itself, which got unrolled in $\mathbb{TG}$, $\mathbb{TM}$, and $\mathbb{TGM}$. Moreover, *B. ovatus* is present in the gut microbiome, and plays a crucial role in the dysbiosis of the gut health. This anerobic bacteria has been found to have significantly elevated abundance in patients suffering from IBD. Findings suggest that some species of *Bacteroides* injure gut tissue and induce inflammation (76). This bacterium does carry the gene *dnaB*, which expresses the protein DNA helicase, an enzyme responsible in unpacking genes in an organism and DNA repair. The production of the metabolite pyridoxine has been found in great proportion when there is an abundance of *B. ovatus* (77). However, evidence suggesting the consumption of pyridoxine by the taxa could not be found. When pyriodoxine is present in great abundance, it is involved in many biochemical pathways that lead to the synthesis or metabolism of nucleic acids, immune modulatory metabolites and many others (77). However, when scarce, it leads to inflammation. We consider this as another example of a "partial" validation of our unrolling strategy.

Uncovering de-confounded biological relationships We focus next on the de-confounding actions performed by METALICA on the networks obtained using the dataset containing all diseases. The edge: thymidylate synthase → glutamate dehydrogenase was inferred in the $\mathbb{G}$ network but disappeared in the $\mathbb{TG}$ network, possibly because both genes are present in the taxon *Haemophilus parainfluenzae*. This suggests that the suggested relationship between the two genes is spurious and the taxon is the confounder. *H. parainfluenza* is an opportunistic pathogen that has been found in elevated levels in patients suffering from many diseases including pneumonia and conjunctivitis. Recent studies have shown that high abundance of this pathogen was found in patients suffering from IBD. Different dynamics have been noted for the abundance of *H. parainfluenza* in the literature. For instance, when IBD patients enter remission, there is a steep decline in this pathogen (78). Additionally, the two genes that are present in *H. parainfluenzae* were found to produce proteins that help drive diseases including colon cancer.

Limitations and future work The methods used by METALICA are only applicable to multi-omic datasets, which are relatively uncommon. However, this is expected to change in the near future with the increased effort to understand the underlying mechanisms within biological processes. Second, these methods do not provide definitive evidence for the causal chains, but rather lend support to generate hypotheses that would have to be proved with experiments in the laboratory. We argue that as larger data sets become more and more commonplace, METALICA will become increasingly useful.

CONCLUSION

We have developed METALICA, which consists of two novel *post hoc* network analysis algorithms, namely *unrolling* and *de-confounding*. We first learned biological networks from a longitudinal multi-omic IBD dataset with three state-of-the-art network and causal discovery tools. We then applied METALICA to the networks learned by the tools (DBN/PALM, tsGFCI/TETRAD, and Tigramite), and compared their predictive performance. The networks produced using DBN/PALM produced the most number of unrollings, suggesting that even though the tool was not explicitly built for causal discovery, its conditional probability underpinnings produce edges that have a reasonable chance of representing causal relationships and to lead to further biological discoveries as outlined above. The top findings by our algorithms were analyzed, and relevant biological interpretations were presented for specific network-inferred interactions.

Data availability. All code, networks, and longitudinal microbiome data sets will be made available upon publication.

Data citation. All data analyzed in this work are derived from the iHMP IBD website: https://www.ibdmdb.org (18).

ACKNOWLEDGMENTS

This work was partially supported by NIH 1R15AI128714-01 (GN), and the FIU Dissertation Year Fellowship (DR-P). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

REFERENCES

1. **Riesenfeld CS, Schloss PD, Handelsman J**. 2004. Metagenomics: genomic analysis of microbial communities. Annu Rev Genet 38:525–552.
2. **Fernandez M, Aguiar-Pulido V, Riveros J, Huang W, Segal J, Zeng E, Campos M, Mathee K, Narasimhan G**. 2016. Microbiome analysis: State of the art and future trends. Comput Methods for Next Gener Seq Data Anal p 401–424.
3. **Bashiardes S, Zilberman-Schapira G, Elinav E**. 2016. Use of metatranscriptomics in microbiome research. Bioinform Biol Insights 10:BBI–S34610.
4. **Turnbaugh PJ, Gordon JI**. 2008. An invitation to the marriage of metagenomics and metabolomics. Cell 134 (5):708–713.
5. **Stebliankin V, Sazal M, Valdes C, Mathee K, Narasimhan G**. 2022. A novel approach for combining the metagenome, metaresistome, metareplicome and causal inference to determine the microbes and their antibiotic resistance gene repertoire that contribute to dysbiosis. Microb Genom 8 (12):mgen000899.
6. **Castro-Nallar E, Shen Y, Freishtat RJ, Pérez-Losada M, Manimaran S, Liu G, Johnson WE, Crandall KA**. 2015. Integrating microbial and host transcriptomics to characterize asthma-associated microbial communities. BMC Med Genom 8 (1):50.
7. **Aguiar-Pulido V, Huang W, Suarez-Ulloa V, Cickovski T, Mathee K, Narasimhan G**. 2016. Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis: supplementary issue: bioinformatics methods and applications for big metagenomics data. Evol Bioinform 12:EBO–S36436.
8. **Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, Xia LC, Xu ZZ, Ursell L, Alm EJ, et al.**. 2016. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. The ISME journal 10 (7):1669–1681.
9. **Fernandez M, Riveros JD, Campos M, Mathee K, Narasimhan G**. 2015. Microbial" social networks". BMC genomics 16 (11):S6.
10. **Sazal M, Mathee K, Ruiz-Perez D, Cickovski T, Narasimhan G**. 2020. Inferring directional relationships in microbial communities using signed Bayesian networks. BMC genomics 21:1–11.
11. **Sazal M, Stebliankin V, Mathee K, Yoo C, Narasimhan G**. 2021. Causal effects in microbiomes using interventional calculus. Sci Reports 11 (1):5724.
12. **Palsson B, Zengler K**. 2010. The challenges of integrating multi-omic data sets. Nat Chem Biol 6 (11):787–789.
13. **Beale DJ, Karpe AV, Ahmed W**. 2016. Beyond metabolomics: a review of multi-omics-based approaches, p 289–312. *In* Microbial metabolomics. Springer, Cham.
14. **Yugi K, Kubota H, Hatano A, Kuroda S**. 2016. Trans-omics: how to reconstruct biochemical networks across multiple 'omic' layers. Trends Biotech 34 (4):276–290.
15. **Madhavan S, Bender RJ, Petricoin EF**. 2019. Integration of multi-omic data into a single scoring model for input into a treatment recommendation ranking. Google Patents US Patent App. 16/405,640.
16. **Xiao H**. 2019. Network-based approaches for multi-omic data integration. PhD thesis. University of Cambridge.
17. **Zhou W, Sailani MR, Contrepois K, Zhou Y, Ahadi S, Leopold SR, Zhang MJ, Rao V, Avina M, Mishra T, et al.**. 2019. Longitudinal multi-omics of host–microbe dynamics in prediabetes. Nature 569 (7758):663–671.
18. **Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E, Ajami NJ, Bonham KS, Brislawn CJ, et al.**. 2019. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature 569 (7758):655.
19. **Boekel J, Chilton JM, Cooke IR, Horvatovich PL, Jagtap PD, Käll L, Lehtiö J, Lukasse P, Moerland PD, Griffin TJ**. 2015. Multi-omic data analysis using Galaxy. Nat Biotechnol 33 (2):137–139.
20. **Sangaralingam A, Dayem U AZ, Marzec J, Gadaleta E, Nagano A, Ross-Adams H, Wang J, Lemoine NR, Chelala C**. 2017. 'Multi-omic' data analysis using O-miner. Brief Bioinform 20 (1):130–143.
21. **Ruiz-Perez D, Lugo-Martinez J, Bourguignon N, Mathee K, Lerner B, Bar-Joseph Z, Narasimhan G**. 2021. Dynamic Bayesian Networks for Integrating Multi-omics Time Series Microbiome Data. mSystems 6 (2):e01105–20.
22. **Canzler S, Schor J, Busch W, Schubert K, Rolle-Kampczyk UE, Seitz H, Kamp H, von Bergen M, Buesen R, Hackermüller J**. 2020. Prospects and challenges of multi-omics data integration in toxicology. Arch Toxicol p 1–18.
23. **Ulfenborg B**. 2019. Vertical and horizontal integration of multi-omics data with miodin. BMC Bioinform 20 (1):649.
24. **Ma T, Zhang A**. 2019. Integrate multi-omics data with biological interaction networks using Multi-view Factorization AutoEncoder (MAE). BMC Genom 20 (11):1–11.
25. **Morton JT, Aksenov AA, Nothias LF, Foulds JR, Quinn RA, Badri MH, Swenson TL, Van Goethem MW, Northen TR, Vazquez-Baeza Y, et al.**. 2019. Learning representations of microbe–metabolite interactions. Nat Methods 16 (12):1306–1314.
26. **Fabres PJ, Collins C, Cavagnaro TR, Rodríguez-López CM**. 2017. A concise review on multi-omics data integration for terroir analysis in *Vitis vinifera*. Front Plant Sci 8:1065.
27. **Ruiz-Perez D, Guan H, Madhivanan P, Mathee K, Narasimhan G**. 2020. So you think you can PLS-DA? BMC Bioinform In Press.
28. **Gerber GK**. 2014. The dynamic microbiome. FEBS Lett 588 (22):4131–4139.
29. **La Rosa PS, Warner BB, Zhou Y, Weinstock GM, Sodergren E, Hall-Moore CM, Stevens HJ, Bennett WE, Shaikh N, Linneman LA, Hoffmann JA, Hamvas A, Deych E, Shands BA, Shannon WD, Tarr PI**. 2014. Patterned progression of bacterial populations in the premature infant gut. Proc Natl Acad Sci 111 (34):12522–12527.
30. **Stein RR, Bucci V, Toussaint NC, Buffie CG, Rätsch G, Pamer EG, Sander C, Xavier JB**. 2013. Ecological modeling from time-series inference: Insight into dynamics and stability of intestinal microbiota. PLoS Comput Biol 9 (12):1–11.
31. **Gibson TE, Gerber GK**. 2018. Robust and scalable models of microbiome dynamics. *In* Proc. 35th International Conference on Machine Learning PMLR 80, p 1763–1772.
32. **Lugo-Martinez J, Ruiz-Perez D, Narasimhan G, Bar-Joseph Z**. 2019. Dynamic interaction network inference from longitudinal microbiome data. Microbiome 7 (1):54.
33. **Hughes DA, Bacigalupe R, Wang J, Rühlemann MC, Tito RY, Falony G, Joossens M, Vieira-Silva S, Henckaerts L, Rymenans L, et al.**. 2020. Genome-wide associations of human gut microbiome variation and implications for causal inference analyses. Nat Microbiol 5

(9):1079–1087.

34. **Lynch KE, Parke EC, O'Malley MA**. 2019. How causal are microbiomes? A comparison with the *Helicobacter pylori* explanation of ulcers. Biol & Philos 34 (6):62.

35. **Sanna S, van Zuydam NR, Mahajan A, Kurilshikov A, Vila AV, Võsa U, Mujagic Z, Masclee AA, Jonkers DM, Oosting M, et al.**. 2019. Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. Nat genetics 51 (4):600–605.

36. **Relman DA**. 2020. Thinking about the microbiome as a causal factor in human health and disease: philosophical and experimental considerations. Curr Opin Microbiol 54:119–126.

37. **Scheines R, Spirtes P, Glymour C, Meek C, Richardson T**. 1998. The TETRAD project: Constraint based aids to causal model specification. Multivar Behav Res 33 (1):65–117.

38. **Ramsey JD, Zhang K, Glymour M, Romero RS, Huang B, Ebert-Uphoff I, Samarasinghe S, Barnes EA, Glymour C**. 2018. TETRAD—-A toolbox for causal discovery. *In* 8th international workshop on climate informatics.

39. **TETRAD**. 2015. CMU Philosophy Group. GitHub: https://github.com/cmu-phil/tetrad .

40. **Runge J, Nowack P, Kretschmer M, Flaxman S, Sejdinovic D**. 2019. Detecting and quantifying causal associations in large nonlinear time series datasets. Sci Adv 5 (11):eaau4996.

41. **Kanehisa M, Goto S**. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research 28 (1):27–30.

42. **Dagum P, Galper A, Horvitz E**. 1992. Dynamic network models for forecasting. *In* Uncertainty in artificial intelligence Elsevier, p 41–48.

43. **Dagum P, Galper A, Horvitz E, Seiver A**. 1995. Uncertain reasoning and forecasting. Int J Forecast 11 (1):73–87.

44. **Lähdesmäki H, Hautaniemi S, Shmulevich I, Yli-Harja O**. 2006. Relationships between probabilistic Boolean networks and dynamic Bayesian networks as models of gene regulatory networks. Signal processing 86 (4):814–834.

45. **Murphy KP**. 2002. Dynamic Bayesian networks: representation, inference and learning. PhD thesis. University of California, Berkeley Berkeley, CA.

46. **Malinsky D, Spirtes P**. 2018. Causal structure learning from multivariate time series in settings with unmeasured confounding. *In* Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery p 23–47.

47. **Causal P**. 2016. by Chirayul. GitHub .

48. **Entner D, Hoyer PO**. 2010. On causal discovery from time series data using FCI. Probabilistic graphical models p 121–128.

49. **Colombo D, Maathuis MH**. 2012. A modification of the PC algorithm yielding order-independent skeletons. Prepr arXiv:1211.3295 .

50. **Spirtes P, Glymour CN, Scheines R, Heckerman D**. 2000. Causation, prediction, and search. MIT press.

51. **Schwarz G, et al.**. 1978. Estimating the dimension of a model. The annals statistics 6 (2):461–464.

52. **Aurora R**. 2019. Confounding factors in the effect of gut microbiota on bone density. Rheumatology 58 (12):2089–2090.

53. **Wang Y, Blei DM**. 2019. The blessings of multiple causes. J Am Stat Assoc 114 (528):1574–1596.

54. **KEGG**. Accessed: 2020-10-20. *Eubacterium siraeum* V10Sc8a: ES1_08270. KEGG .

55. **KEGG**. Accessed: 2020-10-20. *Bacteroides thetaiotaomicron* 7330: Btheta7330_03179. KEGG .

56. **Valentin-Hansen P**. 1978. [39] Uridine-cytidine kinase from *Escherichia coli*, p 308–314. *In* Methods in enzymology, vol 51. Elsevier.

57. **Orengo A**. 1969. Regulation of Enzymic Activity by Metabolites I. URIDINE-CYTIDINE KINASE OF NOVIKOFF ASCITES RAT TUMOR. J Biol Chem 244 (8):2204–2209.

58. **Sköld O**. 1960. Uridine kinase from Ehrlich ascites tumor: purification and properties. J Biol Chem 235 (11):3273–3279.

59. **UniProt**. Accessed: 2020-10-20. UniProtKB - R9HQ62 (R9HQ62_BACT4). UniProt .

60. **Vincenzetti S, Cambi A, Neuhard J, Schnorr K, Grelloni M, Vita A**. 1999. Cloning, expression, and purification of cytidine deaminase from *Arabidopsis thaliana*. Protein expression purification 15 (1):8–15.

61. **Song BH, Neuhard J**. 1989. Chromosomal location, cloning and nucleotide sequence of the *Bacillus subtilis* cdd gene encoding cytidine/deoxycytidine deaminase. Mol Gen Genet MGG 216 (2-3):462–468.

62. **Wang T, Sable H, Lampen J**. 1950. Enzymatic deamination of cytosine nucleosides. J Biol Chem 184 (1):17–28.

63. **Glowacki RW, Pudlo NA, Tuncil Y, Luis AS, Sajjakulnukit P, Terekhov AI, Lyssiotis CA, Hamaker BR, Martens EC**. 2020. A Ribose-Scavenging System Confers Colonization Fitness on the Human Gut Symbiont *Bacteroides thetaiotaomicron* in a Diet-Specific Manner. Cell host & microbe 27 (1):79–92.

64. **Delday M, Mulder I, Logan ET, Grant G**. 2019. *Bacteroides thetaiotaomicron* ameliorates colon inflammation in preclinical models of Crohn's disease. Inflamm bowel diseases 25 (1):85–96.

65. **Borody TJ**. 2003. Treatment of gastro-intestinal disorders. Google Patents US Patent 6,645,530.

66. **Takai A, Marusawa H, Minaki Y, Watanabe T, Nakase H, Kinoshita K, Tsujimoto G, Chiba T**. 2012. Targeting activation-induced cytidine deaminase prevents colon cancer development despite persistent colonic inflammation. Oncogene 31 (13):1733–1742.

67. **NCBI**. Accessed: 2020-10-20. BACSTE_RS07450 uridine kinase [*Bacteroides stercoris* ATCC 43183]. NCBI .

68. **NCBI**. Accessed: 2020-10-20. BACSTE_RS03560 cytidine deaminase [*Bacteroides stercoris* ATCC 43183]. NCBI .

69. **Liu Z, Cao AT, Cong Y**. 2013. ; Elsevier. Microbiota regulation of inflammatory bowel disease and colorectal cancer. Semin Cancer Biol 23 (6):543–552.

70. **Walters SS, Quiros A, Rolston M, Grishina I, Li J, Fenton A, DeSantis TZ, Thai A, Andersen GL, Papathakis P, et al.**. 2014. Analysis of gut microbiome and diet modification in patients with Crohn's disease. SOJ microbiology & infectious diseases 2 (3):1.

71. **Kappler K, Lasanajak Y, Smith DF, Opitz L, Hennet T**. 2020. Increased antibody response to fucosylated oligosaccharides and fucose-carrying Bacteroides species in Crohn's disease. Front microbiology 11:1553.

72. **Bakir MA, Kitahara M, Sakamoto M, Matsumoto M, Benno Y**. 2006. *Bacteroides finegoldii* sp. nov., isolated from human faeces. Int journal systematic evolutionary microbiology 56 (5):931–935.

73. **Craig SA**. 2004. Betaine in human nutrition. The Am journal clinical nutrition 80 (3):539–549.

74. **Imhoff JF, Rodriguez-Valera F**. 1984. Betaine is the main compatible solute of halophilic eubacteria. J bacteriology 160 (1):478–479.

75. **Watkins AJ, Roussel EG, Parkes RJ, Sass H**. 2014. Glycine betaine as a direct substrate for methanogens (Methanococcoides spp.). Appl Environ Microbiol 80 (1):289–293.

76. **Saitoh S, Noda S, Aiba Y, Takagi A, Sakamoto M, Benno Y, Koga Y**. 2002. *Bacteroides ovatus* as the predominant commensal intestinal microbe causing a systemic antibody response in inflammatory bowel disease. Clin diagnostic laboratory immunology 9 (1):54–59.

77. **Selhub J, Byun A, Liu Z, Mason JB, Bronson RT, Crott JW**. 2013. Dietary vitamin B6 intake modulates colonic inflammation in the IL10-/- model of inflammatory bowel disease. The J nutritional biochemistry 24 (12):2138–2143.

78. **Schirmer M, Denson L, Vlamakis H, Franzosa EA, Thomas S, Got-

man NM, Rufo P, Baker SS, Sauer C, Markowitz J, et al.. 2018. Compositional and temporal changes in the gut microbiome of pediatric ulcerative colitis patients are linked to disease course. Cell host & microbe 24 (4):600–610.