ONENET – ONE NETWORK TO RULE THEM ALL: CONSENSUS NETWORK INFERENCE FROM MICROBIOME DATA

CAMILLE CHAMPION
Université Paris-Saclay, INRAE, MGP, 78350, Jouy-en-Josas, France
RAPHAELLE MOMAL
Université Paris-Saclay, INRAE, MGP, 78350, Jouy-en-Josas, France
EMMANUELLE LE CHATELIER
Université Paris-Saclay, INRAE, MGP, 78350, Jouy-en-Josas, France.
MAHENDRA MARIADASSOU
Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France.
mahendra.mariadassou@inrae.fr
MAGALI BERLAND
Université Paris-Saclay, INRAE, MGP, 78350, Jouy-en-Josas, France.
magali.berland@inrae.fr

ABSTRACT

Modeling microbial interactions as sparse and reproducible networks is a major challenge in microbial ecology. Direct interactions between the microbial species of a biome can help to understand the mechanisms through which microbial communities influence the system. Most state-of-the art methods reconstruct networks from abundance data using Gaussian Graphical Models, for which several statistically grounded and computationnally efficient inference approaches are available. However, the multiplicity of existing methods, when applied to the same dataset, generates very different networks. In this article, we present OneNet, a consensus network inference method that combines seven methods based on stability selection. This resampling procedure is used to tune a regularization parameter by computing how often edges are selected in the networks. We modified the stability selection framework to use edge selection frequencies directly and combine them in the inferred network to ensure that only reproducible edges are included in the consensus. We demonstrated on synthetic data that our method generally led to slightly sparser networks while achieving much higher precision than any single method. We further applied the method to gut microbiome data from liver-cirrothic patients and demonstrated that the resulting network exhibited a microbial guild that was meaningful in term of human health.

Keywords Network inference · stability selection · microbial ecology · microbial guild · Gaussian Graphical Models

1 Introduction

The human gut microbiota is a complex ecosystem consisting of trillions of microorganisms, mainly viruses, bacteria, archeae and microbial eucaryotes, that play critical roles in host physiology including digestion, immune function and metabolism [Belkaid and Hand, 2014, Chatelier et al., 2013]. Recent advances in sequencing technologies have enabled the characterization of gut microbiota composition and function at a fine scale, providing opportunities to understand the microbial communities that reside within the human gastrointestinal tract. However, despite these technological advancements, understanding the interactions within the bacteria of the gut microbiota remains a major challenge. These interactions are complex as microorganisms can interact with each other in a multitude of ways: through mutualism, parasitism, commensalism and competition to only cite a few [Weiss et al., 2017, Faust and Raes, 2016].

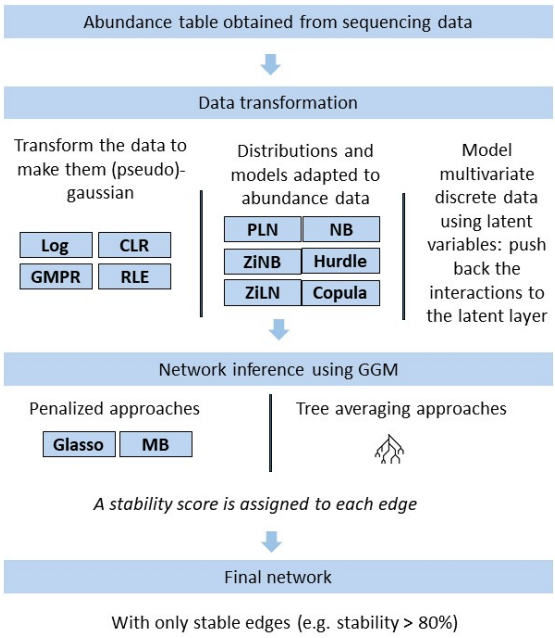


Figure 1: The classical network inference pipeline.

To address this challenge, network-based approaches have been developed to infer microbial interactions and construct microbial interaction networks. The resulting networks can reveal potential interactions between microbial taxa and support the identification of microbial guilds. Those guilds are defined as groups of microorganisms that co-occur and may interact with each other. Identifying microbial guilds is crucial for understanding the ecological dynamics of the gut microbiota and can provide insights into the role of the microbiota in health and disease [Wu et al., 2021, Xiao et al., 2022].

Formally, microbial interaction networks consist of nodes, which correspond to microbial species, and edges, which correspond to interactions between those species. Positive and negative interactions are rarely observed directly. They are instead often reconstructed from abundance data, using either longitudinal data (see the generalized Lotka-Volterra model in Bucci et al. [2016]) or co-occurrence data. We focus here on the latter suite of methods.

The simplest way to identify microbial interactions is to perform a correlation analysis. However, correlation-based methods model total dependencies and are therefore prone to confusion by environmental factors (*e.g.* shared habitat preferences or susceptibility to the same abiotic factors) and do not lend themselves to a clear separation between indirect and direct effects [Friedman and Alm, 2012]. By contrast, conditional dependency-based method eliminate indirect correlations from direct interactions and lead to sparser and easier to interpret networks, at the cost of increased computational burden and more sophisticated models. The problem of network inference is complicated by the adverse characteristics of microbial abundance data, which are sparse, heterogeneous, heteroscedastic and show extreme variability. These data are thus tricky to model, which leads to poorly reproducible and/or sparse networks, with many missing edges [Peschel et al., 2021].

The most common framework for the estimation of the conditional dependencies is Gaussian Graphical Models (GGM) [Lauritzen, 1996], which describe the conditional dependency structure of multivariate Gaussian distributions. As microbiome abundance data don't directly fit within the gaussian framework, three main workarounds are commonly used: data transformation, models based on alternative distributions and models based on latent variables ; the whole strategy is also illustrated on Fig 1, and each step is detailed in the Supplementary Methods (Section 6).

The complexity of reconstructing networks from co-occurrence data has spawned a rich literature with many methods relying on the solutions exposed above, including (i) approaches based on correlation as `SparCC` [Friedman and Alm, 2012], `CoNet` [Faust and Raes, 2016], (ii) approaches based on probabilistic graphical models as `SpiecEasi` [Kurtz et al., 2015], `gCoda` [Fang et al., 2017], `SPRING` [Yoon et al., 2020], `PLNetwork` [Chiquet et al., 2018], `ZiLN` [Prost et al., 2021], `cozine` [Ha et al., 2020], `Magma` [Cougoul et al., 2019], `EMtree` [Momal et al., 2020] and (iii) approaches based on the inference of the latent correlation structure as `CCLasso` [Fang et al., 2015], `SparCC` [Friedman and Alm,

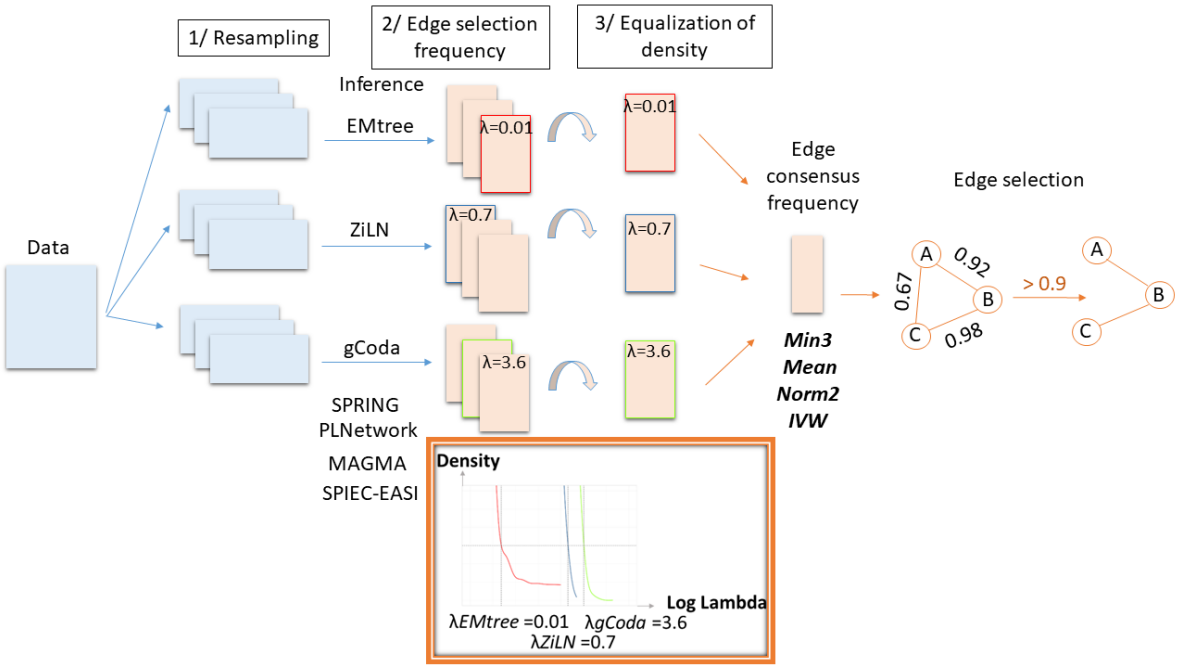


Figure 2: High level summary of the OneNet pipeline: (i) bootstrap subsamples are constructed from the original abundance matrice, (ii) each inference method is applied on the bootstrap subsamples to compute edge selection frequencies using a fixed λ grid, (iii) a different λ is selected for each method to achieve the same density in all methods, (iv) edge selection frequencies are summarized and (v) thresholded to compute the consensus graph.

2012]. The most recent methods include: `mixPLN` and `ms-mixPLN` [Tavakoli and Yooseph, 2019, Yooseph and Tavakoli, 2022] which consider the problem of inferring multiple microbial networks (one per host condition) from a given sample-taxa abundance matrix when microbial associations are impacted by host factors. Jiang et al. [2020] proposed Hybrid Approach foR MicrobiOme Network Inferences via Exploiting Sparsity (`HARMONIES`) which addresses some critical aspects of abundance data (compositionality due to fixed sampling depth, over-dispersion and zero-inflation of the abundances) while maintaining computational scalability and sparsity of the interaction network, in contrast to `mixPLN` and `ms-mixPLN`. Finally, Network Construction and comparison for Microbiome data (`NetCoMi`) [Peschel et al., 2021], provides a one-step platform for inference and comparison of microbial networks, by implementing many existing methods for abundance data preprocessing, network inference and edge selection in single package.

All these methods have been designed to infer networks based on different mathematical hypotheses and thus have different strengths and weaknesses when modeling microbiome data. Each microbial network inference algorithm usually returns distinct edges to connect the taxa together, as many facets of the same reality. In this article, we present OneNet, an ensemble method that generate robust and reliable consensus network that will facilitate the identification of microbial guilds and generation of new hypotheses.

2 Methods

2.1 Overview of OneNet

We developed OneNet, a three-step procedure for robust consensus network reconstruction, illustrated on Figure 2.

We included seven inference methods in OneNet, all of which rely on Gaussian Graphical Models (GGM) to estimate conditional dependencies networks: Magma, SpiecEasi, gCoda, PLNetwork, EMtree, SPRING and ZiLN. We excluded `mixPLN` and `ms-mixPLN` as they do not reconstruct a single network but rather a collection of networks and `NetCoMi` as it collects already existing methods rather than introducing a new one. We left out `HARMONIES` from the comparison as its implementation doesn't allow the user to specify the regularization grid, a crucial step in the ensemble method, and achieved worse performance than included methods in preliminary tests. We also excluded `cozine` from OneNet because its implementation doesn't rely on resampling and prevents it from being integrated, but we nonetheless

Method	Normalization	Distribution	Inference approach	Covariates	Reference
SpiecEasi	CLR	Multivariate gaussian	MB	No	Kurtz et al. [2015]
gCoda	CLR	Multivariate gaussian	glasso	No	Fang et al. [2017]
SPRING	CLR	Copulas	MB	No	Yoon et al. [2020]
Magma	GMPR + RLE	Copulas + ZINB	MB	Yes	Cougoul et al. [2019]
PLNetwork	GMPR + RLE	PLN + Latent variables	glasso	Yes	Chiquet et al. [2018]
EMtree	GMPR + RLE	Latent variables	Tree averaging	Yes	Momal et al. [2020]
ZiLN	CLR	Latent variables	MB	No	Prost et al. [2021]
cozine	CLR	Hurdle gaussian	MB	Yes	Ha et al. [2020]

Table 1: Characteristics of the network inference methods: general methods, abundances normalization (Centred Log Ratio (CLR), Geometric Mean of Pairwise Ratios (GMPR), Relative Log Expression (RLE)), distribution transformations, inference approaches (Meinshausen-Bühlmann (MB), glasso, tree averaging), covariates integration and references. OneNet is based on all methods but cozine.

included it in the benchmark as it compared favorably to others methods in preliminary tests. Table 1 summarizes the inference strategies adopted by each method and the potential integration of covariables in the model.

2.2 Step 1: Assign to each edge a sequence of selection frequency values for each inference method.

Each inference method assigns a score to the edges: either a probability (for the tree averaging method) or the maximal penalty level λ below which the edge is selected in the network. An optimal penalty λ^* on these scores is then needed for an edge to be selected in the final network. Several approaches exist but the concept of stability selection [Liu et al., 2010] is the most widely used and the one considered in this work as it yields a compromise between precision and recall, while fostering reproducibility. The associated method, called Stability Approach to Regularization Selection (StARS) uses a resampling strategy to select the value of λ^* leading to the most stable graph. We describe briefly the StARS algorithm formally in the following and the modification we propose in this work.

2.2.1 StARS algorithm.

The original data X is subsampled B times and the network inference is conducted on each sub-sample for each value of λ in a grid $(\lambda_1, \ldots, \lambda_K)$ to obtain a graph $G^{b,k}$, with $k \in \{1, ...K\}$ and $b \in \{1, \ldots, B\}$. The selection frequency of edge e for parameter λ_k, is computed as its selection frequency across the subgraphs:

$$f_e^k = \sum_{b=1}^{B} \mathbf{1}_{\{e \in G^{b,k}\}}.$$

The selection frequency over resamples gives an idea of edge reproducibility: frequency and robustness of the edges are clearly related. StARS aggregate those frequencies to construct a network-level measure of edge variability defined as:

$$S^k = 1 - 4\frac{1}{q}\sum_e f_e^k(1 - f_e^k)$$

where $q = p(p-1)/2$ is the total number of possible edges and S^k can be thought of as the mean of (edge-level) Bernoulli variances. Each value λ_k is associated to a single selection frequency vector, and a resulting stability value. Finding the right edge frequency is therefore equivalent to finding the right stability level. Classical choices for stability are $stab = 80\%$ or $stab = 90\%$ to have a good compromise between recall and precision and the optimal level λ^* is chosen as $\lambda^* = \min_{\lambda} S(\lambda) \geq stab$. Once the optimal level λ^* is fixed, one solution is to return the refit coefficient (1 if the edge is selected and 0 otherwise) computed by running inference methods on the complete dataset, to get the final network.

2.2.2 Using edge-level selection frequencies rather than network-level stability.

Instead of computing the network-level stability S^k and doing a final refit step, we propose instead to select directly the edges with high selection frequency for penalty λ as $E^\lambda(c)$ to create the set $E^\lambda(c) = \{e, \ \ f_e^\lambda > c\}$, where c is a constant close to 1. In this way, we guarantee both high precision and high reproducibility for edges in $E^\lambda(c)$ as they

are selected many times in the resampling. Smaller values of λ give rise to larger sets $E^{\lambda}(c)$ and higher recalls. Two advantages of using frequencies rather than refitting the network are (i) filtering out edges with low support that could be included in the refit graph and (ii) making it easier to combine the edges inferred by the different methods.

2.3 Step 2: Equalize of the densities of the networks.

In order to include the best edges in the consensus network, we must choose one λ per method. A natural choice would be the value $\lambda_m^{\star}$ selected by StARS for the method m. However, we observed that StARS computes a very different precision/recall for each method. We select instead the smallest λ_m such that (i) the sets $E_m^{\lambda_m}(c)$ are roughly of equal sizes and (ii) the mean stability is above a given threshold: $\frac{1}{M}\sum_{m=1}^{M} S_m(\lambda_m) \geq stab$. It forces all methods to contribute with a similar number of edges to the consensus while ensuring that each edge set is reproducible. In practice, to match edge set sizes, we worked with edge density rather than with λ values as the two are monotonically related.

2.4 Step 3: Summarize the sets of selection frequencies across methods.

Build a consensus network from the sets of edges $E^{\lambda}(c)$ produced by the different methods is akin to an ensemble procedure where many methods are combined together.

In order to produce a stable and accurate consensus network, we define several summary metrics in the objective to mitigate the drawback of each method while benefiting from their strength. The consensus is obtained by summarizing edges frequencies across the methods. Denoting by f_m the selection frequency of a given edge with method $m \in \{1, ..., M\}$, we define:

- *mean*: average selection frequency $\sum_m f_m \Big/ M$,
- *norm2*: euclidean norm (2-norm) $(\sum_m f_m^2)^{1/2} \Big/ M^{1/2}$,
- *IVW*: inverse-variance weighted average $\left(\sum_m f_m \times \frac{1}{f_m(1-f_m)}\right) \Big/ \left(\sum_m \frac{1}{f_m(1-f_m)}\right)$, where f_m follows a Bernoulli with $\widehat{Var}(f_m) = f_m(1-f_m)$,
- *min3*: high frequency for at least 3 methods $\mathbf{1}_{\{\sum_m (f_m > c) \geq 3\}}$.

Note that *min3* is the only one that returns a binary summary, all the other ones take value in $[0, 1]$, like the original selection frequencies.

2.5 Datasets

2.5.1 Simulated dataset

In this work we simulate data using the methodology described in Yoon et al. [2020] which is based on gaussian copula to control the network structure followed by sampling from the species marginal distributions to preserve the peculiarities of abundance data. This method yields synthetic data with marginal distributions that are closer to the original empirical dataset, while enforcing a given correlation structure between the species.

To generate the simulated dataset, we use in input the empirical dataset described below restricted to diseased individuals. The dataset is simulated in the framework of an unknown undirected graph $G(V, E)$, with no retroactive loop, consisting of p vertices $V = \{1, \ldots, p\}$ and a set of edges $E \subseteq V \times V$ connecting each pair of vertices. The graph G is represented by its adjacency matrix $A = (A_{ij})_{(i,j)\in E}$ of size $p \times p$, defined as:

$$\forall (i,j) \in [\![1,p]\!]^2, \;\; A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

The package 'EMtree' v.1.1.0 [Momal, 2021] is used to generate a precision matrix Ω defined as the graphical Laplacian A of a cluster graph. Ω is inverted to create the correlation matrix Σ and the idea was then to simulate variables with arbitrary marginal distributions from multivariate normal variables with correlation structure given by Σ using gaussian copula. Specifically, we generate a $n \times p$ matrix Z with independent normal rows $Z_i \sim \mathcal{N}(0, \Sigma)$. We then get uniform random vectors by applying standard normal cdf transformation to each column of Z, $u^j = \phi(Z^j / \sqrt{\Sigma_{jj}})$ element-wise, and finally apply the quantile functions of the empirical data marginal distributions to each u^j. The function *synthData_from_ecdf* from the 'SPRING' package v.1.0.4 [Yoon, 2022] is used for these simulation steps. To assess the effect of sample size, we simulate datasets of size $n \in \{50, 100, 500, 1000\}$.

2.6 Evaluation criteria

Each methods is evaluated by comparing the inferred network structure to the known simulated network structure using the following metrics:

- Precision (positive predictive value): PPV= TP/(TP+FP),
- Recall (true positive rate): TPR= TP/(TP+FN),

where TP stands for True Positive (a correctly detected edge), FP for False Positive (an edge detected where none should be) and FN for False negative (an undetected edge). The precision measures the proportion of real edges among what the detected ones, whereas the recall measures the proportion of real edges which are detected.

2.6.1 Empirical dataset

The empirical dataset, studied in Qin et al. [2014], corresponds to stool samples from 216 Chinese individuals sequenced using whole-metagenome sequencing techniques. The raw sequences are available as BioProject PRJEB6337 in the European Nucleotide Archive (ENA). Among this population, 102 individuals are healthy and 114 suffer from liver cirrhosis. Abundances of all microbial species (metagenomic species or MSP) detected using 10.4 million IGC2 gut gene catalogue [Wen et al., 2017] are extracted using the Meteor software suite that create a gene abundance table by mapping high quality reads onto the gene catalogue, using Bowtie2. Abundance of each MSP is computed as the mean abundance of 100 marker genes selected for that MSP, where the gene abundance is the read abundance normalized by the gene length. The final table of size 1990 MSP by 216 individuals records all the normalized abundances [Champion et al., 2023]. In the Application section, we used only the 114 cirrhotic patients.

3 Results

In this section we evaluated the performances of both OneNet and the network inference methods on the simulated dataset.

3.1 Influence of the stability level on the inferred graphs

We first evaluated the effect of stability level on the performance of the inference methods. Instead of fixing a target stability at 0.8 or 0.9, we studied the relationship between the precision and recall of the inferred edges by each method for different stability levels. Because interactions between highly prevalent species are easier to reconstruct, we only kept metagenomic species with a prevalence greater than 50% (159 species). We let the sample size vary in $\{50, 100, 500, 1000\}$ and we considered $B = 40$ resamples each time.

Methods have distinct precisions for a given stability level. Figure S1 shows the relationship between the performance obtained with the edge set $E^{\lambda^*}(0.90)$ (precision PPV90 and recall TPR90), and the corresponding stability. The difference in patterns grows with sample size n, revealing peculiarities inherent to each method. Clearly, methods have distinct performances for the same stability level. We observed that methods cluster in groups (glasso, neighorhood selection, tree aggregation) with different precision/recall tradeoffs. As a result, they produce edge sets that greatly differ both in size and quality. This suggests that the stability value is not a good indicator of the precision level achieved by each method.

Methods have comparable precision and recall for a given density level. Unlike precision, which is unavailable when dealing with empirical datasets, the density, or number of detected edges, can always be computed. Figure S2 shows the relationship between precision (resp. recall) and density for all methods at increasing sample sizes. The curves are almost superimposed for values of n up to 100, after which different behaviors appeared. However, the gap in performance between methods stayed small when imposing the same density, rather than the same stability. This also meant that, whatever the method used, the m first edges included in the graph achieve similar graph reconstruction quality, although they correspond to different stabilities. We thus selected individual graphs based on density, rather than stability, to include only graphs with similar precision and recall in the consensus phase.

Mean stability as a proxy of the density level. The previous observation prompted us to explore the link between density and stability for different values of λ. Figure S3 shows how stability decreases with increasing density. We set a target mean stability value (e.g. 90%) for each value of n (here between 50 and 1000). As n grows, we observed that the density increased from 90 to 170, as well as the spread between methods. For $n = 1000$, stabilities ranged from 0.8

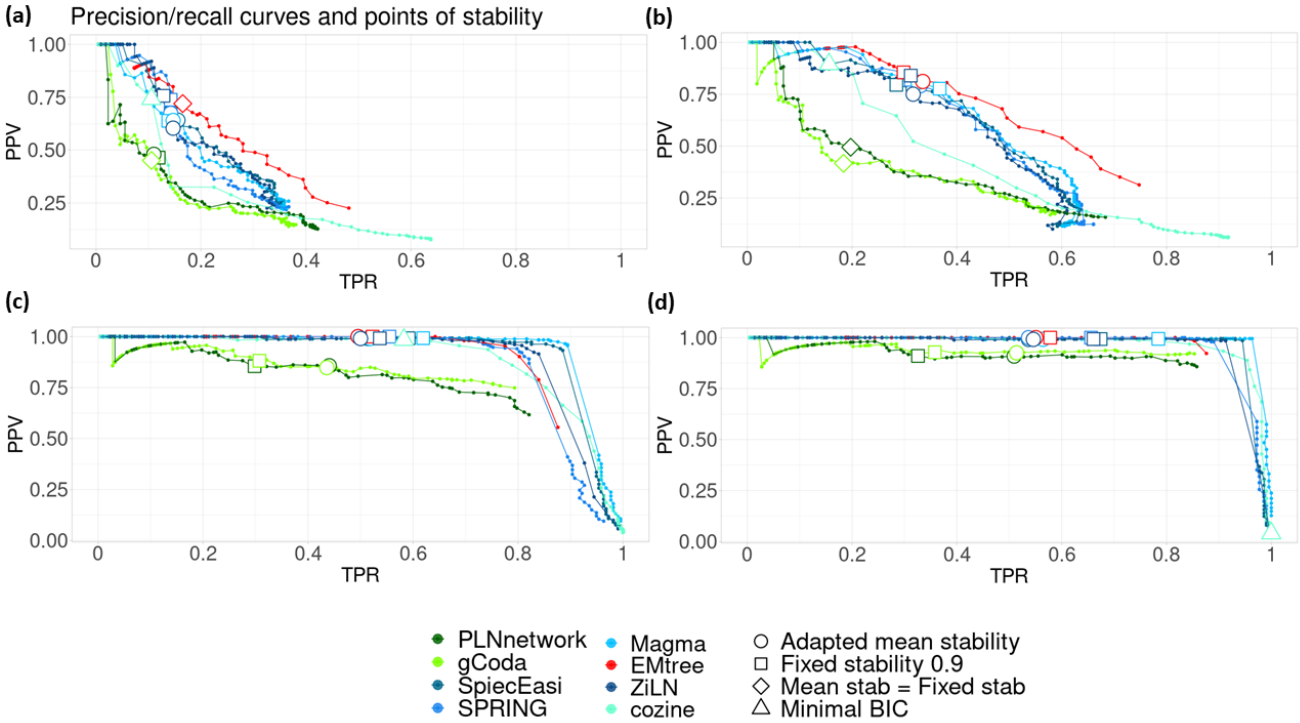


Figure 3: Precision - recall curves of each inference method for different sample sizes: (a) 50 (b) 100 (c) 500 (d) 1000. The TPR/PPV compromise achieved for λ^* corresponding to a stability of 0.9 is shown with a circle, the one achieved by a mean stability across methods of 0.9 is shown with a square. Whenever the selected λ is the same, the circle and the square are replaced with a diamond. Finally, note that cozine relies on minimization of a BIC criteria (shown with a triangle) rather than on the resampling-based stability selection to choose the regularization parameter.

for EMtree to 1 for SPRING. We can see how targeting the mean stability rather than the same stability for all methods allows to adapt the precision level of each method through density to make them more similar.

Mean stability increases the consistency between the performances of the network inference methods. We compared, for different sample sizes, the precision - recall tradeoff achieved by the mean stability to the ones achieved by a fixed stability (e.g. 0.9). Figure 3 shows that the sample size has a major impact on the precision and stability. The ROC curves stabilize to near-perfection starting at $n = 500$ (Figure 3 c). It is also noticeable that the adapted stabilities reduced the range of the method's precision. Furthermore, for glasso-based methods (gcoda, PLNnetwork), the new target led to a 20 points improvement in TPR at almost no cost in PPV, for large sample sizes.

3.2 OneNet versus the classical network inference methods

We computed, for a frequency threshold of 90%, the precision and recall values obtained by the classical network inference methods (cozine, gCoda, PLNnetwork, SPRING, Magma, SpeicEasi, ZiLM and EMtree) and we compared them to the OneNet networks (with the summay metrics mean, norm2, IVW ans min3). Note that because of the similar density, each method provided roughly the same number of edges to OneNet. Figure 4 shows that OneNet with the mean and norm2 consensus methods, achieved the best precision levels but the worse recall values. OneNet with the min3 summary has comparatively lower precision but higher precision and OneNet with the IVW summary has both worse recall and worse precision. This illustrates how OneNet generally leads to sparser networks with higher precision.

We also observed that the MB-based ones tend to outperform glasso-based ones. However, OneNet still demonstrated precision levels equivalent to that of the best methods (Figures S4 and S5). This reflects the inherent robustness of consensus measures to methods with outlying performance.

On top of that, sample size n has a dramatic effect on all criteria and methods. For large sample sizes ($n \geq 500$), most methods exhibit precision above 99% and recall between 30% and 60% (Fig. S6 and S7, panels (c) and (d)). By contrast (Fig. S6a and Fig S6b) for smaller but more realistic sample sizes like $n = 50$ (resp. $n = 100$), the median precision

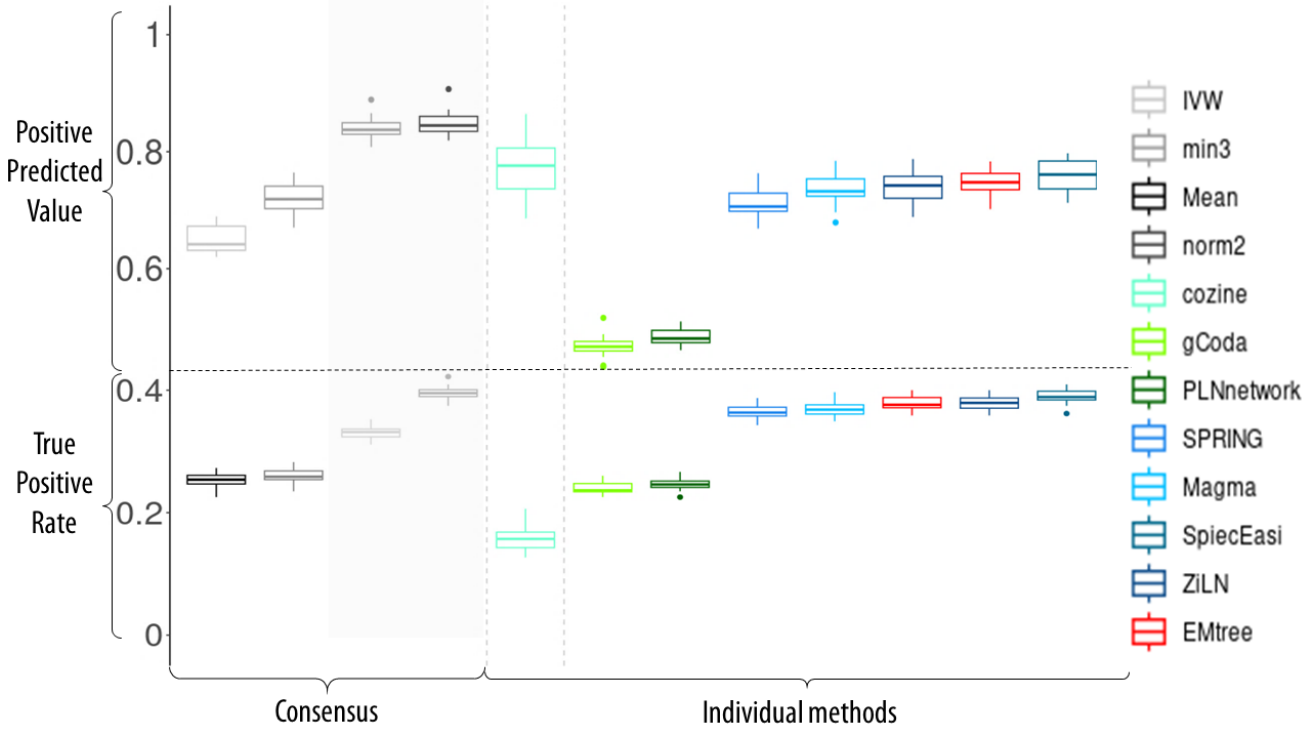


Figure 4: Quality of both individual and aggregated frequencies in terms of PPV (top row, values above 0.6) and TPR (bottom row, values below 0.5) for $n = 100$

drops below 60% (resp. 80%) for all methods except OneNet - Mean and OneNet - norm2 which both remain above 75% (resp. 85%). Likewise, the recall drops below 40% for $n = 100$ (Fig. S7b) and below 20% for $n = 50$ (Fig. S7a).

We observed a discrepancy in terms of precision and recall for the cozine method between figures S6 and S7. We hypothesized that it's due to the original cozine procedure (BIC criteria) used to select the optimal network, which selects to a dense graph (very high recall, very low precision, and therefore many spurious edges, see Figure 3). By contrast, the Figures S6 and S7 show the precision and recall values obtained with the resampling approach. This is an extreme example of lack of robustness, where the network reconstructed from the full dataset differs drastically from the ones reconstructed on random subsets of the data and illustrates the benefits of combining the resamples rather than doing a refit.

4 Application to liver cirrhosis

To investigate the potentiality of OneNet relative to the other methods, we inferred all the networks from the microbiome dataset of cirrhotic patients presented in the Methods (Qin et al. [2014]). Because of the small size of the dataset (114 samples), only metagenomic species with a prevalence greater than 50% were kept (155 species). The networks have been clustered using the CORE-clustering algorithm to reconstruct microbial guilds (Champion et al. [2021]). Following the guidelines of this paper, we fixed the number of clusters between 10 and 19.

Figure 5 illustrates the inferred and clustered networks. We note that the OneNet network, with the mean summary, is the sparsest one and is closely related to the SpiecEasi and the SPRING's one. This result is in line with the results on the simulated data and with Figures 4 and S8.

The second objective of this application was to characterize microbial species responsible for liver cirrhosis using OneNet. One guild – the "cirrhotic guild", represented in Figure 6, is notable as it contains species of the genera *Anaeroglobus*, *Campylobacter*, *Haemophilus*, *Prevotella*, *Streptococcus* known in the literature to be associated to the following diseases: liver cirrhosis [Qin et al., 2014], obesity after weight-loss intervention [Liu et al., 2017], schizophrenia [Zhu et al., 2020], atherosclerotic cardiovascular disease [Jie et al., 2017] and Crohn disease [He et al., 2017].

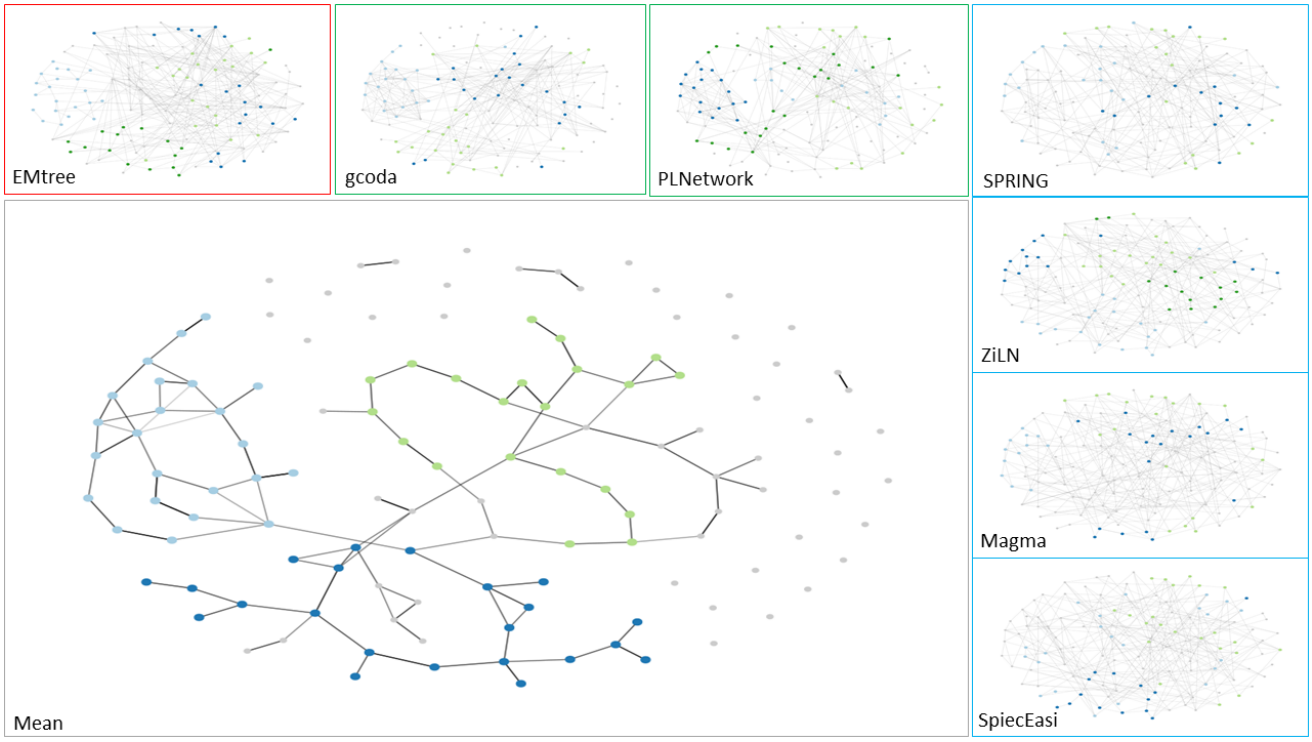


Figure 5: Networks inferred on the liver cirrhosis data set with EMtree, gCoda, PLNnetwork, SPRING, ZiLN, Magma, SpiecEasi and OneNet - mean followed by Core-clustering algorithm to identify the microbial guilds. All graphs are shown using the same node layout for ease of comparison and the nodes are coloured by cluster. Methods are grouped based on the underlying inference technique (tree aggregation, glasso, neighborhood selection).

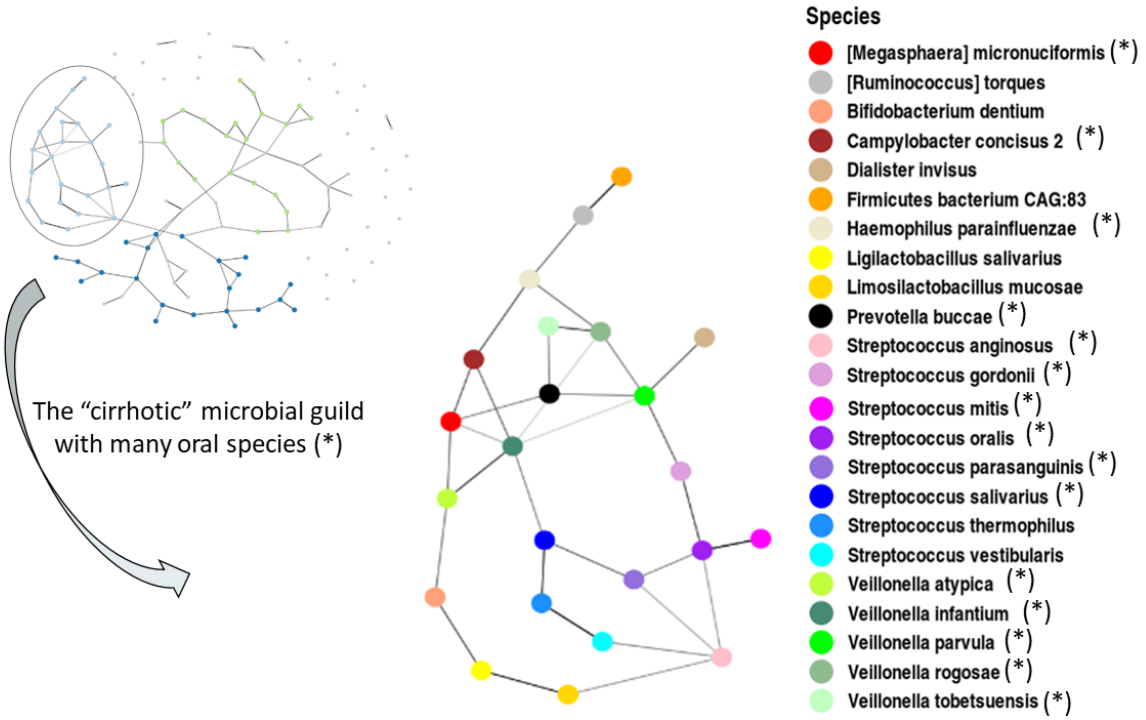


Figure 6: Detailed view of the cirrhotic guild identified in the OneNet - mean network with taxonomic information on the nodes. All species known to be associated with diseases are annotated with a star.

5 Discussion

The proposed framework, with a microbial consensus network inference method, offers new insights about inferring robust and sparse microbial networks. OneNet is robust in the sense that i) it uses GGM adapted to deal with the peculiarities of microbial abundance properties (inclusion of environmental effects as covariates, stabilization of data variability, adaptation to abundances with high proportion of zeros, etc), ii) it depends on seven network inference methods aiming for sparse and reproducible microbial network using either glasso, neighborhood selection or tree averaging approaches, iii) it relies on a three-steps procedure to improve the precision and reproducibility of both inference methods and OneNet. Indeed, the selection of edges with high inclusion frequencies and harmonization of stability selection achieve similar precision levels accross methods. The resulting consensus network uses a summary of the edge inclusion frequencies.

Results from the studies on synthetic and real data illustrated the first major and reassuring fact, that the inference methods overall agree with one another and with the truth (Figures S9 and S10). It then showed the effectiveness of OneNet compared to the inference methods. Among the different summaries considered, the mean or norm2 are preferred since they lead to slightly sparser networks but achieved much higher precision than any inference method, especially for sample sizes around $n = 100$, which is typical in microbiome studies. By contrast, min3 and IVW summaries gave a significant additional quantity of edges compared to the other summary metrics, yielding TPR levels that are comparable to those obtained with glasso-based methods without increasing the PPV, especially when the number of samples is small ($n \leq 100$) (Figures S9).

In all numerical experiments, we showed that a minimal sample size to maintain high robustness was $n = 100$. In this scenario we suggest to use the mean summary in OneNet, as it proved to be more robust to small sample sizes. Obviously, the precision is affected by both the number of samples and microbial species in the system, the latter being controlled by the prevalence threshold imposed at the very beginning of the analysis. As illustrated on Figures S11, S12 and S13, the prevalence threshold can be adjusted to increase the precision of the method depending on the number of samples ($prev = 0.50$ for $n = 100$ and $n = 500$, $prev = 0.20$ for $n = 1000$). From $n = 1000$, when considered individually, the neighborhood selection and the tree averaging approaches showed performances that were similar to OneNet. In this context, it could be possible to select one of these three approaches.

An advantage of OneNet is it's ability to easily incorporate new inference methods as soon as they are amenable to the modified stability selection framework used here. This is the case for all the methods considered in this work but `cozine`. Indeed, `cozine` relies on the BIC criteria to tune the regularization parameter λ: it doesn't allow the user to provide a fixed λ grid for comparisons with other methods and doesn't produce the table of edge selection frequencies required to compute summaries. This is however due to implementation choices (using BIC instead of StARS for selecting λ) rather than to fundamental incompatibilities with the OneNet framework.

6 Supplementary Methods

We detail here each step of the whole network inference strategy, illustrated on Fig 1. As microbiome abundance data don't directly fit within the gaussian framework, three main workarounds are commonly used: data transformation, models based on alternative distributions and models based on latent variables:

Transformations. A small constant is added to each abundance before log-transforming them. However, this transformation does not stabilize data variability because the log-transformed abundances scale with sequencing depths and covary with it, making dependencies modeling tricky. On the contrary, the Centered Log Ratio (CLR) transformation [Aitchison, 1982] guarantees the study of dependencies. It is however highly criticized when data contain a high proportion of zeros: proportions higher than 90% are typical in whole-metagenome or amplicon sequencing data. To circumvent this problem, Yoon et al. [2020] introduced a modified version of the CLR transformation (mCLR) that respects the original ordering of the data but doesn't account anymore for the compositional nature of the data, the primary motivation of the CLR transformation. An alternative to compositional transformations is to use a normalization factor, such as Geometric Mean of Pairwise Ratios (GMPR) [Chen et al., 2018], Relative Log Expression (RLE) [Anders and Huber, 2010] and others like Wrench normalization factors [Kumar et al., 2018] or Cumulative Sum Scaling [Paulson et al., 2013]. The GMPR normalization is designed for abundances with a high proportion of zero values. It compares pairs of samples based only on the species they share, and considers the geometric mean of the median ratio. This makes this technique robust to both differentially abundant species and extreme values. When two samples do not share any species, the computation of GMPR fails (this happens when samples come from very contrasted conditions with no or limited species overlap). In this case, Relative Log Expression (RLE) normalization method can be used. This method is based on the assumption that most of the species are not differentially abundant.

However, this normalization factor fails when no single species is shared across all samples, which is frequently the case in microbiome data. A modified version of RLE only considers positive abundances to avoid this drawback.

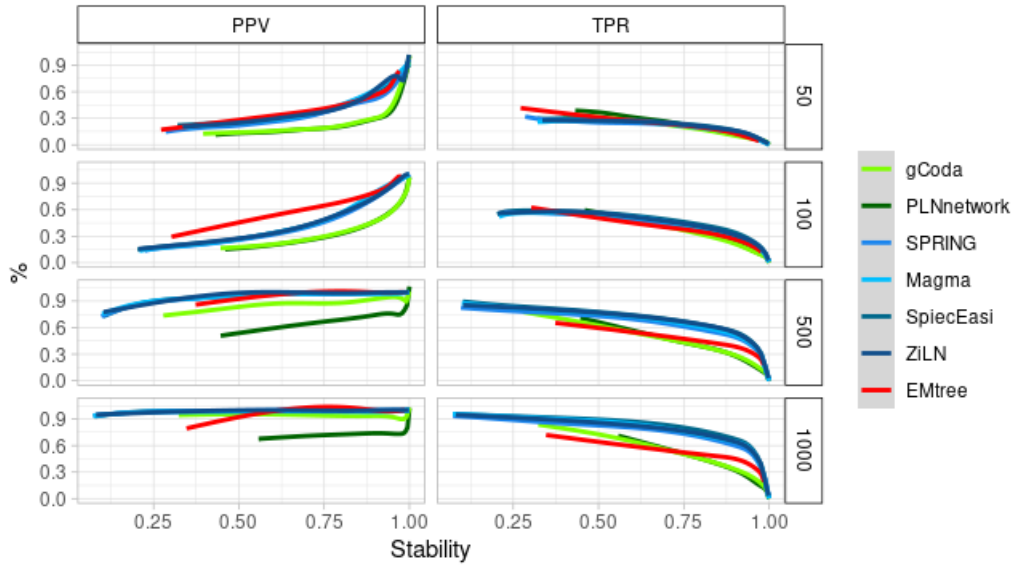
Distributions and models. The second workaround is to use models adapted to abundance data characteristics : overdispersion (excess of variability in the data) and zero-inflation (excess of zeros). The **Poisson-log normal (PLN) model** [Chiquet et al., 2021] is designed for the analysis of abundance tables. It accounts for both structuring factors and potential interactions between the species. In the presence of overdispersion, the Poisson regression model is not adequate and can lead to biased parameter estimates and unreliable standard errors estimates. The **Negative Binomial (NB) model** is then often used [Forbes et al., 2010]. Both models can be seen as compound Poisson model (with a lognormal for the PLN distribution and Gamma for the NB) that are overdispersed compared to a *vanilla* Poisson distribution but the PLN is multivariate and can account for correlations between abundances. Contrary to the NB model, the **zero-inflated model** [Greene, 1994] is often motivated by an excess of zeros in the data, but less flexible than the zero outcome model. An intuitive approach to analyzing zero-inflated abundance data is to view the data as arising from a mixture distribution of a point mass distribution at zero and an abundance distribution. **Hurdle models** [Cragg, 1971] are a class of models for abundance data that help handle excess zeros and overdispersion. In contrast to Zero inflated-models, hurdle models capture both an excess or a lack of zeros in the dataset. The **zero-inflated negative binomial (ZINB) model** [Cheung, 2002], obtained by applying ZI to NB model, takes into account both overdispersion and excess of zeros. Finaly, **copulas** are a multivariate cumulative distribution functions for which the marginal probability distribution of each variable is uniform on the interval $[0, 1]$. As they fully describe the dependency structure, models with copulas allow to separate the modeling of marginal distributions (*e.g.* overdispersed, with excess zeros, etc) from the modeling of dependencies. Recent developments used gaussian copula coupled with arbitrary discrete marginal distributions to study multivariate abundance data [Anderson et al., 2019]. Popovic et al. [2019] showed that Gaussian copulas are a relevant and promising approach to the problem of network inference from abundance data, even if the computational cost is higher than for other methods. One way of taking advantage of the copula theory without having to actually estimate the joint distribution is to use copulas as a sophisticated data transformation technique to transform abundances into pseudo-Gaussian data.

Latent variables. The third popular idea is to model multivariate discrete data using latent variables and push the depdendency back to the latent layer. Latent variables models have recently received increasing attention as they provide a convenient way to model the dependence structure between species. Two specifications of latent variables stand out in community ecology [Warton et al., 2015]: the **Multivariate Generalized Linear Mixed Model (GLMM)** [Ovaskainen et al., 2010, Tingley et al., 2014], and the **Latent Variable Model** (LVM) [Ovaskainen et al., 2016, 2017]. The difference between these models lies in the dimension of their respective random effects: there are as many latent variables as there are species in the GLMM, whereas in the LVM their number is a parameter of the model.

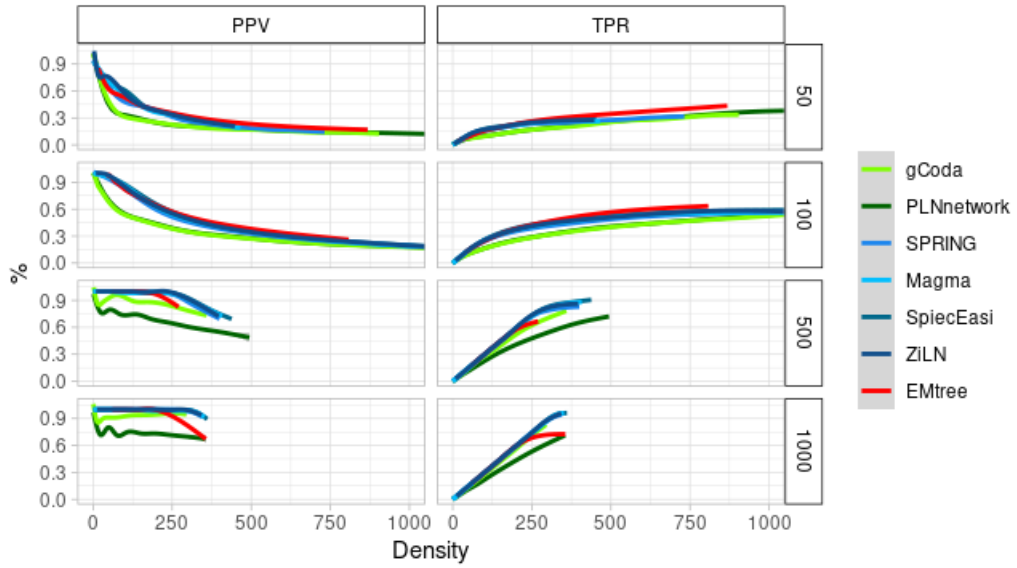
Most methodologies to infer networks from abundance data first use a rationale (data transformation, latent variable modeling, etc.) to solve the problem of network inference in the Gaussian setting. There, they take advantage of the GGM framework to perform network inference using penalized likelihood or tree-based approaches to estimate the precision matrix, from which is finally derived the network.

Penalized likelihood approaches. There exist two main penalized approaches for the estimation of GGM: the graphical LASSO (glasso) [Friedman et al., 2008], and the neighborhood selection, also called the Meinshausen-Bühlmann approach (MB) [Meinshausen and Bühlmann, 2006]. Both are penalized likelihood approaches which perform a sparse estimation of the precision matrix, either all at once for the glasso or row by row for MB.

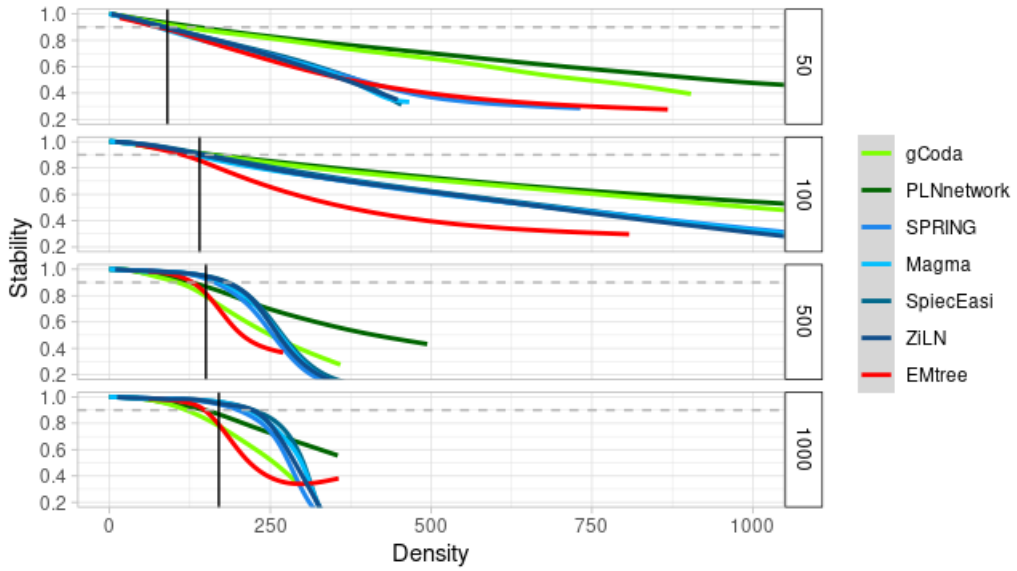
Tree averaging approach. Another GGM inference method considered in this article is the tree averaging approach [Meila and Jordan, 2000], which leverages specific algebraic properties to perform a complete and efficient exploration of the space of spanning tree structures. Note that this approach does not require the GGM Markov faithful property to hold. Each edge is given a posterior probability of being present in the network and those probabilities are thresholded to build the network.



Supplementary Figure S1: PPV - Stability and TPR - Stability curves of the edge set $E^{\lambda}(0.90)$ according to the sample size and inference method. Each point in the curve corresponds to a different value of λ.



Supplementary Figure S2: PPV - Density and TPR - Density curves of the edge set $E^{\lambda}(0.90)$ according to the sample size and inference method. Each point in the curve corresponds to a different value of λ.



Supplementary Figure S3: Stability - Density curves of the edge set $E^{\lambda}(0.90)$ according to the sample size and inference method. Each point in the curve corresponds to a different value of λ.

7 Supplementary Material

7.1 Illustrating the influence of the sample size on the stability

7.2 Illustrating the influence of the sample size on the density

7.3 Illustrating the stability selection

7.4 Illustrating the quality with and without glasso methods

7.5 Illustrating the common number of edges selected

7.6 Illustrating the performances according to the prevalence threshold

References

Y. Belkaid and T.W. Hand. Role of the microbiota in immunity and inflammation. *Cell*, 157:121–141, 2014.

E. Le Chatelier, T. Nielsen, J. Qin, E. Prifti, F. Hildebrand, and G. Falony. Richness of human gut microbiome correlates with metabolic markers. *Nature*, 500:541–546, 2013.

S. Weiss, Z.Z. Xu, S. Peddada, A. Amir, K. Bittinger, A. Gonzalez, and C. Lozupone. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1), 2017.

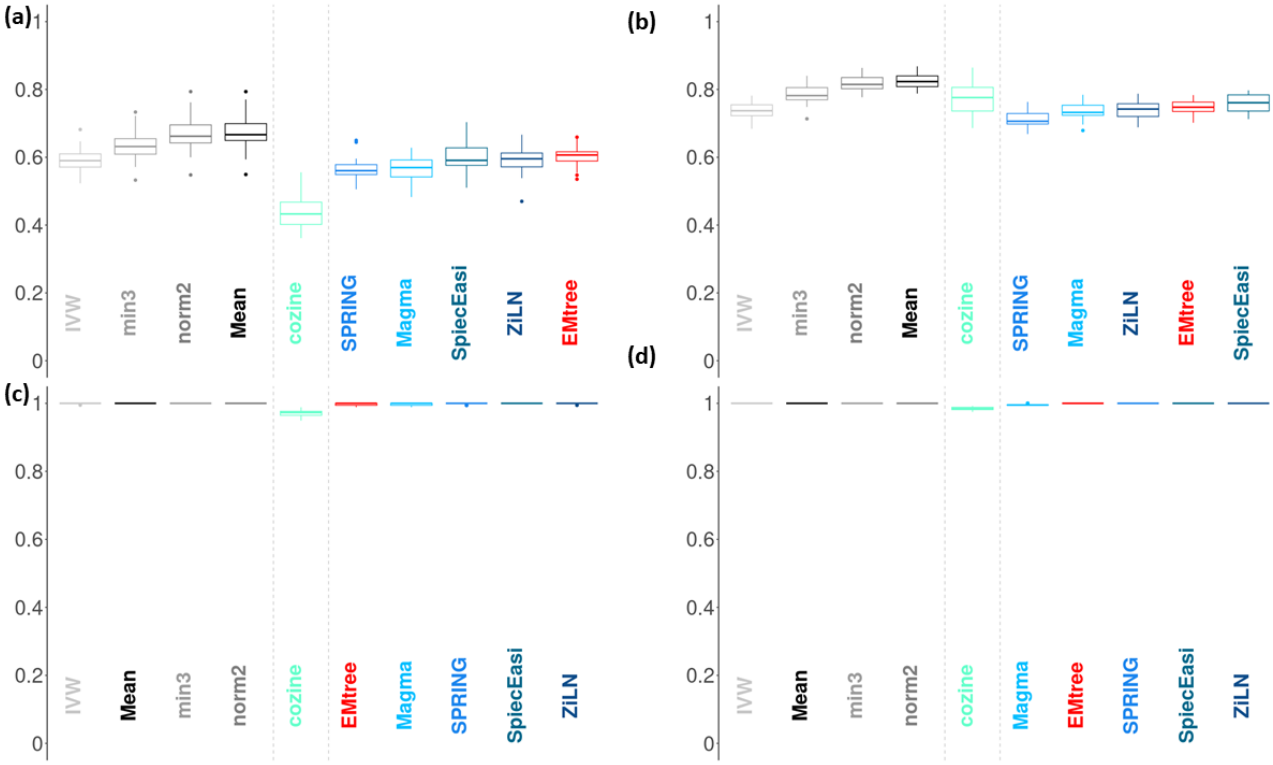
K. Faust and J. Raes. Conet app: inference of biological association networks using cytoscape. *F1000Research*, (5), 2016.

G. Wu, N. Zhao, C. Zhang, Y.Y. Lam, and L. Zhao. Guild-based analysis for understanding gut microbiome in human health and diseases. *Genome Medicine*, 13(22), 2021.

L. Xiao, F.K Zhang, and F. Zhao. Large-scale microbiome data integration enables robust biomarker identification. *Nat Comput Sci*, 2:307–316, 2022.

V. Bucci, B.Tzen, N. Li, M. Simmons, T. Tanoue, E. Bogart, L. Deng, V. Yeliseyev, M.L. Delaney, Q. Liu, B. Olle, R.R. Stein, K. Honda, L. Bry, and G.K. Gerber. Mdsine: Microbial dynamical systems inference engine for microbiome time-series analyses. *Genome Biol*, 17(121), 2016.

J. Friedman and E. Alm. Inferring correlation networks from genomic survey data. *PLoS computational biology*, 8, 2012.



Supplementary Figure S4: Compared precision (PPV) of inference methods and OneNet-* variants after removing glasso-based methods from the set of methods, for different samples sizes: (a) $n = 50$ (b) $n = 100$ (c) $n = 500$ (d) $n = 1000$

S. Peschel, C.L. Müller, E. Von Mutius, A.L. Boulesteix, and M. Depner. Netcomi: network construction and comparison for microbiome data in r. *Brief Bioinformatics*, 22(4), 2021.

S.L. Lauritzen. *Graphical Models*. Oxford Statistical Science Series, 1996.

Z.D. Kurtz, C.L. Müller, E.R Miraldi, D.R. Littman M.J. Blaser, and R.A. Bonneau. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Computational Biology*, 11(5), 2015.

H. Fang, C. Huang, H. Zhao, and M. Deng. gcoda: Conditional dependence network inference for compositional data. *Journal of Computational Biology*, 24(7):699–708, 2017.

G. Yoon, R.J Carroll, and I. Gaynanova. Sparse semiparametric canonical correlation analysis for data of mixed types. *Biometrika*, 107(3):609–625, 2020.

J. Chiquet, S. Robin, and M. Mariadassou. Variational inference for sparse network reconstruction from count data. In *International Conference on Machine Learning*. PMLR, 2018.

V. Prost, S. Gazut, and T. Brüls. A zero inflated log-normal model for inference of sparse microbial association networks. *PLOS Computational Biology*, 17(6), 2021.

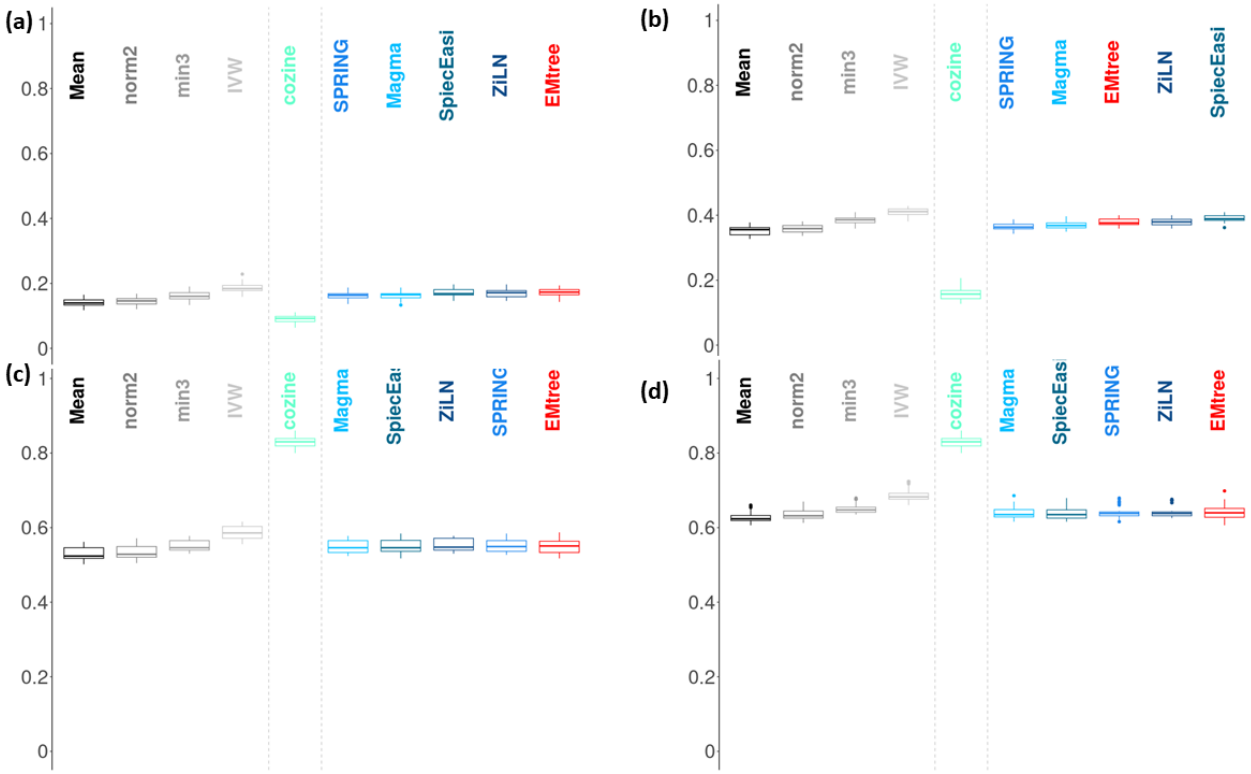
M.J Ha, J. Kim, J. Galloway-Pena, K.A. Do, and C.B. Peterson. Compositional zero-inflated network estimation for microbiome data. *BMC Bioinformatics*, 21(581), 2020.

A. Cougoul, X. Bailly, and E.C. Wit. Magma: Inference of sparse microbial association networks. (538579), 2019. URL `https://www.biorxiv.org/content/10.1101/538579v1.full.pdf`.

R. Momal, S. Robin, and C. Ambroise. Tree-based inference of species interaction networks from abundance data. *Methods in Ecology and Evolution*, 11(5):621–632, 2020.

H. Fang, C. Huang, H. Zhao, and M. Deng. Cclasso: correlation inference for compositional data through lasso. *Bioinformatics*, 31:3172–3180, 2015.

S. Tavakoli and S. Yooseph. Learning a mixture of microbial networks using minorization-maximization. *Bioinformatics*, 35(14):23–30, 2019.



Supplementary Figure S5: Compared recall (TPR) of inference methods and OneNet-* variants after removing glasso-based methods from the set of methods, for different samples sizes: (a) $n = 50$ (b) $n = 100$ (c) $n = 500$ (d) $n = 1000$

S. Yooseph and S. Tavakoli. Variational approximation-based model selection for microbial network inference. *Journal of computational biology*, 29(0), 2022.

S. Jiang, G. Xiao, A.Y. Koh, Y. Chen, B. Yao, Q.Li, and X.Zhan. Harmonies: A hybrid approach for microbiome networks inference via exploiting sparsity. *Front. Genet.*, 11(445), 2020.

H. Liu, K. Roeder, and L. Wasserman. Stability approach to regularization selection (stars) for high dimensional graphical models. *Advances in Neural Information Processing Systems*, 24(2), 2010.

R. Momal. Emtree: Infers direct species association networks using tree averaging. `https://rdrr.io/github/Rmomal/EMtree/`, 2021.

G. Yoon. Semi-parametric rank-based approach for inference in graphical model (spring). `https://rdrr.io/github/GraceYoon/SPRING/`, 2022.

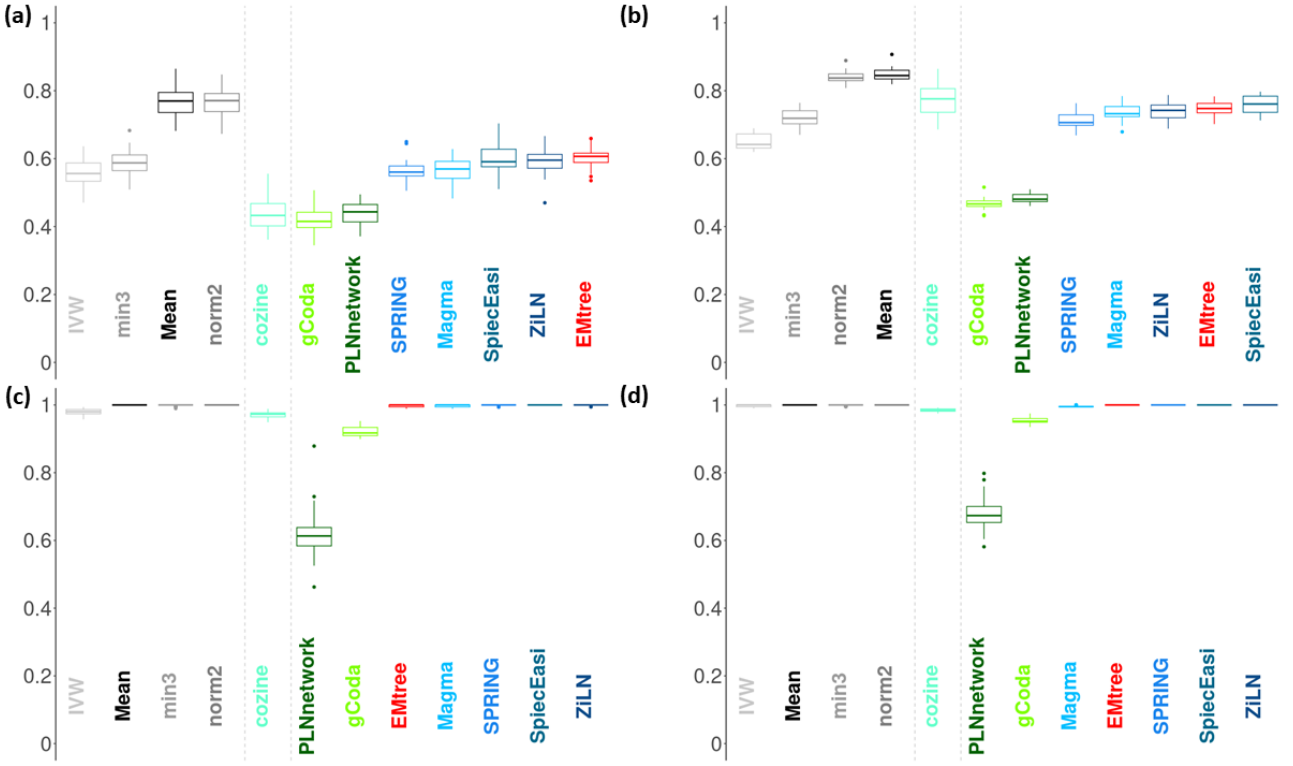
N. Qin, F. Yang, A. Li, E. Prifti, Y. Chen, L. Shao, and J.Guo. Alterations of the human gut microbiome in liver cirrhosis. *Nature*, 513(7516):59–64, 2014.

C. Wen, Z. Zheng, T. Shao, L. Liu, Z. Xie, E. Le Chatelier, Z. He, W. Zhong, Y. Fan, L. Zhang, H. Li, C. Wu, Q. Xu, J. Zhou, S. Cai, D. Wang, Y. Huang, M. Breban N. Qin, and S.D. Ehrlich. Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome biology*, 1(18):255–261, 2017.

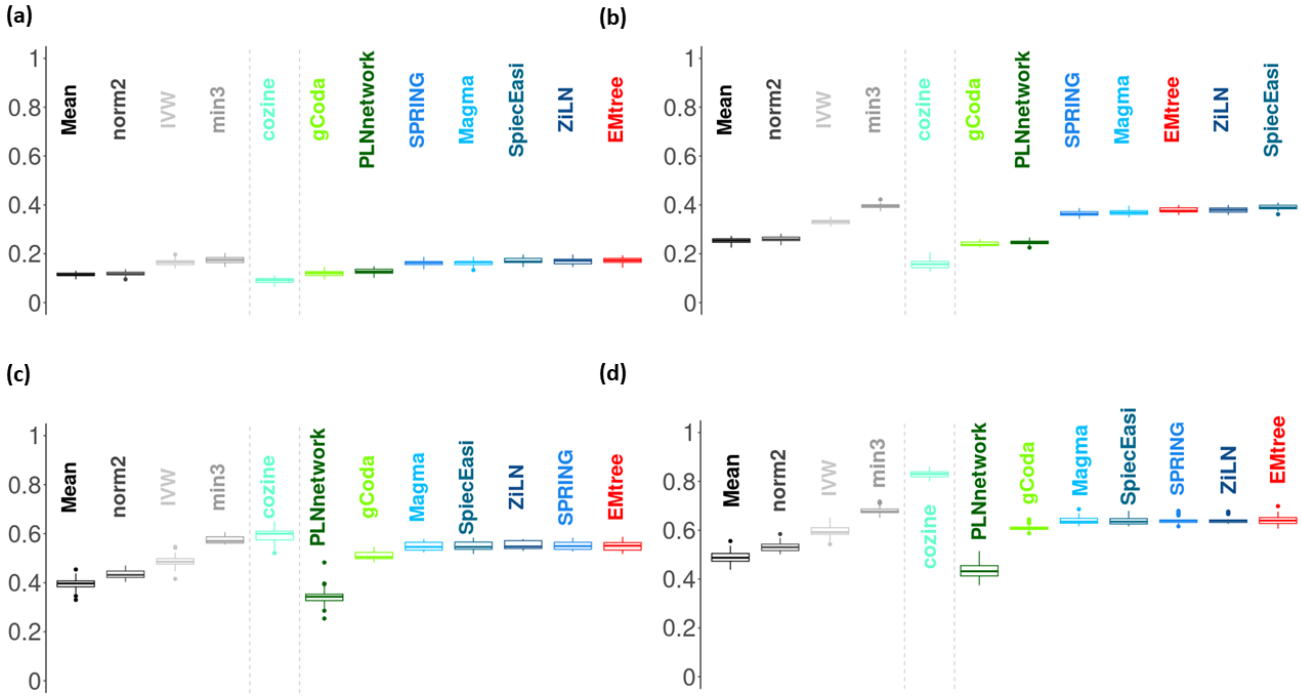
C. Champion, R. Momal, E. Le Chatelier, M. Mariadassou, and M. Berland. Microbial species abundances from public project PRJEB6337 on liver cirrhosis. `https://doi.org/10.57745/5YXXN1`, 2023.

C. Champion, A.C. Brunet, R. Burcelin, J.M. Loubes, and L. Risser. Detection of representative variables in complex systems with interpretable rules using core-clusters. *Algorithms*, 2(14), 2021.

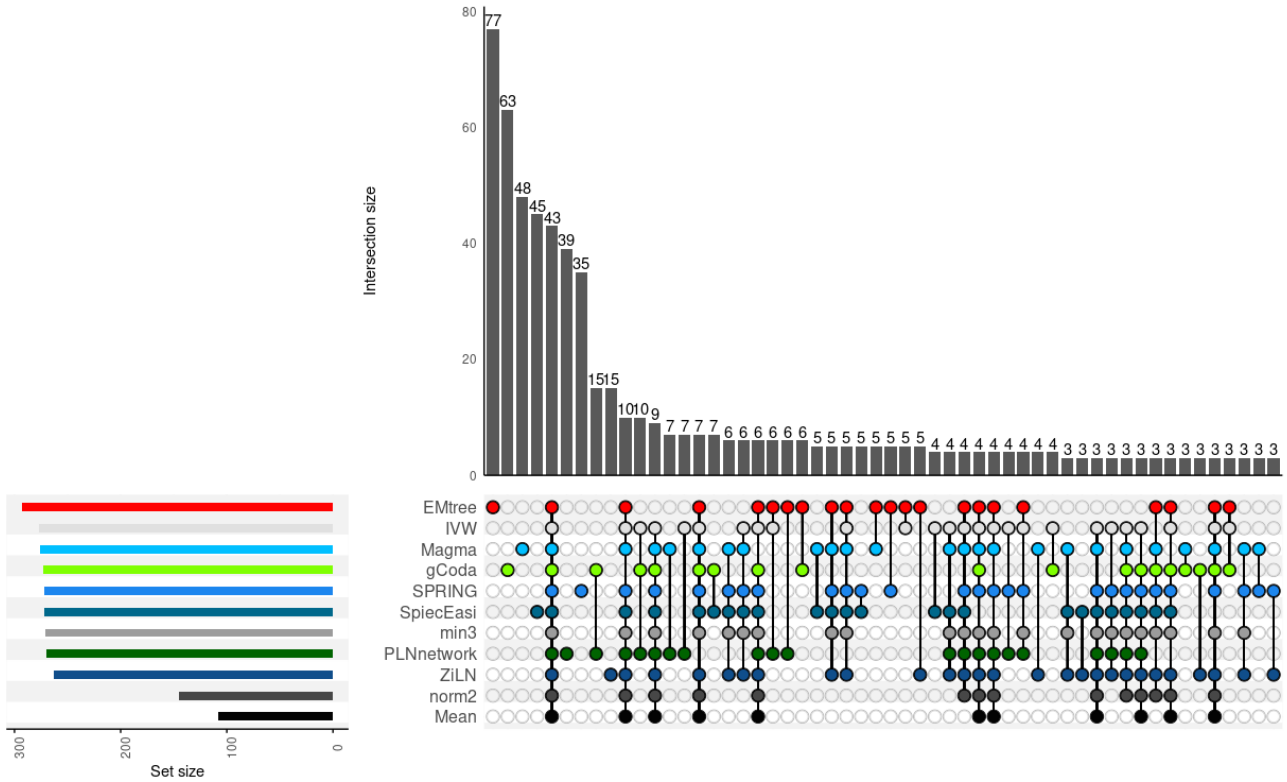
R. Liu, J. Hong, X. Xu, Q. Feng, D. Zhang, Y. Gu, J. Shi, S. Zhao, W. Liu, X. Wang, H. Xia, Z. Liu, B. Cui, P. Liang, L. Xi, J. Jin, X. Ying, X. Wang, X. Zhao, W. Li, H. Jia, Z. Lan, F. Li, R. Wang, Y. Sun, M. Yang, Y. Shen, Z. Jie, J. Li, X. Chen, H. Zhong, H. Xie, Y. Zhang, W. Gu, X. Deng, B. Shen, X. Xu, H. Yang, G. Xu, Y. Bi, S. Lai, J. Wang, L. Qi, L. Madsen, J. Wang, G. Ning, K. Kristiansen, and W. Wang. Gut microbiome and serum metabolome alterations in obesity and after weight-loss intervention. *Nature Medicine*, 23(7):859–868, 2017.



Supplementary Figure S6: Compared precision (PPV) of inference methods and OneNet-* variants when including all 7 methods in the set of methods, for different samples sizes: (a) $n = 50$ (b) $n = 100$ (c) $n = 500$ (d) $n = 1000$



Supplementary Figure S7: Compared recall (TPR) of inference methods and OneNet-* variants when including all 7 methods in the set of methods, for different samples sizes: (a) $n = 50$ (b) $n = 100$ (c) $n = 500$ (d) $n = 1000$



Supplementary Figure S8: Upset plots of the edges identified by the inference methods and the OneNet-* variants applied to the liver cirrhosis dataset.

F. Zhu, Y. Ju, W. Wang, Q. Wang, R. Guo, Q. Ma, Q. Sun, Y. Fan, Y. Xie, Z. Yang, Z. Jie, B. Zhao, L. Xiao, L. Yang, T. Zhang, J. Feng, L. Guo, X. He, Y. Chen, C. Chen, C. Gao, X. Xu, H. Yang, J. Wang, Y. Dang, L. Madsen, S. Brix, K. Kristiansen, H. Jia, and X. Ma. Metagenome-wide association of gut microbiome features for schizophrenia. *Nature Communications*, 11(1612), 2020.

Z. Jie, H. Xia, S.L. Zhong, Q. Feng, S. Li, S. Liang, H. Zhong, Z. Liu, Y. Gao, H. Zhao, D. Zhang, Z. Su, Z. Fang Z. Lan, J. Li, L. Xiao, J. Li, R. Li, X. Li, F. Li, H. Ren, Y. Huang, Y. Peng, G. Li, B. Wen, B. Dong, J.-Y. Chen, Q.S. Geng, Z.W. Zhang, H. Yang, J. Wang, J. Wang, X. Zhang, L. Madsen, S. Brix, G. Ning, X. Xu, X. Liu, Y. Hou, H. Jia, K. He, and K. Kristiansen. The gut microbiome in atherosclerotic cardiovascular disease. *Nature Communications*, 8 (845), 2017.

Q. He, Y. Gao, Z. Jie, X. Yu, J.M. Laursen, L. Xiao, Y. Li, L. Li, F. Zhang, Q. Feng, X. Li, J. Yu, C. Liu, P. Lan, T. Yan, X. Liu, X. Xu, H. Yang, J. Wang, L. Madsen, S. Brix, J. Wang, K. Kristiansen, and H. Jia. Two distinct metacommunities characterize the gut microbiota in crohn's disease patients. *Nature Communications*, 6:1–11, 2017.

J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

L. Chen, J. Reeve, L. Zhang, S. Huang, X. Wang, and J. Chen. Gmpr: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ*, 6, 2018.

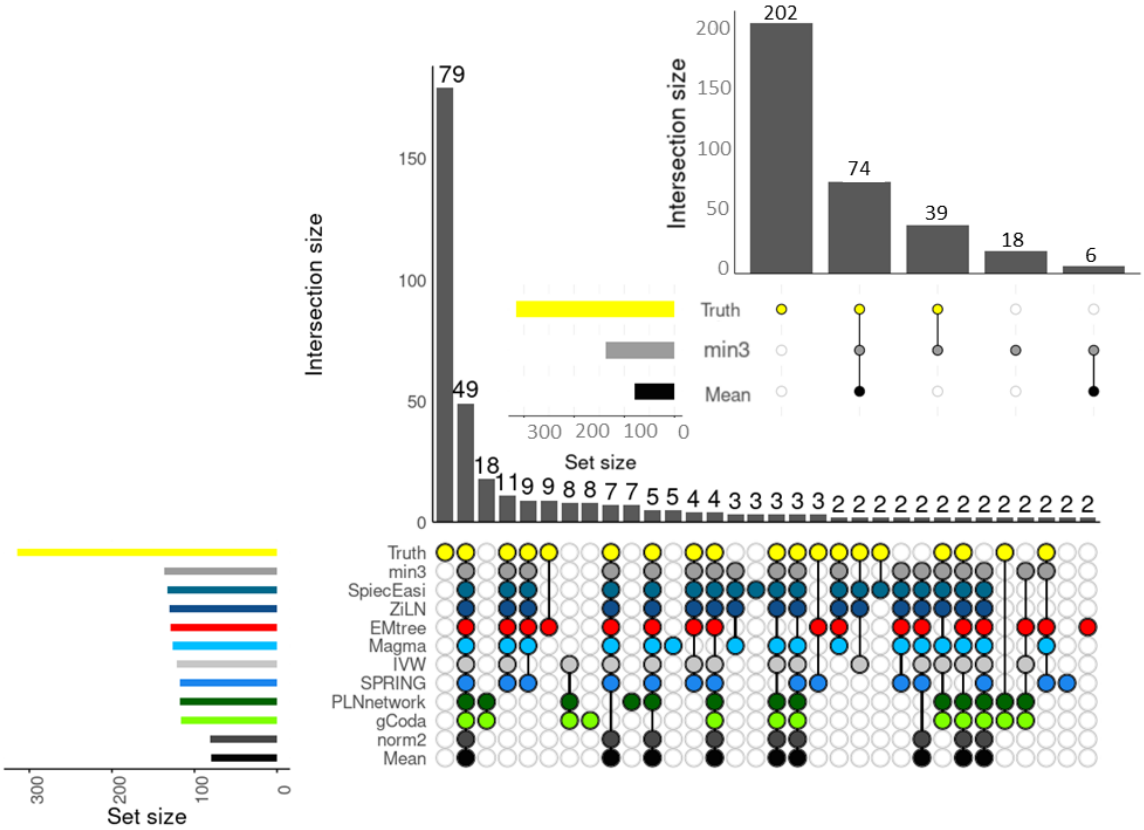
S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10), 2010.

M. Senthil Kumar, E.V. Slud, K. Okrah, S.C. Hicks, S. Hannenhalli, and H. Corrada Bravo. Analysis and correction of compositional bias in sparse sequencing count data. *BMC Genomics*, 19(1):799, 2018.

J.N. Paulson, O. Colin Stine, H. Corrada Bravo, and M. Pop. Differential abundance analysis for microbial marker-gene surveys. *Nat Methods*, 10:1200–1202, 2013.

J. Chiquet, M. Mariadassou, and S. Robin. The poisson-lognormal model as a versatile framework for the joint analysis of species abundances. *Frontiers in Ecology and Evolution*, 9, 2021.

C. Forbes, M. Evans, N. Hastings, and B. Peacock. *Statistical distributions*. John Wiley and Sons, 2010.



Supplementary Figure S9: Upset plot of the edges identified by the inference methods, the OneNet-* variants and the ground truth on the synthetic dataset for $n = 100$.

W.H. Greene. Accounting for excess zeros and sample selection in poisson and negative binomial regression models. *Research Papers in Economics*, 1994.

J.G. Cragg. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, 39:829–844, 1971.

Y.B. Cheung. Zero-inflated models for regression analysis of count data: a study of growth and development. *Statistics in Medicine*, 21, 2002.

M.J. Anderson, P. de Valpine, A. Punnett, and A. E. Miller. A pathway for multivariate analysis of ecological communities using copulas. *Ecology and evolution*, 9(6):3276–3294, 2019.

G. C. Popovic, D. I. Warton, F. J. Thomson, F. K. C. Hui, and A. T. Moles. Untangling direct species associations from indirect mediator species effects with graphical models. *Methods in Ecology and Evolution*, 10(9):1571–1583, 2019.

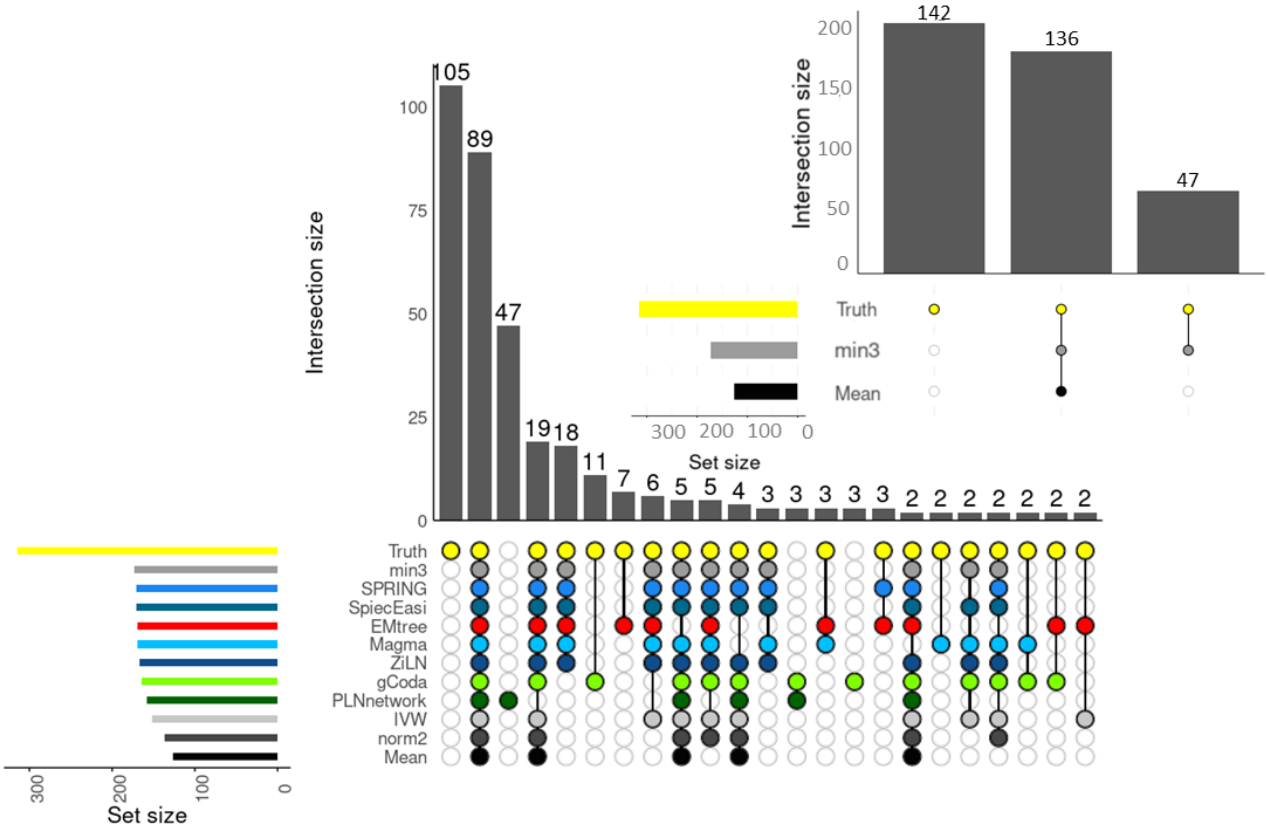
D. I. Warton, F. G. Blanchet, R. B. O'Hara, O. Ovaskainen, S. Taskinen, S. C Walker, and F. KC. Hui. So many variables: joint modeling in community ecology. *Trends in Ecology and Evolution*, 30(12):766–779, 2015.

O. Ovaskainen, J. Hottola, and J. Siitonen. Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, 91(9):2514–2521, 2010.

L. J. Pollockand R. Tingley, W. K. Morris, N. Golding, R. B. O'Hara, K. M. Parris, P. A. Vesk, and M. A. McCarthy. Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (jsdm). *Methods in Ecology and Evolution*, 5(5):397–406, 2014.

O. Ovaskainen, N. Abrego, P. Halme, and D. Dunson. Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods in Ecology and Evolution*, 7(5):549–555, 2016.

O. Ovaskainen, G. Tikhonov, A. Norberg, F Guillaume B., L. Duan, D. Dunson, T. Roslin, and N. Abrego. How to make more out of community data? a conceptual framework and its implementation as models and software. *Ecology Letters*, 20(5):561–576, 2017.

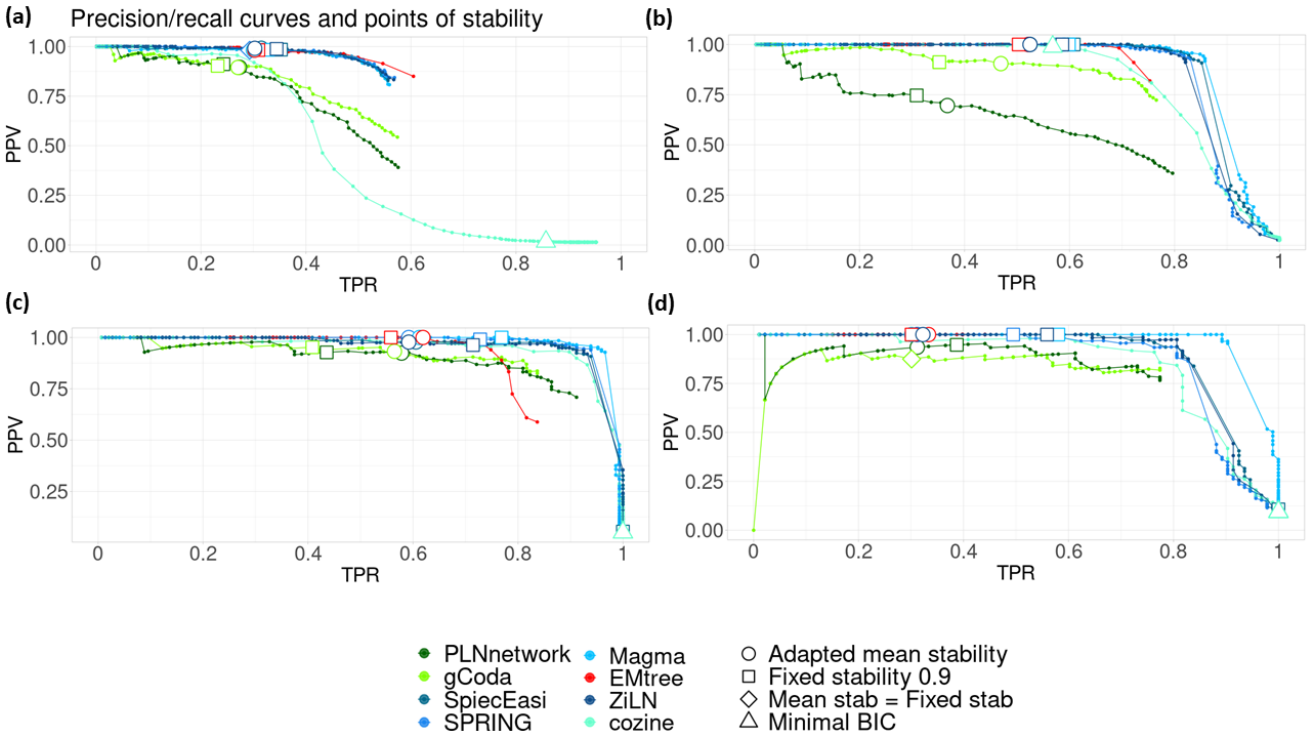


Supplementary Figure S10: Upset plot of the edges identified by the inference methods, the OneNet-* variants and the ground truth on the synthetic dataset for $n = 500$.

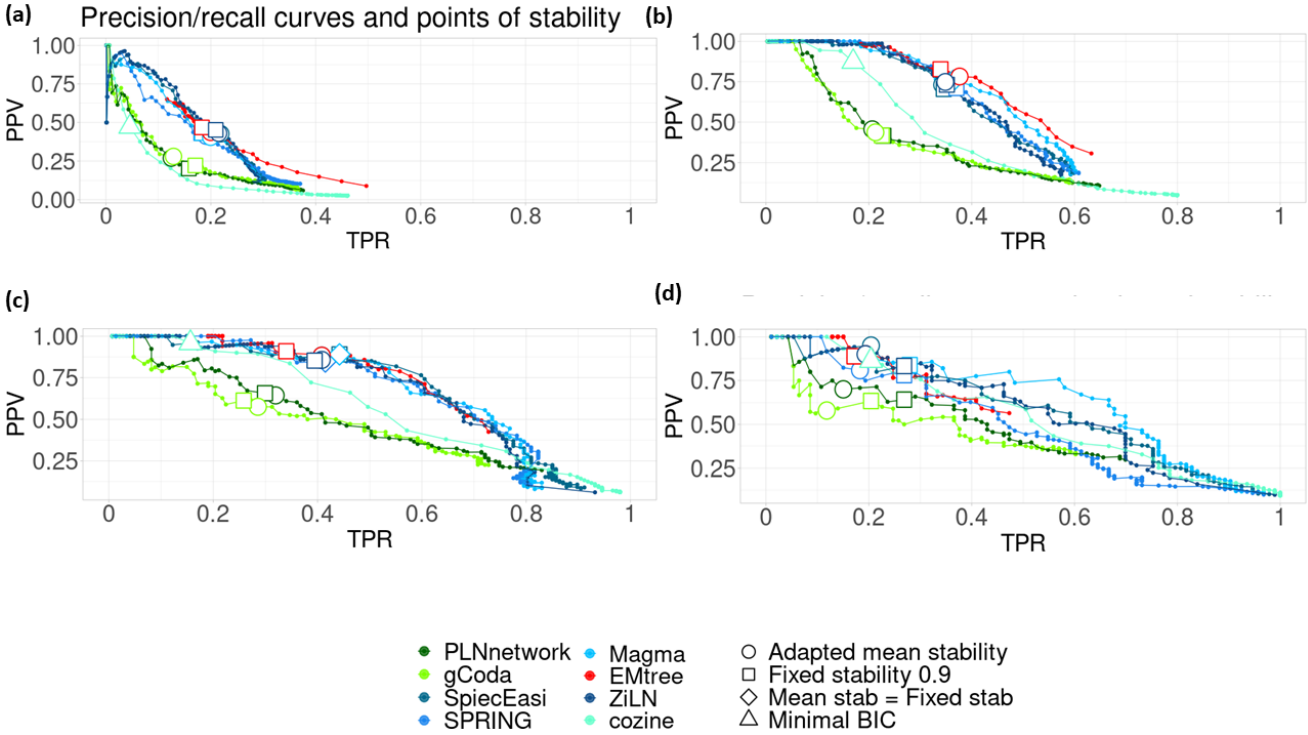
J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9 (3):432–441, 2008.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462, 2006.

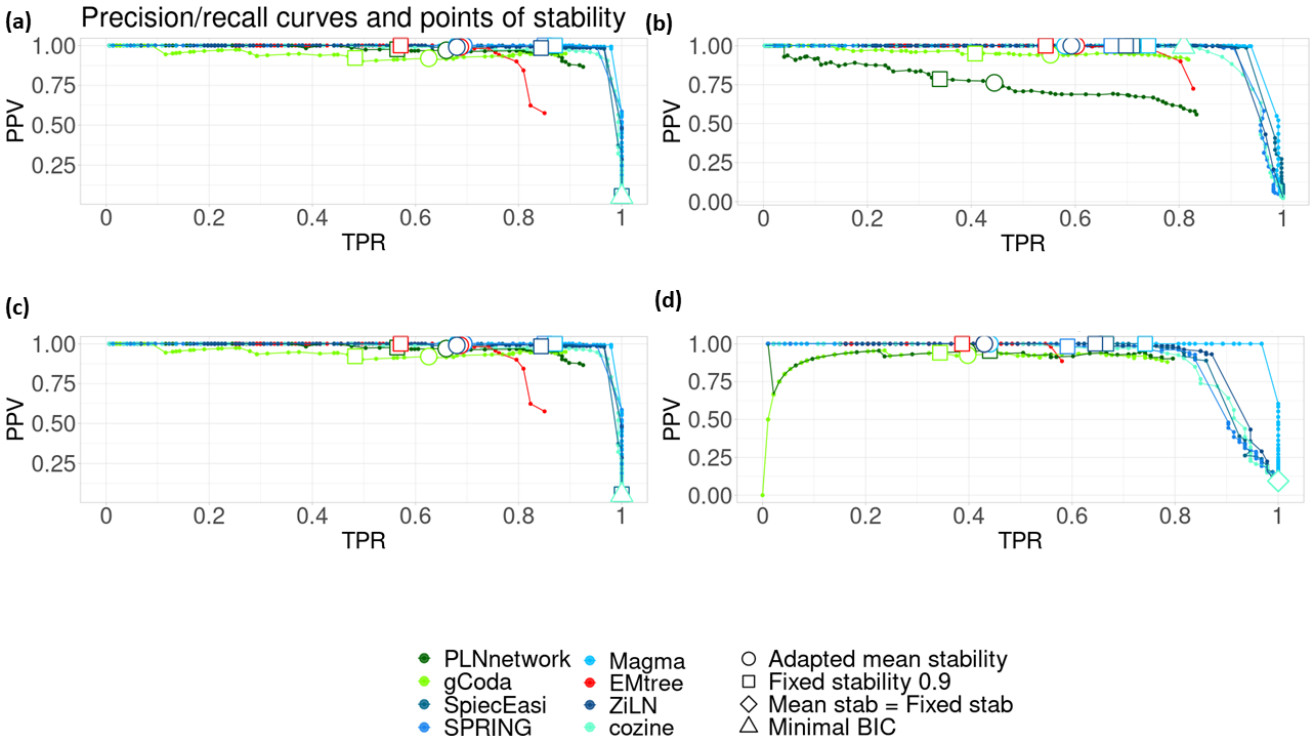
M. Meila and M. I. Jordan. Learning with mixtures of trees. *Journal of Machine Learning Research*, pages 1–48, 2000.



Supplementary Figure S11: Precision - recall curves of each method and TPR/PPV compromise chosen by stability, mean stability and BIC (see figure 3 for details) when $n = 500$ after filtering the dataset to keep only species with prevalence higher than a given threshold: (a)0.20 (b)0.50 (c)0.8 (d)0.9



Supplementary Figure S12: Same as Fig. S11 but for size $n = 100$.



Supplementary Figure S13: Same as Fig. S11 but for size $n = 1000$.