

Disentangling environmental effects in microbial association networks

Ina Maria Deutschmann (✉ ina.m.deutschmann@gmail.com)

Institute of Marine Sciences (ICM-CSIC) <https://orcid.org/0000-0002-3512-261X>

Gipsi Lima-Mendez

Louvain Institute of Biomolecular Science and Technology (IBST), University catholique de Louvain

Anders K. Krabberød

Department of Biosciences/Section for Genetics and Evolutionary Biology (EVOGENE), University of Oslo

Jeroen Raes

VIB Center for Microbiology, KU Leuven Department of Microbiology, Immunology and Transplantation, Rega Institute, Laboratory of Molecular Bacteriology

Sergio M. Vallina

Spanish Institute of Oceanography (IEO)

Karoline Faust

KU Leuven Department of Microbiology, Immunology and Transplantation, Rega Institute, Laboratory of Molecular Bacteriology

Ramiro Logares

Institute of Marine Sciences (ICM-CSIC)

Methodology

Keywords: microbial interactions, association network, effect of indirect dependencies, environmentally-driven edge detection

Posted Date: August 17th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-57387/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Microbiome on November 26th, 2021. See the published version at <https://doi.org/10.1186/s40168-021-01141-7>.

1 Disentangling environmental effects in microbial association
2 networks

3 Ina Maria Deutschmann^{1*} (ina.m.deutschmann@gmail.com), Gipsi Lima-Mendez²
4 (gipsi.lima-mendez@uclouvain.be), Anders K. Krabberød³ (a.k.krabberod@ibv.uio.no),
5 Jeroen Raes^{4,5} (jeroen.raes@kuleuven.vib.be), Sergio M. Vallina⁶
6 (sergio.vallina@oceanglobe.org), Karoline Faust^{5*†} (karoline.faust@kuleuven.be) and
7 Ramiro Logares^{1*} (ramiro.logares@icm.csic.es)
8

¹ Institute of Marine Sciences, CSIC, Passeig Marítim de la Barceloneta, 37, 08003, Barcelona, Spain.

² Louvain Institute of Biomolecular Science and Technology (IBST), Université catholique de Louvain, Croix du sud 4-5/L7.07.02, 1348, Louvain-la-Neuve, Belgium.

³ Department of Biosciences/Section for Genetics and Evolutionary Biology (EVOGENE), University of Oslo, p.b. 1066 Blindern, N-0316, Oslo, Norway.

⁴ VIB Center for Microbiology, Herestraat 49-1028, 3000, Leuven, Belgium.

⁵ KU Leuven Department of Microbiology, Immunology and Transplantation, Rega Institute, Laboratory of Molecular Bacteriology, Leuven, Belgium, Herestraat 49, 3000, Leuven, Belgium.

⁶ Spanish Institute of Oceanography (IEO), Ave Principe de Asturias 70 Bis, 33212, Gijon, Spain.

9

10 * Corresponding authors

11 † Shared last authors

12

13

14 Manuscript for: BMC Microbiome as Methodology paper

15

16

17 Abstract

18 **Background**

19 Ecological interactions among microorganisms are fundamental for ecosystem function, yet
20 they are mostly unknown or poorly understood. High-throughput-omics can indicate
21 microbial interactions by associations across time and space, which can be represented as
22 association networks. Links in these networks could result from either ecological interactions
23 between microorganisms, or from environmental selection, where the association is
24 environmentally-driven. Therefore, before downstream analysis and interpretation, we need
25 to distinguish the nature of the association, particularly if it is due to environmental selection
26 or not.

27

28 **Results**

29 We present EnDED (Environmentally-Driven Edge Detection), an implementation of four
30 approaches as well as their combination to predict which links between microorganisms in
31 an association network are environmentally-driven. The four approaches are Sign Pattern,
32 Overlap, Interaction Information, and Data Processing Inequality. We tested EnDED on
33 networks from simulated data of 50 microorganisms. The networks contained on average 50
34 nodes and 1,087 edges, of which 60 were true interactions but 1,026 false associations (i.e.
35 environmentally-driven or due to chance). Applying each method individually, we detected
36 a moderate to high number of environmentally-driven edges—87% Sign Pattern and Overlap,
37 67% Interaction Information, and 44% Data Processing Inequality. Combining these methods
38 in an intersection approach resulted in retaining more interactions, both true and false (32%
39 of environmentally-driven associations). The addition of noise to the simulated datasets did

40 not alter qualitatively these results. After validation with the simulated datasets, we applied
41 EnDED on a marine microbial network inferred from 10 years of monthly observations of
42 microbial-plankton abundance. The intersection combination predicted that 14.2% of the
43 associations were environmentally-driven, while individual methods predicted 31.4% (Data
44 Processing Inequality), 38.3% (Interaction Information), and up to 83.4% (Sign Pattern as
45 well as Overlap).

46

47 **Conclusions**

48 To reach accurate hypotheses about ecological interactions, it is important to determine,
49 quantify, and remove environmentally-driven associations in marine microbial association
50 networks. For that, EnDED offers up to four individual methods as well as their combination.
51 However, especially for the intersection combination, we suggest to use EnDED with other
52 strategies to reduce the number of false associations and consequently the number of potential
53 interaction hypotheses.

54

55 **Keywords:** microbial interactions; association network; effect of indirect dependencies;
56 environmentally-driven edge detection

57

58 **Background**

59 **Association networks to generate microbial interaction hypotheses**

60 There is a myriad of microorganisms on Earth: current estimates indicate $\approx 10^{12}$
61 microbial species [Locey and Lennon \(2016\)](#), and $\approx 10^{30}$ microbial cells [Kallmeyer et al.
62 \(2012\)](#); [Whitman et al. \(1998\)](#). Microorganisms have crucial roles in the biosphere by

63 contributing to global biogeochemical cycles [Falkowski et al. \(2008\)](#) and underpinning
64 diverse food webs. The importance of microbes for the functioning of ecosystems cannot be
65 understood without considering their ecological interactions [DeLong \(2009\)](#); [Krabberød et](#)
66 [al. \(2017\)](#). These allow transferring carbon and energy to upper trophic levels, and the
67 recycling of nutrients and energy [Worden et al. \(2015\)](#). Furthermore, ecological interactions
68 influence microbial community turnover and composition. These interactions include win-
69 win (e.g. mutual cross-feeding and cooperation), win-loss (e.g. predator-prey and host-
70 parasite), and loss-loss (e.g. resource competition) relationships [Faust and Raes \(2012\)](#).
71 Although microbial communities are highly interconnected [Layeghifard et al. \(2017\)](#), our
72 knowledge about ecological interactions in the microbial world is still limited [Bjorbækmo et](#)
73 [al. \(2019\)](#); [Krabberød et al. \(2017\)](#).

74 Previous studies have shown relationships between a restricted number of
75 microorganisms. However, we need a large number of interactions to understand the
76 functioning of such complex ecosystems. This is challenging, in part, due to the vast number
77 of possible interactions—given n microorganisms, there are $\binom{n}{2} = n(n-1)/2$ potential
78 pairwise interactions. Thus, it is unfeasible to test them experimentally within a reasonable
79 amount of time and cost. The problem of having a large number of potential interactions can
80 be partially circumvented with omics technologies coupled to network analyses.

81 Omics can identify and quantify a large number of microorganisms from a given
82 sample. Typically, the relative abundance for each identified organism per sample is
83 determined. There are multiple methods to determine associations (normally based on
84 correlations) between microorganisms using their abundances (e.g. eLSA [Xia et al. \(2011,](#)
85 [2012\)](#), CoNet [Faust and Raes \(2016\)](#), SPIEC-EASI [Kurtz et al. \(2015\)](#), or FlashWeave

86 [Tackmann et al. \(2019\)](#)). These abundance-based associations compose a network, where
87 nodes represent microorganisms and edges represent either co-presence (positive
88 association) or mutual exclusion (negative association) relationships, which constitute
89 microbial interaction hypotheses.

90

91 **Challenges in using networks as a representation of the microbial ecosystem**

92 Although networks play an essential role in understanding complex systems,
93 microbial ecological networks are not yet as developed in terms of inference and biological
94 interpretation [Lv et al. \(2019\)](#). Network inference from -omics data is difficult [Layeghifard](#)
95 [et al. \(2017\)](#); [Li et al. \(2016\)](#) because of both technical and interpretation challenges.

96 One common technical challenge is the compositional effect—microbial counts
97 become interdependent if normalized by the sample's total number of counts. There are
98 several tools [Li et al. \(2016\)](#) that correct for the compositional effect (e.g. SPIEC-EASI [Kurtz](#)
99 [et al. \(2015\)](#)). Other difficulties include data based on a small number of samples relative to
100 the number of microorganisms they contain (i.e. a low sample-to-microorganisms ratio); plus
101 sparse data—too many zeros in the dataset that can wrongly associate microorganisms
102 [Aitchison \(1981\)](#). A zero indicates either the absence of a microorganism (structural zero),
103 or an insufficient detection level or sequencing depth (sampling zero). Thus, we should
104 remove microorganisms that appear in just a few samples.

105 Interpretation of association networks is challenging because they are not equivalent
106 to ecological networks. Edges in ecological networks represent observed ecological
107 interactions between different microorganisms like parasitism or competition [Xiao et al.](#)
108 [\(2017\)](#). Ecological networks are directed graphs, where the directed edges (arcs) point from
109 a start node (source) to an end node (target). In contrast, association networks are undirected.

110 Although association networks provide ecological insight, they do not necessarily encode
111 causal relationships or observed ecological interactions. Unless we can verify edges, with
112 experiments or additional information, one should be careful when attributing biological
113 meaning to network properties [Röttjers and Faust \(2018\)](#). In sum, association networks
114 inferred from omics data are a great tool to generate microbial interaction hypotheses that
115 must be investigated experimentally. Interpretation and analysis are problematic if inferred
116 association networks are dense, i.e. have too many edges (so-called “hairball” networks). We
117 can obtain lower density networks when lowering the corrected p -value for inferred edges
118 [Weiss et al. \(2016\)](#), or increasing the cut-off for other criteria such as the association strength,
119 prevalence, or abundance filtering [Röttjers and Faust \(2018\)](#). Another strategy is
120 agglomeration using taxonomic or ecological (functional) groupings [Lima-Mendez et al.](#)
121 [\(2015\)](#).

122 The interpretation challenge addressed in this study is the so-called effect of indirect
123 dependencies caused by environmental factors. For most microbial association networks, an
124 edge indicates one of the following three alternatives:

- 125 1. ecological interaction between two microorganisms,
- 126 2. similar or contrary dependence (i.e. preference) to environmental factor/s or a third
127 microorganisms,
- 128 3. association by chance.

129 The effect of indirect dependencies occurs when two microorganisms are indirectly
130 associated because both are dependent on an abiotic environmental factor (e.g. have the same
131 requirements of nutrients and temperature) or biotic environmental factor (e.g. have the same
132 prey or predator). Here, indirect association describes the computational effect of indirect
133 dependencies, and observing an association when in fact there is none.

134

135 **Removing indirect dependencies including environmental effects**

136 To distinguish between direct and indirect interactions, several network construction
137 tools use a probabilistic graphical model [Kurtz et al. \(2015\)](#); [Yang et al. \(2017\)](#), e.g. SPIEC-
138 EASI [Kurtz et al. \(2019, 2015\)](#), miic [Verny et al. \(2017\)](#), or FlashWeave [Tackmann et al.](#)
139 [\(2019\)](#). FlashWeave can also integrate metadata to remove indirect associations driven by
140 environmental factors. The tool ARACNE [Margolin et al. \(2006\)](#) aims to eliminate indirect
141 associations by using an information theoretic property (the *Data Processing Inequality*, DPI,
142 in Methods). The extension TimeDelay-ARACNE [Zoppoli et al. \(2010\)](#) tries to extract
143 dependencies between different times. Another approach including time-delay is
144 implemented in the tool MIDER [Villaverde et al. \(2014\)](#), which combines mutual
145 information-based distances and entropy reduction to detect indirect interactions (*Mutual*
146 *Information*, MI, in Methods). PREMER [Villaverde et al. \(2018\)](#), an evolution of MIDER,
147 allows to include previous knowledge, e.g. known non-existent associations.

148 There are also several approaches to reduce indirect edges that are applied prior to
149 network construction, e.g. a high prevalence filter that preserves microorganisms present in
150 many samples [Pascual-García et al. \(2014\)](#). However, this will keep generalist
151 microorganisms while removing specialist microorganisms. Another approach to reduce
152 environmental effects is to divide datasets displaying a great environmental heterogeneity
153 into sub datasets of similar environmental conditions [Röttjers and Faust \(2018\)](#). For example,
154 a previous work [Mandakovic et al. \(2018\)](#) constructed two microbial co-occurrence networks
155 representing bacterial soil communities from two different sections of a pH, temperature, and
156 humidity gradient. Another work [Lima-Mendez et al. \(2015\)](#) constructed ocean depth-
157 specific networks to account for environmental differences between the surface layer and the

158 deep chlorophyll maximum layer. In addition to dividing samples, an algorithm that aims to
159 correct for habitat filtering effects [Brisson et al. \(2019\)](#), subtracts, for a given habitat, the
160 mean abundance from each microorganisms within each sample. However, this approach is
161 limited to the identified habitat groups that should have a similar sample size.

162 In contrast to the above, there are methods accounting for indirect dependencies after
163 network construction. For instance, global silencing, [Barzel and Barabási \(2013\)](#) and network
164 deconvolution [Feizi et al. \(2013\)](#) aim to recover true direct associations from observed
165 correlations. Both techniques are sensitive to missing variables [Alipanahi and Frey \(2013\)](#).
166 Another post network construction method, called *Sign Pattern*, SP, uses environmental
167 triplets [Lima-Mendez et al. \(2015\)](#). An environmental triplet contains two microorganisms
168 and one environmental factor, which are associated to each other. SP combines the signs of
169 association scores (positive or negative) to determine if a microbial association should be
170 classified as indirect (SP in Methods). Its major drawback is edge removal where
171 microorganisms with similar environmental preference interact. Along SP and network
172 deconvolution, the *Interaction Information*, II, was applied in [Lima-Mendez et al. \(2015\)](#).
173 Within an environmental triplet, the II method aims to distinguish whether an edge is due
174 entirely to shared environmental preferences ($II < 0$) or whether environmental preferences
175 and true interactions are entangled ($II > 0$). However, II cannot determine which of the three
176 associations in a triplet is indirect (II in Methods). Here, we study several indirect edge
177 detection methods: SP, *Overlap*, OL (developed here), II, DPI, as well as their combination.

178

179 **EnDED is an implementation of four methods and their combination**

180 This article presents EnDED, which implements four approaches, and their
181 combination, to disentangle the type of association represented by an edge in order to remove

182 the environmentally-driven (indirect) associations from the network. The four methods are:
183 Sign Pattern [Lima-Mendez et al. \(2015\)](#), Overlap (developed here), Interaction Information
184 [Ghassami and Kiyavash \(2017\)](#); [Lima-Mendez et al. \(2015\)](#), and Data Processing Inequality
185 [Cover and Thomas \(2001\)](#); [Margolin et al. \(2006\)](#). SP requires an association score that
186 represents co-occurrence when it is positive, and mutual-exclusion when it is negative. OL
187 requires temporal data with a known start and end of the association to determine whether
188 the microbial association occurs in a time window when both microorganism are associated
189 to the same environmental factor. The II method indicates the existence of one indirect
190 dependency between three components that are associated with each other. The DPI method
191 states that the association with the smallest mutual information is the indirect association.
192 Here, we evaluate each method as well as their combination in an intersection approach on
193 how well they detect environmentally-driven associations. By combining methods in an
194 intersection approach, we have retained more true positives than using each method on its
195 own. A union approach was discarded because it would have retained the smallest number
196 of true interactions. We are able to disentangle and filter environmentally-driven edges from
197 microbial association networks (0.95-0.96 in positive predictive value and 0.35-0.83 in
198 accuracy). EnDED contributed to both, generating more reliable hypotheses on microbial
199 interactions, and facilitating network analysis by removing edges from dense “hairball”
200 networks. EnDED is publicly available [Deutschmann et al. \(2019\)](#).

201

202 Results

203 Simulated data

204 To evaluate EnDED’s performance in removing environmentally-driven

205 associations, we simulated 1,000 abundance time-series datasets with 50 microorganisms and
206 known true interactions between them. We obtained another 1,000 datasets by introducing
207 noise to these time-series with Poisson distributions. We constructed the networks (below
208 called simulated networks) with the tool eLSA [Xia et al. \(2011, 2012\)](#) (see methods). The
209 simulated networks contained on average (computed as the median) 50 nodes and 1,087
210 edges (1,063 for data with noise; hereafter dwn), of which 60 (59 dwn) were true interactions
211 (edges present in the inferred and true network) and 1,026 (1,005 dwn) false associations
212 (edges present in the inferred but absent in the true network). A simple approach to
213 discriminate true interactions (desired) from false associations (undesired) would be to use a
214 threshold for the association strength, which could be suitable if the values for true
215 interactions and false associations are i) following different distributions, and ii) the
216 distributions are mainly non-overlapping. We tested the former requirement with a two-
217 sample Kolmogorov-Smirnov test with the R [R Core Team \(2019\)](#) function `ks.test`. Using a
218 95% (99%, 99.9%) confidence level, the distributions were significantly different for 358
219 (192, 66) simulated datasets and 355 (173, 68) simulated datasets with noise, which is slightly
220 more than one third of them. This indicates that an association strength cut-off is unsuitable
221 to separate true interactions from false associations. More sophisticated approaches than a
222 simple threshold include the methods implemented in EnDED: SP, OL, II, DPI, as well as
223 their combination.

224 Combining the methods in an intersection approach (hereafter referred to as
225 intersection combination), we classified on average 348 (228 dwn), that is 32% (22% dwn)
226 of the associations, to be environmentally-driven. The number of correctly detected false
227 associations was on average 332 (219 dwn), i.e. 96% of the removed edges. The resulting
228 networks contained on average 737 (828 dwn) edges. When each method was individually

229 applied more edges were removed: 87% (86% dwn) for SP and OL, 67% (60% dwn) for II,
230 and 44% (32% dwn) for DPI. The fraction of correctly removed edges was on average 95%
231 for each individual method. Individual methods removed more edges from the network than
232 the intersection combination, where all methods must agree. However, a method's
233 performance is not solely determined by the number of removed edges.

234 To evaluate the removal of environmentally-driven edges, we scored the different
235 approaches based on the true positive rate, TPR, true negative rate, TNR, false positive rate,
236 FPR, positive predicted value, PPV, and accuracy, ACC, (evaluation measurements see
237 Methods). In order to determine these measurements, we first determined true and false
238 positives, as well as true and false negatives. A true positive is a false association in the
239 network that is correctly removed by a method, and a false negative is a false association that
240 incorrectly is not removed. A false positive is a true interaction in the network that is
241 incorrectly removed by a method, and a true negative is a true interaction that correctly is not
242 removed by a method. The ideal method maximizes true positives and true negatives and
243 minimizes false positives and false negatives. The intersection combination under-performed
244 compared to each individual method when considering the TPR, FPR and ACC as shown in
245 Figure 1. However, applying each method individually has the drawback of removing more
246 true interactions. On average there are 60 (59 dwn) true interactions in the simulated
247 networks. The individual methods remove 86% (85% dwn) (SP), 85% (84% dwn) (OL), 60%
248 (51% dwn) (II), and 38% (28% dwn) (DPI). Therefore, although the intersection combination
249 removes fewer edges, it outperforms the others according to the TNR because it eliminates
250 fewer of the true interactions, 25% (16% dwn). We summarized the performance of EnDED
251 in Additional file Table S1. In Figure 1, we plotted the TPR against the FPR for each
252 simulated network and environmentally-driven edge detection methods as well as their

253 intersection combination. According to the PPV, intersection combination performs best and
254 SP and OL perform worst. SP and OL perform best according to TPR, FPR, and ACC. II
255 performs better than DPI according to TPR, FPR, and ACC. Regarding PPV, the former two
256 methods perform similarly on average. All methods have high PPV values with half of all
257 measured PPV above ≈ 0.95 .

258

259 **Real data**

260 After testing EnDED's performance on simulated networks, we applied it to a real
261 microbial association network, which was constructed from 10 years of monthly samples
262 from January 2004 to December 2013 at the Blanes Bay Microbial Observatory (BBMO)
263 [Gasol et al. \(2016\)](#). These samples included bacteria and eukaryotes of two size-fractions:
264 picoplankton (0.2-3 μm) and nanoplankton (3-20 μm). We estimated community
265 composition via metabarcoding of the 16S and 18S rRNA gene, and inferred an association
266 network, hereafter referred to as BBMO network (See Methods). The BBMO network
267 contained 844 nodes and 33, 832 edges before applying EnDED. The network contained
268 more positive than negative microbial associations (Figure 2).

269 By applying EnDED, we found that 28,210 of the network edges ($\approx 83.4\%$ of all
270 edges) were in at least one and in maximum nine environmental triplets (see Additional file
271 Table S2). The set of environmental factors included abiotic factors, e.g. temperature,
272 nutrients as well as cell counts (cells/ml) of heterotrophic prokaryotes, *Synechococcus*,
273 *Cryptomonas*, *Micromonas*, photosynthetic and heterotrophic nanoflagellates. Overall, we
274 detected 66,964 environmental triplets within the BBMO network. Of the 16 considered
275 environmental factors, PO_4^{3-} and salinity were not associated to any microorganism in the
276 network, and turbidity, NH_4^+ and *Cryptomonas* were not found within a triplet. The influence

277 of the remaining 11 environmental factors affecting microbial associations is displayed in
278 Figure 2. Temperature and day length (hours of light) are the top two environmental factors
279 affecting microbial associations followed by *Micromonas*, photosynthetic and heterotrophic
280 nanoflagellates.

281 The intersection combination of the four methods removed 4,806 ($\approx 14.2\%$)
282 associations from the BBMO network(see Figure 2). When analysing these indirect edges,
283 we discovered that over 51.6% are between bacteria, 38.5% between bacteria and eukaryotes,
284 and 10% between eukaryotes. Figure 2 shows the number of edges and fraction of
285 environmentally-driven edges between the two domains. Considering size fractions, these
286 environmentally-driven edges correspond 19.5% to picoplankton, 44.2% to nanoplankton,
287 and 36.3% to edges between size fractions. We have summarized the number of associations
288 in Additional file Table S3. Compared to the intersection combination approach, each method
289 individually removed more edges: 83.4% (SP and OL removing all microbial edges present
290 in a triplet), 38.3% (II), and 31.4% (DPI); that is, between a factor of x2 and x6 larger
291 removal.

292 We also determined for each association the Jaccard index, also known as Jaccard
293 similarity coefficient (See Methods), which indicates how often two microorganisms appear
294 together in the dataset. Although we are aware of time-delayed interactions and eLSA [Xia et](#)
295 [al. \(2011, 2012\)](#) could account for them, we did not take them into consideration for our
296 BBMO dataset, as we consider our sampling interval is too large (1 month) for inferring
297 associations that may have a solid ecological basis. Thus, in our study, we focused on
298 contemporary interactions between co-occurring microbes. We found that only 29.8% of
299 indirect associations have a Jaccard index above 0.5, i.e. microorganisms appeared together
300 over 50% of the time, compared to 63.3% of the associations that appeared in at least one

301 triplet but were not removed from the BBMO network (see Table 1). We assume that two
 302 microbes that appear together < 50% of the time are less likely to have true contemporary
 303 ecological interactions and the corresponding association is more likely to be false. The fact
 304 that over 70% of environmentally-driven associations have a Jaccard index equal or below
 305 0.5 strengthens the decision of their removal. When considering the sign of an association,
 306 we found that only 3% of negative associations obtain a Jaccard index over 0.5, compared to
 307 68% of the positive associations.

308 In the BBMO network, the intersection combination approach removed roughly the
 309 same number of negative and positive edges, 2,263 and 2,543, respectively (see Figure 2).
 310 The pre-EnDED network contained 81.9% positive and only 18.1% negative edges, so the
 311 method removed 41.5% of the negative and only 8.2% of the positive edges. If we randomly
 312 removed 4,806 edges, we would expect 18.1% to be negative (i.e. 870) and 81.9% of them
 313 to be positive (i.e. 3,936). If we restrict these calculations to the 28,210 microbial associations
 314 that were found in at least one environmental triplet, with 22,492 of them being positive and
 315 5,718 being negative, we would expect to remove 20.3% (i.e. 976) of negative and 79.7%
 316 (i.e. 3, 830) of positive edges. The probability of randomly removing an equal number of
 317 positive and negative associations is nearly zero, since it follows a multivariate
 318 hypergeometric distribution:

$$P(k_{neg}, k_{pos}) = \frac{\binom{N_{neg}}{k_{neg}} \cdot \binom{N_{pos}}{k_{pos}}}{\binom{N}{n}}, \quad \text{Eq. (1)}$$

319 where N_{pos} and N_{neg} are the number of positive and negative associations in the network,
 320 respectively, k_{pos} is the number of removed positive and k_{neg} the removed negative
 321 associations from the network, N is the number of associations in the network, and n is the

322 number of removed associations from the network. The intersection combination removing
323 an equal number of positive and negative edges, indicates that the removal was not random
324 and that the removal is biased towards negative associations.

325 In order to evaluate the performance of EnDED on the BBMO network, we
326 considered interactions described in literature and collected in the Protist Interaction
327 Database (PIDA) [Bjorbækmo et al. \(2019\)](#). In order to use these known interactions, we
328 taxonomically classified the microbes in the BBMO network (see methods). Studies typically
329 compare the associations of a network to those reported in the literature at the genus level
330 [Lima-Mendez et al. \(2015\)](#). The ambiguity in taxonomic classification and the large number
331 of edges challenge this comparison. Thus, we implemented a function to compare strings and
332 match the taxonomic classification of a microorganism in the BBMO network to those in the
333 scientific literature (PIDA). We found that only 31 (< 0.09%) associations were supported
334 by interactions described in the literature (see Table 2). That is, 99.91% of associations in the
335 BBMO network (before applying EnDED) could not be used to evaluate EnDED's
336 performance. These 31 associations describe nine unique interactions between 9
337 microorganisms, and 18 edges were in an environmental triplet to which each method as well
338 as their combination were applied (see summary in Table 2). Ideally none of these described
339 associations should be removed by EnDED. Yet, the intersection combination removed 8
340 associations: one between a diatom (*Thalassiosira*) and a dinoflagellate (*Heterocapsa*), and
341 seven associations between a diatom (*Thalassiosira*) and an unknown *Flavobacteriia*. In
342 contrast and even worse, SP and OL removed all 18 edges, II 11 and DPI 14 edges. The
343 additionally removed edges by individual methods are associations between a diatom
344 (*Thalassiosira*) and an unknown *Flavobacteriia*. Considering only the genus level, there
345 were 179 unique microbes (level genus) in the BBMO network, and 700 in PIDA, combined

346 there were 843 microorganisms, and 36 microorganisms are in both. Thus, 20.1% of
347 microorganisms found in the BBMO network were also in PIDA, and 5.1% of
348 microorganisms found in PIDA were also found in the BBMO network. Regarding
349 interactions, and not considering prokaryote-prokaryote associations, there were 1,266
350 unique interactions in PIDA, but 3,422 in the BBMO network (considering only genus). Only
351 5 unique interactions are in both, i.e. 0.39% of PIDA interactions were found in the BBMO
352 network, and 0.15% of BBMO associations were found in PIDA.

353

354 Discussion

355 Using EnDED to disentangle environmental effects in microbial association networks

356 EnDED

357 EnDED makes several indirect-edge removal techniques accessible to microbial
358 ecologists and does not require previous programming experience. These techniques can be
359 used individually or combined. In addition, this work systematically evaluates the different
360 techniques and their combination to remove indirect edges from microbial association
361 networks. Here, we tested only the union and intersection combination of all four methods,
362 but other combination strategies are possible to obtain with EnDED. EnDED requires the
363 data of the environmental factors in order to predict if an association is environmentally-
364 driven, but we understand that it may be impossible to consider all environmental factors [Ly](#)
365 [et al. \(2019\)](#). Despite this limitation, EnDED can perform well if the major environmental
366 factors, such as, e.g. temperature and nutrient concentrations for marine microbes, are
367 provided. Moreover, knowledge of microbial interactions in nature is rather limited and
368 therefore determining the performance of EnDED for real networks is challenging and carries

369 some degree of uncertainty. Thus, the analysis of EnDED's results without previous
370 validation should be interpreted with care. Here, we first applied EnDED on simulated
371 networks in order to measure the performance of individual methods as well as their
372 intersection combination. Then, we applied EnDED to an association network constructed
373 from observational data. We used monthly data from the BBMO marine time-series [Gasol et](#)
374 [al. \(2016\)](#), which provided 10 years of microbial abundance data and 16 environmental
375 factors.

376 For the simulated networks (and simulated networks with noise), we found that each
377 method individually removed on average a moderate to high number of edges: 44% (32%
378 dwn) DPI, 67% (60% dwn) II, and 87% (86% dwn) SP and OL. The intersection combination
379 of the four methods removed on average 32% (22% dwn) of the edges, while also keeping
380 more true interactions. To understand the impact of the environment, Röttjers and Faust
381 simulated an increasing environmental influence and observed a decrease in retrieving true
382 interactions from inferred associations, i.e. a decrease in precision [Röttjers and Faust \(2018\)](#).
383 They also compared several microbial correlation network construction methods for cross-
384 sectional data, including CoNet [Faust et al. \(2012\)](#), SparCC [Friedman and Alm \(2012\)](#),
385 SPIEC-EASI [Kurtz et al. \(2015\)](#), and Spearman correlations—all exhibited a reduced
386 precision on data with simulated environmental effects. For our simulation networks, we
387 observed a slight increase in precision when removing environmentally-driven associations.
388 In summary, intersection combination removed the smallest edge number, outperforming
389 individual methods in terms of number of remaining true interactions, TNR, and PPV.
390 Regarding TPR and ACC, all methods performed better on simulated networks without noise.
391 Fewer edges have been removed from these networks along with fewer true interactions. All
392 approaches performed similarly according to the PPV, which could be a result of the removal

393 of fewer edges and also fewer true interactions.

394 In our BBMO dataset, the intersection combination removed 14.2% of the edges—
395 41.5% of the negative and only 8.2% of the positive edges. We argue that several negative
396 associations are probably due to different environmental preference (different niches) of
397 microbes. The Jaccard index representing a level of microbial co-occurrence, scored equal or
398 below 50% for 97% of the negative associations. These may partially represent microbes
399 adapted to different seasons. Previous work on the eukaryotic pico- and nano-plankton at the
400 BBMO, using the same basal 10-year dataset used here, indicated a strong seasonality at the
401 community level [Giner et al. \(2019\)](#).

402

403 **Comparisons of indirect edge detection on other datasets**

404 In our BBMO network we found 28,210 (83.4%) microbial edges that were within at
405 least one environmental triplet. Thus, the fraction was 2.6 times higher than what was found
406 for an association network called “global plankton interactome” containing 29,912 (32.3%)
407 edges, associated to microbes as well as small metazoans, that were within at least one
408 environmental triplet [Lima-Mendez et al. \(2015\)](#). In the previous study 29,900 ($\approx 100\%$ of
409 triplets and 32% of all edges) were attributed

410 to environmental factors by SP. II indicated 11,043 environmentally-driven edges ($\approx 37\%$ of
411 triplets and 12% of all edges) with p -value below 0.05 in a permutation test with 500
412 iterations. Network deconvolution suggested 22,439 environmentally-driven edges ($\approx 75\%$ of
413 triplets and 24% of all edges). These three methods agreed for 8,209 edges ($\approx 27\%$ of triplets
414 and 8.9% of all edges). In comparison, we found more environmentally-driven associations
415 for the BBMO network (14.2% of all edges)

416 based on a decade of temporal data from one location and one depth including two size

417 fractions than what was detected for the global ocean interactome covering two depths, 68
418 stations around the world and various size fractions [Lima-Mendez et al. \(2015\)](#). These
419 differences suggest that microbial temporal turnover may induce more indirect edges than
420 spatial turnover. Thus, the effects of indirect dependencies may depend on dataset type (e.g.
421 temporal vs. spatial), and this should be further investigated.

422 Using II for the BBMO network, we removed 38.3% of the edges (45.9% when
423 considering only triplets), which would indicate a moderate number of associations explained
424 by an environmental factor. The DPI also identified a moderate number (31.4%, 37.6% when
425 considering only triplets) of environmentally-driven associations in the BBMO network,
426 whereas SP or OL identified a ubiquitous number of environmentally-driven edges (83.4%,
427 100% when considering only triplets). This indicates that SP and OL are strict and should be
428 used in combination with other methods in an intersection approach. In another study, the
429 tool FlashWeave [Tackmann et al. \(2019\)](#) predicted direct microbial interactions in the human
430 microbiome using the Human Microbiome Project (HMP) dataset, including heterogeneous
431 microbial abundance data of 68, 818 samples [The Human Microbiome Project Consortium:
432 Huttenhower et al. \(2012\)](#). The inferred networks (with and without metadata) were sparser
433 than our networks. The network with metadata contained 10.7% fewer associations compared
434 to the network without metadata, which suggests a minor number of environmentally-driven
435 edges in the tested dataset, slightly less than in our results from BBMO. Considering the
436 previously mentioned comparison between the number of indirect edges detected in BBMO
437 (higher) vs. the ocean interactome (lower), it remains to be tested whether FlashWeave would
438 detect more indirect edges in temporal than in spatial or in environmental vs. host-associated
439 datasets.

440

441 **Factors causing indirect microbial associations**

442 From the simulated networks, we found that using the intersection combination
443 instead of each method individually, we maintained more true interactions at the cost of more
444 false associations in the network—more when considering simulated networks including
445 noise. Comparing our simulated network against the BBMO network, the intersection
446 combination classified a higher number of edges as environmentally-driven in the simulated
447 networks 32% (22% dwn) than in the BBMO network (14.2%). For the simulated data, we
448 previously knew the environmental factor influencing pairwise microbial associations. For
449 the BBMO data, we used 16 available environmental factors, but not all factors that could
450 affect microbial dynamics. Even though the most important factors influencing microbial
451 seasonal dynamics at BBMO were considered [Giner et al. \(2019\)](#), there are several other
452 environmental factors that were not measured and that could generate indirect edges. The
453 indirect edges associated to these factors were not detected in our analyses. Similarly, indirect
454 edges associated to biotic interactions (e.g. two bacteria sharing a positive edge as they are
455 symbionts in the same protists) were not considered. Future sampling for microbial
456 interaction research should expand metadata collection in order to detect more abiotic and
457 biotic factors that could generate indirect edges. While we identified temperature and day
458 length (hours of light) to be the top two environmental factors affecting microbial
459 associations in the BBMO network, followed by *Micromonas*, photosynthetic and
460 heterotrophic nanoflagellates, the frequent environmental factors in the global plankton
461 interactome [Lima-Mendez et al. \(2015\)](#) were phosphate concentration, and temperature,
462 followed by nitrite concentration, and mixed-layer depth. Although we considered PO_4^{3-} it
463 was not associated to any microorganism in the network along with salinity, which could be
464 explained by the fact that these environmental factors were more homogenous in our BBMO

465 dataset. For instance, the standard deviation in BBMO dataset was < 1 for PO_4^{3-} and salinity
466 in contrast to Tara samples [Lima-Mendez et al. \(2015\)](#), where it was about 20-30 when
467 considering all samples. During the Malaspina 2010 Circumnavigation Expedition, the
468 concentrations of trace metals were determined for 110 surface water samples [Pinedo-](#)
469 [González et al. \(2015\)](#). The previous study indicates relationships between primary
470 productivity and trace nutrients, more specifically for the Indian Ocean Cd, the Atlantic
471 Ocean Co, Fe, Cd, Cu, V and Mo, and the Pacific Ocean Fe, Cd, and V. Thus, trace metals
472 may play an important role in regulating oceanic primary productivity.

473

474 **Limitations of EnDED**

475 It may be promising to update EnDED—currently using environmental triplets—to
476 allow any closed triplet, (i.e. three microorganisms being all connected), as done with gene
477 triplets [Margolin et al. \(2006\)](#). A recent update of the network construction tool eLSA [Xia et](#)
478 [al. \(2011, 2012\)](#) permits to examine how a factor, such as a microorganism or environmental
479 variable, mediates the association of two other factors [Ai et al. \(2019\)](#), which allows the study
480 of interactions between three factors. Furthermore, triplets limit the study to first-order
481 indirect dependencies, neglecting higher-order indirect dependencies. Such limitation was
482 solved for the DPI method by examining associations in quadruplets, quintuplets, and
483 sextuplets [Jang et al. \(2013\)](#). Implementing higher-order DPI and adjusting the other three
484 methods to account for higher-order indirect dependencies may be promising but one need
485 to be aware that incorporating higher-order dependencies will also increase the risk of over-
486 fitting. Further, all relevant environmental factors could be incorporated into the calculation
487 of II, which would combine several environmental triplets. However, we reason that such
488 adjustments would require a larger sample size. Both II and DPI methods calculate MI that

489 measures the dependence between two random variables. EnDED is limited by including one
490 function to estimate the MI. A comparison of four different estimates of MI revealed that
491 obtaining the true value of MI is not straightforward, and minor variations of assumptions
492 yield different estimates [Fernandes and Gloor \(2010\)](#). Lastly, the conditional mutual
493 information, CMI, which quantifies nonlinear direct relationships among variables, can be
494 underestimated if variables have tight associations in a network [Zhao et al. \(2016\)](#). The so-
495 called part mutual information, PMI, measurement can help overcome CMI's
496 underestimations. Although using PMI instead of CMI looks promising, calculating PMI is
497 computationally more demanding [Zhao et al. \(2016\)](#).

498

499 **Future Perspective**

500 In this study, we have shown that EnDED with an intersection combination approach
501 provides less dense networks, but still with many potential interactions. Specific associations
502 may be validated with experiments or microscopy [Krabberød et al. \(2017\)](#); [Lima-Mendez et al. \(2015\)](#).
503 However, we suggest to first further reduce the set of potential interaction
504 hypotheses. To improve the selection of interaction hypotheses, we propose to score
505 associations based on re-occurrence: in time, as done with microbial abundance seasonality
506 [Giner et al. \(2019\)](#), or space, where an association appears in different networks based on
507 different datasets, or different regions of the world. In a previous study using 313 samples,
508 including seven size-fractions, four domains (Bacteria, Archaea, Eukarya, and viruses), and
509 two depths from 68 stations across eight oceanic provinces, 14% of the 81,590 predicted
510 biotic interactions were identified as local [Lima-Mendez et al. \(2015\)](#). Thus, re-occurrent
511 associations suggest a higher likelihood that the association represents a true ecological
512 interaction, reducing the number of interaction hypotheses to the strongest ones. Another

513 strategy to shortlist interaction hypotheses is to incorporate additional data into the network
514 and use a multi-layer network approach. Such data could be environmental preferences such
515 as temperature or salinity optima, size of cells, presence of chloroplasts, or more
516 sophisticated data obtained from High-Throughput Cultivation [Faust \(2019\)](#), microbial
517 community transcriptomes that reveal microbes and metabolic pathways [McCarren et al.](#)
518 [\(2010\)](#), or interactions inferred from Single-Cell genome data [Krabberød et al. \(2017\)](#); [Yoon](#)
519 [et al. \(2011\)](#).

520

521 Conclusion

522 In this paper, we present EnDED, an analysis tool to reduce the number of
523 environmentally induced indirect edges in inferred microbial networks. We applied EnDED
524 to networks based on time-series of simulated data and observed marine microbial abundance
525 data. Our simulated networks indicate that false associations, driven by environmental
526 variables instead of true interactions, are ubiquitous in inferred association networks.
527 However, EnDED's intersection combination classified a minority of associations as
528 environmentally-driven in a real (BBMO) network. Depending on the single method used,
529 we classified a moderate to high number of associations as environmentally-driven in the
530 same network. Nevertheless, associations driven by environmental factors must be
531 determined and quantified to generate more accurate insights regarding true microbial
532 interactions. EnDED provides a step forward in this direction.

533

534 Methods

535 **Simulated dataset: time series based on an adjusted generalized Lotka-Volterra model**

536 We simulated a time series using an adjusted version of the standard *generalized*
537 *Lotka-Volterra model*, gLV. The standard gLV generates simulated data for evaluation
538 [Bashan et al. \(2016\)](#); [Berry and Widder \(2014\)](#). The gLV can describe the dynamics of
539 microbial communities, by including a first order approach of the microbial interactions. The
540 model's simplicity arises from the assumption of linear interactions, which facilitates
541 implementation and allows fast numerical simulations. The gLV has, however, several
542 limitations [Gonze et al. \(2018\)](#). For example, gLV neglects higher-order interactions and the
543 additivity of interaction strengths is a weakness because they may be combined in different
544 ways. Also, interactions are often assumed to be constant parameters, but a reducing level of
545 a nutrient may weaken cross-feeding relationships. Moreover, gLV omits the influence of
546 environmental factors, which, for example, can induce oscillations in natural communities
547 [Beninca` et al. \(2011\)](#). Using a model that accounts for nutrients [Kettle et al. \(2018\)](#) is more
548 realistic but also more complex. More elaborate mechanistic models of microbial dynamics
549 than gLV solve explicitly the global cycling of nutrients and are coupled to the oceanic
550 circulation (see [Vallina et al. \(2019\)](#) for a review), but the added complexity can hamper
551 understanding about the ecological interactions among microorganisms when compared to a
552 simpler gLV approach. Thus, we chose to use a simpler extension of the gLV to account for
553 the influence of environmental factors [Dam et al. \(2016\)](#); [Stein et al. \(2013\)](#). In order to allow
554 the growth rates to vary when the environmental variables change, environmental variables
555 can be incorporated directly into the gLV [Dam et al. \(2016\)](#); [Röttjers and Faust \(2018\)](#). We
556 simulated a time series using the Klemm-Eguíluz algorithm [Klemm and Eguíluz \(2002\)](#), and
557 an adjusted gLV. We adjusted the model by defining microorganisms growth rates as a
558 function dependent on one seasonal abiotic environmental factor, and added an abiotic
559 environmental factor in the interaction matrix. We then used the time series generated by the

560 gLV to obtain temporal microbial abundance data. With this simulated data, we inferred a
561 network that contained environmentally-driven associations, needed to evaluate the
562 performance of EnDED. We repeated this procedure 1,000 times to obtain a large set of
563 simulated networks, and then used the determined abundance tables and Poisson distribution
564 to obtain another 1,000 simulated networks including noise. The addition of noise was done
565 by randomly drawing an abundance from the Poisson distribution with λ equaling the original
566 abundance of a specific microorganisms to a specific time.

567

568 Adjusting the gLV

569 To evaluate EnDED, we simulated a time series of microbial abundances with a gLV
570 including true pairwise interactions between 50 microorganisms and adjusted it by
571 incorporating two environmental factors:

$$\frac{dy(t)}{dt} = y(t)[b + Ay(t)], \quad \text{Eq. (2)}$$

572 where t is time, $dy(t)/dt$ is the rate of change of microbial abundances as a column vector,
573 $y(t)$ is the vector of microbial abundance at time t , b is the growth rate vector determined
574 through microorganisms specific growth rate functions that depend on an environmental
575 factor (see equation (4)), and A is the interaction matrix.

576

577 Interaction matrix

578 In the interaction matrix A , each coefficient a_{ji} provides the linear effect that a change
579 in the abundance of microorganism i has on the growth of microorganism j [Novak et al.](#)
580 [\(2016\)](#). We simulated the interaction coefficients a_{ji} with the Klemm-Eguíluz algorithm
581 [Klemm and Eguíluz \(2002\)](#), which generates a modular and scale-free matrix. We also set

582 the interaction probability to 0.01, the percentage of positive coefficients to 30%, and
 583 diagonal coefficients to zero. Negative diagonal coefficients a_{ji} (i.e. the interaction of a
 584 microorganism with itself) can represent intra-specific competition and provides the carrying
 585 capacity for each microorganism, preventing its explosive growth [Haydon \(1994\)](#). We set the
 586 diagonal coefficients $a_{ii} = -0.5$ to avoid excessive microbial abundances in the simulations.

587

588 Two abiotic environmental factors

589 We adjusted the gLV by including two environmental factors. For simplicity, we
 590 assume no feedback between the microorganisms and the environmental factors. That is, the
 591 environmental factors affect the growth of the microorganisms but not vice-versa. The first
 592 environmental factor affects the specific growth rate of each microorganism by interacting
 593 with two of their traits: optimal environmental value for growth and tolerance range of
 594 environmental values. We simulated the environmental factor using a periodic sinusoidal
 595 function (see equation (3)), rounded to 3 digits:

$$\epsilon(t) \triangleq \text{round}(\sin(\omega \cdot t), \text{digits} = 3), \quad \text{Eq. (3)}$$

596 where t is the time axis (months), $\omega = (-2\pi/T)$ is the signal frequency (radians) and $T =$
 597 12 is the signal periodicity (months); resulting in a signal phase shift of $T/4$ (months). While
 598 the first environmental factor is considered to be “external” to the microbial community, the
 599 second environmental factor is considered to be “internal”, and therefore it is included in the
 600 interaction matrix. The interaction coefficients between the microorganisms and the second
 601 environmental factor were generated by splitting the microorganisms into two groups: the
 602 second abiotic environmental factor influenced positively one half and negatively the other
 603 half of the microorganisms. We obtained the interaction coefficients from two uniform

604 distributions defined to range between [-0.8, -0.2] and [0.2, 0.8] respectively. As the
 605 microorganisms did not influence the abiotic factor, the corresponding interaction
 606 coefficients were set to zero.

607

608 Species growth rate

609 The external seasonal abiotic environmental variable affects the growth rate, g , of
 610 each microorganism. This dependency is given by:

$$g(t) \triangleq g_{max}^2 \exp\left(-\frac{1}{2} \frac{(\epsilon_{opt} - \epsilon(t))^2}{\sigma^2}\right), \quad \text{Eq. (4)}$$

611 where $E(t)$ is the environmental parameter that affects the microorganisms growth rate $g(t)$
 612 at time t , g_{max} is the microorganism' specific maximum growth rate that determines the
 613 amplitude of the growth-rate curve, ϵ_{opt} is the microorganism' specific optimal
 614 environmental value that determines the peak of the growth-rate curve, and σ is the
 615 microorganism' specific ecological tolerance (niche width) determining the environmental
 616 range in which the microorganism grows, which determines the length (niche spread) of the
 617 growth-rate curve. We obtained the two constant parameters g_{max} , and σ for each
 618 microorganism from a uniform distribution ranging between 0.3 and 1 to assure positive
 619 values. The values ϵ_{opt} were drawn from a uniform distribution ranging between the minimal
 620 and maximal value of the seasonal environmental factor. We defined the internal abiotic
 621 environmental factor, which is included in the interaction matrix, through the same function
 622 with $g_{max} = 0.8$, $\epsilon_{opt} = 0.5$, and $\sigma = 0.5$. Since the growth rates depend on the
 623 environmental factor, they vary seasonally. Different microorganisms will grow better or
 624 worse at different times of the year following their environmental niches. This will lead to

625 an asynchrony of their growth rate responses to the environment that will translate into an
626 asynchrony of their abundances in time.

627

628 Initial abundances

629 To obtain the microbial abundances in time with the adjusted gLV, we simulated the
630 initial microbial abundances with a stick-breaking process such that abundances add up to 1,
631 using the function `bstick` [Jackson \(1993\)](#); [Legendre and Legendre \(2012\)](#), and the package
632 `vegan` [Oksanen et al. \(2019\)](#). We generated uneven initial microbial abundances without
633 introducing zeros and set the initial value for the internal abiotic environmental factor
634 included in the interaction matrix to 0.001.

635

636 Species abundances in time

637 Once we have set the initial conditions, we simulated microbial abundances over time
638 by solving the equations given in the adjusted gLV (see equation (2)). Start time was 0, end
639 time 49.5, and sample resolution 0.5 resulting in 100 samples. We used the solver function
640 `lsoda` [Soetaert et al. \(2010\)](#). The simulated abundances in time were used to construct an
641 association network, which is referred to as the simulated network.

642

643 **Real dataset: Blanes Bay Microbial Observatory time series**

644 Microbial abundances

645 Surface water (< 1m depth) was sampled monthly from January 2004 to December
646 2013 at the BBMO in the North-Western Mediterranean Sea [Gasol et al. \(2016\)](#). About 6L
647 of seawater were filtered and separated into two size fractions, picoplankton (0.2-3 μm) and
648 nanoplankton (3-20 μm), as described in [Giner et al. \(2019\)](#). Community DNA was extracted,

649 and the 18S ribosomal RNA-gene (V4 region) was amplified [Giner et al. \(2019\)](#). The 16S
650 ribosomal RNA-gene (V4 region) was also amplified from the same DNA extracts using the
651 primers Bakt 341F [Herlemann et al. \(2011\)](#) and 806R [Apprill et al. \(2015\)](#). DADA2 v1.10.1
652 [Callahan et al. \(2016\)](#) was used for read quality control, trimming and inference of
653 Operational Taxonomic Units (OTUs) as Amplicon Sequence Variants (ASVs). In both
654 microbial eukaryotes and prokaryotes, the maximum number of expected errors (MaxEE)
655 was set to 2 and 4 for the forward and reverse reads, respectively. OTU tables were generated
656 for both the 16S and 18S rRNA genes. Before network construction all samples were
657 individually subsampled using the function `rrarefy`, in R package `vegan` [Oksanen et al.](#)
658 [\(2019\)](#), to the size of the sample with the lowest sequencing depth (4,907 reads). Due to
659 suboptimal sequencing of the amplicons from some months, we did not use nanoplankton
660 samples from the period May 2011 to July 2012 (27 samples) as well as 4 additional samples.
661 OTU abundances for the missing samples were estimated using seasonally aware missing
662 value imputation by weighted moving average for time series as implemented in the R
663 package `imputeTS` [Moritz and Gatscha \(2017\)](#).

664

665 Taxonomic classification

666 The taxonomic classification of each OTU was inferred with the naïve Bayesian
667 classifier method [Wang et al. \(2007\)](#) together with the SILVA version 132 [Quast et al. \(2012\)](#)
668 database as implemented in DADA2 [Callahan et al. \(2016\)](#). In addition eukaryotic
669 microorganisms were BLASTed [Altschul et al. \(1990\)](#) against the Protist Ribosomal
670 Reference database [PR2, version 4.10.0; [Guillou et al. \(2012\)](#)]. If the taxonomic assignment
671 for eukaryotes disagreed between SILVA and PR2, we used the PR2 classification. We
672 removed microorganisms identified as either Metazoa, or Streptophyta, plastids and

673 mitochondria. In addition, we removed Archaeas since primers were not optimal for
674 recovering this domain.

675

676 Environmental factors

677 We measured environmental factors that may affect the ecosystem's dynamics. We
678 considered a total of 16 contextual abiotic and biotic variables: temperature (C_4°), turbidity
679 (Secchi depth m), salinity, total chlorophyll ($\mu\text{g/l}$), inorganic nutrients— PO_4^{3-} (μM), NH_4^+
680 (μM), NO_2^- (μM), NO_3^- (μM), and SiO_2 (μM)—heterotrophic prokaryotes (cells/ml),
681 *Synechococcus* (cells/ml), *Cryptomonas* (cells/ml), *Micromonas* (cells/ml), photosynthetic
682 nanoflagellates (cells/ml), heterotrophic nanoflagellates (cells/ml), day length (hours of light)
683 [Giner et al. \(2019\)](#). Water temperature and salinity were sampled in situ with a CTD
684 (Conductivity, Temperature, and Depth) measuring device. Inorganic nutrients were
685 measured with an Alliance Evolution II autoanalyzer [Grasshoff et al. \(2009\)](#). See [Gasol et al.](#)
686 [\(2016\)](#) for specific details on how other variables were measured.

687

688 **Network construction**

689 For both the simulated and BBMO dataset, we constructed association networks from
690 microbial abundance tables and environmental parameters using eLSA [Xia et al. \(2011,](#)
691 [2012\)](#). We included default normalization and a z-score transformation using median and
692 median absolute deviation. We estimated the p -value with a mixed model that performs a
693 random permutation test of a co-occurrence if the theoretical p -values for the comparison are
694 below 0.05; the number of iterations was 2,000, and we considered no delay. For the BBMO
695 dataset, the Bonferroni false discovery rate, q , was calculated for all edges based on the p -
696 values using the R function `p.adjust` [R Core Team \(2019\)](#). For the network inference, we used

697 the significance threshold for the p and q value of 0.001 as suggested in other works [Weiss](#)
698 [et al. \(2016\)](#).

699

700 **Intersection combination of EnDED—Environmentally-Driven Edge Detection**

701 **methods**

702 For the intersection combination approach implemented in EnDED, to determine if a
703 microbial association is environmentally-driven or not, all four individual methods (i.e. Sp,
704 OL, II, and DPI, described below) must converge to the same solution. We applied the
705 methods to find environmentally-driven associations of microorganisms that were within an
706 environmental triplet, as already used in [Lima-Mendez et al. \(2015\)](#). An environmental triplet
707 is a special case of a closed triplet where one of the nodes corresponds to an environmental
708 component and the other two nodes correspond to microorganisms. We define the closed
709 triplet, where there is an edge between each pair of three nodes, as $T = \{v, w, f\}$ where v
710 and w are two microorganisms, and f is an environmental component (see Figure 3). For an
711 environmental triplet, if all methods classify the microbial edge as environmentally-driven,
712 the edge is removed from the network. If a microbial association is within several
713 environmental triplets, at least one of them must indicate the association as environmentally-
714 driven in order to remove the edge from the network. In sum, the intersection combination
715 retains an association in the network if no triplet classifies the association as environmentally-
716 driven.

717

718 Sign Pattern

719 The SP method [Lima-Mendez et al. \(2015\)](#) filters environmentally-driven edges from
720 a network in which a positive association score indicates co-occurrence, and a negative

721 association score indicates mutual exclusion. Let s_{vw} be the sign of the association score of
 722 the association between v and w (i.e. $s_{vw} = +$ or $s_{vw} = -$). A closed triplet T has eight SP
 723 combinations that group into two sets (see Figure 3). If the product of the three association
 724 scores is positive, then the SP suggests that the edge between the two microorganisms is
 725 environmentally-driven. Otherwise, if the product of the three association scores is negative,
 726 SP does not suggest that the association is environmentally-driven.

727

728 Overlap

729 We have developed the OL method to support the SP for temporal data: a microbial
 730 edge should be disregarded as environmentally-driven when the associations are misaligned
 731 in time. Thus, OL requires the time when the association begins as well as how long the
 732 associations lasts, i.e. duration or length of association in time, both determined by the
 733 network construction tool eLSA [Xia et al. \(2011, 2012\)](#). Given an association between v
 734 and w , let b_{vw}^v be the beginning of the association for v , b_{vw}^w the beginning of the
 735 association for w , and d_{vw} be the duration of the association between v and w . Although
 736 not used in the BBMO network, OL can consider time-delays by assuming that the
 737 beginning of the association is the minimum of the two beginnings, $b_{vw} = \min(b_{vw}^v, b_{vw}^w)$,
 738 and the end of the association is the maximum, $e_{vw} = \max(b_{vw}^v + d_{vw}, b_{vw}^w + d_{vw})$. We
 739 indicate two microorganisms with v and w , and the factor by f . The OL method calculates
 740 the overlap O of the microbial association with the two microorganism-environment
 741 associations through equation (5). As depicted in Figure 3, if $O > 60\%$, the microbial
 742 association is considered environmentally-driven.

$$O = 100 \frac{\min(e_{vw}, e_{vf}, e_{wf}) - \max(b_{vw}, b_{vf}, b_{wf})}{e_{vw} - b_{vw}} \quad \text{Eq. (5)}$$

743 Mutual Information and Conditional Mutual Information

744 The method II employs two measurements: MI and CMI. The former is also used by
 745 DPI. Thus, before describing the methods, we first describe the two measurements. MI is a
 746 measure of the degree of statistical dependency between two variables [Margolin et al. \(2006\)](#).
 747 We first consider $\mathbf{v} = v_1, \dots, v_n$, $\mathbf{w} = w_1, \dots, w_n$, and $\mathbf{f} = f_1, \dots, f_n$ as discrete random
 748 variables. The marginal probability of each discrete state (value) of the variable is denoted
 749 by $p(v_i) = P(\mathbf{v} = v_i)$, the joint probability by $p(v_i, w_j)$, and $p(v_i, w_j, f_k)$, and the
 750 conditional probability by $p(v_i|f_k)$, and $p(v_i, w_j|f_k)$. To obtain MI, we calculate the entropy
 751 of \mathbf{v} as

$$S(\mathbf{v}) = - \sum_{i=1}^n p(v_i) \log(p(v_i)), \quad \text{Eq. (6)}$$

752 and the joint entropy of \mathbf{v} and \mathbf{w} as

$$S(\mathbf{v}, \mathbf{w}) = - \sum_{i=1, j=1}^n p(v_i, w_j) \log(p(v_i, w_j)), \quad \text{Eq. (7)}$$

753 using the natural logarithm. The MI of \mathbf{v} and \mathbf{w} is defined through the sum of their entropies
 754 subtracted by their joint entropy:

$$\text{MI}(\mathbf{v}; \mathbf{w}) = S(\mathbf{v}) + S(\mathbf{w}) - S(\mathbf{v}, \mathbf{w}) \quad \text{Eq. (8)}$$

$$= \sum_{i=1}^n \sum_{j=1}^n p(v_i, w_j) \log\left(\frac{p(v_i, w_j)}{p(v_i)p(w_j)}\right), \quad \text{Eq. (9)}$$

755 with marginal probabilities $p(v_i) = \sum_{j=1}^n p(v_i, w_j)$, and $p(w_j) = \sum_{i=1}^n p(v_i, w_j)$.

756 The measurement CMI is the expected value of the MI of two random variables given

757 a third random variable. It is defined as

$$\text{CMI}(\mathbf{v}; \mathbf{w}|\mathbf{f}) = S(\mathbf{v}, \mathbf{f}) + S(\mathbf{w}, \mathbf{f}) - S(\mathbf{v}, \mathbf{w}, \mathbf{f}) - S(\mathbf{f}) \quad \text{Eq. (10)}$$

$$\begin{aligned} &= \sum_{k=1}^n p(f_k) \sum_{i=1}^n \sum_{j=1}^n p(v_i, w_j | f_k) \log \left(\frac{p(v_i, w_i | f_k)}{p(v_i | f_k) p(w_j | f_k)} \right) \quad \text{Eq. (11)} \\ &= \sum_{k=1}^n \sum_{i=1}^n \sum_{j=1}^n p(v_i, w_j, f_k) \log \left(\frac{p(f_k) p(v_i, w_i, f_k)}{p(v_i, f_k) p(w_j, f_k)} \right). \end{aligned}$$

758

759 Interaction Information

760 The II is calculated with microbial abundance and environmental data. In this study,
761 as in [Lima-Mendez et al. \(2015\)](#), II is computed as the difference of the CMI and MI:

$$\text{II} = \text{CMI} - \text{MI} . \quad \text{Eq. (12)}$$

762 In other works [Ghassami and Kiyavash \(2017\)](#), the II is defined with a different sign
763 convention: $\text{II} = \text{MI} - \text{CMI}$. In our study, if II is positive, the method suggests that the microbial
764 association is not environmentally-driven. If II is negative, there is an environmentally-
765 driven association within the closed triplet. However, this method cannot detect which of the
766 three associations is indirect. In other works [Lima-Mendez et al. \(2015\)](#), the microbial
767 association is assumed to be environmentally-driven if II is negative, but here we suggest to
768 combine it with DPI (see below).

769

770 Significance of Interaction Information

771 We determined the significance of II following a strategy from Ref. [North et al.](#)
772 [\(2002\)](#); [Veech \(2012\)](#). We used a null model without prior assumptions on the distribution

773 and computed the p -value by randomizing the environmental vector f . Since the MI is
 774 independent of the environmental factor and therefore remains constant, the significance of
 775 the II is the same as the CMI. Thus, we determined the significance of CMI with 1,000
 776 permutations: we randomized the environmental vector f and recalculated the CMI 1,000×,
 777 obtaining a CMI_i with $i \in \{1, \dots, 1,000\}$. Afterwards, we quantified with c how many
 778 random CMI_i were at least as small as the original CMI_i : $c = |\{i: CMI_i \leq CMI_{original}, i \in$
 779 $\{1, \dots, 1,000\}\}|$. We calculate the p -value as

$$p = \frac{c + 1}{1,000 + 1}. \quad \text{Eq. (13)}$$

780

781 Data Processing Inequality

782 As mentioned above, the II method can detect if an indirect association exists within
 783 a triplet but cannot determine which of the three associations is indirect. Thus, we added DPI
 784 to EnDED. DPI states that if two components v and w interact only through a third
 785 component f (i.e. in a network v and w are connected through a path containing f and there
 786 is no alternative path between v and w), then the MI of v and w , $MI(v; w)$ is smaller than
 787 $MI(v; f)$ and $MI(w; f)$ [Cover and Thomas \(2001\)](#):

$$MI(v; w) \leq \min \{MI(v; f), MI(w; f)\}. \quad \text{Eq. (14)}$$

788 While DPI has been used in previous works on gene triplets [Margolin et al. \(2006\)](#), we used
 789 the DPI method for environmental triplets. We compared the MI between the two
 790 microorganisms with the MI between a microorganism and the environmental factor. If the
 791 MI between the microorganisms is the smallest, then the method suggests that the edge is
 792 environmentally-driven. This method complements the II method.

793

794 Equal Width Discretization

795 To compute the MI, CMI, and subsequently II, we discretized the abundance data and
 796 environmental parameters. EnDED uses the equal width discretization algorithm, which
 797 creates equal sized ranges (also called bins or buckets) for an abundance vector $\mathbf{v} =$
 798 (v_1, \dots, v_n) between the lowest value (v_{min}) and highest value (v_{max}). It is a procedure
 799 implemented in other works [Meyer et al. \(2008\)](#). Given vector \mathbf{v} of length n (that is sample
 800 size) and number of bins $|B| = \lfloor \sqrt{n} \rfloor$, the discretized value v_d of \mathbf{v} is:

$$v_d = \left\lceil \frac{(v - v_{min}) \cdot |B|}{v_{max}} \right\rceil. \quad \text{Eq. (15)}$$

801 This equation assumes positive values. However, if \mathbf{v} contains negative values, $v_{min} < 0$,
 802 we adjust equation (15) by substituting v_{max} for $v'_{max} = v_{max} - v_{min}$. This method does not
 803 fill in missing values, and it is limited in the presence of outliers as most values would go
 804 within the same bin. We can solve this problem with a different discretization method (where
 805 bins have the same number of elements) but we have not implemented it in the current version
 806 of EnDED.

807

808 Applying EnDED to networks constructed from simulated and real data

809 We applied EnDED to association networks constructed from time series of simulated
 810 abundances and observed microbial abundances. The simulated networks were based on a
 811 gLV, while the real network was based on data from the BBMO. For the methods II and DPI
 812 we also included the corresponding abundance tables, and environmental factors. EnDED
 813 was run with the OL threshold of 60%. We set the significance threshold for the II score to
 814 0.05 and used 1,000 iterations.

815

816 **Evaluation of EnDED's performance**

817 Simulated network

818 We evaluated EnDED with the simulated interaction matrices, which revealed the
819 number of true positives (TP), true negatives (TN), false negatives (FN), and false positives
820 (FP) before and after removing associations that were classified as environmentally-driven.
821 We have assumed that all associations not present in the interaction matrices, are
822 environmentally-driven. We consider P as the number of all false associations, both true
823 positive and false negative detected environmentally-driven edges: $P = TP + FN$, and N as
824 the number of all true interactions, i.e. all true negative and false positive detected
825 environmentally-driven edges: $N = TN + FP$. Then, we can calculate the true positive rate
826 (sensitivity) by dividing the number of true positives by the number of all real positives:
827 $TPR = (TP)/(P)$. Equivalently, we can also calculate the true negative rate (specificity) by
828 dividing the number of true negatives by the number of all real negatives, $TNR = (TN)/(N)$.
829 The false positive rate (fall out) is the complementary to TNR, i.e. $FPR = 1 - TNR$. The
830 positive predictive value can be calculated by dividing the number of true positives by the
831 sum of all predicted positives, $PPV = (TP)/(TP + FP)$. Note PPV is also called precision.
832 In the Discussion section we used precision to refer to the networks ability in retrieving true
833 interactions. Here we use PPV in EnDEDs ability to remove false associations. The accuracy
834 is calculated by dividing the sum of true positives and true negatives by the sum of all real
835 positives and real negatives, $ACC = (TP + TN)/(P + N)$.

836

837 Real Dataset

838 *Literature based database*

839 Real network evaluation is limited since the true interactions and the microorganisms that do
840 not interact with each other are poorly known. We assessed true interactions known in the
841 literature based on the genus, which are compiled within the Protist Interaction Database,
842 PIDA [Bjorbækmo et al. \(2019\)](#). On October 15th 2019, PIDA contained 2,448 interactions.
843 Although our dataset contains protists as well as bacteria, we were unable to evaluate
844 interactions between bacteria.

845

846 *Jaccard index*

847 In ecology, the Jaccard index (Jaccard similarity coefficient) is normally used for
848 communities. Here, for each pair of microorganisms in the BBMO network, we computed
849 the Jaccard index as the number of samples in which both microorganisms occur, divided by
850 the number of samples in which at least one of the two microorganisms is present.

851

852 **Ethics approval and consent to participate**

853 Not applicable.

854

855 **Consent for publication**

856 Not applicable.

857

858 **Availability of data and material**

859 EnDED is available here: <https://github.com/InaMariaDeutschmann/EnDED>. It also contains
860 the file `FromDataSimulationToEvaluatingEnDED.RMD`, which contains R code to generate
861 simulated abundance tables, commands to run eLSA network construction and EnDED, as
862 well as the command to run a C++ program (included as well) and R code used for evaluation.
863 The folder BBMO data contains the BBMO abundance table, the taxonomic classification
864 table, and the BBMO network including results of EnDED.

865

866 **Competing interests**

867 The authors declare that they have no competing interests.

868

869 **Funding**

870 This project and IMD received funding from the European Union's Horizon 2020 research
871 and innovation program under the Marie Skłodowska-Curie grant agreement no. 675752
872 (Singek: <http://www.singek.eu>). RL was supported by a Ramón y Cajal fellowship (RYC-

873 2013-12554, MINECO, Spain). This work was also supported by the projects
874 INTERACTOMICS (CTM2015-69936-P, MINECO, Spain) and MicroEcoSystems
875 (240904, RCN, Norway) to RL.

876

877 **Author's contributions**

878 IMD, GLM, JR, KF and RL designed and conceived the project. IMD performed data
879 analysis, data simulation, and implementation of EnDED. IMD received substantial feedback
880 on established indirect detection methods from GLM and KF, on data simulation from SMV
881 and KF, on network construction from AKK, and on evaluation of EnDED from AKK
882 (literature based database for real dataset), GLM and KF (measurements for simulation
883 dataset). RL processed the amplicon data from BBMO generating OTU tables. AKK ran the
884 eLSA network construction tool for the BBMO dataset and IMD ran the tool for the
885 simulation datasets. RL provided funding for the project. The original draft was written by
886 IMD. IMD, GLM, AKK, SMV, KF and RL contributed substantially to manuscript revisions.
887 All authors approved the final version of the manuscript.

888

889 **Acknowledgements**

890 We thank all members of the Blanes Bay Microbial Observatory sampling team and the
891 multiple projects funding this collaborative effort over the years. Thanks to collaborators at
892 www.thepapermill.eu for help with critical reading in the early stages of the manuscript. Part
893 of the analyses have been performed at the Marbits bioinformatics core at ICM-CSIC
894 (<https://marbits.icm.csic.es>).

895

896 **References**

897 Ai, D., Li, X., Pan, H., Chen, J., Cram, J.A., Xia, L.C.: Explore mediated co-varying
898 dynamics in microbial community using integrated local similarity and liquid association
899 analysis. *BMC Genomics* 20(2), 185 (2019). doi:[10.1186/s12864-019-5469-8](https://doi.org/10.1186/s12864-019-5469-8)

900

901 Aitchison, J.: A new approach to null correlations of proportions. *Journal of the International
902 Association for Mathematical Geology* 13(2), 175–189 (1981). doi:[10.1007/BF01031393](https://doi.org/10.1007/BF01031393)

903

904 Alipanahi, B., Frey, B.J.: Network cleanup. *Nature Biotechnology* 31(8), 714–715 (2013).
905 doi:[10.1038/nbt.2657](https://doi.org/10.1038/nbt.2657)

906

907 Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search
908 tool. *Journal of Molecular Biology* 215(3), 403–410 (1990). doi:[10.1016/S0022-
909 2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)

910

911 Apprill, A., McNally, S., Parsons, R., Weber, L.: Minor revision to v4 region ssu rRNA 806r
912 gene primer greatly increases detection of sar11 bacterioplankton. *Aquatic Microbial
913 Ecology* 75(2), 129–137 (2015). doi:[10.3354/ame01753](https://doi.org/10.3354/ame01753)

914

915 Barzel, B., Barabási, A.-L.: Network link prediction by global silencing of indirect
916 correlations. *Nature Biotechnology* 31(8), 720–725 (2013). doi:[10.1038/nbt.2601](https://doi.org/10.1038/nbt.2601)

917

- 918 Bashan, A., Gibson, T.E., Friedman, J., Carey, V.J., Weiss, S.T., Hohmann, E.L., Liu, Y.-Y.:
919 Universality of human microbial dynamics. *Nature* 534(7606), 259–262 (2016).
920 doi:[10.1038/nature18301](https://doi.org/10.1038/nature18301)
- 921
- 922 Benincà, E., Dakos, V., Van Nes, E.H., Huisman, J., Scheffer, M.: Resonance of plankton
923 communities with temperature fluctuations. *The American Naturalist* 178(4), 85–95 (2011).
924 doi:[10.1086/661902](https://doi.org/10.1086/661902). PMID: 21956036
- 925
- 926 Berry, D., Widder, S.: Deciphering microbial interactions and detecting keystone species
927 with co-occurrence networks. *Frontiers in Microbiology* 5, 219 (2014).
928 doi:[10.3389/fmicb.2014.00219](https://doi.org/10.3389/fmicb.2014.00219)
- 929
- 930 Bjorbækmo, M.F.M., Evenstad, A., Røsæg, L.L., Krabberød, A.K., Logares, R.: The
931 planktonic protist interactome: where do we stand after a century of research? *The ISME*
932 *Journal* (2019). doi:[10.1038/s41396-019-0542-5](https://doi.org/10.1038/s41396-019-0542-5)
- 933
- 934 Brisson, V., Schmidt, J., Northen, T.R., Vogel, J.P., Gaudin, A.: A new method to correct for
935 habitat filtering in microbial correlation networks. *Frontiers in Microbiology* 10, 585 (2019).
936 doi:[10.3389/fmicb.2019.00585](https://doi.org/10.3389/fmicb.2019.00585)
- 937
- 938 Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P.:
939 Dada2: High-resolution sample inference from illumina amplicon data. *Nature Methods*
940 13(7), 581–583 (2016). doi:[10.1038/nmeth.3869](https://doi.org/10.1038/nmeth.3869)
- 941
- 942 Cover, T.M., Thomas, J.A.: Inequalities in information theory. In: Schilling, D.L., Cover,
943 T.M., Thomas, J.A. (eds.) *Elements of Information Theory*, pp. 482–509. John Wiley & Sons,
944 Ltd, (2001). Chap. 16. doi:[10.1002/0471200611.ch16](https://doi.org/10.1002/0471200611.ch16).
945 <https://onlinelibrary.wiley.com/doi/abs/10.1002/0471200611.ch16>
- 946
- 947 Dam, P., Fonseca, L.L., Konstantinidis, K.T., Voit, E.O.: Dynamic models of the complex
948 microbial metapopulation of lake mendota. *npj Systems Biology and Applications* 2(1),
949 16007 (2016). doi:[10.1038/npjbsa.2016.7](https://doi.org/10.1038/npjbsa.2016.7)
- 950
- 951 DeLong, E.F.: The microbial ocean from genomes to biomes. *Nature* 459(7244), 200–206
952 (2009). doi:[10.1038/nature08059](https://doi.org/10.1038/nature08059)
- 953
- 954 Deutschmann, I.M., Lima-Mendez, G., Krabberød, A.K., Raes, J., Faust, K., Logares, R.:
955 Environmentally-Driven Edge Detection Program. (2019). doi: [10.5281/zenodo.3271729](https://doi.org/10.5281/zenodo.3271729).
956 EnDED v1.0.1, 15. June 2019. <https://github.com/InaMariaDeutschmann/EnDED>
- 957
- 958 Falkowski, P.G., Fenchel, T., Delong, E.F.: The microbial engines that drive earth's
959 biogeochemical cycles. *Science* 320(5879), 1034–1039 (2008).
960 doi:[10.1126/science.1153213](https://doi.org/10.1126/science.1153213)
- 961
- 962 Faust, K.: Towards a better understanding of microbial community dynamics through high-
963 throughput cultivation and data integration. *mSystems* 4(3) (2019).
964 doi:[10.1128/mSystems.00101-19](https://doi.org/10.1128/mSystems.00101-19)

- 965
966 Faust, K., Raes, J.: Microbial interactions: from networks to models. *Nature Reviews*
967 *Microbiology* 10(8), 538–550 (2012). doi:[10.1038/nrmicro2832](https://doi.org/10.1038/nrmicro2832)
968
- 969 Faust, K., Raes, J.: Conet app: inference of biological association networks using cytoscape
970 [version 2; peer review: 2 approved]. *F1000Research* 5(1519) (2016).
971 doi:[10.12688/f1000research.9050.2](https://doi.org/10.12688/f1000research.9050.2)
972
- 973 Faust, K., Sathirapongsasuti, J.F., Izard, J., Segata, N., Gevers, D., Raes, J., Huttenhower, C.:
974 Microbial co-occurrence relationships in the human microbiome. *PLoS Computational*
975 *Biology* 8(7), 1–17 (2012). doi:[10.1371/journal.pcbi.1002606](https://doi.org/10.1371/journal.pcbi.1002606)
976
- 977 Feizi, S., Marbach, D., Médard, M., Kellis, M.: Network deconvolution as a general method
978 to distinguish direct dependencies in networks. *Nature Biotechnology* 31(8), 726–733
979 (2013). doi:[10.1038/nbt.2635](https://doi.org/10.1038/nbt.2635)
980
- 981 Fernandes, A.D., Gloor, G.B.: Mutual information is critically dependent on prior
982 assumptions: would the correct estimate of mutual information please identify itself?
983 *Bioinformatics* 26(9), 1135–1139 (2010). doi:[10.1093/bioinformatics/btq111](https://doi.org/10.1093/bioinformatics/btq111)
984
- 985 Friedman, J., Alm, E.J.: Inferring correlation networks from genomic survey data. *PLOS*
986 *Computational Biology* 8(9), 1–11 (2012). doi:[10.1371/journal.pcbi.1002687](https://doi.org/10.1371/journal.pcbi.1002687)
987
- 988 Gasol, J.M., Cardelús, C., G Morán, X.A., Balagué, V., Forn, I., Marrasé, C., Massana, R.,
989 Pedrós-Alió, C., Montserrat Sala, M., Simó, R., Vaqué, D., Estrada, M.: Seasonal patterns in
990 phytoplankton photosynthetic parameters and primary production at a coastal nw
991 mediterranean site. *Scientia Marina* 80(S1), 63–77 (2016). doi:[10.3989/scimar.04480.06E](https://doi.org/10.3989/scimar.04480.06E)
992
- 993 Ghassami, A., Kiyavash, N.: Interaction information for causal inference: The case of
994 directed triangle. In: 2017 IEEE International Symposium on Information Theory (ISIT), pp.
995 1326–1330 (2017)
996
- 997 Giner, C.R., Balagué, V., Krabberød, A.K., Ferrera, I., Reñé, A., Garcés, E., Gasol, J.M.,
998 Logares, R., Massana, R.: Quantifying long-term recurrence in planktonic microbial
999 eukaryotes. *Molecular Ecology* 28(5), 923–935 (2019). doi:[10.1111/mec.14929](https://doi.org/10.1111/mec.14929)
1000
- 1001 Gonze, D., Coyte, K.Z., Lahti, L., Faust, K.: Microbial communities as dynamical systems.
1002 *Current Opinion in Microbiology* 44, 41–49 (2018). doi:[10.1016/j.mib.2018.07.004](https://doi.org/10.1016/j.mib.2018.07.004).
1003 Microbiota
1004
- 1005 Grasshoff, K., Kremling, K., Ehrhardt, M.: *Methods of Seawater Analysis*. John Wiley &
1006 Sons, (2009)
1007
- 1008 Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G.,
1009 de Vargas, C., Decelle, J., del Campo, J., Dolan, J.R., Dunthorn, M., Edvardsen, B.,
1010 Holzmann, M., Kooistra, W.H.C.F., Lara, E., Le Bescot, N., Logares, R., Mahé, F., Massana,
1011 R., Montresor, M., Morard, R., Not, F., Pawlowski, J., Probert, I., Sauvadet, A.-L., Siano, R.,

- 1012 Stoeck, T., Vaultot, D., Zimmermann, P., Christen, R.: The protist ribosomal reference
1013 database (pr2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated
1014 taxonomy. *Nucleic Acids Research* 41(D1), 597–604 (2012). doi:[10.1093/nar/gks1160](https://doi.org/10.1093/nar/gks1160)
1015
- 1016 Haydon, D.: Pivotal assumptions determining the relationship between stability and
1017 complexity: An analytical synthesis of the stability-complexity debate. *The American*
1018 *Naturalist* 144(1), 14–29 (1994). doi:[10.1086/285658](https://doi.org/10.1086/285658)
1019
- 1020 Herlemann, D.P., Labrenz, M., Jürgens, K., Bertilsson, S., Waniek, J.J., Andersson, A.F.:
1021 Transitions in bacterial communities along the 2000km salinity gradient of the Baltic Sea. *The*
1022 *ISME Journal* 5(10), 1571–1579 (2011). doi:[10.1038/ismej.2011.41](https://doi.org/10.1038/ismej.2011.41)
1023
- 1024 Jackson, D.A.: Stopping rules in principal components analysis: A comparison of heuristical
1025 and statistical approaches. *Ecology* 74(8), 2204–2214 (1993). doi:[10.2307/1939574](https://doi.org/10.2307/1939574)
1026
- 1027 Jang, I.S., Margolin, A., Califano, A.: haracne: improving the accuracy of regulatory model
1028 reverse engineering via higher-order data processing inequality tests. *Interface Focus* 3(4),
1029 20130011 (2013). doi:[10.1098/rsfs.2013.0011](https://doi.org/10.1098/rsfs.2013.0011)
1030
- 1031 Kallmeyer, J., Pockalny, R., Adhikari, R.R., Smith, D.C., D’Hondt, S.: Global distribution
1032 of microbial abundance and biomass in subseafloor sediment. *Proceedings of the National*
1033 *Academy of Sciences* 109(40), 16213–16216 (2012). doi:[10.1073/pnas.1203849109](https://doi.org/10.1073/pnas.1203849109)
1034
- 1035 Kettle, H., Holtrop, G., Louis, P., Flint, H.J.: micropop: Modelling microbial populations and
1036 communities in R. *Methods in Ecology and Evolution* 9(2), 399–409 (2018).
1037 doi:[10.1111/2041-210X.12873](https://doi.org/10.1111/2041-210X.12873)
1038
- 1039 Klemm, K., Eguíluz, V.M.: Growing scale-free networks with small-world behavior.
1040 *Physical Review E* 65(5), 057102 (2002). doi:[10.1103/PhysRevE.65.057102](https://doi.org/10.1103/PhysRevE.65.057102)
1041
- 1042 Krabberød, A.K., Bjorbækmo, M.F.M., Shalchian-Tabrizi, K., Logares, R.: Exploring the
1043 oceanic microeukaryotic interactome with metaomics approaches. *Aquatic Microbial*
1044 *Ecology* 79(1), 1–12 (2017). doi:[10.3354/ame01811](https://doi.org/10.3354/ame01811)
1045
- 1046 Kurtz, Z.D., Bonneau, R., Müller, C.L.: Disentangling microbial associations from hidden
1047 environmental and technical factors via latent graphical models. *bioRxiv* (2019).
1048 doi:[10.1101/2019.12.21.885889](https://doi.org/10.1101/2019.12.21.885889)
1049
- 1050 Kurtz, Z.D., Müller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J., Bonneau, R.A.: Sparse
1051 and compositionally robust inference of microbial ecological networks. *PLOS*
1052 *Computational Biology* 11(5), 1–25 (2015). doi:[10.1371/journal.pcbi.1004226](https://doi.org/10.1371/journal.pcbi.1004226)
1053
- 1054 Layeghifard, M., Hwang, D.M., Guttman, D.S.: Disentangling interactions in the
1055 microbiome: A network perspective. *Trends in Microbiology* 25(3), 217–228 (2017).
1056 doi:[10.1016/j.tim.2016.11.008](https://doi.org/10.1016/j.tim.2016.11.008)
1057
- 1058 Legendre, P., Legendre, L.F.: *Numerical Ecology* vol. 24. Elsevier, (2012)

- 1059
1060 Li, C., Lim, K.M.K., Chng, K.R., Nagarajan, N.: Predicting microbial interactions through
1061 computational approaches. *Methods* 102, 12–19 (2016). doi:[10.1016/j.ymeth.2016.02.019](https://doi.org/10.1016/j.ymeth.2016.02.019).
1062 Pan-omics analysis of biological data
- 1063
1064 Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., Chaffron, S.,
1065 Ignacio-Espinosa, J.C., Roux, S., Vincent, F., Bittner, L., Darzi, Y., Wang, J., Audic, S.,
1066 Berline, L., Bontempi, G., Cabello, A.M., Coppola, L., Cornejo-Castillo, F.M., d'Ovidio, F.,
1067 De Meester, L., Ferrera, I., Garet-Delmas, M.-J., Guidi, L., Lara, E., Pesant, S., Royo-Llonch,
1068 M., Salazar, G., Sánchez, P., Sebastian, M., Souffreau, C., Dimier, C., Picheral, M., Searson,
1069 S., Kandels-Lewis, S., Tara Oceans coordinators, Gorsky, G., Not, F., Ogata, H., Speich, S.,
1070 Stemmann, L., Weissenbach, J., Wincker, P., Acinas, S.G., Sunagawa, S., Bork, P., Sullivan,
1071 M.B., Karsenti, E., Bowler, C., de Vargas, C., Raes, J.: Determinants of community structure
1072 in the global plankton interactome. *Science* 348(6237), 1262073 (2015).
1073 doi:[10.1126/science.1262073](https://doi.org/10.1126/science.1262073)
- 1074
1075 Locey, K.J., Lennon, J.T.: Scaling laws predict global microbial diversity. *Proceedings of the*
1076 *National Academy of Sciences* 113(21), 5970–5975 (2016). doi:[10.1073/pnas.1521291113](https://doi.org/10.1073/pnas.1521291113)
- 1077
1078 Lv, X., Zhao, K., Xue, R., Liu, Y., Xu, J., Ma, B.: Strengthening insights in microbial
1079 ecological networks from theory to applications. *mSystems* 4(3), 00124–19 (2019).
1080 doi:[10.1128/mSystems.00124-19](https://doi.org/10.1128/mSystems.00124-19)
- 1081
1082 Mandakovic, D., Rojas, C., Maldonado, J., Latorre, M., Travisany, D., Delage, E., Bihouée,
1083 A., Jean, G., Díaz, F.P., Fernández-Gómez, B., Cabrera, P., Gaete, A., Latorre, C., Gutiérrez,
1084 R.A., Maass, A., Cambiazo, V., Navarrete, S.A., Eveillard, D., González, M.: Structure and
1085 co-occurrence patterns in microbial communities under acute environmental stress reveal
1086 ecological factors fostering resilience. *Scientific Reports* 8(1), 5875 (2018).
1087 doi:[10.1038/s41598-018-23931-0](https://doi.org/10.1038/s41598-018-23931-0)
- 1088
1089 Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R.D.,
1090 Califano, A.: Aracne: An algorithm for the reconstruction of gene regulatory networks in a
1091 mammalian cellular context. *BMC Bioinformatics* 7(1), 7 (2006). doi:[10.1186/1471-2105-7-](https://doi.org/10.1186/1471-2105-7-S1-S7)
1092 [S1-S7](https://doi.org/10.1186/1471-2105-7-S1-S7)
- 1093
1094 McCarren, J., Becker, J.W., Repeta, D.J., Shi, Y., Young, C.R., Malmstrom, R.R., Chisholm,
1095 S.W., DeLong, E.F.: Microbial community transcriptomes reveal microbes and metabolic
1096 pathways associated with dissolved organic matter turnover in the sea. *Proceedings of the*
1097 *National Academy of Sciences* 107(38), 16420–16427 (2010).
1098 doi:[10.1073/pnas.1010732107](https://doi.org/10.1073/pnas.1010732107)
- 1099
1100 Meyer, P.E., Lafitte, F., Bontempi, G.: minet: A r/bioconductor package for inferring large
1101 transcriptional networks using mutual information. *BMC Bioinformatics* 9(1), 461 (2008).
1102 doi:[10.1186/1471-2105-9-461](https://doi.org/10.1186/1471-2105-9-461)
- 1103
1104 Moritz, S., Gatscha, S.: imputeTS: Time Series Missing Value Imputation. (2017). R package
1105 version 2.8. <https://github.com/SteffenMoritz/imputeTS>

- 1106
1107 North, B.V., Curtis, D., Sham, P.C.: A note on the calculation of empirical p values from
1108 monte carlo procedures. *The American Journal of Human Genetics* 71(2), 439–441 (2002).
1109 doi:[10.1086/341527](https://doi.org/10.1086/341527)
1110
- 1111 Novak, M., Yeakel, J.D., Noble, A.E., Doak, D.F., Emmerson, M., Estes, J.A., Jacob, U.,
1112 Tinker, M.T., Wootton, J.T.: Characterizing species interactions to understand press
1113 perturbations: What is the community matrix? *Annual Review of Ecology, Evolution, and*
1114 *Systematics* 47(1), 409–432 (2016). doi:[10.1146/annurev-ecolsys-032416-010215](https://doi.org/10.1146/annurev-ecolsys-032416-010215)
1115
- 1116 Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin,
1117 P.R., O’Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., Wagner, H.:
1118 *Vegan: Community Ecology Package*. (2019). R package version 2.5.-6 (network
1119 simulation), and 2.4-2 (BBMO data).
1120 <https://CRAN.R-project.org/package=vegan>
1121
- 1122 Pascual-García, A., Tamames, J., Bastolla, U.: Bacteria dialog with santa rosalia: Are
1123 aggregations of cosmopolitan bacteria mainly explained by habitat filtering or by ecological
1124 interactions? *BMC Microbiology* 14, 284 (2014). doi:[10.1186/s12866-014-0284-5](https://doi.org/10.1186/s12866-014-0284-5)
1125
- 1126 Pinedo-González, P., West, A.J., Tovar-Sánchez, A., Duarte, C.M., Marañón, E., Cermeño,
1127 P., González, N., Sobrino, C., Huete-Ortega, M., Fernández, A., López-Sandoval, D.C.,
1128 Vidal, M., Blasco, D., Estrada, M., Sañudo-Wilhelmy, S.A.: Surface distribution of dissolved
1129 trace metals in the oligotrophic ocean and their influence on phytoplankton biomass and
1130 productivity. *Global Biogeochemical Cycles* 29(10), 1763–1781 (2015).
1131 doi:[10.1002/2015GB005149](https://doi.org/10.1002/2015GB005149)
1132
- 1133 Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner,
1134 F.O.: The silva ribosomal rna gene database project: improved data processing and web-
1135 based tools. *Nucleic Acids Research* 41(D1), 590–596 (2012). doi:[10.1093/nar/gks1219](https://doi.org/10.1093/nar/gks1219)
1136
- 1137 R Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for
1138 Statistical Computing, Vienna, Austria (2019). R Foundation for Statistical Computing.
1139 <https://www.R-project.org/>
1140
- 1141 Röttjers, L., Faust, K.: From hairballs to hypotheses—biological insights from microbial
1142 networks. *FEMS Microbiology Reviews* 42(6), 761–780 (2018). doi:[10.1093/femsre/fuy030](https://doi.org/10.1093/femsre/fuy030)
1143
- 1144 Soetaert, K., Petzoldt, T., Setzer, R.W.: Solving differential equations in R: Package deSolve.
1145 *Journal of Statistical Software* 33(9), 1–25 (2010). doi:[10.18637/jss.v033.i09](https://doi.org/10.18637/jss.v033.i09)
1146
- 1147 Stein, R.R., Bucci, V., Toussaint, N.C., Buffie, C.G., Räscher, G., Pamer, E.G., Sander, C.,
1148 Xavier, J.B.: Ecological modeling from time-series inference: Insight into dynamics and
1149 stability of intestinal microbiota. *PLOS Computational Biology* 9(12), 1–11 (2013).
1150 doi:[10.1371/journal.pcbi.1003388](https://doi.org/10.1371/journal.pcbi.1003388)
1151
- 1152 Tackmann, J., Rodrigues, J.a.F.M., von Mering, C.: Rapid inference of direct interactions in

1153 large-scale ecological networks from heterogeneous microbial sequencing data. *Cell Systems*
1154 9(3), 286–2968 (2019). doi:[10.1016/j.cels.2019.08.002](https://doi.org/10.1016/j.cels.2019.08.002)

1155

1156 The Human Microbiome Project Consortium: Huttenhower, C., Gevers, D., Knight, R.,
1157 Abubucker, S., Badger, J.H., Chinwalla, A.T., Creasy, H.H., Earl, A.M., FitzGerald, M.G.,
1158 Fulton, R.S., Giglio, M.G., Hallsworth-Pepin, K., Lobos, E.A., Madupu, R., Magrini, V.,
1159 Martin, J.C., Mitreva, M., Muzny, D.M., Sodergren, E.J., Versalovic, J., Wollam, A.M.,
1160 Worley, K.C., Wortman, J.R., Young, S.K., Zeng, Q., Aagaard, K.M., Abolude, O.O., Allen-
1161 Vercoe, E., Alm, E.J., Alvarado, L., Andersen, G.L., Anderson, S., Appelbaum, E., Arachchi,
1162 H.M., Armitage, G., Arze, C.A., Ayvaz, T., Baker, C.C., Begg, L., Belachew, T., Bhonagiri,
1163 V., Bihan, M., Blaser, M.J., Bloom, T., Bonazzi, V., Paul Brooks, J., Buck, G.A., Buhay,
1164 C.J., Busam, D.A., Campbell, J.L., Canon, S.R., Cantarel, B.L., Chain, P.S.G., Chen, I.-M.A.,
1165 Chen, L., Chhibba, S., Chu, K., Ciulla, D.M., Clemente, J.C., Clifton, S.W., Conlan, S.,
1166 Crabtree, J., Cutting, M.A., Davidovics, N.J., Davis, C.C., DeSantis, T.Z., Deal, C.,
1167 Delehaunty, K.D., Dewhirst, F.E., Deych, E., Ding, Y., Dooling, D.J., Dugan, S.P., Michael
1168 Dunne, W., Scott Durkin, A., Edgar, R.C., Erlich, R.L., Farmer, C.N., Farrell, R.M., Faust,
1169 K., Feldgarden, M., Felix, V.M., Fisher, S., Fodor, A.A., Forney, L.J., Foster, L., Di
1170 Francesco, V., Friedman, J., Friedrich, D.C., Fronick, C.C., Fulton, L.L., Gao, H., Garcia,
1171 N., Giannoukos, G., Giblin, C., Giovanni, M.Y., Goldberg, J.M., Goll, J., Gonzalez, A.,
1172 Griggs, A., Gujja, S., Kinder Haake, S., Haas, B.J., Hamilton, H.A., Harris, E.L., Hepburn,
1173 T.A., Herter, B., Hoffmann, D.E., Holder, M.E., Howarth, C., Huang, K.H., Huse, S.M.,
1174 Izard, J., Jansson, J.K., Jiang, H., Jordan, C., Joshi, V., Katancik, J.A., Keitel, W.A., Kelley,
1175 S.T., Kells, C., King, N.B., Knights, D., Kong, H.H., Koren, O., Koren, S., Kota, K.C., Kovar,
1176 C.L., Kyrpides, N.C., La Rosa, P.S., Lee, S.L., Lemon, K.P., Lennon, N., Lewis, C.M.,
1177 Lewis, L., Ley, R.E., Li, K., Liolios, K., Liu, B., Liu, Y., Lo, C.-C., Lozupone, C.A., Dwayne
1178 Lunsford, R., Madden, T., Mahurkar, A.A., Mannon, P.J., Mardis, E.R., Markowitz, V.M.,
1179 Mavromatis, K., McCorrison, J.M., McDonald, D., McEwen, J., McGuire, A.L., McInnes,
1180 P., Mehta, T., Mihindukulasuriya, K.A., Miller, J.R., Minx, P.J., Newsham, I., Nusbaum, C.,
1181 O’Laughlin, M., Orvis, J., Pagani, I., Palaniappan, K., Patel, S.M., Pearson, M., Peterson, J.,
1182 Podar, M., Pohl, C., Pollard, K.S., Pop, M., Priest, M.E., Proctor, L.M., Qin, X., Raes, J.,
1183 Ravel, J., Reid, J.G., Rho, M., Rhodes, R., Riehle, K.P., Rivera, M.C., Rodriguez-Mueller,
1184 B., Rogers, Y.-H., Ross, M.C., Russ, C., Sanka, R.K., Sankar, P., Fah Sathirapongsasuti, J.,
1185 Schloss, J.A., Schloss, P.D., Schmidt, T.M., Scholz, M., Schriml, L., Schubert, A.M., Segata,
1186 N., Segre, J.A., Shannon, W.D., Sharp, R.R., Sharpton, T.J., Shenoy, N., Sheth, N.U.,
1187 Simone, G.A., Singh, I., Smillie, C.S., Sobel, J.D., Sommer, D.D., Spicer, P., Sutton, G.G.,
1188 Sykes, S.M., Tabbaa, D.G., Thiagarajan, M., Tomlinson, C.M., Torralba, M., Treangen, T.J.,
1189 Truty, R.M., Vishnivetskaya, T.A., Walker, J., Wang, L., Wang, Z., Ward, D.V., Warren,
1190 W., Watson, M.A., Wellington, C., Wetterstrand, K.A., White, J.R., Wilczek-Boney, K., Wu,
1191 Y., Wylie, K.M., Wylie, T., Yandava, C., Ye, L., Ye, Y., Yooseph, S., Youmans, B.P., Zhang,
1192 L., Zhou, Y., Zhu, Y., Zoloth, L., Zucker, J.D., Birren, B.W., Gibbs, R.A., Highlander, S.K.,
1193 Methé, B.A., Nelson, K.E., Petrosino, J.F., Weinstock, G.M., Wilson, R.K., White, O.:
1194 Structure, function and diversity of the healthy human microbiome. *Nature* 486(7402), 207–
1195 214 (2012). doi:[10.1038/nature11234](https://doi.org/10.1038/nature11234)

1196

1197 Vallina, S.M., Martinez-Garcia, R., Smith, S.L., Bonachela, J.A.: Models in microbial
1198 ecology. In: Schmidt, T.M. (ed.) *Encyclopedia of Microbiology (Fourth Edition)*, Fourth
1199 edition edn., pp. 211–246. Academic Press, Oxford (2019). doi:[10.1016/B978-0-12-809633-](https://doi.org/10.1016/B978-0-12-809633-)

- 1200 [8.20789-9. http://www.sciencedirect.com/science/article/pii/B9780128096338207899](http://www.sciencedirect.com/science/article/pii/B9780128096338207899)
1201
- 1202 Veech, J.A.: Significance testing in ecological null models. *Theoretical Ecology* 5(4), 611–
1203 616 (2012). doi:[10.1007/s12080-012-0159-z](https://doi.org/10.1007/s12080-012-0159-z)
1204
- 1205 Verny, L., Sella, N., Affeldt, S., Singh, P.P., Isambert, H.: Learning causal networks with
1206 latent variables from multivariate information in genomic data. *PLOS Computational*
1207 *Biology* 13(10), 1–25 (2017). doi:[10.1371/journal.pcbi.1005662](https://doi.org/10.1371/journal.pcbi.1005662)
1208
- 1209 Villaverde, A.F., Becker, K., Banga, J.R.: Premer: A tool to infer biological networks.
1210 *IEEE/ACM Trans. Comput.Biol. Bioinformatics* 15(4), 1193–1202 (2018).
1211 doi:[10.1109/TCBB.2017.2758786](https://doi.org/10.1109/TCBB.2017.2758786)
1212
- 1213 Villaverde, A.F., Ross, J., Morán, F., Banga, J.R.: Mider: Network inference with mutual
1214 information distance and entropy reduction. *PLOS ONE* 9(5), 1–15 (2014).
1215 doi:[10.1371/journal.pone.0096732](https://doi.org/10.1371/journal.pone.0096732)
1216
- 1217 Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R.: Naïve bayesian classifier for rapid
1218 assignment of rna sequences into the new bacterial taxonomy. *Applied and Environmental*
1219 *Microbiology* 73(16), 5261–5267 (2007). doi:[10.1128/AEM.00062-07](https://doi.org/10.1128/AEM.00062-07)
1220
- 1221 Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L.C., Xu,
1222 Z.Z., Ursell, L., Alm, E.J., Birmingham, A., Cram, J.A., Fuhrman, J.A., Raes, J., Sun, F.,
1223 Zhou, J., Knight, R.: Correlation detection strategies in microbial data sets vary widely in
1224 sensitivity and precision. *The ISME Journal* 10(7), 1669–1681 (2016).
1225 doi:[10.1038/ismej.2015.235](https://doi.org/10.1038/ismej.2015.235)
1226
- 1227 Whitman, W.B., Coleman, D.C., Wiebe, W.J.: Prokaryotes: The unseen majority.
1228 *Proceedings of the National Academy of Sciences* 95(12), 6578–6583 (1998).
1229 doi:[10.1073/pnas.95.12.6578](https://doi.org/10.1073/pnas.95.12.6578)
1230
- 1231 Worden, A.Z., Follows, M.J., Giovannoni, S.J., Wilken, S., Zimmerman, A.E., Keeling, P.J.:
1232 Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes.
1233 *Science* 347(6223) (2015). doi:[10.1126/science.1257594](https://doi.org/10.1126/science.1257594)
1234
- 1235 Xia, L.C., Steele, J.A., Cram, J.A., Cardon, Z.G., Simmons, S.L., Vallino, J.J., Fuhrman,
1236 J.A., Sun, F.: Extended local similarity analysis (elsa) of microbial community and other time
1237 series data with replicates. *BMC Systems Biology* 5(2), 15 (2011). doi:[10.1186/1752-0509-
1238 5-S2-S15](https://doi.org/10.1186/1752-0509-5-S2-S15)
1239
- 1240 Xia, L.C., Ai, D., Cram, J., Fuhrman, J.A., Sun, F.: Efficient statistical significance
1241 approximation for local similarity analysis of high-throughput time series data.
1242 *Bioinformatics* 29(2), 230–237 (2012). doi:[10.1093/bioinformatics/bts668](https://doi.org/10.1093/bioinformatics/bts668)
1243
- 1244 Xiao, Y., Angulo, M.T., Friedman, J., Waldor, M.K., Weiss, S.T., Liu, Y.-Y.: Mapping the
1245 ecological networks of microbial communities. *Nature Communications* 8(1), 2042 (2017).
1246 doi:[10.1038/s41467-017-02090-2](https://doi.org/10.1038/s41467-017-02090-2)

1247
1248 Yang, Y., Chen, N., Chen, T.: Inference of environmental factor-microbe and microbe-
1249 microbe associations from metagenomic data using a hierarchical bayesian statistical model.
1250 Cell Systems 4(1), 129–1375 (2017). doi:[10.1016/j.cels.2016.12.012](https://doi.org/10.1016/j.cels.2016.12.012)

1251
1252 Yoon, H.S., Price, D.C., Stepanauskas, R., Rajah, V.D., Sieracki, M.E., Wilson, W.H., Yang,
1253 E.C., Duffy, S., Bhattacharya, D.: Single-cell genomics reveals organismal interactions in
1254 uncultivated marine protists. Science 332(6030), 714–717 (2011)

1255
1256 Zhao, J., Zhou, Y., Zhang, X., Chen, L.: Part mutual information for quantifying direct
1257 associations in networks. Proceedings of the National Academy of Sciences 113(18), 5130–
1258 5135 (2016). doi:[10.1073/pnas.1522586113](https://doi.org/10.1073/pnas.1522586113)

1259
1260 Zoppoli, P., Morganella, S., Ceccarelli, M.: Timedelay-aracne: Reverse engineering of gene
1261 networks from time-course data by an information theoretic approach. BMC Bioinformatics
1262 11(1), 154 (2010). doi:[10.1186/1471-2105-11-154](https://doi.org/10.1186/1471-2105-11-154)

1263

1264 Figures

1265 **Figure 1: Evaluation of EnDED: intersection combination and individual methods on**
1266 **simulated networks** Using 1,000 simulated networks, and 1,000 simulated networks
1267 incorporating noise, we evaluate EnDED’s performance. Plot A) displays the evaluation
1268 measurements true positive rate (TRP), true negative rate (TNR), accuracy (ACC), and
1269 positive predictive value (PPV) for each individual method, i.e. Sign Pattern (SP), Overlap
1270 (OL), Interaction Information (II), and Data Processing Inequality (DPI), as well as the
1271 intersection combination (COMBI). SP and OL perform best according to TRP and ACC,
1272 while the intersection combination performs best according to TNR. All methods performed
1273 well according to PPV. The intersection combination, DPI and II performed better on noisy
1274 data according to TNR because less edges were removed along with less true interactions.
1275 Plot B) displays the ROC curve for each environmentally-driven edge detection method as
1276 well as their intersection combination.

1277

1278 **Figure 2: Quantification of environmentally-driven associations in the BBMO network**

1279 For A) to D), the upper left figure shows the number (or fraction) of microbial associations
1280 divided by domain: Bacteria-Bacteria associations (B), Bacteria-Eukaryote associations
1281 (BE), and Eukaryote-Eukaryote associations (E). The upper right figure shows the number
1282 (or fraction) of associations divided by size-fractions: association within the nano size
1283 fraction (n), within the pico size fraction (p), and between these two size fractions (np). The
1284 figure in the middle shows all triplets connected to an environmental parameter: Temperature
1285 (Temp), Day length (Day), *Micromonas* (Mic), photosynthetic Nanoflagellates (PNF),
1286 heterotrophic Nanoflagellates (HNF), Chlorophyll (Chl), *Synechococcus* (Syn), inorganic
1287 nutrients NO_3^- (NO3), SiO_2 (Si), and NO_2^- (NO2), as well as heterotrophic Prokaryotes
1288 (H.Pro). The figure at the bottom shows the number of edges divided in how many triplets
1289 they have been found ranging from no triplets (0) to nine triplets. Figure A) and B) display
1290 the number of microbial associations of the BBMO network before applying EnDED.
1291 Positive associations are indicated with black, negative associations with red. Figure C) and
1292 D) indicate in blue the fraction of environmentally-driven edges among the positive (C) and
1293 negative (D) microbial associations. E) The upper left network shows in black the positive
1294 and in red the negative associations. The upper right network shows in green the number of
1295 triplets a microbial edge is in ranging from one (light green) to nine (dark green), and no
1296 triplet (black). The lower network shows in blue the environmentally-driven associations that
1297 were detected by the intersection combination of the four methods Sign Pattern, Overlap,
1298 Interaction Information, and Data Processing Inequality.

1299

1300 **Figure 3: EnDED Methods Overview** EnDED is an implementation of four methods aiming
1301 to determine whether an edge between two microorganisms is indirect through the action of
1302 an environmental factor. The four methods are: Sign Pattern, Overlap, Interaction

1303 Information, and Data Processing Inequality (see Methods). Each method can be used
 1304 individually or in combination. Here, we show the intersection combination approach, i.e.
 1305 only if all methods classify an edge as indirect, it is removed from the network. Otherwise,
 1306 the edge is classified as not indirect and kept in the network.

1307

1308 Tables

1309 **Table 1 Jaccard index of edges** The BBMO network before applying EnDED contained 33,
 1310 832 edges of which 4,806 (14.2%) are environmentally-driven (indirect). Considering the
 1311 Jaccard index for these indirect edges, 1,433 (29.8% of indirect edges) score above 50%, and
 1312 3, 373 (70.2%) score below or equal to 50%. In contrast, 60.6% of edges not considered as
 1313 indirect have a Jaccard index above 50%, and 39.4% of all not indirect edges have a Jaccard
 1314 index equal or below 50%.

	All	Jaccard index>50	Jaccard index≤50
BBMO network	33, 832 (100%)	19, 015 (56.2%)	14, 817 (43.8%)
positive edges	27, 700 (81.9%)	18, 832 (68.0%)	8, 868 (32.0%)
negative edges	6, 132 (18.1%)	183 (3.0%)	5, 949 (97.0%)
indirect (intersection combination)	4, 806 (14.2%)	1,433 (29.8%)	3, 373 (70.2%)
not indirect (all)	29, 026 (85.8%)	17, 582 (60.6%)	11,444 (39.4%)
not indirect (min 1 triplet)	23, 404 (69.2%)	14, 822 (63.3%)	8, 582 (36.7%)
not indirect (no triplet)	5, 622 (16.7%)	2, 760 (49.1%)	2, 862 (50.9%)
Sign Pattern	28, 210 (83.4%)	16, 255 (57.6%)	11,955 (42.4%)
Overlap	28, 210 (83.4%)	16, 255 (57.6%)	11,955 (42.4%)
Interaction Information	12, 960 (38.3%)	7, 808 (60.2%)	5, 152 (39.8%)
Data Processing Inequality	10, 610 (31.4%)	3, 328 (31.4%)	7, 282 (68.6%)

1315

1316 **Table 2 Interactions found in the BBMO network** These that appear in the literature,
 1317 including whether or not the associations were removed or kept by EnDED. For example, the
 1318 interaction between OTUs classified as *Dia. Thalassiosira* and OTUs classified as F.
 1319 unknown *Flavobacteriia* has been found 18 times in the network: 7 were removed and 11

1320 were kept.

Microorganisms	EnDED	ID in PIDA
<i>Dia. Thalassiosira</i> - <i>Dino. Heterocapsa</i>	1 removed	1665
<i>Dia. Thalassiosira</i> - F. unknown <i>Flavobacteriia</i>	7 removed	2199
	11 kept	
<i>Dino. Heterocapsa</i> - <i>Dino. Prorocentrum</i>	1 kept	1501, 1511
<i>Dino. Gymnodinium</i> - <i>C. Strombidium</i>	1 kept	1209
<i>Dino. Gyrodinium</i> - <i>Dino. Heterocapsa</i>	1 kept	1313, 1314, 1780, 1783
<i>Dino. Prorocentrum</i> - <i>Dino. Gymnodinium</i>	2 kept	1499
<i>Dino. Prorocentrum</i> - <i>Dino. Prorocentrum</i>	4 kept	1509, 1510
<i>Dino. Prorocentrum</i> - <i>Dino. Scrippsiella</i>	2 kept	1513
F. unknown <i>Flavobacteriia</i> - <i>Dia. Pseudo-nitzschia</i>	1 kept	2196

1321 Abbreviations indicate Dia - Diatomea; Dino - Dinoflagellata; C - Ciliophora; F - *Flavobacteriia*; ID in PIDA
 1322 refers to the number PIDA gave an interaction described in literature.

Figures

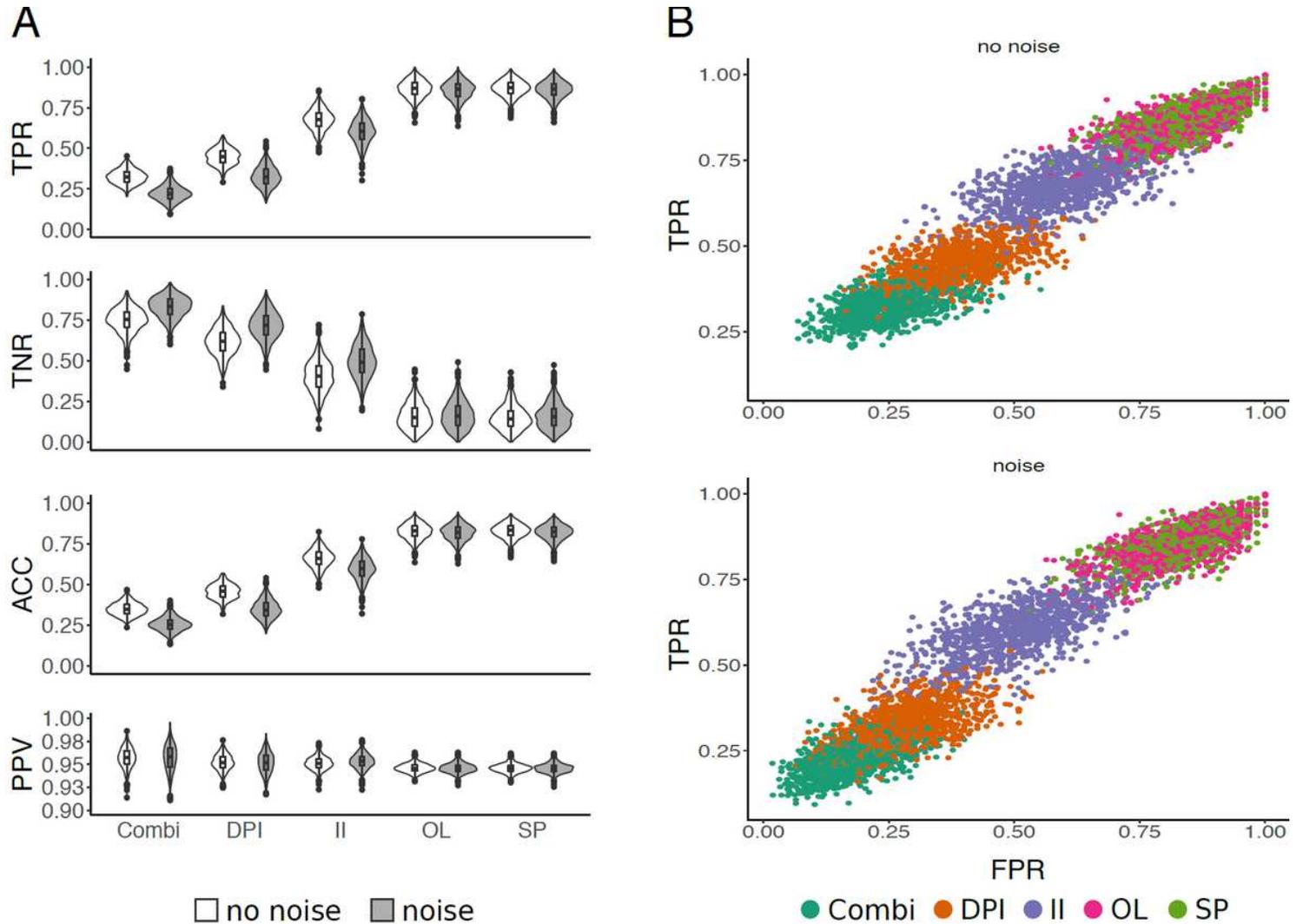


Figure 1

Evaluation of EnDED: intersection combination and individual methods on simulated networks Using 1,000 simulated networks, and 1,000 simulated networks incorporating noise, we evaluate EnDED's performance. Plot A) displays the evaluation measurements true positive rate (TPR), true negative rate (TNR), accuracy (ACC), and positive predictive value (PPV) for each individual method, i.e. Sign Pattern (SP), Overlap (OL), Interaction Information (II), and Data Processing Inequality (DPI), as well as the intersection combination (COMBI). SP and OL perform best according to TPR and ACC, while the intersection combination performs best according to TNR. All methods performed well according to PPV. The intersection combination, DPI and II performed better on noisy data according to TNR because less edges were removed along with less true interactions. Plot B) displays the ROC curve for each environmentally-driven edge detection method as well as their intersection combination.

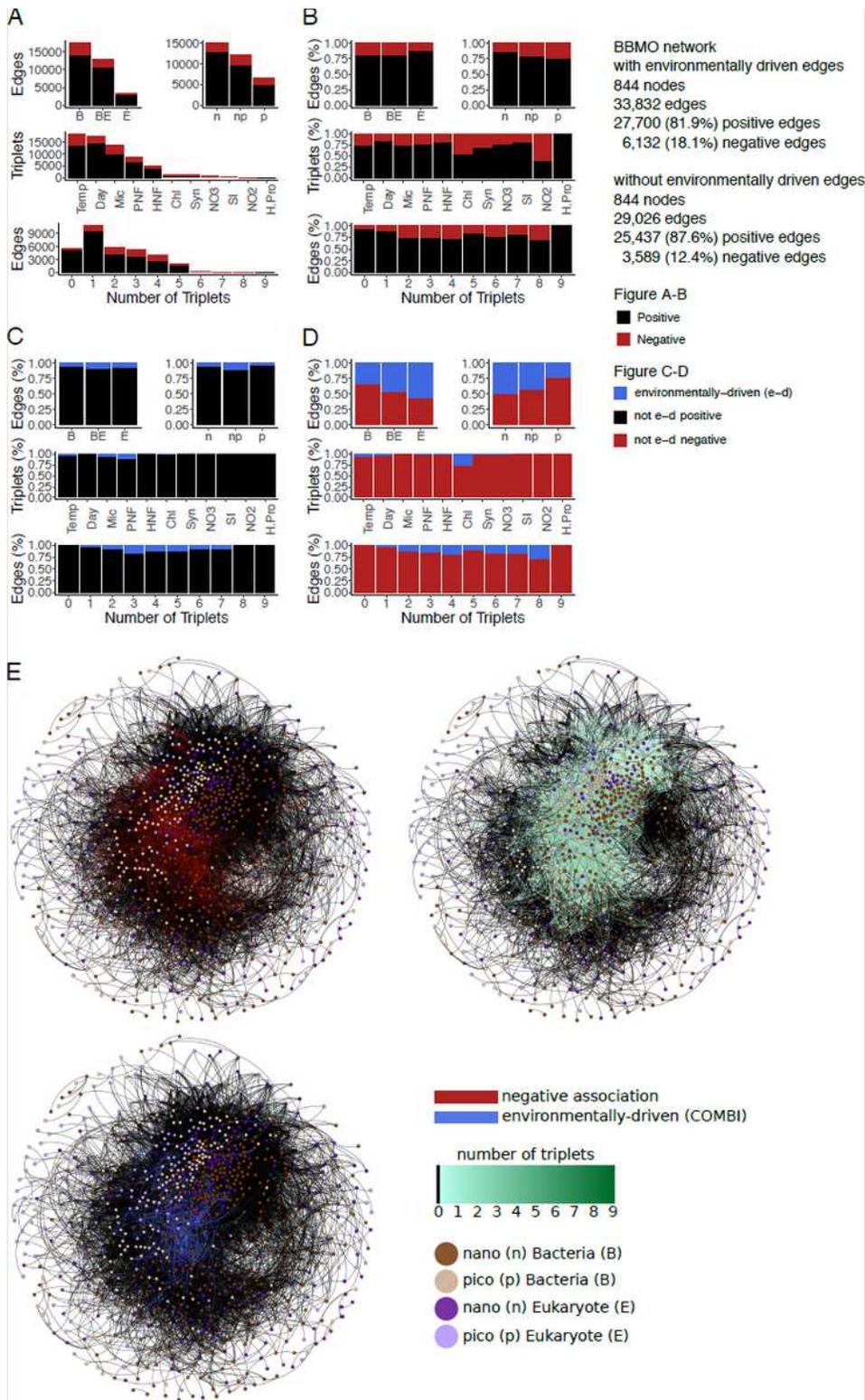


Figure 2

Quantification of environmentally-driven associations in the BBMO network For A) to D), the upper left figure shows the number (or fraction) of microbial associations divided by domain: Bacteria-Bacteria associations (B), Bacteria-Eukaryote associations (BE), and Eukaryote-Eukaryote associations (E). The upper right figure shows the number (or fraction) of associations divided by size-fractions: association within the nano size fraction (n), within the pico size fraction (p), and between these two size fractions

(np). The figure in the middle shows all triplets connected to an environmental parameter: Temperature (Temp), Day length (Day), Micromonas (Mic), photosynthetic Nanoflagellates (PNF), heterotrophic Nanoflagellates (HNF), Chlorophyll (Chl), Synechococcus (Syn), inorganic nutrients NO₃⁻ (NO₃), SiO₂ (Si), and NO₂⁻ (NO₂), as well as heterotrophic Prokaryotes (H.Pro). The figure at the bottom shows the number of edges divided in how many triplets they have been found ranging from no triplets (0) to nine triplets. Figure A) and B) display the number of microbial associations of the BBMO network before applying EnDED. Positive associations are indicated with black, negative associations with red. Figure C) and D) indicate in blue the fraction of environmentally-driven edges among the positive (C) and negative (D) microbial associations. E) The upper left network shows in black the positive and in red the negative associations. The upper right network shows in green the number of triplets a microbial edge is in ranging from one (light green) to nine (dark green), and no triplet (black). The lower network shows in blue the environmentally-driven associations that were detected by the intersection combination of the four methods Sign Pattern, Overlap, Interaction Information, and Data Processing Inequality.

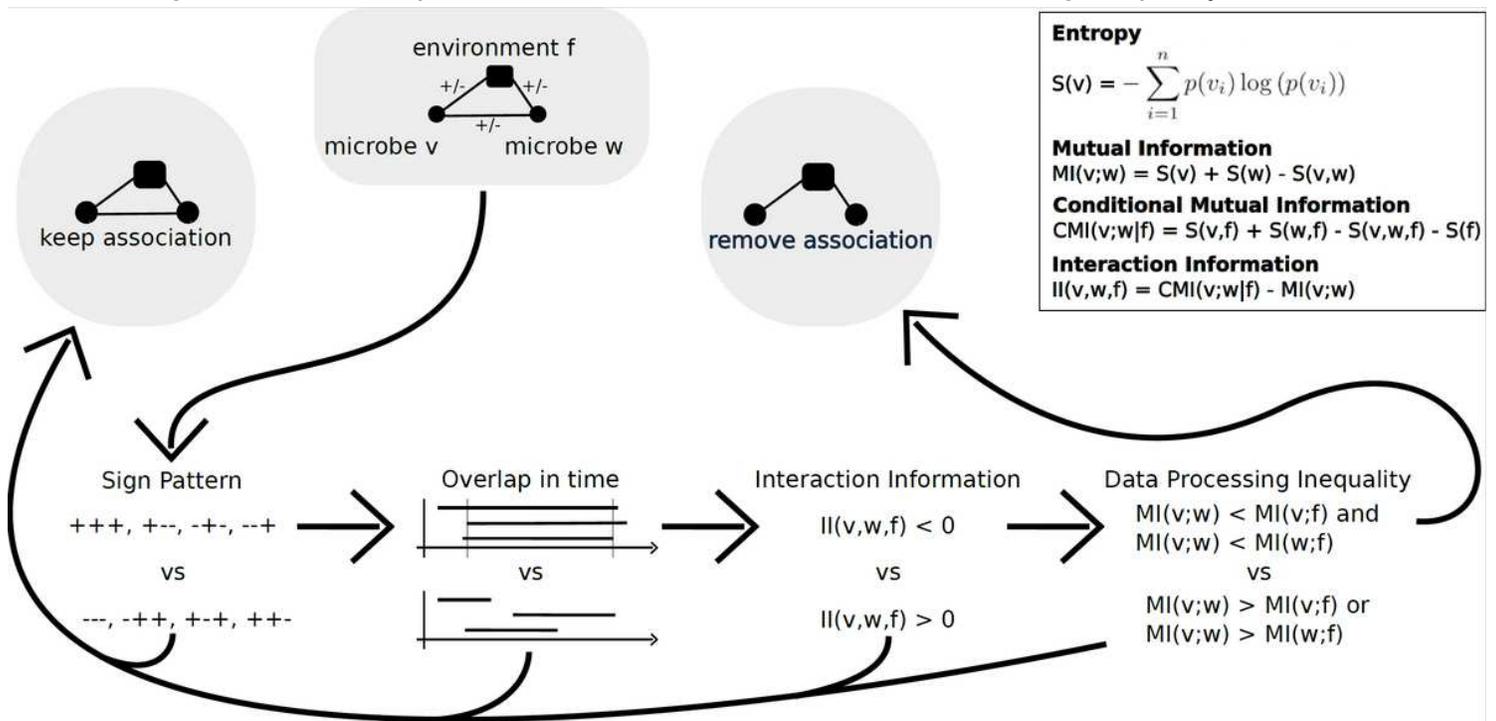


Figure 3

EnDED Methods Overview EnDED is an implementation of four methods aiming to determine whether an edge between two microorganisms is indirect through the action of an environmental factor. The four methods are: Sign Pattern, Overlap, Interaction Information, and Data Processing Inequality (see Methods). Each method can be used individually or in combination. Here, we show the intersection combination approach, i.e. only if all methods classify an edge as indirect, it is removed from the network. Otherwise, the edge is classified as not indirect and kept in the network.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AdditionalFilesTableS1S2S3.docx](#)