

Food web reconstruction through phylogenetic transfer of low-rank network representation

Tanya Strydom^{1,2,‡} Salomé Bouskila^{1,‡} Francis Banville^{1,3,2} Ceres Barros⁴ Dominique Caron^{5,2}
Maxwell J Farrell⁶ Marie-Josée Fortin⁶ Victoria Hemming⁷ Benjamin Mercier^{3,2} Laura
J. Pollock^{5,2} Rogini Runghen⁸ Giulio V. Dalla Riva⁹ Timothée Poisot^{1,2}

¹ Département de Sciences Biologiques, Université de Montréal, Montréal, Canada ² Quebec Centre for Biodiversity Science, Montréal, Canada ³ Département de Biologie, Université de Sherbrooke, Sherbrooke, Canada ⁴ Department of Forest Resources Management, University of British Columbia, Vancouver, B.C., Canada ⁵ Department of Biology, McGill University, Montréal, Canada ⁶ Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, Canada ⁷ Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, Canada ⁸ Centre for Integrative Ecology, School of Biological Sciences, University of Canterbury, Canterbury, New Zealand ⁹ School of Mathematics and Statistics, University of Canterbury, Canterbury, New Zealand

[‡] These authors contributed equally to the work

Correspondance to:

Timothée Poisot — timothee.poisot@umontreal.ca

1. Despite their importance in many ecological processes, collecting data and information on ecological interactions is an exceedingly challenging task. For this reason, large parts of the world have a data deficit when it comes to species interactions, and how the resulting networks are structured. As data collection alone is unlikely to be sufficient, community ecologists must adopt predictive methods.
2. We present a methodological framework that uses graph embedding and transfer learning to assemble a predicted list of trophic interactions of a species pool for which their interactions are unknown. Specifically, we ‘learn’ the information from a known interaction network by inferring the latent traits of species and infer the latent traits of a species pool for which we have no *a priori* interaction data based on their phylogenetic relatedness to species from the known network. The latent traits can then be used to predict interactions and construct an interaction network.
3. Here we assembled a metaweb for Canadian mammals derived from interactions in the European food web, despite only 4% of common species being shared between the two locations. The results of the predictive model are compared against databases of recorded pairwise interactions, showing that we correctly recover 91% of known interactions.
4. The framework itself is robust even when the known network is incomplete or contains spurious interactions making it an ideal candidate as a tool for filling gaps when it comes to species interactions. We provide guidance on how this framework can be adapted by substituting some approaches or predictors in order to make it more generally applicable.

1 Introduction

2 There are two core challenges we are faced with in furthering our understanding of ecological networks
3 across space, particularly at macro-ecologically relevant scales (e.g. Trøjelsgaard & Olesen, 2016). First,
4 ecological networks within a location are difficult to sample properly (Jordano, 2016a, 2016b), resulting in
5 a widespread “Eltonian shortfall” (Hortal et al., 2015), *i.e.* a lack of knowledge about inter- and intra-
6 specific relationships. This first challenge has been, in large part, addressed by the recent emergence of a
7 suite of methods aiming to predict interactions within *existing* networks, many of which are reviewed in
8 Strydom, Catchen, et al. (2021). Second, recent analyses based on collected data (Poisot, Bergeron, et al.,
9 2021) or metadata (Cameron et al., 2019) highlight that ecological networks are currently studied in a
10 biased subset of space and bioclimates, which impedes our ability to generalize any local understanding of
11 network structure. Meaning that, although the framework to address incompleteness *within* networks
12 exists, there would still be regions for which, due to a *lack* of local interaction data, we are unable to infer
13 potential species interactions.

14 Here, we present a general method to infer potential trophic interactions, relying on the transfer learning
15 of network representations, specifically by using similarities of species in a biologically/ecologically
16 relevant proxy space (e.g. shared morphology or ancestry). Transfer learning is a machine learning
17 methodology that uses the knowledge gained from solving one problem and applying it to a related
18 (destination) problem (Pan & Yang, 2010; Torrey & Shavlik, 2010). In this instance, we solve the problem
19 of predicting trophic interactions between species, based on knowledge extracted from another species
20 pool for which interactions are known by using phylogenetic structure as a medium for transfer. There is a
21 plurality of measures of species similarities that can be used for inferring *potential* species interactions *i.e.*
22 metaweb reconstruction (see e.g. Morales-Castilla et al., 2015); however, phylogenetic proximity has
23 several desirable properties when working at large scales. Gerhold et al. (2015) made the point that
24 phylogenetic signal captures diversification of characters (large macro-evolutionary process), but not
25 necessarily community assembly (fine ecological process); Dormann et al. (2010) previously found very
26 similar conclusions. Interactions tend to reflect a phylogenetic signal because they have a conserved
27 pattern of evolutionary convergence that encompasses a wide range of ecological and evolutionary
28 mechanisms (Cavender-Bares et al., 2009; Mouquet et al., 2012), and - most importantly - retain this signal
29 even if it is obscured at the community scale due to e.g. local conditions (Hutchinson et al., 2017; Poisot &

30 Stouffer, 2018). Finally, species interactions at macro-ecological scales seem to respond mostly to
31 macro-evolutionary processes (Price, 2003); which is evidenced by the presence of conserved backbones in
32 food webs (Dalla Riva & Stouffer, 2016; Mora et al., 2018), strong evolutionary signature on prey choice
33 (Stouffer et al., 2012), and strong phylogenetic signature in food web intervality (Eklöf & Stouffer, 2016).
34 Phylogenetic reconstruction has also previously been used within the context of ecological networks,
35 namely understanding ancestral plant-insect interactions (Braga et al., 2021). Taken together, these
36 considerations suggest that phylogenies can reliably be used to transfer knowledge on species interactions.

37 [Figure 1 about here.]

38 In fig. 1, we provide a methodological overview based on learning the embedding of a metaweb of trophic
39 interactions for European mammals (known interactions; Maiorano et al., 2020a, 2020b) and, based on
40 phylogenetic relationships between mammals globally (*i.e.*, phylogenetic tree Upham et al., 2019), infer a
41 metaweb for the Canadian mammalian species pool (using only a species list *i.e.* interactions are treated as
42 unknown in this instance). Our case study shows that phylogenetic transfer learning is an effective
43 approach to the generation of probabilistic metawebs. This showcases that although the components
44 (species) that make up the Canadian and European communities may be *minimally* shared (the overall
45 species overlap is less than 4%), if the medium (proxy space) selected in the transfer step is biologically
46 plausible, we can still effectively learn from the known network and make biologically relevant
47 predictions of interactions. Indeed, as we detail in the results, when validated against known but
48 fractional data of trophic interactions between Canadian mammals, our model achieves a predictive
49 accuracy of approximately 91%.

50 **Method description**

51 The core point of our method is the transfer of knowledge of a known ecological network to predict
52 interactions between species for another location for which the network is unknown (or partially known)
53 and is summarized in the grey text boxes in fig. 1. The method we develop is, ecologically speaking, a
54 “black box,” *i.e.* an algorithm that can be understood mathematically, but whose component parts are not
55 always directly tied to ecological processes. There is a growing realization in machine learning that
56 (unintentional) black box algorithms are not necessarily a bad thing (Holm, 2019), as long as their

57 constituent parts can be examined (which is the case with our method). But more importantly, data hold
58 more information than we might think; as such, even algorithms that are disconnected from a model can
59 make correct guesses most of the time (Halevy et al., 2009); in fact, in an instance of ecological forecasting
60 of spatio-temporal systems, model-free approaches (*i.e.* drawing all of their information from the data)
61 outperformed model-informed ones (Perretti et al., 2013).

62 **Data used for the case study**

63 We use data from the European metaweb assembled by Maiorano et al. (2020a). This was assembled using
64 data extracted from scientific literature (including published papers, books, and grey literature) from the
65 last 50 years and includes all terrestrial tetrapods (mammals, breeding birds, reptiles and amphibians)
66 occurring on the European sub-continent (and Turkey) - with the caveat that only species introduced in
67 historical times and currently naturalized being included. The European metaweb was filtered using the
68 Global Biodiversity Information Facility (GBIF) taxonomic backbone (GBIF Secretariat, 2021) so as to
69 contain only terrestrial and semi-aquatic mammals. As all species had valid matches to the GBIF
70 taxonomy it was used as the backbone for the remaining reconciliation steps namely, the mammalian
71 consensus supertree by Upham et al. (2019) (which is used for the knowledge transfer step) and for the
72 Canadian species list—which was extracted from the International Union for Conservation of Nature
73 (IUCN) checklist, and corresponds to the same selection criteria that was applied by Maiorano et al.
74 (2020a) in the European metaweb. After taxonomic cleaning and reconciliation the European metaweb
75 has 260 species, and the Canadian species pool 163; of these, 17 (about 4% of the total) are shared, and 89
76 species from Canada (54%) had at least one congeneric species in Europe. The similarity for both species
77 pools predictably increases with higher taxonomic order, with 19% of shared genera, 47% of shared
78 families, and 75% of shared orders; for the last point, Canada and Europe each had a single unique order
79 (*Didelphimorphia* for Canada, *Erinaceomorpha* for Europe).

80 **Implementation and code availability**

81 The entire pipeline is implemented in *Julia* 1.6 (Bezanson et al., 2017) and is available under the
82 permissive MIT License at <https://osf.io/2zwqm/>. The taxonomic cleanup steps are done using *GBIF.jl*
83 (Dansereau & Poisot, 2021). The network embedding and analysis is done using *EcologicalNetworks.jl*

84 (Banville et al., 2021; Poisot et al., 2019). The phylogenetic simulations are done using PhyloNetworks.jl
85 (Solís-Lemus et al., 2017) and Phylo.jl (Reeve et al., 2016). A complete Project.toml file specifying the
86 full tree of dependencies is available alongside the code. This material also includes a fully annotated copy
87 of the entire code required to run this project (describing both the intent of the code and discussing some
88 technical implementation details), a vignette for every step of the process, and a series of Jupyter
89 notebooks with the text and code. The pipeline can be executed on a laptop in a matter of minutes, and
90 therefore does not require extensive computational power.

91 **Step 1: Learning the origin network representation**

92 The first step in transfer learning is to learn the structure of the original dataset. In order to do so, we rely
93 on an approach inspired from representational learning, where we learn a *representation* of the metaweb
94 (in the form of the latent subspaces), rather than a list of interactions (species *a* eats *b*). This approach is
95 conceptually different from other metaweb-scale predictions (e.g. Albouy et al., 2019), in that the metaweb
96 representation is easily transferable. Specifically, we use a Random Dot Product Graph model (hereafter
97 RDPG; S. J. Young & Scheinerman, 2007) to create a number of latent variables that can be combined into
98 an approximation of the network adjacency matrix. RDPG is known to capture the evolutionary backbone
99 of food webs (Dalla Riva & Stouffer, 2016), resulting in strong phylogenetic signal in RDPG results; in
100 other words, the latent variables of an RDPG can be mapped onto a phylogenetic tree, and
101 phylogenetically similar predators should share phylogenetically similar preys. In addition, recent
102 advances show that the latent variables produced this way can be used to predict *de novo* interactions.
103 Interestingly, the latent variables do not need to be produced by decomposing the network itself; in a
104 recent contribution, Runghen et al. (2021) showed that deep artificial neural networks are able to
105 reconstruct the left and right subspaces of an RDPG, in order to predict human movement networks from
106 individual/location metadata and opens up the possibility of using additional metadata as predictors.
107 The latent variables are created by performing a truncated Singular Value Decomposition (t-SVD; Halko et
108 al., 2011) on the adjacency matrix. SVD is an appropriate embedding of ecological networks, which has
109 recently been shown to both capture their complex, emerging properties (Strydom, Dalla Riva, et al., 2021)
110 and to allow highly accurate prediction of the interactions within a single network (Poisot, Ouellet, et al.,
111 2021). Under SVD, an adjacency matrix \mathbf{A} (where $\mathbf{A}_{m,n} \in \mathbb{B}$ where 1 indicates predation and 0 an absence
112 thereof) is decomposed into three components resulting in $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}'$. Here, Σ is a $m \times n$ diagonal matrix

113 and contains only singular (σ) values along its diagonal, \mathbf{U} is a $m \times m$ unitary matrix, and \mathbf{V}' a $n \times n$
114 unitary matrix. Truncating the SVD removes additional noise in the dataset by omitting non-zero and/or
115 smaller σ values from Σ using the rank of the matrix. Under a t-SVD $\mathbf{A}_{m,n}$ is decomposed so that Σ is a
116 square $r \times r$ diagonal matrix (with $1 \leq r \leq r_{full}$ where r_{full} is the full rank of \mathbf{A} and r the rank at which we
117 truncate the matrix) containing only non-zero σ values. Additionally, \mathbf{U} is now a $m \times r$ semi unitary
118 matrix and \mathbf{V}' a $n \times r$ semi-unitary matrix.

119 The specific rank at which the SVD ought to be truncated is a difficult question. The purpose of SVD is to
120 remove the noise (expressed at high dimensions) and to focus on the signal, (expressed at low dimensions).
121 In datasets with a clear signal/noise demarcation, a scree plot of Σ can show a sharp drop at the rank where
122 noise starts (Zhu & Ghodsi, 2006). Because the European metaweb is almost entirely known, the amount
123 of noise (uncertainty) is low; this is reflected in fig. 2 (left), where the scree plot shows no important drop,
124 and in fig. 2 (right) where the proportion of variance explained increases smoothly at higher dimensions.
125 For this reason, we default back to a threshold that explains 60% of the variance in the underlying data,
126 corresponding to 12 dimensions - i.e. a tradeoff between accuracy and a reduced number of features.

127 An RDPG estimates the probability of observing interactions between nodes (species) as a function of the
128 nodes' latent variables, and is a way to turn an SVD (which decompose one matrix into three) into two
129 matrices that can be multiplied to provide an approximation of the network. The latent variables used for
130 the RDPG, called the left and right subspaces, are defined as $\mathcal{L} = \mathbf{U}\sqrt{\Sigma}$, and $\mathcal{R} = \sqrt{\Sigma}\mathbf{V}'$ – using the full
131 rank of \mathbf{A} , $\mathcal{L}\mathcal{R} = \mathbf{A}$, and using any smaller rank results in $\mathcal{L}\mathcal{R} \approx \mathbf{A}$. Using a rank of 1 for the t-SVD
132 provides a first-order approximation of the network. One advantage of using an RDPG rather than an SVD
133 is that the number of components to estimate decreases; notably, one does not have to estimate the
134 singular values of the SVD. Furthermore, the two subspaces can be directly multiplied to yield a network.

135 [Figure 2 about here.]

136 Because RDPG relies on matrix multiplication, the higher dimensions essentially serve to make specific
137 interactions converge towards 0 or 1; therefore, for reasonably low ranks, there is no guarantee that the
138 values in the reconstructed network will be within the unit range. In order to determine what constitutes
139 an appropriate threshold for probability, we performed the RDPG approach on the European metaweb,
140 and evaluated the probability threshold by treating this as a binary classification problem, specifically
141 assuming that both 0 and 1 in the European metaweb are all true. Given the methodological details given

142 in Maiorano et al. (2020a) and O'Connor et al. (2020), this seems like a reasonable assumption, although
143 one that does not hold for all metawebs. We used the thresholding approach presented in Poisot, Ouellet,
144 et al. (2021), and picked a cutoff that maximized Youden's J statistic (a measure of the informedness
145 (trust) of predictions; Youden (1950)); the resulting cutoff was 0.22, and gave an accuracy above 0.99. In
146 Supp. Mat. 1, we provide several lines of evidence that using the entire network to estimate the threshold
147 does not lead to overfitting; that using a subset of species would yield the same threshold; that decreasing
148 the quality of the original data by adding or removing interactions would minimally affect the predictive
149 accuracy of RDPG applied to the European metaweb; and that the networks reconstructed from artificially
150 modified data are reconstructed with the correct ecological properties.

151 The left and right subspaces for the European metaweb, accompanied by the threshold for prediction,
152 represent the knowledge we seek to transfer. In the next section, we explain how we rely on phylogenetic
153 similarity to do so.

154 **Steps 2 and 3: Transfer learning through phylogenetic relatedness**

155 In order to transfer the knowledge from the European metaweb to the Canadian species pool, we
156 performed ancestral character estimation using a Brownian motion model, which is a conservative
157 approach in the absence of strong hypotheses about the nature of phylogenetic signal in the network
158 decomposition (Litsios & Salamin, 2012). This uses the estimated feature vectors for the European
159 mammals to create a state reconstruction for all species (conceptually something akin to a trait-based
160 mammalian phylogeny using latent generality and vulnerability traits) and allows us to impute the
161 missing (latent) trait data for the Canadian species that are not already in the European network; as we are
162 focused on predicting contemporary interactions, we only retained the values for the tips of the tree. We
163 assumed that all traits (*i.e.* the feature vectors for the left and right subspaces) were independent, which is
164 a reasonable assumption as every trait/dimension added to the t-SVD has an *additive* effect to the one
165 before it. Note that the Upham et al. (2019) tree itself has some uncertainty associated to inner nodes of
166 the phylogeny. In this case study we have decided to not propagate this uncertainty as it would complexify
167 the process. The Brownian motion algorithm returns the *average* value of the trait, and its upper and
168 lower bounds. Because we do not estimate other parameters of the traits' distributions, we considered that
169 every species trait is represented as a uniform distribution between these bounds. The choice of the
170 uniform distribution was made because the algorithm returns a minimum and maximum point estimate

171 for the value, and given this information, the uniform distribution is the one with maximum entropy. Had
172 all mean parameters estimates been positive, the exponential distribution would have been an alternative,
173 but this is not the case for the subspaces of an RDPG. In order to examine the consequences of the choice
174 of distribution, we estimated the variance per latent variable per node to use a Normal distribution; as we
175 show in Supp. Mat. 2, this decision results in dramatically over-estimating the number and probability of
176 interactions, and therefore we keep the discussions in the main text to the uniform case. The inferred left
177 and right subspaces for the Canadian species pool ($\hat{\mathcal{L}}$ and $\hat{\mathcal{R}}$) have entries that are distributions,
178 representing the range of values for a given species at a given dimension. These objects represent the
179 transferred knowledge, which we can use for prediction of the Canadian metaweb.

180 **Step 4: Probabilistic prediction of the destination network**

181 The phylogenetic reconstruction of $\hat{\mathcal{L}}$ and $\hat{\mathcal{R}}$ has an associated uncertainty, represented by the breadth of
182 the uniform distribution associated to each of their entries. Therefore, we can use this information to
183 assemble a *probabilistic* metaweb in the sense of Poisot et al. (2016), *i.e.* in which every interaction is
184 represented as a single, independent, Bernoulli event of probability p .

185 [Figure 3 about here.]

186 Specifically, we have adopted the following approach. For every entry in $\hat{\mathcal{L}}$ and $\hat{\mathcal{R}}$, we draw a value from
187 its distribution. This results in one instance of the possible left ($\hat{\ell}$) and right (\hat{r}) subspaces for the
188 Canadian metaweb. These can be multiplied, to produce one matrix of real values. Because the entries in
189 $\hat{\ell}$ and \hat{r} are in the same space where \mathcal{L} and \mathcal{R} were originally predicted, it follows that the threshold ρ
190 estimated for the European metaweb also applies. We use this information to produce one random
191 Canadian metaweb, $N = \hat{\mathcal{L}}\hat{\mathcal{R}}' \geq \rho$. As we can see in (fig. 3), the European and Canadian metawebs are
192 structurally similar (as would be expected given the biogeographic similarities) and the two (left and right)
193 subspaces are distinct *i.e.* capturing predation (generality) and prey (vulnerability) latent traits.

194 Because the intervals around some trait values can be broad (in fact, probably broader than what they
195 would actually be, see *e.g.* Garland et al., 1999), we repeat the above process 2×10^5 times, which results in
196 a probabilistic metaweb P , where the probability of an interaction (here conveying our degree of trust that
197 it exists given the inferred trait distributions) is given by the number of times where it appears across all

198 random draws N , divided by the number of samples. An interaction with $P_{i,j} = 1$ means that these two
199 species were predicted to interact in all 2×10^5 random draws.

200 It must be noted that despite bringing in a large amount of information from the European species pool
201 and interactions, the Canadian metaweb has distinct structural properties. Following an approach similar
202 to Vermaat et al. (2009), we show in Supp. Mat. 3 that not only can we observe differences in a
203 multivariate space between the European and Canadian metaweb, we can also observe differences in the
204 same space between random subgraphs from these networks. These results line up with the studies
205 spatializing metawebs that have been discussed in the introduction: changes in the species pool are
206 driving local structural changes in the networks.

207 **Data cleanup, discovery, validation, and thresholding**

208 Once the probabilistic metaweb for Canada has been produced, we followed a number of data inflation
209 steps to finalize it. This step is external to the actual transfer learning framework but rather serves as a
210 way to augment and validate the predicted metaweb.

211 [Figure 4 about here.]

212 First, we extracted the network corresponding to the 17 species shared between the European and
213 Canadian pools and replaced these interactions with a probability of 0 (non-interaction) or 1 (interaction),
214 according to their value in the European metaweb. This represents a minute modification of the inferred
215 network (about 0.8% of all species pairs from the Canadian web), but ensures that we are directly re-using
216 knowledge from Europe.

217 Second, we looked for all species in the Canadian pool known to the Global Biotic Interactions (GloBI)
218 database (Poelen et al., 2014), and extracted their known interactions. Because GloBI aggregates observed
219 interactions, it is not a *networks* data source, and therefore the only information we can reliably extract
220 from it is that a species pair *was reported to interact at least once*. This last statement should yet be taken
221 with caution, as some sources in GloBI (e.g. Thessen & Parr, 2014) are produced through text analysis,
222 and therefore may not document direct evidence of the interaction. Nevertheless, should the predictive
223 model work, we would expect that a majority of interactions known to GloBI would also be predicted. We
224 retrieved 366 interactions between mammals from the Canadian species pool from GloBI, 33 of which

225 were not predicted by the model; this results in a success rate of 91%. After performing this check, we set
226 the probability of all interactions known to GLoBI to 1.

227 Finally, we downloaded the data from Strong & Leroux (2014), who mined various literature sources to
228 identify trophic interactions in Newfoundland. This dataset documented 25 interactions between
229 mammals, only two of which were not part of our (Canada-level) predictions, resulting in a success rate of
230 92%. These two interactions were added to our predicted metaweb with a probability of 1. A table listing
231 all interactions in the predicted Canadian metaweb can be found in the supplementary material.

232 [Figure 5 about here.]

233 Because the confidence intervals on the inferred trait space are probably over-estimates, we decided to
234 apply a thresholding step to the interactions after data inflation (see fig. 5 showing the effect of varying the
235 cutoff on $P(i \rightarrow j)$). Cirtwill & Hambäck (2021) proposed a number of strategies to threshold probabilistic
236 networks. Their methods assume the underlying data to be tag-based sequencing, which represents
237 interactions as co-occurrences of predator and prey within the same tags; this is conceptually identical to
238 our Bernoulli-trial based reconstruction of a probabilistic network. We performed a full analysis of the
239 effect of various cutoffs, and as they either resulted in removing too few interactions, or removing enough
240 interactions that species started to be disconnected from the network, we set this threshold for a
241 probability equivalent to 0 to the largest possible value that still allowed all species to have at least one
242 interaction with a non-zero probability. The need for this slight deviation from the Cirtwill & Hambäck
243 (2021) method highlights the need for additional development on network thresholding.

244 Results and discussion of the case study

245 [Figure 6 about here.]

246 Using a transfer learning framework we were able to construct a probabilistic metaweb and (as per
247 Dunne, 2006) it is a list of potential interactions and does not mean that they will exist wherever the two
248 species co-occur. The t-SVD embedding is able to learn relevant ecological features for the network. fig. 6
249 shows that the first rank correlates linearly with generality and vulnerability (Schoener, 1989), i.e. the
250 number of preys and predators for each species. Importantly, this implies that a rank 1 approximation

251 represents the configuration model for the metaweb, *i.e.* a set of random networks generated from a given
252 degree sequence (Park & Newman, 2004). Accounting for the probabilistic nature of the degrees, the rank
253 1 approximation also represents the *soft* configuration model (van der Hoorn et al., 2018). Both models are
254 maximum entropy graph models (Garlaschelli et al., 2018), with sharp (all network realizations satisfy the
255 specified degree sequence) and soft (network realizations satisfy the degree sequence on average) local
256 constraints, respectively. The (soft) configuration model is an unbiased random graph model widely used
257 by ecologists in the context of null hypothesis significance testing of network structure (*e.g.* Bascompte et
258 al., 2003) and can provide informative priors for Bayesian inference of network structure (*e.g.* J.-G. Young
259 et al., 2021). It is noteworthy that for this metaweb, the relevant information was extracted at the first
260 rank. Because the first rank corresponds to the leading singular value of the system, the results of fig. 6
261 have a straightforward interpretation: degree-based processes are the most important in structuring the
262 mammalian food web.

263 Discussion

264 One important aspect in which Europe and Canada differ (despite their comparable bioclimatic
265 conditions) is the degree of the legacy of human impacts, which have been much longer in Europe.
266 Nenzén et al. (2014) showed that even at small scales (the Iberian peninsula), mammal food webs retain
267 the signal of both past climate change and human activity, even when this human activity was orders of
268 magnitude less important than it is now. Similarly, Yeakel et al. (2014) showed that changes in human
269 occupation over several centuries can lead to food web collapse. Megafauna in particular seems to be very
270 sensitive to human arrival (Pires et al., 2015). In short, there is well-substantiated support for the idea that
271 human footprint affects more than the risk of species extinction (Marco et al., 2018), and can lead to
272 changes in interaction structure.

273 Cirtwill et al. (2019) showed that network inference techniques based on Bayesian approaches would
274 perform far better in the presence of an interaction-level informative prior; the desirable properties of such
275 a prior would be that it is expressed as a probability, preferably representing a Bernoulli event, the value of
276 which would be representative of relevant biological processes (probability of predation in this case). We
277 argue that the probability returned at the very last step of our framework may serve as this informative
278 prior; indeed, the output of our analysis can be used in subsequent steps, also possibly involving expert

279 elicitation to validate some of the most strongly recommended interactions. One important *caveat* to keep
280 in mind when working with interaction inference is that interactions can never really be true negatives (in
281 the current state of our methodological framework and data collection limitations); this renders the task of
282 validating a model through the usual application of binary classification statistics very difficult (although
283 see Strydom, Catchen, et al., 2021 for a discussion of alternative suggestions). The other way through
284 which our framework can be improved is by substituting the predictors that are used for transfer. For
285 example, in the presence of information on species traits that are known to be predictive of species
286 interactions, one might want to rely on functional rather than phylogenetic distances – in food webs, body
287 size (and allometrically related variables) has been established as such a variable (Brose et al., 2006); the
288 identification of relevant functional traits is facilitated by recent methodological developments (Rosado et
289 al., 2013).

290 Finally, it should be noted that the framework we have presented is amenable to changes lending to
291 applicability to a broad range of potential scenarios. For example in this case study we have embedded the
292 original metaweb using t-SVD, because it lends itself to an RDPG reconstruction, which is known to
293 capture the consequences of evolutionary processes (Dalla Riva & Stouffer, 2016); this being said, there are
294 other ways to embed graphs (Arsov & Mirceva, 2019; Cai et al., 2017; Cao et al., 2019), which can be used
295 as alternatives. Regarding the transfer step it is possible to use distinct trees if working with distinct clades
296 (such as pollination networks) or an alternative measure of similarity (transfer medium) such as
297 information on foraging (Beckerman et al., 2006), cell-level mechanisms (Boeckaerts et al., 2021), or a
298 combination of traits and phylogenetic structure (Stock, 2021). Most importantly, although we focus on a
299 trophic system, it is an established fact that different (non-trophic) interactions do themselves interact with
300 and influence the outcome of trophic interactions (see e.g. Kawatsu et al., 2021; Kéfi et al., 2012). Future
301 development of metaweb inference techniques should cover the prediction of multiple interaction types.

302 **Acknowledgements:** We acknowledge that this study was conducted on land within the traditional
303 unceded territory of the Saint Lawrence Iroquoian, Anishinabewaki, Mohawk, Huron-Wendat, and
304 Omàmiwininiwak nations. TP, TS, DC, and LP received funding from the Canadian Institute for Ecology &
305 Evolution. FB is funded by the Institute for Data Valorization (IVADO). TS, SB, and TP are funded by a
306 donation from the Courtois Foundation. CB was awarded a Mitacs Elevate Fellowship no. IT12391, in
307 partnership with fRI Research, and also acknowledges funding from Alberta Innovates and the Forest
308 Resources Improvement Association of Alberta. M-JF acknowledges funding from NSERC Discovery

309 Grant and NSERC CRC. RR is funded by New Zealand's Biological Heritage Ngā Koiora Tuku Iho
310 National Science Challenge, administered by New Zealand Ministry of Business, Innovation, and
311 Employment. BM is funded by the NSERC Alexander Graham Bell Canada Graduate Scholarship and the
312 FRQNT master's scholarship. LP acknowledges funding from NSERC Discovery Grant (NSERC
313 RGPIN-2019-05771). TP acknowledges financial support from NSERC through the Discovery Grants and
314 Discovery Accelerator Supplement programs.

315 **Conflict of interest:** The authors have no conflict interests to disclose

316 **Authors' contributions:** TS, SB, and TP designed the study and performed the analysis; GVDR, MF, and
317 RR provided additional feedback on the analyses. DC, BM, and FB helped with data collection. All
318 authors contributed to writing and editing the manuscript.

319 **Data availability:** All code and data used in this manuscript is publicly available and archived on OSF
320 <https://osf.io/2zwqm/> and is currently referenced in the manuscript.

321 References

- 322 Albouy, C., Archambault, P., Appeltans, W., Araújo, M. B., Beauchesne, D., Cazelles, K., Cirtwill, A. R.,
323 Fortin, M.-J., Galiana, N., Leroux, S. J., Pellissier, L., Poisot, T., Stouffer, D. B., Wood, S. A., & Gravel, D.
324 (2019). The marine fish food web is globally connected. *Nature Ecology & Evolution*, 3(8, 8),
325 1153–1161. <https://doi.org/10.1038/s41559-019-0950-y>
- 326 Arsov, N., & Mirceva, G. (2019, November 26). *Network Embedding: An Overview*.
327 <http://arxiv.org/abs/1911.11726>
- 328 Banville, F., Vissault, S., & Poisot, T. (2021). Mangal.jl and EcologicalNetworks.jl: Two complementary
329 packages for analyzing ecological networks in Julia. *Journal of Open Source Software*, 6(61), 2721.
330 <https://doi.org/10.21105/joss.02721>
- 331 Bascompte, J., Jordano, P., Melian, C. J., & Olesen, J. M. (2003). The nested assembly of plant-animal
332 mutualistic networks. *Proceedings of the National Academy of Sciences*, 100(16), 9383–9387.
333 <https://doi.org/10.1073/pnas.1633576100>
- 334 Beckerman, A. P., Petchey, O. L., & Warren, P. H. (2006). Foraging biology predicts food web complexity.
335 *Proceedings of the National Academy of Sciences*, 103(37), 13745–13749.

- 336 <https://doi.org/10.1073/pnas.0603039103>
- 337 Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. (2017). Julia: A Fresh Approach to Numerical
338 Computing. *SIAM Review*, 59(1), 65–98. <https://doi.org/10.1137/141000671>
- 339 Boeckaerts, D., Stock, M., Criel, B., Gerstmans, H., De Baets, B., & Briers, Y. (2021). Predicting
340 bacteriophage hosts based on sequences of annotated receptor-binding proteins. *Scientific Reports*,
341 11(1, 1), 1467. <https://doi.org/10.1038/s41598-021-81063-4>
- 342 Braga, M. P., Janz, N., Nylin, S., Ronquist, F., & Landis, M. J. (2021). Phylogenetic reconstruction of
343 ancestral ecological networks through time for pierid butterflies and their host plants. *Ecology Letters*,
344 n/a(n/a). <https://doi.org/10.1111/ele.13842>
- 345 Brose, U., Jonsson, T., Berlow, E. L., Warren, P., Banasek-Richter, C., Bersier, L.-F., Blanchard, J. L., Brey,
346 T., Carpenter, S. R., Blandenier, M.-F. C., Cushing, L., Dawah, H. A., Dell, T., Edwards, F.,
347 Harper-Smith, S., Jacob, U., Ledger, M. E., Martinez, N. D., Memmott, J., ... Cohen, J. E. (2006).
348 ConsumerResource Body-Size Relationships in Natural Food Webs. *Ecology*, 87(10), 2411–2417.
349 [https://doi.org/10.1890/0012-9658\(2006\)87%5B2411:CBRINF%5D2.0.CO;2](https://doi.org/10.1890/0012-9658(2006)87%5B2411:CBRINF%5D2.0.CO;2)
- 350 Cai, H., Zheng, V. W., & Chang, K. C.-C. (2017). *A Comprehensive Survey of Graph Embedding: Problems,*
351 *Techniques and Applications*. <http://arxiv.org/abs/1709.07604>
- 352 Cameron, E. K., Sundqvist, M. K., Keith, S. A., CaraDonna, P. J., Mousing, E. A., Nilsson, K. A., Metcalfe,
353 D. B., & Classen, A. T. (2019). Uneven global distribution of food web studies under climate change.
354 *Ecosphere*, 10(3), e02645. <https://doi.org/10.1002/ecs2.2645>
- 355 Cao, R.-M., Liu, S.-Y., & Xu, X.-K. (2019). Network embedding for link prediction: The pitfall and
356 improvement. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(10), 103102.
357 <https://doi.org/10.1063/1.5120724>
- 358 Cavender-Bares, J., Kozak, K. H., Fine, P. V. A., & Kembel, S. W. (2009). The merging of community
359 ecology and phylogenetic biology. *Ecology Letters*, 12(7), 693–715.
360 <https://doi.org/10.1111/j.1461-0248.2009.01314.x>
- 361 Cirtwill, A. R., Ekl, A., Roslin, T., Wootton, K., & Gravel, D. (2019). A quantitative framework for
362 investigating the reliability of empirical network construction. *Methods in Ecology and Evolution*, 0.
363 <https://doi.org/10.1111/2041-210X.13180>

- 364 Cirtwill, A. R., & Hambäck, P. (2021). Building food networks from molecular data: Bayesian or
365 fixed-number thresholds for including links. *Basic and Applied Ecology*, 50, 67–76.
366 <https://doi.org/10.1016/j.baae.2020.11.007>
- 367 Dalla Riva, G. V., & Stouffer, D. B. (2016). Exploring the evolutionary signature of food webs' backbones
368 using functional traits. *Oikos*, 125(4), 446–456. <https://doi.org/10.1111/oik.02305>
- 369 Dansereau, G., & Poisot, T. (2021). SimpleSDMLayers.jl and GBIF.jl: A Framework for Species
370 Distribution Modeling in Julia. *Journal of Open Source Software*, 6(57), 2872.
371 <https://doi.org/10.21105/joss.02872>
- 372 Dormann, C. F., Gruber, B., Winter, M., & Herrmann, D. (2010). Evolution of climate niches in European
373 mammals? *Biology Letters*, 6(2), 229–232. <https://doi.org/10.1098/rsbl.2009.0688>
- 374 Dunne, J. A. (2006). The Network Structure of Food Webs. In J. A. Dunne & M. Pascual (Eds.), *Ecological
375 networks: Linking structure and dynamics* (pp. 27–86). Oxford University Press.
- 376 Eklöf, A., & Stouffer, D. B. (2016). The phylogenetic component of food web structure and intervality.
377 *Theoretical Ecology*, 9(1), 107–115. <https://doi.org/10.1007/s12080-015-0273-9>
- 378 Garland, T., JR., Midford, P. E., & Ives, A. R. (1999). An Introduction to Phylogenetically Based Statistical
379 Methods, with a New Method for Confidence Intervals on Ancestral Values1. *American Zoologist*,
380 39(2), 374–388. <https://doi.org/10.1093/icb/39.2.374>
- 381 Garlaschelli, D., Hollander, F. den, & Roccaverde, A. (2018). Covariance structure behind breaking of
382 ensemble equivalence in random graphs. *Journal of Statistical Physics*, 173(3-4), 644–662.
383 <https://doi.org/10.1007/s10955-018-2114-x>
- 384 GBIF Secretariat. (2021). *GBIF Backbone Taxonomy*. <https://doi.org/10.15468/39omei>
- 385 Gerhold, P., Cahill, J. F., Winter, M., Bartish, I. V., & Prinzing, A. (2015). Phylogenetic patterns are not
386 proxies of community assembly mechanisms (they are far better). *Functional Ecology*, 29(5), 600–614.
387 <https://doi.org/10.1111/1365-2435.12425>
- 388 Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent
389 Systems*, 24(2), 8–12. <https://doi.org/10.1109/MIS.2009.36>
- 390 Halko, N., Martinsson, P. G., & Tropp, J. A. (2011). Finding Structure with Randomness: Probabilistic
391 Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Review*, 53(2), 217–288.

- 392 <https://doi.org/10.1137/090771806>
- 393 Holm, E. A. (2019). In defense of the black box. *Science*, 364(6435), 26–27.
- 394 <https://doi.org/10.1126/science.aax0162>
- 395 Hortal, J., de Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M., & Ladle, R. J. (2015). Seven
396 Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annual Review of Ecology, Evolution, and*
397 *Systematics*, 46(1), 523–549. <https://doi.org/10.1146/annurev-ecolsys-112414-054400>
- 398 Hutchinson, M. C., Cagua, E. F., & Stouffer, D. B. (2017). Cophylogenetic signal is detectable in pollination
399 interactions across ecological scales. *Ecology*, n/a–n/a. <https://doi.org/10.1002/ecy.1955>
- 400 Jordano, P. (2016a). Chasing Ecological Interactions. *PLOS Biol*, 14(9), e1002559.
401 <https://doi.org/10.1371/journal.pbio.1002559>
- 402 Jordano, P. (2016b). Sampling networks of ecological interactions. *Functional Ecology*, 30(12), 1883–1893.
403 <https://doi.org/10.1111/1365-2435.12763>
- 404 Kawatsu, K., Ushio, M., van Veen, F. J. F., & Kondoh, M. (2021). Are networks of trophic interactions
405 sufficient for understanding the dynamics of multi-trophic communities? Analysis of a tri-trophic
406 insect food-web time-series. *Ecology Letters*, 24(3), 543–552. <https://doi.org/10.1111/ele.13672>
- 407 Kéfi, S., Berlow, E. L., Wieters, E. A., Navarrete, S. A., Petchey, O. L., Wood, S. A., Boit, A., Joppa, L. N.,
408 Lafferty, K. D., Williams, R. J., Martinez, N. D., Menge, B. A., Blanchette, C. A., Iles, A. C., & Brose, U.
409 (2012). More than a meal... integrating non-feeding interactions into food webs: More than a meal
410 *Ecology Letters*, 15(4), 291–300. <https://doi.org/10.1111/j.1461-0248.2011.01732.x>
- 411 Litsios, G., & Salamin, N. (2012). Effects of Phylogenetic Signal on Ancestral State Reconstruction.
412 *Systematic Biology*, 61(3), 533–538. <https://doi.org/10.1093/sysbio/syr124>
- 413 Maiorano, L., Montemaggioli, A., Ficetola, G. F., O'Connor, L., & Thuiller, W. (2020a). TETRA-EU 1.0: A
414 species-level trophic metaweb of European tetrapods. *Global Ecology and Biogeography*, 29(9),
415 1452–1457. <https://doi.org/10.1111/geb.13138>
- 416 Maiorano, L., Montemaggioli, A., Ficetola, G. F., O'Connor, L., & Thuiller, W. (2020b). *Data from:
417 Tetra-EU 1.0: A species-level trophic meta-web of European tetrapods* (Version 3, pp. 16596876 bytes)
418 [Data set]. Dryad. <https://doi.org/10.5061/DRYAD.JM63XSJ7B>
- 419 Marco, M. D., Venter, O., Possingham, H. P., & Watson, J. E. M. (2018). Changes in human footprint drive

- 420 changes in species extinction risk. *Nature Communications*, 9(1), 4621.
- 421 <https://doi.org/10.1038/s41467-018-07049-5>
- 422 Mora, B. B., Gravel, D., Gilarranz, L. J., Poisot, T., & Stouffer, D. B. (2018). Identifying a common backbone
423 of interactions underlying food webs from different ecosystems. *Nature Communications*, 9(1), 2603.
- 424 <https://doi.org/10.1038/s41467-018-05056-0>
- 425 Morales-Castilla, I., Matias, M. G., Gravel, D., & Araújo, M. B. (2015). Inferring biotic interactions from
426 proxies. *Trends in Ecology & Evolution*, 30(6), 347–356.
- 427 <https://doi.org/10.1016/j.tree.2015.03.014>
- 428 Mouquet, N., Devictor, V., Meynard, C. N., Munoz, F., Bersier, L.-F., Chave, J., Couteron, P., Dalecky, A.,
429 Fontaine, C., Gravel, D., Hardy, O. J., Jabot, F., Lavergne, S., Leibold, M., Mouillot, D., Münkemüller,
430 T., Pavoine, S., Prinzing, A., Rodrigues, A. S. L., ... Thuiller, W. (2012). Ecophylogenetics: Advances
431 and perspectives. *Biological Reviews*, 87(4), 769–785.
- 432 <https://doi.org/10.1111/j.1469-185X.2012.00224.x>
- 433 Nenzén, H. K., Montoya, D., & Varela, S. (2014). The Impact of 850,000 Years of Climate Changes on the
434 Structure and Dynamics of Mammal Food Webs. *PLOS ONE*, 9(9), e106651.
- 435 <https://doi.org/10.1371/journal.pone.0106651>
- 436 O'Connor, L. M. J., Pollock, L. J., Braga, J., Ficetola, G. F., Maiorano, L., Martinez-Almoyna, C.,
437 Montemaggioli, A., Ohlmann, M., & Thuiller, W. (2020). Unveiling the food webs of tetrapods across
438 Europe through the prism of the Eltonian niche. *Journal of Biogeography*, 47(1), 181–192.
- 439 <https://doi.org/10.1111/jbi.13773>
- 440 Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data
441 Engineering*, 22(10), 1345–1359. <https://doi.org/10.1109/TKDE.2009.191>
- 442 Park, J., & Newman, M. E. J. (2004). Statistical mechanics of networks. *Physical Review E*, 70(6), 066117.
- 443 <https://doi.org/10.1103/PhysRevE.70.066117>
- 444 Perretti, C. T., Munch, S. B., & Sugihara, G. (2013). Model-free forecasting outperforms the correct
445 mechanistic model for simulated and experimental data. *Proceedings of the National Academy of
446 Sciences*, 110(13), 5253–5257. <https://doi.org/10.1073/pnas.1216076110>
- 447 Pires, M. M., Koch, P. L., Fariña, R. A., de Aguiar, M. A. M., dos Reis, S. F., & Guimarães, P. R. (2015).

- 448 Pleistocene megafaunal interaction networks became more vulnerable after human arrival.
- 449 *Proceedings of the Royal Society B: Biological Sciences*, 282(1814), 20151367.
- 450 <https://doi.org/10.1098/rspb.2015.1367>
- 451 Poelen, J. H., Simons, J. D., & Mungall, C. J. (2014). Global biotic interactions: An open infrastructure to
452 share and analyze species-interaction datasets. *Ecological Informatics*, 24, 148–159.
- 453 <https://doi.org/10.1016/j.ecoinf.2014.08.005>
- 454 Poisot, T., Belisle, Z., Hoebeke, L., Stock, M., & Szefer, P. (2019). EcologicalNetworks.jl - analysing
455 ecological networks. *Ecography*. <https://doi.org/10.1111/ecog.04310>
- 456 Poisot, T., Bergeron, G., Cazelles, K., Dallas, T., Gravel, D., MacDonald, A., Mercier, B., Violet, C., &
457 Vissault, S. (2021). Global knowledge gaps in species interaction networks data. *Journal of
458 Biogeography*, n/a(n/a). <https://doi.org/10.1111/jbi.14127>
- 459 Poisot, T., Cirtwill, A. R., Cazelles, K., Gravel, D., Fortin, M.-J., & Stouffer, D. B. (2016). The structure of
460 probabilistic networks. *Methods in Ecology and Evolution*, 7(3), 303–312.
461 <https://doi.org/10.1111/2041-210X.12468>
- 462 Poisot, T., Ouellet, M.-A., Mollentze, N., Farrell, M. J., Becker, D. J., Albery, G. F., Gibb, R. J., Seifert, S. N.,
463 & Carlson, C. J. (2021, May 31). *Imputing the mammalian virome with linear filtering and singular
464 value decomposition*. <http://arxiv.org/abs/2105.14973>
- 465 Poisot, T., & Stouffer, D. B. (2018). Interactions retain the co-phylogenetic matching that communities lost.
466 *Oikos*, 127(2), 230–238. <https://doi.org/10.1111/oik.03788>
- 467 Price, P. W. (2003). *Macroevolutionary theory on macroecological patterns*. Cambridge University Press.
- 468 Reeve, R., Leinster, T., Cobbold, C. A., Thompson, J., Brummitt, N., Mitchell, S. N., & Matthews, L. (2016,
469 December 8). *How to partition diversity*. <http://arxiv.org/abs/1404.6520>
- 470 Rosado, B. H. P., Dias, A., & de Mattos, E. (2013). Going Back to Basics: Importance of Ecophysiology
471 when Choosing Functional Traits for Studying Communities and Ecosystems. *Natureza &
472 Conservaç~ao Revista Brasileira de Conservaç~ao Da Natureza*, 11, 15–22.
473 <https://doi.org/10.4322/natcon.2013.002>
- 474 Runghen, R., Stouffer, D. B., & Dalla Riva, G. V. (2021). *Exploiting node metadata to predict interactions in
475 large networks using graph embedding and neural networks*.

- 476 <https://doi.org/10.1101/2021.06.10.447991>
- 477 Schoener, T. W. (1989). Food webs from the small to the large. *Ecology*, 70(6), 1559–1589.
- 478 Solís-Lemus, C., Bastide, P., & Ané, C. (2017). PhyloNetworks: A Package for Phylogenetic Networks. *Molecular Biology and Evolution*, 34(12), 3292–3298. <https://doi.org/10.1093/molbev/msx235>
- 480 Stock, M. (2021). Pairwise learning for predicting pollination interactions based on traits and phylogeny. *Ecological Modelling*, 14.
- 481
- 482 Stouffer, D. B., Sales-Pardo, M., Sirer, M. I., & Bascompte, J. (2012). Evolutionary Conservation of Species' Roles in Food Webs. *Science*, 335(6075), 1489–1492. <https://doi.org/10.1126/science.1216556>
- 483
- 484 Strong, J. S., & Leroux, S. J. (2014). Impact of Non-Native Terrestrial Mammals on the Structure of the Terrestrial Mammal Food Web of Newfoundland, Canada. *PLOS ONE*, 9(8), e106264. <https://doi.org/10.1371/journal.pone.0106264>
- 485
- 486
- 487 Strydom, T., Catchen, M. D., Banville, F., Caron, D., Dansereau, G., Desjardins-Proulx, P., Forero-Muñoz, N. R., Higino, G., Mercier, B., Gonzalez, A., Gravel, D., Pollock, L., & Poisot, T. (2021). A roadmap towards predicting species interaction networks (across space and time). *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1837), 20210063. <https://doi.org/10.1098/rstb.2021.0063>
- 488
- 489
- 490
- 491
- 492 Strydom, T., Dalla Riva, G. V., & Poisot, T. (2021). SVD Entropy Reveals the High Complexity of Ecological Networks. *Frontiers in Ecology and Evolution*, 9. <https://doi.org/10.3389/fevo.2021.623141>
- 493
- 494 Thessen, A. E., & Parr, C. S. (2014). Knowledge extraction and semantic annotation of text from the encyclopedia of life. *PloS One*, 9(3), e89550.
- 495
- 496 Torrey, L., & Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques* (pp. 242–264). IGI global.
- 497
- 498 Trøjelsgaard, K., & Olesen, J. M. (2016). Ecological networks in motion: Micro- and macroscopic variability across scales. *Functional Ecology*, 30(12), 1926–1935. <https://doi.org/10.1111/1365-2435.12710>
- 499
- 500
- 501 Upham, N. S., Esselstyn, J. A., & Jetz, W. (2019). Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLOS Biology*, 17(12), e3000494. <https://doi.org/10.1371/journal.pbio.3000494>
- 502
- 503

- 504 van der Hoorn, P., Lippner, G., & Krioukov, D. (2018). Sparse Maximum-Entropy Random Graphs with a
505 Given Power-Law Degree Distribution. *Journal of Statistical Physics*, 173(3-4), 806–844.
- 506 <https://doi.org/10.1007/s10955-017-1887-7>
- 507 Vermaat, J. E., Dunne, J. A., & Gilbert, A. J. (2009). Major dimensions in food-web structure properties.
508 *Ecology*, 90(1), 278–282. <http://www.ncbi.nlm.nih.gov/pubmed/19294932>
- 509 Yeakel, J. D., Pires, M. M., Rudolf, L., Dominy, N. J., Koch, P. L., Guimarães, P. R., & Gross, T. (2014).
510 Collapse of an ecological network in Ancient Egypt. *PNAS*, 111(40), 14472–14477.
- 511 <https://doi.org/10.1073/pnas.1408471111>
- 512 Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32–35.
513 [https://doi.org/10.1002/1097-0142\(1950\)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3](https://doi.org/10.1002/1097-0142(1950)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3)
- 514 Young, J.-G., Cantwell, G. T., & Newman, M. E. J. (2021). Bayesian inference of network structure from
515 unreliable data. *Journal of Complex Networks*, 8(6). <https://doi.org/10.1093/comnet/cnaa046>
- 516 Young, S. J., & Scheinerman, E. R. (2007). Random Dot Product Graph Models for Social Networks. In A.
517 Bonato & F. R. K. Chung (Eds.), *Algorithms and Models for the Web-Graph* (pp. 138–149). Springer.
518 https://doi.org/10.1007/978-3-540-77004-6_11
- 519 Zhu, M., & Ghodsi, A. (2006). Automatic dimensionality selection from the scree plot via the use of profile
520 likelihood. *Computational Statistics & Data Analysis*, 51(2), 918–930.
521 <https://doi.org/10.1016/j.csda.2005.09.010>

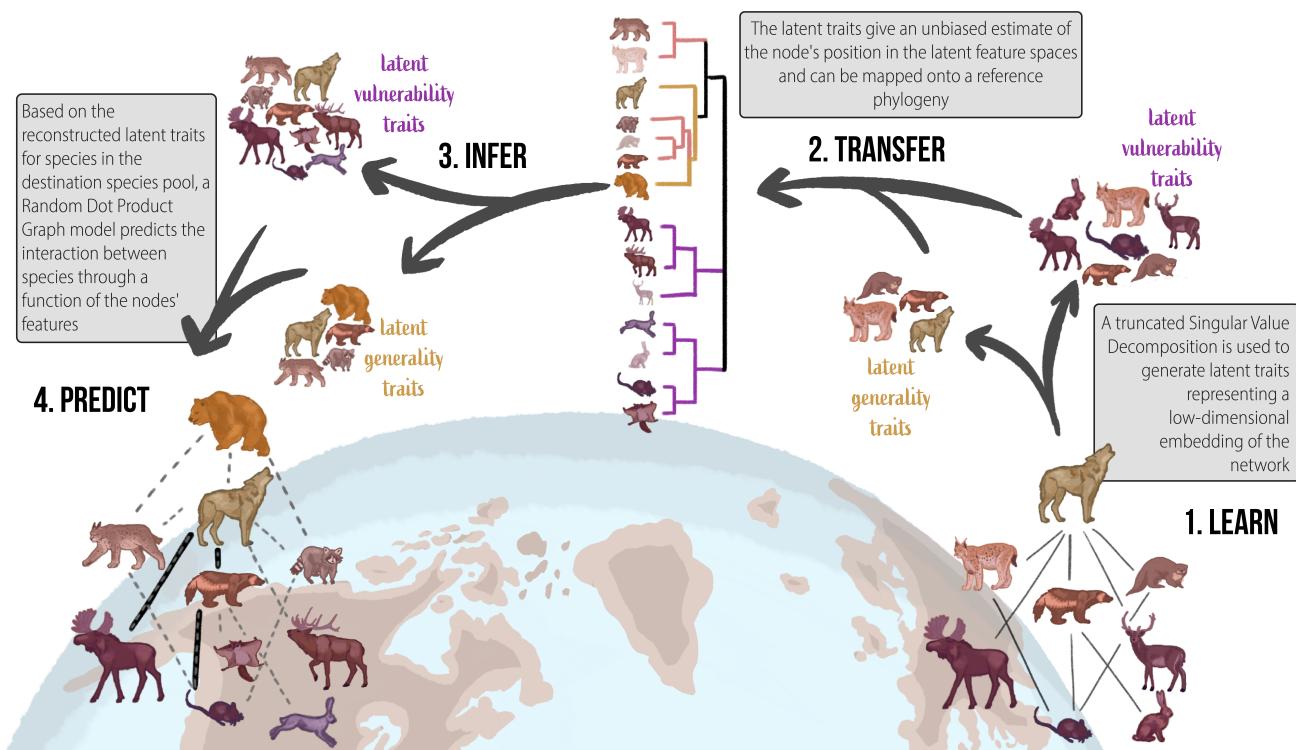


Figure 1: Overview of the phylogenetic transfer learning (and prediction) of species interaction networks. Starting from an initial, known, network, we learn its representation through a graph embedding step (here, a truncated Singular Value Decomposition; Step 1), yielding a series of latent traits (latent vulnerability traits are more representative of species at the lower trophic-level and latent generality traits are more representative of species at higher trophic-levels; *sensu* Schoener (1989)); second, for the destination species pool, we perform ancestral character estimation using a phylogeny (here, using a Brownian model for the latent traits; Step 2); we then sample from the reconstructed distribution of latent traits (Step 3) to generate a probabilistic metaweb at the destination (here, assuming a uniform distribution of traits), and threshold it to yield the final list of interactions (Step 4).

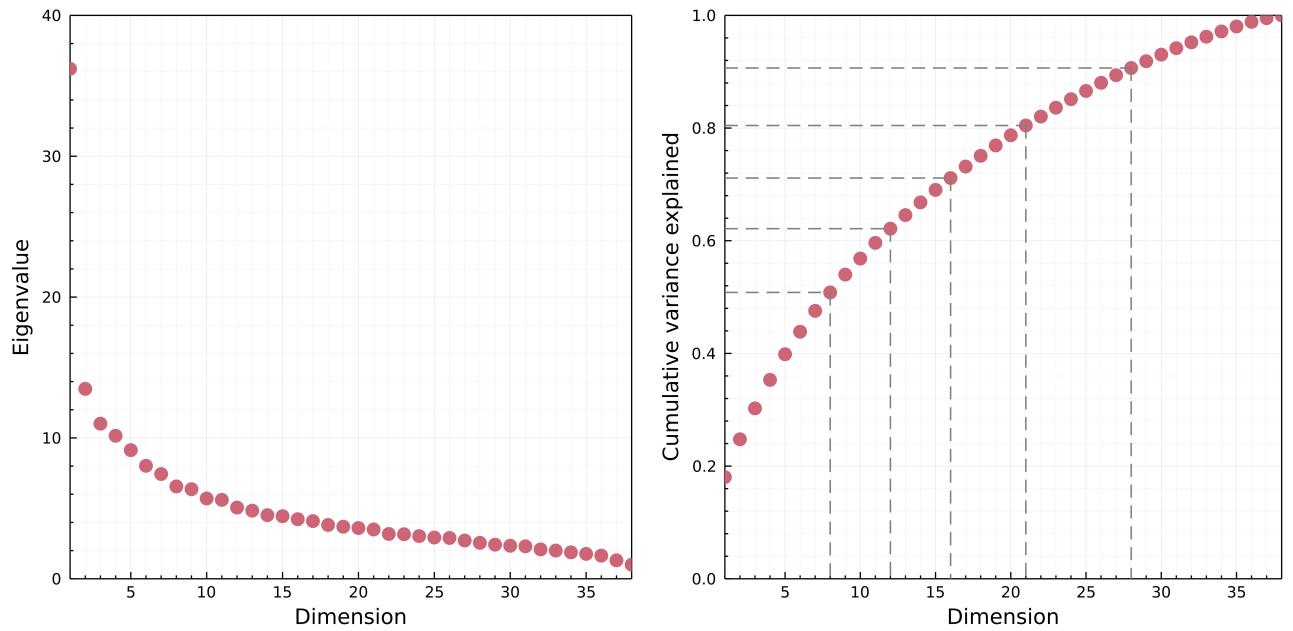


Figure 2: Left: representation of the scree plot of the singular values from the t-SVD on the European metaweb. The scree plot shows no obvious drop in the singular values that may be leveraged to automatically detect a minimal dimension for embedding, after e.g. Zhu & Ghodsi (2006). Right: cumulative fraction of variance explained by each dimension up to the rank of the European metaweb. The grey lines represent cutoffs at 50, 60, ..., 90% of variance explained. For the rest of the analysis, we reverted to an arbitrary threshold of 60% of variance explained, which represented a good tradeoff between accuracy and reduced number of features.

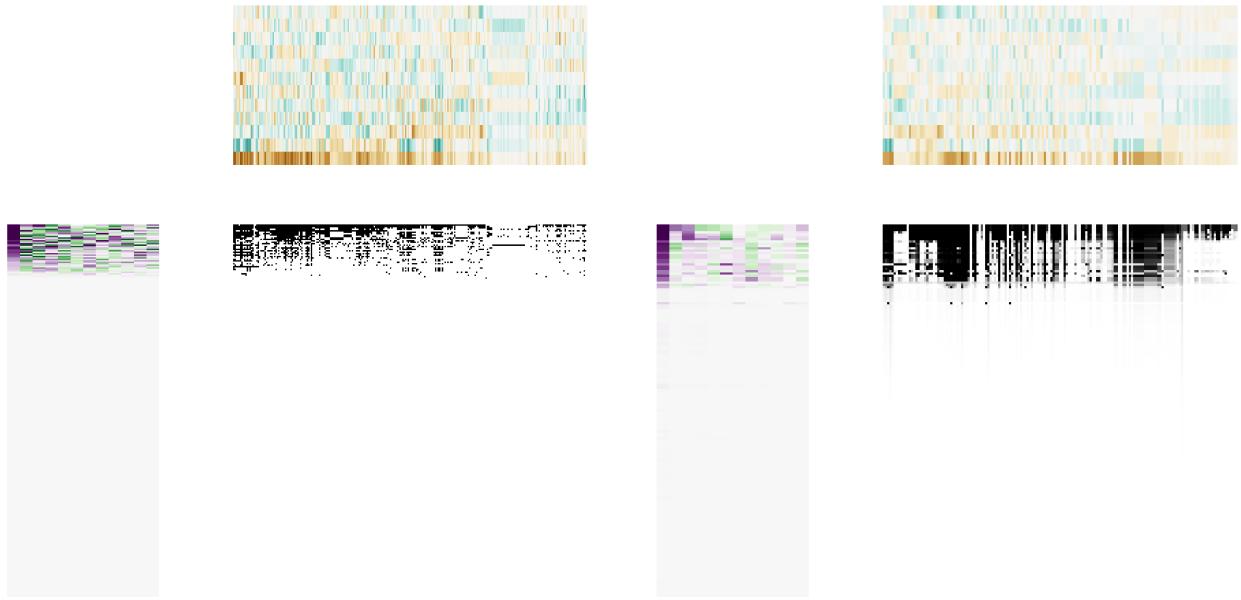


Figure 3: Visual representation of the left (green/purple) and right (green/brown) subspaces, alongside the adjacency matrix of the food web they encode (greyscale). The European metaweb is on the left, and the imputed Canadian metaweb (before data inflation) on the right. This figure illustrates how much structure the left subspace captures. As we show in fig. 6, the species with a value of 0 in the left subspace are species without any prey.

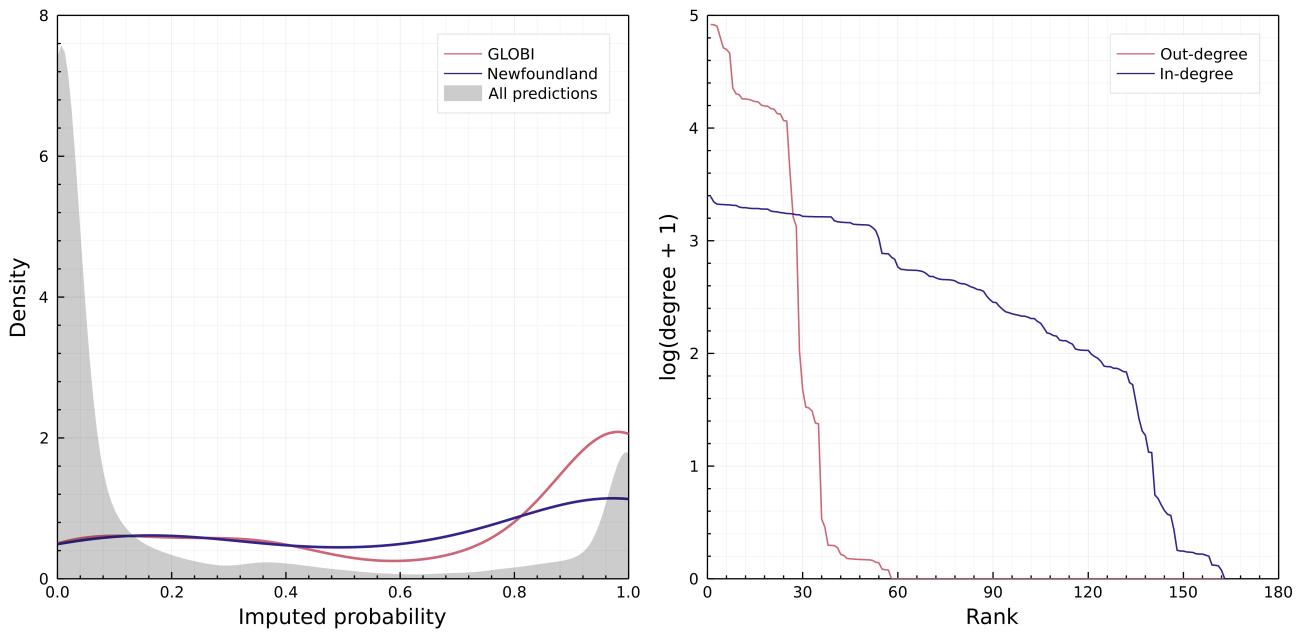


Figure 4: Left, comparison of the probabilities of interactions assigned by the model to all interactions (grey curve), the subset of interactions found in GLOBI (red), and in the Strong & Leroux (2014) Newfoundland dataset (blue). The model recovers more interactions with a low probability compared to data mining, which can suggest that collected datasets are biased towards more common or easy to identify interactions. Right, distribution of the in-degree and out-degree of the mammals from Canada in the reconstructed metaweb. This figure describes a flat, relatively short food web, in which there are few predators but a large number of preys.

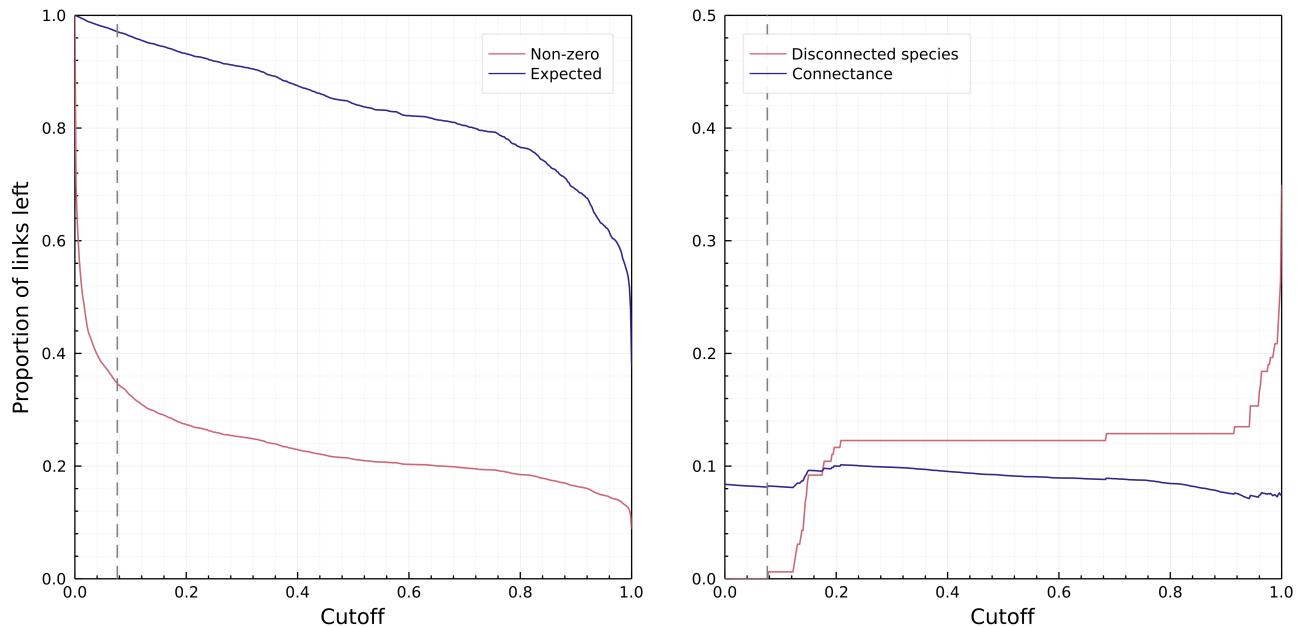


Figure 5: Left: effect of varying the cutoff for probabilities to be considered non-zero on the number of unique links and on \hat{L} , the probabilistic estimate of the number of links assuming that all interactions are independent. Right: effect of varying the cutoff on the number of disconnected species, and on network connectance. In both panels, the grey line indicates the cutoff $P(i \rightarrow j) \approx 0.08$ that resulted in the first species losing all of its interactions.

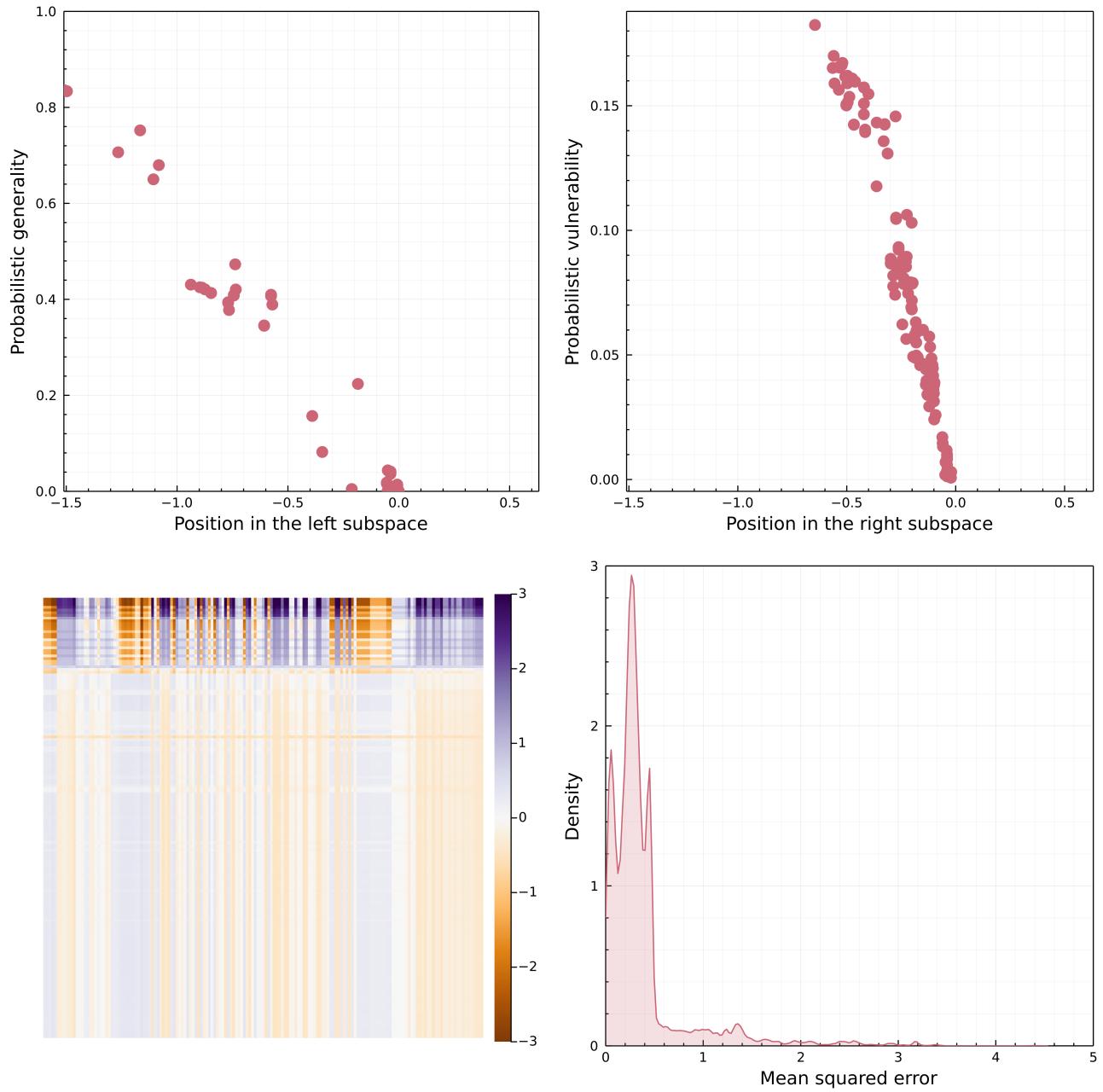


Figure 6: Top: biological significance of the first dimension. Left: there is a linear relationship between the values on the first dimension of the left subspace and the generality, *i.e.* the relative number of preys, *sensu* Schoener (1989). Species with a value of 0 in this subspace are at the bottom-most trophic level. Right: there is, similarly, a linear relationship between the position of a species on the first dimension of the right subspace and its vulnerability, *i.e.* the relative number of predators. Taken together, these two figures show that the first-order representation of this network would capture its degree distribution. Bottom: topological consequences of the first dimension. Left: differences in the z-score of the actual configuration model for the reconstructed network, and the prediction based only on the first dimension. Right: distribution of the differences in the left panel.