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

Tanya Strydom [1,2,‡] Salomé Bouskila [1,‡] Francis Banville [1,3,2] Ceres Barros [4] Dominique Caron [5,2] Maxwell J Farrell [6] Marie-Josée Fortin [6] Victoria Hemming [7] Benjamin Mercier [3,2] Laura J. Pollock [5,2] Rogini Runghen [8] Giulio V. Dalla Riva [9] 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

Last revision: *January 12, 2022*

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.

Introduction

There are two core challenges we are faced with in furthering our understanding of ecological networks across space, particularly at macro-ecologically relevant scales (*e.g.* Trøjelsgaard & Olesen, 2016). First, ecological networks within a location are difficult to sample properly (Jordano, 2016a, 2016b), resulting in a widespread "Eltonian shortfall" (Hortal et al., 2015), *i.e.* a lack of knowledge about inter- and intra-specific relationships. This first challenge has been, in large part, addressed by the recent emergence of a suite of methods aiming to predict interactions within *existing* networks, many of which are reviewed in Strydom, Catchen, et al. (2021). Second, recent analyses based on collected data (Poisot, Bergeron, et al., 2021) or metadata (Cameron et al., 2019) highlight that ecological networks are currently studied in a biased subset of space and bioclimates, which impedes our ability to generalize any local understanding of network structure. Meaning that, although the framework to address incompleteness *within* networks exists, there would still be regions for which, due to a *lack* of local interaction data, we are unable to infer potential species interactions. Having a general solution for inferring *potential* interactions (despite the unavailability of interaction data) could be the catalyst for significant breakthroughs in our ability to start thinking about species interaction networks over large spatial scales. In a recent overview of the field of ecological network prediction, Strydom, Catchen, et al. (2021) identified two challenges of interest to the prediction of interactions at large scales. First, there is a relative scarcity of relevant data in most places globally – paradoxically, this restricts our ability to infer interactions to locations where inference is perhaps the least required; second, accurate predictions often demand accurate predictors, and the lack of methods that can leverage small amount of data is a serious impediment to our predictive ability globally.

Here, we present a general method to infer potential trophic interactions, relying on the transfer learning of network representations, specifically by using similarities of species in a biologically/ecologically relevant proxy space (*e.g.* shared morphology or ancestry). Transfer learning is a machine learning methodology that uses the knowledge gained from solving one problem and applying it to a related (destination) problem (Pan & Yang, 2010; Torrey & Shavlik, 2010). In this instance, we solve the problem of predicting trophic interactions between species, based on knowledge extracted from another species pool for which interactions are known by using phylogenetic structure as a medium for transfer. There is a plurality of measures of species similarities that can be used for inferring *potential* species interactions *i.e.* metaweb reconstruction (see *e.g.* Morales-Castilla et al., 2015); however, phylogenetic proximity has

several desirable properties when working at large scales. Gerhold et al. (2015) made the point that phylogenetic signal captures diversification of characters (large macro-evolutionary process), but not necessarily community assembly (fine ecological process); Dormann et al. (2010) previously found very similar conclusions. Interactions tend to reflect a phylogenetic signal because they have a conserved pattern of evolutionary convergence that encompasses a wide range of ecological and evolutionary mechanisms (Cavender-Bares et al., 2009; Mouquet et al., 2012), and - most importantly - retain this signal even if it is obscured at the community scale due to *e.g.* local conditions (Hutchinson et al., 2017; Poisot & Stouffer, 2018). Finally, species interactions at macro-ecological scales seem to respond mostly to macro-evolutionary processes (Price, 2003); which is evidenced by the presence of conserved backbones in food webs (Dalla Riva & Stouffer, 2016; Mora et al., 2018), strong evolutionary signature on prey choice (Stouffer et al., 2012), and strong phylogenetic signature in food web intervality (Eklöf & Stouffer, 2016). Phylogenetic reconstruction has also previously been used within the context of ecological networks, namely understanding ancestral plant-insect interactions (Braga et al., 2021). Taken together, these considerations suggest that phylogenies can reliably be used to transfer knowledge on species interactions.

[Figure 1 about here.]

In fig. 1, where we provide a methodological overview based on learning the embedding of a metaweb of trophic interactions for European mammals (known interactions; Maiorano et al., 2020a, 2020b) and, based on phylogenetic relationships between mammals globally (*i.e.*, phylogenetic tree Upham et al., 2019), infer a metaweb for the Canadian mammalian species pool (using only a species list *i.e.* interactions are 'unknown' in this instance). Following the definition of Dunne (2006), a metaweb is a network analogue to the regional species pool; specifically, it is an inventory of all *potential* interactions between species from a spatially delimited area (and so captures the γ diversity of interactions). The metaweb is, therefore, *not* a prediction of the food web at a specific locale within the spatial area it covers, and will have a different structure (notably by having a larger connectance; see *e.g.* Wood et al., 2015). These local food webs (which captures the α diversity of interactions) are a subset of the metaweb's species and interactions, and have been called "metaweb realizations" (Poisot et al., 2015). Differences between local food web and their metaweb are due to chance, species abundance and co-occurrence, local environmental conditions, and local distribution of functional traits, among others.

Because the metaweb represents the joint effect of functional, phylogenetic, and macroecological

processes (Morales-Castilla et al., 2015), it holds valuable ecological information. Specifically, it is the "upper bounds" on what the composition of the local networks can be (see *e.g.* McLeod et al., 2021). These local networks, in turn, can be reconstructed given appropriate knowledge of local species composition, providing information on structure of food webs at finer spatial scales. This has been done for example for tree-galler-parasitoid systems (Gravel et al., 2018), fish trophic interactions (Albouy et al., 2019), tetrapod trophic interactions (O'Connor et al., 2020), and crop-pest networks (Grünig et al., 2020). Whereas the original metaweb definition, and indeed most past uses of metawebs, was based on the presence/absence of interactions, we focus on *probabilistic* metawebs where interactions are represented as the chance of success of a Bernoulli trial (see *e.g.* Poisot et al., 2016); therefore, not only does our method recommend interactions that may exist, it gives each interaction a score, allowing us to properly weigh them.

Our case study shows that phylogenetic transfer learning is an effective approach to the generation of probabilistic metawebs. This showcases that although the components (species) that make up the Canadian and European communities may be *minimally* shared (the overall species overlap is less than 4%), if the medium (proxy space) selected in the transfer step is biologically plausible, we can still effectively learn from the known network and make biologically relevant predictions of interactions. Indeed, as we detail in the results, when validated against known but fractional data of trophic interactions between Canadian mammals, our model achieves a predictive accuracy of approximately 91%. It should be reiterated that the framework presented in fig. 1 is amenable to changes *e.g.* it is possible to use distinct trees if working with distinct clades (such as pollination networks) or, alternatively, the measure of similarity may not be phylogeny, and can be replaced by information on foraging (Beckerman et al., 2006), cell-level mechanisms (Boeckaerts et al., 2021), or a combination of traits and phylogenetic structure (Stock, 2021). Most importantly, although we focus on a trophic system, it is an established fact that different (non-trophic) interactions do themselves interact with and influence the outcome of trophic interactions (see *e.g.* Kawatsu et al., 2021; Kéfi et al., 2012). Future development of metaweb inference techniques should cover the prediction of multiple interaction types.

Data used for the case study

We use data from the European metaweb assembled by Maiorano et al. (2020a). This was assembled using data extracted from scientific literature (including published papers, books, and grey literature) from the

last 50 years and includes all terrestrial tetrapods (mammals, breeding birds, reptiles and amphibians) occurring on the European sub-continent (and Turkey) - with the caveat that only species introduced in historical times and currently naturalized being included. The European metaweb was filtered using the Global Biodiversity Information Facility (GBIF) taxonomic backbone (GBIF Secretariat, 2021) so as to contain only terrestrial and semi-aquatic mammals. As all species had valid matches to the GBIF taxonomy it was used as the backbone for the remaining reconciliation steps namely, the mammalian consensus supertree by Upham et al. (2019) (which is used for the knowledge transfer step) and for the Canadian species list—which was extracted from the International Union for Conservation of Nature (IUCN) checklist, and corresponds to the same selection criteria that was applied by Maiorano et al. (2020a) in the European metaweb.

After taxonomic cleaning and reconciliation the mammalian European metaweb has 260 species, and the Canadian species pool 163; of these, 17 (about 4% of the total) are shared, and 89 species from Canada (54%) had at least one congeneric species in Europe. The similarity for both species pools predictably increases with higher taxonomic order, with 19% of shared genera, 47% of shared families, and 75% of shared orders; for the last point, Canada and Europe each had a single unique order (*Didelphimorphia* for Canada, *Erinaceomorpha* for Europe).

Method description

The core point of our method is the transfer of knowledge of a known ecological network, in order to predict interactions between species from another location for which the network is unknown (or partially known). In fig. 1 the grey text boxes give a high-level overview of the approach; in the example around which this manuscript is built The method we develop is, ecologically speaking, a "black box," *i.e.* an algorithm that can be understood mathematically, but whose component parts are not always directly tied to ecological processes. There is a growing realization in machine learning that (unintentional) black box algorithms are not necessarily a bad thing (Holm, 2019), as long as their constituent parts can be examined (which is the case with our method). But more importantly, data hold more information than we might think; as such, even algorithms that are disconnected from a model can make correct guesses most of the time (Halevy et al., 2009); in fact, in an instance of ecological forecasting of spatio-temporal systems, model-free approaches (*i.e.* drawing all of their information from the data) outperformed

model-informed ones (Perretti et al., 2013).

Implementation and code availability

The entire pipeline is implemented in *Julia* 1.6 (Bezanson et al., 2017) and is available under the permissive MIT License at `https://osf.io/2zwqm/`. The taxonomic cleanup steps are done using `GBIF.jl` (Dansereau & Poisot, 2021). The network embedding and analysis is done using `EcologicalNetworks.jl` (Banville et al., 2021; Poisot et al., 2019). The phylogenetic simulations are done using `PhyloNetworks.jl` (Solís-Lemus et al., 2017) and `Phylo.jl` (Reeve et al., 2016). A complete `Project.toml` file specifying the full tree of dependencies is available alongside the code. This material also includes a fully annotated copy of the entire code required to run this project (describing both the intent of the code and discussing some technical implementation details), a vignette for every step of the process, and a series of Jupyter notebooks with the text and code. The pipeline can be executed on a laptop in a matter of minutes, and therefore does not require extensive computational power.

Step 1: Learning the origin network representation

The first step in transfer learning is to learn the structure of the original dataset. In order to do so, we rely on an approach inspired from representational learning, where we learn a *representation* of the metaweb (in the form of the latent subspaces), rather than a list of interactions (species a eats b). This approach is conceptually different from other metaweb-scale predictions (*e.g.* Albouy et al., 2019), in that the metaweb representation is easily transferable. Specifically, we use Random Dot Product Graph model (hereafter RDPG; S. J. Young & Scheinerman, 2007) to create a number of latent variables that can be combined into an approximation of the network adjacency matrix. RDPG is known to capture the evolutionary backbone of food webs (Dalla Riva & Stouffer, 2016), resulting in strong phylogenetic signal in RDPG results; in other words, the latent variables of an RDPG can be mapped onto a phylogenetic tree, and phylogenetically similar predators should share phylogenetically similar preys. In addition, recent advances show that the latent variables produced this way can be used to predict *de novo* interactions. Interestingly, the latent variables do not need to be produced by decomposing the network itself; in a recent contribution, Runghen et al. (2021) showed that deep artificial neural networks are able to reconstruct the left and right subspaces of an RDPG, in order to predict human movement networks from

individual/location metadata and opens up the possibility of using additional metadata as predictors.

The latent variables are created by performing a truncated Singular Value Decomposition (t-SVD; Halko et al., 2011) on the adjacency matrix. SVD is an appropriate embedding of ecological networks, which has recently been shown to both capture their complex, emerging properties (Strydom, Dalla Riva, et al., 2021) and to allow highly accurate prediction of the interactions within a single network (Poisot, Ouellet, et al., 2021). Under SVD, an adjacency matrix $\mathbf{A}$ (where $\mathbf{A}_{m,n} \in \mathbb{B}$ where 1 indicates predation and 0 an absence thereof) is decomposed into three components resulting in $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'$. Here, $\mathbf{\Sigma}$ is a $m \times n$ diagonal matrix and contains only singular (σ) values along its diagonal, $\mathbf{U}$ is a $m \times m$ unitary matrix, and $\mathbf{V}'$ a $n \times n$ unitary matrix. Truncating the SVD removes additional noise in the dataset by omitting non-zero and/or smaller σ values from $\mathbf{\Sigma}$ using the rank of the matrix. Under a t-SVD $\mathbf{A}_{m,n}$ is decomposed so that $\mathbf{\Sigma}$ is a 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 truncate the matrix) containing only non-zero σ values. Additionally, $\mathbf{U}$ is now a $m \times r$ semi unitary matrix and $\mathbf{V}'$ a $n \times r$ semi-unitary matrix.

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

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

[Figure 2 about here.]

Because RDPG relies on matrix multiplication, the higher dimensions essentially serve to make specific interactions converge towards 0 or 1; therefore, for reasonably low ranks, there is no guarantee that the values in the reconstructed network will be within the unit range. In order to determine what constitutes an appropriate threshold for probability, we performed the RDPG approach on the European metaweb, and evaluated the probability threshold by treating this as a binary classification problem, specifically assuming that both 0 and 1 in the European metaweb are all true. Given the methodological details given in Maiorano et al. (2020a) and O'Connor et al. (2020), this seems like a reasonable assumption, although one that does not hold for all metawebs. We used the thresholding approach presented in Poisot, Ouellet, et al. (2021), and picked a cutoff that maximized Youden's J statistic (a measure of the informedness (trust) of predictions; Youden (1950)); the resulting cutoff was 0.22, and gave an accuracy above 0.99. In Supp. Mat. 1, we provide several lines of evidence that using the entire network to estimate the threshold does not lead to overfitting; that using a subset of species would yield the same threshold; that decreasing the quality of the original data by adding or removing interactions would minimally affect the predictive accuracy of RDPG applied to the European metaweb; and that the networks reconstructed from artificially modified data are reconstructed with the correct ecological properties.

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

Steps 2 and 3: Transfer learning through phylogenetic relatedness

In order to transfer the knowledge from the European metaweb to the Canadian species pool, we performed ancestral character estimation using a Brownian motion model, which is a conservative approach in the absence of strong hypotheses about the nature of phylogenetic signal in the network decomposition (Litsios & Salamin, 2012). This uses the estimated feature vectors for the European mammals to create a state reconstruction for all species (conceptually something akin to a trait-based mammalian phylogeny using latent generality and vulnerability traits) and allows us to impute the missing (latent) trait data for the Canadian species that are not already in the European network; as we are focused on predicting contemporary interactions, we only retained the values for the tips of the tree. We

assumed that all traits (*i.e.* the feature vectors for the left and right subspaces) were independent, which is a reasonable assumption as every trait/dimension added to the t-SVD has an *additive* effect to the one before it. Note that the Upham et al. (2019) tree itself has some uncertainty associated to inner nodes of the phylogeny. In this case study, we have decided to not propagate this uncertainty, as it would complexify the process. The Brownian motion algorithm returns the *average* value of the trait, and its upper and lower bounds. Because we do not estimate other parameters of the traits' distributions, we considered that every species trait is represented as a uniform distribution between these bounds. The choice of the uniform distribution was made because the algorithm returns a minimum and maximum point estimate for the value, and given this information, the uniform distribution is the one with maximum entropy. Had all mean parameters estimates been positive, the exponential distribution would have been an alternative, but this is not the case for the subspaces of an RDPG. In order to examine the consequences of the choice of distribution, we estimated the variance per latent variable per node to use a Normal distribution; as we show in Supp. Mat. 2, this decision results in dramatically over-estimating the number and probability of interactions, and therefore we keep the discussions in the main text to the uniform case. The inferred left and right subspaces for the Canadian species pool ($\hat{\mathscr{L}}$ and $\hat{\mathscr{R}}$) have entries that are distributions, representing the range of values for a given species at a given dimension.

These objects represent the transferred knowledge, which we can use for prediction of the Canadian metaweb.

Step 4: Probabilistic prediction of the destination network

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

[Figure 3 about here.]

Specifically, we have adopted the following approach. For every entry in $\hat{\mathscr{L}}$ and $\hat{\mathscr{R}}$, we draw a value from its distribution. This results in one instance of the possible left ($\hat{\ell}$) and right ($\hat{r}$) subspaces for the Canadian metaweb. These can be multiplied, to produce one matrix of real values. Because the entries in

$\hat{\ell}$ and $\hat{r}$ are in the same space where $\mathscr{L}$ and $\mathscr{R}$ were originally predicted, it follows that the threshold ρ estimated for the European metaweb also applies. We use this information to produce one random Canadian metaweb, $N = \hat{\mathscr{L}}\hat{\mathscr{R}}' \geq \rho$. As we can see in (fig. 3), the European and Canadian metawebs are structurally similar (as would be expected given the biogeographic similarities) and the two (left and right) subspaces are distinct *i.e.* capturing predation (generality) and prey (vulnerability) latent traits.

Because the intervals around some trait values can be broad (in fact, probably broader than what they would actually be, see *e.g.* Garland et al., 1999), we repeat the above process 2×10^5 times, which results in a probabilistic metaweb P, where the probability of an interaction (here conveying our degree of trust that it exists given the inferred trait distributions) is given by the number of times where it appears across all random draws N, divided by the number of samples. An interaction with $P_{i,j} = 1$ means that these two species were predicted to interact in all 2×10^5 random draws.

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

Data cleanup, discovery, validation, and thresholding

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

[Figure 4 about here.]

First, we extracted the subgraph corresponding to the 17 species shared between the European and Canadian pools and replaced these interactions with a probability of 0 (non-interaction) or 1 (interaction), according to their value in the European metaweb. This represents a minute modification of the inferred

network (about 0.8% of all species pairs from the Canadian web), but ensures that we are directly re-using knowledge from Europe.

Second, we looked for all species in the Canadian pool known to the Global Biotic Interactions (GLoBI) database (Poelen et al., 2014), and extracted their known interactions. Because GLoBI aggregates observed interactions, it is not a *networks* data source, and therefore the only information we can reliably extract from it is that a species pair *was reported to interact at least once*. This last statement should yet be taken with caution, as some sources in GLoBI (*e.g.* Thessen & Parr, 2014) are produced through text analysis, and therefore may not document direct evidence of the interaction. Nevertheless, should the predictive model work, we would expect that a majority of interactions known to GLoBI would also be predicted. We retrieved 366 interactions between mammals from the Canadian species pool from GLoBI, 33 of which were not predicted by the model; this results in a success rate of 91%. After performing this check, we set the probability of all interactions known to GLoBI to 1.

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

[Figure 5 about here.]

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

Results and discussion of the case study

[Figure 6 about here.]

The t-SVD embedding is able to learn relevant ecological features for the network. fig. 6 shows that the first rank correlates linearly with generality and vulnerability (Schoener, 1989), *i.e.* the number of preys and predators for each species. Importantly, this implies that a rank 1 approximation represents the configuration model for the metaweb, *i.e.* a set of random networks generated from a given degree sequence (Park & Newman, 2004). Accounting for the probabilistic nature of the degrees, the rank 1 approximation also represents the *soft* configuration model (van der Hoorn et al., 2018). Both models are maximum entropy graph models (Garlaschelli et al., 2018), with sharp (all network realizations satisfy the specified degree sequence) and soft (network realizations satisfy the degree sequence on average) local constraints, respectively. The (soft) configuration model is an unbiased random graph model widely used by ecologists in the context of null hypothesis significance testing of network structure (*e.g.* Bascompte et al., 2003) and can provide informative priors for Bayesian inference of network structure (*e.g.* J.-G. Young et al., 2021). It is noteworthy that for this metaweb, the relevant information was extracted at the first rank. Because the first rank corresponds to the leading singular value of the system, the results of fig. 6 have a straightforward interpretation: degree-based processes are the most important in structuring the mammalian food web.

Discussion

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

documented instances of food webs undergoing rapid collapse/recovery cycles over short periods of time (Pedersen et al., 2017). The embedding of a network, in a sense, embeds its macro-evolutionary history, especially as RDPG captures ecological signal (Dalla Riva & Stouffer, 2016); at this point, it is important to recall that a metaweb is intended as a catalogue of all potential interactions, which should then be filtered (Morales-Castilla et al., 2015). In practice (and in this instance) the reconstructed metaweb will predict interactions that are plausible based on the species' evolutionary history, however some interactions would/would not be realized due to human impact.

Dallas et al. (2017) suggested that most links in ecological networks may be cryptic, *i.e.* uncommon or otherwise hard to observe. This argument essentially echoes Jordano (2016b): the sampling of ecological interactions is difficult because it requires first the joint observation of two species, and then the observation of their interaction. In addition, it is generally expected that weak or rare links would be more common in networks (Csermely, 2004), compared to strong, persistent links; this is notably the case in food chains, wherein many weaker links are key to the stability of a system (Neutel et al., 2002). In the light of these observations, the results in fig. 4 are not particularly surprising: we expect to see a surge in these low-probability interactions under a model that has a good predictive accuracy. Because the predictions we generate are by design probabilistic, then one can weigh these rare links appropriately. In a sense, that most ecological interactions are elusive can call for a slightly different approach to sampling: once the common interactions are documented, the effort required in documenting each rare interaction may increase exponentially. Recent proposals suggest that machine learning algorithms, in these situations, can act as data generators (Hoffmann et al., 2019): in this perspective, high quality observational data can be supplemented with synthetic data coming from predictive models, which increases the volume of information available for inference. Indeed, Strydom, Catchen, et al. (2021) suggested that knowing the metaweb may render the prediction of local networks easier, because it fixes an "upper bound" on which interactions can exist; indeed, with a probabilistic metaweb, we can consider that the metaweb represents an aggregation of informative priors on the interactions.

Related to the last point, Cirtwill et al. (2019) showed that network inference techniques based on Bayesian approaches would perform far better in the presence of an interaction-level informative prior; the desirable properties of such a prior would be that it is expressed as a probability, preferably representing a Bernoulli event, the value of which would be representative of relevant biological processes (probability of predation in this case). We argue that the probability returned at the very last step of our

framework may serve as this informative prior; indeed, the output of our analysis can be used in subsequent steps, also possibly involving expert elicitation to validate some of the most strongly recommended interactions. One important *caveat* to keep in mind when working with interaction inference is that interactions can never really be true negatives (in the current state of our methodological framework and data collection limitations); this renders the task of validating a model through the usual application of binary classification statistics very difficult (although see Strydom, Catchen, et al., 2021 for a discussion of alternative suggestions). The other way through which our framework can be improved is by substituting the predictors that are used for transfer. For example, in the presence of information on species traits that are known to be predictive of species interactions, one might want to rely on functional rather than phylogenetic distances – in food webs, body size (and allometrically related variables) has been established as such a variable (Brose et al., 2006); the identification of relevant functional traits is facilitated by recent methodological developments (Rosado et al., 2013). It should be noted that Xing & Fayle (2021) highlight phylogenetic relatedness as one of the core components of network comparison at the global scale. In this case study, we have embedded the original metaweb using t-SVD, because it lends itself to an RDPG reconstruction, which is known to capture the consequences of evolutionary processes (Dalla Riva & Stouffer, 2016); this being said, there are other ways to embed graphs (Arsov & Mirceva, 2019; Cai et al., 2017; Cao et al., 2019), which can be used as alternatives.

As Herbert (1965) rightfully pointed out, "[y]ou can't draw neat lines around planet-wide problems"; in this regard, our approach (and indeed, any inference of a metaweb at large scales) must contend with several interesting and interwoven families of problems. The first is the limit of the metaweb to embed and transfer. If the initial metaweb is too narrow in scope, notably from a taxonomic point of view, the chances of finding another area with enough related species to make a reliable inference decreases; this would likely be indicated by large confidence intervals during ancestral character estimation, but the lack of well documented metawebs is currently preventing the development of more concrete guidelines. The question of phylogenetic relatedness and dispersal is notably true if the metaweb is assembled in an area with mostly endemic species, and as with every predictive algorithm, there is room for the application of our best ecological judgement. Conversely, the metaweb should be reliably filled, which assumes that the S^2 interactions in a pool of S species have been examined, either through literature surveys or expert elicitation. Supp. Mat. 1 provides some guidance as to the type of sampling effort that should be prioritized. While RDPG was able to maintain very high predictive power when interactions were missing,

the addition of false positive interactions was immediately detected; this suggests that it may be appropriate to err on the side of "too many" interactions when constructing the initial metaweb to be transferred. The second series of problems are related to determining which area should be used to infer the new metaweb in, as this determines the species pool that must be used. In our application, we focused on the mammals of Canada. The upside of this approach is that information at the country level is likely to be required by policy makers and stakeholders for their biodiversity assessment, as each country tends to set goals at the national level (Buxton et al., 2021) for which quantitative instruments are designed (Turak et al., 2017), with specific strategies often enacted at smaller scales (Ray et al., 2021). And yet, we do not really have a satisfying answer to the question of "where does a food web stop?"; the current most satisfying solutions involve examining the spatial consistency of network area relationships (Fortin et al., 2021; see *e.g.* Galiana et al., 2018, 2019, 2021), which is of course impossible in the absence of enough information about the network itself. This suggests that an *a posteriori* refinement of the results may be required, based on a downscaling of the metaweb. The final family of problems relates less to the availability of data or quantitative tools, and more to the praxis of spatial ecology. Operating under the context of national divisions, in large parts of the world, reflects nothing more than the legacy of settler colonialism. Indeed, the use of ecological data is not an apolitical act (Nost & Goldstein, 2021), as data infrastructures tend to be designed to answer questions within national boundaries, and their use both draws upon and reinforces territorial statecraft; as per Machen & Nost (2021), this is particularly true when the output of "algorithmic thinking" (*e.g.* relying on machine learning to generate knowledge) can be re-used for governance (*e.g.* enacting conservation decisions at the national scale). We therefore recognize that methods such as we propose operate under the framework that contributed to the ongoing biodiversity crisis (Adam, 2014), reinforced environmental injustice (Choudry, 2013; Domínguez & Luoma, 2020), and on Turtle Island especially, should be replaced by Indigenous principles of land management (Eichhorn et al., 2019; No'kmaq et al., 2021). As we see AI/ML being increasingly mobilized to generate knowledge that is lacking for conservation decisions (*e.g.* Lamba et al., 2019; Mosebo Fernandes et al., 2020), our discussion of these tools need to go beyond the technical, and into the governance consequences they can have.

Acknowledgements: We acknowledge that this study was conducted on land within the traditional unceded territory of the Saint Lawrence Iroquoian, Anishinabewaki, Mohawk, Huron-Wendat, and Omàmiwininiwak nations. TP, TS, DC, and LP received funding from the Canadian Institue for Ecology &

Evolution. FB is funded by the Institute for Data Valorization (IVADO). TS, SB, and TP are funded by a donation from the Courtois Foundation. CB was awarded a Mitacs Elevate Fellowship no. IT12391, in partnership with fRI Research, and also acknowledges funding from Alberta Innovates and the Forest Resources Improvement Association of Alberta. M-JF acknowledges funding from NSERC Discovery Grant and NSERC CRC. RR is funded by New Zealand's Biological Heritage Ngā Koiora Tuku Iho National Science Challenge, administered by New Zealand Ministry of Business, Innovation, and Employment. BM is funded by the NSERC Alexander Graham Bell Canada Graduate Scholarship and the FRQNT master's scholarship. LP acknowledges funding from NSERC Discovery Grant (NSERC RGPIN-2019-05771). TP acknowledges financial support from NSERC through the Discovery Grants and Discovery Accelerator Supplement programs.

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

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

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

References

Adam, R. (2014). *Elephant treaties: The Colonial legacy of the biodiversity crisis*. UPNE.

Albouy, C., Archambault, P., Appeltans, W., Araújo, M. B., Beauchesne, D., Cazelles, K., Cirtwill, A. R., Fortin, M.-J., Galiana, N., Leroux, S. J., Pellissier, L., Poisot, T., Stouffer, D. B., Wood, S. A., & Gravel, D. (2019). The marine fish food web is globally connected. *Nature Ecology & Evolution*, *3*(8, 8), 1153–1161. https://doi.org/10.1038/s41559-019-0950-y

Arsov, N., & Mirceva, G. (2019, November 26). *Network Embedding: An Overview*. http://arxiv.org/abs/1911.11726

Banville, F., Vissault, S., & Poisot, T. (2021). Mangal.jl and EcologicalNetworks.jl: Two complementary packages for analyzing ecological networks in Julia. *Journal of Open Source Software*, *6*(61), 2721. https://doi.org/10.21105/joss.02721

Bascompte, J., Jordano, P., Melian, C. J., & Olesen, J. M. (2003). The nested assembly of plant-animal mutualistic networks. *Proceedings of the National Academy of Sciences*, *100*(16), 9383–9387. https://doi.org/10.1073/pnas.1633576100

Beckerman, A. P., Petchey, O. L., & Warren, P. H. (2006). Foraging biology predicts food web complexity. *Proceedings of the National Academy of Sciences*, *103*(37), 13745–13749. https://doi.org/10.1073/pnas.0603039103

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, *59*(1), 65–98. https://doi.org/10.1137/141000671

Boeckaerts, D., Stock, M., Criel, B., Gerstmans, H., De Baets, B., & Briers, Y. (2021). Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins. *Scientific Reports*, *11*(1, 1), 1467. https://doi.org/10.1038/s41598-021-81063-4

Braga, M. P., Janz, N., Nylin, S., Ronquist, F., & Landis, M. J. (2021). Phylogenetic reconstruction of ancestral ecological networks through time for pierid butterflies and their host plants. *Ecology Letters*, *n/a*(n/a). https://doi.org/10.1111/ele.13842

Brose, U., Jonsson, T., Berlow, E. L., Warren, P., Banasek-Richter, C., Bersier, L.-F., Blanchard, J. L., Brey, T., Carpenter, S. R., Blandenier, M.-F. C., Cushing, L., Dawah, H. A., Dell, T., Edwards, F., Harper-Smith, S., Jacob, U., Ledger, M. E., Martinez, N. D., Memmott, J., … Cohen, J. E. (2006). ConsumerResource Body-Size Relationships in Natural Food Webs. *Ecology*, *87*(10), 2411–2417. https://doi.org/10.1890/0012-9658(2006)87%5B2411:CBRINF%5D2.0.CO;2

Buxton, R. T., Bennett, J. R., Reid, A. J., Shulman, C., Cooke, S. J., Francis, C. M., Nyboer, E. A., Pritchard, G., Binley, A. D., Avery-Gomm, S., Ban, N. C., Beazley, K. F., Bennett, E., Blight, L. K., Bortolotti, L. E., Camfield, A. F., Gadallah, F., Jacob, A. L., Naujokaitis-Lewis, I., … Smith, P. A. (2021). Key information needs to move from knowledge to action for biodiversity conservation in Canada. *Biological Conservation*, *256*, 108983. https://doi.org/10.1016/j.biocon.2021.108983

Cai, H., Zheng, V. W., & Chang, K. C.-C. (2017). *A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications*. http://arxiv.org/abs/1709.07604

Cameron, E. K., Sundqvist, M. K., Keith, S. A., CaraDonna, P. J., Mousing, E. A., Nilsson, K. A., Metcalfe, D. B., & Classen, A. T. (2019). Uneven global distribution of food web studies under climate change. *Ecosphere*, *10*(3), e02645. https://doi.org/10.1002/ecs2.2645

Cao, R.-M., Liu, S.-Y., & Xu, X.-K. (2019). Network embedding for link prediction: The pitfall and improvement. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *29*(10), 103102. https://doi.org/10.1063/1.5120724

Cavender-Bares, J., Kozak, K. H., Fine, P. V. A., & Kembel, S. W. (2009). The merging of community ecology and phylogenetic biology. *Ecology Letters*, *12*(7), 693–715. https://doi.org/10.1111/j.1461-0248.2009.01314.x

Choudry, A. (2013). Saving biodiversity, for whom and for what? Conservation NGOs, complicity, colonialism and conquest in an era of capitalist globalization. In *NGOization: Complicity, contradictions and prospects* (pp. 24–44). Bloomsbury Publishing.

Cirtwill, A. R., Ekl, A., Roslin, T., Wootton, K., & Gravel, D. (2019). A quantitative framework for investigating the reliability of empirical network construction. *Methods in Ecology and Evolution*, *0*. https://doi.org/10.1111/2041-210X.13180

Cirtwill, A. R., & Hambäck, P. (2021). Building food networks from molecular data: Bayesian or fixed-number thresholds for including links. *Basic and Applied Ecology*, *50*, 67–76. https://doi.org/10.1016/j.baae.2020.11.007

Csermely, P. (2004). Strong links are important, but weak links stabilize them. *Trends in Biochemical Sciences*, *29*(7), 331–334. https://doi.org/10.1016/j.tibs.2004.05.004

Dalla Riva, G. V., & Stouffer, D. B. (2016). Exploring the evolutionary signature of food webs' backbones using functional traits. *Oikos*, *125*(4), 446–456. https://doi.org/10.1111/oik.02305

Dallas, T., Park, A. W., & Drake, J. M. (2017). Predicting cryptic links in host-parasite networks. *PLOS Computational Biology*, *13*(5), e1005557. https://doi.org/10.1371/journal.pcbi.1005557

Dansereau, G., & Poisot, T. (2021). SimpleSDMLayers.jl and GBIF.jl: A Framework for Species Distribution Modeling in Julia. *Journal of Open Source Software*, *6*(57), 2872. https://doi.org/10.21105/joss.02872

Domínguez, L., & Luoma, C. (2020). Decolonising Conservation Policy: How Colonial Land and Conservation Ideologies Persist and Perpetuate Indigenous Injustices at the Expense of the Environment. *Land*, *9*(3, 3), 65. https://doi.org/10.3390/land9030065

Dormann, C. F., Gruber, B., Winter, M., & Herrmann, D. (2010). Evolution of climate niches in European

mammals? *Biology Letters*, *6*(2), 229–232. https://doi.org/10.1098/rsbl.2009.0688

Dunne, J. A. (2006). The Network Structure of Food Webs. In J. A. Dunne & M. Pascual (Eds.), *Ecological networks: Linking structure and dynamics* (pp. 27–86). Oxford University Press.

Eichhorn, M. P., Baker, K., & Griffiths, M. (2019). Steps towards decolonising biogeography. *Frontiers of Biogeography*, *12*(1), 1–7. https://doi.org/10.21425/F5FBG44795

Eklöf, A., & Stouffer, D. B. (2016). The phylogenetic component of food web structure and intervality. *Theoretical Ecology*, *9*(1), 107–115. https://doi.org/10.1007/s12080-015-0273-9

Fortin, M.-J., Dale, M. R. T., & Brimacombe, C. (2021). Network ecology in dynamic landscapes. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1949), rspb.2020.1889, 20201889. https://doi.org/10.1098/rspb.2020.1889

Galiana, N., Barros, C., Braga, J., Ficetola, G. F., Maiorano, L., Thuiller, W., Montoya, J. M., & Lurgi, M. (2021). The spatial scaling of food web structure across European biogeographical regions. *Ecography*, *n/a*(n/a). https://doi.org/10.1111/ecog.05229

Galiana, N., Hawkins, B. A., & Montoya, J. M. (2019). The geographical variation of network structure is scale dependent: Understanding the biotic specialization of hostparasitoid networks. *Ecography*, *42*(6), 1175–1187. https://doi.org/10.1111/ecog.03684

Galiana, N., Lurgi, M., Claramunt-López, B., Fortin, M.-J., Leroux, S., Cazelles, K., Gravel, D., & Montoya, J. M. (2018). The spatial scaling of species interaction networks. *Nature Ecology & Evolution*, *2*(5), 782–790. https://doi.org/10.1038/s41559-018-0517-3

Garland, T., JR., Midford, P. E., & Ives, A. R. (1999). An Introduction to Phylogenetically Based Statistical Methods, with a New Method for Confidence Intervals on Ancestral Values1. *American Zoologist*, *39*(2), 374–388. https://doi.org/10.1093/icb/39.2.374

Garlaschelli, D., Hollander, F. den, & Roccaverde, A. (2018). Covariance structure behind breaking of ensemble equivalence in random graphs. *Journal of Statistical Physics*, *173*(3-4), 644–662. https://doi.org/10.1007/s10955-018-2114-x

GBIF Secretariat. (2021). *GBIF Backbone Taxonomy*. https://doi.org/10.15468/39omei

Gerhold, P., Cahill, J. F., Winter, M., Bartish, I. V., & Prinzing, A. (2015). Phylogenetic patterns are not proxies of community assembly mechanisms (they are far better). *Functional Ecology*, *29*(5), 600–614.

https://doi.org/10.1111/1365-2435.12425

Gravel, D., Baiser, B., Dunne, J. A., Kopelke, J.-P., Martinez, N. D., Nyman, T., Poisot, T., Stouffer, D. B., Tylianakis, J. M., Wood, S. A., & Roslin, T. (2018). Bringing Elton and Grinnell together: A quantitative framework to represent the biogeography of ecological interaction networks. *Ecography*, *0*(0). https://doi.org/10.1111/ecog.04006

Grünig, M., Mazzi, D., Calanca, P., Karger, D. N., & Pellissier, L. (2020). Crop and forest pest metawebs shift towards increased linkage and suitability overlap under climate change. *Communications Biology*, *3*(1, 1), 1–10. https://doi.org/10.1038/s42003-020-0962-9

Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, *24*(2), 8–12. https://doi.org/10.1109/MIS.2009.36

Halko, N., Martinsson, P. G., & Tropp, J. A. (2011). Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Review*, *53*(2), 217–288. https://doi.org/10.1137/090771806

Herbert, F. (1965). *Dune* (1st ed.). Chilton Book Company.

Hoffmann, J., Bar-Sinai, Y., Lee, L. M., Andrejevic, J., Mishra, S., Rubinstein, S. M., & Rycroft, C. H. (2019). Machine learning in a data-limited regime: Augmenting experiments with synthetic data uncovers order in crumpled sheets. *Science Advances*, *5*(4), eaau6792. https://doi.org/10.1126/sciadv.aau6792

Holm, E. A. (2019). In defense of the black box. *Science*, *364*(6435), 26–27. https://doi.org/10.1126/science.aax0162

Hortal, J., de Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M., & Ladle, R. J. (2015). Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, *46*(1), 523–549. https://doi.org/10.1146/annurev-ecolsys-112414-054400

Hutchinson, M. C., Cagua, E. F., & Stouffer, D. B. (2017). Cophylogenetic signal is detectable in pollination interactions across ecological scales. *Ecology*, n/a–n/a. https://doi.org/10.1002/ecy.1955

Jordano, P. (2016a). Chasing Ecological Interactions. *PLOS Biol*, *14*(9), e1002559. https://doi.org/10.1371/journal.pbio.1002559

Jordano, P. (2016b). Sampling networks of ecological interactions. *Functional Ecology*, *30*(12), 1883–1893. https://doi.org/10.1111/1365-2435.12763

Kawatsu, K., Ushio, M., van Veen, F. J. F., & Kondoh, M. (2021). Are networks of trophic interactions sufficient for understanding the dynamics of multi-trophic communities? Analysis of a tri-trophic insect food-web time-series. *Ecology Letters*, *24*(3), 543–552. https://doi.org/10.1111/ele.13672

Kéfi, S., Berlow, E. L., Wieters, E. A., Navarrete, S. A., Petchey, O. L., Wood, S. A., Boit, A., Joppa, L. N., Lafferty, K. D., Williams, R. J., Martinez, N. D., Menge, B. A., Blanchette, C. A., Iles, A. C., & Brose, U. (2012). More than a meal… integrating non-feeding interactions into food webs: More than a meal …. *Ecology Letters*, *15*(4), 291–300. https://doi.org/10.1111/j.1461-0248.2011.01732.x

Lamba, A., Cassey, P., Segaran, R. R., & Koh, L. P. (2019). Deep learning for environmental conservation. *Current Biology*, *29*(19), R977–R982. https://doi.org/10.1016/j.cub.2019.08.016

Litsios, G., & Salamin, N. (2012). Effects of Phylogenetic Signal on Ancestral State Reconstruction. *Systematic Biology*, *61*(3), 533–538. https://doi.org/10.1093/sysbio/syr124

Machen, R., & Nost, E. (2021). Thinking algorithmically: The making of hegemonic knowledge in climate governance. *Transactions of the Institute of British Geographers*, *46*(3), 555–569. https://doi.org/10.1111/tran.12441

Maiorano, L., Montemaggiori, A., Ficetola, G. F., O'Connor, L., & Thuiller, W. (2020a). TETRA-EU 1.0: A species-level trophic metaweb of European tetrapods. *Global Ecology and Biogeography*, *29*(9), 1452–1457. https://doi.org/10.1111/geb.13138

Maiorano, L., Montemaggiori, A., Ficetola, G. F., O'Connor, L., & Thuiller, W. (2020b). *Data from: Tetra-EU 1.0: A species-level trophic meta-web of European tetrapods* (Version 3, pp. 16596876 bytes) [Data set]. Dryad. https://doi.org/10.5061/DRYAD.JM63XSJ7B

Marco, M. D., Venter, O., Possingham, H. P., & Watson, J. E. M. (2018). Changes in human footprint drive changes in species extinction risk. *Nature Communications*, *9*(1), 4621. https://doi.org/10.1038/s41467-018-07049-5

McLeod, A., Leroux, S. J., Gravel, D., Chu, C., Cirtwill, A. R., Fortin, M.-J., Galiana, N., Poisot, T., & Wood, S. A. (2021). Sampling and asymptotic network properties of spatial multi-trophic networks. *Oikos*, *n/a*(n/a). https://doi.org/10.1111/oik.08650

Mora, B. B., Gravel, D., Gilarranz, L. J., Poisot, T., & Stouffer, D. B. (2018). Identifying a common backbone of interactions underlying food webs from different ecosystems. *Nature Communications*, *9*(1), 2603. https://doi.org/10.1038/s41467-018-05056-0

Morales-Castilla, I., Matias, M. G., Gravel, D., & Araújo, M. B. (2015). Inferring biotic interactions from proxies. *Trends in Ecology & Evolution*, *30*(6), 347–356. https://doi.org/10.1016/j.tree.2015.03.014

Mosebo Fernandes, A. C., Quintero Gonzalez, R., Lenihan-Clarke, M. A., Leslie Trotter, E. F., & Jokar Arsanjani, J. (2020). Machine Learning for Conservation Planning in a Changing Climate. *Sustainability*, *12*(18, 18), 7657. https://doi.org/10.3390/su12187657

Mouquet, N., Devictor, V., Meynard, C. N., Munoz, F., Bersier, L.-F., Chave, J., Couteron, P., Dalecky, A., Fontaine, C., Gravel, D., Hardy, O. J., Jabot, F., Lavergne, S., Leibold, M., Mouillot, D., Münkemüller, T., Pavoine, S., Prinzing, A., Rodrigues, A. S. L., … Thuiller, W. (2012). Ecophylogenetics: Advances and perspectives. *Biological Reviews*, *87*(4), 769–785. https://doi.org/10.1111/j.1469-185X.2012.00224.x

Nenzén, H. K., Montoya, D., & Varela, S. (2014). The Impact of 850,000 Years of Climate Changes on the Structure and Dynamics of Mammal Food Webs. *PLOS ONE*, *9*(9), e106651. https://doi.org/10.1371/journal.pone.0106651

Neutel, A.-M., Heesterbeek, J. A. P., & de Ruiter, P. C. (2002). Stability in Real Food Webs: Weak Links in Long Loops. *Science*, *296*(5570), 1120–1123. https://doi.org/10.1126/science.1068326

No'kmaq, M., Marshall, A., Beazley, K. F., Hum, J., joudry, shalan, Papadopoulos, A., Pictou, S., Rabesca, J., Young, L., & Zurba, M. (2021). "Awakening the sleeping giant": Re-Indigenization principles for transforming biodiversity conservation in Canada and beyond. *FACETS*, *6*(1), 839–869.

Nost, E., & Goldstein, J. E. (2021). A political ecology of data. *Environment and Planning E: Nature and Space*, 25148486211043503. https://doi.org/10.1177/25148486211043503

O'Connor, L. M. J., Pollock, L. J., Braga, J., Ficetola, G. F., Maiorano, L., Martinez-Almoyna, C., Montemaggiori, A., Ohlmann, M., & Thuiller, W. (2020). Unveiling the food webs of tetrapods across Europe through the prism of the Eltonian niche. *Journal of Biogeography*, *47*(1), 181–192. https://doi.org/10.1111/jbi.13773

Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Park, J., & Newman, M. E. J. (2004). Statistical mechanics of networks. *Physical Review E*, *70*(6), 066117. https://doi.org/10.1103/PhysRevE.70.066117

Pedersen, E. J., Thompson, P. L., Ball, R. A., Fortin, M.-J., Gouhier, T. C., Link, H., Moritz, C., Nenzen, H., Stanley, R. R. E., Taranu, Z. E., Gonzalez, A., Guichard, F., & Pepin, P. (2017). Signatures of the collapse and incipient recovery of an overexploited marine ecosystem. *Royal Society Open Science*, *4*(7), 170215. https://doi.org/10.1098/rsos.170215

Perretti, C. T., Munch, S. B., & Sugihara, G. (2013). Model-free forecasting outperforms the correct mechanistic model for simulated and experimental data. *Proceedings of the National Academy of Sciences*, *110*(13), 5253–5257. https://doi.org/10.1073/pnas.1216076110

Pires, M. M., Koch, P. L., Fariña, R. A., de Aguiar, M. A. M., dos Reis, S. F., & Guimarães, P. R. (2015). Pleistocene megafaunal interaction networks became more vulnerable after human arrival. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1814), 20151367. https://doi.org/10.1098/rspb.2015.1367

Poelen, J. H., Simons, J. D., & Mungall, C. J. (2014). Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecological Informatics*, *24*, 148–159. https://doi.org/10.1016/j.ecoinf.2014.08.005

Poisot, T., Belisle, Z., Hoebeke, L., Stock, M., & Szefer, P. (2019). EcologicalNetworks.jl - analysing ecological networks. *Ecography*. https://doi.org/10.1111/ecog.04310

Poisot, T., Bergeron, G., Cazelles, K., Dallas, T., Gravel, D., MacDonald, A., Mercier, B., Violet, C., & Vissault, S. (2021). Global knowledge gaps in species interaction networks data. *Journal of Biogeography*, *n/a*(n/a). https://doi.org/10.1111/jbi.14127

Poisot, T., Cirtwill, A. R., Cazelles, K., Gravel, D., Fortin, M.-J., & Stouffer, D. B. (2016). The structure of probabilistic networks. *Methods in Ecology and Evolution*, *7*(3), 303–312. https://doi.org/10.1111/2041-210X.12468

Poisot, T., Ouellet, M.-A., Mollentze, N., Farrell, M. J., Becker, D. J., Albery, G. F., Gibb, R. J., Seifert, S. N., & Carlson, C. J. (2021, May 31). *Imputing the mammalian virome with linear filtering and singular*

value decomposition. http://arxiv.org/abs/2105.14973

Poisot, T., & Stouffer, D. B. (2018). Interactions retain the co-phylogenetic matching that communities lost. *Oikos*, *127*(2), 230–238. https://doi.org/10.1111/oik.03788

Poisot, T., Stouffer, D. B., & Gravel, D. (2015). Beyond species: Why ecological interaction networks vary through space and time. *Oikos*, *124*(3), 243–251. https://doi.org/10.1111/oik.01719

Price, P. W. (2003). *Macroevolutionary theory on macroecological patterns*. Cambridge University Press.

Ray, J. C., Grimm, J., & Olive, A. (2021). The biodiversity crisis in Canada: Failures and challenges of federal and sub-national strategic and legal frameworks. *FACETS*, *6*, 1044–1068. https://doi.org/10.1139/facets-2020-0075

Reeve, R., Leinster, T., Cobbold, C. A., Thompson, J., Brummitt, N., Mitchell, S. N., & Matthews, L. (2016, December 8). *How to partition diversity*. http://arxiv.org/abs/1404.6520

Rosado, B. H. P., Dias, A., & de Mattos, E. (2013). Going Back to Basics: Importance of Ecophysiology when Choosing Functional Traits for Studying Communities and Ecosystems. *Natureza & Conservaç~ao Revista Brasileira de Conservaç~ao Da Natureza*, *11*, 15–22. https://doi.org/10.4322/natcon.2013.002

Runghen, R., Stouffer, D. B., & Dalla Riva, G. V. (2021). *Exploiting node metadata to predict interactions in large networks using graph embedding and neural networks*. https://doi.org/10.1101/2021.06.10.447991

Schoener, T. W. (1989). Food webs from the small to the large. *Ecology*, *70*(6), 1559–1589.

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

Stock, M. (2021). Pairwise learning for predicting pollination interactions based on traits and phylogeny. *Ecological Modelling*, 14.

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

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

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

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

Thessen, A. E., & Parr, C. S. (2014). Knowledge extraction and semantic annotation of text from the encyclopedia of life. *PloS One*, *9*(3), e89550.

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.

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

Turak, E., Brazill-Boast, J., Cooney, T., Drielsma, M., DelaCruz, J., Dunkerley, G., Fernandez, M., Ferrier, S., Gill, M., Jones, H., Koen, T., Leys, J., McGeoch, M., Mihoub, J.-B., Scanes, P., Schmeller, D., & Williams, K. (2017). Using the essential biodiversity variables framework to measure biodiversity change at national scale. *Biological Conservation*, *213*, 264–271. https://doi.org/10.1016/j.biocon.2016.08.019

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

van der Hoorn, P., Lippner, G., & Krioukov, D. (2018). Sparse Maximum-Entropy Random Graphs with a Given Power-Law Degree Distribution. *Journal of Statistical Physics*, *173*(3-4), 806–844. https://doi.org/10.1007/s10955-017-1887-7

Vermaat, J. E., Dunne, J. A., & Gilbert, A. J. (2009). Major dimensions in food-web structure properties. *Ecology*, *90*(1), 278–282. http://www.ncbi.nlm.nih.gov/pubmed/19294932

Wood, S. A., Russell, R., Hanson, D., Williams, R. J., & Dunne, J. A. (2015). Effects of spatial scale of

sampling on food web structure. *Ecology and Evolution*, *5*(17), 3769–3782. https://doi.org/10.1002/ece3.1640

Xing, S., & Fayle, T. M. (2021). The rise of ecological network meta-analyses: Problems and prospects. *Global Ecology and Conservation*, *30*, e01805. https://doi.org/10.1016/j.gecco.2021.e01805

Yeakel, J. D., Pires, M. M., Rudolf, L., Dominy, N. J., Koch, P. L., Guimarães, P. R., & Gross, T. (2014). Collapse of an ecological network in Ancient Egypt. *PNAS*, *111*(40), 14472–14477. https://doi.org/10.1073/pnas.1408471111

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35. https://doi.org/10.1002/1097-0142(1950)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3

Young, J.-G., Cantwell, G. T., & Newman, M. E. J. (2021). Bayesian inference of network structure from unreliable data. *Journal of Complex Networks*, *8*(6). https://doi.org/10.1093/comnet/cnaa046

Young, S. J., & Scheinerman, E. R. (2007). Random Dot Product Graph Models for Social Networks. In A. Bonato & F. R. K. Chung (Eds.), *Algorithms and Models for the Web-Graph* (pp. 138–149). Springer. https://doi.org/10.1007/978-3-540-77004-6_11

Zhu, M., & Ghodsi, A. (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, *51*(2), 918–930. https://doi.org/10.1016/j.csda.2005.09.010

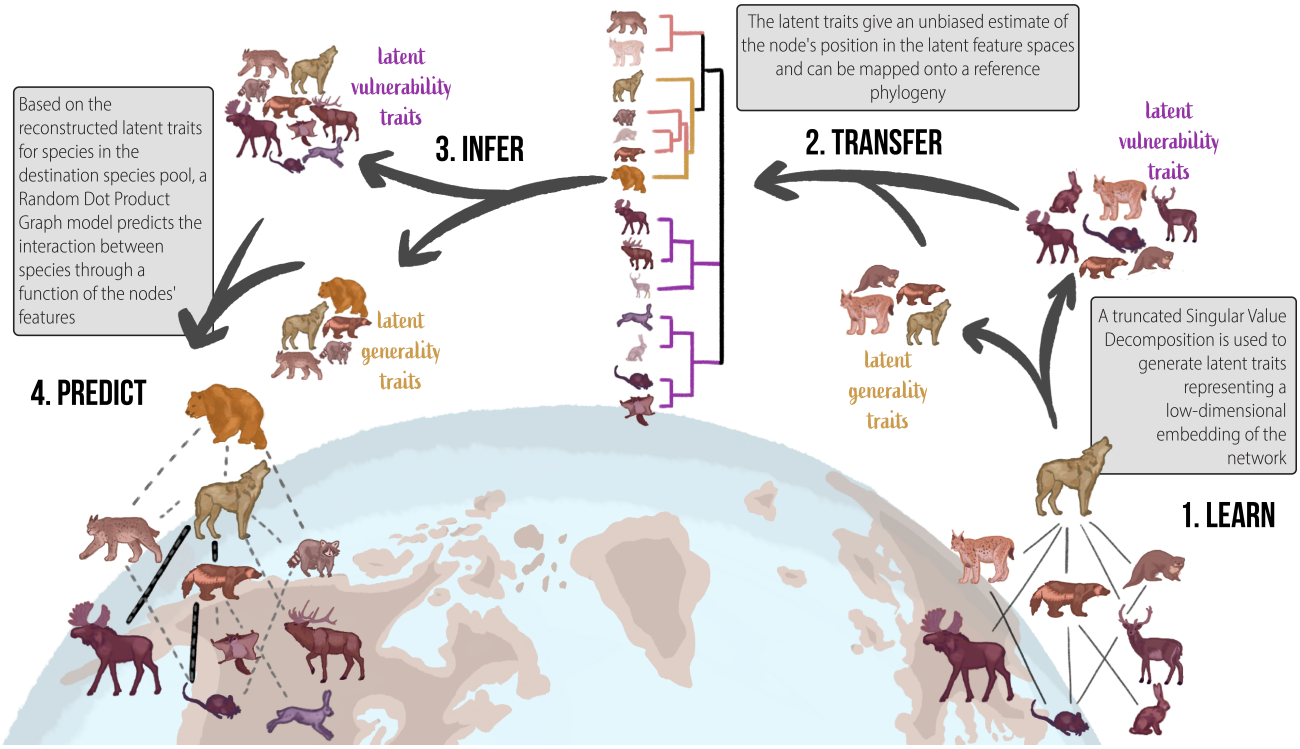

Figure 1: Overview of the phylogenetic transfer learning (and prediction) of species interactions networks. Starting from an initial, known, network, we learn its representation through a graph embedding step (here, a truncated Singular Value Decomposition; Step 1), yielding a series of latent traits (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).

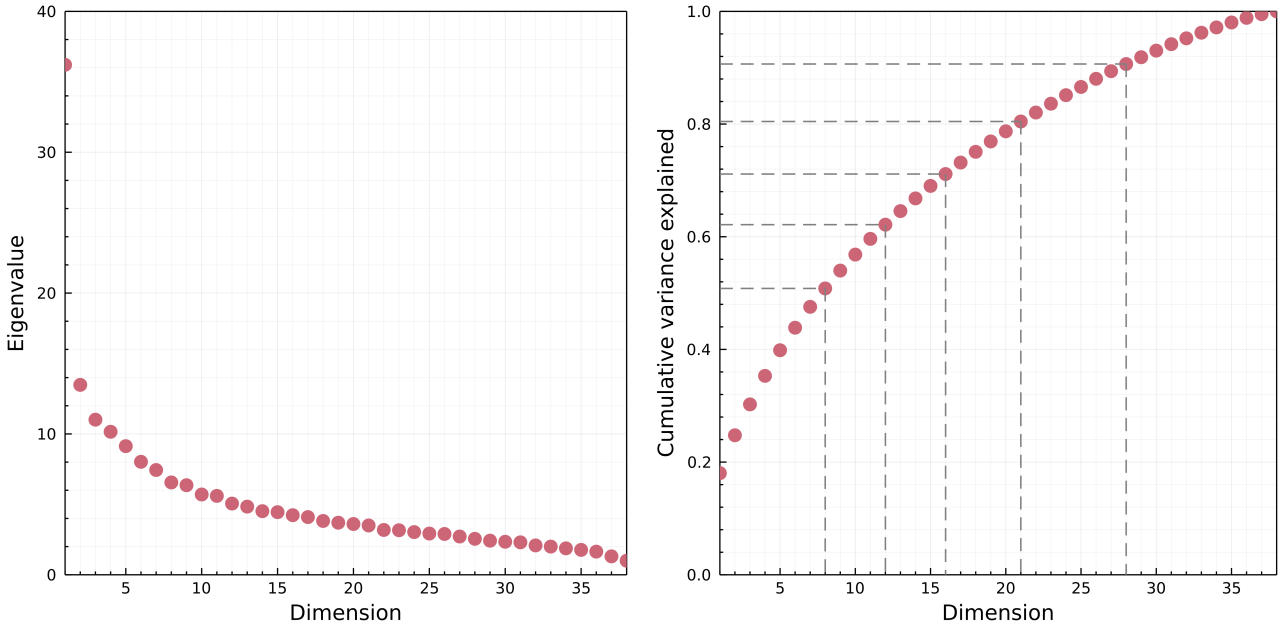

Figure 2: Left: representation of the scree plot of the singular values from the t-SVD on the European metaweb. The scree plot shows no obvious drop in the singular values that may be leveraged to automatically detect a minimal dimension for embedding, after *e.g.* Zhu & Ghodsi (2006). Right: cumulative 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.

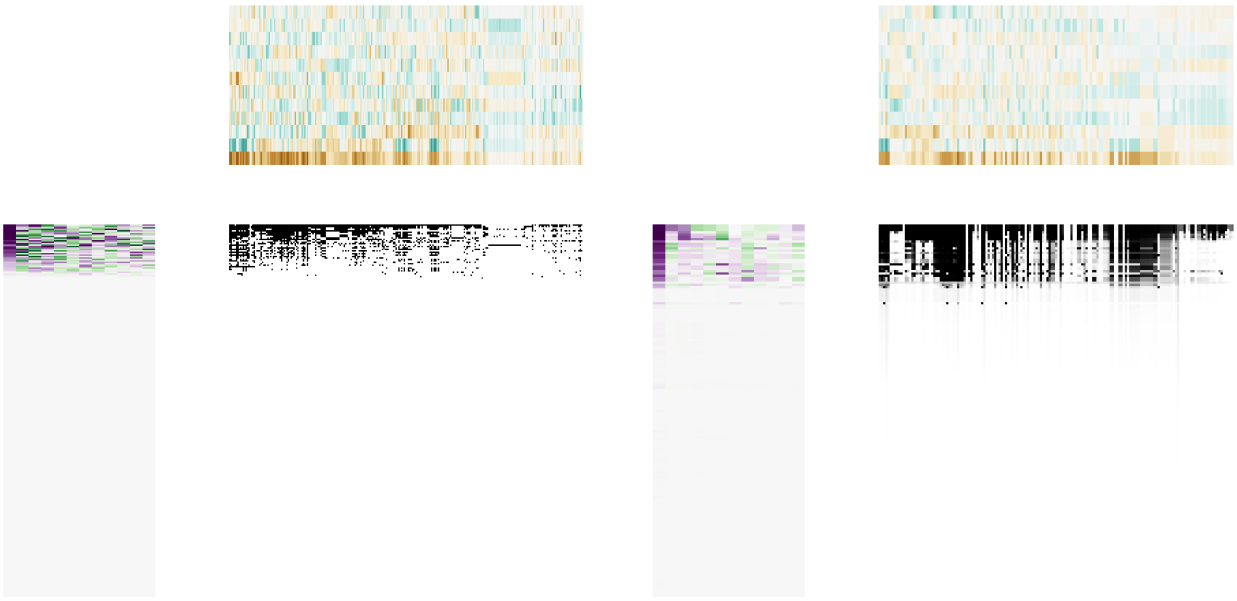

Figure 3: Visual representation of the left (green/purple) and right (green/brown) subspaces, alongside the adjacency matrix of the food web they encode (greyscale). The European metaweb is on the left, and the imputed Canadian metaweb (before data inflation) on the right. This figure illustrates how 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.

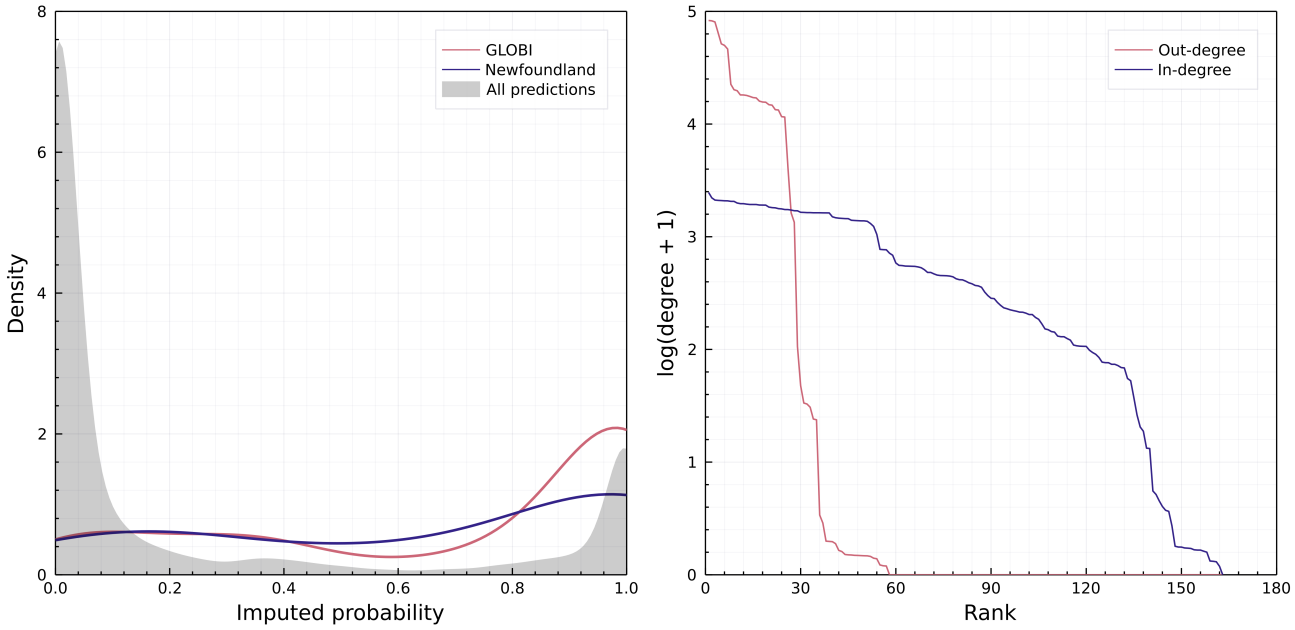

Figure 4: Left, comparison of the probabilities of interactions assigned by the model to all interactions (grey curve), the subset of interactions found in GLOBI (red), and in the Strong & Leroux (2014) Newfoundland dataset (blue). The model recovers more interactions with a low probability compared 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.

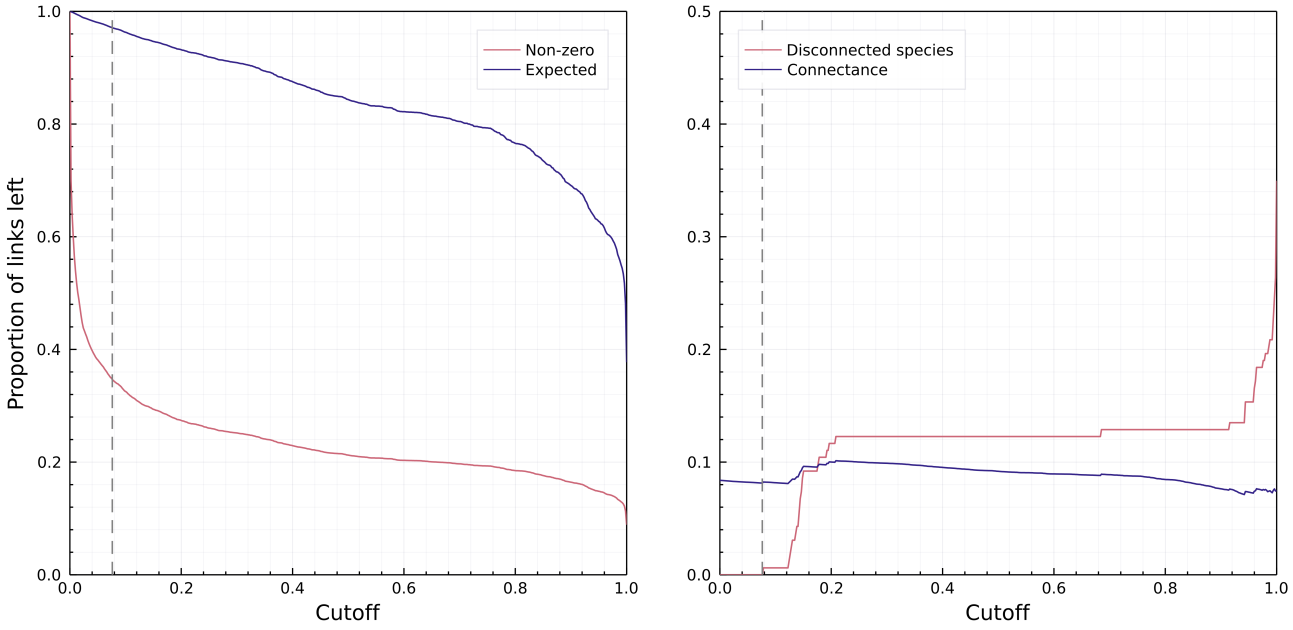

Figure 5: Left: effect of varying the cutoff for probabilities to be considered non-zero on the number of unique links and on $\hat{L}$, the probabilistic estimate of the number of links assuming that all interactions are independent. Right: effect of varying the cutoff on the number of disconnected 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.

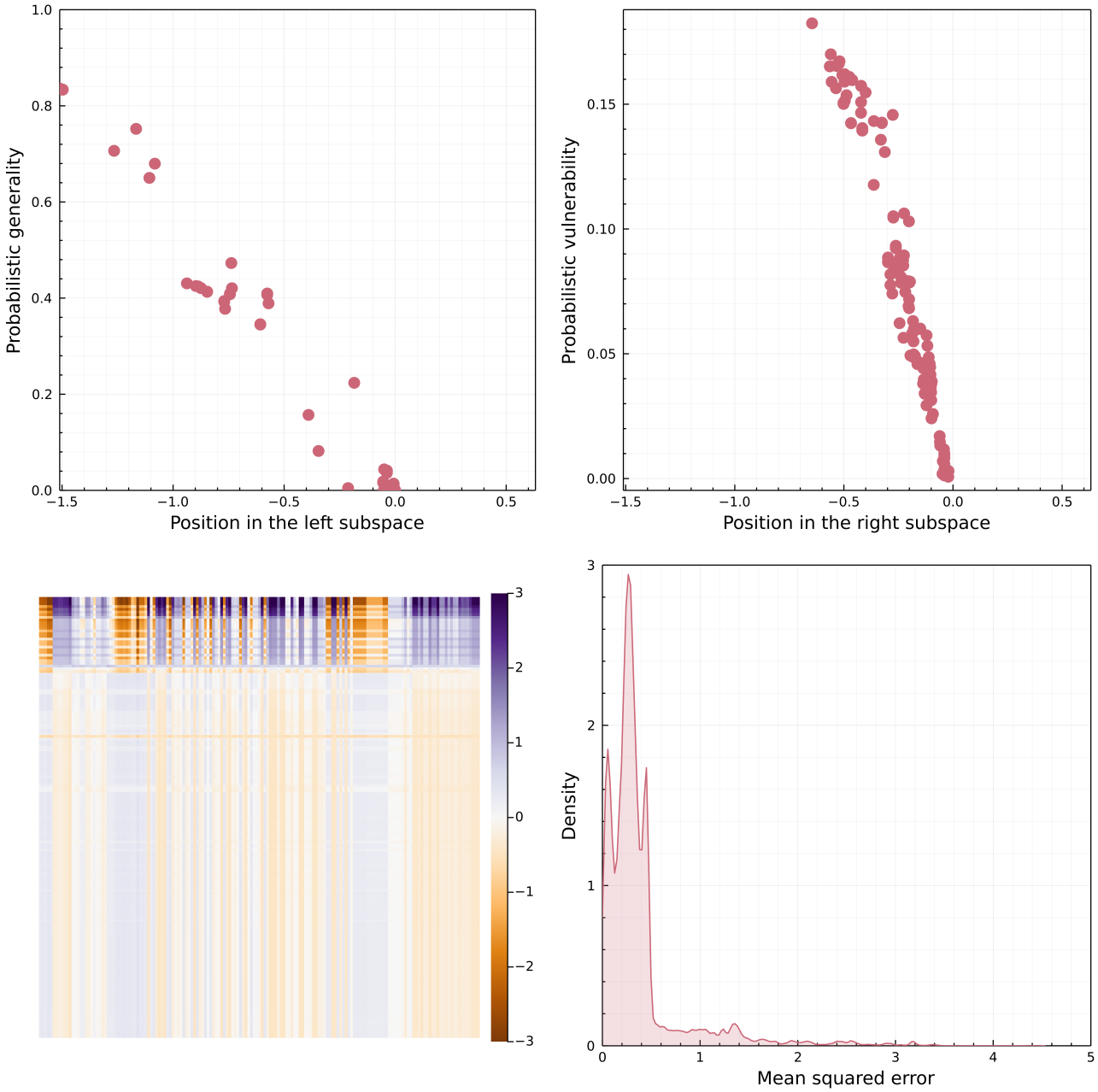

Figure 6: Top: biological significance of the first dimension. Left: there is a linear relationship between the values on the first dimension of the left subspace and the generality, *i.e.* the relative number of preys, *sensu* Schoener (1989). Species with a value of 0 in this subspace are at the bottom-most 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.