Université de Montréal

Tools 'O the Times: Understanding the common properties of species interaction networks across space

par

Tanya Strydom

Unité académique : Département de sciences biologiques,
Faculté des arts et des sciences

Thèse présentée en vue de l'obtention du grade de
Philosophiæ Doctor (Ph.D.)
en sciences biologiques,

8 Novembre 2023

Université de Montréal

Unité académique : Département de sciences biologiques,
Faculté des arts et des sciences

Cette thèse intitulée

Tools 'O the Times: Understanding the common properties of species interaction networks across space

présentée par

Tanya Strydom

a été évaluée par un jury composé des personnes suivantes :

Christopher Cameron
(président-rapporteur)

Timothée Poisot
(directeur de recherche)

Mariana P. Braga
(membre du jury)

Jacquelyn Gill
(examinateur externe)

Miklos Csűrös
(représentant du doyen de la FESP)

Résumé

Le domaine de l'écologie des réseaux est encore limité dans sa capacité à faire des inférences mondiales à grande échelle. Ce défi est principalement dû à la difficulté d'échantillonnage des interactions sur le terrain, entraînant de nombreuses « lacunes » en ce qui concerne la couverture mondiale des données. Cette thèse adopte une approche « centrée sur les méthodes » de l'écologie des réseaux et se concentre sur l'idée de développer des outils pour aider à combler les lacunes en matière de données en présentant la prédiction comme une alternative accessible à l'échantillonnage sur le terrain et introduit deux « outils » différents qui sont prêts à poser des questions à l'échelle mondiale.

Le chapitre 1 présente les outils que nous pouvons utiliser pour faire des prédictions de réseaux et est motivé par l'idée selon laquelle avoir la capacité de prédire les interactions entre les espèces grâce à l'utilisation d'outils de modélisation est impératif pour une compréhension plus globale des réseaux écologiques. Ce chapitre comprend une preuve de concept (dans laquelle nous montrons comment un simple modèle de réseau neuronal est capable de faire des prédictions précises sur les interactions entre espèces), une évaluation des défis et des opportunités associés à l'amélioration des prédictions d'interaction et une feuille de route conceptuelle concernant l'utilisation de modèles prédictifs pour les réseaux écologiques.

Les chapitres 2 et 3 sont étroitement liés et se concentrent sur l'utilisation de l'intégration de graphiques pour la prédiction de réseau. Essentiellement, l'intégration de graphes nous permet de transformer un graphe (réseau) en un ensemble de vecteurs, qui capturent une propriété écologique du réseau et nous fournissent une abstraction simple mais puissante d'un réseau d'interaction et servent de moyen de maximiser les informations disponibles. disponibles à partir des réseaux d'interactions d'espèces. Parce que l'intégration de graphes nous permet de « décoder » les informations au sein d'un réseau, elle est conçue comme un outil de prédiction de réseau, en particulier lorsqu'elle est utilisée dans un cadre d'apprentissage par transfert. Elle s'appuie sur l'idée que nous pouvons utiliser les connaissances acquises en résolvant un problème connu. et l'utiliser pour résoudre un problème étroitement lié. Ici, nous avons utilisé le métaweb européen (connu) pour prédire un métaweb pour les espèces

canadiennes en fonction de leur parenté phylogénétique. Ce qui rend ce travail particulièrement passionnant est que malgré le faible nombre d'espèces partagées entre ces deux régions, nous sommes capables de récupérer la plupart (91%) des interactions.

Le chapitre 4 approfondit la réflexion sur la complexité des réseaux et les différentes manières que nous pourrions choisir de définir la complexité. Plus spécifiquement, nous remettons en question les mesures structurelles plus traditionnelles de la complexité en présentant l'entropie SVD comme une mesure alternative de la complexité. Adopter une approche physique pour définir la complexité nous permet de réfléchir aux informations contenues dans un réseau plutôt qu'à leurs propriétés émergentes. Il est intéressant de noter que l'entropie SVD révèle que les réseaux bipartites sont très complexes et ne sont pas nécessairement conformes à l'idée selon laquelle la complexité engendre la stabilité.

Enfin, je présente le package Julia `SpatialBoundaries.jl`. Ce package permet à l'utilisateur d'implémenter l'algorithme de wombling spatial pour des données disposées de manière uniforme ou aléatoire dans l'espace. Étant donné que l'algorithme de wombling spatial se concentre à la fois sur le gradient et sur la direction du changement pour un paysage donné, il peut être utilisé à la fois pour détecter les limites au sens traditionnel du terme ainsi que pour examiner de manière plus nuancée la direction des changements. Cette approche pourrait être un moyen bénéfique de réfléchir aux questions liées à la détection des limites des réseaux et à leur relation avec les limites environnementales.

Mots-clés: réseaux écologiques, décomposition des valeurs singulières, apprentissage par transfert, wombling spatial

Abstract

The field of network ecology is still limited in its ability to make large-scale, global inferences. This challenge is primarily driven by the difficulty of sampling interactions in the field, leading to many 'gaps' with regards to global coverage of data. This thesis takes a 'methods-centric' approach to network ecology and focuses on the idea of developing tools to help with filling in the the data gaps by presenting prediction as an accessible alternative to sampling in the field and introduces two different 'tools' that are primed for asking questions at global scales.

Chapter 1 maps out tools we can use to make network predictions and is driven by the idea that having the ability to predict interactions between species through the use of modelling tools is imperative for a more global understanding of ecological networks. This chapter includes a proof-of-concept (where we show how a simple neural network model is able to make accurate predictions about species interactions), an assessment of the challenges and opportunities associated with improving interaction predictions, and providing a conceptual roadmap concerned with the use of predictive models for ecological networks.

Chapters 2 and 3 are closely intertwined and are focused on the use of graph embedding for network prediction. Essentially graph embedding allows us to transform a graph (network) into a set of vectors, which capture an ecological property of the network and provides us with a simple, yet powerful abstraction of an interaction network and serves as a way to maximise the available information available from species interaction networks. Because graph embedding allows us to 'decode' the information within a network it is primed as a tool for network prediction, specifically when used in a transfer learning framework, this builds on the idea that we can take the knowledge gained from solving a known problem and using it to solve a closely related problem. Here we used the (known) European metaweb to predict a metaweb for Canadian species based on their phylogenetic relatedness. What makes this work particularly exciting is that despite the low number of species shared between these two regions we are able to recover most (91%) of interactions.

Chapter 4 delves into thinking about the complexity of networks and the different ways we might choose to define complexity. More specifically we challenge the more traditional structural measures of complexity by presenting SVD entropy as an alternative measure of

complexity. Taking a physical approach to defining complexity allows us to think about the information contained within a network as opposed to their emerging properties. Interestingly, SVD entropy reveals that bipartite networks are highly complex and do not necessarily conform to the idea that complexity begets stability.

Finally, I present the Julia package `SpatialBoundaries.jl`. This package allows the user to implement the spatial wombling algorithm for data arranged uniformly or randomly across space. Because the spatial wombling algorithm focuses on both the gradient as well as the direction of change for the given landscape it can be used both for detecting boundaries in the traditional sense as well as a more nuanced look at at the direction of changes. This approach could be a beneficial way with which to think about questions which relate to boundary detection for networks and how these relate to environmental boundaries.

Keywords: ecological networks, singular value decomposition, transfer learning, spatial wombling

Contents

List of tables

List of figures

List of abbreviations

RDPG	Random Dot Product Graph
SVD	Singular Value Decomposition
t-SVD	truncated Singular Value Decomposition

For those with the messy notebooks.
To those always seeking a way

Acknowledgements

To my many, many (awesome and all around great) collaborators. To the roadmap team; Michael Catchen, Francis Banville, Dominique Caron, Gabriel Dansereau, Philippe Desjardins-Proulx, Norma Forero-Muñoz, Gracielle Higino, Benjamin Mercier, Andrew Gonzalez, Dominique Gravel, and Laura Pollock — thank you for bringing your diverse scientific backgrounds, thoughts, and ideas to the table. To the metaweb team; Salomé Bouskila, Francis Banville, Ceres Barros, Dominique Caron, Maxwell Farrell, Marie-Josée Fortin, Victoria Hemming, Benjamin Mercier, Laura Pollock, and Rogini Runghen — thanks for the (seemingly) endless rounds of feedback and tweaking to make the manuscripts more reader friendly and robust - I *think* we're almost there! A special shout out to Giulio Dalla Riva for bringing the endless energy and enthusiasm when it comes to anything SVD related!

To my advisor Timothée Poisot. Thank you for taking in the field ecologist who wanted to dip their toes into the world of thinking boxes and species interaction networks. (Turns out that the need to have a little cry before continuing with work is not unique to field work but also extends to trying to make the code go brrr). Thanks for affording me the space to grow not only as a scientist but also as an artist (*sensu lato*). The chance to experiment with visual ways to communicate our science may not necessarily have resulted in more succinct manuscripts but it has for sure changed the way I interact with my research and makes me think about the ways we communicate our science maybe a bit too much!

To the (past and present) members of the Poisot Lab. The last few years may not have been the best environment for nurturing a collaborative environment but you made the best of it. Despite the lag-y online calls and weird time differences you managed to make 'lab-life' feel somewhat normal. Thanks for being a first port of call when trying to navigate university administration. Sorry for always running overtime in my 1:1's (despite my best efforts). Look after Ahsoka!

A special nod to Gracielle Higino. Thanks for always bringing the *ENERGY* and for your continued mentorship and commitment to making science a KINDER place.

To mom and dad. Thanks for letting me draw on the windows when the notebook space just wasn't enough. I'm sure the funny letters will wash off one day...

Thank you to those that are the driving force behind the Living Data Project and BIOS^2 training programs. The exposure and training opportunities related to 'real world' science outside of school has been invaluable.

This work would not have been possible without funding from the Courtois Foundation, the Canadian Institute for Ecology & Evolution (CIEE), the Viral Emergence Research Initiative (VERENA), and support provided by Calcul Québec (www.calculquebec.ca) and Compute Canada (www.computecanada.ca).

Introduction

0.1. A case for tools and methods

The way that species interact with one another provides us with a 'point of departure' from which to study or understand biodiversity and the environment at a range of scales (Jordano, 2016a). This ranges from understanding how interactions can shape and drive population dynamics, the maintenance and functioning of ecosystems, as well as long-term evolutionary dynamics (Albrecht et al., 2018; Landi et al., 2018). Species interactions (and the resulting networks) can be formalised and viewed under the lens of graph theory (Dale and Fortin, 2010) - with species being nodes and interactions being edges. This provides us with a robust framework built on a mathematical foundation from which to approach network analysis and quantify various measures of network structure and behaviour (Delmas et al., 2018).

In the process of assembling ecological networks as graphs we are also 'encoding' an 'ecological fingerprint' for that community. This raises the question of how far we can take the idea of 'decoding' networks by leveraging the mathematical framework to better understand the information that they contain. In particular by leaning on the mathematical properties (and the ecological information they represent) to make network predictions, and as a means to provide us with more information as to how networks may vary over time or spatial scales.

Although the field of network ecology might have a strong conceptual and theoretical basis from which to work with, we are still at somewhat of a loss when is comes to our ability to leverage this framework to make any generalised or macroecological conclusions

about the properties of networks over larger geographic scales (although see Baiser et al., 2019; Pinheiro et al., 2023 who explicitly try and tie networks to classical macroecological theories/laws). This limited understanding can (at least in a large part) be attributed to the sparse global coverage of interaction data (Cameron et al., 2019; Poisot, Bergeron, et al., 2021), which itself is driven by the immense challenges associated with observing and recording interactions in the field (Bennett et al., 2019; Jordano, 2016b). Given the limited feasibility of being able to curate interaction datasets in a way that will result in a global coverage it makes sense to turn to predictive methods as a way to begin filling in the 'gaps' of the global map of interaction data. Although this may seem a daunting task we can lean on the mathematical formalisation and the information that networks contain to make this a possibility, once we have crossed that bridge (*i.e.,* filled the global gaps) we may then find ourselves in a position to be able to ask more global-minded questions (Thuiller et al., 2023; Windsor et al., 2023).

This pipeline from prediction to global questions is shown in Figure 1 and is the mainstay of this thesis document *i.e.,* the thesis itself can be thought of as two parts. The first part is addressing the need for predictive tools and discusses as well as develops methods we can use to begin filling in the global map. The second phase of the thesis briefly touches on some new 'tools' we can use when we start to think about large scale questions pertaining to network properties, specifically the question of network complexity (and how the definition thereof matters), as well as detecting boundaries between networks.

0.1.1. Prediction for gap-filling

Current methods for network prediction are often conceptualised around and focused on a single facet of species interactions, such phylogenetic matching (Elmasri et al., 2020; Pomeranz et al., 2018), or functional traits (Bartomeus et al., 2016). More recently applications of ensemble modelling (Becker et al., 2020) and discussions on the potential of machine learning methods (Desjardins-Proulx et al., 2019) show promise in addressing methodological constraints to prediction and the growth of open tools and data may mitigate some data

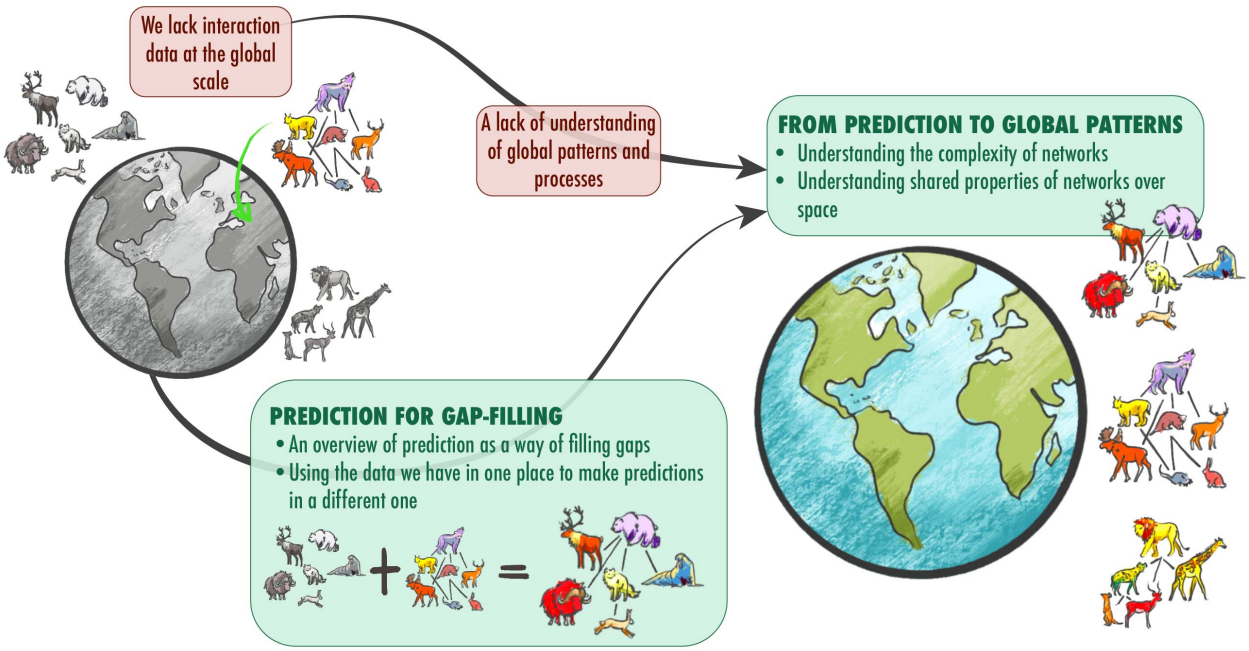


Fig. 1. One of the biggest factors limiting our ability to ask global questions about ecological networks is the lack of global data. This figure provides a high-level overview of how the development and adoption of predictive methods will equip us to begin asking and answering large-scale questions.

constraints in the coming years. However, we still lack a clear path forward or research agenda as to how we can maximise and integrate these resources to allow for ecologically plausible, and accurate, predictions.

The task of trying to predict networks is discussed in chapters 1 and 2, where we map out and discuss some of the methodological considerations we are faced with when trying to approach the task of network prediction. Chapter 1 provides a more scoping discussion on these methods, whereas Chapter 2 represents a more detailed discussion on the prospect of using graph embedding and transfer learning for network prediction. These specific methods are also the framework presented and used in Chapter 3. This section acts as a 'proof-of-concept' showcasing that the task of network prediction is both attainable and capable of producing ecologically plausible networks. Prediction is thus a way in which we can move from a scenario where we have incomplete global coverage of interaction data (*i.e.,* a 'grey scaled' world) to one that exists with less gaps in the data (*i.e.,* is more 'colourful'; Figure 1).

0.1.2. From prediction to global patterns

Although prediction is a powerful tool in the immediate/local sense (*e.g.,* it can allow local land managers/custodians to have a first approximation of how species may be interacting in that given area) it is of course also a feasible way to fill in the global interaction network map. A 'filled map' will put us in the position to develop a more mechanistic, global-scaled, understanding of networks. Specifically, we need tools that will allow us to use the *correct* methodology when comparing networks form different regions (chapter 4) and we can begin to leverage those data to understand the spatial structure of networks (chapter 5). Chapter 4, which presents a different, more information theory approach to defining complexity using the singular value vector component of an SVD (Shannon, 1948), and the final chapter of this thesis (chapter 5) is a `Julia` package that allows users to implement the Wombling algorithm (an edge detection mechanism; Womble, 1951).

Ecological networks have always been deemed to be "complex", and an interest in the notion of complexity has (in part) been tied to network stability (Landi et al., 2018). However the relationship between complexity and stability remains inconsistent when rigorously tested on empirical datasets (Jacquet et al., 2016), and although ecological networks may be complex, the ways that we currently define complexity do not translate into predictions about their stability. Traditionally network ecology readily assumes that because a system has more components (*e.g.,* links) it means that the system itself is complex. In chapter 4 we challenge the more traditional structural ('behavioural') measures of complexity and present SVD entropy as an alternative ('physical') measure of complexity.

Being able to subdivide networks into patches within a landscape will help us to better understand the boundaries of (and between) networks as well as how these may relate to species or community changes and boundaries - such as when transitioning across habitat 'boundaries' (Hackett et al., 2019). Wombling has been discussed as a useful tool for spatial analyses in ecology (Fortin and Dale, 2005) and has been used to detect transitions across a landscape (Philibert et al., 2008), changes in biological variables in communities (Barbujani et al., 1989) and to analyse the spread of invasive species (Fitzpatrick et al., 2010).

0.1.3. Objectives

Being able to understand, quantify, and work with ecological networks is important from a conservation and land management perspective as this will have cascading implications with regards to ecosystem functioning and stability. Yet we are severely hindered by a lack of high-quality, usable data as well as an appropriate set of tools that can be used to contextualise and understand ecological networks. There is a need for tools that can help us construct networks for where there are no data *i.e.,* make predictions, as well as developing tools (or ideas) that can be used to help further our mechanistic understanding of networks once we are at a point where we have the large scale data to do so. My work will help address these two issues in the context of developing tools that will either directly enable us to make predictions (Chapters 1, 2, and 3), or present methods that are aligned with global (large-scale) questions that will allow us to compare networks (Chapter 4) or attempt to delineate them (Chapter 5).

0.2. Overview of key methodological approaches

0.2.1. Transfer learning for network prediction

Transfer learning is a machine learning methodology that uses the knowledge gained when solving a known problem and applying it to solve a (related) problem by transferring the knowledge across a shared medium (space; Pan and Yang, 2010; Torrey and Shavlik, 2010). The concept of transfer learning is an approach that is particularly well suited for the problem of network prediction as it allows us to lean on the data that are available to enable us to make *de novo* interaction network predictions. This could be as simple as pinpointing missing interactions in the existing data (*e.g.,* pairwise learning has been used to predict plant-pollinator interactions; Stock, 2021) as well as a way to predict novel interactions (*i.e.,* fill in those global gaps) in a different location. This, in a sense, allows us to bring knowledge with us from an area for which we *have* data to an area where it is *lacking*. In the case of predicting species interactions, transfer learning is useful because

interactions are phylogenetically conserved and thus phylogenetic relatedness can be used to predict interactions (Davies, 2021; Elmasri et al., 2020; Gómez et al., 2010). Chapter 3 presents a transfer learning framework and uses the task of constructing a Canadian metaweb (a list of all possible interactions for a species pool) using the European metaweb assembled by (Maiorano et al., 2020) as a proof-of-concept. Below is a high-level summary of that framework, and a more detailed description of the workflow can be found `here`.

0.2.1.1. Learning using graph embedding. Before one can transfer any knowledge we must first learn something about the system using known interaction network. Since ecological networks can be represented by their adjacency matrices we can turn to graph theory to help us find a way to learn something about the known interaction network. Graph embedding is a low dimensional representation of the graph (interaction network) but, importantly, still preserves its topology (Yan et al., 2005). This process essentially allows us to learn something about where species (nodes) are situated within the network - which (in an abstract way) informs us of the role a species plays in the community (*e.g.,* the 'predatorness' or 'prey-ness' of a species). There are multiple embedding approaches discussed in Chapter 2, but in the context of the framework developed in Chapter 3 we will focus on the use of SVD as an embedding technique. SVD presents an appropriate embedding of ecological networks, having been shown to both capture their complex, emerging properties (Strydom et al., 2021) and allow for the highly accurate prediction of the interactions within a single network (Poisot, Ouellet, et al., 2021).

0.2.1.2. Graph embedding using SVD. Singular Value Decomposition (Forsythe and Moler, 1967; Golub and Reinsch, 1971) is the factorisation of an adjacency matrix $\mathbf{A}$ (where $\mathbf{A}_{m,n} \in \mathbb{B}$) into the form:

$$\mathbf{U} \cdot \mathbf{\Sigma} \cdot \mathbf{V}^T$$

Where $\mathbf{U}$ is an $m \times m$ orthogonal matrix and $\mathbf{V}$ an $n \times n$ orthogonal matrix. The columns in these matrices are, respectively, the left- and right-singular vectors of $\mathbf{A}$. $\mathbf{\Sigma}$ is a diagonal matrix that contains only non-negative σ values.

An SVD can be truncated so as to remove 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 (whith $1 \leq r \leq r_{full}$ where r_{full} is the full rank of $\mathbf{A}$ and r the rank at which we truncate the matrix). Additionally, $\mathbf{U}_t$ is now a $m \times r$ semi unitary matrix and $\mathbf{V}'_t$ an $n \times r$ semi-unitary matrix.

In the context of 'learning using embedding' the learned information is captured using an SVD, however for the task of network prediction we modified the products of the SVD so that they could be used for an RDPG. An RDPG estimates the probability of observing interactions between nodes (species) as a function of the latent variables of the nodes. An RDPG allows us to turn an SVD (which consists of three matrices) 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 thus constructed from an SVD and 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 $\mathscr{L}\mathscr{R} \approx \mathbf{A}$. These subspaces are ecologically informative and tell us about the 'generality' (think predator capacity, *sensu* Schoener, 1989) and 'vulnerability' (think capacity to be prey, *sensu* Schoener, 1989) of the species in the European network. This in essence provides us with an idea of where a species is likely to occur within a network/the space it occupies in the network.

0.2.1.3. Transferring and inferring using phylogenetic relatedness. In order to transfer the knowledge (the generality and vulnerability values) from a known network to the destination species pool (*i.e.,* a community for which we have no interaction data), we performed ancestral character estimation using a Brownian motion model and the Upham et al., 2019 mammalian phylogeny. This uses the estimated feature vectors (left and right subspaces) for the species from the known network to create a state reconstruction for all species and allows us to impute the missing generality and vulnerability values for the destination species pool that are not already in the known network. Essentially this allows us to infer where in the two subspaces the destination species are located.

0.2.1.4. Novel Prediction using RDPG. As we now essentially have the left and right subspaces for the destination species pool we can directly multiply these to yield the metaweb, specifically using an RDPG. Because of how the phylogenetic reconstruction was implemented the left and right subspaces have an associated uncertainty, therefore, we can assemble a *probabilistic* metaweb, *sensu* Poisot et al., 2016, *i.e.,* in which every interaction is represented as a single, independent, Bernoulli event of probability p.

0.2.2. SVD entropy: a measure of network complexity

We can also use SVD as a way to define the complexity of a network. Two potential candidate measures of complexity can be derived based on the 'physical structure' of (*i.e.,* information within) a network. The first measure is the rank of the matrix. The rank of $\mathbf{A}$ (noted as $r = \text{rk}(\mathbf{A})$) is the dimension of the vector space spanned by the matrix and corresponds to the number of linearly independent rows or columns, which works as an estimate of its 'external complexity', since it describes the dimensionality of the vector space of the matrix. Looking at this from an ecological standpoint, we can think of the this as quantifying the number of unique 'strategies' within a network.

The second measure is to calculate the entropy of the matrix obtained through SVD by using the singular values Shannon, 1948. This so-called SVD entropy measures the extent to which each rank encodes an equal amount of information (as the singular values capture the importance of each rank to reconstruct the original matrix) this approach therefore serves as a measure of 'internal complexity'.

Intuitively, the singular value i (σ_i) measures how much of the dataset is (proportionally) explained by each vector - therefore, one can measure the entropy of σ following Shannon, 1948. High values of SVD entropy reflects that all vectors are equally important, *i.e.,* that the structure of the ecological network cannot be efficiently compressed, and therefore indicates a high complexity (Gu and Shao, 2016). Because networks have different dimensions, we can use Pielou's evenness (Pielou, 1975) to ensure that values are lower than unity, and quantify SVD entropy, using $s_i = \sigma_i / \text{sum}(\sigma)$ as:

$$J = -\frac{1}{\ln(k)} \sum_{i=1}^{k} s_i \cdot \ln(s_i)$$

Where $k = \mathrm{rk}(\mathbf{A})$ *i.e.,* the rank of the matrix, which is equal to the number of non-zero entries in $\mathbf{\Sigma}$ as per the Eckart-Young-Mirsky theorem (Eckart and Young, 1936; Golub et al., 1987).

0.2.3. Spatial wombling for edge detection

Spatial wombling (an edge detection algorithm; Womble, 1951). Chapter 5 presents a `Julia` package that implements both the lattice and triangulation wombling algorithms. Broadly, wombling interpolates between a given set of points but in addition to looking at the deference between said points it also looks at the direction (slope) of the difference between points. First we can calculate the rate of change m which is calculated as:

$$m = \sqrt{\frac{\partial f(x,y)}{\partial x}^2 + \frac{\partial f(x,y)}{\partial y}^2}$$

This can be used to find the zones of rapid change across the landscape and identify potential candidate boundaries (which would be where change is occurring most rapidly). It is also possible to calculate the direction (θ) for each rate of change. This is calculated as:

$$\theta = \arctan\left(\frac{\partial f(x,y)}{\partial y} \Big/ \frac{\partial f(x,y)}{\partial x}\right) + \Delta$$

$$\text{where} \quad \Delta = \left\{ \begin{array}{ll} 0 & \text{if} \quad \frac{\partial f(x,y)}{\partial x} \geq 0 \\ 180 & \text{if} \quad \frac{\partial f(x,y)}{\partial x} < 0 \end{array} \right\}$$

Both m and θ are an approximation on the 'topology' of a certain metric (z, *e.g.,* number of species) between a collection of points in a landscape. Similarity between the z values indicates a uniformity between those points and thus a low rate of change whereas a high degree of difference between points is indicative of rapid change *i.e.,* a boundary as we transition from one zone to the next.

0.2.3.1. Lattice wombling. For a lattice of points where one will have sampling locations arranged the 'topology' *i.e.* function of the landscape as determined by z ($f(x,y)$) can be defined as:

$$f(x,y) = z_1(1-x)(1-y) + z_2x(1-y) + z_3xy + z_4(1-x)y$$

0.2.3.2. Triangulation wombling. When working with points that are irregularly distributed across the landscape it is possible to use triangulation wombling (Fortin and Drapeau, 1995). The three nearest neighbours can be determined using a Delaunay triangulation algorithm Delaunay, 1934 and $f(x,y)$ can be defined as:

$$f(x,y) = ax + by + c$$

where:

$$\begin{bmatrix} a & b & c \end{bmatrix} = \begin{bmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{bmatrix}^{-1} \cdot \begin{bmatrix} z_1 & z_2 & z_3 \end{bmatrix}$$

and the position of the centroid between points is calculated as follows:

$$\left(\frac{x_1 + x_2 + x_3}{3}\right), \left(\frac{y_1 + y_2 + y_3}{3}\right)$$

0.2.3.3. Boundary detection. Detecting boundaries *i.e.,* areas where the angle of the landscape transitions sharply is surprisingly simple. After having calculated the rate of change (m) for the geographical area it is possible to use these values to identify and assign potential boundaries (Fortin and Dale, 2005; Fortin and Drapeau, 1995; Oden et al., 1993). Following the approach outlined in Fortin and Dale, 2005, a threshold value (or percentile class) can be set and will determine what proportion of cells will be retained as potential boundaries.

0.3. Chapter summaries

0.3.1. Chapter 1: A roadmap for predicting ecological networks

Chapter 1 maps out a series of questions and considerations with regards to approaching the challenge of predicting species interactions across space and time. This chapter represents a scoping overview of the 'gap-filling' portion of this thesis and strongly focuses on the idea of networks as predictable objects. Section 1.3 starts by outlining the challenges that might limit our ability to predict networks (which is primarily due to a limitation in data) but also looks into the opportunities we have to try and circumvent or overcome these limitations. Overall this section highlights the fact that one of the most limiting factors for prediction (a lack of data) is also the reason we need predictive tools (to overcome the lack of data), and that we are (methodologically and computationally) in a place where we can start to make feasible predictions, particularly if we think about combining different data sources.

Section 1.2 does exactly this by providing a proof-of-concept showcasing the use of co-occurrence and known interactions to predict novel interactions. This is done using the Hadfield et al., 2014 dataset, which describes 51 host-parasite networks sampled across space. Essentially this showcases how we can extract features for each species based on co-occurrence, use said features to train an artificial neural network to predict interactions, and apply this classifier to the original features to predict potential interactions across the entire species pool. This framework essentially allows us to 'correct' any false negatives (interactions recorded as missing but are actually plausible) within the existing data. This is particularly meaningful as interactions intrinsically vary across space and time, and given the number of species that compose ecological communities, it can be tough to distinguish between a true negative (where two species never interact) from a false negative (where two species have not been observed interacting even though they actually do).

The final part of this manuscript (section 1.4) aims to provide a practical overview of different components to think about and take into consideration when wanting to predict networks. This body of work is intended to be something that can be taken and used as

a brief primer, and as such focuses on discussing some fundamental ideas and concepts. This includes breaking down aspects of the modelling process, aspects of species interaction networks (and the interactions within them), and predicting networks over space and time. As a whole, this chapter really serves to sketch out the 'nuts and bolts' of wanting to take on the task of network prediction and serve as a useful roadmap for those wishing to find a more practical and attainable approach to addressing the global interaction data shortage.

0.3.2. Chapter 2: Graph embedding for network prediction

This chapter should be viewed as a prospective companion piece to the more 'tangible' methods presented in chapter 3, and is strongly rooted in the realm of thinking about network prediction. This chapter is still related to the idea of transfer learning for network prediction, however it focuses more on the expanding how we think about (and can use) a metaweb, as well as potential (alternative) graph embedding methodologies. This chapter thus pushes forward the 'we need gap-filling methods' agenda of this thesis and helps in providing a larger discussion as to the alternative ways we can modify and approach the transfer learning framework from chapter 3.

Section 2.2 provides a discussion on how we can push the original definition of a metaweb developed by Dunne, 2006 in a new 'prediction friendly' direction. As the term 'metaweb' has been used multiple times it is perhaps useful to define it with regards to its original function – to act as an inventory of all possible interactions for a given community. This means that as a concept a metaweb is a realistic, and attainable object to try and predict, however, it is beneficial to move away from thinking of the interactions in a metaweb as binary and rather define the interactions as Bernoulli events. Fundamentally this will allow us greater flexibility in how we weight rare interactions, provide a more nuanced overview of how the community is actually interacting, or factor in a sense of 'uncertainty' into our predictions.

Section 2.3 provides a more detailed overview and discussion of how graph embedding works and why it is a useful way to approach network prediction. This section also includes

examples of different graph embedding techniques, and (where possible) their applications to network ecology. The fundamental argument in favour of using network embedding for network prediction is that it is capturing elements of the *structure* of a network (as opposed to pairwise learning of species a eats b) and thus provides a more powerful abstraction of a network that can be used for predicting networks for other, non-related, communities. There is also an illustration of the embedding process, which is discussed in section 2.4 and acts as a way to showcase how the embedding process captures ecological processes. A more in-depth tutorial/breakdown of the analysis can be accessed through Appendix B.

The final section (subsection 2.5.1) is more focused on the limitations and scope of network embedding and prediction. There is a particular focus on the limitations of taxonomic overlap and the need for 'just the right amount' of species to be shared between known and target communities. There is of course also the challenge of political scale and how the construction of metawebs are at regional scales that may not be ecologically relevant (but are relevant for policy making). Although we are not able to confidently provide a solution for this problem (as we do not even know what an 'ecologically relevant scale' is) it is still important to think about and grapple with these topics.

0.3.3. Chapter 3: Prediction in action: The Canadian Metaweb

Building on the ideas in chapter 1, work on the use of transfer learning for predicting *de novo* interactions (Runghen et al., 2021), and the applicability of phylogenetic reconstruction within the context of ecological networks *e.g.,* (Braga et al., 2021), we set out to create a probabilistic metaweb for terrestrial Canadian mammals in chapter 3. 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. A key premise of this chapter is the idea of being able to take the information that we do have and bring it with us to predict networks in an area where we have no information. This is fundamentally a chance to 'put our money where our mouth is' and provide a *tangible* way

to approach (and round out) the gap-filling portion of this thesis. Specifically, this framework allows us to 'learn' the information (latent traits) of species from a known interaction network (in this case, the European metaweb) and infer the latent traits of another species pool for which we have no *a priori* interaction data (in this case Canadian species) based on their phylogenetic relatedness to species from the known network (see section 0.2.1 for a more detailed summary of the methodology).

Using the prediction of the Canadian metaweb as a way to test the methodology presented in this chapter is useful as we have existing datasets with which to test the validity of our predictions. What is perhaps most exciting about this chapter is that despite sharing about only 4% of species between Canada and Europe we were able to construct a metaweb that correctly predicted about 91% of the species interactions in Canada. It should also be noted that when comparing the European and Canadian metawebs we see a difference in their structures (section A.3), implying that the embedding process is not 'copy pasting' the European network and filling in Canadian species but rather capturing an ecological process.

In addition to testing the validity of the predicted interactions within the Canadian metaweb we also did some additional tests using just the European metaweb. In this instance interactions within the European network were modified (either removed or new interactions were added) and the modified network was used to predict a 'new' European network. This allowed us to compare how well the model could recover the original network despite being 'given' erroneous information. Overall the model is robust to both the addition as well as removal of interactions, although the removal of interactions does have a more negative effect on the ability of the model to recover interactions (subsection A.1.3)

Overall it appears that the transfer learning framework presented in this chapter is quite robust and has potential applicability in a variety of settings (*e.g.,* generating metawebs that can be used as 'informative priors' from which more localised/spatially explicit networks can be constructed, Cirtwill et al., 2019), can be given to a local expert for more refined validation, and overall presents a potential mechanism to begin filling in the global gaps.

0.3.4. Chapter 4: SVD entropy: a measure of network complexity

In chapter 4 we present SVD entropy as a starting point to unifying (and standardising) how we define the complexity of ecological networks. In the perspective of 'global questions about patterns' this of course presents a way in which we can ask a simple question - do different networks (in the case of this chapter different bipartite networks) have differences in their complexity. What makes SVD entropy a compelling metric for quantifying complexity (when comparing to the more 'standard', structural measures such as nestedness, connectance, and spectral radius) is that it focuses more on the 'physical' complexity of the network as opposed to the complexity of the behaviour of the system. This is because the structural measures of complexity are capturing an emerging property of the network, whereas SVD entropy captures the information contained in the the network (one can think of this as the 'compressibility' of the network, more complex networks are harder to compress).

The primary take away from this chapter is that (at least bipartite networks) are exceptionally complex. In subsection 4.3.1 we can see that networks have a relative rank deficiency of zero (*i.e.,* they have a maximal 'external complexity') and all networks have an SVD entropy value greater than 0.80, *i.e.,* near maximal 'internal complexity' (for context the way that SVD entropy is calculated means that values are constrained between zero and one). In subsection 4.3.2 we also looked at the corresponding connectance, nestedness, and spectral radius of these networks . Although there is a correlation between the calculated entropy and these other metrics the story that is told by the different metrics is different. Namely, for the structural metrics the 'complexity' spans the entire potential range of of values, and there is the potential for 'misinterpreting' what could be considered complex. For example networks that have a maximal nestedness have the lowest SVD entropy (*i.e.,* the lowest physical complexity), this is not necessarily the most intuitive way to interpret a maximal nestedness. This 'breakdown' of what complexity means is also echoed in subsection 4.3.3, here we simulated extinctions to get a measure of network resilience (since a common adage is that complexity begets stability), however we do not see a strong relationship between

SVD entropy and resilience. This again highlights that 'structural' and 'physical' complexity metrics are capturing different facets of a network, and although it is not to say that SVD entropy is a 'better' way to measure the complexity of a network it does highlight that we need to be mindful of how we are defining 'complexity' and particularly how that might impact on how we interpret results based on the complexity of networks.

An additional interesting result discussed in subsection 4.3.6 is that although the complexity of ecological networks is indeed *immense* they are still not reaching their *maximum* potential complexity, which implies that *something* might be constraining network complexity. This result is echoed in subsection 4.3.5, which looks at the relationship between network size, connectance and complexity. Results point to the potential constraint of network size on complexity. One possible explanation is that networks at the early assembly stages tend to be severely constrained (Barbier et al., 2018; Saravia et al., 2018) due to conditions needed for the persistence of multiple species. As networks grow larger, these constraints may "relax", leading to networks with more redundancy, and therefore a lower complexity.

0.3.5. Chapter 5: SpatialBoundaries.jl: a software for boundary detection

In this chapter we present a `Julia` package `SpatialBoundaries.jl`, (the documentation is available here) that has the functionality to implement the spatial wombling algorithm across both a uniform landscape *i.e.,* lattice wombling as well as irregular/random landscapes *i.e.,* triangulation wombling. These two methods still calculate the rate of change (m) and directionality (θ) in the same manner but differ in how the aggregate and quantify the surface for a set of points (Fortin and Dale, 2005). These two algorithms provide functionality for most use cases when data are quantitative. `SpatialBoundaries.jl` has also been developed so as to integrate with other packages such as `SimpleSDMLayers.jl`.

Overall this chapter is 'simple' in its content (a software package that can implement the spatial wombling algorithm) however it has been developed with the forward-scoping idea of being used within the context of thinking about boundaries between networks (or if they

are even present) and thus aligns well with the idea of developing tools for understanding global/large scale network patterns. Some ideas for implementing this package are presented in Appendix C, and primarily rest on the idea of using a combination of a metacommunity model and simulated landscapes to see if networks, species, and environmental boundaries show a high degree of fidelity or not. The work presented in Appendix C should be treated as a speculative outline of what we can do with the `SpatialBoundaries.jl` in the context of network analysis and could be viewed as an rough first draft on trying to understand 'where networks stop?', which echoes one of the challenges discussed in chapter 2, particularly in subsection 2.5.2.

0.4. Conclusion

As a whole this thesis should be viewed as a computational toolbox for network ecology that addresses both the issue of data scarcity through the use of predictive tools (addressing the 'Eltonian shortfall' highlighted by Hortal et al., 2015) as well as presenting methods/ideas geared towards thinking about networks at global scales. This means that we would *i*) have 'tangible' networks from which we can begin to work with in various contexts or situations and *ii*) have new methods/tools to begin asking questions about networks at a global scale. In other words adding more building blocks from which we can begin to take network ecology to the next level, *i.e.,* bridging the gap from 'local-level network understanding' to 'tools for global network analysis'.

References

Albrecht, J., Classen, A., Vollstädt, M. G. R., Mayr, A., Mollel, N. P., Schellenberger Costa, D., Dulle, H. I., Fischer, M., Hemp, A., Howell, K. M., Kleyer, M., Nauss, T., Peters, M. K., Tschapka, M., Steffan-Dewenter, I., Böhning-Gaese, K., & Schleuning, M. (2018). Plant and animal functional diversity drive mutualistic network assembly across an elevational gradient. *Nature Communications*, *9*(1), 3177. https://doi.org/10.1038/s41467-018-05610-w

Baiser, B., Gravel, D., Cirtwill, A. R., Dunne, J. A., Fahimipour, A. K., Gilarranz, L. J., Grochow, J. A., Li, D., Martinez, N. D., McGrew, A., Poisot, T., Romanuk, T. N., Stouffer, D. B., Trotta, L. B., Valdovinos, F. S., Williams, R. J., Wood, S. A., & Yeakel, J. D. (2019). Ecogeographical rules and the macroecology of food webs. *Global Ecology and Biogeography*, *28*(9), 1204–1218. https://doi.org/10.1111/geb.12925

Barbier, M., Arnoldi, J.-F., Bunin, G., & Loreau, M. (2018). Generic assembly patterns in complex ecological communities. *Proceedings of the National Academy of Sciences*, 201710352. https://doi.org/10.1073/pnas.1710352115

Barbujani, G., Oden, N. L., & Sokal, R. R. (1989). Detecting Regions of Abrupt Change in Maps of Biological Variables. *Systematic Zoology*, *38*(4), 376–389. https://doi.org/10.2307/2992403

Bartomeus, I., Gravel, D., Tylianakis, J. M., Aizen, M. A., Dickie, I. A., & Bernard-Verdier, M. (2016). A common framework for identifying linkage rules across different types of interactions. *Functional Ecology*, *30*(12), 1894–1903. https://doi.org/10.1111/1365-2435.12666

Becker, D. J., Albery, G. F., Sjodin, A. R., Poisot, T., Dallas, T. A., Eskew, E. A., Farrell, M. J., Guth, S., Han, B. A., Simmons, N. B., & Carlson, C. J. (2020). Predicting wildlife hosts of betacoronaviruses for SARS-CoV-2 sampling prioritization. *bioRxiv*, 2020.05.22.111344. https://doi.org/10.1101/2020.05.22.111344

Bennett, A. E., Evans, D. M., & Powell, J. R. (2019). Potentials and pitfalls in the analysis of bipartite networks to understand plant–microbe interactions in changing environments. *Functional Ecology*, *33*(1), 107–117. https://doi.org/10.1111/1365-2435.13223

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

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

Cirtwill, A. R., Eklöf, 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*(ja). https://doi.org/10.1111/2041-210X.13180

Dale, M., & Fortin, M.-J. (2010). From Graphs to Spatial Graphs. *Annual Review of Ecology, Evolution, and Systematics*, *41*, 21–38.

Davies, T. J. (2021). Ecophylogenetics redux. *Ecology Letters*, *n/a*. https://doi.org/10.1111/ele.13682

Delaunay, B. (1934). Sur la sphere vide. *Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques*, *6*, 793–800.

Delmas, E., Besson, M., Brice, M.-H., Burkle, L. A., Dalla Riva, G. V., Fortin, M.-J., Gravel, D., Guimarães, P. R., Hembry, D. H., Newman, E. A., Olesen, J. M., Pires, M. M., Yeakel, J. D., & Poisot, T. (2018). Analysing ecological networks of species interactions. *Biological Reviews*, 112540. https://doi.org/10.1111/brv.12433

Desjardins-Proulx, P., Poisot, T., & Gravel, D. (2019). Artificial Intelligence for Ecological and Evolutionary Synthesis. *Frontiers in Ecology and Evolution*, *7*. https://doi.org/10.3389/fevo.2019.00402

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.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, *1*(3), 211–218. https://doi.org/10.1007/BF02288367

Elmasri, M., Farrell, M. J., Davies, T. J., & Stephens, D. A. (2020). A hierarchical Bayesian model for predicting ecological interactions using scaled evolutionary relationships. *The Annals of Applied Statistics*, *14*(1), 221–240. https://doi.org/10.1214/19-AOAS1296

Fitzpatrick, M. C., Preisser, E. L., Porter, A., Elkinton, J., Waller, L. A., Carlin, B. P., & Ellison, A. M. (2010). Ecological boundary detection using Bayesian areal wombling. *Ecology*, *91*(12), 3448–3455. https://doi.org/10.1890/10-0807.1

Forsythe, G., & Moler, C. (1967). *Computer Solution of Linear Algebraic Systems*. Prentice Hall.

Fortin, M.-J., & Dale, M. R. T. (2005). *Spatial analysis: A guide for ecologists*. Cambridge University Press.

Fortin, M.-J., & Drapeau, P. (1995). Delineation of Ecological Boundaries: Comparison of Approaches and Significance Tests. *Oikos*, *72*(3), 323–332. https://doi.org/10.2307/3546117

Golub, G. H., Hoffman, A., & Stewart, G. W. (1987). A generalization of the Eckart-Young-Mirsky matrix approximation theorem. *Linear Algebra and its Applications*, *88-89*, 317–327. https://doi.org/10.1016/0024-3795(87)90114-5

Golub, G. H., & Reinsch, C. (1971). Singular value decomposition and least squares solutions. In *Linear Algebra* (pp. 134–151). Springer.

Gómez, J. M., Verdú, M., & Perfectti, F. (2010). Ecological interactions are evolutionarily conserved across the entire tree of life. *Nature*, *465*(7300), 918–921. https://doi.org/10.1038/nature09113

Gu, R., & Shao, Y. (2016). How long the singular value decomposed entropy predicts the stock market? — Evidence from the Dow Jones Industrial Average Index. *Physica A: Statistical Mechanics and its Applications*, *453*(100), 150–161.

Hackett, T. D., Sauve, A. M. C., Davies, N., Montoya, D., Tylianakis, J. M., & Memmott, J. (2019). Reshaping our understanding of species' roles in landscape-scale networks. *Ecology Letters*, *22*(9), 1367–1377. https://doi.org/10.1111/ele.13292

Hadfield, J. D., Krasnov, B. R., Poulin, R., & Nakagawa, S. (2014). A Tale of Two Phylogenies: Comparative Analyses of Ecological Interactions. *The American Naturalist*, *183*(2), 174–187. https://doi.org/10.1086/674445

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

Jacquet, C., Moritz, C., Morissette, L., Legagneux, P., Massol, F., Archambault, P., & Gravel, D. (2016). No complexity–stability relationship in empirical ecosystems. *Nature Communications*, *7*, 12573. https://doi.org/10.1038/ncomms12573

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

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

Landi, P., Minoarivelo, H. O., Brännström, Å., Hui, C., & Dieckmann, U. (2018). Complexity and stability of ecological networks: A review of the theory. *Population Ecology*, *60*(4), 319–345. https://doi.org/10.1007/s10144-018-0628-3

Maiorano, L., Montemaggiori, A., Ficetola, G. F., O'Connor, L., & Thuiller, W. (2020). 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

Oden, N. L., Sokal, R. R., Fortin, M.-J., & Goebl, H. (1993). Categorical Wombling: Detecting Regions of Significant Change in Spatially Located Categorical Variables. *Geographical Analysis*, *25*(4), 315–336. https://doi.org/10.1111/j.1538-4632.1993.tb00301.x

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

Philibert, M. D., Fortin, M.-J., & Csillag, F. (2008). Spatial structure effects on the detection of patches boundaries using local operators. *Environmental and Ecological Statistics*, *15*(4), 447–467. https://doi.org/10.1007/s10651-007-0061-9

Pielou, E. C. (1975). *Ecological diversity*. Wiley.

Pinheiro, R. B. P., Felix, G. M. F., Bell, J. A., & Fecchio, A. (2023). The latitudinal specialization gradient of bird–malarial parasite networks in South America: Lower connectance, but more evenly distributed interactions towards the equator. *Ecography*, *n/a*, e06763. https://doi.org/10.1111/ecog.06763

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*, jbi.14127. 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). Imputing the mammalian virome with linear filtering and singular value decomposition. *arXiv:2105.14973 [q-bio]*.

Pomeranz, J. P., Thompson, R. M., Poisot, T., & Harding, J. S. (2018). Inferring predator-prey interactions in food webs. *Methods in Ecology and Evolution*, *0*(ja). https://doi.org/10.1111/2041-210X.13125

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

Saravia, L. A., Marina, T. I., Troch, M. D., & Momo, F. R. (2018). Ecological Network assembly: How the regional meta web influence local food webs. *bioRxiv*, 340430. https://doi.org/10.1101/340430

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

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

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

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

Thuiller, W., Calderon-Sanou, I., Chalmandrier, L., Gaüzere, P., Ohlmann, M., Poggiato, G., & Münkemüller, T. (2023). Navigating the integration of Biotic Interactions in Biogeography. *Journal of Biogeography*. https://doi.org/10.1111/jbi.14734

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.

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

Windsor, F. M., van den Hoogen, J., Crowther, T. W., & Evans, D. M. (2023). Using ecological networks to answer questions in global biogeography and ecology. *Journal of Biogeography*, *50*(1), 57–69. https://doi.org/10.1111/jbi.14447

Womble, W. H. (1951). Differential Systematics. *Science*, *114*(2961), 315–322. https://doi.org/10.1126/science.114.2961.315

Yan, S., Xu, D., Zhang, B., & Zhang, H.-J. (2005). Graph embedding: A general framework for dimensionality reduction. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, *2*, 830–837 vol. 2. https://doi.org/10.1109/CVPR.2005.170

Chapter 1 First article

A roadmap towards predicting species interaction networks (across space and time)

by

Tanya Strydom[1], Michael D. Catchen[2], Francis Banville[3], Dominique Caron[4], Gabriel Dansereau[5], Philippe Desjardins-Proulx[6], Norma R. Forero-Muñoz[7], Gracielle Higino[8], Benjamin Mercier[9], Andrew Gonzalez[10], Dominique Gravel[11], Laura Pollock[12], and Timothée Poisot[13]

([1]) Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([2]) McGill University, Montréal, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([3]) Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada
Université de Sherbrooke, Sherbrooke, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([4]) McGill University, Montréal, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([5]) Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([6]) Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([7]) Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([8]) Universidade Federal de Goiás, Goiâna, Brasil

([9]) Université de Sherbrooke, Sherbrooke, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([10]) McGill University, Montréal, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([11]) Université de Sherbrooke, Sherbrooke, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([12]) McGill University, Montréal, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([13]) Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

This article was submitted in Philosophical Transactions of the Royal Society B, and can be accessed at https://doi.org/10.1098/rstb.2021.0063.

The main contributions of Tanya Strydom for this articles are presented. All authors contributed to the drafting, writing and editing of the manuscript.

Résumé. Les réseaux d'interactions entre espèces sous-tendent de nombreux processus écosystémiques, mais il est difficile d'échantillonner de manière exhaustive ces interactions. Les interactions varient intrinsèquement dans l'espace et dans le temps, et étant donné le nombre d'espèces qui composent les communautés écologiques, il peut être difficile de distinguer un vrai négatif (dans lequel deux espèces n'interagissent jamais) d'un faux négatif (dans lequel deux espèces n'ont même pas été observées en interaction). bien qu'ils le fassent réellement). Évaluer la probabilité d'interactions entre espèces est un impératif pour plusieurs domaines de l'écologie. Cela signifie que pour prédire les interactions entre les espèces – et pour décrire la structure, la variation et l'évolution des réseaux écologiques qu'elles forment – nous devons nous appuyer sur des outils de modélisation. Nous fournissons ici une preuve de concept, dans laquelle nous montrons comment un simple modèle de réseau neuronal fait des prédictions précises sur les interactions entre espèces avec des données limitées. Nous évaluons ensuite les défis et les opportunités associés à l'amélioration des prédictions d'interaction et fournissons une feuille de route conceptuelle vers des modèles prédictifs de réseaux écologiques explicitement spatiaux et temporels. Nous concluons par une brève introduction aux méthodes et outils pertinents nécessaires pour commencer à construire ces modèles, qui, nous l'espérons, guideront ce programme de recherche.

Mots clés : réseaux écologiques, apprentissage automatique, apprentissage profond, prévisions écologiques, biogéographie

Abstract. Networks of species interactions underpin numerous ecosystem processes, but comprehensively sampling these interactions is difficult. Interactions intrinsically vary across space and time, and given the number of species that compose ecological communities, it can be tough to distinguish between a true negative (where two species never interact) from a false negative (where two species have not been observed interacting even though they actually do). Assessing the likelihood of interactions between species is an imperative for several fields of ecology. This means that to predict interactions between species—and to describe the structure, variation, and change of the ecological networks they form—we need to rely on modelling tools. Here, we provide a proof-of-concept, where we show how a simple neural network model makes accurate predictions about species interactions given limited data. We then assess the challenges and opportunities associated with improving interaction predictions, and provide a conceptual roadmap forward towards predictive models of ecological networks that is explicitly spatial and temporal. We conclude with a brief primer on the relevant methods and tools needed to start building these models, which we hope will guide this research programme forward.

Keywords: ecological networks, machine learning, deep learning, ecological forecasting, biogeography

1.1. Introduction

Ecosystems are, in large part, constructed by the interactions within them — organisms interact with one-another and with their environment, either directly or indirectly. Interactions between individuals, populations, and species create networks of interactions that drive ecological and evolutionary dynamics and maintain the coexistence, diversity, and functioning of ecosystems (Albrecht et al., 2018; Delmas et al., 2018; Landi et al., 2018). Species interaction networks underpin our understanding of numerous ecological processes (Heleno et al., 2014; Pascual and Dunne, 2006). Yet, even basic knowledge of species interactions (like being able to list them, or guess which ones may exist) remains one of the most severe biodiversity shortfalls (Hortal et al., 2015), in large part due to the tedious, time-consuming, and expensive process of collecting species interaction data. Comprehensively sampling every possible interaction is not feasible given the sheer number of species on Earth, and the data we can collect about interactions tend to be biased and noisy (de Aguiar et al., 2019).

This is then compounded as species interactions are typically measured as a binary variable (present or absent) even though it is evident interactions are not all-or-nothing. Empirically we know species interactions occur probabilistically due to variation in species abundances in space and time (Poisot et al., 2015). Different types of interactions vary in their intrinsic predictability (e.g. some fungal species engage in opportunistic saprotrophy Smith et al., 2017, obligate parasites are more deterministic in their interactions than facultative parasites Luong and Mathot, 2019; Poisot et al., 2013). In addition to this variance in predictability, networks from different systems are structured by different mechanisms.

Still, like all of Earth's systems, species interaction networks have entered their "long now" (Carpenter, 2002), where anthropogenic change will have long-term, low-predictability consequences (Burkle et al., 2013) for our planet's ecology. Therefore, our field needs a roadmap towards models that enable prediction (for the present) and forecasting (for the future) of species interactions and the networks they form, and which accounts for their spatial and temporal variation (McCann, 2007; Seibold et al., 2018). As an example, in disease ecology, predicting potential hosts of novel disease (recently notably the search for wildlife hosts of betacoronaviruses; Becker et al., 2020; Wardeh et al., 2021) has received much attention. Network approaches have been used for the prediction of risk and dynamics of dengue (Zhao et al., 2020), Chagas disease (Rengifo-Correa et al., 2017), Rickettsiosis (Morand et al., 2020), Leishmaniasis Stephens, 2009, and a myriad infectious diseases in livestock and wildlife (Craft, 2015). Additionally, prediction of interaction networks is a growing imperative for next-generation biodiversity monitoring, requiring a conceptual framework and a flexible set of tools to predict interactions that is explicitly spatial and temporal in perspective (Edwards et al., 2021; Magioli and Ferraz, 2021; Zhang and He, 2021). Developing better models for prediction of these interactions will rely on integration of data from many sources, and the sources for this data may differ depending on the type of interaction we wish to predict (Gibb et al., 2021).

Interactions between species can be conceptualised in a multitude of ways (mutualistic vs. antagonistic, strong vs. weak, symmetric vs. asymmetric, direct vs. indirect) (Jordano,

2016a; Morales-Castilla et al., 2015). What is common among definitions of species interactions is that *at least* one of the species is affected by the presence of another (Morales-Castilla et al., 2015). Networks can be used to represent a variety of interaction types, including: *unipartite networks*: where each species can be linked to other species (often food webs), *bipartite networks*: where there are two pools of species and all interactions occur between species in each pool (typically used for pairwise interactions; e.g. hosts and parasites), and *k-partite networks,*: which expand to more than two discrete sets of interacting species (e.g., some parasitoid webs, seed dispersal networks, and pollination networks; Pocock et al., 2012).

Methods for predicting interactions between species exist, but at the moment are difficult to generalise as they are typically based around a single mechanism at a single scale: position in the trophic niche (Gravel et al., 2013; Petchey et al., 2008), phylogenetic distance (Elmasri et al., 2020; Pomeranz et al., 2018), functional trait matching (Bartomeus et al., 2016), interaction frequency (Vázquez et al., 2005; Weinstein and Graham, 2017), or other network properties (Stock et al., 2017; Terry and Lewis, 2020). Species interaction networks, as we observe them on Earth today, are the product of ecological and evolutionary mechanisms interacting across spatial, temporal and organisational scales. The interwoven nature of these processes imposes structure on biodiversity data which is invisible when examined only through the lens of a single scale, however machine learning (ML) methods have enormous potential to find this structure in data (Desjardins-Proulx et al., 2019), and have the potential to be used together with mechanistic models in order to make prediction of ecological dynamics more robust (Rackauckas et al., 2020).

Here we use a case study to show how machine-learning models (specifically a deep neural network) can enable prediction of species interactions: we construct a metaweb of host-parasite interactions across space, using predictors extracted from empirical data and accounting for the structure of co-occurrence between species. We use this case study to illustrate a roadmap for improving predictions using open data and ML methods; specifically, we focus on how emerging tools from ML can be used to deliver more accurate and more efficient predictions of ecological systems, and how the potential of these approaches will

be magnified with increased data access. We then provide a non-exhaustive primer on the literature on interaction prediction, and identify the tools and methods most suited for the future of interaction network prediction models, covering the spatial, temporal, and climatic dimensions of network prediction (Burkle and Alarcon, 2011). Both the case study and primer are largely geared towards binary (interactions are either present or absent) networks; there are limitations in data and tools that make it a more reasonable starting approach. First, most ecological networks do not have estimates of interaction strength, and particularly not estimates that are independent from relative abundances. Second, the methodological toolkit to analyse the structure of networks is far more developed for binary interactions (Delmas et al., 2018), meaning that the predictions of binary interactions can be more readily interpreted.

We argue that adopting a more predictive approach to complex ecological systems (like networks) will establish a positive feedback loop with our understanding of these systems (Houlahan et al., 2017): the tasks of understanding and predicting are neither separate nor opposed (Maris et al., 2017); instead, ML tools have the ability to capture a lot of our understanding into working assumptions, and comparing predictions to empirical data gives us better insights about how much we ignore about the systems we model (see for example Borowiec et al., 2021, who provide an overview of deep learning techniques and concepts in ecology and evolution). Although data on species interaction networks are currently limited in the size and spatial coverage, machine learning approaches have a demonstrated track record of revealing the "unreasonable effectiveness" of data (Halevy et al., 2009); we argue that with a clear roadmap guiding the use of these methods, the task of predicting species interaction networks will become more attainable.

1.2. A case study: deep learning of spatially sparse host-parasite interactions

The premise of this manuscript is that we can predict interactions between species. In this section we provide a proof-of-concept, where we use data from Hadfield et al., 2014

describing 51 host-parasite networks sampled across space. In this data, as in most spatially distributed ecological networks, not all species co-occur across sites. As a direct consequence there are pairs of species that may or may not be able to interact for which we have no data; furthermore there are pairs of species that may interact, but have only been documented in a single location where the interaction was not detected. In short, there are ecological reasons to believe that a number of negative associations in the metaweb (*sensu* Dunne, 2006) are false negatives.

Without any species-level information, we resort to using both co-occurrence and known interactions to predict novel interactions. To do this we (i) extract features (equivalent to explanatory variables in a statistical model) for each species based on co-occurrence, (ii) use these features to train an artificial neural network to predict interactions, and (iii) apply this classifier (an algorithm that assigns a categorical output based on input features) to the original features to predict potential interactions across the entire species pool. Machine learning relies on a lexicon that shares some terms with statistics, albeit with different meaning; we expand on the precise meanings in the "How to validate a predictive model" section below. The outputs of the analysis are presented in Figure 1, and the code to reproduce it is available at `https://osf.io/6jp4b/`; the entire example was carried out in `Julia 1.6.2` (Bezanson et al., 2017), using the *Flux* machine learning framework (Innes, 2018).

We first aggregate all species into a co-occurrence matrix A which represents whether a given pair of species (i,j) was observed coexisting across any location. We then transform this co-occurrence matrix A via probabilistic PCA (Tipping and Bishop, 1999) and use the first 15 values from this PCA space as the features vector for each species i. For each pair of (host, parasite) species (i,j), we then feed the features vectors (v_i, v_j) into a neural network. The neural network uses four feed-forward layers (each layer is independent from the one before and after); the first layer uses the RELU activation function (which ignores input below a threshold), the rest use a σ function (which transforms linear activation energies

into logistic responses). All layers have appropriate dropout rates (in order to avoid overfitting, only a fraction of the network is updated on each iteration: $1 - 0.8$ for the first layer, $1 - 0.6$ for the subsequent ones). This produces an output layer with a single node, which is the probability-score for interaction between species i and j.

We then train (equivalent to *fit*) this neural network by dividing the original dataset into testing and training sets (split 80-20 for training and testing respectively). During the training of this neural network (using the ADAM optimiser), the 5×10^4 batches of 64 items used for training were constrained to have at least 25% of positive interactions, as (Poisot, Ouellet, et al., 2021) show slightly inflating the dataset with positive interactions enables us to counterbalance sampling biases. Furthermore, setting a minimum threshold of response balance is an established approach for datasets with strong biases (Lemaître et al., 2017). Validating this model on the test data shows our model provides highly effective prediction of interactions between pairs of species not present in the training data (Figure 1). The behaviour of the model was, in addition, checked by measuring the training and testing loss (difference between the actual value and the prediction, here using mean-squared error) and stopping well before they diverged (to avoid overfitting).

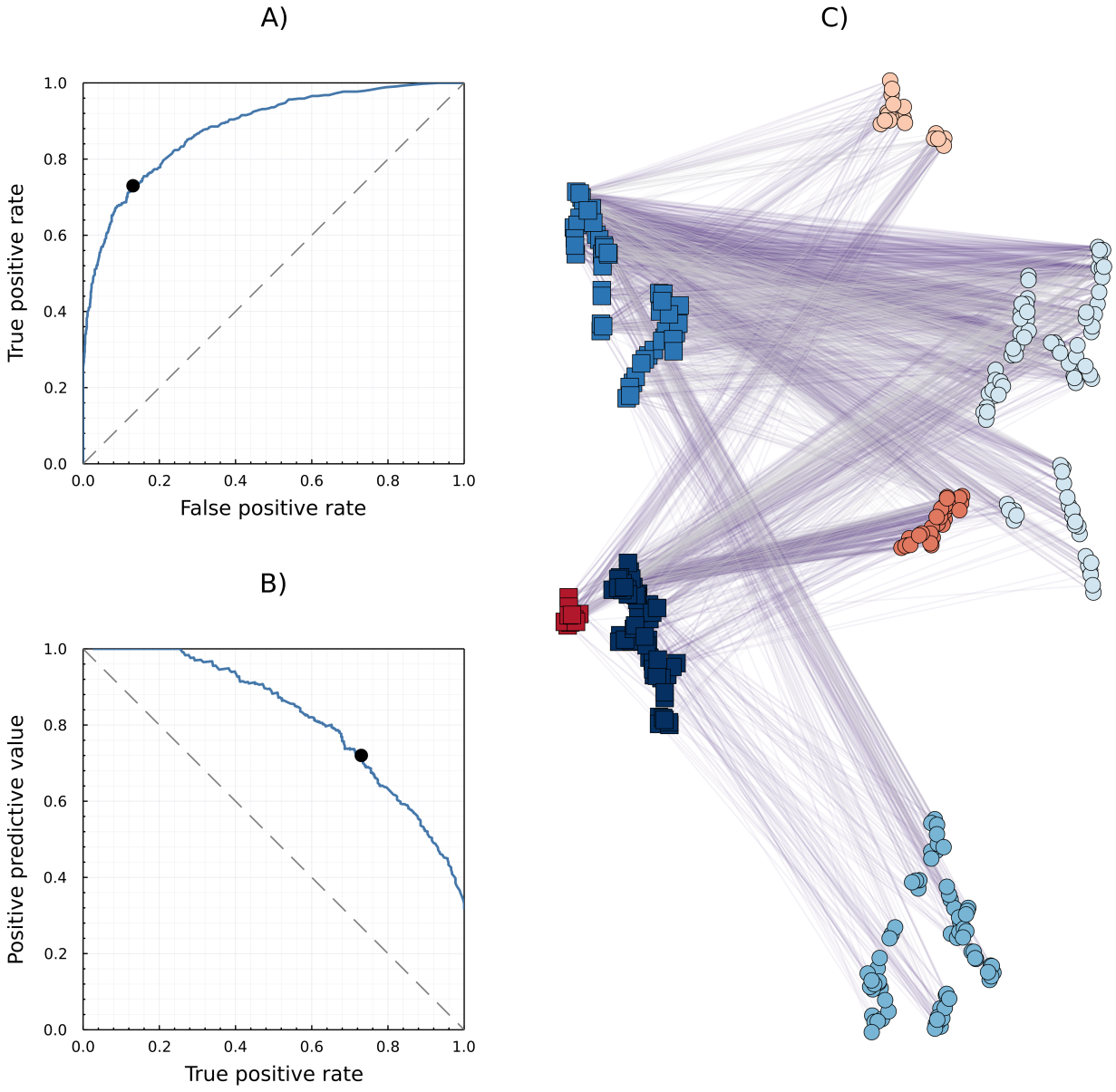


Fig. 1. Proof-of-Concept: An empirical metaweb (from Hadfield et al., 2014, i.e. a list of known possible interactions within a species pool, is converted into latent features using probabilistic PCA, then used to train a deep neural network to predict species interactions. Panels A and B represent, respectively, the ROC curve and the precision-recall curve, with the best classifier (according to Youden's J) represented by a black dot. The expected performance of a neutral "random-guessing" classifier is shown with a dashed line. Panel C shows the imputed using t-distributed stochastic neighbour embedding (tSNE), and the colours of nodes are the cluster to which they are assigned based on a k-means clustering of the tSNE output. Empirical interactions are shown in purple, and imputed interactions in grey.

This case study shows that a simple neural network can be very effective in predicting species interactions even without additional species-level data. Applying this model to the entire dataset (including species pairs never observed to co-occur) identified 1546 new possible interactions – 746 (48%) of which were between pairs of species for which no co-occurrence was observed in the original dataset. This model reaches similar levels of predictive efficacy as previous studies that use far more species-level data and mechanistic assumptions (Gravel et al., 2013), which serves to highlight the potential for including external sources of data for *improving* our prediction of interaction networks even further. For example, Krasnov et al., 2016 collected traits data for this system that could be added to the model, in addition or in substitution to latent variables derived from observed interactions.

1.3. Predicting species interaction networks across space: challenges and opportunities

Here we present a conceptual roadmap (Figure 2) which shows a conceptual path from data to prediction of species interaction networks, incorporating several modelling frameworks. We envisage this roadmap to be one conceptual path toward incorporating space in to our prediction of interaction networks, and developing spatially explicit models of networks and their properties. In the following sections we discuss the challenges and opportunities for this path forward, and highlight two specific areas where it can have a strong impact: the temporal forecasting of species interaction networks structure, and the use of predicted networks for applied ecology and conservation biology.

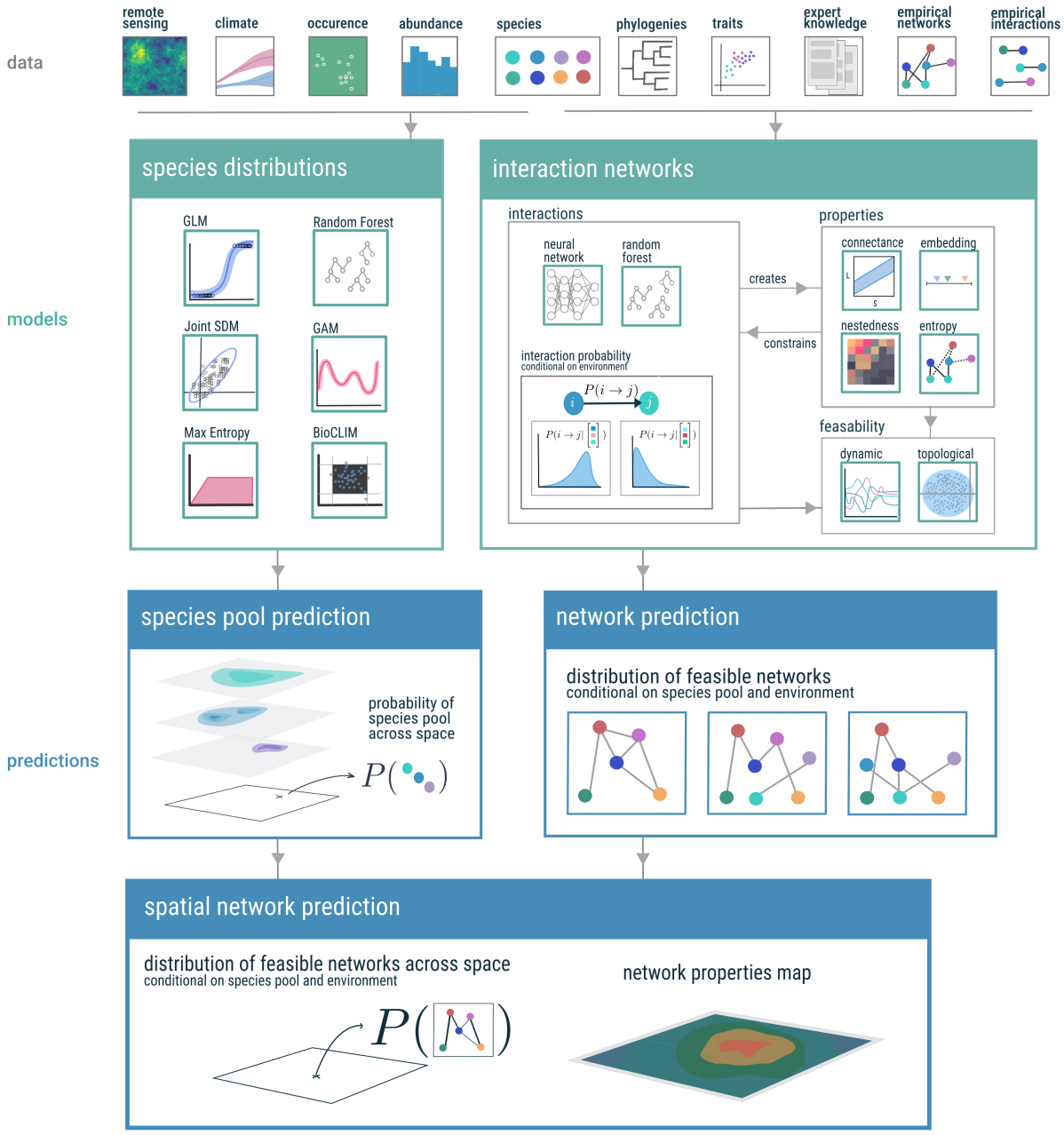


Fig. 2. A conceptual roadmap highlighting key areas for the prediction of ecological networks. Starting with the input of data from multiple sources, followed by a modelling framework for ecological networks and the landscape, which are then ultimately combined to allow for the prediction of spatially explicit networks.

1.3.1. Challenges: constraints on predictions

1.3.1.1. Ecological network data are scarce and hard to obtain. At the moment, prediction of species interactions is made difficult by the limited availability of data. Although we have seen a growth in species occurrence data, this growth is much slower for ecological interactions because species interactions are challenging to sample comprehensively (Bennett et al., 2019; Jordano, 2016b) and sampling methodology has strong effects on the resulting data (de Aguiar et al., 2019). In turn, the difficulty of sampling interactions can lead to biases in our understanding of network structure (de Aguiar et al., 2019). This knowledge gap has motivated a variety of approaches to deal with interactions in ecological research based on assumptions that do not always hold, such as the assumption that co-occurrence is equivalent to meaningful interaction strength (Blanchet et al., 2020). Spatial biases in data coverage are prevalent at the global scale (with South America, Africa and Asia being under-represented) and different interaction types show biases towards different biomes (Poisot, Bergeron, Cazelles, Dallas, Gravel, et al., 2021). These "spatial gaps" serve as a limitation to our ability to confidently make predictions when accounting for real-world environmental conditions, especially in environments for which there are no analogous data.

Further, empirical estimation of interaction *strength* is highly prone to bias as existing data are usually summarised at the taxonomic scale of the species or higher, thereby losing information that differentiates the strength in per-individual interactions from the strength of a whole species interaction (Wells and O'Hara, 2013). Empirical estimations of interaction strength are still crucial (Novak and Wootton, 2008), but are a hard task to quantify in natural communities (Sala and Graham, 2002; Wootton, 1997; Wootton and Emmerson, 2005), especially as the number of species composing communities increases, compounded by the possibility of higher-order interactions or non-linear responses in interactions (Wootton and Emmerson, 2005). Further, interaction strength is often variable and context dependent and can be influenced by density-dependence and spatio-temporal variation in community composition (Wootton and Emmerson, 2005).

1.3.1.2. Powerful predictive tools work better on large data volumes. This scarcity of data limits the range of computational tools that can be used by network ecologists. Most deep learning methods, for instance, are very data expensive. The paucity of data is compounded by a collection of biases in existing datasets. Species interaction data are typically dominated by food webs, pollination, and host-parasite networks (Ings et al., 2009; Poisot et al., 2020). This could prove to be a limiting factor when trying to understand or predict networks of underrepresented interaction types or when trying to integrate networks of different types (Fontaine et al., 2011), especially given their inherent structural variation (Michalska-Smith and Allesina, 2019). This stresses the need for an integrated, flexible, and data-efficient set of computational tools which will allow us to predict ecological networks accurately from existing and imperfect datasets, but also enable us to perform model validation and comparison with more flexibility than existing tools. We argue that Figure 1 is an example of the promise of these tools *even* when facing datasets of small size. The ability to extract and engineer features also serves to bolster our predictive power. Although it may be tempting to rely on approaches like bootstrapping to estimate the consistency of the predictions, we are confronted with the issues of low data volume and data bias—that we are more likely to observe interactions between some pairs of species (*i.e.,* those that co-occur often, e.g. Cazelles et al., 2015, and those with higher relative abundance, e.g. Vazquez et al., 2009). This introduces risk in training models on pseudo-replicated data. In short, the current lack of massive datasets must not be an obstacle to prediction; it is an ideal testing ground to understand how little data is sufficient to obtain actionable predictions, and how much we can rely on data inflation procedures to reach this minimal amount.

1.3.1.3. Scaling-up predictions requires scaled-up data. We are also currently limited by the level of biological organisation at which we can describe ecological networks. For instance, our understanding of individual-based networks (*e.g.,* M. S. Araújo et al., 2008; Tinker et al., 2012) is still in its infancy (Guimarães, 2020) and acts as a resolution-limit. Similarly, the resolution of environmental (or landscape) data also limits our ability to predict networks at small scales, although current trends in remote sensing would suggest that

this will become less of a hindrance with time (Makiola et al., 2020). Ecosystems are a quintessential complex-adaptive-system (Levin, 1998) with a myriad of processes at different spatial, temporal, and organisational scales that influence and respond to one another. Understanding how the product of these different processes drive the properties of ecosystems across different scales remains a central challenge of ecological research, and we should strive to work on methods that will integrate different empirical "snapshots" of this larger system.

1.3.2. Opportunities: an emerging ecosystem of open tools and data

1.3.2.1. Data are becoming more interoperable. The acquisition of biodiversity and environmental data has tremendously increased over the past decades thanks to the rise of citizen science (J. L. Dickinson et al., 2010) and of novel technology (Stephenson, 2020), including wireless sensors (Porter et al., 2005), next-generation DNA sequencing (Creer et al., 2016), and remote sensing (Lausch et al., 2016; Skidmore and Pettorelli, 2015). Open access databases, such as GBIF (for biodiversity data), NCBI (for taxonomic and genomics data), TreeBASE (for phylogenetics data), CESTE (Jeliazkov et al., 2020) (for metacommunity ecology and species traits data), and WorldClim (for bioclimatic data) contain millions of data points that can be integrated to monitor and model biodiversity at the global scale. For species interactions data, at the moment Mangal is the most comprehensive open database of published ecological networks (Poisot et al., 2016), and GloBI is an extensive database of realised and potential species interactions (Poelen et al., 2014). Developing standard practices in data integration and quality control (Kissling et al., 2018) and in next-generation biomonitoring (NGB; Makiola et al., 2020) would improve our ability to make reliable predictions of ecosystem properties on increasing spatial and temporal scales. The advancement of prediction techniques coupled with a movement towards standardising data collection protocols (e.g. Pérez-Harguindeguy et al., 2013 for plant functional traits) and metadata (e.g. DarwinCore)—which facilitates interoperability and integration of datasets—as well as

a growing interest at the government level (Scholes et al., 2012) paints a positive picture for data access and usability in the coming years.

1.3.2.2. Machine learning tools are becoming more accessible. This effort is also supported by a thriving ecosystem of data sources and novel tools. ML methods can often be more flexible and perform better than classical statistical methods, and can achieve a very high level of accuracy in many predictive and classification tasks in a relatively short amount of time (e.g., Cutler et al., 2007; Krizhevsky et al., 2017). Increasing computing power combined with recent advances in machine learning techniques and applications shows promise in ecology and environmental science (see Christin et al., 2019 for an overview). Moreover, ongoing developments in deep learning are aimed at improvement in low-data regimes and with unbalanced datasets (Antoniou et al., 2018; Chawla, 2010). Considering the current biases in network ecology (Poisot, Bergeron, Cazelles, Dallas, Gravel, et al., 2021) and the scarcity of data of species interactions, the prediction of ecological networks will undoubtedly benefit from these improvements. Machine learning methods are emerging as the new standard in computational ecology in general (Christin et al., 2019; Olden et al., 2008), and in network ecology in particular (Bohan et al., 2017), as long as sufficient, relevant data are available. Many studies have used machine learning models specifically with ecological interactions. Relevant examples include species traits used to predict interactions and infer trait-matching rules (Desjardins-Proulx et al., 2017; Pichler et al., 2020), automated discovery of food webs (Bohan et al., 2011), reconstruction of ecological networks using next-generation sequencing data (Bohan et al., 2017), and network inference from presence-absence data (Sander et al., 2017). As many ecological and evolutionary processes underlie species interactions and the structure of their ecological networks (*e.g.,* Segar et al., 2020; Vazquez et al., 2009, it can be difficult to choose relevant variables and model species interactions networks explicitly. A promising application of machine learning in natural sciences is Scientific-Machine Learning (SciML), a framework that combines machine learning with mechanistic models (Chuang and Keiser, 2018; Rackauckas et al., 2020).

1.4. A primer on predicting ecological networks

Within the constraints outlined in the previous section, we now provide a primer on the background concepts necessary to build predictive models of species interaction networks, with a focus on using machine learning approaches in the modelling process. As Figure 2 illustrates, this involves a variety of numerical and computational approaches; therefore, rather than an exhaustive summary, we aim to convey a high-level understanding that translates the core concepts into their application to ecological networks.

1.4.1. Models

1.4.1.1. What is a predictive model? Models are used for many purposes, and the term "model" itself embodies a wide variety of meanings in scientific discourse. All models can be thought of as a function, f, that takes a set of inputs x (also called features, descriptors, or independent variables) and parameters θ, and maps them to predicted output states y (also called label, response, or dependent variable) based on the input to the model: $y = f(x,\theta)$.

A given model f can be used for either descriptive or predictive purposes. Many forms of scientific inquiry are based around using models *descriptively*, a practice also called inference, the inverse problem, fitting a model, or training a model (Stouffer, 2019). In this context, the goal of using a model is to estimate the parameters, θ, that best explain a set of empirical observations, $\{\hat{x}, \hat{y}\}$. In some cases, these parameter values are themselves of interest (e.g., the strength of selection, intrinsic growth rate, dispersal distance), but in others cases, the goal is to compare a set of competing models $f_1, f_2, \ldots$ to determine which provides the most parsimonious explanation for a dataset. The quantitative representation of "effects" in these models—the influence of each input on the output—is often assumed to be linear, and within the frequentist world-view, the goal is often to determine if the coefficient corresponding with an input is non-zero to determine its "significance" (often different from its ecological relevance Martínez-Abraín, 2008) in influencing the outcome.

Models designed for inference have utility—descriptive models of networks can reveal underlying mechanisms that structure ecological communities, given a proper null model

(Connor et al., 2017). However, in order for ecology to develop as a predictive science (Evans et al., 2012), interest has grown in developing models that are used not just for description of data, but also for prediction. Predictive models are based in *the forward problem*, where the aim is to predict new values of the output y given an input x and our estimate value of θ (Stouffer, 2019). Because the forward problem relies on an estimate of θ, then, the problem of inference is nested within the forward problem (Figure 3): working towards a predictive view of ecological networks will give us the needed tools to further our understanding of them.

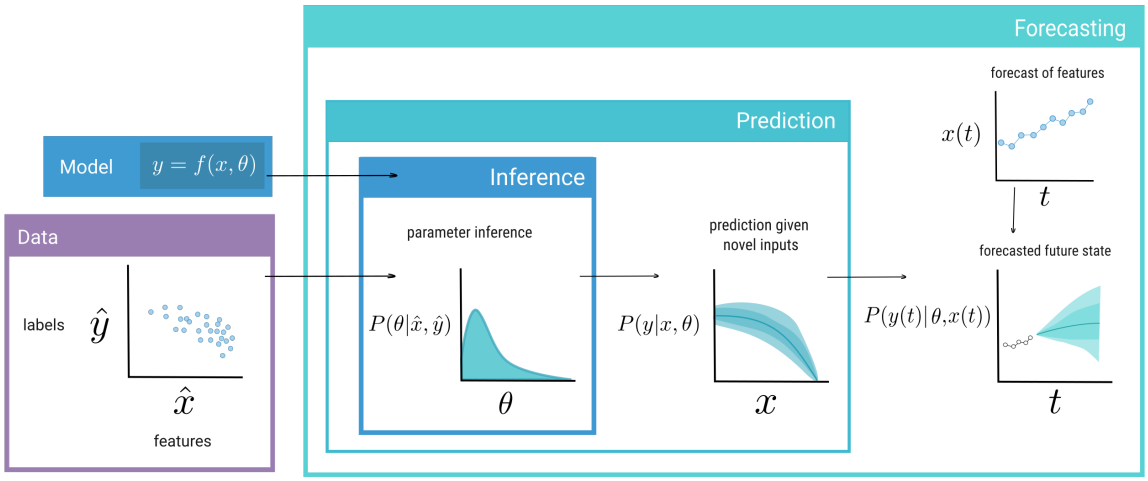


Fig. 3. The nested nature of developing predictive and forecasting models, showcases the *forward problem* and how this relies on a hierarchical structure of the modelling process. The choice of a specific modelling technique and framework, as well as the data retained to be part of this model, proceeds directly from our assumptions about which ecological mechanisms are important in shaping both extant and future data.

1.4.1.2. What do you need to build a predictive model? To build a predictive model, one needs the following: first, **data**, split into features $\hat{x}$ and labels $\hat{y}$ (Figure 3). Second, a **model** f, which maps features x to labels y as a function of parameters θ, i.e. $y = f(x, \theta)$. Third, a **loss function** $L(\hat{y}, y)$, which describes how far a model's prediction y is from an empirical value $\hat{y}$. Lastly, **priors** on parameters, $P(\theta)$, which describe the modeller's *a priori* belief about the value of the parameters; rather than making an analysis implicit, specifying priors has the merit of making the modeller's assumptions explicit, which is a most desirable feature when communicating predictions to stakeholders (Spiegelhalter et al., 2000). Often an important step before fitting a model is feature engineering: adjusting and reworking the features to better uncover feature-label relationships (Kuhn and Johnson, 2019). This can include projecting the features into a lower dimensional space, as we did through a probabilistic PCA in the case study, or removing the covariance structure using a Whitening approach. Then, when a model is fitted (synonymous with parameter inference or the inverse problem, see Figure 3), a fitting algorithm attempts to estimate the values of θ that minimises the mean value of loss function $L(\hat{y},y)$ for all labels $\hat{y}$ in the provided data Y. In a Bayesian approach, this typically relys on drawing candidate parameter values from priors and applying some form of sampling to generate a posterior estimate of parameters, $P(\theta|\hat{x}, \hat{y})$. In the training of neural networks, this usually involves some form of error back-propagation across the edges in order to tune their weights, and the biases of each nodes.

1.4.1.3. How do we validate a predictive model? After we fit a model, we inevitably want to see how "good" (meaning, "fit for purpose") it is. This process can be divided into two parts: (i)) model selection, where the modeller chooses from a set of possible models and (ii) model assessment, where the modeller determines the performance characteristics of the chosen model (Hastie et al., 2009).

In the context of *model selection*, a naïve initial approach is to simply compute the average error between the model's prediction and the true data we have, and choose the model with the smallest error—however this approach inevitably results in *overfitting*. One approach to avoid overfitting is using information criteria (e.g., AIC, BIC, MDL) based

around the heuristic that good models maximise the ratio of information provided by the model to the number of parameters it has. However, when the intended use-case of a model is prediction the relevant form of validation is *predictive accuracy*, which should be tested with *cross-validation*. Cross-validation methods divide the original dataset into two—one which is used to fit the model (called the *training* set) and one used to validate its predictive accuracy on the data that it hasn't "seen" yet (called the *test* set) (Bishop, 2006). This procedure is often repeated across different test and training subdivisions of the dataset (either picked randomly or stratified by some criteria, like balance between positive and negative interactions in the case study) to determine the uncertainty associated with our measurement due to our choice of test and training sets (Arlot and Celisse, 2010), in the same conceptual vein as data bootstrapping: the mean value of the validation metric gives an overall estimate of its performance, and the variance around this mean represents the effect of using different data for training and testing. In a robust model/dataset combination, we expect this variance to be low, although there are no prescriptive guidelines as to how little variance is acceptable; the choice of whether to use a model is often left to the best judgement of the modeller.

We still have to define what *predictive accuracy* means in the context of interaction network prediction. In the proof-of-concept, we used a neural-network to perform binary classification by predicting the presence/absence of an interaction between any two species. There are two ways for the model to be right: the model predicts an interaction and there is one (a *true positive* (TP)), or the model predicts no interaction and there isn't one (a *true negative* (TN)). Similarly, there are two ways for the model to be wrong: the model predicts an interaction which does not exist (a *false positive* (FP)), or the model predicts no interaction but it does exist (a *false negative* (FN)).

A naïve initial approach to measure how well a model does is *accuracy*, i.e. the proportion of values it got correct. However, consider what we know about interaction networks: they are often very sparse, with connectance usually below a third (Cohen et al., 1990). If we build a model that always guesses there will be no interaction between two species, it

will be correct in the majority of cases because the majority of potential interactions in a network typically do not exist. Therefore this "empty-matrix" model would always have an *accuracy* of $1 - C$, where C is the observed connectance, which would almost always be greater than 50%. Understanding model performance within sensitivity-specificity space may be more informative, where sensitivity evaluates how good the model is at predicting true interactions (True Positive Rate) and specificity refers to the prediction of true "non-interactions" (True Negative Rate). It must be noted that in ecological networks, there is no guarantee that the "non-interactions" (assumed true negatives) in the original dataset are indeed true negatives (Jordano, 2016a, 2016b). This can result in the positive/negative values, and the false omission/discovery being artificially worse, and specifically decrease our confidence in predicted interactions.

In response to the general problem of biases in classifiers, many metrics have been proposed to measure binary-classifiers (Drummond and Holte, 2006; Gu et al., 2009) and are indicative of how well the model performs with regards to some aspect of accuracy, sensitivity, specificity and/or precision (Table 1). Ultimately the choice of metric will depend on the intended use of the model: there is not a single definition of "success", but rather different interpretation of what sources of error are acceptable for a given application.

In the machine learning literature, a common way of visualising this extensive list of possible metrics is through the use of ROC (receiver-operating-characteristic; False Positive Rate on the x-axis, and True Positive Rate on the y-axis) and PR (precision-recall; True-Positive-Rate on the x-axis, Positive-predictive-value on the y-axis) curves (see Figure 1). These curves are generated by considering a continuum of thresholds of classifier acceptance, and computing the values of ROC/PR metrics for each value of the threshold. The area-under-the-curve (AUC) is then used as a validation metric and are typically called AUC-ROC (Area-Under-the-Curve Receiver-Operator-Curve) and AUC-PR (Area-Under-the-Curve Precision-Recall) (e.g. ROC-AUC in Table 1). These measures have the unstated assumption that the training and testing set are "correct", or at least correct enough that the number of true/false positive/negatives are meaningful; although should this assumption

Name	Value	Success	Description
Random accuracy	0.56		Fraction of correct predictions if the classifier is random
Accuracy	0.81	$\to 1$	Observed fraction of correct predictions
Balanced accuracy	0.80	$\to 1$	Average fraction of correct positive and negative predictions
True Positive Rate	0.77	$\to 1$	Fraction of interactions predicted
True Negative Rate	0.83	$\to 1$	Fraction of non-interactions predicted
False Positive Rate	0.16	$\to 0$	Fraction of non-interactions predicted as interactions
False Negative Rate	0.22	$\to 0$	Fraction of interactions predicted as non-interactions
ROC-AUC	0.86	$\to 1$	Proximity to a perfect prediction (ROC-AUC=1)
Youden's J	0.60	$\to 1$	Informedness of predictions (trust in individual prediction)
Cohen's κ	0.58	≥ 0.5	
Positive Predictive Value	0.66	$\to 1$	Confidence in predicted interactions
Negative Predictive Value	0.89	$\to 1$	Confidence in predicted non-interactions
False Omission Rate	0.10	$\to 0$	Expected proportion of missed interactions
False Discovery Rate	0.33	$\to 0$	Expected proportion of wrongly imputed interactions

Table 1. Overview of the validation statistics applied to the case study, alongside the criteria indicating a successful classifier and a guide to interpretation of the values. Taken together, these validation measures indicate that the model performs well, especially considering that it is trained from a small volume of data.

be true, there would be no need for any predictive approach – but it is a well established fact that machine learning systems are resilient to even relatively high uncertainties in the data (Halevy et al., 2009).

1.4.2. Networks and interactions as predictable objects

1.4.2.1. Why predict networks and interactions at the same time? Ecological networks are quite sparse, and larger networks tend to get sparser (MacDonald et al., 2020); in other words, although networks are composed of a set of interactions between species pairs, they also form a much larger set of species pairs that do not interact. If we aim to predict the structure of networks from the "bottom-up"— by considering each pairwise combination of S different species—we are left with S^2 interaction values to estimate, a majority of which will be 0. Instead, we can use our existing understanding of the mechanisms that structure ecological networks to whittle down the set of feasible adjacency matrices, thereby reducing the amount of information we must predict, and making the problem of predicting interactions less daunting. The processes that structure ecological networks do not only occur at

the scale of interactions—there are also processes at the network level which limit what interactions (or how many) are realistic. The realised structure of a network is the synthesis of the interactions forming the basis for network structure, and the network structure refining the possible interactions—"Part makes whole, and whole makes part" (Levins and Lewontin, 1987).

Another argument for the joint prediction of networks and interactions is to reduce circularity and biases in the predictions. As an example, models like linear filtering (Stock et al., 2017) generate probabilities of non-observed interactions existing, but do so based on measured network properties. Some recent models make interaction-level predictions (*e.g.,* Gravel et al., 2019); these are not unlike stacked species distribution models, which are individually fit, but collectively outperformed by joint models or rule-based models (Zurell et al., 2020). By relying on adequate testing of model performance of biases (*i.e.,* optimising not only accuracy, but paying attention to measures like false discovery and false omission rates), and developing models around a feedback loop between network and interaction prediction, it is likely that the quality of the predicted networks will be greatly improved compared to current models.

1.4.2.2. What network properties should we use to inform our predictions of interactions? There are many dimensions of network structure (Delmas et al., 2018), yet there are two arguments to support basing network prediction around a single property: *connectance* (the ratio of actual edges to possible edges in the network). First, connectance is ecologically informative—it relates to resilience to invasion (Baiser et al., 2010; Smith-Ramesh et al., 2016), can increase robustness to extinction in food webs (Dunne et al., 2002a), while decreasing it in mutualistic networks (Vieira and Almeida-Neto, 2015), and connectance relates to network stability (Landi et al., 2018). Second, most (if not all) network properties covary with connectance (Dunne et al., 2002b; Poisot and Gravel, 2014).

Within the network science literature, there are numerous methods for predicting edges based on network properties (*e.g.,* block models (Yen and Larremore, 2020) based on modularity, hierarchical models (Kawakatsu et al., 2021) based on embedding, etc.). However, in

the context of species interaction networks, these properties often co-vary with connectance. As a result we suggest that using connectance as the primary property of interest is most likely to be practical to formulate at the moment. We have models to estimate species richness over space (Jenkins et al., 2013), and because we can predict connectance from species richness alone (MacDonald et al., 2020), we can then derive distributions of network properties from richness estimates, that can serve to penalise further models that formulate their predictions at the scale of each possible interaction.

1.4.2.3. How do we predict how species that we have never observed together will interact? A neutral approach to ecological interactions would assume the probability of an interaction to mirror the relative abundance of both species, and would be unaffected by trait variation (Pichler et al., 2020; Poisot et al., 2015); more accurately, a neutral assumption states that the relative abundances are sufficient to predict the structure of networks, and this view is rather well supported in empirical and theoretical systems (Canard et al., 2014; Canard et al., 2012). However, functional-trait based proxies could enable better predictions of ecological interactions (Bartomeus, 2013; Bartomeus et al., 2016; Cirtwill and Eklöf, 2018; Cirtwill et al., 2019). Selection on functional traits could cause interactions to be conserved at some evolutionary scales, and therefore predictions of interaction could be informed by phylogenetic analyses (Davies, 2021; Elmasri et al., 2020; Gómez et al., 2010). Phylogenetic matching in bipartite networks is consistent across scales (Poisot and Stouffer, 2018), even in the absence of strong selective pressure (Coelho et al., 2017).

A separate family of methods are based on network embedding (as in the proof-of-concept). A network embedding projects each node of the network into a lower-dimensional latent space. Previous explorations of the dimensionality of food webs have revealed that a reduced number of dimensions (7) was sufficient to capture most of their structure (Eklöf et al., 2013); however, recent quantifications of the complexity of the embedding space of bipartite ecological networks found a consistent high complexity (Strydom et al., 2021), suggesting that the precise depth of embedding required may vary considerably across systems. Embeddings enables us to represent the structure of a network, which previously required

the S^2 dimensions of an adjacency matrix, with a smaller number of dimensions. The position of each node in this lower dimensional space is then treated as a latent measurement corresponding to the role of that species in the network (*e.g.,* (Poisot, Ouellet, et al., 2021), where a network of about 1500 species was most accurately described using 12 dimensions). Species close together in the latent space should interact with similar set of species (Rohr et al., 2010; Rossberg et al., 2006). However, these models are sensitive to sampling biases as they are limited to species for which there is already interaction data, and as a result a methodological breakthrough is needed to extend these models to species for which there is little or no interaction data.

1.4.2.4. How do we quantify interaction strength? Species interaction networks can also be used as a means to quantify and understand *interaction strength*. Interaction strength, unlike the qualitative presence or absence of an interaction, is a continuous measurement which attempts to quantify the effect of one species on another. This results in weighted networks representing different patterns of 'flows' between nodes – which can be modelled in a variety of ways (Borrett and Scharler, 2019). Interaction strength can generally be divided into two main categories (as suggested by Berlow et al., 2004): 1) the strength of an interaction between individuals of each species, or 2) the effect that changes in one species population has on the dynamics of the other species. It can be measured as the effect over a period of time (in the units of biomass or energy flux Barnes et al., 2018; Brown et al., 2004) or the relative importance of one species on another (Berlow et al., 2004; Heleno et al., 2014; Wootton and Emmerson, 2005). One recurring observation is that networks are often composed of many weak interactions and few strong interactions (Berlow et al., 2004). The distribution of interaction strength within a network effects its stability (de Ruiter et al., 1995; Neutel et al., 2002) and functioning (Duffy, 2002; Montoya et al., 2003), and serves to benefit multi-species models (Wootton and Emmerson, 2005). Alternatively, understanding flow in modules within networks can aid in understanding the organisation of networks (Farage et al., 2021; Montoya and Solé, 2002) or the cascading effects of perturbations (Gaiarsa and Guimarães, 2019).

In some systems, quantifying interaction strength is relatively straightforward; this includes a lot of host-parasite systems. For example, freshwater cyprinid fish can be divided in micro-habitats (fins, skin, digestive system, gill subsections) and the parasites counted in each of these micro-habitats, giving within-host resolution (Simková et al., 2002); marine sparids and labrids have similarly been studied this way, see notably (Desdevises, 2006; Morand et al., 2002; Sasal et al., 1999). In some cases, within-host assessments of interaction strengths can reveal macro-ecological events, like in the conservatism of micro-habitat use in amphibian hosts by helminths (Badets et al., 2011). Even ectoparasites can provide reliable assessments of interaction strength; for example, when rodent hosts are minimally disturbed during capture, fine combing of their fur will result in exhaustive ectoparasites inventories (E. R. Dickinson et al., 2020; Hadfield et al., 2014; Karbowiak et al., 2019; Matthee et al., 2020; Sánchez et al., 2014). Parasites have the desirable property of usually remaining intact within their host during the interaction, as opposed to prey items as can be recovered through *e.g.,* gut content analysis or stable isotopes (Macías-Hernández et al., 2018; Schmid-Araya et al., 2016). As network ecology is starting to explore the use of predictive models, leading up to forecasting, we argue that host-parasite systems can provide data that are reliable and trustworthy enough that they can become the foundations for methodological development and benchmark studies, thereby providing more information about host-parasite systems and supporting the technical development of the field.

Yet in most situations, much like quantifying the occurrence of an interaction, quantifying interaction *strength* in the field is challenging in the majority of systems, and one must often rely on proxies. In some contexts, interaction strength can be estimated via functional foraging (Portalier et al., 2019), where the primary basis for inferring interaction is foraging behaviour like searching, capture and handling times. In food-webs, metabolic based models use body mass, metabolic demands, and energy loss to infer energy fluxes between organisms (Berlow et al., 2009; Yodzis and Innes, 1992). In addition, food-web energetics models can be incorporated at various resolutions for a specific network, ranging from individual-based data to more lumped data at the species level or trophic group, depending on data availability(

Barnes et al., 2018; Berlow et al., 2009). Taken together, these considerations impose too many constraints on predicting continuous interaction strength at the moment, resulting in our primary focus in binary present/absent interactions within this manuscript.

1.4.2.5. How do we determine what interaction networks are feasible? For several decades, ecologists have aimed to understand how networks of many interacting species persist through time. The diversity-stability paradox, first explored by (May, 1974), shows that under a neutral set of assumptions ecological networks should become decreasingly stable as the number of species increases. Yet, in the natural world we observe networks of interactions that consist of far more species than May's model predicts (Albouy et al., 2019). As a result, understanding what aspects of the neutral assumptions of May's model are incorrect has branched many investigations into the relationship between ecological network structure and persistence (Allesina and Tang, 2012). These assumptions can be split into dynamical assumptions and topological assumptions. Topologically, we know that ecological networks are not structured randomly. Some properties, like the aforementioned connectance, are highly predictable (MacDonald et al., 2020). Generative models of food-webs (based on network embeddings) fit empirical networks more effectively than random models (Allesina et al., 2008). These models have long used allometry as a single-dimensional niche space—naturally we want to extend this to traits in general. The second approach to stability is through *dynamics*. Early models of community dynamics rely on the assumption of linear interaction effects, but in recent years models of bioenergetic community dynamics have shown promise in basing our understanding of energy flow in food-webs in the understood relationship between allometry and metabolism (Delmas et al., 2017). An additional consideration is the multidimensional nature of "stability" and "feasibility" (*e.g.,* resilience to environmental change vs extinctions; Domínguez-García et al., 2019) and how different disturbances propagate across levels of biological organisation (Gravel et al., 2016; Kéfi et al., 2019). Recent approaches such as structural stability (Ferrera et al., 2016; Saavedra et al., 2017) allow us to think of network feasibility in rigorous mathematical terms, which may end up as usable parameters to penalise network predictions.

1.4.2.6. What taxonomic scales are suitable for the prediction of species interactions? If we use different trait-based proxies to predict potential interactions between species the choice of such proxies should be theoretically linked to the taxonomic and spatial scale we are using in our prediction (Wiens, 1989). At some scales we can use morphological traits of co-occurring species to assess the probability of interaction between them (Bartomeus et al., 2016). On broader taxonomic scales we can infer interaction probability through the phylogenetic distance, assuming that functional traits themselves are conserved (Gómez et al., 2010). In this case, we can think of the probability that one species will interact with another as the distance between them in niche-space (Desjardins-Proulx et al., 2017), and this can be modelled by simulating neutral expectations of trait variation on phylogenetic trees (Davies, 2021). At the narrowest scales, we may be interested in predicting behavioural traits like foraging behaviour (Bartomeus et al., 2016), and at this scale we may need to consider abundance's effect on the probability of an encounter (Wells and O'Hara, 2013).

1.4.2.7. What about indirect and higher-order interactions? Although network ecology often assumes that interactions go strictly from one node to the other, the web of life is made up of a variety of interactions. Indirect interactions—either higher-order interactions between species, or interaction strengths that themselves interact — have gained interest in recent years(Golubski and Abrams, 2011; Golubski et al., 2016). One mathematical tool to describe these situations is hypergraphs: hypergraphs are the generalisation of a graph, allowing a broad yet manageable approach to complex interactions (Carletti et al., 2020), by allowing for particular interactions to occur beyond a pair of nodes. An additional degree of complexity is introduced by multi-layer networks (Hutchinson et al., 2019). Multi-layer networks include edges across "variants" of the networks (timepoints, locations, or environments). These can be particularly useful to account for the metacommunity structure (Gross et al., 2020), or to understand how dispersal can inform conservation action (Albert et al., 2017). Ecological networks are intrinsically multi-layered (Pilosof et al., 2017). However, *prima facie*, increasing the dimensionality of the object we need to predict (the multiple

layers rather than a single network) makes the problem more complicated. Yet, multi-layer approaches improve prediction in social networks (Jalili et al., 2017; Najari et al., 2019; Yasami and Safaei, 2018), and they may prove useful in network ecology going forward.

1.4.3. Space

Although networks were initially used to describe the interactions *within* a community, interest in the last decade has shifted towards understanding their structure and variation over space (Baiser et al., 2019; Trøjelsgaard and Olesen, 2016), and has established network ecology as an important emerging component of biogeography and macroecology.

1.4.3.1. How much do networks vary over space? Networks can vary across space either in their structural properties (*e.g.,* connectance or degree distribution) or in their composition (identity of nodes and edges). Interestingly, variation in the structural properties of ecological networks primarily responds to changes in the size of the network. The number of links in ecological networks scales with the number of species (Brose et al., 2004; MacDonald et al., 2020), and connectance and size drive the rest of network structure (Dunne et al., 2002b; Poisot and Gravel, 2014; Riede et al., 2010). Species turnover in space results in changes in the composition of ecological networks. But, this is not the only reason network composition varies (Poisot et al., 2015). Intraspecific variation can result in interaction turnovers without changes in species composition Bolnick et al., 2011. Similarly, changes in species abundances can lead to variation in interaction strengths (Canard et al., 2014; Vázquez et al., 2007). Variation in the abiotic environment and indirect interactions Golubski et al., 2016 could modify the occurrence and strength of individual interactions. Despite this, empirical networks tend to share a common backbone (Bramon Mora et al., 2018) and functional composition (Dehling et al., 2020) across space.

1.4.3.2. How do we predict what the species pool at a particular location is? As the species pool forms the basis for network structure, predicting which species are present at a particular location is essential to predict networks across space. Species distribution models (SDMs) are increasingly ubiquitous in macroecology— these models predict the range of a

species based on known occurrences and environmental conditions, such as climate and land cover (Elith et al., 2006; Guisan and Thuiller, 2005). Including interactions or co-occurrences in SDMs generally improves predictive performance (Wisz et al., 2013). Several approaches exist to combine multiple SDMs: community assemblage at a particular site can be predicted either by combining independent single-species SDMs (stacked-SDMs, SSDMs) or by directly modelling the entire species assemblage and multiple species at the same time (joint SDMs, JSDMs) (Norberg et al., 2019). Building on the JSDM framework, hierarchical modelling of species communities (Ovaskainen et al., 2017) has the advantage of capturing processes that structure communities. Spatially Explicit Species Assemblage Modelling (SESAM) constrains SDM predictions using macro-ecological models (Guisan and Rahbek, 2011) — for example, variation in species richness across space can constrain assemblage predictions (D'Amen et al., 2015).

The next step is to constrain distribution predictions using network properties. This builds on previous calls to adopt a probabilistic view: a probabilistic species pool (Karger et al., 2016), and probabilistic interactions through Bayesian networks (Staniczenko et al., 2017). Blanchet et al., 2020 argue that the probabilistic view avoids confusion between interactions and co-occurrences, but that it requires prior knowledge of interactions. This could potentially be solved through our framework of predicting networks first, interactions next, and finally the realised species pool.

1.4.3.3. How do we combine spatial and network predictions? In order to predict networks across space, we need to combine multiple models—one which predicts what the species pool will be at a given location, and one to predict what interaction networks composed from this species pool are likely to be (see Figure 2). Both of these models contain uncertainty, and when we combine them the uncertainty from each model should be propagated into the combined model. The Bayesian paradigm provides a convenient solution to this—if we have a chain of models where each model feeds into the next, we can sample from the posterior of the input models. A different approach is *ensemble modelling* which combines the predictions made by several models, where each model is predicting the same thing

(Parker, 2013). Error propagation, an important step in building any ecological model, describes the effect of the uncertainty of input variables on the uncertainty of output variables (Draper, 1995; Parysow et al., 2000). Benke et al., 2018 identifies two broad approaches to model error propagation: analytically using differential equations or stochastically using Monte-Carlo simulation methods. Errors induced by the spatial or temporal extrapolation of data also need to be taken into account when estimating the uncertainty of a model's output (Peters and Herrick, 2004).

1.4.4. Time

1.4.4.1. Why should we forecast species interaction networks? Forecasting species interactions are critical for informing ecosystem management (Harvey et al., 2017) and systematic conservation prioritisation (Pollock et al., 2020), and for anticipating extinctions and their consequences (McDonald-Madden et al., 2016; McWilliams et al., 2019). Ecological interactions shape species distributions at both local and broad spatial scales, and including interactions in SDM models typically improves predictive performance (M. B. Araújo and Luoto, 2007; Pigot and Tobias, 2013; Wisz et al., 2013). However, these tend to rely on approaches involving estimating pairwise dependencies based on co-occurrence, using surrogates for biotic-interaction gradients, and hybridising SDMs with dynamic models (Wisz et al., 2013). Most existing models to predict the future distribution of species ignore interactions (Urban et al., 2016). Changes in species ranges and phenology will inevitably create spatiotemporal mismatches and affect encounter rates between species (Gilman et al., 2010), which will further shift the distribution of species across space. New interactions will also appear between species that are not currently co-occurring (Gilman et al., 2010). Only by forecasting how species will interact can we hope to have an accurate portrait of how biodiversity will be distributed under the future climate.

Forecasting how climate change will alter biodiversity is also crucial for maximising conservation outcomes. Improving SDMs through interactions is crucial for conservation, as nearly 30% of models in SDM studies are used to assess population declines or landscape

ability to support populations (M. B. Araújo et al., 2019). Reliable predictions about how ecological networks will change over time will give us critical information that could be communicated to decision-makers and the scientific community about what future environmental risks we are awaiting and how to mitigate them (Kindsvater et al., 2018). Not only this, but how biodiversity is structured influences the functioning of the whole ecosystem, community stability and persistence (Stouffer and Bascompte, 2010; Thompson et al., 2012). Will climate change impact the distribution of network properties (*e.g.,* connectance)? If so, which regions or species groups need special conservation efforts? These overarching questions are yet to be answered (but see Albouy et al., 2013; Hattab et al., 2016; Kortsch et al., 2015). We believe that the path toward forecasting ecological networks provides useful guidelines to ultimately better predict how climate change will affect the different dimensions of biodiversity and ecosystem functioning.

1.4.4.2. How do we turn a predictive model into a forecasting model? On some scales, empirical time-series encode enough information about ecological processes for machine-learning approaches to make accurate forecasts. However, there is an intrinsic limit to the predictability of ecological time-series (Pennekamp et al., 2019). A forecast inherently has a *resolution limit* in space, time, and organisation. For example, one could never hope to predict the precise abundance of every species on Earth on every day hundreds of years into the future. There is often a trade-off between the resolution and horizon of forecast, *e.g.,* a lower resolution forecast, like primary production will be at a maximum in the summer, is likely to be true much further into the future than a higher resolution forecast. If we want to forecast the structure of ecological networks beyond the forecasting horizon of time-series based methods, we need forecasts of our predictive model's inputs—a forecast of the distribution of both environmental conditions and the potential species pool across space (Figure 3).

1.4.4.3. How can we validate a forecasting model? Often the purpose of building a forecasting model is to inform *present* action (Dietze et al., 2018). Yet, the nature of forecasting—trying to predict the future—is that you can only know if a forecast is "right"

once it is too late to change it. If we want to maximise the chance that reality falls within a forecasting model's predictions, there are two directions to approach this problem: the first is to extend model validation techniques to a forecasting context, and the second is to attempt to maximise the amount of uncertainty in the forecast without compromising its resolution. Cross-validation (see subsubsection 1.4.1.3) can be used to test the efficacy of a forecasting model. Given a time-series of N observations, a model can iteratively be trained on the first n time-points of data, and the forecasting model's accuracy can be evaluated on the remaining time-points it hasn't "seen" (Bishop, 2006). This enables us to understand both how much temporal data is required for a model to be robust, and also enables us to explore the *forecasting horizon* of a process. Further, this approach can also be applied in the opposite temporal direction— if we have reliable data from the past, "hindcasting" can also be used to test a forecast's robustness.

However, these methods inevitably bump into a hard-limitation on what is feasible for a forecasting model. The future is uncertain. Any empirical time-series we use to validate a model was collected in past conditions that may not persist into the future. Any system we wish to forecast will undergo only one of many possible scenarios, yet we can only observe the realised outcome of the system under the scenario that actually unfolds. It is therefore impossible to assess the quality of a forecasting model in scenarios that remain hypothetical. If the goal is to maximise the probability that reality will fall within the forecast's estimates, forecasts should incorporate as much uncertainty about the future scenario as possible— one way to do this is ensemble modelling (Parker, 2013). However, as we increase the amount of uncertainty we incorporate into a forecasting model, the resolution of the forecast's predictions could shrink (Lei and Whitaker, 2017), and therefore the modeller should be mindful of the trade-off between resolution and accuracy when developing any forecast. Finally, ensemble models are not guaranteed to give more accurate results: for example, Becker et al., 2020 noted that the ensemble model outperforms the best-in-class models, which should be taken as an indication that careful model building and selection is of the

utmost importance when dealing with a problem as complex as the prediction of species interactions.

1.5. Conclusion: why should we predict species interaction networks?

Because we almost can, and because we definitely should.

A better understanding of species interactions, and the networks they form, would help unify the fields of community, network, and spatial ecology; improve the quantification of the functional relationships between species (Dehling and Stouffer, 2018; O'Connor et al., 2020); re-evaluate metacommunities in light of network structure (Guzman et al., 2019); and enable a new line of research into the biogeography of species interactions (Braga et al., 2019; Massol et al., 2017) which incorporates a synthesis of both Eltonian and Grinnellian niche (Gravel et al., 2019). Further, the ability to reliably predict and forecast species interactions would inform conservation efforts for protecting species, communities, and ecosystems. Integration of species interactions into the assessment of vulnerability to climate change is a needed methodological advancement (Foden and Young, 2016). International panels draw on models to establish scientific consensus (M. B. Araújo et al., 2019), and they can be improved through more effective prediction of species distributions and interactions (Syfert et al., 2014). Further, recent studies argue for a shift in focus from species to interaction networks for biodiversity conservation to better understand ecosystem processes (Harvey et al., 2017).

We should invest in network prediction because the right conditions to do so reliably and rapidly are beginning to emerge. Given the possible benefits to a variety of ecological disciplines that would result from an increased ability to predict networks, we feel strongly that the research agenda we outline here should be picked up by the community. Although novel technologies are bringing massive amounts of data to some parts of ecology (primarily environmental DNA and remote sensing, but now more commonly image analysis and bioacoustics), it is even more important to be intentional about *reconciling* data. This involves

not only the work of understanding the processes encoded within data, but also the groundwork of developing pipelines to bridge the ever-expanding gap between "high-throughput" and "low-throughput" sampling methods. An overall increase in the volume of data will not result in an increase of our predictive capacity as long as this data increase is limited to specific aspects of the problem. In the areas we highlight in Figure 2, many data steps are still limiting: documenting empirical interactions is natural history work that doesn't lend itself to systematic automation; expert knowledge is by design a social process that may be slightly accelerated by text mining and natural language processing (but is not yet, or not routinely or at scale). These limitations are affecting our ability to reconstruct networks.

But the tools to which we feed these data, incomplete as they may be, are gradually getting better; that is, they can do predictions faster, they handle uncertainty and propagate it well, and they can accommodate data volumes that are lower than we may expect (Pichler et al., 2020). It is clear attempting to predict the structure of ecological networks at any scale is a methodological and ecological challenge; yet it will result in qualitative changes in our understanding of complex adaptive systems, as well as changes to our ability to leverage information about network structure for conservation decision. It is perhaps even more important to forecast the structure of ecological networks because it is commonly neglected as a facet of biodiversity that can (and should) be managed. In fact, none of the Aichi targets mention biostructure or its protection, despite this being recognised as an important task (McCann, 2007), either implicitly or explicitly. Being able to generate reliable datasets on networks in space or time will make this information more actionable.

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. TS, NF, TP are funded by a donation from the Courtois Foundation; FB, NF, and TP are funded by IVADO; BM is funded by the NSERC Alexander Graham Bell Canada Graduate Scholarship and the FRQNT master's scholarship; FB, GD, NF, and GH are funded by the NSERC BIOS^2 CREATE program; GD

is funded by the FRQNT doctoral scholarship; DC, TS, LP, and TP are funded by the Canadian Institute of Ecology & Evolution; this research was enabled in part by support provided by Calcul Québec (www.calculquebec.ca) and Compute Canada (www.computecanada.ca). This work was supported by funding to the Viral Emergence Research Initiative (VERENA) consortium including NSF BII 2021909. AG and MDC are supported in part by the Liber Ero Chair.

References

Albert, C. H., Rayfield, B., Dumitru, M., & Gonzalez, A. (2017). Applying network theory to prioritize multispecies habitat networks that are robust to climate and land-use change: Prioritizing a network for biodiversity. *Conservation Biology*, *31*(6), 1383–1396. https://doi.org/10.1111/cobi.12943

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), 1153–1161. https://doi.org/10.1038/s41559-019-0950-y

Albouy, C., Guilhaumon, F., Leprieur, F., Lasram, F. B. R., Somot, S., Aznar, R., Velez, L., Le Loc'h, F., & Mouillot, D. (2013). Projected climate change and the changing biogeography of coastal Mediterranean fishes (P. Pearman, Ed.). *Journal of Biogeography*, *40*(3), 534–547. https://doi.org/10.1111/jbi.12013

Albrecht, J., Classen, A., Vollstädt, M. G. R., Mayr, A., Mollel, N. P., Schellenberger Costa, D., Dulle, H. I., Fischer, M., Hemp, A., Howell, K. M., Kleyer, M., Nauss, T., Peters, M. K., Tschapka, M., Steffan-Dewenter, I., Böhning-Gaese, K., & Schleuning, M. (2018). Plant and animal functional diversity drive mutualistic network assembly across an elevational gradient. *Nature Communications*, *9*(1), 3177. https://doi.org/10.1038/s41467-018-05610-w

Allesina, S., Alonso, D., & Pascual, M. (2008). A General Model for Food Web Structure. *Science*, *320*(5876), 658–661. https://doi.org/10.1126/science.1156269

Allesina, S., & Tang, S. (2012). Stability criteria for complex ecosystems. *Nature*, *483*(7388), 205–208. https://doi.org/10.1038/nature10832

Antoniou, A., Storkey, A., & Edwards, H. (2018). Data Augmentation Generative Adversarial Networks. *arXiv:1711.04340 [cs, stat]*.

Araújo, M. S., Guimarães, P. R., Svanbäck, R., Pinheiro, A., Guimarães, P., dos Reis, S. F., & Bolnick, D. I. (2008). Network Analysis Reveals Contrasting Effects of Intraspecific Competition on Individual Vs. Population Diets. *Ecology*, *89*(7), 1981–1993. https://doi.org/10.1890/07-0630.1

Araújo, M. B., Anderson, R. P., Barbosa, A. M., Beale, C. M., Dormann, C. F., Early, R., Garcia, R. A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R. B., Zimmermann, N. E., & Rahbek, C. (2019). Standards for distribution models in biodiversity assessments. *Science Advances*, *5*(1), eaat4858. https://doi.org/10.1126/sciadv.aat4858

Araújo, M. B., & Luoto, M. (2007). The importance of biotic interactions for modelling species distributions under climate change. *Global Ecology and Biogeography*, *16*(6), 743–753. https://doi.org/10.1111/j.1466-8238.2007.00359.x

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79. https://doi.org/10.1214/09-SS054

Badets, M., Whittington, I., Lalubin, F., Allienne, J.-F., Maspimby, J.-L., Bentz, S., Du Preez, L. H., Barton, D., Hasegawa, H., Tandon, V., Imkongwapang, R., Imkongwapang, R., Ohler, A., Combes, C., & Verneau, O. (2011). Correlating early evolution of parasitic platyhelminths to Gondwana breakup. *Systematic Biology*, *60*(6), 762–781. https://doi.org/10.1093/sysbio/syr078

Baiser, B., Gravel, D., Cirtwill, A. R., Dunne, J. A., Fahimipour, A. K., Gilarranz, L. J., Grochow, J. A., Li, D., Martinez, N. D., McGrew, A., Poisot, T., Romanuk, T. N., Stouffer, D. B., Trotta, L. B., Valdovinos, F. S., Williams, R. J., Wood, S. A., & Yeakel, J. D. (2019). Ecogeographical rules and the macroecology of food webs. *Global Ecology and Biogeography*, *28*(9), 1204–1218. https://doi.org/10.1111/geb.12925

Baiser, B., Russell, G. J., & Lockwood, J. L. (2010). Connectance determines invasion success via trophic interactions in model food webs. *Oikos*, *119*(12), 1970–1976. https://doi.org/10.1111/j.1600-0706.2010.18557.x

Barnes, A. D., Jochum, M., Lefcheck, J. S., Eisenhauer, N., Scherber, C., O'Connor, M. I., de Ruiter, P., & Brose, U. (2018). Energy Flux: The Link between Multitrophic Biodiversity and Ecosystem Functioning. *Trends in Ecology & Evolution*, *33*(3), 186–197. https://doi.org/10.1016/j.tree.2017.12.007

Bartomeus, I. (2013). Understanding linkage rules in plant-pollinator networks by using hierarchical models that incorporate pollinator detectability and plant traits. *PloS one*, *8*(7), e69200. Retrieved 2017-02-21, from http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0069200

Bartomeus, I., Gravel, D., Tylianakis, J. M., Aizen, M. A., Dickie, I. A., & Bernard-Verdier, M. (2016). A common framework for identifying linkage rules across different types of interactions. *Functional Ecology*, *30*(12), 1894–1903. https://doi.org/10.1111/1365-2435.12666

Becker, D. J., Albery, G. F., Sjodin, A. R., Poisot, T., Dallas, T. A., Eskew, E. A., Farrell, M. J., Guth, S., Han, B. A., Simmons, N. B., & Carlson, C. J. (2020). Predicting wildlife hosts of betacoronaviruses for SARS-CoV-2 sampling prioritization. *bioRxiv*, 2020.05.22.111344. https://doi.org/10.1101/2020.05.22.111344

Benke, K. K., Norng, S., Robinson, N. J., Benke, L. R., & Peterson, T. J. (2018). Error propagation in computer models: Analytic approaches, advantages, disadvantages and constraints. *Stochastic Environmental Research and Risk Assessment*, *32*(10), 2971–2985. https://doi.org/10.1007/s00477-018-1555-8

Bennett, A. E., Evans, D. M., & Powell, J. R. (2019). Potentials and pitfalls in the analysis of bipartite networks to understand plant–microbe interactions in changing environments. *Functional Ecology*, *33*(1), 107–117. https://doi.org/10.1111/1365-2435.13223

Berlow, E. L., Dunne, J., Martinez, N. D., Stark, P. B., Williams, R. J., & Brose, U. (2009). Simple prediction of interaction strengths in complex food webs. *Proceedings of the*

National Academy of Sciences, *106*(1), 187–191. https://doi.org/10.1073/pnas.0806823106

Berlow, E. L., Neutel, A.-M., Cohen, J. E., de Ruiter, P. C., Ebenman, B., Emmerson, M., Fox, J. W., Jansen, V. A. A., Iwan Jones, J., Kokkoris, G. D., Logofet, D. O., McKane, A. J., Montoya, J. M., & Petchey, O. (2004). Interaction strengths in food webs: Issues and opportunities. *Journal of Animal Ecology*, *73*(3), 585–598. https://doi.org/10.1111/j.0021-8790.2004.00833.x

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

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blanchet, F. G., Cazelles, K., & Gravel, D. (2020). Co-occurrence is not evidence of ecological interactions. *Ecology Letters*.

Bohan, D. A., Caron-Lormier, G., Muggleton, S., Raybould, A., & Tamaddoni-Nezhad, A. (2011). Automated discovery of food webs from ecological data using logic-based machine learning. *PLOS ONE*, *6*(12), e29028. https://doi.org/10.1371/journal.pone.0029028

Bohan, D. A., Vacher, C., Tamaddoni-Nezhad, A., Raybould, A., Dumbrell, A. J., & Woodward, G. (2017). Next-generation global biomonitoring: Large-scale, automated reconstruction of ecological networks. *Trends in Ecology & Evolution*, *32*(7), 477–487. https://doi.org/10.1016/j.tree.2017.03.001

Bolnick, D. I., Amarasekare, P., Araújo, M. S., Bürger, R., Levine, J. M., Novak, M., Rudolf, V. H., Schreiber, S. J., Urban, M. C., & Vasseur, D. A. (2011). Why intraspecific trait variation matters in community ecology. *Trends in Ecology & Evolution*, *26*(4), 183–192. https://doi.org/10.1016/j.tree.2011.01.009

Borowiec, M. L., Frandsen, P., Dikow, R., McKeeken, A., Valentini, G., & White, A. E. (2021). Deep learning as a tool for ecology and evolution. https://doi.org/10.32942/osf.io/nt3as

Borrett, S. R., & Scharler, U. M. (2019). Walk partitions of flow in Ecological Network Analysis: Review and synthesis of methods and indicators. *Ecological Indicators*, *106*, 105451. https://doi.org/10.1016/j.ecolind.2019.105451

Braga, J., Pollock, L. J., Barros, C., Galiana, N., Montoya, J. M., Gravel, D., Maiorano, L., Montemaggiori, A., Ficetola, G. F., Dray, S., & Thuiller, W. (2019). Spatial analyses of multi-trophic terrestrial vertebrate assemblages in Europe. *Global Ecology and Biogeography*, *28*(11), 1636–1648. https://doi.org/10.1111/geb.12981

Bramon Mora, 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

Brose, U., Ostling, A., Harrison, K., & Martinez, N. D. (2004). Unified spatial scaling of species and their trophic interactions. *Nature*, *428*(6979), 167–171. https://doi.org/10.1038/nature02297

Brown, J. H., Gillooly, J. F., Allen, A. P., Savage, V. M., & West, G. B. (2004). Toward a Metabolic Theory of Ecology. *Ecology*, *85*(7), 1771–1789. https://doi.org/10.1890/03-9000

Burkle, L. A., & Alarcon, R. (2011). The future of plant-pollinator diversity: Understanding interaction networks across time, space, and global change. *American Journal of Botany*, *98*(3), 528–538. https://doi.org/10.3732/ajb.1000391

Burkle, L. A., Marlin, J. C., & Knight, T. M. (2013). Plant-Pollinator Interactions over 120 years: Loss of Species, Co-Occurrence, and Function. *Science*, *339*(6127), 1611–1615. https://doi.org/10.1126/science.1232728

Canard, E., Mouquet, N., Mouillot, D., Stanko, M., Miklisova, D., & Gravel, D. (2014). Empirical Evaluation of Neutral Interactions in Host-Parasite Networks. *The American Naturalist*, *183*(4), 468–479. https://doi.org/10.1086/675363

Canard, E., Mouquet, N., Marescot, L., Gaston, K. J., Gravel, D., & Mouillot, D. (2012). Emergence of Structural Patterns in Neutral Trophic Networks. *PLOS ONE*, *7*(8), e38295. https://doi.org/10.1371/journal.pone.0038295

Carletti, T., Fanelli, D., & Nicoletti, S. (2020). Dynamical systems on hypergraphs. *Journal of Physics: Complexity*, *1*(3), 035006. https://doi.org/10.1088/2632-072X/aba8e1

Carpenter, S. R. (2002). Ecological futures: Building an ecology of the long now. *Ecology*, *83*(8), 2069–2083.

Cazelles, K., Araújo, M. B., Mouquet, N., & Gravel, D. (2015). A theory for species co-occurrence in interaction networks. *Theoretical Ecology*, *9*(1), 39–48. https://doi.org/10.1007/s12080-015-0281-9

Chawla, N. V. (2010). Data Mining for Imbalanced Datasets: An Overview. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 875–886). Springer US. https://doi.org/10.1007/978-0-387-09823-4_45

Christin, S., Hervet, é., & Lecomte, N. (2019). Applications for deep learning in ecology. *Methods in Ecology and Evolution*, *10*(10), 1632–1644. https://doi.org/10.1111/2041-210X.13256

Chuang, K. V., & Keiser, M. J. (2018). Adversarial Controls for Scientific Machine Learning. *Acs Chemical Biology*, *13*(10), 2819–2821. https://doi.org/10.1021/acschembio.8b00881

Cirtwill, A. R., & Eklöf, A. (2018). Feeding environment and other traits shape species' roles in marine food webs. *Ecology Letters*, *21*(6), 875–884. https://doi.org/10.1111/ele.12955

Cirtwill, A. R., Eklöf, 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*(ja). https://doi.org/10.1111/2041-210X.13180

Coelho, M. T. P., Rodrigues, J. F. M., & Rangel, T. F. (2017). Neutral Biogeography of Phylogenetically Structured Interaction Networks. *Ecography*, *40*(12), 1467–1474. https://doi.org/10.1111/ecog.02780

Cohen, J. E., Briand, F., & Newman, C. (1990). *Community Food Webs: Data and Theory*. Springer-Verlag.

Connor, N., Barberán, A., & Clauset, A. (2017). Using null models to infer microbial co-occurrence networks. *PLOS ONE*, *12*(5), e0176751. https://doi.org/10.1371/journal.pone.0176751

Craft, M. E. (2015). Infectious disease transmission and contact networks in wildlife and livestock. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1669), 20140107. https://doi.org/10.1098/rstb.2014.0107

Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W. K., Potter, C., & Bik, H. M. (2016). The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution*, *7*(9), 1008–1018. https://doi.org/10.1111/2041-210X.12574

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., & Hess, K. T. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783–2792. https://doi.org/10.1890/07-0539.1

D'Amen, M., Dubuis, A., Fernandes, R. F., Pottier, J., Pellissier, L., & Guisan, A. (2015). Using species richness and functional traits predictions to constrain assemblage predictions from stacked species distribution models. *Journal of Biogeography*, *42*(7), 1255–1266. https://doi.org/10.1111/jbi.12485

Davies, T. J. (2021). Ecophylogenetics redux. *Ecology Letters*, *n/a*. https://doi.org/10.1111/ele.13682

de Aguiar, M. A., Newman, E. A., Pires, M. M., Yeakel, J. D., Boettiger, C., Burkle, L. A., Gravel, D., Guimarães, P. R., O'Donnell, J. L., Poisot, T., Fortin, M.-J., & Hembry, D. H. (2019). Revealing biases in the sampling of ecological interaction networks. *PeerJ*, *7*, e7566. https://doi.org/10.7717/peerj.7566

Dehling, D. M., Peralta, G., Bender, I. M. A., Blendinger, P. G., Böhning-Gaese, K., Muñoz, M. C., Neuschulz, E. L., Quitián, M., Saavedra, F., Santillán, V., Schleuning, M.,

& Stouffer, D. B. (2020). Similar composition of functional roles in Andean seed-dispersal networks, despite high species and interaction turnover. *Ecology*, *101*(7). https://doi.org/10.1002/ecy.3028

Dehling, D. M., & Stouffer, D. B. (2018). Bringing the Eltonian niche into functional diversity. *Oikos*, *127*(12), 1711–1723. https://doi.org/10.1111/oik.05415

Delmas, E., Besson, M., Brice, M.-H., Burkle, L. A., Dalla Riva, G. V., Fortin, M.-J., Gravel, D., Guimarães, P. R., Hembry, D. H., Newman, E. A., Olesen, J. M., Pires, M. M., Yeakel, J. D., & Poisot, T. (2018). Analysing ecological networks of species interactions. *Biological Reviews*, 112540. https://doi.org/10.1111/brv.12433

Delmas, E., Brose, U., Gravel, D., Stouffer, D. B., & Poisot, T. (2017). Simulations of biomass dynamics in community food webs. *Methods in Ecology and Evolution*, *8*(7), 881–886. https://doi.org/10.1111/2041-210X.12713

de Ruiter, P. C., Neutel, A.-M., & Moore, J. C. (1995). Energetics, Patterns of Interaction Strengths, and Stability in Real Ecosystems. *Science*, *269*(5228), 1257–1260. https://doi.org/10.1126/science.269.5228.1257

Desdevises, Y. (2006). Determinants of parasite species richness on small taxonomical and geographical scales: Lamellodiscus monogeneans of northwestern Mediterranean sparid fish. *Journal of Helminthology*, *80*(3), 235–241.

Desjardins-Proulx, P., Laigle, I., Poisot, T., & Gravel, D. (2017). Ecological interactions and the Netflix problem. *PeerJ*, *5*, e3644. https://doi.org/10.7717/peerj.3644

Desjardins-Proulx, P., Poisot, T., & Gravel, D. (2019). Artificial Intelligence for Ecological and Evolutionary Synthesis. *Frontiers in Ecology and Evolution*, *7*. https://doi.org/10.3389/fevo.2019.00402

Dickinson, E. R., Millins, C., & Biek, R. (2020). Sampling scale and season influence the observed relationship between the density of deer and questing Ixodes ricinus nymphs. *Parasites & Vectors*, *13*(1), 493. https://doi.org/10.1186/s13071-020-04369-8

Dickinson, J. L., Zuckerberg, B., & Bonter, D. N. (2010). Citizen Science as an Ecological Research Tool: Challenges and Benefits. *Annual Review of Ecology, Evolution, and*

Systematics, *41*(1), 149–172. https://doi.org/10.1146/annurev-ecolsys-102209-144636

Dietze, M. C., Fox, A., Beck-Johnson, L. M., Betancourt, J. L., Hooten, M. B., Jarnevich, C. S., Keitt, T. H., Kenney, M. A., Laney, C. M., Larsen, L. G., Loescher, H. W., Lunch, C. K., Pijanowski, B. C., Randerson, J. T., Read, E. K., Tredennick, A. T., Vargas, R., Weathers, K. C., & White, E. P. (2018). Iterative near-term ecological forecasting: Needs, opportunities, and challenges. *Proceedings of the National Academy of Sciences*, *115*(7), 1424–1432. https://doi.org/10.1073/pnas.1710231115

Domínguez-García, V., Dakos, V., & Kéfi, S. (2019). Unveiling dimensions of stability in complex ecological networks. *Proceedings of the National Academy of Sciences*, *116*(51), 25714–25720. https://doi.org/10.1073/pnas.1904470116

Draper, D. (1995). Assessment and Propagation of Model Uncertainty. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, *57*(1), 45–97. https://doi.org/10.1111/j.2517-6161.1995.tb02015.x

Drummond, C., & Holte, R. C. (2006). Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, *65*(1), 95–130. https://doi.org/10.1007/s10994-006-8199-5

Duffy, J. E. (2002). Biodiversity and ecosystem function: The consumer connection. *Oikos*, *99*(2), 201–219. https://doi.org/10.1034/j.1600-0706.2002.990201.x

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.

Dunne, J. A., Williams, R. J., & Martinez, N. D. (2002a). Network structure and biodiversity loss in food webs: Robustness increases with connectance. *Ecology Letters*, *5*(4), 558–567. https://doi.org/10.1046/j.1461-0248.2002.00354.x

Dunne, J. A., Williams, R. J., & Martinez, N. D. (2002b). Food-web structure and network theory: The role of connectance and size. *Proceedings of the National Academy of Sciences*, *99*(20), 12917–12922. https://doi.org/10.1073/pnas.192407699

Edwards, F. A., Edwards, D. P., Hamer, K. C., & Fayle, T. M. (2021). Tropical land-use change alters trait-based community assembly rules for dung beetles and birds. *Oecologia*. https://doi.org/10.1007/s00442-020-04829-z

Eklöf, A., Jacob, U., Kopp, J., Bosch, J., Castro-Urgal, R., Chacoff, N. P., Dalsgaard, B., de Sassi, C., Galetti, M., Guimarães, P. R., Lomáscolo, S. B., Martín González, A. M., Pizo, M. A., Rader, R., Rodrigo, A., Tylianakis, J. M., Vázquez, D. P., & Allesina, S. (2013). The dimensionality of ecological networks. *Ecology Letters*, *16*(5), 577–583. https://doi.org/10.1111/ele.12081

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M. M., Peterson, A. T., ... Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Elmasri, M., Farrell, M. J., Davies, T. J., & Stephens, D. A. (2020). A hierarchical Bayesian model for predicting ecological interactions using scaled evolutionary relationships. *The Annals of Applied Statistics*, *14*(1), 221–240. https://doi.org/10.1214/19-AOAS1296

Evans, M. R., Norris, K. J., & Benton, T. G. (2012). Predictive ecology: Systems approaches. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1586), 163–169. https://doi.org/10.1098/rstb.2011.0191

Farage, C., Edler, D., Eklöf, A., Rosvall, M., & Pilosof, S. (2021). Identifying flow modules in ecological networks using Infomap. *Methods in Ecology and Evolution*, *12*(5), 778–786. https://doi.org/10.1111/2041-210X.13569

Ferrera, A., Pascual-García, A., & Bastolla, U. (2016). Effective competition determines the global stability of model ecosystems. *Theoretical Ecology*, 1–11. https://doi.org/10.1007/s12080-016-0322-z

Foden, W. B., & Young, B. E. (2016). *IUCN SSC guidelines for assessing species' vulnerability to climate change*. IUCN Cambridge, England and Gland, Switzerland.

Fontaine, C., Guimarães, P. R., Kéfi, S., Loeuille, N., Memmott, J., van der Putten, W. H., van Veen, F. J. F., & Thébault, E. (2011). The ecological and evolutionary implications of merging different types of networks. *Ecology Letters*, *14*(11), 1170–1181. https://doi.org/10.1111/j.1461-0248.2011.01688.x

Gaiarsa, M. P., & Guimarães, P. R. (2019). Interaction strength promotes robustness against cascading effects in mutualistic networks. *Scientific Reports*, *9*(1), 676. https://doi.org/10.1038/s41598-018-35803-8

Gibb, R., Albery, G. F., Becker, D. J., Brierley, L., Connor, R., Dallas, T. A., Eskew, E. A., Farrell, M. J., Rasmussen, A. L., Ryan, S. J., Sweeny, A., Carlson, C. J., & Poisot, T. (2021). Data Proliferation, Reconciliation, and Synthesis in Viral Ecology. *BioScience*, *71*(11), 1148–1156. https://doi.org/10.1093/biosci/biab080

Gilman, S. E., Urban, M. C., Tewksbury, J., Gilchrist, G. W., & Holt, R. D. (2010). A framework for community interactions under climate change. *Trends in Ecology & Evolution*, *25*(6), 325–331. https://doi.org/10.1016/j.tree.2010.03.002

Golubski, A. J., & Abrams, P. A. (2011). Modifying modifiers: What happens when interspecific interactions interact? *Journal of Animal Ecology*, *80*(5), 1097–1108. https://doi.org/10.1111/j.1365-2656.2011.01852.x

Golubski, A. J., Westlund, E. E., Vandermeer, J., & Pascual, M. (2016). Ecological Networks over the Edge: Hypergraph Trait-Mediated Indirect Interaction (TMII) Structure. *Trends in Ecology & Evolution*, *31*(5), 344–354. https://doi.org/10.1016/j.tree.2016.02.006

Gómez, J. M., Verdú, M., & Perfectti, F. (2010). Ecological interactions are evolutionarily conserved across the entire tree of life. *Nature*, *465*(7300), 918–921. https://doi.org/10.1038/nature09113

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. (2019). Bringing Elton

and Grinnell together: A quantitative framework to represent the biogeography of ecological interaction networks. *Ecography*, *42*(3), 401–415. https://doi.org/10.1111/ecog.04006

Gravel, D., Massol, F., & Leibold, M. A. (2016). Stability and complexity in model meta-ecosystems. *Nature Communications*, *7*, 12457. https://doi.org/10.1038/ncomms12457

Gravel, D., Poisot, T., Albouy, C., Velez, L., & Mouillot, D. (2013). Inferring food web structure from predator–prey body size relationships. *Methods in Ecology and Evolution*, *4*(11), 1083–1090. https://doi.org/10.1111/2041-210X.12103

Gross, T., Allhoff, K. T., Blasius, B., Brose, U., Drossel, B., Fahimipour, A. K., Guill, C., Yeakel, J. D., & Zeng, F. (2020). Modern models of trophic meta-communities. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1814), 20190455. https://doi.org/10.1098/rstb.2019.0455

Gu, Q., Zhu, L., & Cai, Z. (2009). Evaluation Measures of the Classification Performance of Imbalanced Data Sets. In Z. Cai, Z. Li, Z. Kang, & Y. Liu (Eds.), *Computational Intelligence and Intelligent Systems* (pp. 461–471). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04962-0_53

Guimarães, P. R. (2020). The Structure of Ecological Networks Across Levels of Organization. *Annual Review of Ecology, Evolution, and Systematics*, *51*(1), 433–460. https://doi.org/10.1146/annurev-ecolsys-012220-120819

Guisan, A., & Rahbek, C. (2011). SESAM – a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography*, *38*(8), 1433–1444. https://doi.org/10.1111/j.1365-2699.2011.02550.x

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, *8*(9), 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

Guzman, L. M., Germain, R. M., Forbes, C., Straus, S., O'Connor, M. I., Gravel, D., Srivastava, D. S., & Thompson, P. L. (2019). Towards a multi-trophic extension of metacommunity ecology. *Ecology Letters*, *22*(1), 19–33. https://doi.org/10.1111/ele.13162

Hadfield, J. D., Krasnov, B. R., Poulin, R., & Nakagawa, S. (2014). A Tale of Two Phylogenies: Comparative Analyses of Ecological Interactions. *The American Naturalist*, *183*(2), 174–187. https://doi.org/10.1086/674445

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

Harvey, E., Gounand, I., Ward, C. L., & Altermatt, F. (2017). Bridging ecology and conservation: From ecological networks to ecosystem function. *Journal of Applied Ecology*, *54*(2), 371–379. https://doi.org/10.1111/1365-2664.12769

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed). Springer.

Hattab, T., Leprieur, F., Ben Rais Lasram, F., Gravel, D., Loc'h, F. L., & Albouy, C. (2016). Forecasting fine-scale changes in the food-web structure of coastal marine communities under climate change. *Ecography*, *39*(12), 1227–1237. https://doi.org/10.1111/ecog.01937

Heleno, R., Garcia, C., Jordano, P., Traveset, A., Gómez, J. M., Blüthgen, N., Memmott, J., Moora, M., Cerdeira, J., Rodríguez-Echeverría, S., Freitas, H., & Olesen, J. M. (2014). Ecological networks: Delving into the architecture of biodiversity. *Biology Letters*, *10*(1). https://doi.org/10.1098/rsbl.2013.1000

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

Houlahan, J. E., McKinney, S. T., Anderson, T. M., & McGill, B. J. (2017). The priority of prediction in ecological understanding. *Oikos*, *126*(1), 1–7. https://doi.org/10.1111/oik.03726

Hutchinson, M. C., Bramon Mora, B., Pilosof, S., Barner, A. K., Kéfi, S., Thébault, E., Jordano, P., & Stouffer, D. B. (2019). Seeing the forest for the trees: Putting multilayer networks to work for community ecology (O. Godoy, Ed.). *Functional Ecology*, *33*(2), 206–217. https://doi.org/10.1111/1365-2435.13237

Ings, T. C., Montoya, J. M., Bascompte, J., Blüthgen, N., Brown, L., Dormann, C. F., Edwards, F., Figueroa, D., Jacob, U., Jones, J. I., Lauridsen, R. B., Ledger, M. E., Lewis, H. M., Olesen, J. M., van Veen, F. J. F., Warren, P. H., & Woodward, G. (2009). Ecological networks–beyond food webs. *The Journal of Animal Ecology*, *78*(1), 253–269. https://doi.org/10.1111/j.1365-2656.2008.01460.x

Innes, M. (2018). Flux: Elegant machine learning with Julia. *Journal of Open Source Software*, *3*(25), 602. https://doi.org/10.21105/joss.00602

Jalili, M., Orouskhani, Y., Asgari, M., Alipourfard, N., & Perc, M. (2017). Link prediction in multiplex online social networks. *Royal Society Open Science*, *4*(2), 160863. https://doi.org/10.1098/rsos.160863

Jeliazkov, A., Mijatovic, D., Chantepie, S., Andrew, N., Arlettaz, R., Barbaro, L., Barsoum, N., Bartonova, A., Belskaya, E., Bonada, N., Brind'Amour, A., Carvalho, R., Castro, H., Chmura, D., Choler, P., Chong-Seng, K., Cleary, D., Cormont, A., Cornwell, W., ... Chase, J. M. (2020). A global database for metacommunity ecology, integrating species, traits, environment and space. *Scientific Data*, *7*(1), 6. https://doi.org/10.1038/s41597-019-0344-7

Jenkins, C. N., Pimm, S. L., & Joppa, L. N. (2013). Global patterns of terrestrial vertebrate diversity and conservation. *Proceedings of the National Academy of Sciences*, *110*(28), E2602–E2610. https://doi.org/10.1073/pnas.1302251110

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

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

Karbowiak, G., Miklisová, D., Stanko, M., Werszko, J., Hajdul-Marwicz, M., Szewczyk, T., & Rychlik, L. (2019). The Competition Between Immatures of Ixodes ricinus and Dermacentor reticulatus (Ixodida: Ixodidae) Ticks for Rodent Hosts. *Journal of Medical Entomology*, *56*(2), 448–452. https://doi.org/10.1093/jme/tjy188

Karger, D. N., Cord, A. F., Kessler, M., Kreft, H., Kühn, I., Pompe, S., Sandel, B., Cabral, J. S., Smith, A. B., Svenning, J.-C., Tuomisto, H., Weigelt, P., & Wesche, K. (2016). Delineating probabilistic species pools in ecology and biogeography. *Global Ecology and Biogeography*, *25*(4), 489–501. https://doi.org/10.1111/geb.12422

Kawakatsu, M., Chodrow, P. S., Eikmeier, N., & Larremore, D. B. (2021). Emergence of Hierarchy in Networked Endorsement Dynamics. *arXiv:2007.04448 [nlin, physics:physics]*.

Kéfi, S., Domínguez-García, V., Donohue, I., Fontaine, C., Thébault, E., & Dakos, V. (2019). Advancing our understanding of ecological stability. *Ecology Letters*, *22*(9), 1349–1356. https://doi.org/10.1111/ele.13340

Kindsvater, H. K., Dulvy, N. K., Horswill, C., Juan-Jordá, M.-J., Mangel, M., & Matthiopoulos, J. (2018). Overcoming the Data Crisis in Biodiversity Conservation. *Trends in Ecology & Evolution*, *33*(9), 676–688. https://doi.org/10.1016/j.tree.2018.06.004

Kissling, W. D., Ahumada, J. A., Bowser, A., Fernandez, M., Fernandez, N., Alonso Garcia, E., Guralnick, R. P., Isaac, N. J. B., Kelling, S., Los, W., McRae, L., Mihoub, J.-B., Obst, M., Santamaria, M., Skidmore, A. K., Williams, K. J., Agosti, D., Amariles, D., Arvanitidis, C., … Hardisty, A. R. (2018). Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. *Biological Reviews*, *93*(1), 600–625. https://doi.org/10.1111/brv.12359

Kortsch, S., Primicerio, R., Fossheim, M., Dolgov, A. V., & Aschan, M. (2015). Climate change alters the structure of arctic marine food webs due to poleward shifts of boreal generalists. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1814), 20151546. https://doi.org/10.1098/rspb.2015.1546

Krasnov, B. R., Shenbrot, G. I., Khokhlova, I. S., & Degen, A. A. (2016). Trait-based and phylogenetic associations between parasites and their hosts: A case study with small

mammals and fleas in the Palearctic. *Oikos*, *125*(1), 29–38. https://doi.org/10.1111/oik.02178

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection : A practical approach for predictive models*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315108230

Landi, P., Minoarivelo, H. O., Brännström, Å., Hui, C., & Dieckmann, U. (2018). Complexity and stability of ecological networks: A review of the theory. *Population Ecology*, *60*(4), 319–345. https://doi.org/10.1007/s10144-018-0628-3

Lausch, A., Bannehr, L., Beckmann, M., Boehm, C., Feilhauer, H., Hacker, J. M., Heurich, M., Jung, A., Klenke, R., Neumann, C., Pause, M., Rocchini, D., Schaepman, M. E., Schmidtlein, S., Schulz, K., Selsam, P., Settele, J., Skidmore, A. K., & Cord, A. F. (2016). Linking Earth Observation and taxonomic, structural and functional biodiversity: Local to ecosystem perspectives. *Ecological Indicators*, *70*, 317–339. https://doi.org/10.1016/j.ecolind.2016.06.022

Lei, L., & Whitaker, J. S. (2017). Evaluating the trade-offs between ensemble size and ensemble resolution in an ensemble-variational data assimilation system. *Journal of Advances in Modeling Earth Systems*, *9*(2), 781–789. https://doi.org/10.1002/2016MS000864

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, *18*(17), 1–5.

Levin, S. A. (1998). Ecosystems and the Biosphere as Complex Adaptive Systems. *Ecosystems*, *1*(5), 431–436. https://doi.org/10.1007/s100219900037

Levins, R., & Lewontin, R. C. (1987). *The Dialectical Biologist*. Harvard University Press.

Luong, L. T., & Mathot, K. J. (2019). Facultative parasites as evolutionary stepping-stones towards parasitic lifestyles. *Biology Letters*, *15*(4), 20190058. https://doi.org/10.1098/rsbl.2019.0058

MacDonald, A. A. M., Banville, F., & Poisot, T. (2020). Revisiting the Links-Species Scaling Relationship in Food Webs. *Patterns*, *0*(0). https://doi.org/10.1016/j.patter.2020.100079

Macías-Hernández, N., Athey, K., Tonzo, V., Wangensteen, O. S., Arnedo, M., & Harwood, J. D. (2018). Molecular gut content analysis of different spider body parts. *PloS One*, *13*(5), e0196589. https://doi.org/10.1371/journal.pone.0196589

Magioli, M., & Ferraz, K. M. P. M. d. B. (2021). Deforestation leads to prey shrinkage for an apex predator in a biodiversity hotspot. *Mammal Research*. https://doi.org/10.1007/s13364-021-00556-9

Makiola, A., Compson, Z. G., Baird, D. J., Barnes, M. A., Boerlijst, S. P., Bouchez, A., Brennan, G., Bush, A., Canard, E., Cordier, T., Creer, S., Curry, R. A., David, P., Dumbrell, A. J., Gravel, D., Hajibabaei, M., Hayden, B., van der Hoorn, B., Jarne, P., ... Bohan, D. A. (2020). Key Questions for Next-Generation Biomonitoring. *Frontiers in Environmental Science*, *7*. https://doi.org/10.3389/fenvs.2019.00197

Maris, V., Huneman, P., Coreau, A., Kéfi, S., Pradel, R., & Devictor, V. (2017). Prediction in ecology: Promises, obstacles and clarifications. *Oikos*, n/a–n/a. https://doi.org/10.1111/oik.04655

Martínez-Abraín, A. (2008). Statistical significance and biological relevance: A call for a more cautious interpretation of results in ecology. *Acta Oecologica*, *34*(1), 9–11. https://doi.org/10.1016/j.actao.2008.02.004

Massol, F., Dubart, M., Calcagno, V., Cazelles, K., Jacquet, C., Kéfi, S., & Gravel, D. (2017). Chapter Four - Island Biogeography of Food Webs. In D. A. Bohan, A. J. Dumbrell, & F. Massol (Eds.), *Advances in Ecological Research* (pp. 183–262). Academic Press. https://doi.org/10.1016/bs.aecr.2016.10.004

Matthee, S., Stekolnikov, A. A., van der Mescht, L., Froeschke, G., & Morand, S. (2020). The diversity and distribution of chigger mites associated with rodents in the South African savanna. *Parasitology*, *147*(9), 1038–1047. https://doi.org/10.1017/S0031182020000748

May, R. M. (1974). *Stability and Complexity in Model Ecosystems* (Vol. 1). Princeton University Press. https://doi.org/10.2307/j.ctvs32rq4

McCann, K. (2007). Protecting biostructure. *Nature*, *446*(7131), 29–29. https://doi.org/10.1038/446029a

McDonald-Madden, E., Sabbadin, R., Game, E. T., Baxter, P. W. J., Chadès, I., & Possingham, H. P. (2016). Using food-web theory to conserve ecosystems. *Nature Communications*, *7*(1), 10245. https://doi.org/10.1038/ncomms10245

McWilliams, C., Lurgi, M., Montoya, J. M., Sauve, A., & Montoya, D. (2019). The stability of multitrophic communities under habitat loss. *Nature Communications*, *10*(1), 2322. https://doi.org/10.1038/s41467-019-10370-2

Michalska-Smith, M. J., & Allesina, S. (2019). Telling ecological networks apart by their structure: A computational challenge. *PLOS Computational Biology*, *15*(6), e1007076. https://doi.org/10.1371/journal.pcbi.1007076

Montoya, J. M., & Solé, R. V. (2002). Small World Patterns in Food Webs. *Journal of Theoretical Biology*, *214*(3), 405–412. https://doi.org/10.1006/jtbi.2001.2460

Montoya, J. M., Rodríguez, M. A., & Hawkins, B. A. (2003). Food web complexity and higher-level ecosystem services. *Ecology Letters*, *6*(7), 587–593. https://doi.org/10.1046/j.1461-0248.2003.00469.x

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

Morand, S., Chaisiri, K., Kritiyakan, A., & Kumlert, R. (2020). Disease Ecology of Rickettsial Species: A Data Science Approach. *Tropical Medicine and Infectious Disease*, *5*(2), 64. https://doi.org/10.3390/tropicalmed5020064

Morand, S., Simková, A., Matejusová, I., Plaisance, L., Verneau, O., & Desdevises, Y. (2002). Investigating patterns may reveal processes: Evolutionary ecology of ectoparasitic monogeneans. *International Journal for Parasitology*, *32*(2), 111–119. https://doi.org/10.1016/s0020-7519(01)00347-2

Najari, S., Salehi, M., Ranjbar, V., & Jalili, M. (2019). Link prediction in multiplex networks based on interlayer similarity. *Physica A: Statistical Mechanics and its Applications*, *536*, 120978. https://doi.org/10.1016/j.physa.2019.04.214

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

Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., Araújo, M. B., Dallas, T., Dunson, D., Elith, J., Foster, S. D., Fox, R., Franklin, J., Godsoe, W., Guisan, A., O'Hara, B., Hill, N. A., Holt, R. D., Hui, F. K. C., ... Ovaskainen, O. (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, *89*(3), e01370. https://doi.org/10.1002/ecm.1370

Novak, M., & Wootton, J. T. (2008). Estimating Nonlinear Interaction Strengths: An Observation-Based Method for Species-Rich Food Webs. *Ecology*, *89*(8), 2083–2089. https://doi.org/10.1890/08-0033.1

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

Olden, J. D., Lawler, J. J., & Poff, N. L. (2008). Machine Learning Methods Without Tears: A Primer for Ecologists. *The Quarterly Review of Biology*, *83*(2), 171–193. https://doi.org/10.1086/587826

Ovaskainen, O., Tikhonov, G., Norberg, A., Blanchet, F. G., Duan, L., Dunson, D., Roslin, T., & Abrego, N. (2017). How to make more out of community data? A conceptual

framework and its implementation as models and software. *Ecology Letters*, *20*(5), 561–576. https://doi.org/10.1111/ele.12757

Parker, W. S. (2013). Ensemble modeling, uncertainty and robust predictions. *WIREs Climate Change*, *4*(3), 213–223. https://doi.org/10.1002/wcc.220

Parysow, P., Gertner, G., & Westervelt, J. (2000). Efficient approximation for building error budgets for process models. *Ecological Modelling*, *135*(2), 111–125. https://doi.org/10.1016/S0304-3800(00)00347-1

Pascual, M., & Dunne, J. A. (2006). *Ecological Networks: Linking Structure to Dynamics in Food Webs*. Oxford University Press, USA.

Pennekamp, F., Iles, A. C., Garland, J., Brennan, G., Brose, U., Gaedke, U., Jacob, U., Kratina, P., Matthews, B., Munch, S., Novak, M., Palamara, G. M., Rall, B. C., Rosenbaum, B., Tabi, A., Ward, C., Williams, R., Ye, H., & Petchey, O. L. (2019). The intrinsic predictability of ecological time series and its potential to guide forecasting. *Ecological Monographs*, *89*(2), e01359. https://doi.org/10.1002/ecm.1359

Pérez-Harguindeguy, N., Díaz, S., Garnier, E., Lavorel, S., Poorter, H., Jaureguiberry, P., Bret-Harte, M. S., Cornwell, W. K., Craine, J. M., Gurvich, D. E., Urcelay, C., Veneklaas, E. J., Reich, P. B., Poorter, L., Wright, I. J., Ray, P., Enrico, L., Pausas, J. G., de Vos, A. C., … Cornelissen, J. H. C. (2013). New handbook for standardised measurement of plant functional traits worldwide. *Australian Journal of Botany*, *61*(3), 167. https://doi.org/10.1071/BT12225

Petchey, O. L., Beckerman, A. P., Riede, J. O., & Warren, P. H. (2008). Size, foraging, and food web structure. *Proceedings of the National Academy of Sciences*, *105*(11), 4191–4196. https://doi.org/10.1073/pnas.0710672105

Peters, D. P. C., & Herrick, J. E. (2004). Strategies for ecological extrapolation. *Oikos*, *106*(3), 627–636. https://doi.org/10.1111/j.0030-1299.2004.12869.x

Pichler, M., Boreux, V., Klein, A.-M., Schleuning, M., & Hartig, F. (2020). Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. *Methods in Ecology and Evolution*, *11*(2), 281–293. https://doi.org/10.1111/2041-210X.13329

Pigot, A. L., & Tobias, J. A. (2013). Species interactions constrain geographic range expansion over evolutionary time (C. Webb, Ed.). *Ecology Letters*, *16*(3), 330–338. https://doi.org/10.1111/ele.12043

Pilosof, S., Porter, M. A., Pascual, M., & Kéfi, S. (2017). The multilayer nature of ecological networks. *Nature Ecology & Evolution*, *1*(4), 101. https://doi.org/10.1038/s41559-017-0101

Pocock, M. J. O., Evans, D. M., & Memmott, J. (2012). The Robustness and Restoration of a Network of Ecological Networks. *Science*, *335*(6071), 973–977. https://doi.org/10.1126/science.1214915

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., Baiser, B., Dunne, J., Kéfi, S., Massol, F., Mouquet, N., Romanuk, T. N., Stouffer, D. B., Wood, S. A., & Gravel, D. (2016). Mangal – making ecological network analysis simple. *Ecography*, *39*(4), 384–390. https://doi.org/10.1111/ecog.00976

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*, jbi.14127. https://doi.org/10.1111/jbi.14127

Poisot, T., Bergeron, G., Cazelles, K., Dallas, T., Gravel, D., Macdonald, A., Mercier, B., Violet, C., & Vissault, S. (2020). Environmental biases in the study of ecological networks at the planetary scale. *bioRxiv*, 2020.01.27.921429. https://doi.org/10.1101/2020.01.27.921429

Poisot, T., & Gravel, D. (2014). When is an ecological network complex? Connectance drives degree distribution and emerging network properties. *PeerJ*, *2*, e251. https://doi.org/10.7717/peerj.251

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). Imputing the mammalian virome with linear filtering and singular value decomposition. *arXiv:2105.14973 [q-bio]*.

Poisot, T., Stanko, M., Miklisová, D., & Morand, S. (2013). Facultative and obligate parasite communities exhibit different network properties. *Parasitology*, *140*(11), 1340–1345. https://doi.org/10.1017/S0031182013000851

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

Pollock, L. J., O'Connor, L. M. J., Mokany, K., Rosauer, D. F., Talluto, M. V., & Thuiller, W. (2020). Protecting Biodiversity (in All Its Complexity): New Models and Methods. *Trends in Ecology & Evolution*, *35*(12), 1119–1128. https://doi.org/10.1016/j.tree.2020.08.015

Pomeranz, J. P., Thompson, R. M., Poisot, T., & Harding, J. S. (2018). Inferring predator-prey interactions in food webs. *Methods in Ecology and Evolution*, *0*(ja). https://doi.org/10.1111/2041-210X.13125

Portalier, S. M. J., Fussmann, G. F., Loreau, M., & Cherif, M. (2019). The mechanics of predator–prey interactions: First principles of physics predict predator–prey size ratios. *Functional Ecology*, *33*(2), 323–334. https://doi.org/10.1111/1365-2435.13254

Porter, J., Arzberger, P., Braun, H.-W., Bryant, P., Gage, S., Hansen, T., Hanson, P., Lin, C.-C., Lin, F.-P., Kratz, T., Michener, W., Shapiro, S., & Williams, T. (2005). Wireless Sensor Networks for Ecology. *BioScience*, *55*(7), 561–572. https://doi.org/10.1641/0006-3568(2005)055[0561:WSNFE]2.0.CO;2

Rackauckas, C., Ma, Y., Martensen, J., Warner, C., Zubov, K., Supekar, R., Skinner, D., Ramadhan, A., & Edelman, A. (2020). Universal Differential Equations for Scientific Machine Learning.

Rengifo-Correa, L., Stephens, C. R., Morrone, J. J., Téllez-Rendón, J. L., & Gonzalez-Salazar, C. (2017). Understanding transmissibility patterns of Chagas disease through complex vector-host networks. *Parasitology*, *144*(6), 760.

Riede, J. O., Rall, B. C., Banasek-Richter, C., Navarrete, S. A., Wieters, E. A., Emmerson, M. C., Jacob, U., & Brose, U. (2010). Scaling of Food-Web Properties with Diversity and Complexity Across Ecosystems. In *Advances in Ecological Research* (pp. 139–170). Elsevier. https://doi.org/10.1016/B978-0-12-381363-3.00003-4

Rohr, R. P., Scherer, H., Kehrli, P., Mazza, C., & Bersier, L.-F. (2010). Modeling Food Webs: Exploring Unexplained Structure Using Latent Traits. *The American Naturalist*, *176*(2), 170–177. https://doi.org/10.1086/653667

Rossberg, A. G., Matsuda, H., Amemiya, T., & Itoh, K. (2006). Food webs: Experts consuming families of experts. *Journal of Theoretical Biology*, *241*(3), 552–563. https://doi.org/10.1016/j.jtbi.2005.12.021

Saavedra, S., Rohr, R. P., Bascompte, J., Godoy, O., Kraft, N. J. B., & Levine, J. M. (2017). A structural approach for understanding multispecies coexistence. *Ecological Monographs*, *87*(3), 470–486. https://doi.org/10.1002/ecm.1263

Sala, E., & Graham, M. H. (2002). Community-wide distribution of predator–prey interaction strength in kelp forests. *Proceedings of the National Academy of Sciences*, *99*(6), 3678–3683. https://doi.org/10.1073/pnas.052028499

Sánchez, S., Serrano, E., Gómez, M. S., Feliu, C., & Morand, S. (2014). Positive co-occurrence of flea infestation at a low biological cost in two rodent hosts in the Canary archipelago. *Parasitology*, *141*(4), 511–521. https://doi.org/10.1017/S0031182013001753

Sander, E. L., Wootton, J. T., & Allesina, S. (2017). Ecological network inference from long-term presence-absence data. *Scientific Reports*, *7*(1), 7154. https://doi.org/10.1038/s41598-017-07009-x

Sasal, P., Niquil, N., & Bartoli, P. (1999). Community structure of digenean parasites of sparid and labrid fishes of the Mediterranean sea: A new approach. *Parasitology*, *119 (Pt 6)*, 635–648. https://doi.org/10.1017/s0031182099005077

Schmid-Araya, J. M., Schmid, P. E., Tod, S. P., & Esteban, G. F. (2016). Trophic positioning of meiofauna revealed by stable isotopes and food web analyses. *Ecology*, *97*(11), 3099–3109. https://doi.org/10.1002/ecy.1553

Scholes, R. J., Walters, M., Turak, E., Saarenmaa, H., Heip, C. H., Tuama, é. Ó., Faith, D. P., Mooney, H. A., Ferrier, S., Jongman, R. H., Harrison, I. J., Yahara, T., Pereira, H. M., Larigauderie, A., & Geller, G. (2012). Building a global observing system for biodiversity. *Current Opinion in Environmental Sustainability*, *4*(1), 139–146. https://doi.org/10.1016/j.cosust.2011.12.005

Segar, S. T., Fayle, T. M., Srivastava, D. S., Lewinsohn, T. M., Lewis, O. T., Novotny, V., Kitching, R. L., & Maunsell, S. C. (2020). The Role of Evolution in Shaping Ecological Networks. *Trends in Ecology & Evolution*, *35*(5), 454–466. https://doi.org/10.1016/j.tree.2020.01.004

Seibold, S., Cadotte, M. W., MacIvor, J. S., Thorn, S., & Müller, J. (2018). The Necessity of Multitrophic Approaches in Community Ecology. *Trends in Ecology & Evolution*, *33*(10), 754–764. https://doi.org/10.1016/j.tree.2018.07.001

Simková, A., Kadlec, D., Gelnar, M., & Morand, S. (2002). Abundance-prevalence relationship of gill congeneric ectoparasites: Testing the core satellite hypothesis and ecological specialisation. *Parasitology Research*, *88*(7), 682–686. https://doi.org/10.1007/s00436-002-0650-3

Skidmore, A. K., & Pettorelli, N. (2015). Agree on biodiversity metrics to track from space: Ecologists and space agencies must forge a global monitoring strategy. *Nature*, *523*(7561), 403–406.

Smith, G. R., Finlay, R. D., Stenlid, J., Vasaitis, R., & Menkis, A. (2017). Growing evidence for facultative biotrophy in saprotrophic fungi: Data from microcosm tests with 201 species of wood-decay basidiomycetes. *The New Phytologist*, *215*(2), 747–755. https://doi.org/10.1111/nph.14551

Smith-Ramesh, L. M., Moore, A. C., & Schmitz, O. J. (2016). Global synthesis suggests that food web connectance correlates to invasion resistance. *Global Change Biology*, n/a–n/a. https://doi.org/10.1111/gcb.13460

Spiegelhalter, D. J., Myles, J. P., Jones, D. R., & Abrams, K. R. (2000). Bayesian methods in health technology assessment: A review.

Staniczenko, P. P. A., Sivasubramaniam, P., Suttle, K. B., & Pearson, R. G. (2017). Linking macroecology and community ecology: Refining predictions of species distributions using biotic interaction networks. *Ecology Letters*, *20*(6), 693–707. https://doi.org/10.1111/ele.12770

Stephens, C. (2009). Using Biotic interaction Networks for prediction in Biodiversity and emerging diseases. *PLoS ONE 4(5):* https://doi.org/doi:10.1371/journal.pone.0005725

Stephenson, P. (2020). Technological advances in biodiversity monitoring: Applicability, opportunities and challenges. *Current Opinion in Environmental Sustainability*, *45*, 36–41. https://doi.org/10.1016/j.cosust.2020.08.005

Stock, M., Poisot, T., Waegeman, W., & Baets, B. D. (2017). Linear filtering reveals false negatives in species interaction data. *Scientific Reports*, *7*, 45908. https://doi.org/10.1038/srep45908

Stouffer, D. B. (2019). All ecological models are wrong, but some are useful. *Journal of Animal Ecology*, *88*(2), 192–195. https://doi.org/10.1111/1365-2656.12949

Stouffer, D. B., & Bascompte, J. (2010). Understanding food-web persistence from local to global scales. *Ecology Letters*, *13*(2), 154–161. https://doi.org/10.1111/j.1461-0248.2009.01407.x

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

Syfert, M. M., Joppa, L., Smith, M. J., Coomes, D. A., Bachman, S. P., & Brummitt, N. A. (2014). Using species distribution models to inform IUCN Red List assessments. *Biological Conservation*, *177*, 174–184. https://doi.org/10.1016/j.biocon.2014.06.012

Terry, J. C. D., & Lewis, O. T. (2020). Finding missing links in interaction networks. *Ecology*, *101*(7), e03047. https://doi.org/10.1002/ecy.3047

Thompson, R. M., Brose, U., Dunne, J., Hall, R. O., Hladyz, S., Kitching, R. L., Martinez, N. D., Rantala, H., Romanuk, T. N., Stouffer, D. B., & Tylianakis, J. M. (2012). Food webs: Reconciling the structure and function of biodiversity. *Trends in Ecology & Evolution*, *27*(12), 689–697. https://doi.org/10.1016/j.tree.2012.08.005

Tinker, M. T., Guimarães, P. R., Novak, M., Marquitti, F. M. D., Bodkin, J. L., Staedler, M., Bentall, G., & Estes, J. A. (2012). Structure and mechanism of diet specialisation: Testing models of individual variation in resource use with sea otters. *Ecology Letters*, *15*(5), 475–483. https://doi.org/10.1111/j.1461-0248.2012.01760.x

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(3), 611–622. https://doi.org/10.1111/1467-9868.00196

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

Urban, M. C., Bocedi, G., Hendry, A. P., Mihoub, J. B., Peer, G., Singer, A., Bridle, J. R., Crozier, L. G., De Meester, L., Godsoe, W., Gonzalez, A., Hellmann, J. J., Holt, R. D., Huth, A., Johst, K., Krug, C. B., Leadley, P. W., Palmer, S. C. F., Pantel, J. H., … Travis, J. M. J. (2016). Improving the forecast for biodiversity under climate change. *Science*, *353*(6304), aad8466–aad8466. https://doi.org/10.1126/science.aad8466

Vazquez, D. P., Bluethgen, N., Cagnolo, L., & Chacoff, N. P. (2009). Uniting pattern and process in plant-animal mutualistic networks: A review. *Annals of Botany*, *103*(9), 1445–1457. https://doi.org/10.1093/aob/mcp057

Vázquez, D. P., Melián, C. J., Williams, N. M., Blüthgen, N., Krasnov, B. R., & Poulin, R. (2007). Species abundance and asymmetric interaction strength in ecological networks. *Oikos*, *116*(7), 1120–1127. https://doi.org/10.1111/j.0030-1299.2007.15828.x

Vázquez, D. P., Morris, W. F., & Jordano, P. (2005). Interaction frequency as a surrogate for the total effect of animal mutualists on plants. *Ecology Letters*, *8*(10), 1088–1094.

Vieira, M. C., & Almeida-Neto, M. (2015). A simple stochastic model for complex coextinctions in mutualistic networks: Robustness decreases with connectance. *Ecology Letters*, *18*(2), 144–152. https://doi.org/10.1111/ele.12394

Wardeh, M., Baylis, M., & Blagrove, M. S. C. (2021). Predicting mammalian hosts in which novel coronaviruses can be generated. *Nature Communications*, *12*(1), 780. https://doi.org/10.1038/s41467-021-21034-5

Weinstein, B. G., & Graham, C. H. (2017). On comparing traits and abundance for predicting species interactions with imperfect detection. *Food Webs*, *11*, 17–25.

Wells, K., & O'Hara, R. B. (2013). Species interactions: Estimating per-individual interaction strength and covariates before simplifying data into per-species ecological networks. *Methods in Ecology and Evolution*, *4*(1), 1–8. https://doi.org/10.1111/j.2041-210x.2012.00249.x

Wiens, J. A. (1989). Spatial Scaling in Ecology. *Functional Ecology*, *3*(4), 385–397. https://doi.org/10.2307/2389612

Wisz, M. S., Pottier, J., Kissling, W. D., Pellissier, L., Lenoir, J., Damgaard, C. F., Dormann, C. F., Forchhammer, M. C., Grytnes, J.-A., Guisan, A., Heikkinen, R. K., Høye, T. T., Kühn, I., Luoto, M., Maiorano, L., Nilsson, M.-C., Normand, S., Öckinger, E., Schmidt, N. M., ... Svenning, J.-C. (2013). The role of biotic interactions in shaping distributions and realised assemblages of species: Implications for species distribution

modelling. *Biological Reviews*, *88*(1), 15–30. https://doi.org/10.1111/j.1469-185X.2012.00235.x

Wootton, J. T. (1997). Estimates and Tests of Per Capita Interaction Strength: Diet, Abundance, and Impact of Intertidally Foraging Birds. *Ecological Monographs*, *67*(1), 45–64. https://doi.org/10.1890/0012-9615(1997)067[0045:EATOPC]2.0.CO;2

Wootton, J. T., & Emmerson, M. (2005). Measurement of Interaction Strength in Nature. *Annual Review of Ecology, Evolution, and Systematics*, *36*(1), 419–444. https://doi.org/10.1146/annurev.ecolsys.36.091704.175535

Yasami, Y., & Safaei, F. (2018). A novel multilayer model for missing link prediction and future link forecasting in dynamic complex networks. *Physica A: Statistical Mechanics and its Applications*, *492*, 2166–2197. https://doi.org/10.1016/j.physa.2017.11.134

Yen, T.-C., & Larremore, D. B. (2020). Community Detection in Bipartite Networks with Stochastic Blockmodels. *Physical Review E*, *102*(3), 032309. https://doi.org/10.1103/PhysRevE.102.032309

Yodzis, P., & Innes, S. (1992). Body Size and Consumer-Resource Dynamics. *The American Naturalist*, *139*(6), 1151–1175. https://doi.org/10.1086/285380

Zhang, M., & He, F. (2021). Plant breeding systems influence the seasonal dynamics of plant-pollinator networks in a subtropical forest. *Oecologia*. https://doi.org/10.1007/s00442-021-04863-5

Zhao, N., Charland, K., Carabali, M., Nsoesie, E. O., Maheu-Giroux, M., Rees, E., Yuan, M., Garcia Balaguera, C., Jaramillo Ramirez, G., & Zinszer, K. (2020). Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia (M. Choisy, Ed.). *PLOS Neglected Tropical Diseases*, *14*(9), e0008056. https://doi.org/10.1371/journal.pntd.0008056

Zurell, D., Zimmermann, N. E., Gross, H., Baltensweiler, A., Sattler, T., & Wüest, R. O. (2020). Testing species assemblage predictions from stacked and joint species distribution models. *Journal of Biogeography*, *47*(1), 101–113. https://doi.org/10.1111/jbi.13608

Chapter 2 Second article

Graph embedding and transfer learning can help predict potential species interaction networks despite data limitations

by

Tanya Strydom[1], Salomé Bouskila[2], Francis Banville[3], Ceres Barros[4], Dominique Caron[5], Maxwell J. Farrell[6], Marie-Josée Fortin[7], Benjamin Mercier[8], Laura Pollock[9], Rogini Runghen[10], Giulio V. Dalla Riva[11], and Timothée Poisot[12]

([1]) Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([2]) Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([3]) Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada
Université de Sherbrooke, Sherbrooke, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([4]) Department of Forest Resources Management, University of British Columbia, Vancouver, BC, Canada

([5]) McGill University, Montréal, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([6]) Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, ON, Canada

([7]) Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, ON, Canada

([8]) Université de Sherbrooke, Sherbrooke, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([9]) McGill University, Montréal, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([10]) Centre for Integrative Ecology, School of Biological Sciences, University of Canterbury, Canterbury, New Zealand

([11]) School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

([12]) Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

This article was submitted in Methods in Ecology and Evolution and is available at https://doi.org/10.1111/2041-210X.14228.

The main contributions of Tanya Strydom for this articles are presented.

TP and TS designed the study and edited the manuscript for submission. TS performed the analysis and wrote the manuscript. All authors contributed to the article and approved the submitted version.

RÉSUMÉ. 1. Les métawebs (réseaux d'interactions potentielles au sein d'un pool d'espèces) constituent une abstraction puissante pour comprendre comment sont structurés les réseaux d'interactions d'espèces à grande échelle.

2. Étant donné que les métawebs sont généralement exprimés à de grandes échelles spatiales et taxonomiques, leur assemblage est un processus fastidieux et coûteux ; les méthodes prédictives peuvent aider à contourner les limitations liées aux carences des données, en fournissant une première approximation des métawebs.

3. Une façon d'améliorer notre capacité à prédire les métawebs consiste à maximiser les informations disponibles en utilisant des graphiques intégrés, par opposition à une liste exhaustive des interactions entre espèces. L'intégration de graphes est un domaine émergent de l'apprentissage automatique qui recèle un grand potentiel de problèmes écologiques.

4. Ici, nous décrivons comment les défis associés à l'inférence de métawebs s'alignent avec les avantages des intégrations de graphes ; suivi d'une discussion sur la façon dont le choix du pool d'espèces a des conséquences sur le réseau reconstruit, en particulier sur le rôle des frontières créées par l'homme (ou arbitrairement assignées) et comment celles-ci peuvent influencer les hypothèses écologiques.

Mots clés : réseaux écologiques, intégration de réseaux, apprentissage par transfert, macroécologie de réseau

Abstract. 1. Metawebs (networks of potential interactions within a species pool) are a powerful abstraction to understand how large-scale species interaction networks are structured.

2. Because metawebs are typically expressed at large spatial and taxonomic scales, assembling them is a tedious and costly process; predictive methods can help circumvent the limitations in data deficiencies, by providing a first approximation of metawebs.

3. One way to improve our ability to predict metawebs is to maximize available information by using graph embeddings, as opposed to an exhaustive list of species interactions. Graph embedding is an emerging field in machine learning that holds great potential for ecological problems.

4. Here, we outline how the challenges associated with inferring metawebs line-up with the advantages of graph embeddings; followed by a discussion as to how the choice of the species pool has consequences on the reconstructed network, specifically as to the role of human-made (or arbitrarily assigned) boundaries and how these my influence ecological hypotheses.

Keywords: ecological networks, network embedding, transfer learning, netwrok macroecology

2.1. Introduction

The ability to infer *potential* interactions could serve as a significant breakthrough in our ability to conceptualize species interaction networks over large spatial scales (Hortal et al., 2015). Reliable inferences would not only boost our understanding of the structure of species interaction networks, but also increase the amount of information that can be used for biodiversity management. 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 – which, due to the limitations in most predictive methods, restricts the ability to infer interactions to locations where it is least required (*i.e.,* regions where we already have interaction data) leaving us unable to make inference in data scarce regions (where we most need it); second, accurate predictors are important for accurate predictions, and the lack of methods that can leverage a small amount of *accurate* data is a serious impediment

to our predictive ability. In most places, our most reliable biodiversity knowledge is that of a species pool where a set of potentially interacting species in a given area could occur: through the analysis of databases like the Global Biodiversity Information Facility (GBIF) or the International Union for the Conservation of Nature (IUCN), it is possible to construct a list of species for a region of interest; however inferring the potential interactions between these species still remains a challenge.

Following the definition of Dunne, 2006, a metaweb is the ecological network analogue to the species pool; specifically, it inventories all *potential* interactions between species for a spatially delimited area (and so captures the γ diversity of interactions). The metaweb itself is not a prediction of local networks at specific locations within the spatial area it covers: it will have a different structure, notably by having a larger connectance (see *e.g.,* Wood et al., 2015) and complexity (see *e.g.,* Galiana et al., 2022), than any of these local networks. These local networks (which capture the α diversity of interactions) are a subset of the metaweb's species and its realized interactions, and have been called "metaweb realizations" (Poisot et al., 2015). Differences between local networks and their metawebs are due to chance, species abundance and co-occurrence, local environmental conditions, and local distribution of functional traits, among others. Specifically, although co-occurrence can be driven by interactions (Cazelles et al., 2016), co-occurrence alone is not a predictor of interactions (Blanchet et al., 2020; Thurman et al., 2019), and therefore the lack of co-occurrence cannot be used to infer the lack of a feasible interaction. Yet, recent results by Saravia et al., 2021 strongly suggested that local (metaweb) realizations only respond weakly to local conditions: instead, they reflect constraints inherited by the structure of their metaweb. This sets up the core goal of predictive network ecology as the prediction of metaweb structure, as it is required to accurately produce downscaled, local predictions.

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 represents the "upper bounds" on what the composition of the local networks, given a local species pool, can be (see *e.g.,* McLeod et al., 2021); this information can help

evaluate the ability of ecological assemblages to withstand the effects of, for example, climate change (Fricke et al., 2022). These local networks may be reconstructed given an appropriate knowledge of local species composition and provide information on the 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 (J. Braga et al., 2019; O'Connor et al., 2020), and crop-pest networks (Grünig et al., 2020). In this contribution, we highlight the power of viewing (and constructing) metawebs as *probabilistic* objects in the context of low-probability interactions, discuss how a family of machine learning tools (graph embeddings and transfer learning) can be used to overcome data limitations to metaweb inference, and highlight how the use of metawebs introduces important questions for the field of network ecology.

2.2. A metaweb is an inherently probabilistic object

Treating interactions as probabilistic (as opposed to binary) events is a more nuanced and realistic way to represent them. Dallas et al., 2017 suggested that most interactions (links) in ecological networks are cryptic, *i.e.,* uncommon or hard to observe. This argument echoes Jordano, 2016: sampling 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 interactions will be more prevalent in networks than common or strong interactions (Csermely, 2004), compared to strong, persistent interactions; this is notably the case in food chains, wherein many weaker interactions are key to the stability of a system (Neutel et al., 2002). In the light of these observations, we expect to see an over-representation of low-probability (hereafter rare) interactions under a model that accurately predicts interaction probabilities.

Yet, the original metaweb definition, and indeed most past uses of metawebs, was based on the presence/absence of interactions. Moving towards *probabilistic* metawebs, by representing interactions as Bernoulli events (see *e.g.,* Poisot et al., 2016), offers the opportunity

to weigh these rare interactions appropriately. The inherent plasticity of interactions is important to capture: there have been documented instances of food webs undergoing rapid collapse/recovery cycles over short periods of time (*e.g.,* Pedersen et al., 2017). Furthermore, because the structure of the metaweb cannot be known in advance, it is important to rely on predictive tools that do not assume a specific network topology for link prediction (Gaucher et al., 2021), but are able to work on generalizations of the network. These considerations emphasize why metaweb predictions should focus on quantitative (preferentially probabilistic) predictions, and this should constrain the suite of models that are appropriate for prediction.

It is important to recall that a metaweb is intended as a catalogue of all potential (feasible) interactions, which is then filtered for a given application (Morales-Castilla et al., 2015). It is therefore important to separate the interactions that happen "almost surely" (repeated observational data), "almost never" (repeated lack of evidence *or* evidence that the link is forbidden through *e.g.,* trait mis-match), and interactions with a probability that lays somewhere in between (Catchen et al., 2023). 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 will increase exponentially (Jordano, 2016). Recent proposals in other fields relying on machine learning approaches emphasize the idea that algorithms meant to predict, through the assumption that they approximate the process generating the data, can also act as data generators (Hoffmann et al., 2019). High quality observational data can be used to infer core rules underpinning network structure, and be supplemented with synthetic data coming from predictive models trained on them, thereby increasing the volume of information available for analysis. 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. In this context, a probabilistic metaweb represents an aggregation of informative priors on the biological feasibility of interactions, which is usually hard to obtain yet has possibly the most potential to boost our predictive ability of local networks

(Bartomeus, 2013; Bartomeus et al., 2016). This would represent a departure from simple rules expressed at the network scale (*e.g.,* Williams and Martinez, 2000 to a view of network prediction based on learning the rules that underpin interactions *and* their variability (Gupta et al., 2022).

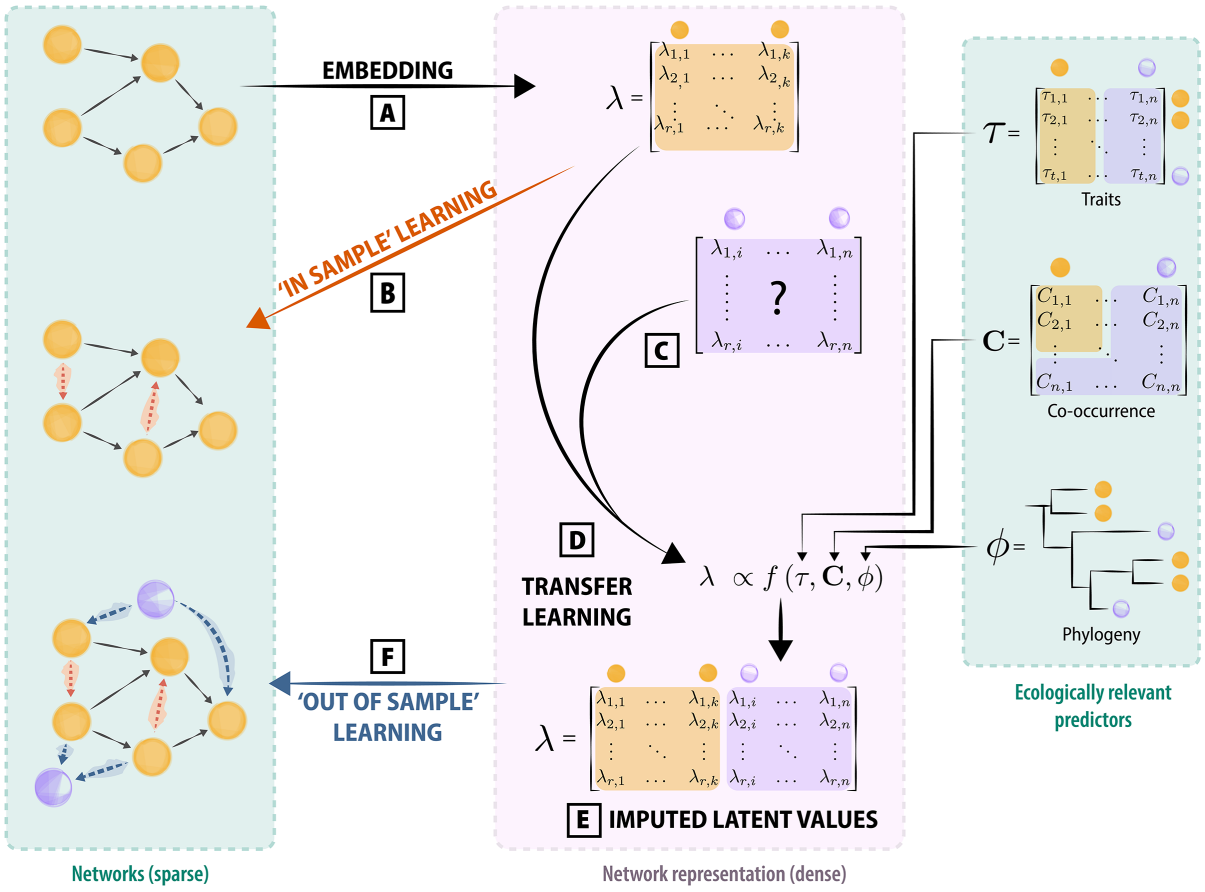


Fig. 1. The embedding process (**A**) can help to identify links (interactions) that may have been missed within the original community (represented by the orange dashed arrows, **B**). Transfer learning (**D**) allows for the prediction links (interactions) even when novel species (**C**) are included alongside the original community. This is achieved by learning using other relevant predictors (*e.g.,* traits) in conjunction with the known interactions to infer latent values (**E**). Ultimately this allows us to predict links (interactions) for species external from the original sample (blue dashed arrows) as well as missing within sample links (**F**). Within this context the predicted (and original) networks as well as the ecological predictors used (green boxes) are products that can be quantified through measurements in the field, whereas the embedded as well as imputed matrices (purple box) are representative of a decomposition of the interaction matrices onto the embedding space

2.3. Graph embedding offers promises for the inference of potential interactions

Graph (or network) embedding (Figure 1) is a family of machine learning techniques, whose main task is to learn a mapping function from a discrete graph to a continuous domain (Arsov and Mirceva, 2019; Chami et al., 2022). Their main goal is to learn a low dimensional vector representation of the graph (embeddings), such that its key properties (*e.g.,* local or global structures) are retained in the embedding space (Yan et al., 2005). The embedding space may, but will not necessarily, have lower dimensionality than the graph. Ecological networks are promising candidates for the routine application of embeddings, as they tend to possess a shared structural backbone (see *e.g.,* Bramon Mora et al., 2018), which hints at structural invariants in empirical data. Assuming that these structural invariants are common enough, they would dominate the structure of networks, and therefore be adequately captured by the first (lower) dimensions of an embedding, without the need to measure derived aspects of their structure (*e.g.,* motifs, paths, modularity, ...).

2.3.1. Graph embedding produces latent variables (but not traits)

Before moving further, it is important to clarify the epistemic status of node values derived from embeddings: specifically, they are *not* functional traits, and therefore should not be interpreted in terms of effects or responses. As per the framework of Malaterre et al., 2019, these values neither derive from, nor result in, changes in organismal performance, and should therefore not be used to quantify *e.g.,* functional diversity. This holds true even when there are correlations between latent values and functional traits: although these enable an ecological discussion of how traits condition the structure of the network, the existence of a statistical relationship does not elevate the latent values to the status of functional traits.

Rather than directly predicting biological rules (see *e.g.,* Pichler et al., 2020 for an overview), which may be confounded by the sparse nature of graph data, learning embeddings works in the low-dimensional space that maximizes information about the network structure. This approach is further justified by the observation, for example, that the

macro-evolutionary history of a network is adequately represented by some graph embeddings (Random dot product graphs (RDPG); see Dalla Riva and Stouffer, 2016). In a recent publication, Strydom et al., 2022 have used an embedding (based on RDPG) to project a metaweb of trophic interactions between European mammals, and transferred this information to mammals of Canada, using the phylogenetic distance between related clades to infer the values in the latent subspace into which the European metaweb was projected. By performing the RDPG step on re-constructed values, this approach yields a probabilistic trophic metaweb for mammals of Canada based on knowledge of European species, despite a limited ($\approx 5\%$) taxonomic overlap, and illustrates how the values derived from an embedding can be used for prediction without being "traits" of the species they represent.

2.3.2. Ecological networks are good candidates for embedding

Food webs are inherently low-dimensional objects, and can be adequately represented with less than ten dimensions (J. Braga et al., 2019; M. P. Braga et al., 2021; Eklöf et al., 2013). Simulation results by Botella et al., 2022 suggested that there is no dominant method to identify architectural similarities between networks: multiple approaches need to be tested and compared to the network descriptor of interest on a problem-specific basis. This matches previous results on graph embedding, wherein different embedding algorithms yield different network embeddings (Goyal and Ferrara, 2018), calling for a careful selection of the problem-specific approach to use. In Table 1, we present a selection of common graph and node embedding methods, alongside examples of their use to predict interactions or statistical associations between species. These methods rely largely on linear algebra or pseudo-random walks on graphs. All forms of embeddings presented in Table 1 share the common property of summarizing their objects into (sets of) dense feature vectors, that capture the overall network structure, pairwise information on nodes, and emergent aspects of the network, in a compressed way (*i.e.,* with some information loss, as we later discuss in the illustration). Node embeddings tend to focus on maintaining pairwise relationships (*i.e.,* species interactions), while graph embeddings focus on maintaining the network structure

(*i.e.,* emergent properties). Nevertheless, some graph embedding techniques (like RDPG, see *e.g.,* Wu et al., 2021) will provide high-quality node-level embeddings while also preserving network structure.

Graph embeddings *can* serve as a dimensionality reduction method. For example, RDPG (Strydom et al., 2022) and t-SVD (truncated Singular Value Decomposition; (Poisot et al., 2021) typically embed networks using fewer dimensions than the original network (the original network has as many dimensions as species, and as many informative dimensions as trophically unique species; Strydom, Dalla Riva, et al., 2021). However, this is not necessarily the case – indeed, one may perform a PCA (a special case of SVD) to project the raw data into a subspace that improves the efficacy of t-SNE (t-distributed stochastic neighbor embedding; van der Maaten, 2009). There are many dimensionality reductions (Anowar et al., 2021) that can be applied to an embedded network should the need for dimensionality reduction (for example for data visualization) arise. In brief, many graph embeddings *can* serve as dimensionality reduction steps, but not all do, neither do all dimensionality reduction methods provide adequate graph embedding capacities. In the next section (and Figure 1), we show how the amount of dimensionality reduction can affect the quality of the embedding.

Method	Object	Technique	Reference	Application
tSNE	nodes	statistical divergence	Hinton and Roweis, 2002	Cieslak et al., 2020, species-environment responses[a] Gibb et al., 2021, host-virus network representation
LINE	nodes	stochastic gradient descent	Tang et al., 2015	
SDNE	nodes	gradient descent	D. Wang et al., 2016	
node2vec	nodes	stochastic gradient descent	Grover and Leskovec, 2016	
HARP	nodes	meta-strategy	H. Chen et al., 2017	
DMSE	joint nodes	deep neural network	D. Chen et al., 2017	D. Chen et al., 2017, species-environment interactions[b]
graph2vec	sub-graph	skipgram network	Narayanan et al., 2017	
RDPG	graph	SVD	Young and Scheinerman, 2007	Dalla Riva and Stouffer, 2016, trophic interactions Poisot et al., 2021, host-virus network prediction
GLEE	graph	Laplacian eigenmap	Torres et al., 2020	
DeepWalk	graph	stochastic gradient descent	Perozzi et al., 2014	Wardeh et al., 2021, host-virus interactions
GraphKKE	graph	stochastic differential equation	Melnyk et al., 2020	Melnyk et al., 2020, microbiome species associations[a]
FastEmbed	graph	eigen decomposition	Ramasamy and Madhow, 2015	
PCA	graph	eigen decomposition	Surendran, 2013	Strydom, Catchen, et al., 2021, host-parasite interactions
Joint methods	multiple graphs	multiple strategies	S. Wang et al., 2021	

Table 1. Overview of some common graph embedding approaches, by type of embedded objects, alongside examples of their use in the prediction of species interactions. These methods have not yet been routinely used to predict species interactions; most examples that we identified were either statistical associations, or analogues to joint species distribution models. See also Box 1 for an additional discussion on Graph Neural Networks. [a]: application is concerned with *statistical* interactions, which are not necessarily direct biotic interactions; [b]:application is concerned with joint-SDM-like approach, which is also very close to statistical associations as opposed to direct biotic interactions. Given the need to evaluate different methods on a problem-specific basis, the fact that a lot of methods have not been used on network problems is an opportunity for benchmarking and method development. Note that the row for PCA also applies to kernel/probabilistic PCA, which are variations on the more general method of SVD. Note further that tSNE has been included because it is frequently used to embed graphs, including of species associations/interactions, despite not being strictly speaking, a graph embedding technique (see *e.g.,* Chami et al., 2022.)

Box 1

Graph Neural Networks

One prominent family of approaches we do not discuss in the present manuscript is Graph Neural Networks (GNN; Zhou et al., 2020). GNN are, in a sense, a method to embed a graph into a dense subspace, but belong to the family of deep learning methods, which has its own set of practices (see *e.g.,* Goodfellow et al., 2016). An important issue with methods based on deep learning is that, because their parameter space is immense, the sample size of the data fed into them must be similarly large (typically thousands of instances). This is a requirement for the model to converge correctly during training, but this assumption is unlikely to be met given the size of datasets currently available for metawebs (or single time/location species interaction networks). This data volume requirement is mostly absent from the techniques we list below. Furthermore, GNN still have some challenges related to their shallow structure, and concerns related to scalability (see Gupta et al., 2021 for a review), which are mostly absent from the methods listed in Table 1. Assuming that the uptake of next-generation biomonitoring techniques does indeed deliver larger datasets on species interactions (Bohan et al., 2017), there is nevertheless the potential for GNN to become an applicable embedding/predictive technique in the coming years.

The popularity of graph embedding techniques in machine learning is more than the search for structural invariants: graphs are discrete objects, and machine learning techniques tend to handle continuous data better. Bringing a sparse graph into a continuous, dense vector space (Xu, 2021) opens up a broader variety of predictive algorithms, notably of the sort that are able to predict events as probabilities (Murphy, 2022). Furthermore, the projection of the graph itself is a representation that can be learned; (Runghen et al., 2021), for example, used a neural network to learn the embedding of a network in which not all interactions were known, based on the nodes' metadata. This example has many parallels in

ecology (see Figure 1 **C**), in which node metadata can be represented by phylogeny, abundance, or functional traits. Using phylogeny as a source of information assumes (or strives to capture) the action of evolutionary processes on network structure, which at least for food webs have been well documented (M. P. Braga et al., 2021; Dalla Riva and Stouffer, 2016; Eklöf and Stouffer, 2016; Stouffer et al., 2007; Stouffer et al., 2012); similarly, the use of functional traits assumes that interactions can be inferred from the knowledge of trait-matching rules, which is similarly well supported in the empirical literature (Bartomeus, 2013; Bartomeus et al., 2016; Goebel et al., 2023; Gravel et al., 2013). Relating this information to an embedding rather than a list of network measures would allow to capture their effect on the more fundamental aspects of network structure; conversely, the absence of a phylogenetic or functional signal may suggest that evolutionary/trait processes are not strong drivers of network structure, therefore opening a new way to perform hypothesis testing.

2.4. An illustration of metaweb embedding

In this section, we illustrate the embedding of a collection of bipartite networks collected by Hadfield et al., 2014, using t-SVD and RDPG. Briefly, an RDPG decomposes a network into two subspaces (left and right), which are matrices that when multiplied give an approximation of the original network. RDPG has the particularly desirable properties of being a graph embedding technique that produces relevant node-level feature vectors, and provides good approximations of graphs with varied structures (Athreya et al., 2017). The code to reproduce this example is available as supplementary material in Appendix B (note, for the sake of comparison, that Strydom, Catchen, et al., 2021 have an example using embedding through PCA followed by prediction using a deep neural network on the same dataset). The resulting (binary) metaweb $\mathcal{M}$ has 2131 interactions between 206 parasites and 121 hosts, and its adjacency matrix has full rank (*i.e.,* it represents a space with 121 dimensions). All analyses were done using `Julia` (Bezanson et al., 2017) version 1.7.2, `Makie.jl` (Danisch and Krumbiegel, 2021), and `EcologicalNetworks.jl` (Poisot et al., 2019).

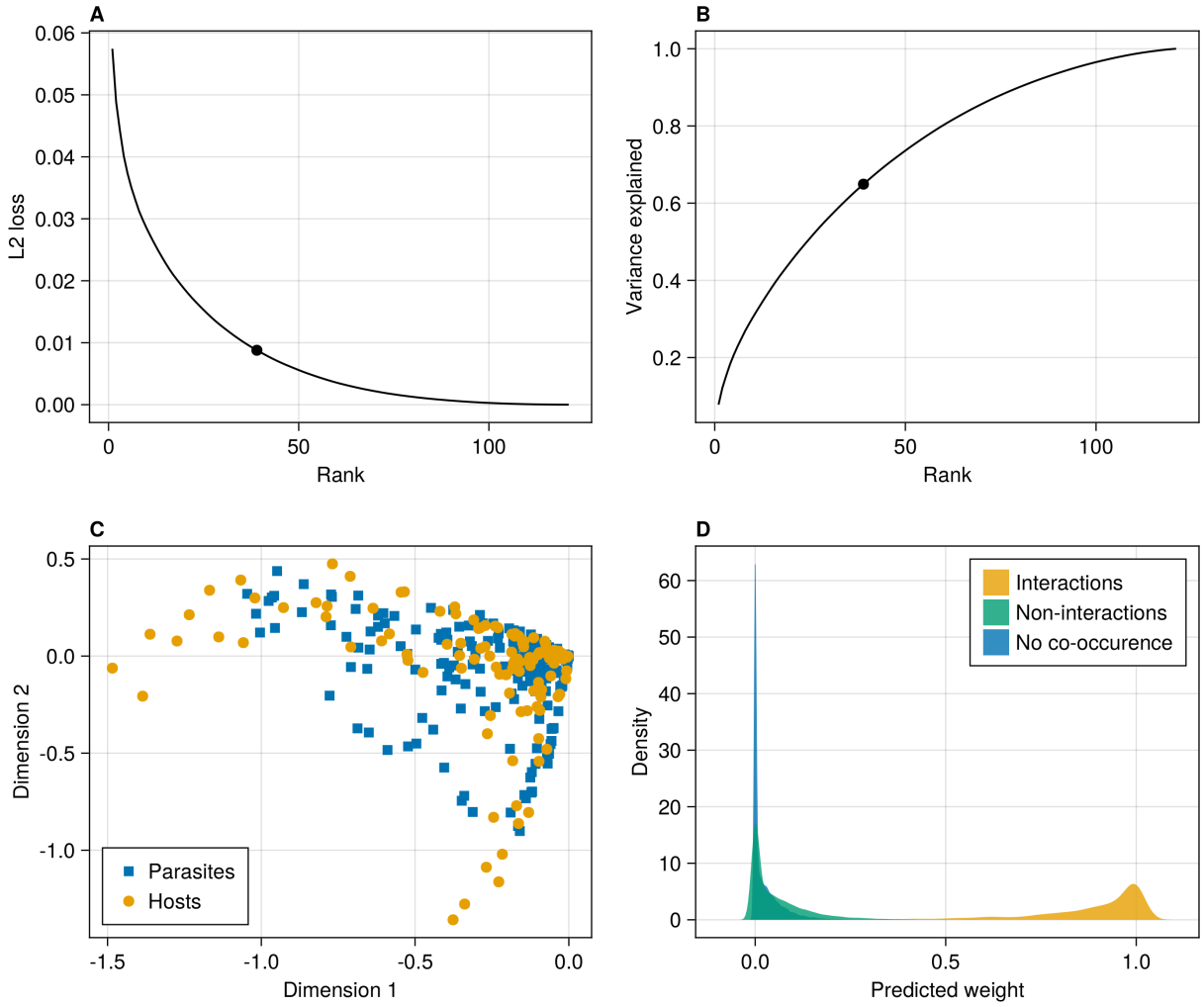


Fig. 2. Validation of an embedding for a host-parasite metaweb, using Random Dot Product Graphs. **A**, decrease in approximation error as the number of dimensions in the subspaces increases. **B**, increase in cumulative variance explained as the number of ranks considered increases; in **A** and **B**, the dot represents the point of inflexion in the curve (at rank 39) estimated using the finite differences method. **C**, position of hosts and parasites in the space of latent variables on the first and second dimensions of their respective subspaces (the results have been clamped to the unit interval). **D**, predicted interaction weight from the RDPG based on the status of the species pair in the metaweb.

In Figure 2, we focus on some statistical checks of the embedding. In panel **A**, we show that the averaged L_2 loss (*i.e.,* the sum of squared errors) between the empirical and reconstructed metaweb decreases when the number of dimensions (rank) of the subspace increases, with an inflection at 39 dimensions (out of 120 initially) according to the finite differences method. As discussed by Runghen et al., 2021, there is often a trade-off between the number of dimensions to use (more dimensions are more computationally demanding) and the quality of the representation. In panel **B**, we show the increase in cumulative variance explained at each rank, and visualize that using 39 ranks explains about 70% of the variance in the empirical metaweb. This is a different information from the L_2 loss (which is averaged across interactions), as it works on the eigenvalues of the embedding, and therefore captures higher-level features of the network. In panel **C**, we show positions of hosts and parasites on the first two dimensions of the left and right subspaces. Note that these values largely skew negative, because the first dimensions capture the coarse structure of the network: most pairs of species do not interact, and therefore have negative values. Finally in panel **D**, we show the predicted weight (*i.e.,* the result of the multiplication of the RDGP subspaces at a rank of 39) as a function of whether the interactions are observed, not-observed, or unknown due to lack of co-occurrence in the original dataset. This reveals that the observed interactions have higher predicted weights, although there is some overlap; the usual approach to identify potential interactions based on this information would be a thresholding analysis, which is outside the scope of this manuscript (and is done in the papers cited in this illustration). Because the values returned from RDPG are not bound to the unit interval, we performed a clamping of the weights to the unit space, showing a one-inflation in documented interactions, and a zero-inflation in other species pairs. This last figure crosses from the statistical into the ecological, by showing that species pairs with no documented co-occurrence have weights that are not distinguishable from species pairs with no documented interactions, suggesting that (as befits a host-parasite model) the ability to interact is a strong predictor of co-occurrence.

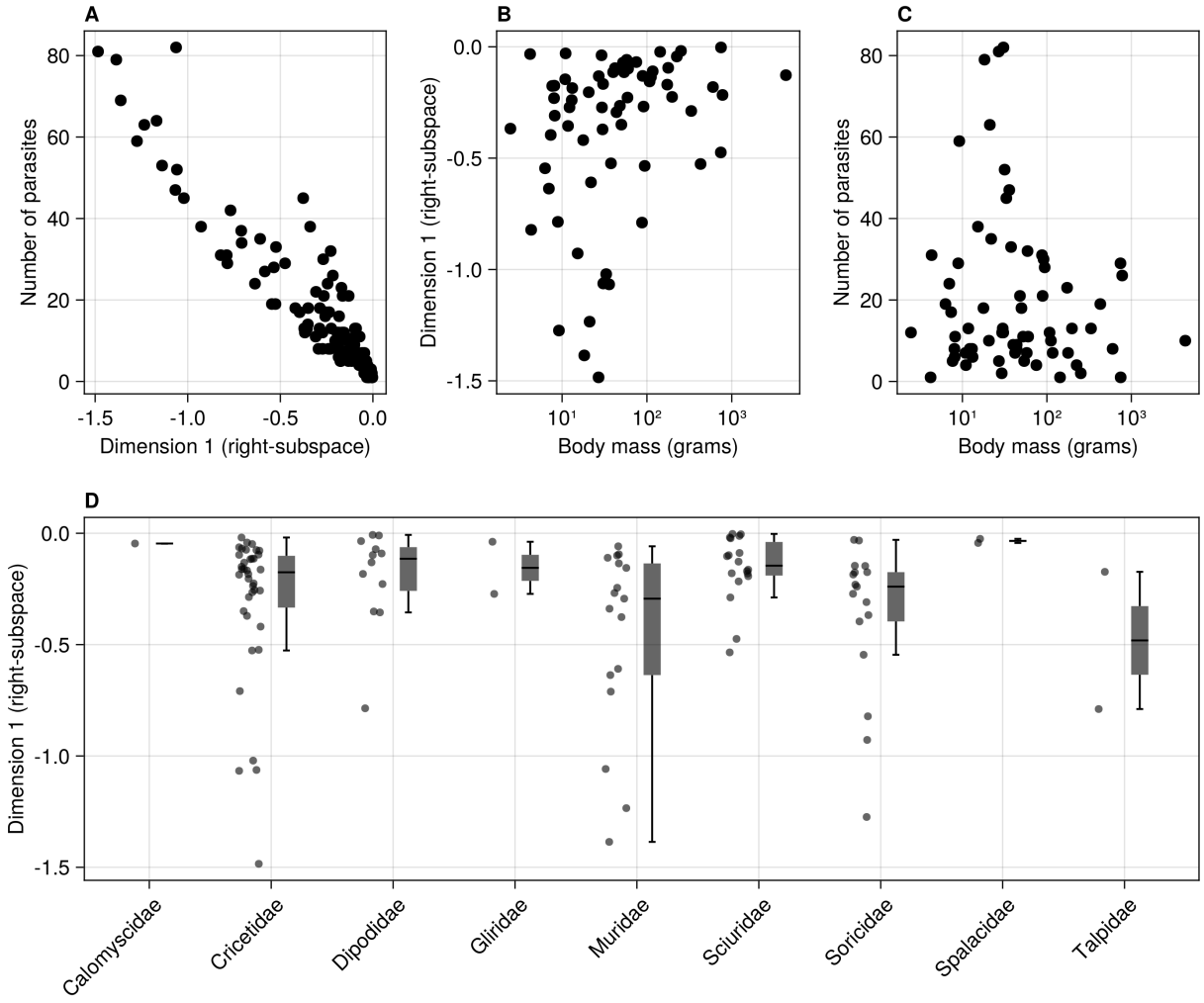


Fig. 3. Ecological analysis of an embedding for a host-parasite metaweb, using Random Dot Product Graphs. **A**, relationship between the number of parasites and position along the first axis of the right-subspace for all hosts, showing that the embedding captures elements of network structure at the species scale. **B**, weak relationship between the body mass of hosts (in grams) and the position alongside the same dimension. **C**, weak relationship between body mass of hosts and parasite richness. **D**, distribution of positions alongside the same axis for hosts grouped by taxonomic family.

The results of Figure 2 show that we can extract an embedding of the metaweb that captures enough variance to be relevant; specifically, this is true for both L_2 loss (indicating that RDPG is able to capture pairwise processes) and the cumulative variance explained (indicating that RDPG is able to capture network-level structure). Therefore, in Figure 3, we relate the values of latent variables for hosts to different ecologically-relevant data. In panel **A**, we show that host with a higher value on the first dimension have fewer parasites. This relates to the body size of hosts in the *PanTHERIA* database (Jones et al., 2009), as shown in panel **B**: interestingly, the position on the first axis is only weakly correlated to body mass of the host; this matches well established results showing that body size/mass is not always a direct predictor of parasite richness in terrestrial mammals (Morand and Poulin, 1998), a result we observe in panel **C**. Finally, in panel **D**, we can see how different taxonomic families occupy different positions on the first axis, with *e.g.,* Sciuridae being biased towards higher values. These results show how we can look for ecological informations in the output of network embeddings, which can further be refined into the selection of predictors for transfer learning.

2.5. The metaweb merges ecological hypotheses and practices

Metaweb inference seeks to provide information about the interactions between species at a large spatial scale, typically a scale large enough to be considered of biogeographic relevance (indeed, many of the examples covered in the introduction span areas larger than a country, some of them global). But as Herbert, 1965 rightfully pointed out, "[y]ou can't draw neat lines around planet-wide problems"; any inference of a metaweb must therefore contend with several novel, interwoven, families of problems. In this section, we outline three that we think are particularly important, and can discuss how they may addressed with subsequent data analysis or simulations, and how they emerge in the specific context of using embeddings; some of these issues are related to the application of these methods at the science-policy interface.

2.5.1. Identifying the properties of the network to embed

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 (through phylogenetic relatedness or similarity of functional traits) to make a reliable inference decreases. This is because transfer requires similarity (Figure 1). A diagnostic for the lack of similar species would likely be large confidence intervals during estimation of the values in the low-rank space. In other words, the representation of the original graph is difficult to transfer to the new problem. Alternatively, if the initial metaweb is too large (taxonomically), then the resulting embeddings would need to represent interactions between taxonomic groups that are not present in the new location. This would lead to a much higher variance in the starting dataset, and to under-dispersion in the target dataset, resulting in the potential under or over estimation of the strength of new predicted interactions. Llewelyn et al., 2022 provided compelling evidence for these situations by showing that, even at small spatial scales, the transfer of information about interactions becomes more challenging when areas rich with endemic species are considered. The lack of well documented metawebs is currently preventing the development of more concrete guidelines. The question of phylogenetic relatedness and distribution is notably relevant if the metaweb is assembled in an area with mostly endemic species (*e.g.,* a system that has undergone recent radiation or that has remained in isolation for a long period of time might not have an analogous system with which to draw knowledge from), and as with every predictive algorithm, there is room for the application of our best ecological judgement. Because this problem relates to distribution of species in the geographic or phylogenetic space, it can certainly be approached through assessing the performance of embedding transfer in simulated starting/target species pools.

2.5.2. Identifying the scope of the prediction to perform

The area for which we seek to predict the metaweb should determine the species pool on which the embedding is performed. Metawebs can be constructed by assigning interactions in a list of species within specific regions. The upside of this approach is that information

relevant for the construction of this dataset is likely to exist, as countries usually set conservation goals at the national level (Buxton et al., 2021), and as quantitative instruments are consequently designed to work at these scales (Turak et al., 2017); specific strategies are often enacted at smaller scales, nested within a specific country (Ray et al., 2021). However, there is no guarantee that these arbitrary boundaries are meaningful. In fact, we do not have a satisfying answer to the question of "where does an ecological network stop?", the answer to which would dictate the spatial span to embed/predict. Recent results by Martins et al., 2022 suggested that networks are shaped within eco-regions, with abrupt structural transitions from an eco-region to the next. Should this trend hold generally, this would provide an ecologically-relevant scale at which metawebs can be downscaled and predicted. Other solutions could leverage network-area relationships to identify areas in which networks are structurally similar (see *e.g.,* Fortin et al., 2021; Galiana et al., 2022; Galiana et al., 2018). Both of these solutions require ample pre-existing information about the network in space. Nevertheless, the inclusion of species for which we have data but that are not in the right spatial extent *may* improve the performance of approaches based on embedding and transfer, *if* they increase the similarity between the target and destination network. This proposal can specifically be evaluated by adding nodes to the network to embed, and assessing the performance of predictive models (see *e.g.,* Llewelyn et al., 2022).

2.6. Conclusion: metawebs, predictions, and people

Predictive approaches in ecology, regardless of the scale at which they are deployed and the intent of their deployment, originate in the framework that contributed to the ongoing biodiversity crisis (Adam, 2014) and reinforced environmental injustice (Choudry, 2013; Domínguez and Luoma, 2020). The risk of embedding this legacy in our models is real, especially when the impact of this legacy on species pools is being increasingly documented. This problem can be addressed by re-framing the way we interact with models, especially when models are intended to support conservation actions. Particularly on territories that

were traditionally stewarded by Indigenous people, we must interrogate how predictive approaches and the biases that underpin them can be put to task in accompanying Indigenous principles of land management (Eichhorn et al., 2019; No'kmaq et al., 2021). The discussion of "algorithm-in-the-loop" approaches that is now pervasive in the machine learning community provides examples of why this is important. Human-algorithm interactions are notoriously difficult and can yield adverse effects (Green and Chen, 2019; Stevenson and Doleac, 2021), suggesting the need to systematically study them for the specific purpose of, here, biodiversity governance. Improving the algorithmic literacy of decision makers is part of the solution (*e.g.,* Lamba et al., 2019; Mosebo Fernandes et al., 2020), as we can reasonably expect that model outputs will be increasingly used to drive policy decisions (Weiskopf et al., 2022). Our discussion of these approaches need to go beyond the technical and statistical, and into the governance consequences they can have. To embed data also embeds historical and contemporary biases that acted on these data, both because they shaped the ecological processes generating them, and the global processes leading to their measurement and publication. For a domain as vast as species interaction networks, these biases exist at multiple scales along the way, and a challenge for prediction is not only to develop (or adopt) new quantitative tools, but to assess the behavior of these tools in the proper context.

Box 2

Minding legacies shaping ecological datasets

In large parts of the world, boundaries that delineate geographic regions are merely a reflection the legacy of settler colonialism, which drives global disparity in capacity to collect and publish ecological data. Applying any embedding to biased data does not debias them, but rather embeds these biases, propagating them to the models using embeddings to make predictions. Furthermore, the use of ecological data itself is not an apolitical act (Nost and Goldstein, 2021): data infrastructures tend to be designed to answer questions within national boundaries (therefore placing contingencies on what is available to be embedded), their use often drawing upon, and reinforcing, territorial statecraft (see *e.g.,* Barrett, 2005). As per Machen and Nost, 2021, these biases are particularly important to consider when knowledge generated algorithmically is used to supplement or replace human decision-making, especially for governance (*e.g.,* enacting conservation decisions on the basis of model prediction). As information on networks is increasingly leveraged for conservation actions (see *e.g.,* Eero et al., 2021; Naman et al., 2022; Stier et al., 2017), the need to appraise and correct biases that are unwittingly propagated to algorithms when embedded from the original data is immense. These considerations are even more urgent in the specific context of biodiversity data. Long-term colonial legacies still shape taxonomic composition to this day (Lenzner et al., 2022; Raja, 2022), and much shorter-term changes in taxonomic and genetic richness of wildlife emerged through environmental racism (Schmidt and Garroway, 2022). Thus, the set of species found at a specific location is not only as the result of a response to ecological processes separate from human influence, but also the result of human-environment interaction as well as the result legislative/political histories.

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 Institute 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 the Fondation Courtois, and NSERC through the Discovery Grants and Discovery Accelerator Supplement programs. MJF is supported by an NSERC PDF and an RBC Post-Doctoral Fellowship.

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), 1153–1161. https://doi.org/10.1038/s41559-019-0950-y

Anowar, F., Sadaoui, S., & Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Computer Science Review*, *40*, 100378. https://doi.org/10.1016/j.cosrev.2021.100378

Arsov, N., & Mirceva, G. (2019). *Network Embedding: An Overview*. Retrieved 2020-06-02, from http://arxiv.org/abs/1911.11726

Athreya, A., Fishkind, D. E., Levin, K., Lyzinski, V., Park, Y., Qin, Y., Sussman, D. L., Tang, M., Vogelstein, J. T., & Priebe, C. E. (2017). Statistical inference on random dot product graphs: A survey. Retrieved 2022-09-28, from http://arxiv.org/abs/1709.05454

Barrett, S. (2005). *Environment and Statecraft: The Strategy of Environmental Treaty-Making* (1st ed.). Oxford University PressOxford. https://doi.org/10.1093/0199286094.001.0001

Bartomeus, I. (2013). Understanding linkage rules in plant-pollinator networks by using hierarchical models that incorporate pollinator detectability and plant traits. *PloS*

one, *8*(7), e69200. Retrieved 2017-02-21, from http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0069200

Bartomeus, I., Gravel, D., Tylianakis, J. M., Aizen, M. A., Dickie, I. A., & Bernard-Verdier, M. (2016). A common framework for identifying linkage rules across different types of interactions. *Functional Ecology*, *30*(12), 1894–1903. https://doi.org/10.1111/1365-2435.12666

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

Blanchet, F. G., Cazelles, K., & Gravel, D. (2020). Co-occurrence is not evidence of ecological interactions. *Ecology Letters*.

Bohan, D. A., Vacher, C., Tamaddoni-Nezhad, A., Raybould, A., Dumbrell, A. J., & Woodward, G. (2017). Next-Generation Global Biomonitoring: Large-scale, Automated Reconstruction of Ecological Networks. *Trends in Ecology & Evolution*. https://doi.org/10.1016/j.tree.2017.03.001

Botella, C., Dray, S., Matias, C., Miele, V., & Thuiller, W. (2022). An appraisal of graph embeddings for comparing trophic network architectures. *Methods in Ecology and Evolution*, *13*(1), 203–216. https://doi.org/10.1111/2041-210X.13738

Braga, J., Pollock, L. J., Barros, C., Galiana, N., Montoya, J. M., Gravel, D., Maiorano, L., Montemaggiori, A., Ficetola, G. F., Dray, S., & Thuiller, W. (2019). Spatial analyses of multi-trophic terrestrial vertebrate assemblages in Europe. *Global Ecology and Biogeography*, *28*(11), 1636–1648. https://doi.org/10.1111/geb.12981

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

Bramon Mora, 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

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

Catchen, M. D., Poisot, T., Pollock, L. J., & Gonzalez, A. (2023). The missing link: Discerning true from false negatives when sampling species interaction networks.

Cazelles, K., Araújo, M. B., Mouquet, N., & Gravel, D. (2016). A theory for species co-occurrence in interaction networks. *Theoretical Ecology*, *9*(1), 39–48. https://doi.org/10.1007/s12080-015-0281-9

Chami, I., Abu-El-Haija, S., Perozzi, B., Ré, C., & Murphy, K. (2022). Machine Learning on Graphs: A Model and Comprehensive Taxonomy. *Journal of Machine Learning Research*, *23*(89), 1–64. Retrieved 2022-06-24, from http://jmlr.org/papers/v23/20-852.html

Chen, D., Xue, Y., Fink, D., Chen, S., & Gomes, C. P. (2017). Deep Multi-species Embedding, 3639–3646. Retrieved 2022-01-12, from https://www.ijcai.org/proceedings/2017/509

Chen, H., Perozzi, B., Hu, Y., & Skiena, S. (2017). *HARP: Hierarchical Representation Learning for Networks*. Retrieved 2022-01-12, from http://arxiv.org/abs/1706.07845

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.

Cieslak, M. C., Castelfranco, A. M., Roncalli, V., Lenz, P. H., & Hartline, D. K. (2020). T-Distributed Stochastic Neighbor Embedding (t-SNE): A tool for eco-physiological

transcriptomic analysis. *Marine Genomics*, *51*, 100723. https://doi.org/10.1016/j.margen.2019.100723

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

Danisch, S., & Krumbiegel, J. (2021). Makie.jl: Flexible high-performance data visualization for Julia. *Journal of Open Source Software*, *6*(65), 3349. https://doi.org/10.21105/joss.03349

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), 65. https://doi.org/10.3390/land9030065

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.

Eero, M., Dierking, J., Humborg, C., Undeman, E., MacKenzie, B. R., Ojaveer, H., Salo, T., & Köster, F. W. (2021). Use of food web knowledge in environmental conservation and management of living resources in the Baltic Sea. *ICES Journal of Marine Science*, *78*(8), 2645–2663. https://doi.org/10.1093/icesjms/fsab145

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., Jacob, U., Kopp, J., Bosch, J., Castro-Urgal, R., Chacoff, N. P., Dalsgaard, B., de Sassi, C., Galetti, M., Guimarães, P. R., Lomáscolo, S. B., Martín González, A. M., Pizo, M. A., Rader, R., Rodrigo, A., Tylianakis, J. M., Vázquez, D. P., & Allesina, S.

(2013). The dimensionality of ecological networks. *Ecology Letters*, *16*(5), 577–583. https://doi.org/10.1111/ele.12081

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

Fricke, E. C., Ordonez, A., Rogers, H. S., & Svenning, J.-C. (2022). The effects of defaunation on plants' capacity to track climate change. *Science*. Retrieved 2022-01-13, from https://www.science.org/doi/abs/10.1126/science.abk3510

Galiana, N., Lurgi, M., Bastazini, V. A. G., Bosch, J., Cagnolo, L., Cazelles, K., Claramunt-López, B., Emer, C., Fortin, M.-J., Grass, I., Hernández-Castellano, C., Jauker, F., Leroux, S. J., McCann, K., McLeod, A. M., Montoya, D., Mulder, C., Osorio-Canadas, S., Reverté, S., … Montoya, J. M. (2022). Ecological network complexity scales with area. *Nature Ecology & Evolution*, 1–8. https://doi.org/10.1038/s41559-021-01644-4

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

Gaucher, S., Klopp, O., & Robin, G. (2021). Outlier detection in networks with missing links. *Computational Statistics & Data Analysis*, *164*, 107308. https://doi.org/10.1016/j.csda.2021.107308

Gibb, R., Albery, G. F., Becker, D. J., Brierley, L., Connor, R., Dallas, T. A., Eskew, E. A., Farrell, M. J., Rasmussen, A. L., Ryan, S. J., Sweeny, A., Carlson, C. J., & Poisot, T. (2021). Data Proliferation, Reconciliation, and Synthesis in Viral Ecology. *BioScience*, *71*(11), 1148–1156. https://doi.org/10.1093/biosci/biab080

Goebel, L. G. A., Vitorino, B. D., Frota, A. V. B., & dos Santos-Filho, M. (2023). Body mass determines the role of mammal species in a frugivore-large fruit interaction network in a Neotropical savanna. *Journal of Tropical Ecology*, *39*, e12. https://doi.org/10.1017/S0266467422000505

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Goyal, P., & Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, *151*, 78–94. https://doi.org/10.1016/j.knosys.2018.03.022

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

Gravel, D., Poisot, T., Albouy, C., Velez, L., & Mouillot, D. (2013). Inferring food web structure from predator-prey body size relationships (R. Freckleton, Ed.). *Methods in Ecology and Evolution*, *4*(11), 1083–1090. https://doi.org/10.1111/2041-210X.12103

Green, B., & Chen, Y. (2019). Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 90–99. https://doi.org/10.1145/3287560.3287563

Grover, A., & Leskovec, J. (2016). Node2vec: Scalable Feature Learning for Networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 855–864. https://doi.org/10.1145/2939672.2939754

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), 1–10. https://doi.org/10.1038/s42003-020-0962-9

Gupta, A., Furrer, R., & Petchey, O. L. (2022). Simultaneously estimating food web connectance and structure with uncertainty. *Ecology and Evolution*, *12*(3), e8643. https://doi.org/10.1002/ece3.8643

Gupta, A., Matta, P., & Pant, B. (2021). Graph neural network: Current state of Art, challenges and applications. *Materials Today: Proceedings*, *46*, 10927–10932. https://doi.org/10.1016/j.matpr.2021.01.950

Hadfield, J. D., Krasnov, B. R., Poulin, R., & Nakagawa, S. (2014). A Tale of Two Phylogenies: Comparative Analyses of Ecological Interactions. *The American Naturalist*, *183*(2), 174–187. https://doi.org/10.1086/674445

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

Hinton, G., & Roweis, S. T. (2002). Stochastic neighbor embedding. *NIPS*, *15*, 833–840.

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

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

Jones, K. E., Bielby, J., Cardillo, M., Fritz, S. A., O'Dell, J., Orme, C. D. L., Safi, K., Sechrest, W., Boakes, E. H., Carbone, C., Connolly, C., Cutts, M. J., Foster, J. K., Grenyer, R., Habib, M., Plaster, C. A., Price, S. A., Rigby, E. A., Rist, J., ... Purvis, A. (2009). PanTHERIA: A species-level database of life history, ecology, and geography of extant and recently extinct mammals: Ecological Archives E090-184 (W. K. Michener, Ed.). *Ecology*, *90*(9), 2648–2648. https://doi.org/10.1890/08-1494.1

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

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

Lenzner, B., Latombe, G., Schertler, A., Seebens, H., Yang, Q., Winter, M., Weigelt, P., van Kleunen, M., Pyšek, P., Pergl, J., Kreft, H., Dawson, W., Dullinger, S., & Essl, F. (2022). Naturalized alien floras still carry the legacy of European colonialism. *Nature Ecology & Evolution*, 1–10. https://doi.org/10.1038/s41559-022-01865-1

Llewelyn, J., Strona, G., Dickman, C. R., Greenville, A. C., Wardle, G. M., Lee, M. S. Y., Doherty, S., Shabani, F., Saltré, F., & Bradshaw, C. J. A. (2022). *Predicting predator-prey interactions in terrestrial endotherms using random forest* (preprint). Ecology. https://doi.org/10.1101/2022.09.02.506446

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

Malaterre, C., Dussault, A. C., Mermans, E., Barker, G., Beisner, B. E., Bouchard, F., Desjardins, E., Handa, I. T., Kembel, S. W., Lajoie, G., Maris, V., Munson, A. D., Odenbaugh, J., Poisot, T., Shapiro, B. J., & Suttle, C. A. (2019). Functional Diversity: An Epistemic Roadmap. *BioScience*, *69*(10), 800–811. https://doi.org/10.1093/biosci/biz089

Martins, L. P., Stouffer, D. B., Blendinger, P. G., Böhning-Gaese, K., Buitrón-Jurado, G., Correia, M., Costa, J. M., Dehling, D. M., Donatti, C. I., Emer, C., Galetti, M., Heleno, R., Jordano, P., Menezes, Í., Morante-Filho, J. C., Muñoz, M. C., Neuschulz, E. L., Pizo, M. A., Quitián, M., … Tylianakis, J. M. (2022). Global and regional ecological boundaries explain abrupt spatial discontinuities in avian frugivory interactions. *Nature Communications*, *13*(1), 6943. https://doi.org/10.1038/s41467-022-34355-w

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

Melnyk, K., Klus, S., Montavon, G., & Conrad, T. O. F. (2020). GraphKKE: Graph Kernel Koopman embedding for human microbiome analysis. *Applied Network Science*, *5*(1), 96. https://doi.org/10.1007/s41109-020-00339-2

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

Morand, S., & Poulin, R. (1998). Density, body mass and parasite species richness of terrestrial mammals. *Evolutionary Ecology*, *12*(6), 717–727. https://doi.org/10.1023/A:1006537600093

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), 7657. https://doi.org/10.3390/su12187657

Murphy, K. P. (2022). *Probabilistic machine learning: An introduction*. MIT Press. probml.ai

Naman, S. M., White, S. M., Bellmore, J. R., McHugh, P. A., Kaylor, M. J., Baxter, C. V., Danehy, R. J., Naiman, R. J., & Puls, A. L. (2022). Food web perspectives and methods for riverine fish conservation. *WIREs Water*, *n/a*(n/a), e1590. https://doi.org/10.1002/wat2.1590

Narayanan, A., Chandramohan, M., Venkatesan, R., Chen, L., Liu, Y., & Jaiswal, S. (2017). *Graph2vec: Learning Distributed Representations of Graphs*. Retrieved 2022-01-12, from http://arxiv.org/abs/1707.05005

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, s., 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

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

Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). DeepWalk: Online learning of social representations. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 701–710. https://doi.org/10.1145/2623330.2623732

Pichler, M., Boreux, V., Klein, A.-M., Schleuning, M., & Hartig, F. (2020). Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. *Methods in Ecology and Evolution*, *11*(2), 281–293. https://doi.org/10.1111/2041-210X.13329

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., 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). Imputing the mammalian virome with linear filtering and singular value decomposition. *arXiv:2105.14973 [q-bio]*.

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

Raja, N. B. (2022). Colonialism shaped today's biodiversity. *Nature Ecology & Evolution*, 1–2. https://doi.org/10.1038/s41559-022-01903-y

Ramasamy, D., & Madhow, U. (2015). Compressive spectral embedding: Sidestepping the SVD. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2015/file/4f6ffe13a5d75b2d6a3923922b3922e5-Paper.pdf

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

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

Saravia, L. A., Marina, T. I., Kristensen, N. P., De Troch, M., & Momo, F. R. (2021). Ecological network assembly: How the regional metaweb influences local food webs. *Journal of Animal Ecology*, *n/a*(n/a). https://doi.org/10.1111/1365-2656.13652

Schmidt, C., & Garroway, C. J. (2022). Systemic racism alters wildlife genetic diversity. *Proceedings of the National Academy of Sciences*, *119*(43), e2102860119. https://doi.org/10.1073/pnas.2102860119

Stevenson, M. T., & Doleac, J. L. (2021). Algorithmic Risk Assessment in the Hands of Humans. https://doi.org/10.2139/ssrn.3489440

Stier, A. C., Samhouri, J. F., Gray, S., Martone, R. G., Mach, M. E., Halpern, B. S., Kappel, C. V., Scarborough, C., & Levin, P. S. (2017). Integrating Expert Perceptions into Food Web Conservation and Management. *Conservation Letters*, *10*(1), 67–76. https://doi.org/10.1111/conl.12245

Stouffer, D. B., Camacho, J., Jiang, W., & Nunes Amaral, L. A. (2007). Evidence for the existence of a robust pattern of prey selection in food webs. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1621), 1931–1940. https://doi.org/10.1098/rspb.2007.0571

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

Strydom, T., Bouskila, S., Banville, F., Barros, C., Caron, D., Farrell, M. J., Fortin, M.-J., Hemming, V., Mercier, B., Pollock, L. J., Runghen, R., Dalla Riva, G. V., & Poisot, T. (2022). Food web reconstruction through phylogenetic transfer of low-rank network representation. *Methods in Ecology and Evolution*, *n/a*(n/a). https://doi.org/10.1111/2041-210X.13835

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

Surendran, S. (2013). Graph Embedding and Dimensionality Reduction - A Survey. *International Journal of Computer Science & Engineering Technology*, *4*(1). Retrieved 2022-06-28, from https://www.semanticscholar.org/paper/Graph-Embedding-and-Dimensionality-Reduction-A-Surendran/3f413d591e4b2b876e033eeb9390e232ad4826ca

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., & Mei, Q. (2015). LINE: Large-scale Information Network Embedding. *Proceedings of the 24th International Conference on World Wide Web*, 1067–1077. https://doi.org/10.1145/2736277.2741093

Thurman, L. L., Barner, A. K., Garcia, T. S., & Chestnut, T. (2019). Testing the link between species interactions and co-occurrence in a trophic network. *Ecography*, *0*. https://doi.org/10.1111/ecog.04360

Torres, L., Chan, K. S., & Eliassi-Rad, T. (2020). GLEE: Geometric Laplacian Eigenmap Embedding. *Journal of Complex Networks*, *8*(2), cnaa007. https://doi.org/10.1093/comnet/cnaa007

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

van der Maaten, L. (2009). Learning a Parametric Embedding by Preserving Local Structure. *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, 384–391. Retrieved 2022-06-19, from https://proceedings.mlr.press/v5/maaten09a.html

Wang, D., Cui, P., & Zhu, W. (2016). Structural Deep Network Embedding. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1225–1234. https://doi.org/10.1145/2939672.2939753

Wang, S., Arroyo, J., Vogelstein, J. T., & Priebe, C. E. (2021). Joint Embedding of Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(4), 1324–1336. https://doi.org/10.1109/TPAMI.2019.2948619

Wardeh, M., Baylis, M., & Blagrove, M. S. C. (2021). Predicting mammalian hosts in which novel coronaviruses can be generated. *Nature Communications*, *12*(1), 780. https://doi.org/10.1038/s41467-021-21034-5

Weiskopf, S. R., Harmáčková, Z. V., Johnson, C. G., Londoño-Murcia, M. C., Miller, B. W., Myers, B. J. E., Pereira, L., Arce-Plata, M. I., Blanchard, J. L., Ferrier, S., Fulton, E. A., Harfoot, M., Isbell, F., Johnson, J. A., Mori, A. S., Weng, E., & Rosa, I. M. D. (2022). Increasing the uptake of ecological model results in policy decisions to improve

biodiversity outcomes. *Environmental Modelling & Software*, *149*, 105318. https://doi.org/10.1016/j.envsoft.2022.105318

Williams, R. J., & Martinez, N. D. (2000). Simple rules yield complex food webs. *Nature*, *404*(6774), 180–183. https://doi.org/10.1038/35004572

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

Wu, D., Palmer, D. R., & Deford, D. R. (2021). Maximum a Posteriori Inference of Random Dot Product Graphs via Conic Programming. Retrieved 2022-09-28, from http://arxiv.org/abs/2101.02180

Xu, M. (2021). Understanding Graph Embedding Methods and Their Applications. *SIAM Review*, *63*(4), 825–853. https://doi.org/10.1137/20M1386062

Yan, S., Xu, D., Zhang, B., & Zhang, H.-J. (2005). Graph embedding: A general framework for dimensionality reduction. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, *2*, 830–837 vol. 2. https://doi.org/10.1109/CVPR.2005.170

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

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open*, *1*, 57–81. https://doi.org/10.1016/j.aiopen.2021.01.001

Chapter 3 Third article

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

by

Tanya Strydom[1], Salomé Bouskila[2], Francis Banville[3], Ceres Barros[4], Dominique Caron[5], Maxwell J. Farrell[6], Marie-Josée Fortin[7], Victoria Hemming[8], Benjamin Mercier[9], Laura Pollock[10], Rogini Runghen[11], Giulio V. Dalla Riva[12], and Timothée Poisot[13]

([1]) Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([2]) Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([3]) Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada
Université de Sherbrooke, Sherbrooke, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([4]) Department of Forest Resources Management, University of British Columbia, Vancouver, BC, Canada

([5]) McGill University, Montréal, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([6]) Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, ON, Canada

([7]) Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, ON, Canada

([8]) Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, BC, Canada

([9]) Université de Sherbrooke, Sherbrooke, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([10]) McGill University, Montréal, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([11]) Centre for Integrative Ecology, School of Biological Sciences, University of Canterbury, Canterbury, New Zealand

([12]) School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

([13]) Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

This article was submitted in Methods in Ecology and Evolution and can be found at https://doi.org/10. 1111/2041-210X.13835.

The main contributions of Tanya Strydom for this articles are presented. T.S., S.B. and T.P. designed the study and performed the analysis; G.V.D.R., M.J.F. and R.R. provided additional feedback on the analyses; D.C., B.M. and F.B. helped with data collection. All authors contributed to writing and editing the manuscript.

RÉSUMÉ. 1. Malgré leur importance dans de nombreux processus écologiques, la collecte de données et d'informations sur les interactions écologiques est une tâche extrêmement complexe. Pour cette raison, de nombreuses régions du monde révèlent un déficit de données en ce qui concerne les interactions entre les espèces et la structure des réseaux qui en résultent. Comme il est peu probable que la collecte de données soit suffisante à elle seule, les écologistes des communautés doivent adopter des méthodes prédictives.

2. Nous présentons un cadre méthodologique qui utilise l'incorporation graphique et l'apprentissage par transfert pour établir une liste prédictive des interactions trophiques d'un bassin d'espèces dont les interactions sont inconnues. Plus précisément, nous « apprenons » l'information (caractères latents) des espèces à partir d'un réseau d'interaction connu et inférons les caractères latents d'un autre bassin d'espèces pour lequel nous n'avons pas de données d'interaction a priori fondées sur leur lien phylogénétique avec les espèces du réseau connu. Les traits latents peuvent ensuite être utilisés pour prédire les interactions et construire un réseau d'interaction.

3. Ici, nous avons assemblé un méta-réseau (metaweb) pour les mammifères canadiens à partir des interactions dans le réseau trophique européen, en dépit d'un partage d'espèces communes de seulement 4% entre les deux sites. Les résultats du modèle prédictif sont comparés aux bases de données répertoriées d'interactions par paires, montrant que nous recouvrons correctement 91% des interactions connues.

4. Le cadre est intrinsèquement robuste, même lorsque le réseau connu est incomplet ou contient des interactions fallacieuses, en faisant un candidat idéal comme outil pour combler les lacunes en ce qui concerne les interactions entre les espèces. Nous fournissons des conseils sur la façon dont ce cadre peut être adapté en remplaçant certaines approches ou certains prédicteurs afin de le rendre plus généralement applicable.

Mots clés : estimation des caractères ancestraux, biogéographie, réseaux écologiques, intégration de réseaux, apprentissage par transfert

Abstract. 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 (latent traits) of species from a known interaction network and infer the latent traits of another 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.

Keywords: ancestral character estimation, biogeography, ecological networks, network embedding, transfer learning

3.1. 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 and 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.

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 and Yang, 2010; Torrey and 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 and 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 (Bramon Mora et al., 2018; Dalla Riva and Stouffer, 2016),

strong evolutionary signature on prey choice (Stouffer et al., 2012), and strong phylogenetic signature in food web intervality (Eklöf and 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.

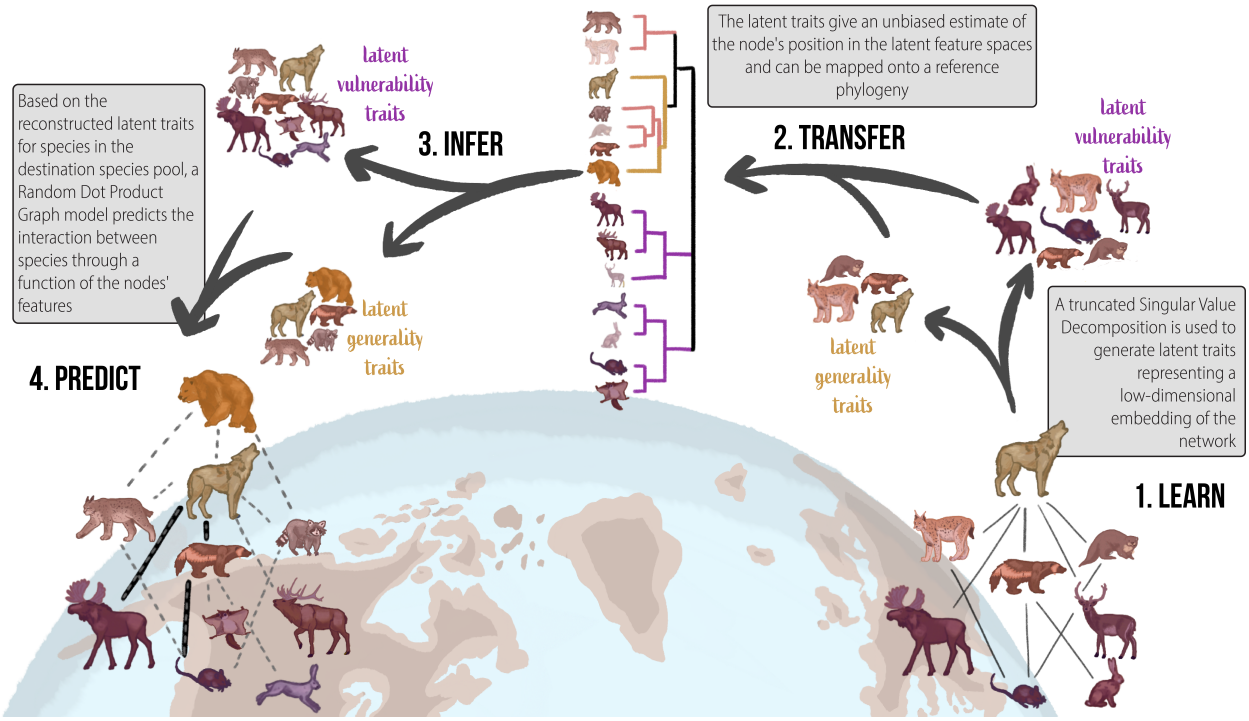


Fig. 1. Overview of the phylogenetic transfer learning (and prediction) of species interactions networks. Starting from an initial, known, network, we learn its representation through a graph embedding step (here, a truncated Singular Value Decomposition; Step 1), yielding a series of latent traits (latent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

In Figure 1, 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.,* we have no prior data on species interaction data for Canada in this instance). 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 the known (but fractional) data of trophic interactions present between Canadian mammals, our model achieves a predictive accuracy of approximately 91%.

3.2. Method description

The core point of our method is the transfer of knowledge of a known ecological network to predict interactions between species for another location for which the network is unknown (or partially known) and is summarized in the grey text boxes in Figure 1. 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).

3.2.1. Data used for the case study

We use data from the European metaweb assembled by Maiorano et al., 2020b. 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., 2020b in the European metaweb. After taxonomic cleaning and reconciliation the 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).

3.2.2. 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 and 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.

3.2.3. 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 a Random Dot Product Graph model (hereafter RDPG; S. J. Young and 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 and 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 an $m \times r$ semi unitary matrix and $\mathbf{V}'$ an $r \times n$ 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 and 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 for the network reconstruction 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.

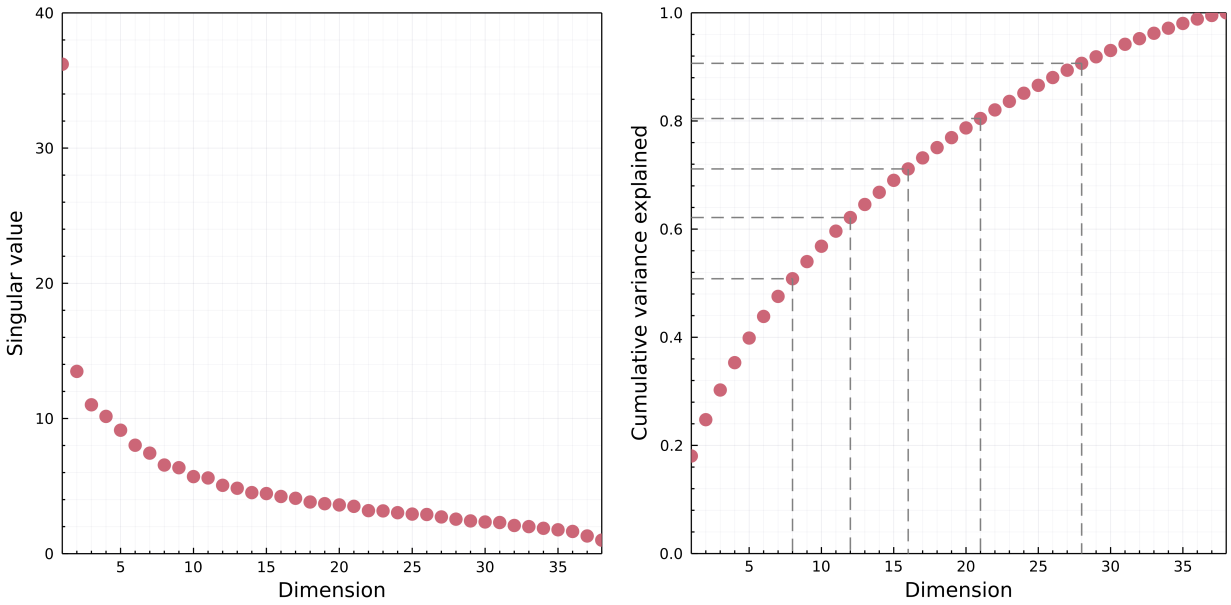


Fig. 2. Left: representation of the scree plot of the singular values from the t-SVD on the European metaweb. The scree plot shows no obvious drop in the singular values that may be leveraged to automatically detect a minimal dimension for embedding, after *e.g.,* Zhu and Ghodsi, 2006. Right: cumulative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

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., 2020b 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 section A.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.

3.2.4. 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 and 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 section A.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.

3.2.5. 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.

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 ($\hat{l}$) and right ($\hat{r}$) subspaces for the Canadian metaweb. These can be multiplied, to produce one matrix of real values. Because the entries in $\hat{l}$ 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

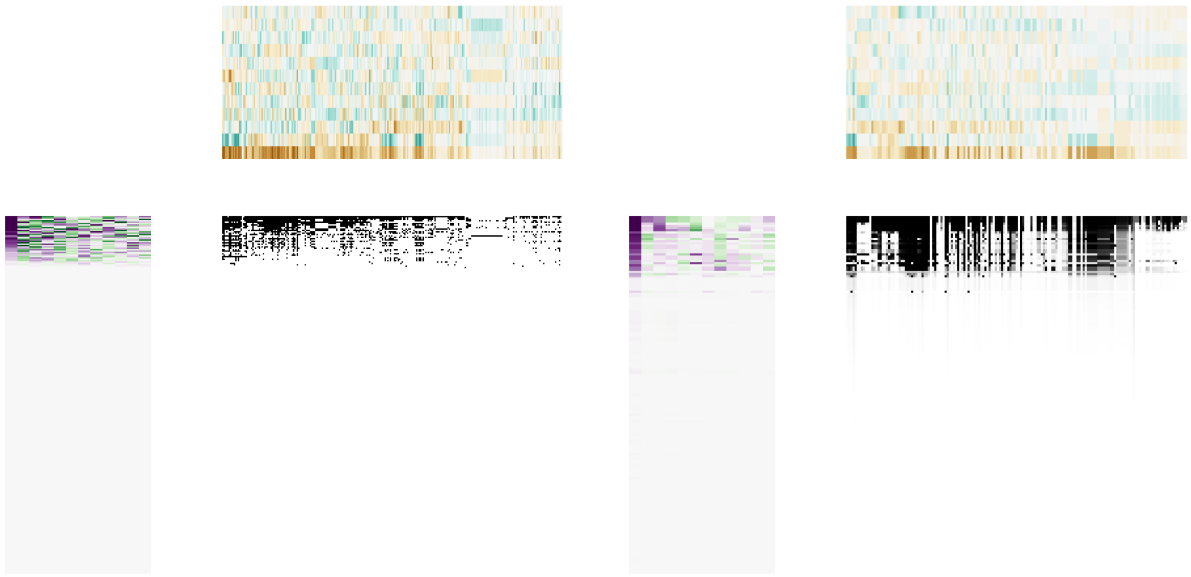


Fig. 3. Visual representation of the left (green/purple; left-side matrix) and right (green/brown; top matrix) subspaces, alongside the adjacency matrix of the food web they encode (grey scale). Where the color saturation is the magnitude of the latent trait value. The European metaweb is on the left, 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 Figure 6, the species with a value of 0 in the left subspace are species without any prey.

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 (Figure 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 section A.3 that not only can we observe differences in the multivariate space between the European and Canadian metawebs, 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.

3.2.6. 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.

First, we extracted the network 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 and 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

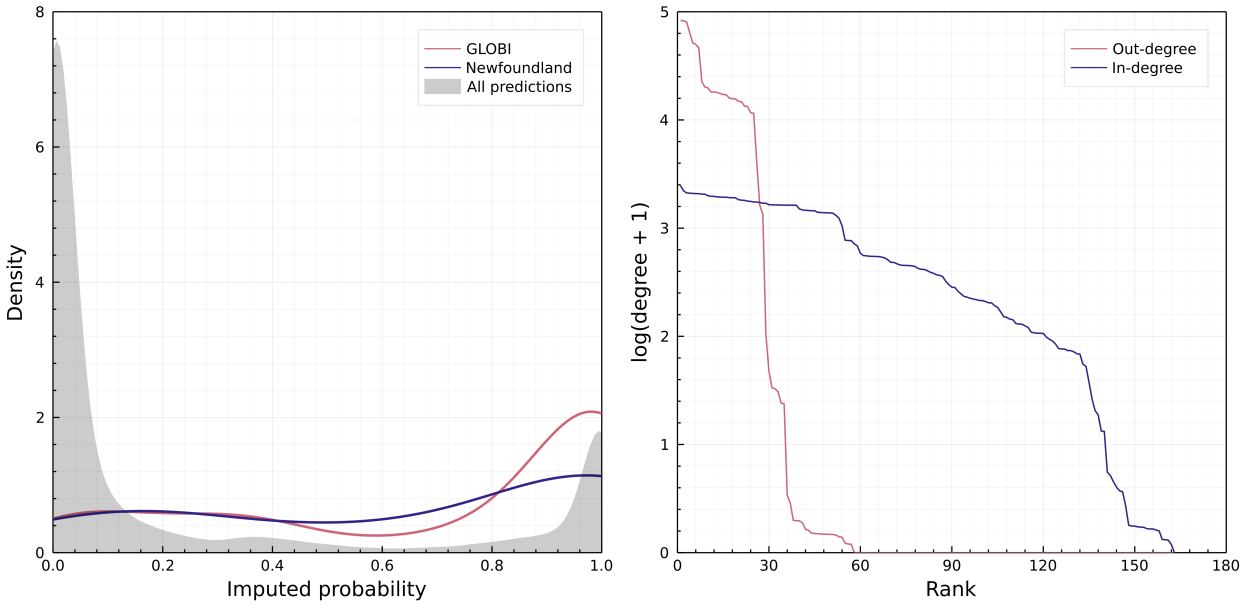


Fig. 4. Left: comparison of the probabilities of interactions assigned by the model to all interactions (grey curve), the subset of interactions found in GloBI (red), and in the Strong and Leroux, 2014 Newfoundland dataset (blue). The model recovers more interactions with a low probability compared to 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, where the rank is the maximal number of linearly independent columns (interactions) in the metaweb. This figure describes a flat, relatively short food web, in which there are few predators but a large number of preys.

91%. After performing this check, we set the probability of all interactions known to GloBI to 1.

Finally, we downloaded the data from Strong and 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 comparison of interaction densities for the inferred metaweb, and the Globi and Newfoundland is shown in Fig.4 and a table listing all interactions in the predicted Canadian metaweb can be found in the supplementary material.

Because the confidence intervals on the inferred trait space are probably over-estimates, we decided to apply a thresholding step to the interactions after data inflation (see Fig.5 showing the effect of varying the cutoff on $P(i \rightarrow j)$). Cirtwill and Hambäck, 2021 proposed

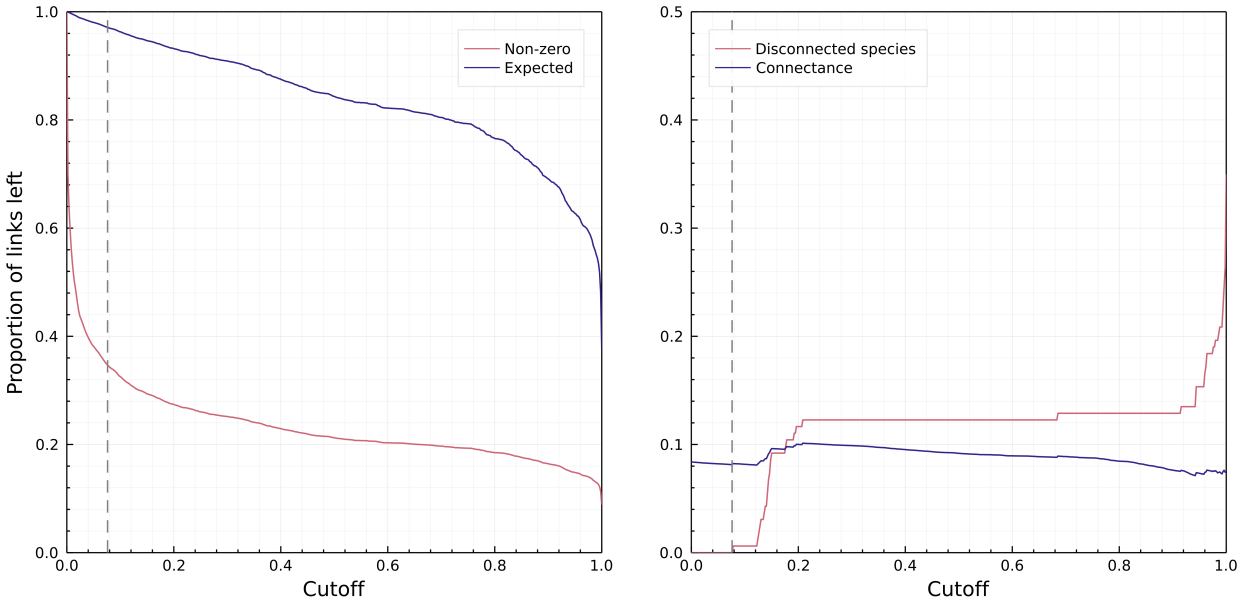


Fig. 5. Left: effect of varying the cutoff for probabilities to be considered non-zero on the number of unique links and on $\hat{L}$, the probabilistic estimate of the number of links assuming that all interactions are independent. Right: effect of varying the cutoff on the number of disconnected species,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

a number of strategies to threshold probabilistic networks. Their methodology assumes 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 and Hambäck, 2021 methodology highlights the need for additional development on network thresholding.

3.3. Results and discussion

Using a transfer learning framework we were able to construct a probabilistic metaweb and (as per Dunne, 2006) is a list of potential interactions, meaning that they will not

necessarily be realized wherever the two species co-occur. 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 and 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.

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.

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).

Finally, it should be noted that the framework we have presented is amenable to changes lending to applicability to a broad range of potential scenarios. For example 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 and Stouffer, 2016); this being said, there are other ways to embed graphs (Arsov and Mirceva, 2019; Cai et al., 2017; Cao et al., 2019), which can be used as alternatives. Regarding the transfer step it is possible to use distinct trees if working with distinct clades (such as pollination networks) or an alternative measure of similarity (transfer medium) such as 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.

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 Institute 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. MJF is supported by an NSERC PDF and an RBC Post-Doctoral Fellowship

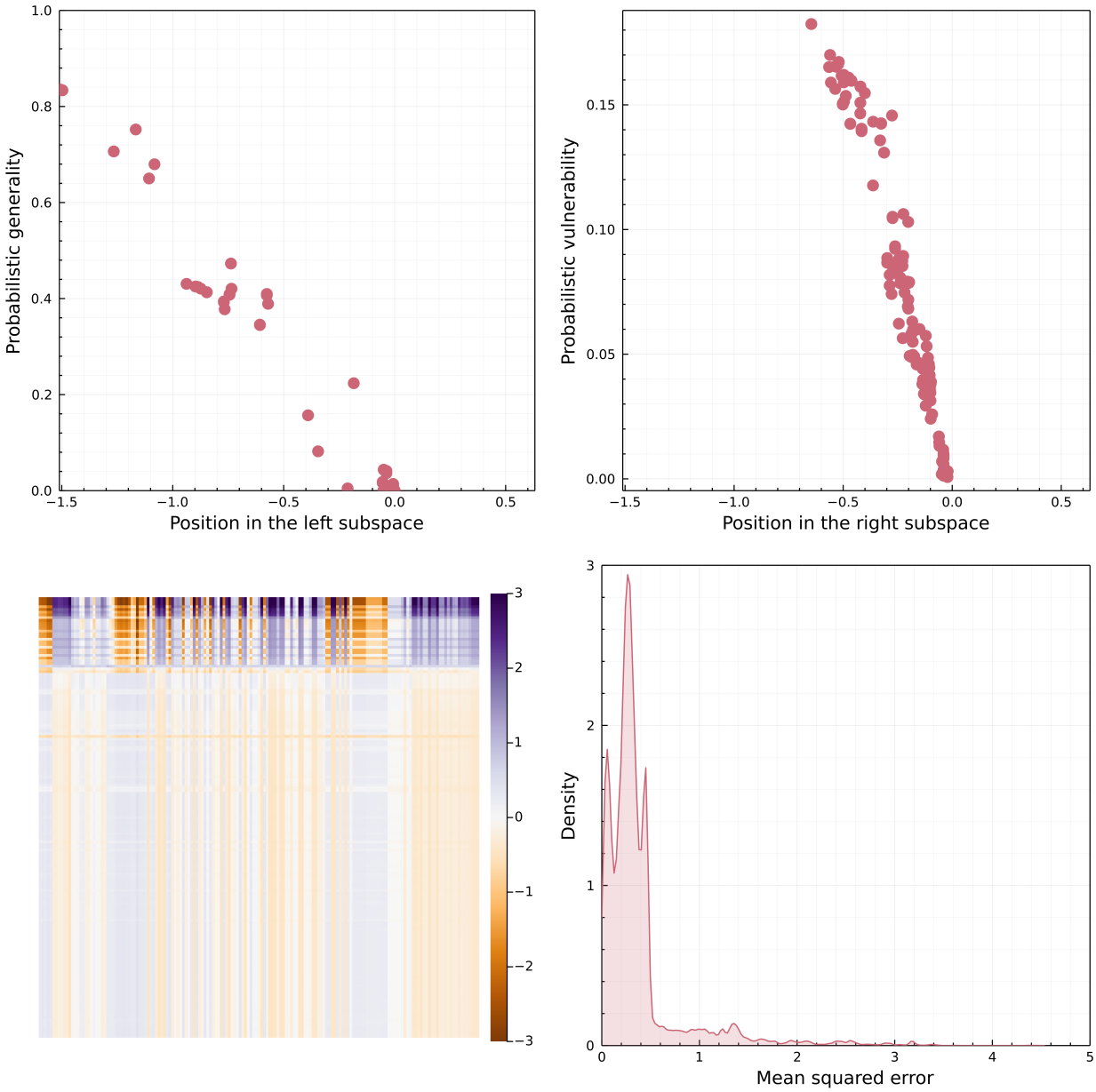


Fig. 6. Top: biological significance of the first dimension. Left: there is a linear relationship between the values on the first dimension of the left subspace and the generality, *i.e.,* the relative number of preys, *sensu* Schoener, 1989. Species with a value of 0 in this subspace are at the bottom-most 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-scores of the actual configuration model for the reconstructed network and the prediction based only on the first dimension (with a deeper saturation indicating a bigger difference in scores). Right: distribution of the differences in the left panel.

References

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), 1153–1161. https://doi.org/10.1038/s41559-019-0950-y

Arsov, N., & Mirceva, G. (2019). *Network Embedding: An Overview*. Retrieved 2020-06-02, from 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. B. (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), 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

Bramon Mora, 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

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). Consumer–Resource Body-Size Relationships in Natural Food Webs. *Ecology*, *87*(10), 2411–2417. https://doi.org/10.1890/0012-9658(2006)87[2411:CBRINF]2.0.CO;2

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

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

Cirtwill, A. R., Eklöf, 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*(ja). 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

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

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

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.

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

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., den Hollander, F., & 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

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

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 Biology*, *14*(9), e1002559. https://doi.org/10.1371/journal.pbio.1002559

Jordano, P. (2016b). Sampling networks of ecological interactions. *Functional Ecology*. 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

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

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

Maiorano, L., Montemaggiori, A., Ficetola, G. F., O'Connor, L., & Thuiller, W. (2020b). 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

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

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

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

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

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*, jbi.14127. 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). Imputing the mammalian virome with linear filtering and singular value decomposition. *arXiv:2105.14973 [q-bio]*.

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

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

Reeve, R., Leinster, T., Cobbold, C. A., Thompson, J., Brummitt, N., Mitchell, S. N., & Matthews, L. (2016). *How to partition diversity*. Retrieved 2021-06-22, from 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

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.

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<32::AID-CNCR2820030106>3.0.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

Chapter 4 Fourth article

SVD Entropy reveals the high complexity of ecological networks

by

Tanya Strydom[1], Giulio V. Dalla Riva[2], and Timothée Poisot[3]

([1]) Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([2]) School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

([3]) Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

This article was submitted in Frontiers in Ecology and Evolution and can be found at https://doi.org/10.3389/fevo.2021.623141.

The main contributions of Tanya Strydom for this articles are presented.

TP and TS designed the study and edited the manuscript for submission. TS performed the analysis and wrote the manuscript. All authors contributed to the article and approved the submitted version.

Résumé. Quantifier la complexité des réseaux écologiques reste difficile à évaluer. Principalement, la complexité a été définie sur la base de la complexité structurelle (ou comportementale) du système. Ces définitions ignorent la notion de « complexité physique », qui peut mesurer la quantité d'informations contenues dans un réseau écologique et la difficulté de les compresser. Nous présentons respectivement le déficit de rang relatif et l'entropie SVD comme mesures de la complexité "externe" et "interne". En utilisant des réseaux écologiques bipartites, nous constatons qu'ils présentent tous une complexité physique très élevée, presque maximale. Les réseaux de pollinisation, en particulier, sont plus complexes que d'autres types d'interactions. De plus, nous constatons que l'entropie SVD est liée à d'autres mesures structurelles de complexité (imbrication, connectivité et rayon spectral), mais ne renseigne pas sur la résilience d'un réseau lors de l'utilisation de cascades d'extinction simulées, ce qui a déjà été rapporté pour des mesures structurelles de complexité. Nous soutenons que l'entropie SVD fournit une mesure fondamentalement plus "correcte" de la complexité des réseaux et devrait être ajoutée à la boîte à outils des descripteurs des réseaux écologiques à l'avenir.

Mots clés : décomposition en valeurs singulières, complexité physique, réseau bipartite, entropie, pollinisation, analyse de réseaux écologiques

Abstract. Quantifying the complexity of ecological networks has remained elusive. Primarily, complexity has been defined on the basis of the structural (or behavioural) complexity of the system. These definitions ignore the notion of "physical complexity," which can measure the amount of information contained in an ecological network, and how difficult it would be to compress. We present relative rank deficiency and SVD entropy as measures of "external" and "internal" complexity, respectively. Using bipartite ecological networks, we find that they all show a very high, almost maximal, physical complexity. Pollination networks, in particular, are more complex when compared to other types of interactions. In addition, we find that SVD entropy relates to other structural measures of complexity (nestedness, connectance, and spectral radius), but does not inform about the resilience of a network when using simulated extinction cascades, which has previously been reported for structural measures of complexity. We argue that SVD entropy provides a fundamentally more "correct" measure of network complexity and should be added to the toolkit of descriptors of ecological networks moving forward.

Keywords: singular value decomposition, physical complexity, bipartite network, entropy, pollination, ecological network analysis

4.1. Introduction

Ecologists have turned to network theory because it offers a powerful mathematical formalism to embrace the complexity of ecological communities (Bascompte and Jordano, 2007). Indeed, analysing ecological systems as networks highlighted how their structure ties into ecological properties and processes (Poulin, 2010; Proulx et al., 2005), and there has been a subsequent explosion of measures that purport to capture elements of network structure, to be related to the ecology of the system they describe (Delmas et al., 2018). Since the early days of network ecology, ecological networks have been called "complex". This sustained interest for the notion of complexity stems, in part, from the strong ties it has to stability (Landi et al., 2018). As such, many authors have looked for clues, in the network structure, as to why the networks do not collapse (Borrelli, 2015; Brose et al., 2006; Gravel et al., 2016; Staniczenko et al., 2013). Yet decades of theoretical refinements on the relationship between complexity and stability had a hard time when rigorously tested on empirical datasets

(Jacquet et al., 2016); although ecological networks may be complex, our current measures of complexity do not translate into predictions about stability.

Surprisingly, *complexity* itself has proven an elusive concept to define in a rigorous way. It has over time been defined as connectance (Rozdilsky and Stone, 2001), as measures of the diversity of species or their interactions (Landi et al., 2018), or as a combination of species richness and trophic diversity (Duffy et al., 2007). In short, network ecology as a field readily assumes that because we have more information about a system, or because this system has more components, or simply because this system can be expressed as a network, it follows that the system is complex. But such a diversity of definitions, for a concept that is so central to our quest to understand network stability, decreases the clarity of what complexity means, and what all of these alternative definitions do actually capture. This is a common thread in some measures of ecological network structure, as has been discussed at length for the various definitions of nestedness (Ulrich et al., 2009).

None of the previous definitions of complexity are formally wrong, in that they do capture an aspect of complexity that ultimately ties to the behaviour of the system, *i.e.,* its low predictability over time. Yet (Adami, 2002) provides a compelling argument for why the complexity of the behaviour does not necessarily reflects the complexity of the system; in fact, one would be very hard pressed to think of a more simple system than the logistic map used by May, 1976 to illustrate how easily complexity of behaviour emerges. Rather than yielding to the easy assumption that a system will be complex because it has many parts, or because it exhibits a complex behaviour, Adami, 2002 suggests that we focus on measuring "physical complexity", *i.e.,* the amount of information required to encode the system, and how much signal this information contains. Complex systems, in this perspective, are those who cannot easily be compressed - and this is a notion we can explore for the structure of ecological networks.

Ecological networks are primarily represented by their adjacency matrices, *i.e.,* a matrix in which every entry represents a pair of species, which can take a value of 1 when the two species interact, and a value of 0 when they do not. These matrices (as any matrices)

can easily be factorised using Singular Value Decomposition (Forsythe and Moler, 1967; Golub and Reinsch, 1971), which offers two interesting candidate measures of complexity for ecological networks (both of which we describe at length in the methods). The first measure is the rank of the matrix, which works as an estimate of "external complexity", in that it describes the dimension of the vector space of this matrix, and therefore the number of linearly independent rows (or columns) of it. From an ecological standpoint, this quantifies the number of unique "strategies" represented in the network: a network with two modules that are distinct complete graphs has a rank of 2. The second measure is an application of the entropy measure of Shannon, 1948 to the non-zero singular values of the matrix obtained through SVD. This so-called SVD entropy measures the extent to which each rank encodes an equal amount of information, as the singular values capture the importance of each rank to reconstruct the original matrix; this approach therefore serves as a measure of "internal complexity".

In this manuscript, we present and evaluate the use of both the rank and SVD entropy of ecological networks as alternative and more robust measures of complexity when compared to traditional approaches to defining complexity. This is done by using a collection of 220 bipartite networks from various types of interaction, sizes, connectances, and environments. We show that while the rank of the adjacency matrix holds little information, SVD entropy functions as an appropriate quantification of the complexity of ecological systems. Notably, SVD entropy is an intuitive, robust, non-structural approach to defining the (surprisingly high) complexity of ecological networks, by relating them to their 'physical' as opposed to 'behavioural' complexity. In this process we showcase a breakdown in the assumption that all measures of complexity of networks are indicative of their robustness to extinctions. Finally, we show that, despite their high complexity, observed networks are less complex when compared to pseudo-random networks, especially for larger networks. We propose that taking a physical approach to quantifying the complexity of ecological networks is a step in the right direction to unifying how we define complexity in the context of ecological

networks, as it restores other measures (like connectance and nestedness) to their original role and signification.

4.2. Data and methods

We used all bipartite networks contained in the `web-of-life.es` database. This database extracted species interaction networks from supplementary materials across all inhabited continents and covers a large array of sampling years, environments, organisms, and sampling methodologies. As such, this dataset is particularly suited to describe general trends across *all* ecological networks. We specifically worked on the version of this dataset distributed with the `EcologicalNetworks.jl` package (Poisot et al., 2019) for the *Julia* (Bezanson et al., 2017) programming language, in which all analyses were conducted. Using bipartite networks means that interacting species are split into two sets (or interacting groups) and along different dimensions in the interaction matrix. Thus, columns in the matrix represent one group (or type) of species and rows represent the other group of species involved in the interaction. Because SVD gives similar results on the matrix and its transpose, it captures the complexity of both sides of the system at once. A summary of the dataset is given in Table 1.

Interaction type	Sample size	Latitude range	Richness (top)	Richness (bottom)
Host-Parasite	51	38.77 → 72.65	20.47	12.23
Plant-Ant	4	-16.11 → -2.40	18.75	21.75
Plant-Herbivore	4	30.20 → 64.91	49.5	29.25
Pollination	134	-43.09 → 81.81	40.22	18.02
Seed Dispersal	33	-28.95 → 53.05	18.75	25.12

Table 1. Overview of the `web-of-life.es` dataset. We used all networks with up to 500 species. Although there are spatial biases in the sampling of interaction types (and some interaction types being under-represented), this dataset covers a range of latitudes from -43 degrees south to 81 degrees north. The average richness of the top and bottom level of the bipartite networks are also given in the last columns.

4.2.1. Estimating complexity with rank deficiency

The rank of $\mathbf{A}$ (noted as $r = \mathrm{rk}(\mathbf{A})$) is the dimension of the vector space spanned by the matrix and corresponds to the number of linearly independent rows or columns; therefore, the maximum rank of a matrix ($M = \mathrm{rk}_{\max}(\mathbf{A})$) will always be equal to the length of the shortest dimension of $\mathbf{A}$, which ecologically speaking is the richness of the least species-rich compartment of the bipartite network (or the richness in the case of unipartite networks). A matrix is "full-ranked" when $r = M$, *i.e.,* all of its rows/columns are unique. Matrices that are not full-ranked are called rank deficient, and we can measure rank deficiency using $d = M - r$. So as to control for the difference in species richness of the different networks, we report the relative rank deficiency, *i.e.,* expressed as a ratio between rank deficiency and the maximal rank:

$$D = 1 - \frac{r}{M}$$

This measure returns values between 0 (the matrix is full ranked) and $1 - M^{-1} \approx 1$ (the matrix has rank 1). This serves as a coarse estimate of complexity, as the more unique columns/rows are in the matrix, the larger this value will be. Yet it may also lack sensitivity, because it imposes a stringent test on uniqueness, which calls for more quantitative approaches to complexity.

4.2.2. Estimating complexity with SVD entropy

Singular Value Decomposition (SVD) is the factorisation of a matrix $\mathbf{A}$ (where $\mathbf{A}_{m,n} \in \mathbb{B}$ in our case, but SVD works for matrices of real numbers as well) into the form $\mathbf{U} \cdot \mathbf{\Sigma} \cdot \mathbf{V}^T$. Where $\mathbf{U}$ is an $m \times m$ orthogonal matrix and $\mathbf{V}$ an $n \times n$ orthogonal matrix. The columns in these matrices are, respectively, the left- and right-singular vectors of $\mathbf{A}$, were $\mathbf{U} = \mathbf{A}\mathbf{A}^T$ and $\mathbf{V} = \mathbf{A}^T\mathbf{A}$. $\mathbf{\Sigma}$ is a matrix that only contains non-negative σ values along its diagonal and all other entries are zero. Where $\sigma_i = \Sigma ii$, which contains the singular values of $\mathbf{A}$. When the values of σ are arranged in descending order, the singular values ($\mathbf{\Sigma}$) are unique, though the singular vectors ($\mathbf{U}$ and $\mathbf{V}$) may not be.

After the Eckart-Young-Mirsky theorem (Eckart and Young, 1936; Golub et al., 1987), the number of non-zero entries (after rounding of small values if required due to numerical precision issues in computing the factorisation) in σ is the rank of matrix $\mathbf{A}$. For the sake of simplicity in notation, we will use $k = \text{rk}(\mathbf{A})$) for the rank of the matrix. Because only the first k elements of σ are non-zero, and that the result of the SVD is a simple matrix multiplication, one can define a truncated SVD containing only the first k singular values.

Intuitively, the singular value i (σ_i) measures how much of the dataset is (proportionally) explained by each vector - therefore, one can measure the entropy of σ following Shannon, 1948. High values of SVD entropy reflects that all vectors are equally important, *i.e.,* that the structure of the ecological network cannot efficiently be compressed, and therefore indicates high complexity (Gu and Shao, 2016). Because networks have different dimensions, we use Pielou's evenness (Pielou, 1975) to ensure that values are lower than unity, and quantify SVD entropy, using $s_i = \sigma_i / \text{sum}(\sigma)$ as:

$$J = -\frac{1}{\ln(k)} \sum_{i=1}^{k} s_i \cdot \ln(s_i)$$

4.3. Results and discussion

4.3.1. Most ecological networks are close to full-rank

The majority (63% of our dataset) of bipartite ecological networks have a relative rank deficiency of 0 (Figure 1), which indicates that all species have different and unique interaction lists. Interestingly, the networks that had a comparatively larger relative rank deficiency tended to be smaller ones. Yet because most of the networks return the same value, matrix rank does not appear to be a useful or discriminant measure of network complexity. Another striking result (from Figure 1) is that the SVD entropy of ecological networks is really large – although the value can range from 0 to 1, all ecological networks had SVD entropy larger than 0.8, which is indicative of a strong complexity.

As expected following the observation that ecological networks are overwhelmingly full ranked, we do not see a relationship between SVD entropy and relative rank deficiency,

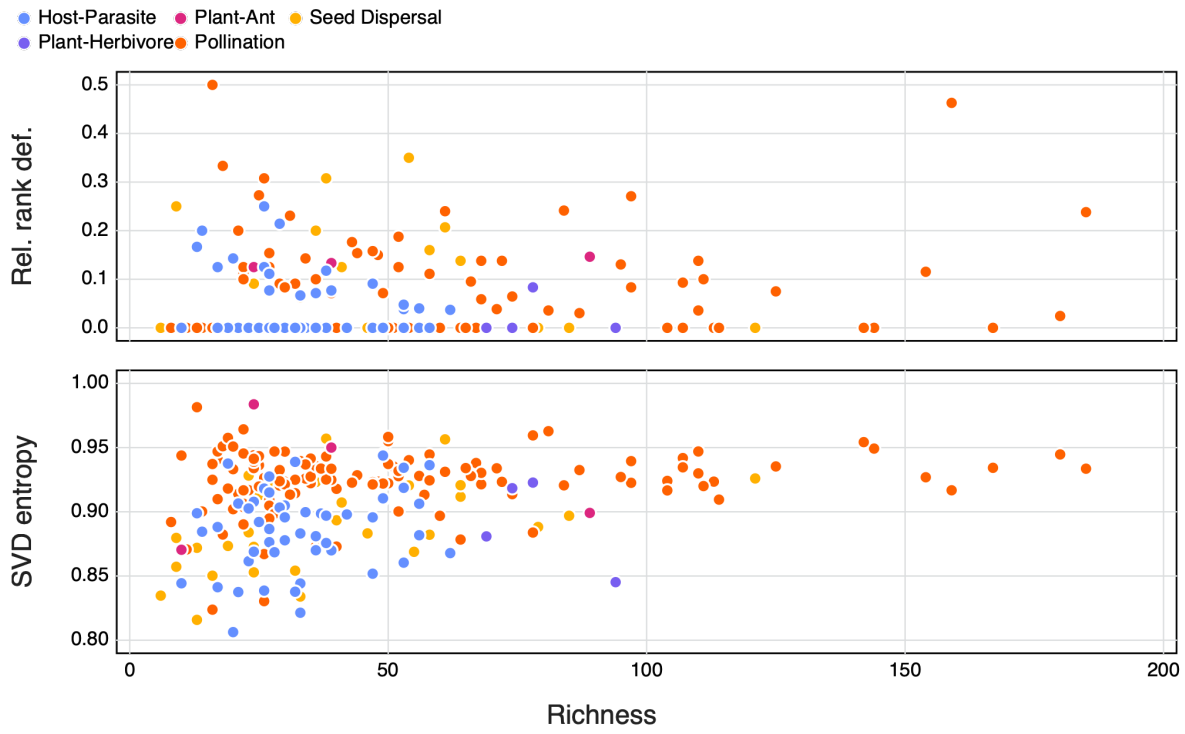


Fig. 1. The relationship between network richness and relative rank deficiency, and SVD entropy. The different types of interactions are indicated by the colours.

neither do we observe differences between interaction types (2). Based on these results, we feel confident that SVD entropy provides a more informative measure of the complexity of ecological networks, and will use it moving forward.

4.3.2. Most elements of network structure capture network complexity

We compared SVD entropy to some of the more common measures of complexity, namely nestedness (η, as per Bastolla et al., 2009), connectance (Co), and the spectral radius of the network (ρ, following Staniczenko et al., 2013). All of these measures are positively correlated, especially over the range of connectances covered by empirical bipartite ecological networks.

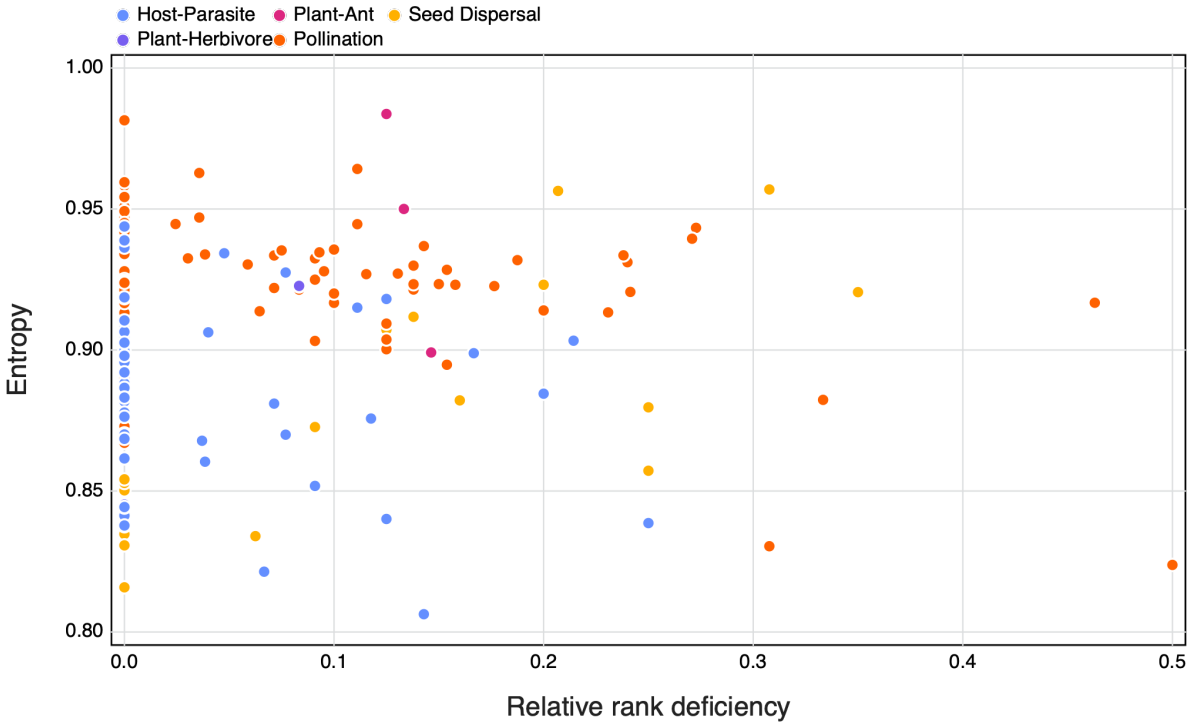


Fig. 2. The relationship between SVD entropy and the relative rank deficiency of different species interaction networks Colours indicate the different interaction types of the networks.

Nestedness is calculated based on the number of interactions shared between species pairs and is a measure of the degree of overlap between species links (or strategies) in the community, where larger assemblages are made up of a subset of smaller ones that share common interactions. Networks with a higher degree of nestedness could be considered simpler when compared to networks with a lower degree of nestedness. Connectance is the realised number of interactions (links) in an ecological network and is calculated as the fraction of the total number of realised interactions (or links) and the maximum number of possible interactions in a network (Martinez, 1992). This has been shown to be a good estimate of a community's resilience to perturbation (Dunne et al., 2002). The spectral radius of a matrix is the largest absolute value of its eigenvalues, which, in addition to being presented as a measure of network complexity has also been suggested as an indicator of the ability of a system to dampen disturbances (Phillips, 2011).

We find that SVD entropy has a clear negative relationship with nestedness, spectral radius, and connectance (Figure 3). As in Figure 5, mutualistic networks tend to be more complex, and they also are both sparser and less nested than other types of networks. Bastolla et al., 2009 give a convincing demonstration that mutualistic networks are shaped to minimise competition – this can be done by avoiding to duplicate overlap in interactions, thereby resulting in a network that is close to full rank, and with high SVD entropy. Interestingly, Figure 3 suggests that both nestedness and connectance measure the *lack* of complexity in an ecological network, which contrasts to how they may commonly be viewed (Landi et al., 2018).

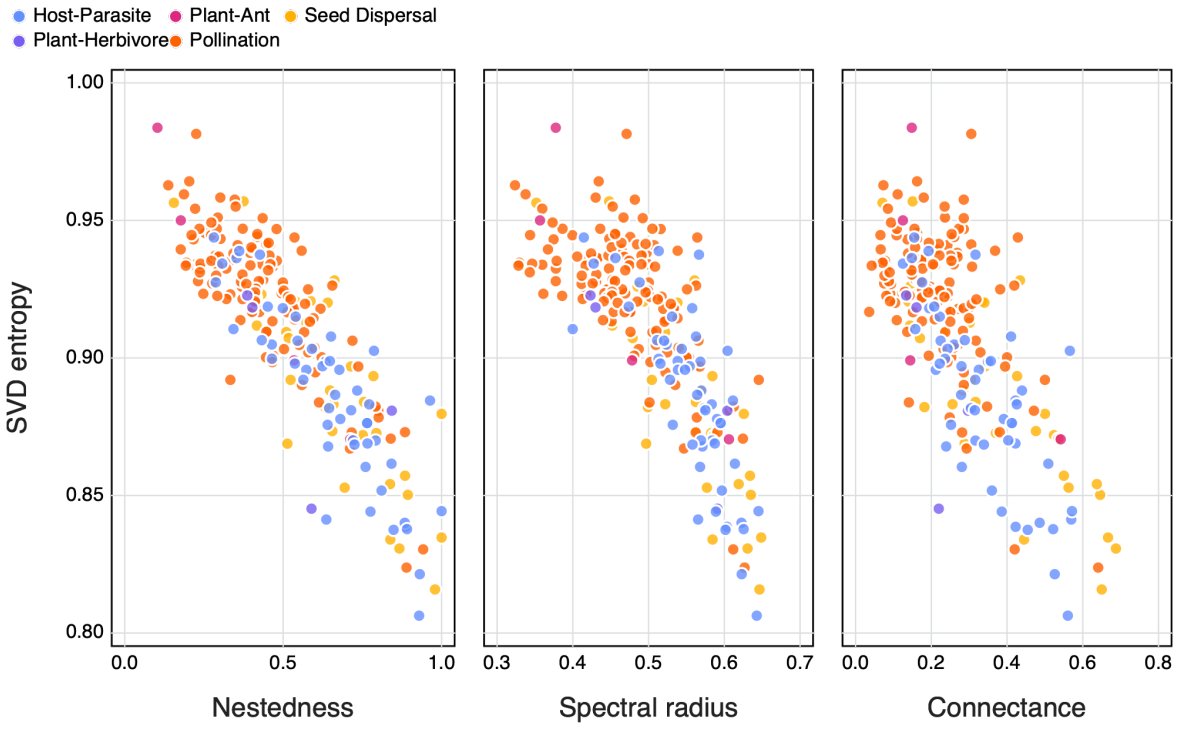


Fig. 3. The relationship between SVD entropy and the nestedness (left panel), spectral radius (central panel) and connectance (right panel) of ecological networks. Colours indicate the different interaction types of the networks.

4.3.3. Complex networks are not more robust to extinction

One approach to calculating the overall structural robustness of an ecological network is by simulating extinction events through the sequential removal of species, which allows constructing an extinction curve that plots the relationship between species removed and cumulative secondary extinctions (Dunne et al., 2002; Memmott et al., 2004). Extinction events can be simulated in a manner of different ways, either by removing 1) a random individual, 2) systematically removing the most connected species (one with the highest number of interactions with other species) and 3) the least connected species (Dunne et al., 2002). After each extinction event, we remove species from the network that no longer have any interacting partners, thereby simulating secondary extinctions. This is then repeated until there are no species remaining in the network. Furthermore, we can restrict extinction events to only one dimension of the interaction matrix, *i.e.,* removing only top-level or bottom-level species, or alternatively removing a species from any dimension of the matrix. Extinction curves are then constructed by plotting the proportion of species remaining against those that have been removed; it stands to reason that a flatter curve 'maintains' its species pool for a longer number of cumulative extinctions, and could be seen as more resilient, when compared to a curve that has a much steeper decline. As per previous studies, we measure the robustness to extinction as the area under the extinction curve (AUC), calculated using the Trapezoidal rule. AUC values close to 0 means that a single extinction is enough to collapse the network almost entirely, and values close to 1 means that most species persist even when the number of extinctions is really high.

When looking at the relationship between SVD entropy and the area under an extinction curve (as a proxy for resilience to extinction) we find differences depending on both the extinction mechanism as well as along which dimension the species removal occurred (Figure 4). As a whole we do not observe any obvious relationships between SVD entropy and resilience, nor for different interaction types. We do however see differences in the resilience of networks depending on how the extinctions were simulated. Generally we see a higher

resilience in networks where species of only a specific group are removed or in networks where species were either randomly removed or based on an increasing number of interactions.

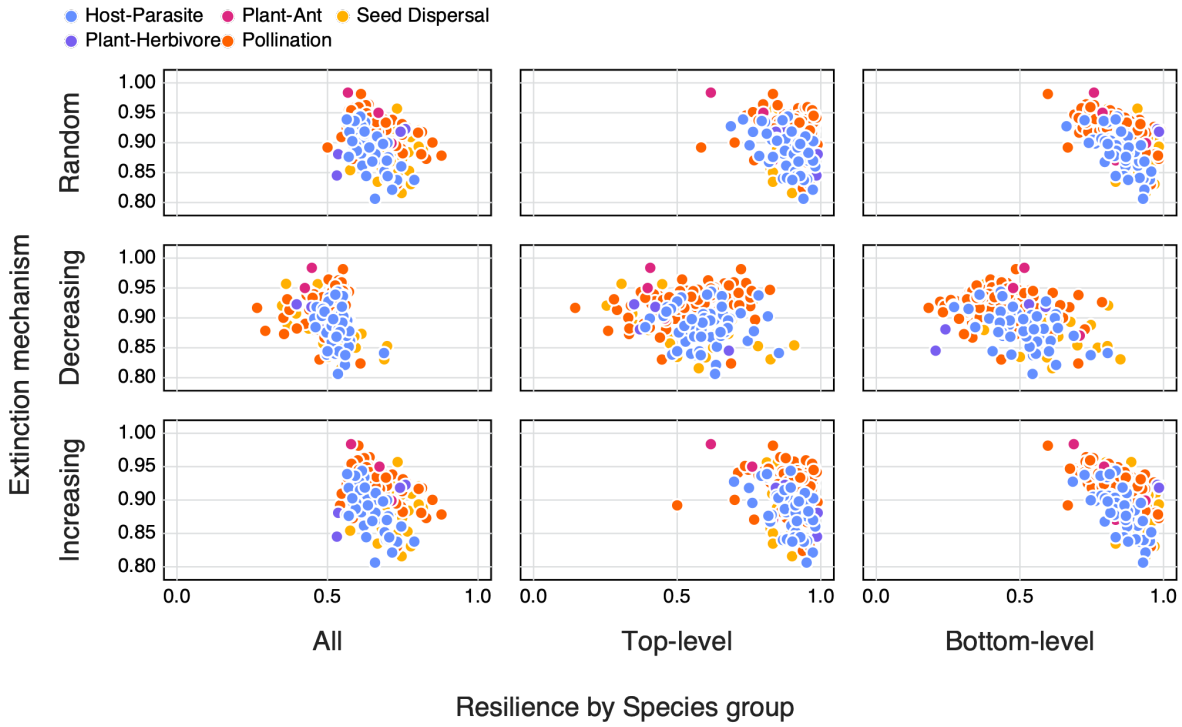


Fig. 4. The relationship between SVD entropy and the area under an extinction curve (as a proxy for resilience to extinction) for both different extinction mechanisms (Random = the removal of a random species, Decreasing = the removal of species in order of decreasing number of interactions (i.e most to least number of interactions), Increasing = the removal of species in order of increasing number of interactions) as well as along different dimensions (species groups) of the network (All = any species, Top-level = only top-level species, and Bottom-level = only bottom- level species) Colours indicate the different interaction types of the networks.

As highlighted in Figure 3 SVD entropy can be used as an additional measure of network complexity. However, as shown in Figure 4, the assumption that network complexity begets resilience to extinction begins to unravel when we use a measure of physical complexity. This is in contrast to previous studies that have shown how connectance plays a role in the resilience of networks to extinctions (Dunne et al., 2002; Memmott et al., 2004). This does not discount the role of using *structural* measures of network complexity (*e.g.,* connectance, nestedness or spectral radius) as indicators of their resilience (although possibly hinting as

to why there is no strong emerging consensus as to how structural complexity relates to this), but rather points to an erroneous assumption as to what aspects of a network we have previously used to define its complexity.

4.3.4. Plant-pollinator networks are slightly more complex

Although we don't observe clear differences in the relationship between different interaction types when comparing amongst various measures of complexity, we do find that different types of interaction networks have differing SVD entropy's. When comparing calculated SVD entropy between interaction types using an ANOVA (after excluding Plant-Ant and Plant-Herbivore interactions due to their small sample size in our dataset) we find a significant difference between group means ($F = 47.047, p < 10^{-3}$). A Tukey's HSD test reveals that plant-pollinator networks ($\mu = .924$) are more complex than both host- parasite networks ($\mu = .885, p < 10^{-3}$) and seed dispersal ($\mu = .888, p < 10^{-3}$). Host-parasite and seed dispersal networks had apparently no difference in average complexity ($p = .889$). These results suggest that mutualistic networks may be more complex, which matches with previous literature: these networks have been shown to minimise competition (Bastolla et al., 2009) and favour unique interactions, thereby increasing network complexity. This specific structure can appear as a side-process of either ecological (Maynard et al., 2018) or evolutionary (Valverde et al., 2018) processes, but nevertheless leaves a profound imprint on the complexity of the networks.

4.3.5. Connectance constrains complexity (but also rank deficiency)

We used simulated annealing (Kirkpatrick, 1984) to generate networks with the highest, or lowest, possible SVD entropy values. From a set network size (30 species, 15 on each side) with a random number of interactions (spanning the entire range from minimally to maximally connected), we reorganised interactions until the SVD entropy was as close to 0 or 1 as possible. We repeated the process 25 times for every number of interactions. We also

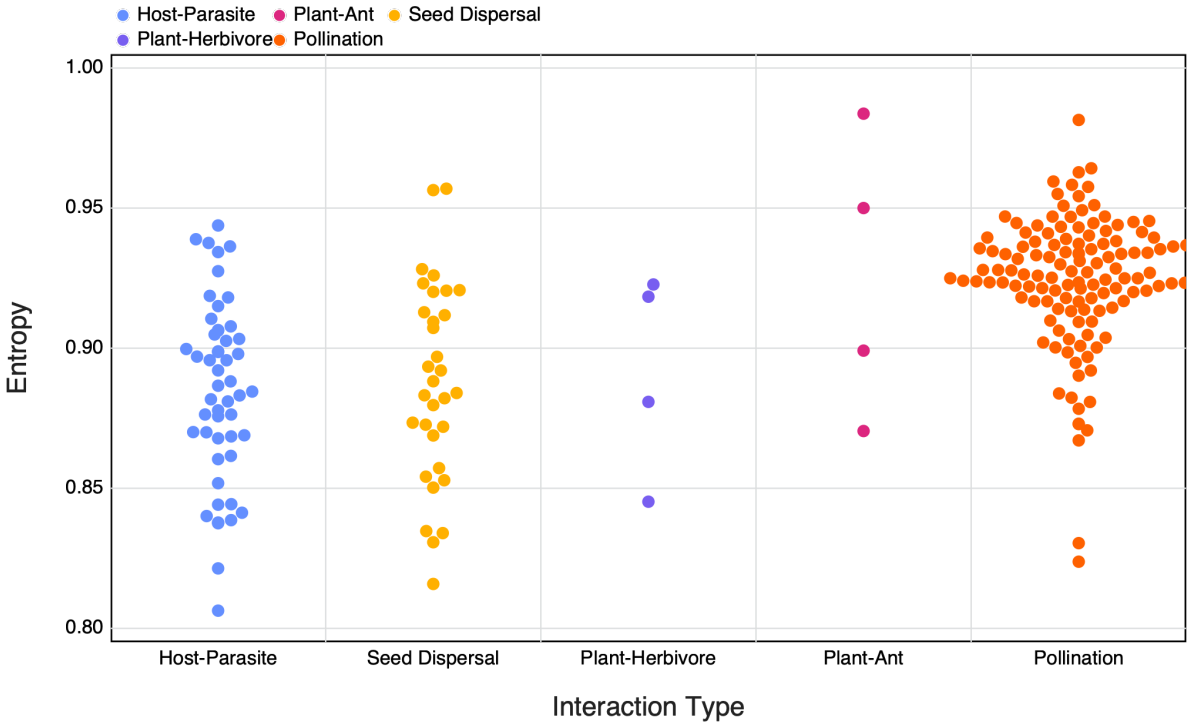


Fig. 5. The calculated SVD entropy of different interaction networks of different interaction types

measured the relative rank deficiency of the generated networks. This allows identifying the boundaries of both measures of complexity. The specific simulated annealing we used is as follows. We set an initial temperature $T_0 = 2$. At every timestep t (up until $t = 10^4$), the temperature is set to $T_t = T_0 \times \lambda^t$, so that is decays exponentially at a rate $\lambda = 1 - 10^{-4}$. At each timestep, we switch two interactions in the network $\mathcal{N}$ at random to generate a proposal network $\mathcal{M}$. The score of this proposal is the difference between the squared error of $\mathcal{N}$ and $\mathcal{M}$ *i.e.*, $\Delta = (f(\mathcal{M}) - \theta)^2 - (f(\mathcal{N}) - \theta)^2$, where f is the SVD entropy and θ is the target for optimisation (either 0 or 1 for respectively minimally or maximally complex). A proposal is accepted with probability $\mathrm{P}(\mathcal{N} \to \mathcal{M}|\Delta) = \exp\left(-\Delta \times T_t^{-1}\right)$.

By exploring the minimal and maximal values of SVD entropy for networks of a given size, we can show that the range of complexity that a network can express varies as a function of connectance (Figure 6). As reported by Poisot and Gravel, 2014, there is no variation

when the networks are either minimally or maximally connected, but any connectance in between can give rise to networks of varying complexities. This being said – minimally connected networks always show the largest complexity, and an increase in connectance will always decrease complexity. Interestingly, this relationship is monotonous, and there is no peak of complexity where the maximal number of possible networks combination exists, *i.e.,* around Co ≈ 0.5 (Poisot and Gravel, 2014). This is an intriguing result – ecological networks are indeed extremely complex, but whereas ecologists have usually interpreted connectance as a measure of complexity, it is in fact sparse networks that are the complex ones, and connectance acts to decomplexify network structure.

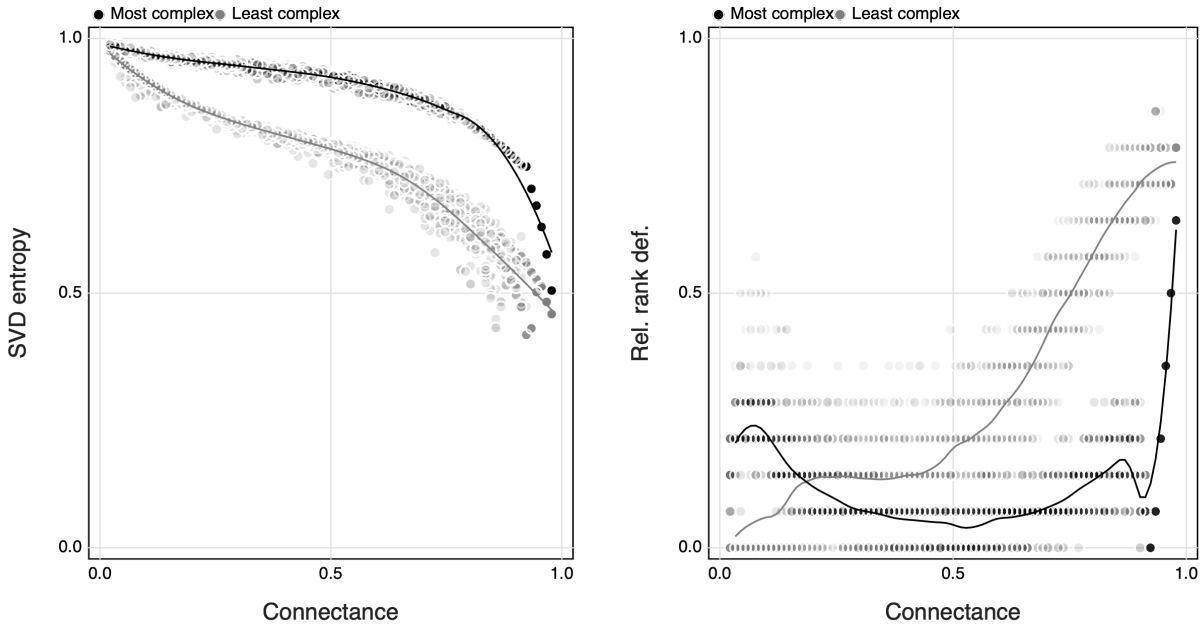


Fig. 6. The relationship between the maximum and minimum value of SVD entropy of a collection of random interaction networks (using simulated annealing) for a given connectance spanning from 0 to 1 (left panel) and how this relates to the relative rank deficiency of networks (right panel)

The right panel of Figure 6 shows the average rank deficiency of networks for which SVD entropy was either maximised or minimised. Complex networks (meaning, maximally complex given their connectance) had a lower deficiency, indicating that except at extreme connectances, there are combinations of networks for which all species can interact in unique

ways – this is a natural consequence of the results reported by Poisot and Gravel, 2014, whereby the number of possible networks is only really constrained at the far ends of the connectance gradient. Minimally complex networks, on the other hand, saw their rank deficiency increase with connectance. This hints at the fact that the decrease in complexity with connectance may be primarily driven by the infeasibility of having enough species for them to all interact uniquely as connectance increases. Because non-unique interactions tend to result in competition Bascompte and Jordano, 2007, this can "push" networks towards the full-rank configuration (as suggested by the results in Figure 1), thereby maximising complexity regardless of connectance.

4.3.6. Larger networks are less complex than they could be

To assess whether ecological networks are more, or less, complex than expected, we applied two null models that generate pseudo-random networks: Type I (Fortuna and Bascompte, 2006), where interactions happen proportionally to connectance, and Type II (Bascompte et al., 2003), where interactions happen proportionally to the joint degree of the two species involved. The models are equivalent to, respectively, the Erdos-Renyi and Configuration models (Newman, 2010), both of which are maximum entropy generative models that reflect global (Type I) or local (Type II) constraints (Park and Newman, 2004). We generated 999 samples for every network in the dataset, and measured the z-score of the empirical network as

$$z_i = \frac{x_i - \mu_i}{\sigma_i}$$

where x_i is the SVD entropy of network i, and μ_i and σ_i are respectively the average and standard deviation of the distribution of SVD entropy under the null model. Negative values of z_i reflect a network that has lower entropy than expected under the assumptions of the null model. In Figure 7, we show that despite high *absolute* values of SVD entropy, ecological networks are not as complex as they *could* be. This is consistently true for both null models, and for the three types of networks that had a sufficient sample size.

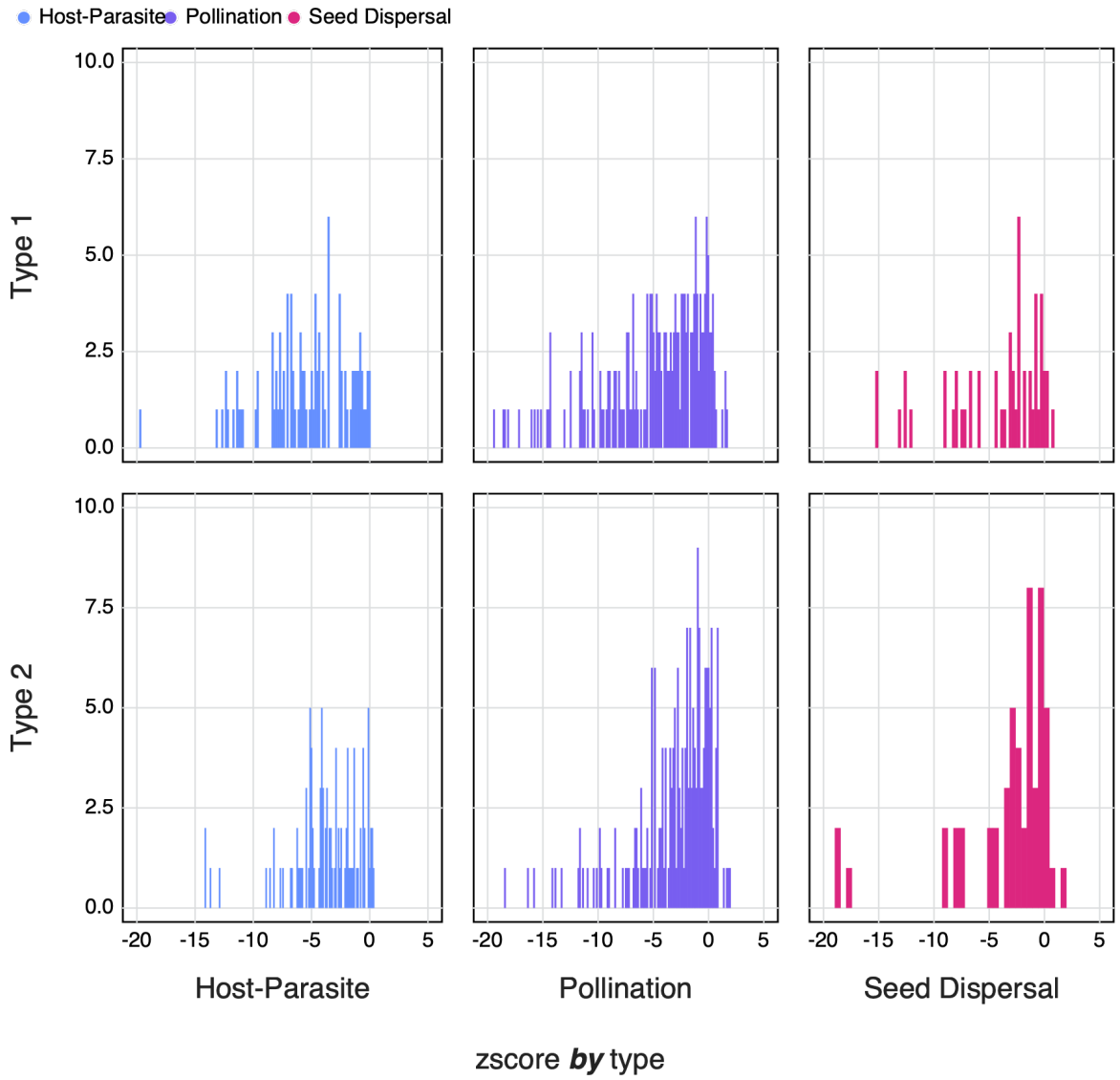


Fig. 7. The counts of the z_i-scores of different types of networks for both Type I and Type II null models. Negative z_i-scores indicate networks with an SVD entropy that is lower *i.e.*, less complex than expected

Previous work on random networks (using a model that is essentially the Type I null model) shows that sufficiently large networks achieve maximal von Neuman entropy (Du et al., 2010; Passerini and Severini, 2011). In Figure 8, we compare the *logistic* of z_i to the richness of the network. Transforming to the logistic smooths out differences in absolute value that are apparent in Figure 7, and projects the values in the unit range, with values

above 0.5 being more complex than expected. It is quite obvious that, across both models and the three types of interactions, only smaller networks achieve higher entropy. Both Barbier et al., 2018 and Saravia et al., 2018 have previously noted that the early stages of network assembly usually result in severely constrained networks, due to the conditions required for multiple species to persist; as networks grow larger, these constraints may "relax", leading in networks with more redundancy, and therefore a lower complexity.

4.4. Conclusion

We present SVD entropy as a starting point to unifying (and standardising) how we should approach defining the complexity of ecological networks. The use of a unified definition will allow us to revisit how complexity relates to the ecological properties of networks using a standardised method. One important result from using SVD entropy is that the complexity of ecological networks is indeed *immense*, yet despite this high complexity networks are still not reaching their *maximum* potential complexity. We suggest that the assembly dynamics of networks may explain this observation but this still raises the question as to why larger (or more mature) networks are not 'maintaining' their expected complexity and prompts further exploration as to the role of ecological assembly in structuring networks.

References

Adami, C. (2002). What is complexity? *BioEssays*, *24*(12), 1085–1094. https://doi.org/10.1002/bies.10192

Barbier, M., Arnoldi, J.-F., Bunin, G., & Loreau, M. (2018). Generic assembly patterns in complex ecological communities. *Proceedings of the National Academy of Sciences*, 201710352. https://doi.org/10.1073/pnas.1710352115

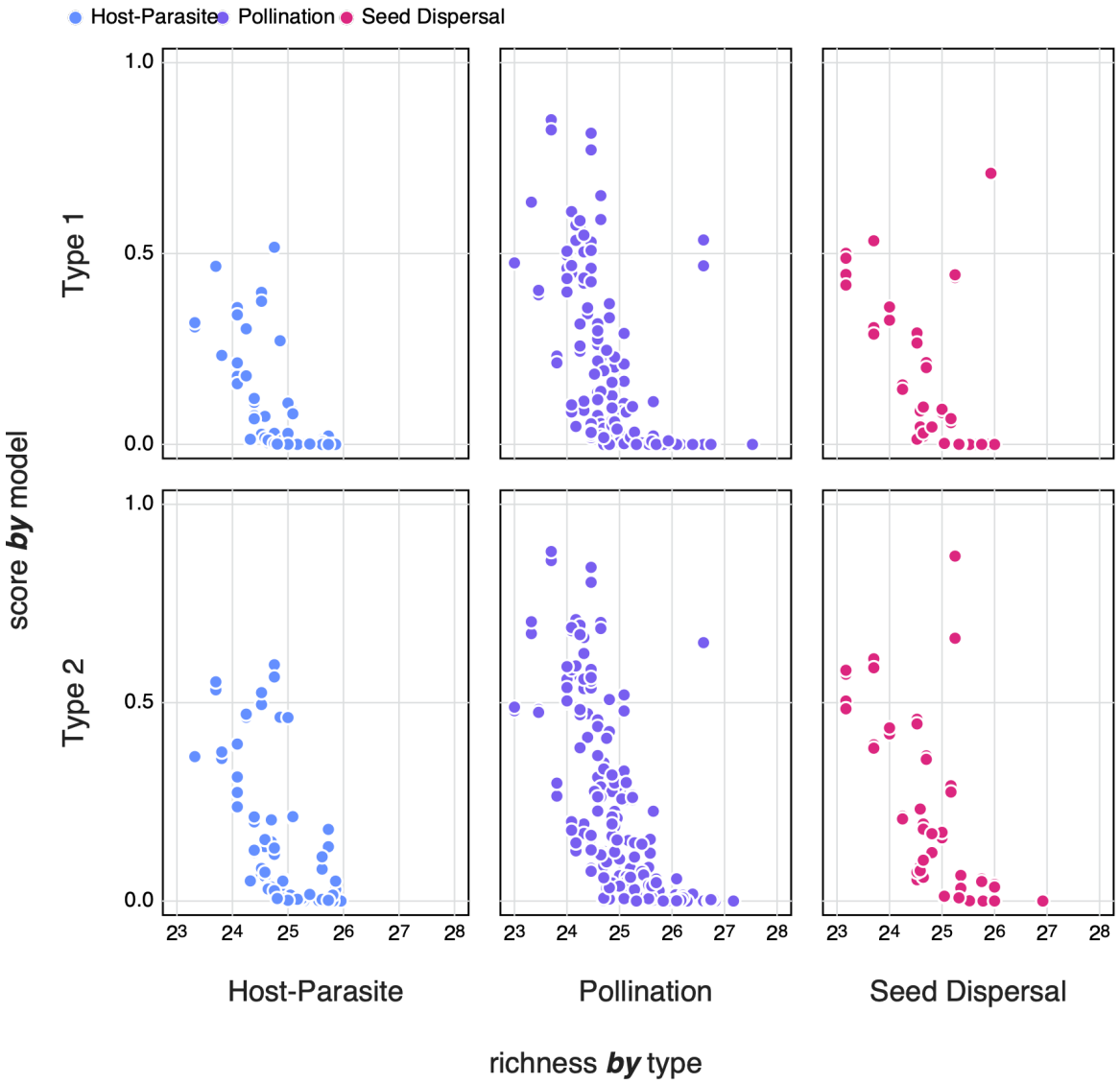


Fig. 8. The logistic z_i-scores of different types of networks for both Type I and Type II null models compared to the species richness of the network. Where z_i-scores below 0.5 indicate networks with an SVD entropy that is lower *i.e.,* less complex than expected

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

Bascompte, J., & Jordano, P. (2007). Plant-Animal Mutualistic Networks: The Architecture of Biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, *38*(1), 567–593. https://doi.org/10.1146/annurev.ecolsys.38.091206.095818

Bastolla, U., Fortuna, M. A., Pascual-García, A., Ferrera, A., Luque, B., & Bascompte, J. (2009). The architecture of mutualistic networks minimizes competition and increases biodiversity. *Nature*, *458*(7241), 1018–1020. https://doi.org/10.1038/nature07950

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

Borrelli, J. J. (2015). Selection against instability: Stable subgraphs are most frequent in empirical food webs. *Oikos*, *124*(12), 1583–1588. https://doi.org/10.1111/oik.02176

Brose, U., Williams, R. J., & Martinez, N. D. (2006). Allometric scaling enhances stability in complex food webs. *Ecology Letters*, *9*(11), 1228–1236. https://doi.org/10.1111/j.1461-0248.2006.00978.x

Delmas, E., Besson, M., Brice, M.-H., Burkle, L. A., Dalla Riva, G. V., Fortin, M.-J., Gravel, D., Guimarães, P. R., Hembry, D. H., Newman, E. A., Olesen, J. M., Pires, M. M., Yeakel, J. D., & Poisot, T. (2018). Analysing ecological networks of species interactions. *Biological Reviews*, 112540. https://doi.org/10.1111/brv.12433

Du, W., Li, X., Li, Y., & Severini, S. (2010). A note on the von Neumann entropy of random graphs. *Linear Algebra and its Applications*, *433*(11), 1722–1725. https://doi.org/10.1016/j.laa.2010.06.040

Duffy, J. E., Cardinale, B. J., France, K. E., McIntyre, P. B., Thébault, E., & Loreau, M. (2007). The functional role of biodiversity in ecosystems: Incorporating trophic complexity. *Ecology Letters*, *10*(6), 522–538. https://doi.org/10.1111/j.1461-0248.2007.01037.x

Dunne, J. A., Williams, R. J., & Martinez, N. D. (2002). Network structure and biodiversity loss in food webs: Robustness increases with connectance. *Ecology Letters*, *5*(4), 558–567. https://doi.org/10.1046/j.1461-0248.2002.00354.x

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, *1*(3), 211–218. https://doi.org/10.1007/BF02288367

Forsythe, G., & Moler, C. (1967). *Computer Solution of Linear Algebraic Systems*. Prentice Hall.

Fortuna, M. A., & Bascompte, J. (2006). Habitat loss and the structure of plant-animal mutualistic networks: Mutualistic networks and habitat loss. *Ecology Letters*, *9*(3), 281–286. https://doi.org/10.1111/j.1461-0248.2005.00868.x

Golub, G. H., Hoffman, A., & Stewart, G. W. (1987). A generalization of the Eckart-Young-Mirsky matrix approximation theorem. *Linear Algebra and its Applications*, *88-89*, 317–327. https://doi.org/10.1016/0024-3795(87)90114-5

Golub, G. H., & Reinsch, C. (1971). Singular value decomposition and least squares solutions. In *Linear Algebra* (pp. 134–151). Springer.

Gravel, D., Massol, F., & Leibold, M. A. (2016). Stability and complexity in model meta-ecosystems. *Nature Communications*, *7*, 12457. https://doi.org/10.1038/ncomms12457

Gu, R., & Shao, Y. (2016). How long the singular value decomposed entropy predicts the stock market? — Evidence from the Dow Jones Industrial Average Index. *Physica A: Statistical Mechanics and its Applications*, *453*(100), 150–161.

Jacquet, C., Moritz, C., Morissette, L., Legagneux, P., Massol, F., Archambault, P., & Gravel, D. (2016). No complexity–stability relationship in empirical ecosystems. *Nature Communications*, *7*, 12573. https://doi.org/10.1038/ncomms12573

Kirkpatrick, S. (1984). Optimization by simulated annealing: Quantitative studies. *Journal of statistical physics*, *34*(5-6), 975–986.

Landi, P., Minoarivelo, H. O., Brännström, Å., Hui, C., & Dieckmann, U. (2018). Complexity and stability of ecological networks: A review of the theory. *Population Ecology*, *60*(4), 319–345. https://doi.org/10.1007/s10144-018-0628-3

Martinez, N. D. (1992). Constant Connectance in Community Food Webs. *The American Naturalist*, *139*(6), 1208–1218.

May, R. M. (1976). Simple mathematical models with very complicated dynamics. *Nature*, *261*(5560), 459. https://doi.org/10.1038/261459a0

Maynard, D. S., Serván, C. A., & Allesina, S. (2018). Network spandrels reflect ecological assembly. *Ecology Letters*, n/a–n/a. https://doi.org/10.1111/ele.12912

Memmott, J., Waser, N. M., & Price, M. V. (2004). Tolerance of pollination networks to species extinctions. *Proceedings of the Royal Society B: Biological Sciences*, *271*(1557), 2605–2611. https://doi.org/10.1098/rspb.2004.2909

Newman, M. E. J. (2010). *Networks. An introduction*. Oxford University Press.

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

Passerini, F., & Severini, S. (2011). The von Neumann entropy of networks. *arXiv:0812.2597 [cond-mat, physics:quant-ph, q-bio]*. https://doi.org/10.4018/978-1-60960-171-3.ch005

Phillips, J. D. (2011). The structure of ecological state transitions: Amplification, synchronization, and constraints in responses to environmental change. *Ecological Complexity*, *8*(4), 336–346. https://doi.org/10.1016/j.ecocom.2011.07.004

Pielou, E. C. (1975). *Ecological diversity*. Wiley.

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., & Gravel, D. (2014). When is an ecological network complex? Connectance drives degree distribution and emerging network properties. *PeerJ*, *2*, e251. https://doi.org/10.7717/peerj.251

Poulin, R. (2010). Network analysis shining light on parasite ecology and diversity. *Trends in Parasitology*, *26*(10), 492–498. https://doi.org/10.1016/j.pt.2010.05.008

Proulx, S. R., Promislow, D. E. L., & Phillips, P. C. (2005). Network thinking in ecology and evolution. *Trends in Ecology & Evolution*, *20*(6), 345–353. https://doi.org/10.1016/j.tree.2005.04.004

Rozdilsky, I. D., & Stone, L. (2001). Complexity can enhance stability in competitive systems. *Ecology Letters*, *4*(5), 397–400. https://doi.org/10.1046/j.1461-0248.2001.00249.x

Saravia, L. A., Marina, T. I., Troch, M. D., & Momo, F. R. (2018). Ecological Network assembly: How the regional meta web influence local food webs. *bioRxiv*, 340430. https://doi.org/10.1101/340430

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Staniczenko, P. P. A., Kopp, J. C., & Allesina, S. (2013). The ghost of nestedness in ecological networks. *Nature Communications*, *4*(1), 1391. https://doi.org/10.1038/ncomms2422

Ulrich, W., Almeida-Neto, M., & Gotelli, N. J. (2009). A consumer's guide to nestedness analysis. *Oikos*, *118*(1), 3–17.

Valverde, S., Piñero, J., Corominas-Murtra, B., Montoya, J., Joppa, L., & Solé, R. (2018). The architecture of mutualistic networks as an evolutionary spandrel. *Nature Ecology & Evolution*, *2*(1), 94. https://doi.org/10.1038/s41559-017-0383-4

Chapter 5 Fifth article

SpatialBoundaries.jl: Edge detection using spatial wombling

by

Tanya Strydom[1], and Timothée Poisot[2]

([1]) Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

([2]) Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada
Québec Centre for Biodiversity Sciences, Montreal, QC, Canada

This article was submitted in Ecography and can be found at https://doi.org/10.1111/ecog.06609.

The main contributions of Tanya Strydom for this articles are presented. TP and TS designed the study and developed the software. TS was the lead on writing and editing.

Résumé. Le wombling spatial est une approche permettant de détecter les contours d'un paysage bidimensionnel défini. Ceci est réalisé en calculant le taux et la direction du changement grâce à l'interpolation de points. Cela donne non seulement une approximation de la forme du paysage, mais peut également être utilisé pour identifier des cellules de limites potentielles qui délimitent un passage d'un état à un autre au sein du paysage. Nous présentons ici le package SpatialBoundaries.jl pour Julia, qui a été développé pour implémenter l'algorithme de wombling pour les ensembles de données référencés spatialement pour des paysages échantillonnés de manière uniforme ou aléatoire. D'un point de vue pratique, la fonctionnalité de wombling permet à l'utilisateur de répondre à deux questions: dans quelle mesure et dans quelle direction la variable d'intérêt change-t-elle ? et la fonctionnalité de limites identifie des cellules limites candidates. Nous concluons en fournissant un exemple fonctionnel du package utilisant les différentes couches de plantes ligneuses pour la Grande-Bretagne et l'Irlande à partir de la base de données EarthEnv.
Mots clés : wombling spatial, détection des contours, limites, écologie spatiale, logiciel, Julia

Abstract. Spatial wombling is an approach for detecting edges within a defined two-dimensional landscape. This is achieved by calculating the rate and direction of change through the interpolation of points. This not only gives an approximation as to the shape of the landscape but can also be used to identify candidate boundaries cells that delimit a shift from one state to another within the landscape. Here we introduce the SpatialBoundaries.jl package for Julia, which has been developed to implement the wombling algorithm for datasets that are spatially referenced for both uniformly or randomly sampled landscapes.From a practical perspective, the wombling functionality allow the user to answer two questions: how much and in which direction does the variable of interest change? and the boundaries functionality identifies candidate boundary cells. We conclude by providing a working example of the package using the various woody plant layers for Britain and Ireland from the EarthEnv database.
Keywords: spatial wombling, edge detection, boundaries, spatial ecology, software, Julia

5.1. Background

There is value in being able to identify boundaries within a landscape as it provides us with a starting point from which to understand changes in species assemblages, ecological

communities, or even simply to delineate areas (based on a shared property) into discrete units, for example ecosystemic regions (Fortin et al., 2000; Post et al., 2007). Here we present a `Julia` (Bezanson et al., 2017) package aimed at detecting boundaries across a specified geographical area by identifying zones of rapid change using the wombling edge detection algorithm. This approach was originally developed by Womble, 1951 in the context of understanding trait variation within a geographic area and was later modified by Barbujani et al., 1989 for the purpose of understanding changes in gene frequencies, although it also has a more general ecological application with regards to spatial data (Fortin and Dale, 2005), serving as a complimentary approach to cluster analysis (Fortin and Drapeau, 1995). Wombling has applicability to a wide range scenarios *e.g.,* trait measurements or genotypes (Barbujani et al., 1989), species interaction networks (Fortin et al., 2021), and has explicitly been used (to list a few examples) to detect transitions within a landscape (Camarero et al., 2000; Philibert et al., 2008), and analyse the spread of invasive species (Fitzpatrick et al., 2010). Although the origins of wombling may be rooted in anthropology and has been extensively used in ecology the potential applicability also extends to other systems such as high-energy experiments in physics (Matchev et al., 2020), or to understand the genetic-linguistic patterns of European populations (Sokal et al., 1990).

Broadly speaking spatial wombling is an edge-detection algorithm which traverses a geographic area (for the purpose of this discussion let's imagine a spatially referenced dataset pertaining to species richness for each location) and defines this area in terms of the rate (m) and corresponding direction of change (θ) through interpolating between nearest neighbours. Although the wombling algorithm (as implemented here) is designed to work with two-dimensional *i.e.,* planar data (as delimited by x and y — which would be the co-ordinates of where species richness was sampled), it is beneficial to view this plane as a three-dimensional object (or series of curves), as shown in Figure 1, panel A. Here the 'amplitude' of the curvature of the plane is determined by the value of z (species richness) and the rate and direction of change is calculated by using the first-order partial derivative ∂ of the surface (curve) as described by $f(x,y)$. This then gives us an indication of how steep the gradient/curve (m) is

between neighbouring cells as well as the direction (from the 'low' to the 'high' point; θ) of the slope (panel B, Figure 1). Large values of m are associated with zones of rapid change in the landscape and are indicative of a shift from one 'state' to another *i.e.,* a potential ecological boundary within the landscape (Fortin and Dale, 2005; dashed line in panel C, Figure 1). One benefit of the wombling approach is that interpolation is not necessarily restricted to a rectangular (2×2) window (that would entail a landscape where points are regularly arranged in space) and can easily be re-written so as to accommodate points that are not regularly arranged across space (as per Fortin, 1994), thereby giving the user more flexibility with regards to how the sampling points are arranged (*i.e.,* sampled) across the landscape.

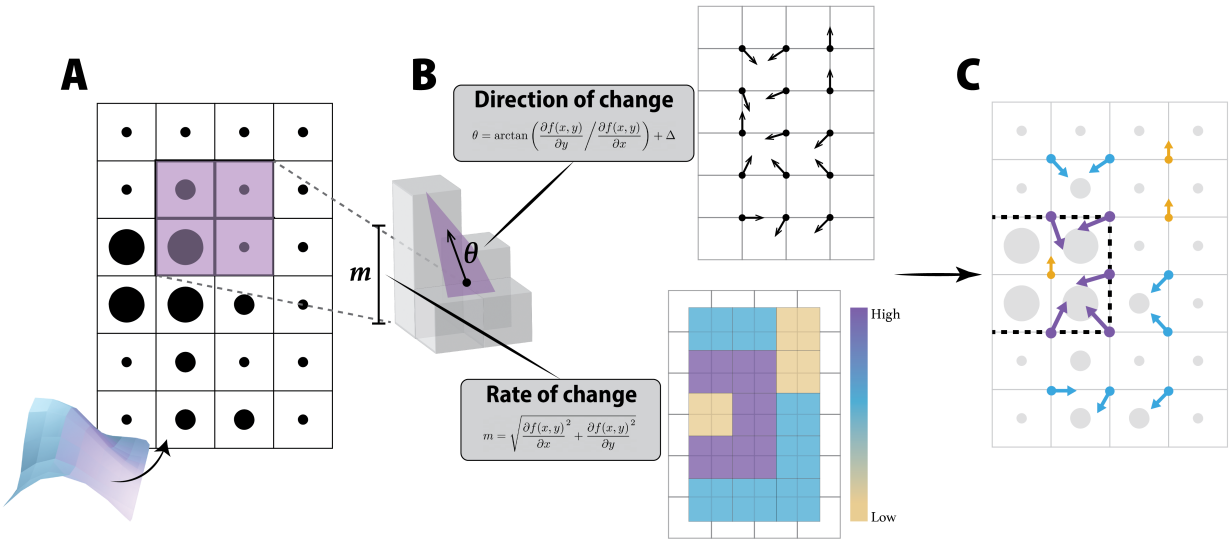


Fig. 1. A visual conceptualisation of how the wombling algorithm interpolates points across a geographical area (in this case the points are regularly arrange in space) for a variable of interest (z) to calculate the rate (m) as well as the direction (θ) of change. Here the sampled landscape is shown in panel A with the size of the points correlating to the magnitude if the variable of interest (z). Panel B shows the two components of the landscape once wombled, which are then combined and superimposed across the original landscape in panel C, with the dashed line indicating a candidate boundary. Here the colours as well as the size of the arrows indicate the rate of change and the direction should be interpreted as moving from the 'low' to the 'high' point. Note that the dimensions of the wombled landscapes (B) will be smaller than the original landscape (A) due to the interpolation process *i.e.,* where we originally had an $n \times r$ grid we now have an $(n - 1)(r - 1)$ sized grid.

5.1.1. Rate of change

The rate of change (m) can be used to find the zones of rapid change within the geographical area — which, in turn, can be used to identify potential candidate boundaries. The rate of change is calculated as follows:

$$m = \sqrt{\frac{\partial f(x,y)}{\partial x}^2 + \frac{\partial f(x,y)}{\partial y}^2} \tag{5.1.1}$$

Where $f(x,y)$ can be expanded as:

$$f(x,y) = z_1(1-x)(1-y) + z_2x(1-y) + z_3xy + z_4(1-x)y$$

For convenience the values of the centroid of the 'search window' *i.e.,* x and y can be standardised to 0.5 when working with points regularly arranged in space. Additionally, as we are interpolating between points, it should also be noted that the original $n * r$ geographical area will now be an $(n-1)(r-1)$ sized grid (*i.e.,* one less row and one less column of values for the wombled landscape as illustrated in panel C of Figure 1).

When we are working with points that are irregularly arranged within the geographical area it is possible to use triangulation wombling (Fortin, 1994; Fortin and Dale, 2005; Fortin et al., 2021). Here the approach to wombling has been modified by Fortin, 1992 so as to interpolate the plane between the three nearest neighbours (as opposed to the usual 2×2 grid). Nearest neighbours are found by using the Delaunay triangulation algorithm (Delaunay, 1934) after which the rate of change is still calculated in the same manner as in Equation 5.1.1, however as we are now only working with a three-point 'window' $f(x,y)$ will be defined as:

$$f(x,y) = ax + by + c$$

where

$$\begin{bmatrix} a & b & c \end{bmatrix} = \begin{bmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{bmatrix}^{-1} \cdot \begin{bmatrix} z_1 & z_2 & z_3 \end{bmatrix}$$

and the x and y co-ordinates of the centroid of the triangle formed by the three points are calculated as follows:

$$\left(\frac{x_1 + x_2 + x_3}{3}\right), \left(\frac{y_1 + y_2 + y_3}{3}\right)$$

5.1.2. Direction of change

It is also possible to calculate a corresponding direction (θ) for each rate of change (noting that the same equation can be used for both lattice and triangulation wombling). This is calculated as:

$$\theta = \arctan\left(\frac{\partial f(x,y)}{\partial y} \Big/ \frac{\partial f(x,y)}{\partial x}\right) + \Delta$$

$$\text{where} \quad \Delta = \left\{ \begin{array}{ccc} 0 & \text{if} & \frac{\partial f(x,y)}{\partial x} \geq 0 \\ 180 & \text{if} & \frac{\partial f(x,y)}{\partial x} < 0 \end{array} \right\}$$

This gives the direction of change which, as the name implies, indicates the direction the rate of change is 'travelling'. The direction of change should be interpreted as wind direction *i.e.,* where the change is coming from and not where it is moving towards. As is the nature of maths when the rate of change is zero it is still possible to calculate a *real* direction for the non-change — which will be 180°. This means it is possible to think of and use the direction of change independently of calculating boundaries *per se* and can be used to inform how the landscape is behaving/changing in a more 'continuous' way as opposed to discrete zones/boundaries. For example if changes in species richness are more gradual (rate of change is near constant) but the the direction of change is consistently East to West (*i.e.,* 90°) we can still infer that species richness is more or less uniformly increasing in a

South-North direction (this is somewhat exemplified in the direction of change landscape of panel B in Figure 1 where the dominant direction of change is East-West).

5.1.3. Candidate boundaries

Detecting boundaries *i.e.,* areas where the angle of the landscape transitions sharply is surprisingly simple. After having calculated the rate of change (m) for the geographical area it is possible to use these values to identify and assign potential boundaries (Fortin and Dale, 2005; Fortin and Drapeau, 1995; Oden et al., 1993). As large rate of change values are indicative of a steep gradient it stands to reason that these points are indicative of a shift from one state to another *i.e.,* indicative of a boundary. Following the approach outlined in Fortin and Dale, 2005 a threshold value (or percentile class) can be set and will determine what proportion of cells will be retained as potential boundaries. For example if using a 0.1 threshold then the highest 10% of points (which are ranked based on m) will be classified as candidate boundaries. Note that points with the same rate of change will be assigned the same rank meaning that more than 10% of the $(n-1)(r-1)$ could potentially be identified as candidate boundary cells. This approach to identifying potential boundary cells is not the sole approach and there are other ways and nuances from which to approach boundary estimation, such as the use of Voronoi tessellations (Fortin and Drapeau, 1995; Matchev et al., 2020; Oden et al., 1993).

5.2. Methods and features

`SpatialBoundaries` v0.0.3 implements the Wombling algorithm within the `Julia` ecosystem and is made available under the permissive MIT license. The source code (along with more extensive documentation) can be found at https://poisotlab.github.io/SpatialBoundaries.jl/. This is an open project and is thus open to contributions. The package itself has two main functions 1) calculate the rate (m) and direction (θ) of change for landscapes for points that are both regularly (*i.e.,* lattice wombling) and irregularly arranged cross space (triangulation wombling) arranged

in space using the `wombling` function (for which layers can also be aggregated using `mean`), and 2) identifying candidate boundary cells based on a user defined threshold value using the `boundaries` function. Objects that have been passed through a wombling function are of the `Womble` abstract type (which has the two sub-types of `LatticeWomble` and `TriangulationWomble`). This package is, to the best of our knowledge, the only package to implement the Wombling algorithm within `Julia`.

5.2.1. Wombling

The `wombling` function calculates and outputs both the rate (m) and direction (θ; where the direction is denoted from the 'low' to the 'high' point) of change for a given geographical area. Leveraging the multiple dispatch within `Julia` this one function will execute either the lattice or triangulation wombling algorithm depending on the structure (spatial referencing) of the input dataset, meaning that it does not need to be user specified. When a *matrix* of z values is used (where the row and column id's act as the co-ordinates) the lattice wombling algorithm will be executed and when three *vectors* (consisting of the co-ordinates (x and y), and z values respectively) the triangulation wombling algorithm is executed. The resulting output object will have two components (m and θ) and will be typed based on the wombling method used. That is a uniform matrix will result in an object of the type `LatticeWomble` and irregularly arranged points will be of the type `TriangulationWomble`.

5.2.2. Overall mean wombling value

The `mean` function calculates the Overall mean wombling value. The methodology stems from the Overall Mean Lattice-Wombling Value ($\bar{m}$) used by Fortin, 1994 in which multiple surfaces (think different z variables) can be overlaid for the purpose of finding the mean rates and directions of change for a composite landscape. The mean rate of change can be defined as the average of m values (for a specific centroid) for the given set of surfaces and the same is done for the direction of change using the θ values. Alongside the mean the standard deviation is also calculated for both the rate and direction of change. Although

Fortin, 1994 only present this approach for lattice wombled landscapes we have extended the functionality to include both lattice and triangulation wombled types. This is provided that the landscape (*i.e.,* co-ordinates) of the different surfaces are *exactly* the same. Note here that the original wombling type is retained and the output data will remain either type `LatticeWomble` or `TriangulationWomble`.

5.2.3. Boundaries

The `boundaries` function takes the rate of change (m) of an object of either of the two `Womble` types (*i.e.,* `LatticeWomble` or `TriangulationWomble`), and identifies potential boundary cells based on a user specified threshold (with the default being 0.1). As opposed to selecting only one threshold value we recommend inputting a range of threshold values into the `boundaries` function to assess how it changes the number of points retained (*i.e.,* boundary cells identified). For example we might see sharp transitions in the number of points that are retained as the threshold value is increased. This inflection point is probably the ideal threshold value to use for boundary selection as a rapid increase in the number of points retained is indicative of a large number of cells with the same rate of change.

5.3. Woody areas of the Hawaiian Islands: a wombling example

Below is an example using the various functions within `SpatialBoundaries` to estimate boundaries for (*i.e.,* patches of) wooded areas on the Southwestern islands of the Hawaiian Islands using landcover data from the EarthEnv project (Tuanmu and Jetz, 2014) as well as integrating some functionality from `SimpleSDMLayers` (Dansereau and Poisot, 2021) for easier work with the spatial nature of the input data. The `SpatialBoundaries` package works really well with `SimpleSDMLayers`, so that you can (i) apply wombling and boundaries finding to a `SimpleSDMLayer` object, and (ii) convert the output of a `Womble` object to a *pair* of `SimpleSDMLayer` corresponding to the rate and direction of change.

Because there are four different layers in the EarthEnv database that represent different types of woody cover we will use the overall mean wombling value. As the data are arranged in a matrix *i.e.,* a lattice this example will focus on lattice wombling, however for triangulation wombling the implementation of functions and workflow would look similar with the exception that the input data would be structured differently (as three vectors of x, y, z) and the output data would be typed as `TriangulationWomble` objects.

```
using SpatialBoundaries
using SpeciesDistributionToolkit
using CairoMakie
import Plots
```

Note that the warning about dependencies is a side-effect of loading some functionalities for `SimpleSDMLayers` as part of `SpatialBoundaries`, and can safely be ignored.

First we can start by defining the extent of the Southwestern islands of Hawaii, which can be used to restrict the extraction of the various landcover layers from the EarthEnv database. We do the actual database querying using `SimpleSDMLayers`.

```
hawaii = (left = -160.2, right = -154.5, bottom = 18.6, top = 22.5)
dataprovider = RasterData(EarthEnv, LandCover)
landcover_classes = SimpleSDMDatasets.layers(dataprovider)
landcover = [SimpleSDMPredictor(dataprovider;
                layer=class, full=true, hawaii...)
                for class in landcover_classes]
```

We can remove all the areas that contain 100% water from the landcover data as our question of interest is restricted to the terrestrial realm. We do this by using the "Open Water" layer to mask over each of the landcover layers individually:

```
ow_index = findfirst(isequal("Open Water"), landcover_classes)
not_water = landcover[ow_index] .!== 0x64
lc = [mask(not_water, layer) for layer in landcover]
```

As layers one through four of the EarthEnv data are concerned with data on woody cover (*i.e.*, "Evergreen/Deciduous Needleleaf Trees", "Evergreen Broadleaf Trees", "Deciduous Broadleaf Trees", and "Mixed/Other Trees") we will work with only these layers. To get a sense of the overall structure of raw landcover components we can sum these four layers and plot the total woody cover for the Southwestern islands (The code for the plot below will give us panel A in Figure 2).

```
classes_with_trees = findall(contains.(landcover_classes, "Trees"))
tree_lc = convert(Float32, reduce(+, lc[classes_with_trees]))
heatmap(tree_lc; colormap=:linear_kbgyw_5_98_c62_n256)
```

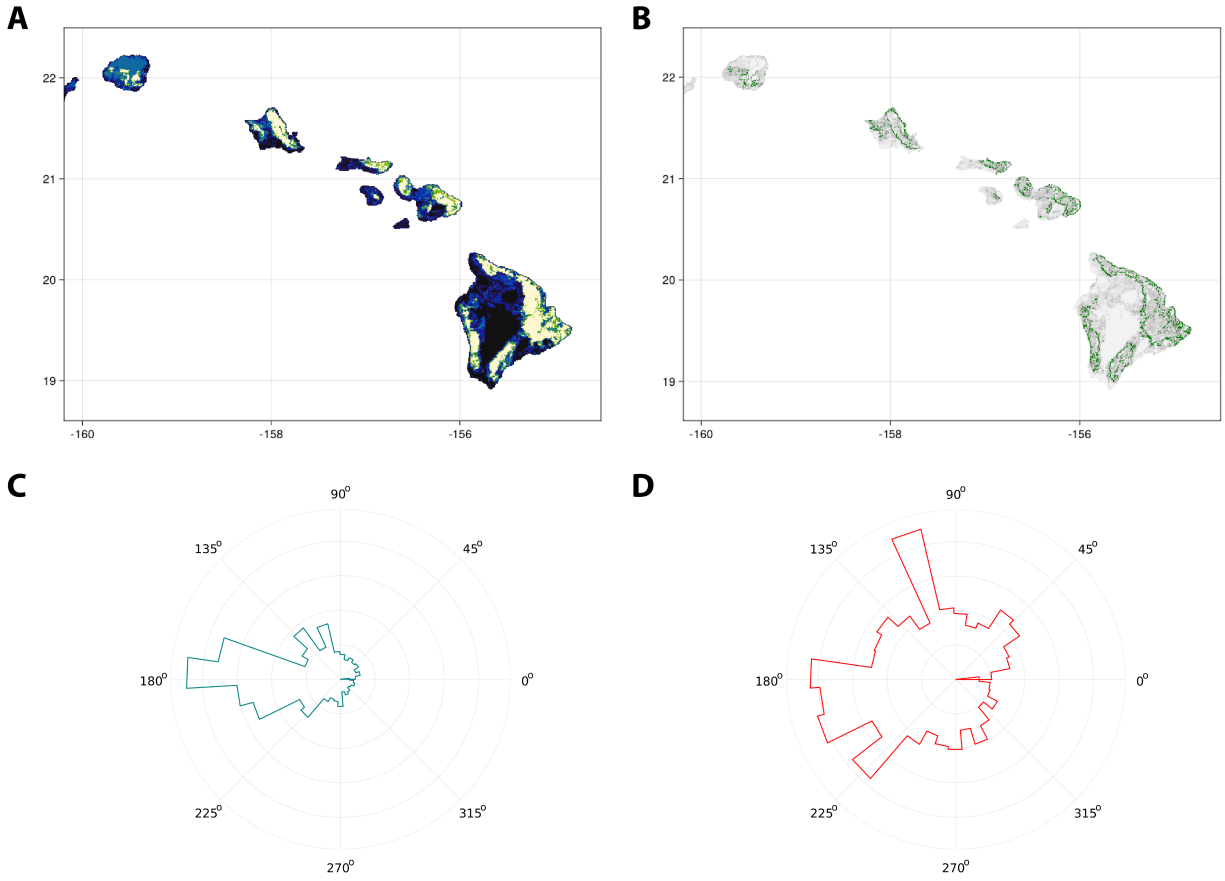


Fig. 2. **A** Woody plant coverage for Southwestern islands of the Hawaiian Islands based on the sum of the cover for layers 1-4 from the EarthEnv project. **B** the overall mean rate of change (*i.e.*, the composite of the wombled layers for layers 1-4) but only for the cells identified as candidate boundary cells when using a 10% threshold, with identified boundaries (shown in green) over the rate of change (shown in levels of grey). The final two panels show the direction of change for all cells (**C**) and only for cells considered to be candidate boundary cells (**D**).

Although we have previously summed the four landcover layers for the actual wombling part we will apply the wombling function to each layer before we calculate the overall mean wombling value. We can broadcast `wombling` in an element-wise fashion to the four different woody cover layers. This will give as a vector containing four `LatticeWomble` objects (since the input data was in the form of a matrix).

```
wombled_layers = wombling.(lc[classes_with_trees])
```

As we are interested in overall woody cover for Southwestern islands we can take the `wombled_layers` vector and use them with the `mean` function to get the overall mean wombling value of the rate and direction of change for woody cover. This will 'flatten' the four wombled layers into a single `LatticeWomble` object.

```
wombled_mean = mean(wombled_layers)
```

From the `wombled_mean` object we can 'extract' the layers for both the mean rate and direction of change. For ease of plotting we will also convert these layers to `SimpleSDMPredictor` type objects. It is also possible to call these matrices directly from the `wombled_mean` object, which has fields for m (the magnitude of change) and θ (the direction of change).

```
rate, direction = SimpleSDMPredictor(wombled_mean)
```

Lastly we can identify candidate boundaries using the `boundaries`. Here we will use a thresholding value (t) of 0.1 and save these candidate boundary cells as `b`. Note that we are now working with a `SimpleSDMResponse` object and this is simply for ease of plotting.

```
b = similar(rate)
b.grid[boundaries(wombled_mean, 0.1; ignorezero = true)] .= 1.0
```

In addition to being used to help find candidate boundary cells we can also use this object (`b`) as masking layer when visualising wombling outputs. In this case we can view the `rate` layer in a similar fashion to the original landcover layer but by masking it with `b` we only plot the candidate boundaries (B in Figure 2) *i.e.,* the cells with the top 10% of highest rate of change values. For visualisation we will overlay the identified boundaries (in green) over the rate of change (in levels of grey)

```
heatmap(rate, colormap=[:grey95, :grey5])
heatmap!(b, colormap=[:transparent, :green])
current_figure()
```

For this example we will plot the direction of change as radial plots (third and fourth panels in Figure 2) to get an idea of the prominent direction of change. Here we will plot *all* the direction values from `direction` for which the rate of change is greater than zero (so as to avoid denoting directions for a slope that does not exist) as well as the `direction` values from only candidate cells using the same masking principle as what we did for the rate of change. It is of course also possible to forgo the radial plots and plot the direction of change in the same manner as the rate of change should one wish.

Before we plot let us create our two 'masked layers'. For all direction values for which there is a corresponding rate of change greater than zero we can use `rate` as a masking layer but first replace all zero values with 'nothing'. For the candidate boundary cells we can simply mask `direction` with `b` as we did for the rate of change.

```
direction_all = mask(replace(rate, 0 => nothing), direction)

direction_candidate = mask(b, direction)
```

Because stephist() requires a vector of radians for plotting we must first collect the cells and convert them from degrees to radians. Then we can start by plotting the direction of change of *all* cells (C in Figure 2).

```
Plots.stephist(
        deg2rad.(values(direction_all));
        proj=:polar,
        lab="",
        c=:teal,
        nbins = 36,
        yshowaxis=false,
        normalize = false,
```

```
        dpi=600)
```

Followed by plotting the direction of change only for cells that are considered as candidate boundary cells (D in Figure 2).

```
Plots.stephist(
                deg2rad.(values(direction_candidate));
                proj=:polar,
                lab="",
                c=:red,
                nbins = 36,
                yshowaxis=false,
                normalize = false,
                dpi=600)
```

5.4. Summary

Edge and boundary detection (as well as their delineation) is an important and valuable concept in spatial ecology (Cadenasso et al., 2003) of which wombling serves as an approach that is flexible in its execution (owing to the non-lattice or triangulation capacity of the function) (Fortin, 1994; Fortin and Dale, 2005) as well as it's capacity to detect more nuanced landscape changes as opposed to being limited to more abrupt discontinuities such as cliffs/ridges by reducing noise in the landscape (Matchev et al., 2020). Wombling sets us up to answer two questions about the geographic area of interest: at what rate and in which direction does the variable of interest change? This of course has value when it comes to evaluating the variation (or uniformity for that matter) of a suite of ecological variables as well as how they may vary with relation to each other.

`SpatialBoundaries.jl` provides the toolset with which to implement both lattice and triangulation wombling using the `wombling` function - multiple dispatch means that the structure of the input dataset will determine exactly which algorithm is implemented. This will simultaneously calculate both the rate and direction of change and if desired multiple sets of

different layers of the same geographic area but defined by different z-variables/surfaces can be aggregated and averaged to calculate the overall mean wombling value. Both `wombling` and `mean` will return objects of the type `Womble` of either the sub-type `LatticeWomble` or `TriangulationWomble` depending on which method was used. An object of any sub-type `Womble` can be input into the `boundaries` function so as to identify cells that can be considered as candidate boundaries based on a user specified threshold.

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. TS and TP are funded by a donation from the Courtois Foundation.

References

Barbujani, G., Oden, N. L., & Sokal, R. R. (1989). Detecting Regions of Abrupt Change in Maps of Biological Variables. *Systematic Zoology*, *38*(4), 376–389. https://doi.org/10.2307/2992403

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

Cadenasso, M. L., Pickett, S. T. A., Weathers, K. C., & Jones, C. G. (2003). A Framework for a Theory of Ecological Boundaries. *BioScience*, *53*(8), 750–758. https://doi.org/10.1641/0006-3568(2003)053[0750:AFFATO]2.0.CO;2

Camarero, J. J., Gutiérrez, E., & Fortin, M.-J. (2000). Boundary Detection in Altitudinal Treeline Ecotones in the Spanish Central Pyrenees. *Arctic, Antarctic, and Alpine Research*, *32*(2), 117–126. https://doi.org/10.1080/15230430.2000.12003347

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

Delaunay, B. (1934). Sur la sphere vide. *Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques*, *6*, 793–800.

Fitzpatrick, M. C., Preisser, E. L., Porter, A., Elkinton, J., Waller, L. A., Carlin, B. P., & Ellison, A. M. (2010). Ecological boundary detection using Bayesian areal wombling. *Ecology*, *91*(12), 3448–3455. https://doi.org/10.1890/10-0807.1

Fortin, M.-J. (1992). *Detection of ecotones: Definition and scaling factors* (Doctoral dissertation). State University of New York at Stony Brook. United States – New York.

Fortin, M.-J. (1994). Edge Detection Algorithms for Two-Dimensional Ecological Data. *Ecology*, *75*(4), 956–965. https://doi.org/10.2307/1939419

Fortin, M.-J., & Dale, M. R. T. (2005). *Spatial analysis: A guide for ecologists*. Cambridge University Press.

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

Fortin, M.-J., & Drapeau, P. (1995). Delineation of Ecological Boundaries: Comparison of Approaches and Significance Tests. *Oikos*, *72*(3), 323–332. https://doi.org/10.2307/3546117

Fortin, M.-J., Olsen, R., Ferson, S., Iverson, L., Hunsaker, C., Edwards, G., Levine, D., Butera, K., & Klemas, V. (2000). Issues related to the detection of boundaries. *Landscape Ecology*, *15*, 453–466.

Matchev, K. T., Roman, A., & Shyamsundar, P. (2020). Finding wombling boundaries in LHC data with Voronoi and Delaunay tessellations. *Journal of High Energy Physics*, *2020*(12), 137. https://doi.org/10.1007/JHEP12(2020)137

Oden, N. L., Sokal, R. R., Fortin, M.-J., & Goebl, H. (1993). Categorical Wombling: Detecting Regions of Significant Change in Spatially Located Categorical Variables. *Geographical Analysis*, *25*(4), 315–336. https://doi.org/10.1111/j.1538-4632.1993.tb00301.x

Philibert, M. D., Fortin, M.-J., & Csillag, F. (2008). Spatial structure effects on the detection of patches boundaries using local operators. *Environmental and Ecological Statistics*, *15*(4), 447–467. https://doi.org/10.1007/s10651-007-0061-9

Post, D. M., Doyle, M. W., Sabo, J. L., & Finlay, J. C. (2007). The problem of boundaries in defining ecosystems: A potential landmine for uniting geomorphology and ecology. *Geomorphology*, *89*(1-2), 111–126. https://doi.org/10.1016/j.geomorph.2006.07.014

Sokal, R. R., Oden, N. L., Legendre, P., Fortin, M.-J., Kim, J., Thomson, B. A., Vaudor, A., Harding, R. M., & Barbujani, G. (1990). Genetics and Language in European Populations. *The American Naturalist*, *135*(2), 157–175.

Tuanmu, M.-N., & Jetz, W. (2014). A global 1-km consensus land-cover product for biodiversity and ecosystem modelling. *Global Ecology and Biogeography*, *23*(9), 1031–1045. https://doi.org/10.1111/geb.12182

Womble, W. H. (1951). Differential Systematics. *Science*, *114*(2961), 315–322. https://doi.org/10.1126/science.114.2961.315

Chapter 6

General Conclusion

As species interaction networks are determined by ecological and evolutionary mechanisms that have played out across spatial and temporal scales the measures that define their structure (properties) are also capturing information about the processes that have played a role in structuring them. Thus the properties of a network are not only representative of its structure but also a measure of *how* different processes have played a role in determining it, meaning that when we are measuring the property of a network we are also capturing information about the network. This information is something that can be used for making predictions about network structure, *e.g.,* it has been shown that it is possible to to make network-level inferences with very low-level data (*e.g.,* Banville et al., 2023; MacDonald et al., 2020), or to help make within network predictions to find missing links (*e.g.,* Poisot et al., 2023; Stock, 2021). The ability to use a 'simple' measure of the species community for a given location (such as species richness) and to have an estimate of the structure of the potential network (such as connectance) is truly amazing and speaks to how much information we are able to extract either from species networks, or alternatively use to make network predictions. This is something is echoed throughout this thesis.

6.1. What we have learnt

6.1.1. Prediction is attainable and feasible

Chapter 3 showcases that with very little 'real world' information we can make accurate predictions by using the information that is 'encoded' in existing interaction networks. This is also echoed in Chapter 4, where we can see the *immense* amount of information that networks contain and that it is a case of finding ways to access and use this information and, in this instance, using it to make network predictions. It is also clear that this information can go a long way, the metaweb for Canada shared only 4% of species with its European counterpart, yet we were able to recover 91% of the interactions. From an ecological perspective this highlights how the laws governing interactions (*sensu* common backbones from Bramon Mora et al., 2018) are being conserved phylogenetically (Davies, 2021; Elmasri et al., 2020; Gómez et al., 2010). From a more practical perspective we show that this transfer learning framework is primed to be adopted as a 'gap-filling' tool as it does not require *extensive* data or computational resources. It is worth noting that in order to predict the Canadian metaweb one only needs to access three different data sources (the species community for Canada, the European metaweb, and a well resolved phylogeny) and that it is possible to execute the required code on a standard laptop, underscoring the lightweight nature of the framework. Finally, the transfer learning framework itself has a lot of scope to be modified should the transfer task have a different set of *e.g.,* data requirements (this is discussed extensively in chapter 2).

As highlighted by the broad, scoping, discussion in Chapter 1 transfer learning is not the only way to approach transfer learning and there are a variety of methodological approaches and data sources that we can tap into when wanting to make network predictions (Figure 2. Having a body of work that explicitly addresses the idea of machine learning for network prediction will be particularly valuable as the popularity and interest in alternative 'non-statistical' methods continues to grow within the field of ecology and evolution (Cuff et al., 2023; Pichler and Hartig, 2023). This chapter also highlights that as our access to the

'auxiliary data' and the computational power we need for network prediction grows we are methodologically and computationally in a prime position to start making feasible network predictions. Hopefully the ideas discussed in these chapters will also allow us to develop even more "unreasonably effective" methods for network prediction, thereby providing us with an even more diverse set of approaches we can use for the different scenarios we will inevitably be faced with in the quest to fill in the global gaps.

6.1.2. Tools for cross-regional comparison

Chapter 4 highlights that we need to be critical of the 'tools' we are using when are trying to make cross-region comparisons, and that quantifying the complexity of networks remains, well, complex (Riva et al., 2023). Taken at face value it appears that (at least bipartite networks) are extremely complex. All of the 226 networks that we looked at are near maximal 'physical complexity', with all networks being near maximal rank and having an SVD entropy greater than 0.8. However, it is the comparison of SVD entropy with the other (structural) measures of complexity (nestedness, connectance, and spectral radius), that highlights the need for us to be critical of the tools we are using. It is clear that structural measures of 'complexity' are in fact capturing a different facet of network complexity than when we are using SVD entropy (Figure 3), this is due to fundamental differences in what aspect of network 'complexity' these measures are trying to capture. One could argue that SVD entropy provides a more fundamentally "correct" measure of complexity as it is quantifying the information within a network as opposed to the number of components/parts a network has and thus should have a place in the toolkit of network descriptors. In addition we show in subsection 4.3.6 that despite their high complexity networks are still not reaching their highest potential complexity. Although we suggest that the assembly dynamics of networks may play a role, it still raises the question as to why larger networks are not maintaining their complexity and opens the door to questions about how assembly (time) shapes ecological networks (Barbier et al., 2018; Saravia et al., 2018).

One of the biggest challenges we will be faced with as network ecology moves from trying to fill in the global gaps to grappling with global scale questions is that we need to be able to delimit them. The software developed in Chapter 5 is primed for this specific challenge once we begin to leverage global-scale data to understand the spatial structure of networks, especially with regards to being able to discretise them. In this sense this chapter is perhaps the most open-ended component of this thesis, but as such it has a great deal of potential applications, particularly addressing a very simple question - where do networks stop? This is particularly meaningful even in the context of network prediction - at what scale should we be making our network predictions? Although there may be methodological constraints that determine how large the (for example) taxonomic scope of the task of network prediction should be (see subsection 2.5.1) there is also the question as to what constitutes the correct biogeographic (and socio-political) extent for prediction. Thus being able to 'draw boundaries' around networks has both theoretical as well as applied significance.

6.1.3. Putting it all together

Arguably the *potential* applicability of the potential applications of chapter 5 represents a culmination of the bulk of the work presented in this thesis. Although Chapters one through three showcase the 'attainability' of network prediction one core aspect that we are still missing is knowing exactly how to delimit the scope (specifically spatial area) of our predictions. In chapter 3 we use 'Canada' as the area for prediction, however Canada is a geopolitical unit and there is no strong ecological reason to have omitted the other parts of the North American landmass from the scope of prediction, as there is no strong environmental boundary on the Canada-US border. There is of course also the inverse argument that then questions what would be the optimal 'area' to have made the predictions for chapter 3 at - it is not feasible, nor ecologically pertinent, to want to construct a metaweb for The Americas in their entirety. There is thus a need, from a practical perspective, to be able to discretise the area (or species community) that we wish to make a network prediction for, and in order

to do that we need to build a theoretical understanding of the boundaries between networks. These ideas are discussed in more depth in subsection 6.2.2, however it is the functionality of `SpatialBoundaries.jl` that can facilitate the advancement and development of theory and ideas related to boundaries between networks and help to guide us when choosing the scale at we should be making our predictions at.

6.2. The direction moving forward

6.2.1. Scrutinising our methods

Although chapters 1 and 2 discuss predictive methods (and chapter 3 provides a tangible example thereof) the job isn't done when it comes to evaluating the data we are using for prediction. More recent work is showing that the imbalances in current data might be a large problem (especially the false negative rate, *i.e.,* interactions that do occur but are missed in the field, and thus viewed as being 'absent' from the system). When reading the work from Brimacombe et al., 2023, Catchen, Poisot, et al., 2023, and Poisot, 2023 one can't help but to be a bit hesitant to adopt a purely predictive framework, however, as we show in subsection A.1.3 the transfer learning model does do quite well, even when we bring "false interactions" into the dataset. Of course this does highlight the need to be critical (or at least cautious) when it comes to using datasets for learning, and highlights the need for identifying priority sampling locations and (maybe) even priority interactions, (*e.g.,* some of the work coming out of the GeoBon group focusing on locating priority sampling locations, "BiodiversityObservationNetworks.Jl", 2021/2022) to help create a 'best subset' of datasets that can be used for additional data curation.

Even the work presented in chapter 3 has room for expansion, and we can (and should) try and push the limits further to see where this transfer learning framework 'breaks'. One tempting challenge would be to try and construct a metaweb for Australian mammals — One is inclined to think that if one were to use the framework from chapter 3 'out of the box' the predictions would have a large degree of uncertainty around them due to the taxonomic relationships between Europe and Australian mammals. But this does make for a case study

to experiment with other transfer mediums (such as traits). An additional 'testing ground' that could prove interesting is to look at rewilding as well as species invasions. Within the context of rewilding one can test how well the predictions are able to 'forecast' the potential impacts of a re-introduced species *a priori* (which one could validate using existing rewilding projects), as well as assess the utility of different candidate ecological surrogates that may be earmarked for introduction in a specific area. The latter point may be particularly useful as one of the main goals is to target the trophic complexity of the area to be rewilded (Perino et al., 2019). For invasions, this can be used to prioritise species that are expected to have a disproportionate impact on the community and flag them in advance as potential threats (David et al., 2017).

6.2.2. Defining ecotrophic zones

Networks are dynamic, and they can vary across space (Golubski et al., 2016; Vázquez et al., 2007) or time (Poisot et al., 2015; Trøjelsgaard and Olesen, 2016) as a function of the environmental conditions. Naturally, we expect network properties to also be dynamic and vary over - in this instance - space. Spatial wombling can be used as a starting point for understanding *how* networks vary across a landscape, particularly if we were to combine this with information on environmental change. Appendix C shows some initial (and by no means well resolved) ideas of how we can use the `SpatialBoundaries.jl` along with the metacommunity model developed by Thompson and Gonzalez, 2017 to look at how the environmental, species community, and network communities boundaries compare within a landscape.

With regards to environmental change it might be interesting to compare species turnover and network changes across the landscape, specifically if we see similar patterns of rates of change at the species or community level and with regards to network structure. This is interesting because there area a myriad of ways we might expect networks to change (or not change) along environmental gradients. Firstly, we might expect network structure to be constant along gradients due to energetic or evolutionary constraints that force networks

to take on a specific shape, *sensu lato* conserved backbones (Bramon Mora et al., 2018). Alternatively network structure does indeed change along a gradient. This could be due to intraspecific variation that causes the re-wiring of interactions (Bolnick et al., 2011) or changes in species composition are also driving changes in the resulting network (Martins et al., 2022).

6.2.2.1. How do the structures within networks vary. There is also the scope to develop a more nuanced understanding of how the landscape structures networks, specifically how the different nodes (*i.e.* species) of the network will perceive the landscape differently. When looking at other fields of ecology *e.g.,* productivity-diversity studies it is clear that the nature of the relationship between productivity and diversity is scale-dependent (Chase and Leibold, 2002; Gillman and Wright, 2006). It stands to reason that this will also be the case when looking at interaction networks, specifically how 'boundaries' may be dependent on the node (species). Which means that we might expect *within* network changes *e.g.,* motifs (specific patterns of linkages in a network) to vary across the landscape even if the larger, regional network structure may remain stable. That being said, there is a compelling argument for the need to 'combine' these smaller functional units with larger spatial networks (Fortin et al., 2021) and that we should also start thinking about the interplay of time and space (Estay et al., 2023). Although deciding exactly what measure might actually be driving differences between local networks and the regional metaweb might not be that simple (Saravia et al., 2022).

6.2.2.2. Boundaries for policy or management. Although this section argues for a more theoretical approach to understanding boundaries in the context of potential assembly patterns/constraints there is also a strong argument for being able to draw lines around communities in the context of having a network (or, more realistically, a metaweb) as an 'object' that can be used in a policy or management context. In section 2.5 and section 2.6 we briefly mention that the scale of prediction should be ecologically relevant, but should also take into consideration the social aspect of why (and how) we are making predictions. To me, there is an argument that this is also the case when thinking about network boundaries.

Given that policy and legislation are enacted at various levels of government or other ruling bodies, being able to identify the boundaries between networks may in fact be a powerful tool at the governance level. Being able to delimit interacting communities (*i.e.,* identify a metaweb) is surely more meaningful than looking solely at species inventories or community composition, particularly if one is truly concerned with conserving ecosystem functions or processes (Thuiller et al., 2023; Wood et al., 2022). However, I feel it is important to stress that the idea of trying to draw boundaries should be approached with caution and sensitivity, especially within in the context of 'doing no harm' (*sensu* Box 2 of section 2.6) and understanding that ecological and socio-political 'boundaries' may in fact have different 'goals' or contexts.

6.2.3. The future collaborative toolbox

On a more contemplative note, I want to discuss the value of thinking about the development of further tools for the toolbox analogy used in this thesis. A significant amount of work in this thesis was only possible with the support and intellectual contribution of many collaborators and there is an argument for continuing this strong network of collaboration for the development of future such tools. From a purely practical perspective the continued push for developing biology-centric packages within the `Julia` language (Roesch et al., 2023) requires that we maintain interoperability between packages and build a collection of tools that build on and fit in amongst each other. Looking at the science/theoretical side of the toolbox, a unified idea or goal for moving the macroecological network 'agenda' forward means that we can build on ideas and thoughts in a more cohesive manner. For example (Banville et al., 2023; Catchen, Lin, et al., 2023; Dansereau et al., 2023) have all already used the work presented in this thesis to take the ideas discussed in new and further directions. This is not to say that we should not also work on developing 'competing' methods (although I would argue 'competing' here is used in the context for finding alternative approaches to solving a similar problem *e.g.,* Caron et al., 2022 takes a more trait-based approach to network prediction), but there is strong evidence that in working together we can get where we

want to be sooner. The 'toolbox' that this thesis represents is but a small step in thinking about interaction networks at a global scale, but it is nevertheless an important step, as it will hopefully lay the groundwork for even more innovation and creation.

References

Banville, F., Gravel, D., & Poisot, T. (2023). What constrains food webs? A maximum entropy framework for predicting their structure with minimal biases. *PLOS Computational Biology*, *19*(9), e1011458. https://doi.org/10.1371/journal.pcbi.1011458

Barbier, M., Arnoldi, J.-F., Bunin, G., & Loreau, M. (2018). Generic assembly patterns in complex ecological communities. *Proceedings of the National Academy of Sciences*, 201710352. https://doi.org/10.1073/pnas.1710352115

BiodiversityObservationNetworks.jl. (2022, October 1). https://github.com/EcoJulia/BiodiversityObservationNetworks.jl/

Bolnick, D. I., Amarasekare, P., Araújo, M. S., Bürger, R., Levine, J. M., Novak, M., Rudolf, V. H., Schreiber, S. J., Urban, M. C., & Vasseur, D. A. (2011). Why intraspecific trait variation matters in community ecology. *Trends in Ecology & Evolution*, *26*(4), 183–192. https://doi.org/10.1016/j.tree.2011.01.009

Bramon Mora, 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

Brimacombe, C., Bodner, K., Michalska-Smith, M., Poisot, T., & Fortin, M.-J. (2023). Shortcomings of reusing species interaction networks created by different sets of researchers. *PLOS Biology*, *21*(4), e3002068. https://doi.org/10.1371/journal.pbio.3002068

Caron, D., Maiorano, L., Thuiller, W., & Pollock, L. J. (2022). Addressing the Eltonian shortfall with trait-based interaction models. *Ecology Letters*, *25*(4), 889–899. https://doi.org/10.1111/ele.13966

Catchen, M. D., Lin, M., Poisot, T., Rolnick, D., & Gonzalez, A. (2023). Improving ecological connectivity assessments with transfer learning and function approximation.

Catchen, M. D., Poisot, T., Pollock, L. J., & Gonzalez, A. (2023). The missing link: Discerning true from false negatives when sampling species interaction networks.

Chase, J. M., & Leibold, M. A. (2002). Spatial scale dictates the productivity–biodiversity relationship. *Nature*, *416*(6879), 427–430. https://doi.org/10.1038/416427a

Cuff, J. P., Deivarajan Suresh, M., Dopson, M. E. G., Hawthorne, B. S. J., Howells, T., Kitson, J. J. N., Miller, K. A., Xin, T., & Evans, D. M. (2023). Chapter One - A roadmap for biomonitoring in the 21st century: Merging methods into metrics via ecological networks. In D. A. Bohan & A. J. Dumbrell (Eds.), *Advances in Ecological Research* (pp. 1–34). Academic Press. https://doi.org/10.1016/bs.aecr.2023.09.002

Dansereau, G., Barros, C., & Poisot, T. (2023). Spatially explicit predictions of food web structure from regional level data.

David, P., Thébault, E., Anneville, O., Duyck, P. .-., Chapuis, E., & Loeuille, N. (2017, January 1). Chapter One - Impacts of Invasive Species on Food Webs: A Review of Empirical Data. In D. A. Bohan, A. J. Dumbrell, & F. Massol (Eds.), *Advances in Ecological Research* (pp. 1–60). Academic Press. https://doi.org/10.1016/bs.aecr.2016.10.001

Davies, T. J. (2021). Ecophylogenetics redux. *Ecology Letters*, *n/a*. https://doi.org/10.1111/ele.13682

Elmasri, M., Farrell, M. J., Davies, T. J., & Stephens, D. A. (2020). A hierarchical Bayesian model for predicting ecological interactions using scaled evolutionary relationships. *The Annals of Applied Statistics*, *14*(1), 221–240. https://doi.org/10.1214/19-AOAS1296

Estay, S. A., Fortin, M.-J., & López, D. N. (2023). Editorial: Patterns and processes in ecological networks over space. *Frontiers in Ecology and Evolution*, *11*.

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

Gillman, L. N., & Wright, S. D. (2006). The Influence of Productivity on the Species Richness of Plants: A Critical Assessment. *Ecology*, *87*(5), 1234–1243. https://doi.org/10.1890/0012-9658(2006)87[1234:TIOPOT]2.0.CO;2

Golubski, A. J., Westlund, E. E., Vandermeer, J., & Pascual, M. (2016). Ecological Networks over the Edge: Hypergraph Trait-Mediated Indirect Interaction (TMII) Structure. *Trends in Ecology & Evolution*, *31*(5), 344–354. https://doi.org/10.1016/j.tree.2016.02.006

Gómez, J. M., Verdú, M., & Perfectti, F. (2010). Ecological interactions are evolutionarily conserved across the entire tree of life. *Nature*, *465*(7300), 918–921. https://doi.org/10.1038/nature09113

MacDonald, A. A. M., Banville, F., & Poisot, T. (2020). Revisiting the Links-Species Scaling Relationship in Food Webs. *Patterns*, *0*(0). https://doi.org/10.1016/j.patter.2020.100079

Martins, L. P., Stouffer, D. B., Blendinger, P. G., Böhning-Gaese, K., Buitrón-Jurado, G., Correia, M., Costa, J. M., Dehling, D. M., Donatti, C. I., Emer, C., Galetti, M., Heleno, R., Jordano, P., Menezes, Í., Morante-Filho, J. C., Muñoz, M. C., Neuschulz, E. L., Pizo, M. A., Quitián, M., … Tylianakis, J. M. (2022). Global and regional ecological boundaries explain abrupt spatial discontinuities in avian frugivory interactions. *Nature Communications*, *13*(1), 6943. https://doi.org/10.1038/s41467-022-34355-w

Perino, A., Pereira, H. M., Navarro, L. M., Fernández, N., Bullock, J. M., Ceauşu, S., Cortés-Avizanda, A., van Klink, R., Kuemmerle, T., Lomba, A., Pe'er, G., Plieninger, T., Rey Benayas, J. M., Sandom, C. J., Svenning, J.-C., & Wheeler, H. C. (2019). Rewilding

complex ecosystems. *Science*, *364*(6438), eaav5570. https://doi.org/10.1126/science.aav5570

Pichler, M., & Hartig, F. (2023). Machine learning and deep learning—A review for ecologists. *Methods in Ecology and Evolution*, *14*(4), 994–1016. https://doi.org/10.1111/2041-210X.14061

Poisot, T. (2023). Guidelines for the prediction of species interactions through binary classification. *Methods in Ecology and Evolution*, *14*(5), 1333–1345. https://doi.org/10.1111/2041-210X.14071

Poisot, T., Ouellet, M.-A., Mollentze, N., Farrell, M. J., Becker, D. J., Brierley, L., Albery, G. F., Gibb, R. J., Seifert, S. N., & Carlson, C. J. (2023). Network embedding unveils the hidden interactions in the mammalian virome. *Patterns*, *4*(6), 100738. https://doi.org/10.1016/j.patter.2023.100738

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

Riva, F., Graco-Roza, C., Daskalova, G. N., Hudgins, E. J., Lewthwaite, J. M. M., Newman, E. A., Ryo, M., & Mammola, S. (2023). Toward a cohesive understanding of ecological complexity. *Science Advances*, *9*(25), eabq4207. https://doi.org/10.1126/sciadv.abq4207

Roesch, E., Greener, J. G., MacLean, A. L., Nassar, H., Rackauckas, C., Holy, T. E., & Stumpf, M. P. H. (2023). Julia for biologists. *Nature Methods*, *20*, 655–664. https://doi.org/https://doi.org/10.1038/s41592-023-01832-z

Saravia, L. A., Marina, T., Kristensen, N. P., De Troch, M., & Momo, F. R. (2022). Ecological network assembly: How the regional metaweb influences local food webs. *Journal of Animal Ecology*, *91*(3), 630–642. https://doi.org/10.1111/1365-2656.13652

Saravia, L. A., Marina, T. I., Troch, M. D., & Momo, F. R. (2018). Ecological Network assembly: How the regional meta web influence local food webs. *bioRxiv*, 340430. https://doi.org/10.1101/340430

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

Thompson, P. L., & Gonzalez, A. (2017). Dispersal governs the reorganization of ecological networks under environmental change. *Nature Ecology & Evolution*, *1*(6), 0162. https://doi.org/10.1038/s41559-017-0162

Thuiller, W., Calderon-Sanou, I., Chalmandrier, L., Gaüzere, P., Ohlmann, M., Poggiato, G., & Münkemüller, T. (2023). Navigating the integration of Biotic Interactions in Biogeography. *Journal of Biogeography*. https://doi.org/10.1111/jbi.14734

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

Vázquez, D. P., Melián, C. J., Williams, N. M., Blüthgen, N., Krasnov, B. R., & Poulin, R. (2007). Species abundance and asymmetric interaction strength in ecological networks. *Oikos*, *116*(7), 1120–1127. https://doi.org/10.1111/j.0030-1299.2007.15828.x

Wood, S. L. R., Martins, K. T., Dumais-Lalonde, V., Tanguy, O., Maure, F., St-Denis, A., Rayfield, B., Martin, A. E., & Gonzalez, A. (2022). Missing Interactions: The Current State of Multispecies Connectivity Analysis. *Frontiers in Ecology and Evolution*, *10*.

Appendix A

Supplementary material for chapter 3

A.1. SVD does not overfit on the European network

In order to ensure that the creation of the RDPG on the European network does not lead to overfitting, we performed two numerical experiments.

First, we estimated the threshold that separates interactions from non-interactions based on a decreasing amount of species; this highlights that removing up to 50% of the total species in the network does not change the estimate of the threshold, suggesting that there is an important amount of information contained in the first 12 ranks of the network.

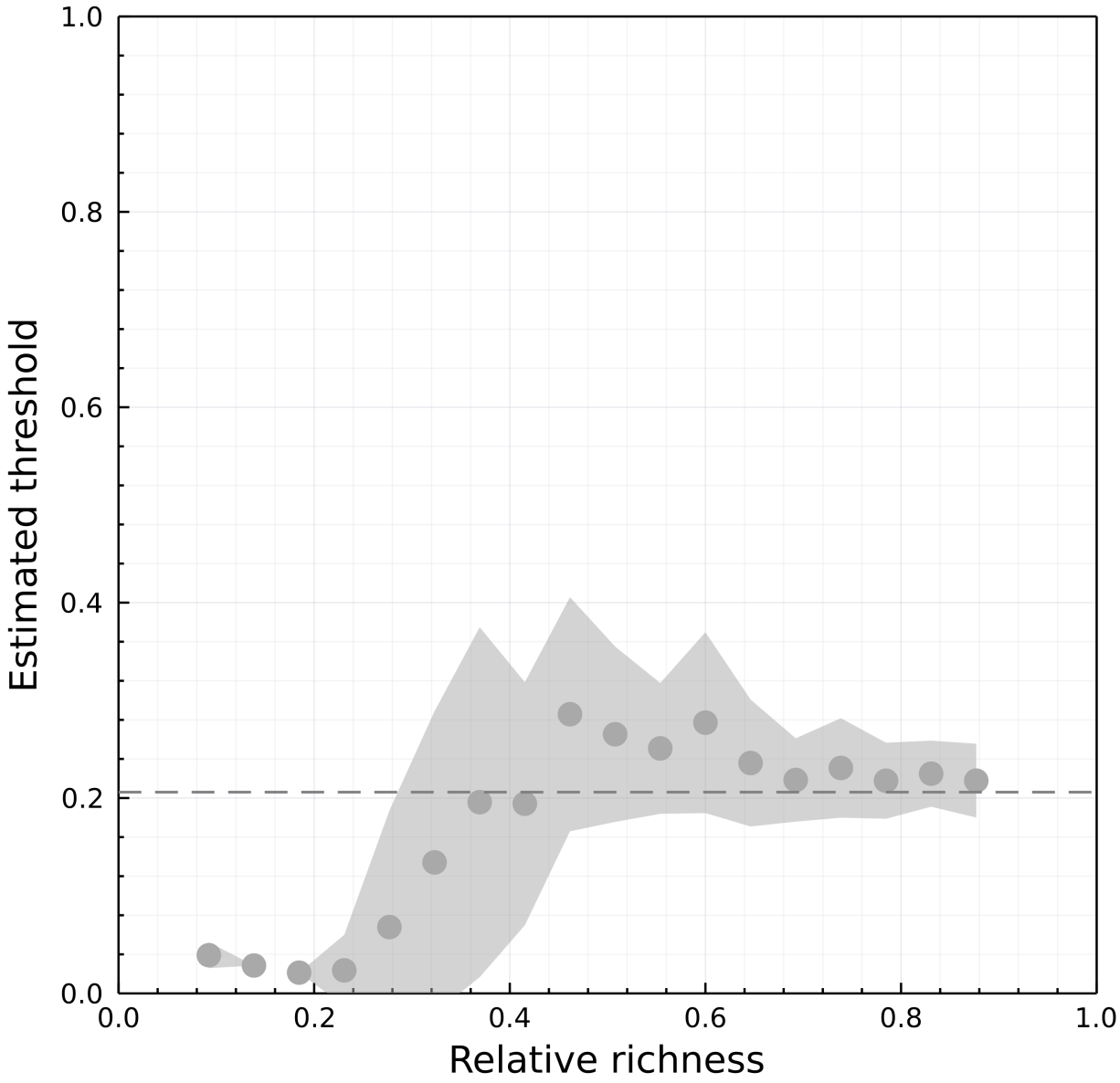
Second, we extracted $\mathcal{L}$ and $\mathcal{R}$, the left and right subspace of the entire network, at rank 12. Then, for every number n of interactions between 10 and links$(M) - 1$ (where M is the European metaweb), we define m as a network in which n interactions have either been randomly removed, randomly added, or both. We then define $\updownarrow$ and ∇ as the left and right subspaces coming from the rank-12 RDPG applied to this network, and compare the original network M to the one that was reconstructed after thresholding $\updownarrow\nabla$ by picking the cutoff that maximizes Youden's J measure (Youden, 1950).

This last experiment allows measuring the response of various measures of fit of the binary classifier to incomplete sampling. We are specifically interested in (i) the ability of RDPG to identify modified interactions, (ii) the ability of RDPG to function as a performant

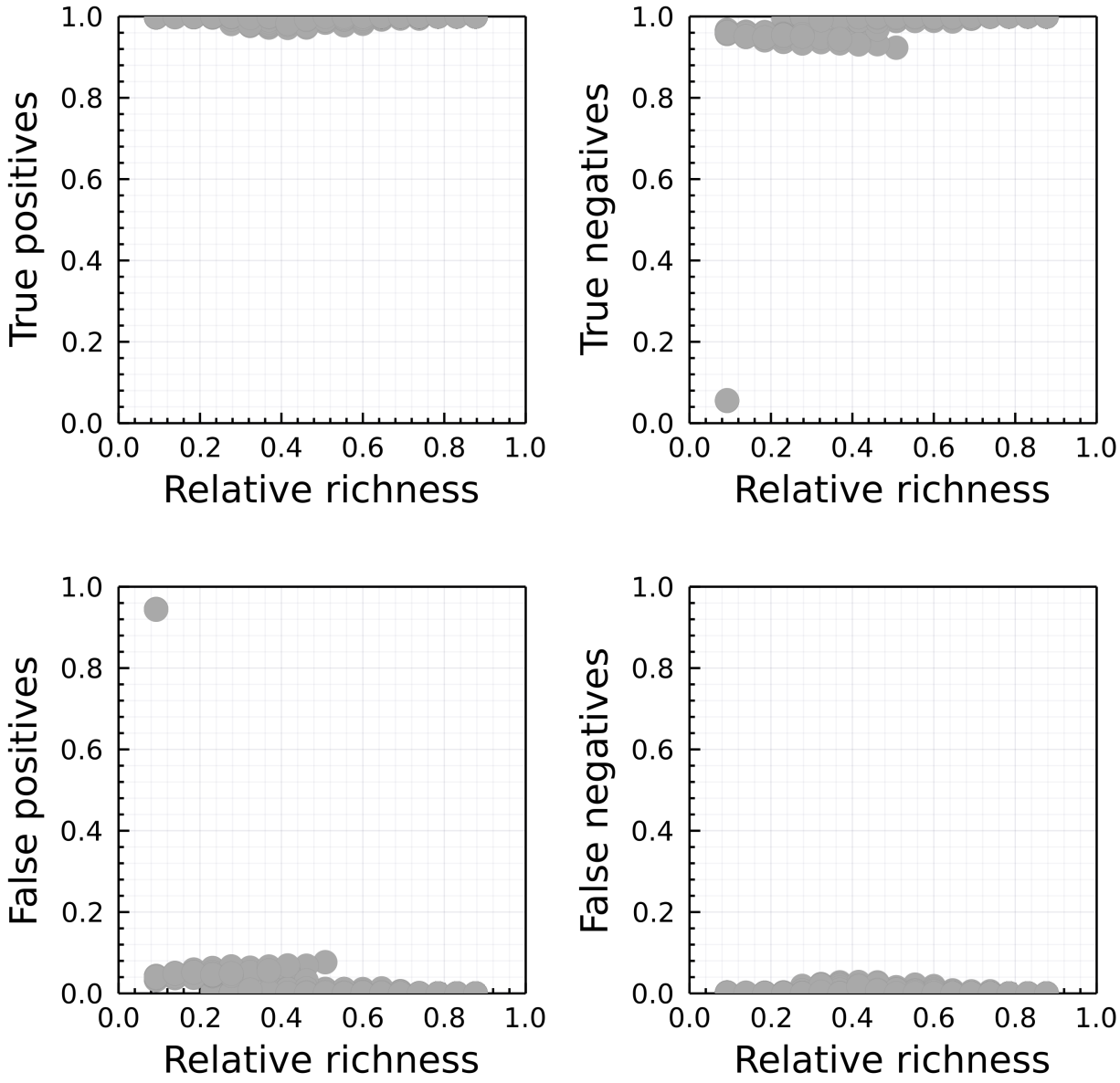
classifier in the presence of uncertainty in the original data, and (iii) the ability of RDPG to reconstruct biologically realistic data when interactions are withheld.

A.1.1. Threshold estimation is robust to species sub-sampling

In the initial experiment, we withheld an increasing number of species from the European metaweb, ranging from 20% for training and 80% for validation, to 90% for training and 10% for validation. Surprisingly, the estimation of the threshold, here presented as the mean and standard deviation of 50 folds for validation, is remarkably robust (and matches the value obtained using the entire network, as a dashed line). Specifically, even using 60% of species to estimate the threshold gave on average the same threshold as would be estimated based on the entire network; therefore, this establishes that the decision in the main text to use the entire European metaweb to set the threshold is correct.



More strikingly, looking at the rates of true/false positive/negative, as illustrated below, it is clear that RDPG can be thresholded in a way that yields an almost perfect classifier:

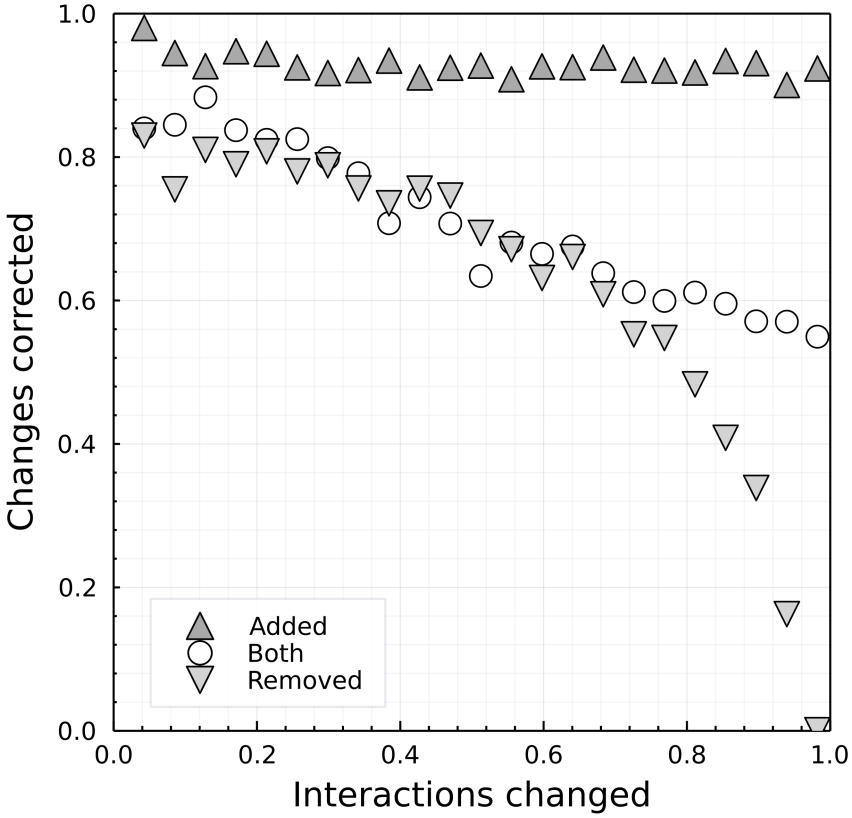


These results may be surprising, given that ecological models usually do not reach this degree of accuracy. That being said, we use the first 12 ranks of the network to approximate it, and this contains a lot of information; in short, the minute discrepancies between the predictions and the actual data can be attributed to leftover noise in the original dataset.

A.1.2. RDPG recovers withheld interactions

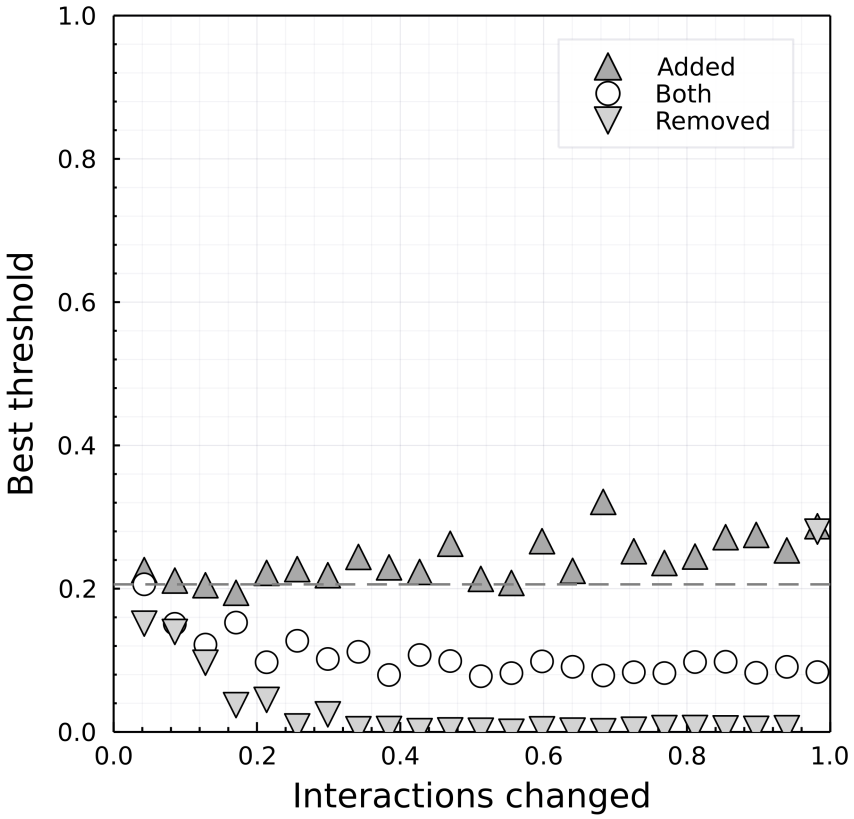
RDPG is able to correct almost all *added* interactions (added interactions were *not* originally present in the European metaweb and could be seen as introducing false positives to

the data), which is very strong evidence that the metaweb produced using it are not going to contain too much spurious interactions. When *removing* interactions (*i.e.* introducing false negatives to the data), even when half are missing, RDPG was able to accurately reconstruct about 75 to 80% of them. Predictably, the performance when both adding and removing interactions is in between the two scenarios.



The stochasticity in the proportion of recovered interactions is larger when a small number of interactions are withheld, which makes sense as the *number* of interactions is far smaller (compared to the overall network size).

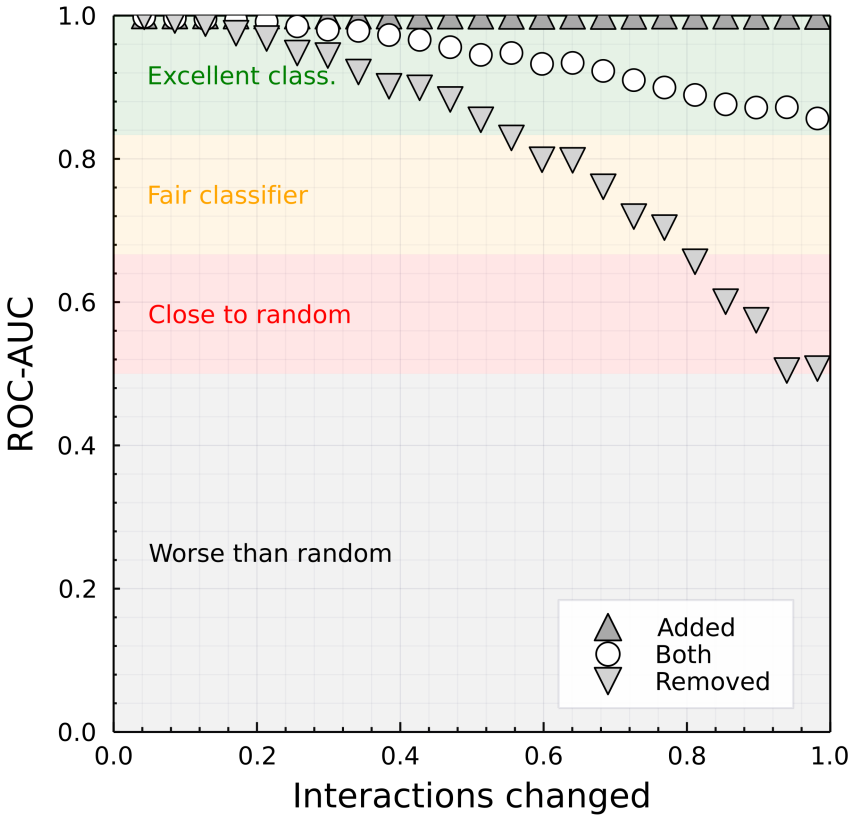
Next, it is interesting to note that the threshold "adapts" to the amount of missing information - the dashed line corresponds to the threshold we used in the manuscript. Adding interactions specifically did not result in an increase in the threshold, further suggesting that RDPG is extremely good at removing spurious interactions.



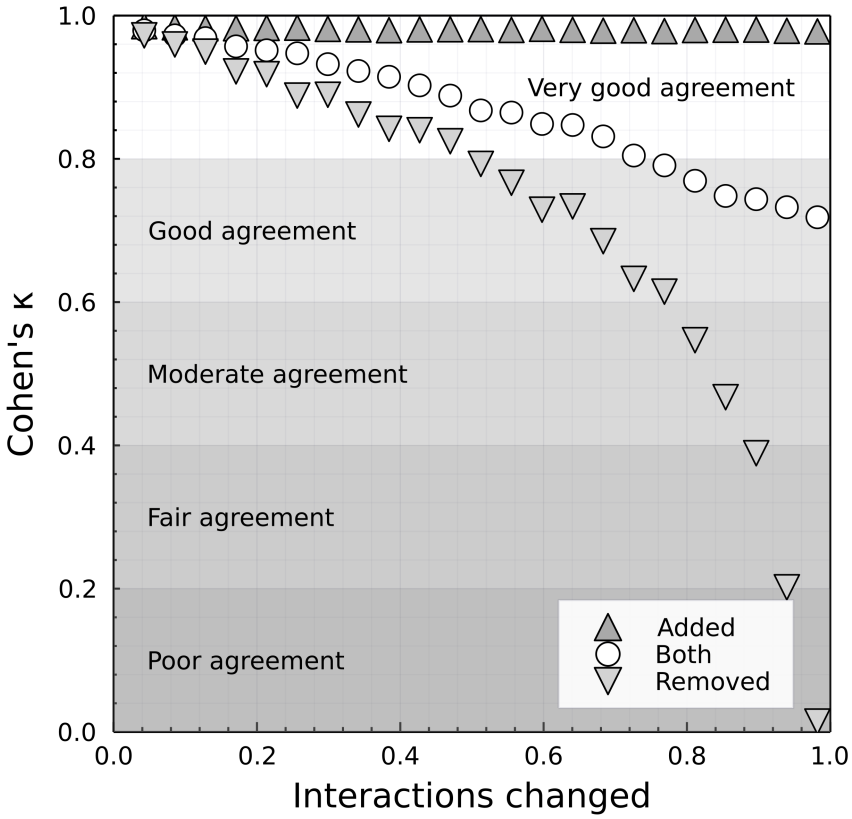
The important consequence of this result is that training the RDPG on a sub-sample of the network (*i.e.* one missing interactions) would result in a lower threshold, thereby potentially creating more false positives when applied to novel data; this further justifies our decision to use the entire evidence to estimate the threshold.

A.1.3. RDPG yields an accurate classifier

More important than the recovery of removed interaction is the fact that the classifier should have a good global performance. One measure to assess this is the area under the receiving operator characteristic curve, or ROC-AUC. By this measure, the RDPG remains an excellent classifier even if 50% of interactions are withheld, and no matter what the amount of changes are made by adding or both adding and removing interactions.



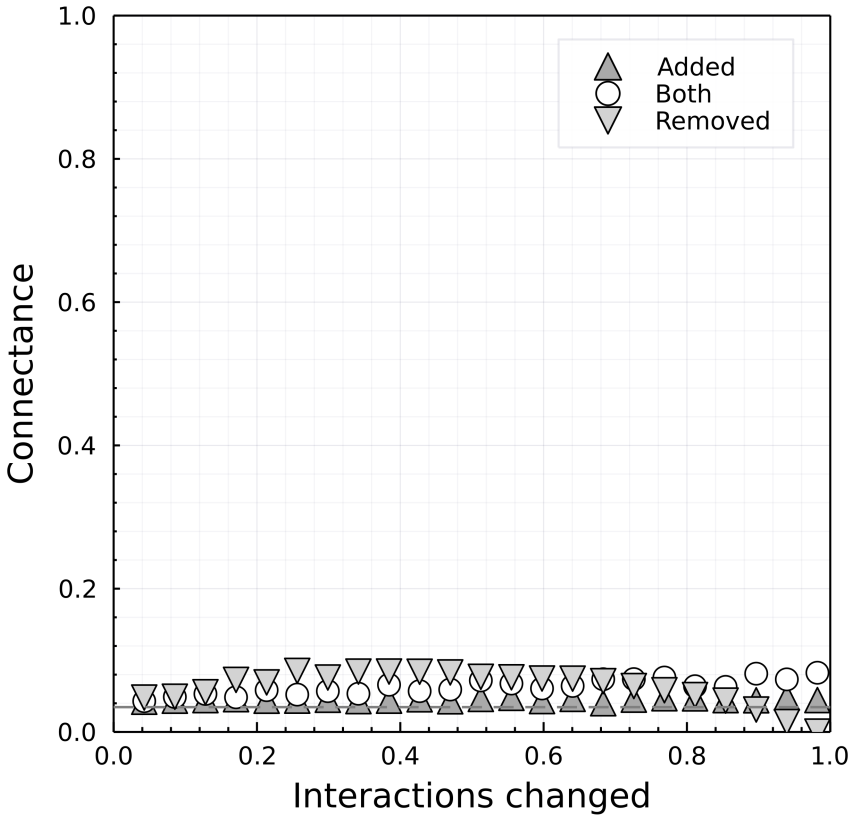
The overall agreement between a classifier and the actual data can be measured by Cohen's κ, which gives a similar result.



These two diagnostic figures reveal that, although we used a probably exhaustive list of interactions to do the initial RDPG, there are chances that the approach would work on less complete datasets.

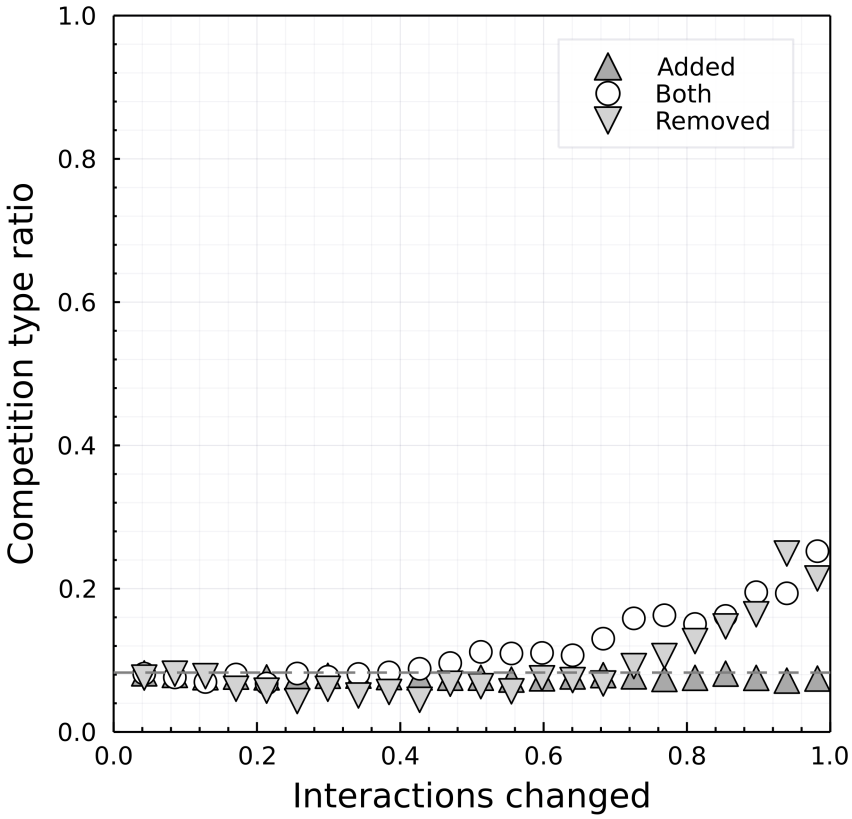
A.1.4. RDPG recreates ecologically realistic networks

In this section, we present the relationship between the empirical measure of the network structure (dashed line) and the reconstructed estimate based on RDPG after the optimal threshold has been applied. We focus on connectance (for its broad relevant to food web structure) first:



Connectance increases slightly when initial information is incomplete, but saturates at a value of around 0.12 – this is still within the bounds of connectances expected for food webs.

Next, we look at the ratio between direct competition ($a \rightarrow (b,c)$) and apparent competition ($(a,b) \rightarrow c$) motifs, as motifs are known to be conserved blocks in food webs:



This ratio remains close to the real one up until 75% of initial interactions are modified.

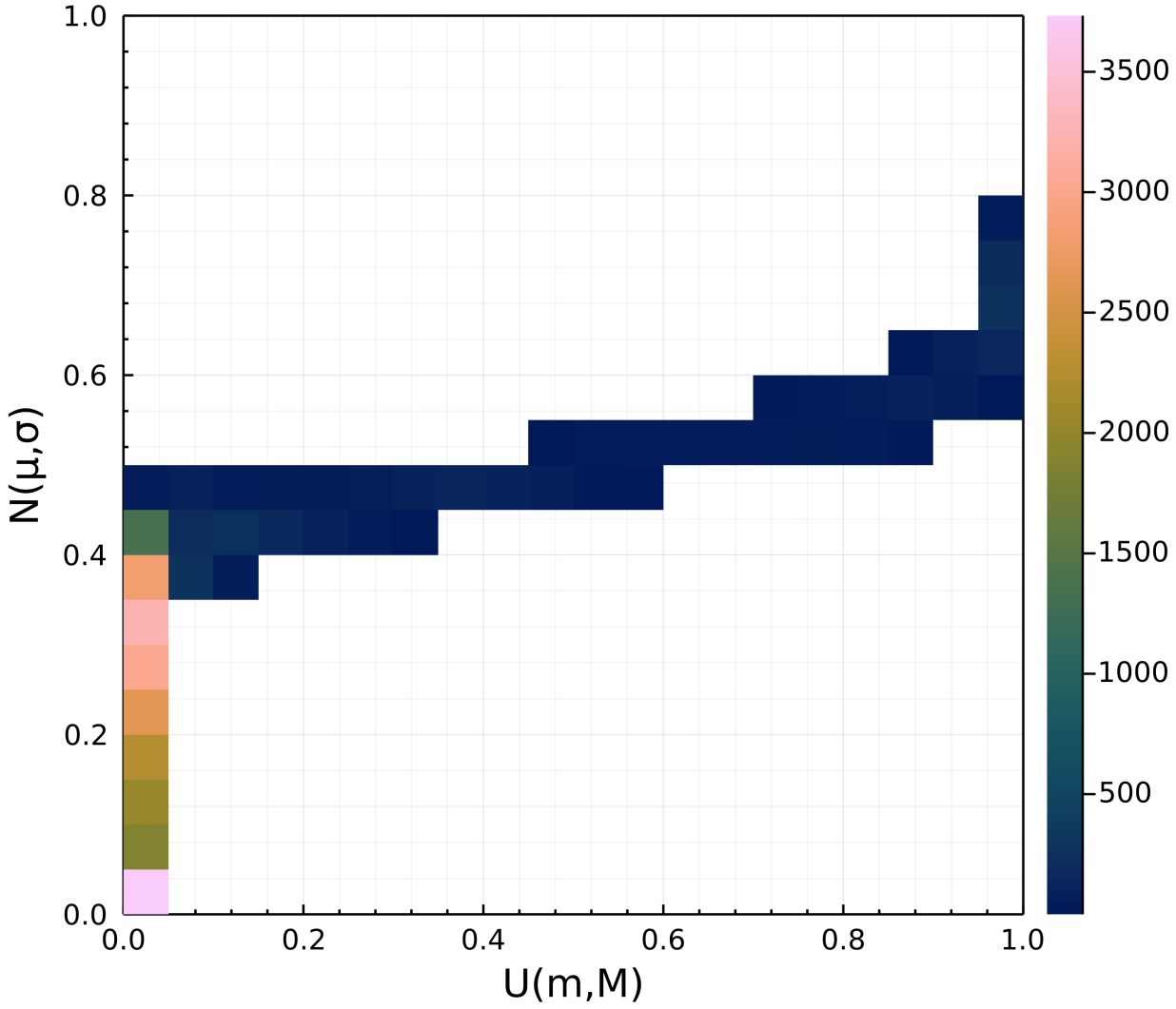
A.1.5. Consequences

Based on these results, applying RDPG on the entire European network is reasonable, especially since (i) the threshold is insensitive to the number of withheld species, and (ii) removing interactions would artificially lower the threshold. Interestingly, the RDPG remains an excellent binary classifier even in the face of strong data modifications, which suggests that our framework can be used even in the absence of a complete metaweb. Even more importantly, the addition of wrong interactions to the original dataset was never an issue for the RDPG classifier, which was almost always able to remove them.

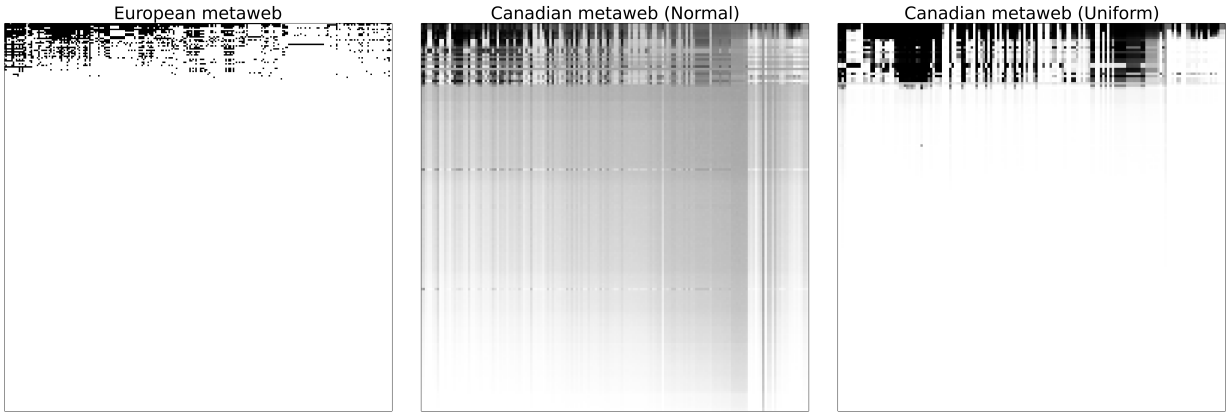
A.2. The Normal model of latent variable evolution over-predicts

In this appendix, we compare the raw predictions made by the Normal and Uniform models of latent variable evolution. The Normal model was created by (i) getting the average μ of the simulated values for each species/variable combination, and (ii) estimating the standard deviation as $(\mu + c - \mu - c)/3.92$, where c is one half of the 95% confidence interval around μ divided by 3.92

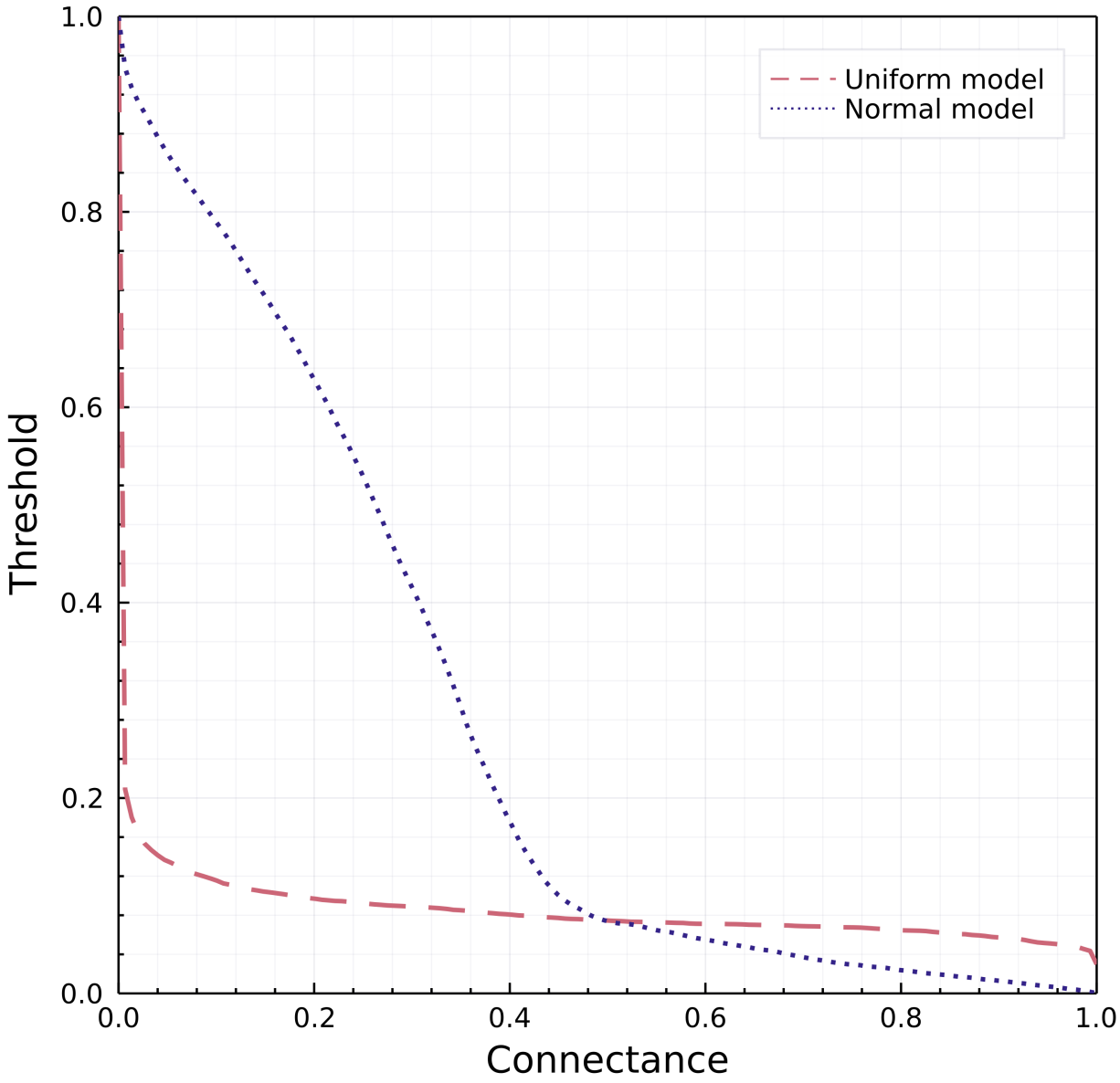
As can be seen on the following figure, the Normal model tends to assign high probabilities (up to $p \approx 0.4$) for interactions that the Uniform model essentially rules out:



This can lead to severe over-estimation of the number of interactions. In fact, the consequences of using a Normal model are obvious from looking at the adjacency matrices below: most of the interactions are predicted between species that occupy the lower trophic level, and are ecologically unrealistic.



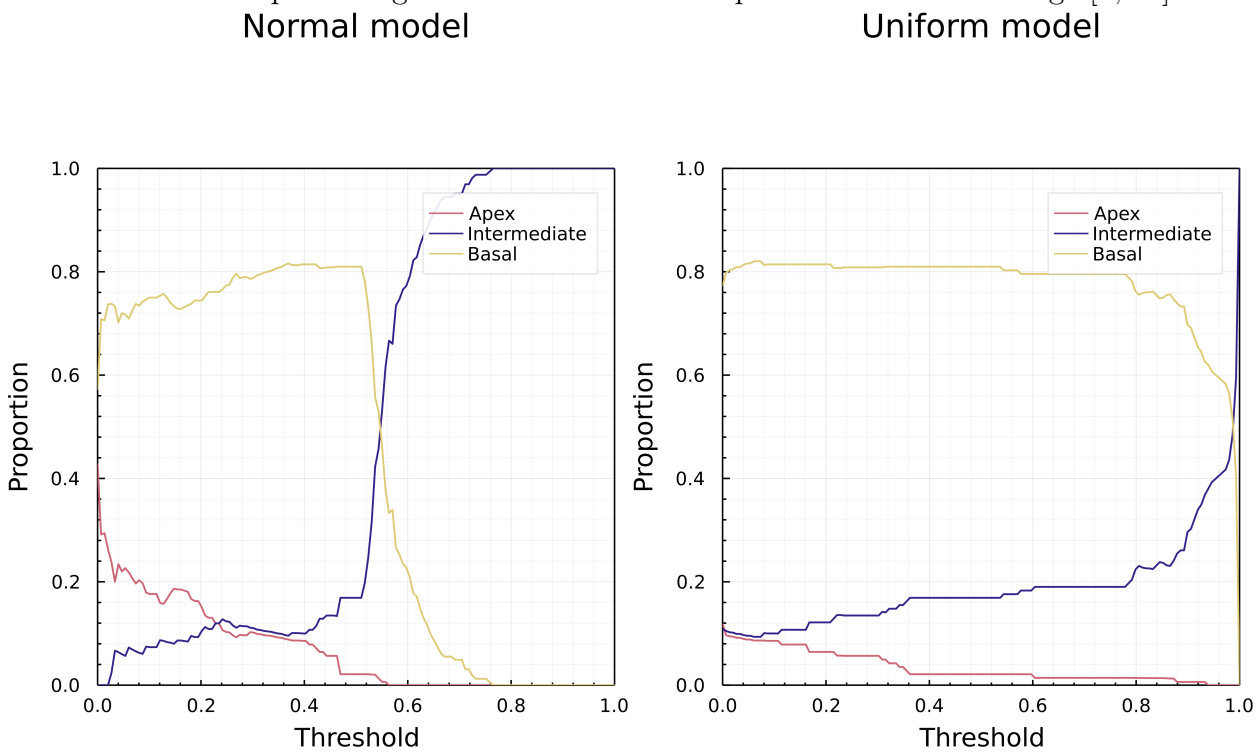
This can be further revealed by looking at the connectance of the networks under different thresholds:



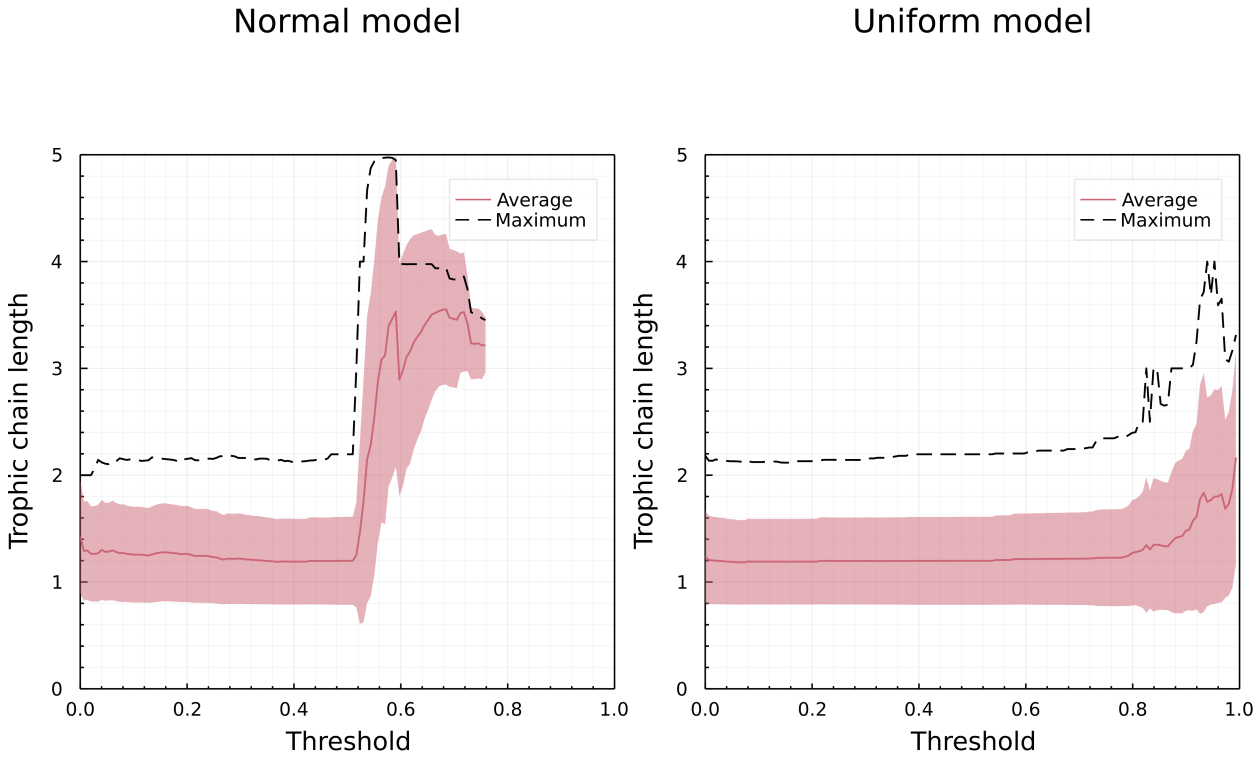
Although the Uniform model predicts a lot of interactions with extremely low probability, that are removed at a low threshold, the distribution of probabilities under the Normal model leads to extremely (abnormally) high connectances even for thresholds that are over twice as large as the optimal threshold determined in main text and Supp. Mat. 1.

This has consequences for the overall network *structure*: specifically, the Normal model predicts a lot more top predators than we expect under the uniform model; rather than there being a progressive change in top-intermediate-bottom proportions as the threshold changes,

there is an abrupt shift at a threshold of about 0.6, which suggests that the Normal model is biased towards over-predicting most interactions with probabilities in the range [0,0.6].



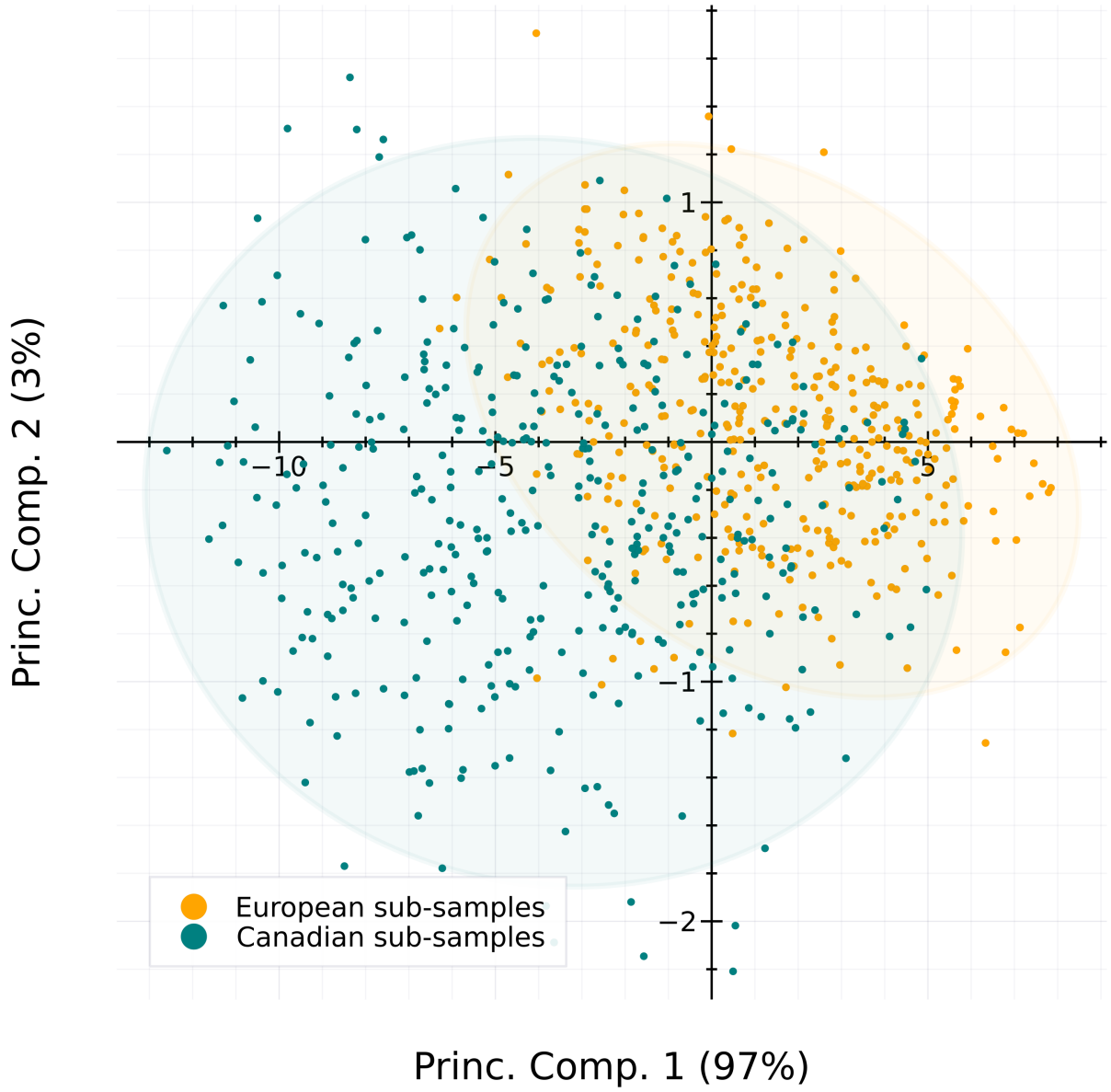
The same "jump" can be observed when looking at the distribution of food chain lengths:



For these reasons, we only use predictions from the Uniform model in the main text.

A.3. RDPG reconstructed networks have diverse structures

In this appendix, we check that the networks reconstructed from the RDPG do keep a variety of structural components, especially when selecting a small species pools from within them. In order to do so, we induced 400 random subgraphs containing between 30 and 70 species, both from the Canadian and European metawebs. For each of these subgraphs, we measured eight variables: the mean and standard deviation of trophic levels, the standard deviation of degree (total, in, and out), and the proportion of top, intermediate, and basal species. We selected a random subset of 300 rows from the network-property matrix to fit a Principal Component Analysis projection matrix (W), which we then used to project all networks into the space formed by the first two principal components.



The first axis (explaining most variance) was strongly correlated to the standard deviation of the number of preys (-0.71), and the second axis to the standard deviation in the number of predators (-0.95). These results match the conclusions in main text, namely that the first dimensions of network embedding capture the degree distribution.

Two things are important to note on this representation; each point is an induced sub-graph, and the ellipses are the 95% confidence interval around the points. First, there is some variations *within* a group (Europe *v.* Canada); second, the two groups do not fully overlap. This suggests that not only the sub-samples of the Canadian metaweb are not equivalent to

the sub-samples of the European metaweb (*i.e.* the two networks have structural differences), realizations (here in the form of random local species pools) of the Canadian metaweb also show some variability; in short, reconstructing a metaweb using a RDPG will not result in homogeneous local networks, and may therefore be suitable for lower-scale predictions.

Appendix B

Supplementary material for chapter 2

The associated materials for this appendix is in the form of a Jupyter notebook file (a web-based interactive computing platform). At the time of writing the published article was not yet available online and thus this appendix will point to the notebook file that was associated with the GitHub repository associated with this chapter. The *.ipynb* file can be found here. Note it is also possible to download this file and open it in a Jupyter application should you wish for a more interactive experience.

Appendix C

Understanding where networks stop

C.1. Why boundaries are interesting

As discussed in both Chapters 2 and 5 there is value in thinking about the existence of boundaries between networks, either from a prediction perspective (*e.g.,* knowing at what scale to make predictions at) or from a more theoretical question of where do networks stop? Although this is discussed in more detail in subsection 6.2.2 there is one question regarding network boundaries that might be a good starting point and that is looking at how environmental and network boundaries relate, more specifically do environmental changes drive changes in networks. Here I present what is more of a methodological framework (as opposed to an actual answer, hence why it has been relegated as an Appendix) that we can use to try and answer this question.

C.2. A metacommunity model for boundary detection

The metacommunity model developed by Thompson and Gonzalez, 2017 is a good starting point to use for this 'case study' as it allows us some flexibility with how we want to parameterise the system. The model (C.2.1) itself is based on a tritrophic community ('plants', 'herbivores', and 'carnivores') and is a collection of modified Lotka–Volterra equations and (broadly) models species abundance as a function of interaction strength, environmental

effect, immigration, and emigration. The metacommunity consists of S species with M environmental patches and looks as follows:

$$X_{ij}(t+1) = X_{ij}(t)exp\left[C_i + \sum_{k=1}^{S} B_{ik}X_{kj}(t) + A_{ij}(t)\right] + I_{ij}(t) - X_{ij}(t)a_i \qquad \text{(C.2.1)}$$

Where $X_{ij}(t)$ is the abundance of species i in patch j at time t. C_i is its intrinsic rate of increase (which we have set to 0.1 for 'plants' and -0.01 for 'herbivores' and 'carnivores'). B_{ik} is the per capita effect of species k on species i. The exact interaction strength for each species pair is drawn from a uniform distribution with the parameters for the interaction pairs listed in Table 1, the values drawn from the uniform distribution are scaled by dividing by $0.33S$ to yield the final interaction strength for each interacting pair.

Interacting pair	Range of uniform distribution
Plant-plant	-1 – 0
Plant-herbivore	0 – 0.1
Plant-carnivore	0
Herbivore-plant	-0.3 – 0
Herbivore-herbivore	-0.2– -0.15
Herbivore-carnivore	0 – 0.08
Carnivore-plant	0
Carnivore-herbivore	-0.1 – 0
Carnivore-carnivore	-0.2– -0.15

Table 1. Intervals used for the uniform distribution from which interaction strengths values are drawn from for the different types of species pair interactions. Note this is represent the effect of species type 1 on species type 2 *i.e.,* herbivore-plant represents the effect of a herbivore species on a plant species

$A_{ij}(t)$ is the effect of the environment in patch j on species i at time t and can be further expanded as follows:

$$A_{ij}(t) = h\left(exp - \frac{(E_j(t) - H_i)^2}{2\sigma^2} - 1\right) \qquad \text{(C.2.2)}$$

Species environmental optima (H_i) are evenly distributed across the entire range of environmental conditions for each trophic level, meaning that species from different trophic levels will be at, or near the same environmental optima. h is a scaling parameter (set to

300), $E_j(t)$ is the environment in patch j at time t and σ is the standard deviation (set to 50).

$I_{ij}(t)$is the abundance of species i immigrating to patch j at time t and can be expanded as follows:

$$I_{ij}(t) = \sum_{l=j}^{M} a_i X_{il}(t) exp(-L d_{jl}) \tag{C.2.3}$$

Where ai is the proportion of the population of species i that disperses at each time step, the dispersal rate is drawn from a normal distribution ($\mu = 0.1$, $\sigma = 0.025$) for each species. The abundance of immigrants to patch j from all other patches is governed by where d_{jl} is the geographic distance between patches j and l, and L (the strength of the exponential decrease in dispersal with distance), which is also drawn from a normal distribution for each species. The parameters used for L are trophic level dependant and are show in Table 2

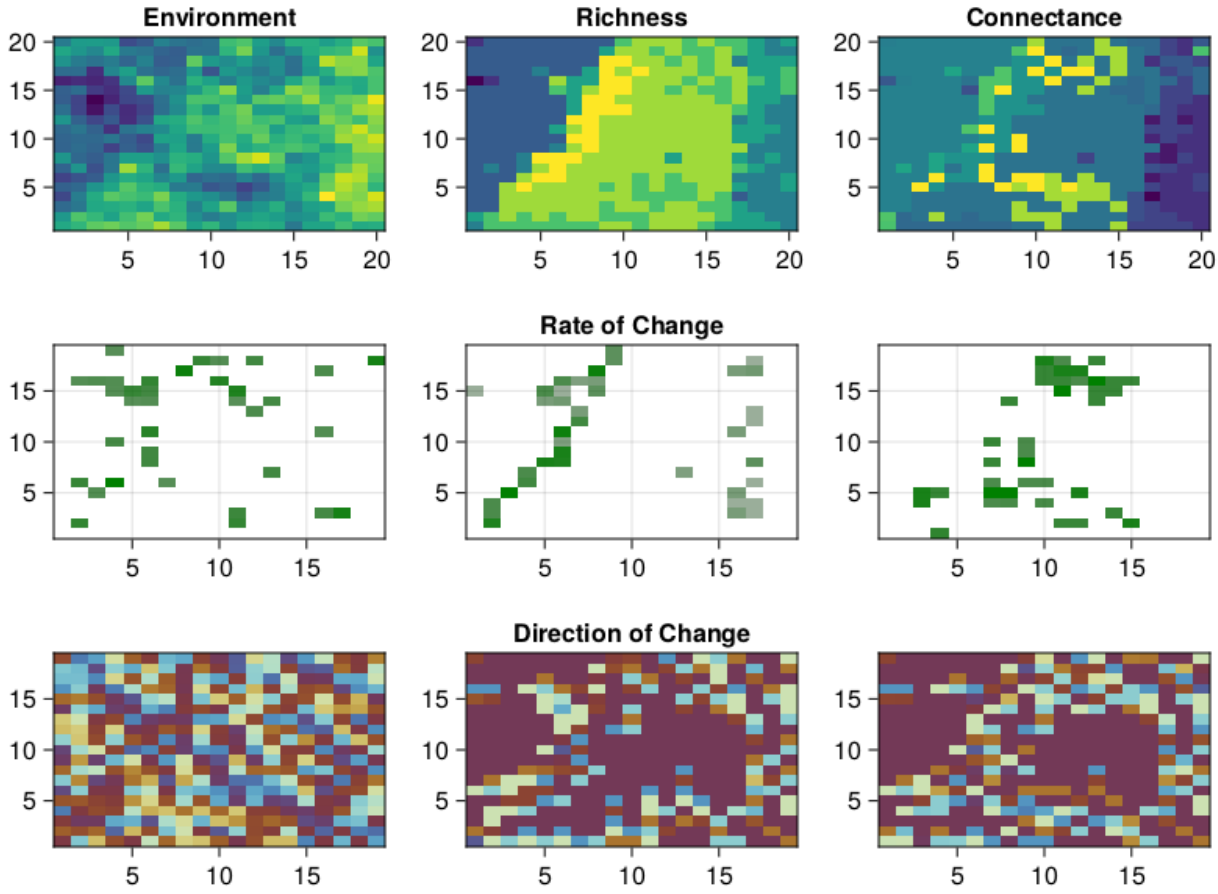
Trophic level	μ	σ
Plant	0.3	0.075
Herbivore	0.2	0.05
Carnivore	0.1	0.025

Table 2. Parameters for the normal distributions used to determine the dispersal decay (L) for each species depending on its trophic level.

C.3. A toy example of boundary detection

The associated code for these simulations was carried out in `Julia 1.8` (Bezanson et al., 2017) using `Makie.jl` (Danisch and Krumbiegel, 2021) and `SimpleSDMLayers` (Dansereau and Poisot, 2021), this can be found in a GitHub repo `here`. Note that the results presented here are supposed to represent a hypothetical result and it is not so much that there is ecological knowledge to be gleaned from the figure below but rather to showcase how we can approach the idea of boundary detection across landscapes. For the initial modelling exercise presented below I have used 80 species ($S = 80$) within a 20 by 20 landscape (*i.e.,* $M = 400$). This landscape is generated using `NeutralLandscapes.jl` ("Neutral Landscapes",

2021/2023), which allows the user to specify different landscape types *e.g.,* one with a clear boundary, the landscape values are used to represent the environmental value for the specific patch.



The top row represents the 'raw' values for the landscape after 500 generations. Note here I have included species richness as it might be interesting to see if species richness is related to network structure, in this instance I have used connectance as that measure of network structure since it is one of the more common network metrics to use.

The second row show the rates of change for the respective metrics, but only limited to the top 90% rate of change values for a cleaner visual. Here the colour intensity indicates the magnitude of the rate of change. The final row is showing the direction of change for the respective metrics and each colour can be thought of as indicating a cardinal point.

What is interesting about this simulation is that the rate of change for environmental, species richness, and connectance do not 'line-up' and that the species community and the

way they interact are responding differently to changes in the environment - although the direction of change seems quite similar for species richness and connectance. This is also interesting since it might suggest that although the exact 'location' of changes might be different the way the change is propagated across the landscape is the same. Overall I would argue that there is evidence that indicates that the idea of 'ecotrophic boundaries' is one worth exploring further.

References

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

Danisch, S., & Krumbiegel, J. (2021). Makie.jl: Flexible high-performance data visualization for Julia. *Journal of Open Source Software*, *6*(65), 3349. https://doi.org/10.21105/joss.03349

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

Neutral Landscapes. (2023, August 31). https://github.com/EcoJulia/NeutralLandscapes.jl

Thompson, P. L., & Gonzalez, A. (2017). Dispersal governs the reorganization of ecological networks under environmental change. *Nature Ecology & Evolution*, *1*(6), 0162. https://doi.org/10.1038/s41559-017-0162