An appraisal of graph embeddings for comparing trophic network architectures

Christophe Botella, Stéphane Dray, Catherine Matias, Vincent Miele, Wilfried Thuiller

An appraisal of graph embeddings for comparing trophic network architectures

Christophe Botella[1,*], Stéphane Dray[2], Catherine Matias[3], Vincent Miele[2], and Wilfried Thuiller[1]

[1]Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LECA, Laboratoire d'Ecologie Alpine, F-38000 Grenoble, France
[2]Université de Lyon, F-69000 Lyon; Université Lyon 1; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Évolutive, F-69622 Villeurbanne, France
[3]Sorbonne Université, Université de Paris, Centre National de la Recherche Scientifique, Laboratoire de Probabilités, Statistique et Modélisation, F-75005 Paris, France
[*]Corresponding author: christophe.botella@gmail.com

Running headline: Compare trophic networks with graph embedding

Abstract

1. Comparing the architecture of interaction networks in space or time is essential for understanding the assembly, trajectory, functioning and persistence of species communities. Graph embedding methods, which position networks into a vector space where nearby networks have similar architectures, could be ideal tools for this purposes.
2. Here, we evaluated the ability of seven graph embedding methods to disentangle architectural similarities of interactions networks for supervised and unsupervised posterior analytic tasks. The evaluation was carried out over a large number of simulated trophic networks representing variations around six ecological properties and size.
3. We did not find an overall best method and instead showed that the performance of the methods depended on the targeted ecological properties and thus on the research questions. We also highlighted the importance of normalizing the embedding for network sizes for meaningful posterior unsupervised analyses.
4. We concluded by orientating potential users to the most suited methods given the question, the targeted network ecological property, and outlined links between those ecological properties and three ecological processes: robustness to extinction, community persistence and ecosystem functioning. We hope this study will stimulate the appropriation of graph embedding methods by ecologists.

Keywords: dimension reduction; ecological interaction networks; evaluation; food-webs; graph embedding; species interactions; trophic networks; trophic groups

1 Introduction

Community ecology is entering a new era, where data are becoming multi-species, multi-trophic, and integrate species interactions [Pellissier et al., 2018]. So far, ecologists have compared community composition data on the basis of species identity, functional traits or phylogenetic similarities (reviewed in Münkemüller et al. [2020]). With the ever-increasing availability of interaction knowledge, we are now facing the opportunity to also compare communities based on their interaction network architecture, i.e. the configuration of the interaction links between species of a community. This might provide crucial insights to describe biodiversity variations across environments [Pellissier et al., 2018], unveil network architecture similarities across communities composed of different species [Ohlmann et al., 2019] or understand the assembly rules behind multi-trophic assemblages [Münkemüller et al., 2020]. In addition, the ecological properties that can be described from the architecture of a trophic network (e.g. degree of omnivory, generalism, compartmentalization, number of trophic levels) are important to characterize ecosystem functioning, resilience and robustness to extinctions [Monteiro and Faria, 2016, Tylianakis et al., 2010]. To address these objectives, graph embedding methods that cast many networks into a common multi-dimensional vector space reflecting many aspects of architectural variations across the networks (e.g. Narayanan et al. [2017]), are appealing. They allow standard multivariate analyses to be applied a posteriori to a set of networks, including descriptive analyses (e.g. dimension reduction techniques for visualization) and supervised learning (i.e. predicting an external characteristic from a network embedding coordinates based on knowledge of its values over a sample of network examples). Despite the diversity of individual network metrics or motifs studied in ecology [Lau et al., 2017], a small number of multi-dimensional graph embedding methods have been applied to trophic networks, and there has been no comparison of their abilities to capture the signatures of ecological processes on network architectures.

The ecological properties of a trophic network partially determines its dynamics, especially its persistence, its robustness to extinctions and other ecological processes. The distribution of species across trophic levels in a network especially impacts its robustness to extinctions [Pimm et al., 1991]. For example, a lower proportion of basal species induces less prey per predators in higher trophic levels and thus increases the likelihood of secondary extinctions and extinction cascades. Regarding community persistence, longer trophic chains are also suggested to decrease the recovering rate of species populations after disturbance [Pimm et al., 1991], which explains shorter chains in fluctuating environments like for insect food webs. The length of trophic chains may also impact the global balance of carbon fluxes in the ecosystem through compensation of primary production and respiration as shown for lake ecosystems [Schindler et al., 1997]. Compartmentalization has been theoretically shown to favor robustness to extinctions in food webs because it limits the effect of extinction cascades across modules [Thébault and Fontaine, 2010, Tylianakis et al., 2010]. Compartmentalization can also impact ecosystem functioning, for instance by decreasing parasitism rate in plant-herbivore-parasite system [Montoya et al., 2003]. Regarding more local ecological properties, predator generalism (number of preys), which is related to connectance, increases robustness to extinctions [Thébault and Fontaine, 2010]. Morever, theory suggests a strong link between generalism and the long-term persistence of community [Pimm et al., 1991, Thébault and Fontaine, 2010, Tylianakis et al., 2010], even though the precise mechanism behind this phenomenon still appears ambiguous. While it is often documented that generalism negatively affects community persistence [Thébault and Fontaine, 2010, Torres-Alruiz and Rodríguez, 2013], it may also provide a buffer in the response of individual predators to stochastic fluctuations of prey abundances [Tylianakis et al., 2010]. This paradox is apparently resolved when high generalism is composed of many weak links which favors persistence [McCann et al., 1998]. While generalism makes a species less sensitive

to varying prey populations, vulnerability increases its population control. When both increase, the biomass transfer rates are optimised at the network scale and may improve ecosystem resilience. For instance, the vulnerability of herbivores to many predators improved their population control in a collard-aphid system [Snyder et al., 2006], but excessive competition can lead to the opposite effect [Montoya et al., 2003]. Omnivory is another local property known to influence parasitism rate [Montoya et al., 2003] and community persistence [Borrelli, 2015, Pimm et al., 1991]. Loop patterns are suggested to destabilize trophic networks and decrease the persistence of species participating in them. Indeed, triangular motifs containing loops are less stable compared to other triangular motifs which was proposed as an explanation of their rarity in empirical food webs [Borrelli, 2015, Monteiro and Faria, 2016]. As those ecological properties can be measured from network architectures and are important for many ecological processes, they provide a solid ground for comparing networks through graph embedding methods.

Amongst the few methods used to analyse the spatial variation of interaction networks, most quantify interaction turnover (i.e. turnover of links) between pairs of species [Ohlmann et al., 2019, Poisot et al., 2012]. However, these methods do not account for the ecological properties that involve more than two species like omnivory and generalism degrees, trophic levels, loops, or compartmentalization. Yet, it is crucial to compare these properties across distant ecosystems, independently of pure species and interaction turnover. Thus, measuring more complex network ecological properties requires to look at sub-structures composed of more than two species, often interlinked and very numerous. These characterizations of network architectures may be simplified by finding a suited vector space where each network will be represented as a vector, which is exactly the purpose of graph embedding methods. Some common network metrics in ecology (e.g. connectance) may be seen as components of graph embedding methods [Braga et al., 2019, Kortsch et al., 2019, Thompson and Townsend, 2005, Wood et al., 2015]. Ecologists have indeed already used node level metrics (e.g. average number of links per node), node distances (e.g. diameter, mean distance), and whole networks metrics (e.g. connectance, modularity or nestedness) that measure specific properties of the network architecture. However, it is notoriously difficult to select the right metric to measure the variability of an ecological property [Braga et al., 2019] and one may also miss variations that are not explicitly dealt with the selected metrics. Counts of subgraphs with fixed number of nodes, called motifs, are also used in ecology. For instance, counts of triangular and bipartite motifs were applied to compare either trophic network architectures [Camacho et al., 2007] and plant-pollinator mutualistic networks [Simmons et al., 2019]. However, these methods are restricted to small motif sizes due to computational complexity (i.e. up to 4-nodes in food webs, see Monteiro and Faria [2016], up to 6-nodes in bipartite mutualistic networks, see Simmons et al. [2019]).

Interestingly, a spectrum of efficient graph embedding methods have been developed in other domains to represent networks based on different types of sub-structures. Some methods may further integrate information on node labels to which we could feed information on trophic groups [Cirtwill et al., 2018], i.e. sets of species sharing similar prey and predators, or any other relevant external information that classifies species into groups. Several sophisticated unsupervised machine learning algorithms (e.g. UGRAPHEMB Bai et al. [2019], Graph2Vec Narayanan et al. [2017]) have been proposed to produce graph embeddings. Particularly, Graph2Vec [Narayanan et al., 2017] is an interesting candidate for ecological applications. In this method, networks composed of similar node neigborhoods are represented by embedding coordinates that are close in terms of Euclidean distance. Shortest-paths lengths have also been used to compare network architectures [Borgwardt and Kriegel, 2005]. Comparing shortest-paths lengths across trophic networks is also interesting from a functional point of view. Indeed, shortest-paths lengths encode information related to trophic chain lengths or energy flows and are at the root of centrality measures [Costa et al., 2019]. These methods might bring different and relevant perspectives to ecological network analyses, but they

require a comprehensive and contextual understanding to be successfully applied.

To compare graph embedding methods for their use in ecology, we need to define their usage scenarios. They may be used for different posterior analytic tasks on networks including supervised and unsupervised learning. An unsupervised learning task would typically consist in identifying gradients or clusters of variation through visualization after reducing the embedding space dimensionality (e.g., 2 dimensions). In this visualization space, a set of networks that appear clustered, are thus neighbors in the embedding and share similarities in their architectures. This task requires depicting, in the embedding matrix, the main architectural variations existing across networks while reducing the effects of non-interesting sources of variability (e.g, network size). Alternatively, supervised learning aims to predict a given property of a network as a function of its embedding coordinates. For instance, one may want to predict the robustness to extinction [Dunne et al., 2002] as a function of the counts of shortest-path lengths. We therefore appraise the ability of graph embedding methods to represent important architectural variations across trophic networks and for posterior supervised and unsupervised analyses. We first introduce five graph embedding methods that are relevant to trophic network analyses and based on different architectural characteristics (e.g. motifs or paths lengths). Since two of these methods can handle node labels information, we thus moreover test the use of known trophic groups as node labels. Second, we illustrate the dimension reduction step for visualization (unspervised learning) with a recent non-linear dimension reduction technique called Uniform Manifold Approximate Projection (UMAP, McInnes et al. [2018]). Third, we detail our simulation procedure of trophic networks where we control the variation of six important categorical ecological properties (maximum trophic length, trophic groups composition, compartmentalization, omnivory, generalism and intra-trophic group predation) and, independently, species richness. Fourth, we introduce several measures to evaluate the relative performances of the embedding methods for supervised and unsupervised scenarios, including robustness to species richness variability for the latter. This methodological workflow is summarized by Figure 1. We finally guide the user towards the most suited method given the general aim of the analysis and the network ecological properties focused on, and further relate these properties to important ecological processes.

2 Material and Methods

2.1 Embedding methods

Among the available graph embedding methods, we selected five of them that should prove useful for ecological data and are relatively easy to use (Table 1). Two of them (Graph2Vec and ShortPaths2Vec) can also use prior information on species (node labels) like the belonging to a trophic group. These methods all apply to directed unweighted graphs and thus account for the asymmetry of interactions between species' pairs. We consider by convention that interactions are directed from prey to predators.

Acronym	Principle	Reference
Groups2Vec	Trophic group proportions	This study
Metrics2Vec	Seventeen classic foodweb metrics	Thompson and Townsend [2005]
Motifs2Vec	Directed triangular motif proportions	Camacho et al. [2007]
Graph2Vec	Decomposition into local neighborhoods	Narayanan et al. [2017]
Graph2Vec_lab	Graph2Vec + trophic groups as node labels	Narayanan et al. [2017]
ShortPaths2Vec	Shortest-paths lengths distribution	This study
ShortPaths2Vec_lab	ShortPaths2Vec + trophic groups as node labels	This study

Table 1: Graph embedding methods tested in this study.

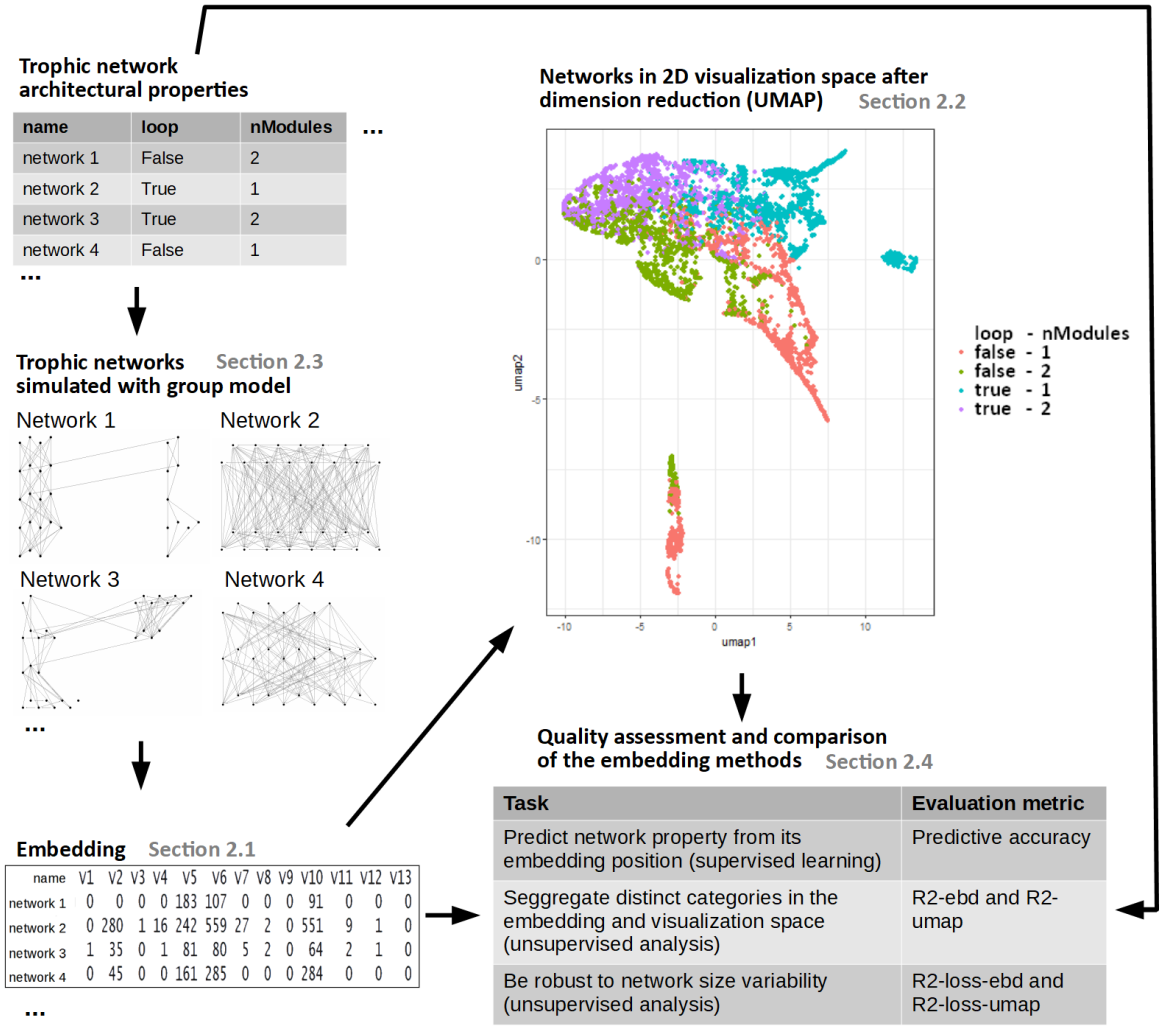

Figure 1: Study workflow. We used a group model to simulate 5000 trophic networks with controlled ecological properties, cast them in a graph embedding matrix using various methods, reduced the embedding matrix dimension to generate a 2D visualization space, and compare the embedding methods qualities for supervised and unsupervised posterior analysis tasks.

Groups2Vec. The underlying idea is that all species of the metanetwork (which describes potential interactions between the union of species over all networks) can be sorted by their belonging to trophic groups, meant to represent their topological role in the metanetwork [Cirtwill et al., 2018]. In general, there exist many ways to define a partition of species into trophic groups. For instance, one could rely on predefined known guilds or trophic groups, build a classification from species functional traits or use a statistical approach that group species behaving in the same way in the network, like a Stochastic Block Model (SBM, see Allesina et al. [2008], Kéfi et al. [2016]) applied on the metanetwork as done by Ohlmann et al. [2019]. In this study, for the sake of simplicity, we rely on the trophic groups used for the construction of the simulated networks. Thus, we do not infer those groups from the data but rather use the already known group structure. Groups2Vec simply builds its embedding matrix by computing the vector of group proportions (species richness in each group divided by total richness) for each network.

Metrics2Vec. Several metrics have long been used to characterize variations of trophic networks in space (see e.g. Braga et al. [2019], Kortsch et al. [2019], Thompson and Townsend [2005], Wood et al. [2015]) or in time [Albouy et al., 2014]. We selected a total of 17 classical metrics (detailed in Table 1 of the Supporting Information) including the average trophic level [Williams and Martinez, 2004], the average generalism, the frequency of omnivore species (defined as species that predate other species across more than one trophic level), proportions of top and basal species, modularity [Newman, 2006] and trophic length. Note that for our following analyses, we centered and scaled each metric (i.e. each column of the embedding matrix) so that metrics had equivalent contributions in analyses based on Euclidean distances.

Motifs2Vec. This embedding method gathers the frequencies of the 13 directed connected triangular motifs (see Figure 3 in Supporting Information), without self-loop (i.e. their count divided by the number of possible species triplets which is $n(n-1)(n-2)/6$ for a network of size n). These small motifs have been regularly used to characterize local architectures in trophic networks [Camacho et al., 2007]. This normalization allows to correct for the effect of network size on motifs counts when comparing different networks. Larger motifs would give more precise representation of the whole network structure, but the computational complexity of motifs count, $O(n^k)$ for k-nodes motifs, is prohibitive [Shervashidze et al., 2009].

Graph2Vec. Graph2Vec [Narayanan et al., 2017] is an approach based on the description of the local neighborhood of each node. Practically, the algorithm decomposes each network into trees rooted at each of its nodes. It takes as input a maximal depth, which corresponds to the distance up to which the neighbors of each node will be explored. Then, the description of the local neighborhood of each node is used to generate an embedding matrix, whose dimensionality is chosen a priori by the user and where networks with similar node neighborhoods tend to be close (more details in Section A.3 of Supporting Information). Graph2Vec has the possibility to account (or not) for prior information on node labels. In the case of no prior information, it takes the node degree as a label. In this study, we compared the embedding method (hereafter called Graph2Vec) with no prior information on node labels, with Graph2Vec_lab that directly uses the species trophic group as node label (the same than in Groups2Vec). Since Graph2Vec was originally designed for undirected graphs, we forced the method to take into account edges direction when building a node neighborhood. By default, it explores the network from prey to predators. We also tested to concatenate this default embedding matrix with the one derived from the network transposed adjacency matrix (exploration from predators to prey) to better represent the directed architecture of the trophic network. We compared the concatenated version to the default embedding matrix and also the effect of the maximal

depth choice (Section A.4 from Supporting Information). In the following application, we used the concatenated version of Graph2Vec (with an embedding matrix dimension of 30 for both the default and transposed version) with a maximal depth of 2.

ShortPaths2Vec. This embedding method gathers the frequencies of the directed shortest paths lengths. For a network, we first computed the length of the shortest-path between all pairs of nodes, following edges direction. When there is no path between two nodes, it is counted as an infinite length. Then, the frequency vector of this set of lengths is constructed (for each length, its occurrence count is divided by $n(n-1)$, the total number of ordered pairs of distinct nodes for a size-n graph). When considering several networks, the columns of this embedding matrix correspond to the set of all lengths observed across the different networks. This embedding method gathers information related to the notion of trophic length, sometimes computed as the length of the longest directed shortest-path. The ratio of shortest-paths with infinite length may also encode information on compartmentalization in the network. As for Graph2Vec, our proposed ShortPaths2Vec approach is generalized to account for prior information on node label in a second version called ShortPaths2Vec_lab. In ShortPaths2Vec_lab, the counts of shortest-paths of any given length is decomposed per combination of source and target node labels. In other words, any column of the embedding matrix corresponds to the count of shortests-paths of a given length k starting from a node with a given label l and going to a node with a given label l' divided by $n(n-1)$, as for ShortPaths2Vec. We again use trophic groups as node labels (the same than in Groups2Vec).

2.2 From embedding space to visualization space using UMAP dimension reduction

Once an embedding matrix is obtained from a set of networks, dimension reduction techniques can be used to visualize the networks in a lower dimensional vector space (here 2 dimensions) called the visualization space. Indeed, our embedding matrices have dimensionality ranging from 13 (Motifs2vec) up to 60 (Graph2Vec in both directions) and above (dimension of ShortPaths2Vec/ShortPaths2Vec_lab depends on the networks shortest paths). We thus used a non-linear dimension reduction technique called Uniform Manifold Approximate Projection (UMAP, McInnes et al. [2018]) that was recently popularized for the analysis of biological data [Becht et al., 2019]. UMAP relies on a user-defined distance metric (e.g. here the Euclidean distance) computed between all pairs of networks using their embedding coordinates. For any embedded network, UMAP considers its closest neighbors (we selected 150 neighbors here), and aims to find a projection into the visualization space such that, in this new space, these neighboring networks are also close to the targeted one. This property has to be satisfied for every embedded network, and a mathematical criterion is numerically optimized to achieve this goal.

2.3 Simulation experiment

To compare these seven embedding methods, we carried out a simulation study to generate 5000 trophic interaction networks (i.e. food webs). Each network was drawn from a parametric random network model (a SBM, see Allesina et al. [2008]), hereafter called a group model, where species are divided into trophic groups and where the probabilities of interaction between species depend on their belonging to trophic groups. A group model is defined by its trophic groups (here, between 2 and 10 groups), the interaction probabilities between groups and the distribution of nodes (species) in groups. For each simulated network, we first randomly draw the network size and six parameters, hereafter called ecological properties (Table 2), determining the group model structure. As illustrated

in Figure 2, our group model splits species in trophic groups which organize the network vertically in trophic levels and horizontally in one or two modules, in which case there are less interactions between modules than inside modules. As ecological properties varied across the networks, our procedure allowed generating networks with contrasted ecological characteristics. More details on the construction of the group models are provided in Section B from Supporting Information. We defined the number of species (network size) in $\{60, 120\}$ and independently of the other properties. We affected one species per group (to avoid emptiness) and then randomly distributed the remaining species among the groups. Lastly, we randomly drew the presence of a directed interaction between each ordered pair of species through the probability of interaction of their groups as defined by the group model. Figure 2 illustrates the variety of our simulated networks and the effects of the variations in ecological properties through some examples.

Property acronym	Effect	Possible values (categories)
nModules	Number of modules (compartments)	$\{1, 2\}$
trophLens	Trophic length in each module	see Section B from Supp. Info.
maxTrophLen	Maximal trophic length across modules	$\{2, 3, 4, 5\}$
omni	Activates interactions between non-successive trophic levels in a module	$\{TRUE, FALSE\}$
generalism	Favors interactions between successive trophic levels in a module	$\{TRUE, FALSE\}$
loop	Allows intra-group interactions and thus favors loops	$\{TRUE, FALSE\}$

Table 2: The six ecological properties controlled in our networks simulation and their categories.

We then applied the seven embedding methods to the 5000 simulated networks. Practically, the 10 trophic groups resulting from the largest possible group model (see details in Section B from Supporting Information) were used as the groups for Groups2Vec and also as the node labels in Graph2Vec_lab and ShortPaths2Vec_lab.

2.4 Quality assessment of the embedding methods

We evaluated the ability of the seven embedding methods to disentangle the network ecological properties (**maxTrophLen**, **trophlens**, **nModules**, **generalism**, **omni** and **loop**) of the simulated networks for posterior supervised and unsupervised analysis tasks with several criteria (detailed in Section D from Supporting Information). In all that follows, we call categories the possible values taken by an ecological property as shown in the right column of Table 2, e.g. 1 and 2 are the categories of the property **nModules**. For posterior supervised learning tasks, we used the predictive accuracy as a measure of quality to evaluate how well each property is predicted from network embedding coordinates. For a given property, we used the classification accuracy of a Random Forest [Breiman, 2001] trained to predict the category of a network from its coordinates in the embedding (e.g. predict the category of **nModules** using the 13 motifs proportions provided by Motifs2Vec). For unsupervised analyses, we aimed at evaluating how well the segregation between categories can be detected by human eyes in the visualization space. We assumed that if each category forms a cluster of networks well separated from other categories in the embedding space, the user has more chances to detect a structure with a clustering method or visually on the visualization space. For that purpose, three criteria were used. First, we measured how well networks of the same category aggregate into clusters, i.e. how well distinct categories were segregated, in (i) the embedding space using a metric called R2-ebd and (ii) in the 2D visualization space obtained from dimension reduction of the embedding matrix using UMAP using a metric called R2-umap (see Sections D.2 and D.3

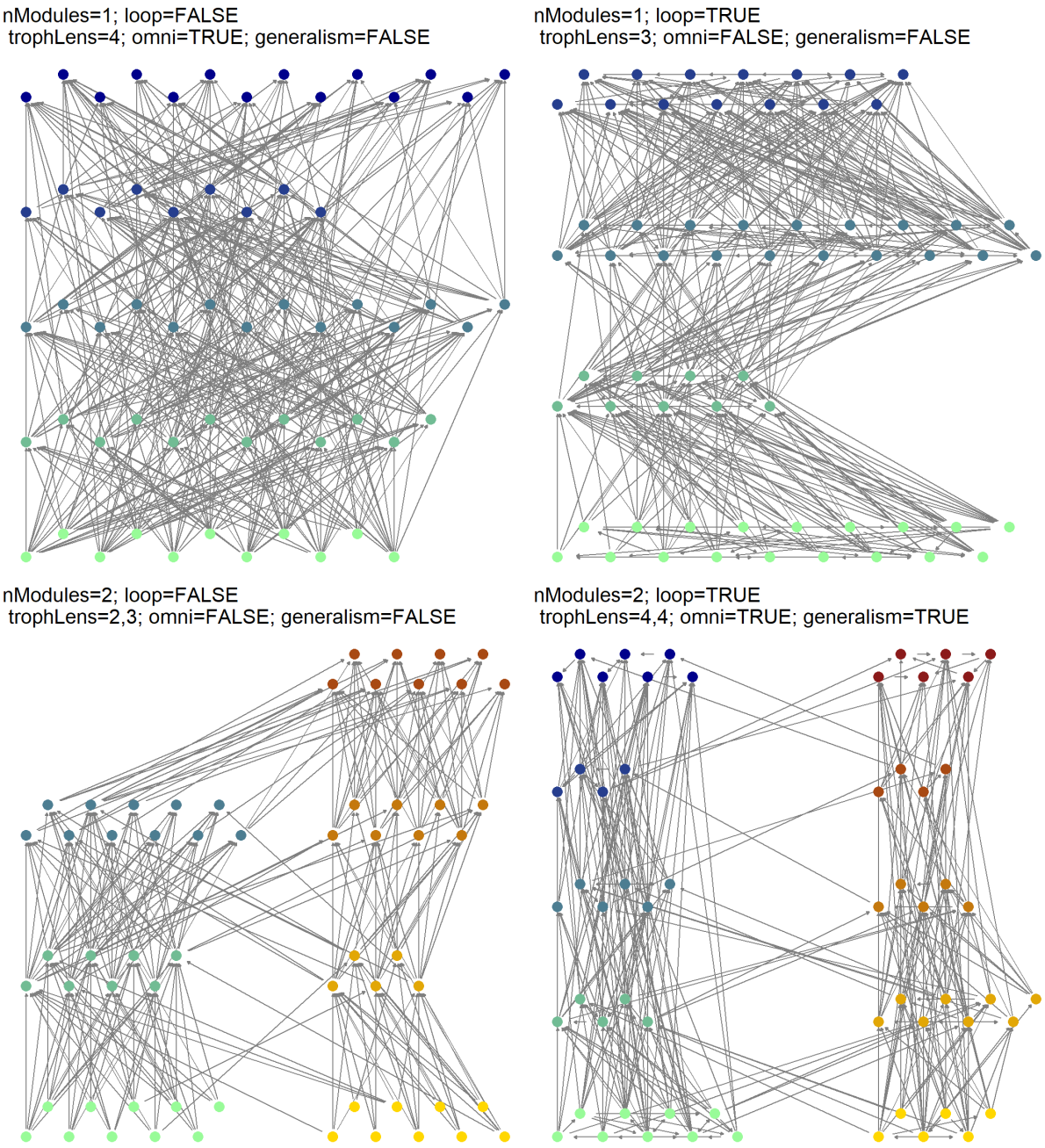

Figure 2: Figure 1. Examples of simulated trophic networks of size 60 and their group model parameters (ecological properties). These networks span the 4 possible combinations of values for **nModules** (compartmentalization) and **loop** (intra-group predation). Nodes are colored according to their trophic group.

from Supporting Information). The third criterion evaluated how the segregation of categories (in the embedding space and in the visualization space, respectively) is blurred due to variation in network sizes (see Section D.4 from Supporting Information for the definitions of the two metrics R2-loss-ebd and R2-loss-umap). Note that, given the simulation design, Groups2Vec, Graph2Vec_lab and ShortPaths2Vec_lab were unfairly favored for all criteria on properties **maxTrophLen**, **trophlens** and **nModules** compared to other embedding methods because they directly encoded trophic groups composition that generated the network. Groups2Vec was also disfavored compared to other methods for the **omni**, **generalism** and **loop** properties because, by our simulation design the trophic

groups proportions were not affected by changes in these three properties.

3 Results

Evaluation for supervised learning task: predictive accuracy. The relative ability of the different embedding methods to predict the categories varied across ecological properties (Table 3). Amongst the different methods, Metrics2Vec had the highest predictive accuracy for **maxTrophLen**, **trophlens** and **nModules** (excluding methods that integrate a priori information on trophic groups), closely followed by ShortPaths2Vec. For **omni**, the methods with the highest predictive accuracy were Graph2Vec_lab, ShortPaths2Vec_lab and Motifs2Vec. Metrics2Vec and ShortPaths2Vec had the best predictive accuracy for **generalism**, and the latter was even more efficient when using trophic groups as node labels (ShortPaths2Vec_lab, Table 3). **loop** was almost perfectly predicted by all embedding methods. Unlike its extension with node labels, we found that Graph2Vec had a relatively weak predictive accuracy for all properties except **loop** (Table 3, also visible on Figure 3D). This relatively low predictive accuracy persisted when applying the algorithm with more iterations (higher order neighborhood exploration, see Section A.4 from Supporting Information).

Evaluation for unsupervised learning task: segregation of categories. Motifs2Vec had the best segregation of categories in the embedding space (highest R2-ebd) for **maxTrophLen**, **trophlens** and **nModules** (Table 4). However, this relative performance was not consistently preserved in the UMAP plane. Indeed, its R2-umap was significantly inferior to the one of Metrics2Vec for **nModules** (Table 4, and see Figure 3B/F). Even though the categories of **maxTrophLen** and **nModules** were well separated in the UMAP plans of Metrics2Vec, ShortPaths2Vec and Motifs2Vec, this separation was always non-linear and would not allow us to distinguish visually the categories without knowing them a priori (see respectively Figures 7 and 8 from Supporting Information). Specifically, Motifs2Vec had a higher R2-ebd than Metrics2Vec (Table 4) whereas it had a lower predictive accuracy for the same three properties (Table 3). These apparent conflicting results showed that the most suited approach clearly depends on the research questions and the task scenario (unsupervised *vs* supervised analysis). Surprisingly, categories of **maxTrophLen** seemed much less segregated in the embedding space than in the visualization space for ShortPaths2Vec (R2-ebd $<$ R2-umap, Table 4). This shows that even though a small portion of this embedding matrix variability was related to **maxTrophLen**, it was well retained after UMAP compression (see Figure 3D). Properties **omni** and **generalism** had an overall much weaker effect on the embedding matrix structures than **nModules**, **loop** and **n** (R2-ebd and R2-umap close to 0, see Table 4), even if they had the same number of categories. Here, the heterogeneity of results between methods was not meaningful enough to recommend one method over another. This result might arise from our simulation model and might not be generalised to other datasets. For **loop**, the segregation of categories was heterogeneous across methods both in the embedding spaces and visualization spaces. ShortPaths2Vec showed the highest performance in the embedding space (R2-ebd, Table 4), but not in the visualization space where Graph2Vec and Graph2Vec_lab had the highest R2-umap. Visually, all embeddings except ShortPaths2Vec_lab exhibited a clean segregation of categories for **loop** in their visualization space, as shown in Figure 3D/E and Figure 3B/C/E. Note, however, that it is certain that ShortPaths2Vec_lab separates linearly **loop** in its embedding space (non-infinite shortest paths lengths exist between species of a same trophic group when **loop** is activated), but its R2-ebd does not reflect it and remains relatively small probably because this embedding space contains mostly dimensions unrelated to **loop**.

Robustness to size variability. Groups2Vec and Motif2Vec were almost insensitive to network sizes variability and were thus the most robust methods both in the embedding and visualization spaces (R2-loss-ebd and R2-loss-umap almost null in Tables 3 and 4 from Supporting Information) for all ecological properties, revealing the efficiency of their normalization for size. For other methods, network size variability decreased more or less the segregation of categories in their embedding and visualization spaces for all properties (R2-loss-ebd>0 and R2-loss-umap>0). Indeed, for Metrics2Vec, ShortPaths2Vec and ShortPaths2Vec_lab, the segregation of categories decreased by around 10% in the embedding space and visualization space (i.e. R2-loss-ebd and R2-loss-umap are around 10 in Tables 3 and 4 from Supporting Information) when networks of size 60 and 120 were all considered together compared to the average segregation of categories when considering only one size at a time. This shows that ShortPaths2Vec normalization for size is not fully working. Finally, Graph2Vec and Graph2Vec_lab appeared as the least robust methods as they showed the highest R2-loss-ebd and R2-loss-umap for all properties (resp. Tables 3 and 4 from Supporting Information). For instance, networks with the same size tended to be clustered in the UMAP plane of Graph2Vec (see Figure 3D). This strongly suggested that Graph2Vec and Graph2Vec_lab are not suitable to compare networks with different sizes (when the effect of size was not a feature of interest) in posterior unsupervised analysis tasks. On the other hand, one can observe how the positions of networks of different size were mixed on the visualization planes of Groups2Vec and Motif2Vec (e.g. Figure 6 from Supporting Information for Groups2Ved and Figure 3A for Motifs2Vec).

method	maxTrophLen	trophlens	nModules	omni	generalism	loop
Groups2Vec	100	100	100	52	50	49
Metrics2Vec	**93**	**71**	**99**	89	95	100
Motifs2Vec	85	64	98	90	93	100
Graph2Vec	67	49	89	83	75	100
Graph2Vec_lab	100	99	100	**92**	88	100
ShortPaths2Vec	**93**	69	**99**	82	94	100
ShortPaths2Vec_lab	100	99	100	91	**96**	100

Table 3: Predictive accuracy percentage for each ecological property and embedding method. This is the Out-Of-Bag classification accuracy of a Random Forest trained to predict the category of the ecological property from the coordinates of the network in the embedding. Some coefficients are shaded because their comparison with other embedding methods would be unfair, see our methodology. Bold values correspond to the best performance for each property (per column).

4 Discussion

This study proposed a critical evaluation of the ability of different graph embedding methods to detect ecological property variations between simulated trophic networks. We introduced two methods that have never been used to analyse ecological networks (Graph2Vec and ShortPaths2Vec) and proposed to use trophic groups as node labels to enrich the description of network architecture for these two methods. We evaluated seven embedding methods for posterior use in supervised and unsupervised analysis tasks focusing on six important network ecological properties, and testing their robustness to network sizes variability for the unsupervised setting.

Depending on the type of task, supervised or not, and the ecological property targeted, the relative performances of the embedding methods differed. Overall, Motifs2Vec and ShortPaths2Vec, which have a relatively small dimensionality (here 13 and 15), interpretable dimensions, and are

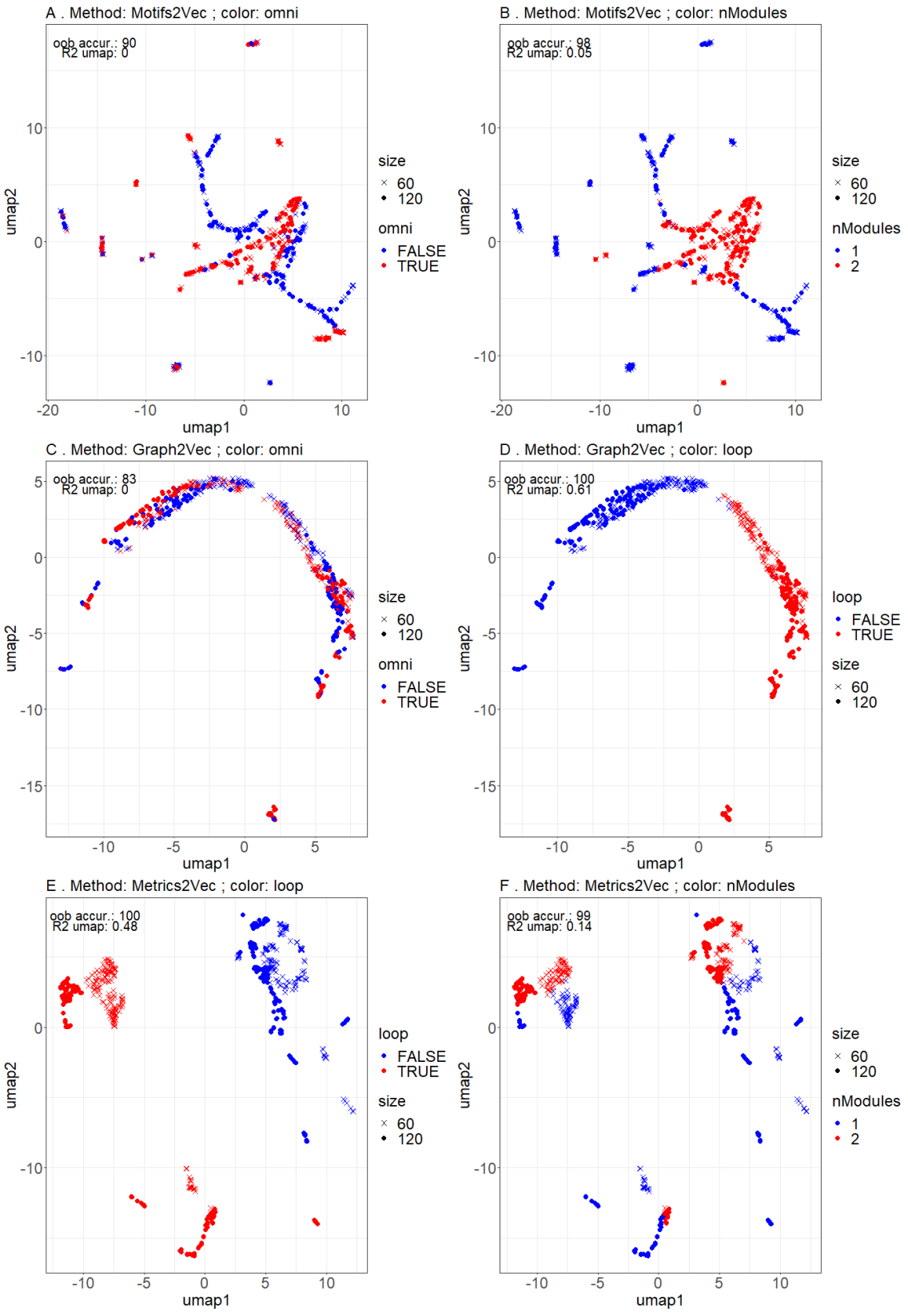

Figure 3: Figure 2. UMAP 2D plans, Part 1. Each plane, i.e. visualization space, is output by UMAP applied to an embedding of the simulated networks. Axes UMAP1 and UMAP2 may be read as the principal axes in multivariate linear analyses, without notion of importance ranking. Each point represents a network and the color (resp. shape) indicates the category of the ecological property at stake (resp. species richness $\mathbf{n} \in \{60, 120\}$). For visualization clarity, we randomly subsampled 600 points out of 5000.

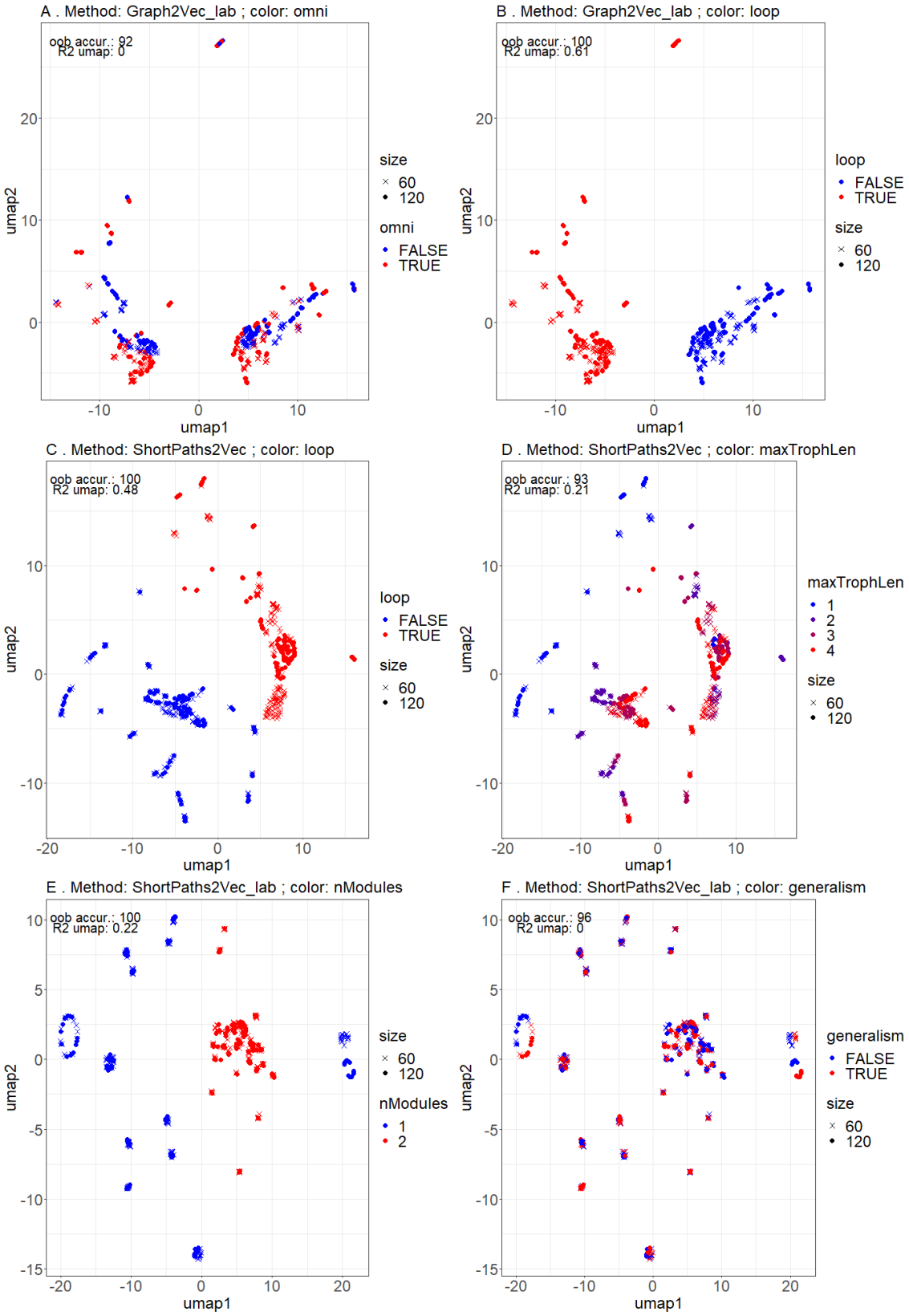

Figure 4: Figure 3. UMAP 2D plans, Part 2.

Method	maxTrophLen		trophlens		nModules		omni		generalism		loop	
	ebd	umap	ebd	umap	ebd	umap	ebd	umap	ebd	umap	ebd	umap
Groups2Vec	0.42	0.60	0.90	0.99	0.38	0.30	0.00	0.00	0.00	0.00	0.00	0.00
Metrics2Vec	0.19	0.10	0.32	0.20	0.20	**0.14**	0.02	0.00	0.02	0.00	0.30	0.48
Motifs2Vec	**0.39**	0.19	**0.62**	**0.25**	**0.32**	0.05	0.01	0.00	**0.07**	**0.01**	0.09	0.25
Graph2Vec	0.06	0.08	0.12	0.13	0.07	0.08	0.00	0.00	0.01	**0.01**	0.09	**0.61**
Graph2Vec_lab	0.10	0.09	0.30	0.13	0.09	0.07	**0.04**	0.00	0.00	0.00	0.08	**0.61**
ShortPaths2Vec	0.03	**0.21**	0.15	0.24	0.13	0.04	0.01	**0.01**	0.01	0.00	**0.61**	0.48
ShortPaths2Vec_lab	0.35	0.04	0.64	0.35	0.22	0.22	0.01	**0.01**	0.00	0.00	0.10	0.15

Table 4: Segregation of categories for each ecological property (column) in each embedding space and 2D visualization space (measured through R2-ebd and R2-umap). This measure reflects the level of clustering of networks of a same category compared to the average distance. Shaded values shouldn't be compared with other embeddings (in the same column) as it would be unfair, see 2.4. Bold values correspond to the best performance for each property (per column).

relatively robust to network sizes variability, often proved to be more suitable for unsupervised analysis. Other embedding methods either poorly captured the properties of interest (e.g. Graph2Vec), or captured them in a non-linear way which reduces their potential interpretability in visualization steps of unsupervised task (e.g. ShortPaths2Vec_lab in Figure 3F). Metrics2Vec, ShortPaths2Vec_lab and Graph2Vec_lab, revealed their potential for posterior supervised learning tasks (Table 3). Moreover, predictive accuracy increased when architecturally relevant information (here trophic groups) was integrated as node labels in the embedding methods (ShortPaths2Vec_lab and Graph2Vec_lab, Table 3). In the following paragraphs, we suggest the most appropriate approach for a given targeted ecological property, and further motivate the study of a given property in the light of the ecological processes of interest (Table 5).

Property	Supervised	Unsupervised	Ecological processes
maxTrophLen	Metrics2Vec	Motifs2Vec	CP, CFB
TrophLens	Metrics2Vec	Motifs2Vec	RE, CP, CFB
nModules	Metrics2Vec	Motifs2Vec	RE, PC
omni	Graph2Vec_lab / ShortPaths2Vec_lab / Motifs2Vec		CP, PC
generalism	ShortPaths2Vec(_lab)		RE, CP, PC
loop		ShortPaths2Vec / Metrics2Vec	

Table 5: Guide table summarizing which embedding method to use according to the network ecological property that is targeted, the supervised or unsupervised nature of the analysis. Ecological processes demonstrably linked to each property are listed in the last column. We used the following acronyms, CP: Community persistence, RE: Robustness to extinctions, CFB: Carbon Flux balance, PC: Population control,

Motifs2Vec and Metrics2Vec proved to be the most suitable, respectively for unsupervised and supervised tasks, for maximum trophic length, trophic groups composition and compartmentalization. As relationships between compartmentalization and robustness to extinction are mainly theoretical [Dunne et al., 2002, Thébault and Fontaine, 2010], supervised methods could be applied to predict the robustness to extinctions from an embedding coordinates that contains compartmentaliza-

tion descriptors (e.g. modularity, clustering coefficient, nestedness). As noted earlier, trophic levels were suggested to impact resilience to perturbations and carbon fluxes balance [Pimm et al., 1991, Schindler et al., 1997]. To test hypotheses on the effect of these environmental factors, we could apply ShortPaths2Vec combined with dimension reduction to a large set of trophic networks spatially distributed in various conditions of perturbation frequency and primary productivity in order to visualize potential patterns of association with trophic chain lengths.

Concerning generalism, ShortPaths2Vec_lab and Metrics2Vec were the most suitable methods for supervised analysis while there was no efficient approach for unsupervised analyses (but Motifs2Vec performed best in R2-ebd, Table 4). This suggests that the effect of generalism appeared as a minor driver of the embedding matrices and one might probably specify another embedding method describing the joint distribution of in/out degrees across species to better capture variations of generalism. It might enable to better understand the still ambiguous relationship between generalism, vulnerability and robustness to extinctions [Dunne et al., 2002, Thébault and Fontaine, 2010]. For instance, we hypothesise that a stronger negative relationship between species generalism and vulnerability within a network, for a fixed connectance, makes it more robust to extinctions. Further, the change of these relationships in trophic networks along long time scales have the potential to reveal signatures of community level selection related to mass extinction events [Roopnarine et al., 2007]. For omnivory, Graph2Vec_lab appears as the most adapted method for supervised learning, while no approach yielded significant performances for posterior unsupervised analyses. We might improve unsupervised analysis applications by combining some triangular motifs proportions related to omnivory and some dedicated metrics (e.g. omnivore proportion, average degree of omnivory) in order to unveil contrasts in omnivory patterns across networks that are known to influence parasitism rate [Montoya et al., 2003] and community persistence [Pimm et al., 1991]. Finally, the predictive accuracy for the loop property was almost perfect for all methods, so that we cannot discriminate between them for a supervised learning application. For the unsupervised case, ShortPaths2Vec turned out to be the most suitable. Lastly, it might be important to note that the tested food web Metrics and Graph2Vec require less computational effort than the Shortest-Paths embeddings. Indeed, the complexity of the former is linear with the number of nodes, while it is cubic for the latter using the Floyd-Warshall algorithm [Floyd, 1962].

Most graph embedding developments concerned supervised learning problems and especially graph classification [Li et al., 2017, Xu et al., 2018]. Even the unsupervised graph embedding methods [Ivanov and Burnaev, 2018, Taheri et al., 2018, Verma and Zhang, 2017] have been mostly evaluated on supervised learning tasks. Further, the evaluation is most often done on multiple benchmark datasets coming from different research domains (e.g. bioinformatics or social networks) and weakly related to research questions of these domains, thus questioning the relevance of the evaluation for the end-users of these methods. Here, we took a different perspective since our comparative analysis and evaluation focused on the usefulness of graph embedding methods to address explicit research questions based on trophic network analyses, especially in the unsupervised context which is probably the most common case in ecology. This perspective comes at the price of some questions regarding the generality of our evaluation methodology and the potential limits of our simulation experiments. Regarding network simulation, omnivory and generalism were especially difficult to segregate for all embedding methods suggesting that our simulation model didn't make their variations salient. Besides, two species belonging to the same trophic group in our simulation model tend to interact with the same other groups and with an equal proportion of species in each of these groups. This is a way to model trophic Eltonian niches [O'Connor et al., 2020], but it induces very similar positions of species belonging to a given group in the network. Regarding the evaluation methodology, the UMAP dimension reduction step has sometimes drastically reduced the segregation of categories of some property in the visualization space compared to the embedding space,

while preserving it for others (e.g. nModules with Motifs2Vec, see Table 4). This behavior might have been different using another dimension reduction technique.

Some lessons have been learnt and some questions have been raised on different embedding methods during this study. First, Groups2Vec did not allow us to segregate omnivory, generalism and loop because these properties were considered independently to generate the simulated trophic groups in our design. However, real studies demonstrated that some trophic groups can highlight differences in omnivory, presence of loops or generalism. For example, Kéfi et al. [2016] fitted a stochastic block model (SBM) on a large interaction network including trophic interactions and showed that some identified groups included more omnivore species than others. Then, networks having higher species richness for these groups would have a higher degree of omnivory. We therefore point out that the relevance of Groups2Vec entirely depends on the ecological properties captured by the trophic groups so that results for this approach are very context-dependent. We also showed that Graph2Vec was strongly affected by network size and thus not suitable for unsupervised analyses. Indeed, the distribution of the local neighborhoods that are present in a network is affected by network size. For instance, deeper node neighborhoods are more likely in larger networks. However, to our knowledge, no suitable size normalization procedure is available for this embedding method and its use in unsupervised analyses should thus be restricted to the comparison of networks with similar size. Even though ShortPaths2Vec was more robust to size variability, finding the right normalisation for this embedding method is not simple either. The normalisation we used revealed to not be fully efficient. This is probably because the frequencies of shortest-paths lengths are affected by network size, and not only their average count. Another important lesson regarding Graph2Vec concerns its parameterization. It is crucial to adapt the number of iterations (depth) to the average size of networks because the diversity of subtrees increases exponentially with the number of iterations. Then, applying a large depth to small graphs will tend to increase similarities between networks and thus induce approximately equally spaced points in the embedding space. In our experiment, we standardized each column of the Metrics2Vec embedding matrix so that they have equal contribution to the Euclidean distance computed in the embedding space. This impacts the arrangement of networks in the UMAP plane and the measure of segregation of categories. Without standardization, we would have favored properties discriminated by the columns with highest variance. Interestingly, we can use standardization to control the relative importance of the metrics by multiplying each column by a specific coefficient depending on the targeted network ecological contrasts that we want to distinguish in the visualization space. Recently, Graph Neural Networks (GNNs, see e.g. Gilmer et al. [2017], Kipf and Welling [2016]) have been proposed to produce more expressive and flexible graph embedding methods. Most GNNs are designed for supervised learning tasks, such as graphs classification, rather than unsupervised graph embedding methods (but see e.g. Bandyopadhyay et al. [2020]). Supervised graph embedding with GNNs may be a way forward to find more general representations of interaction networks, as a GNN embedding may be trained, for instance, to predict several network ecological features such as dynamical behaviors or robustness to extinctions.

5 Acknowledgments

This research was funded by the French Agence Nationale de la Recherche (ANR) through the EcoNet (ANR-18-CE02-0010) project. WT also acknowledge support from MIAI@Grenoble Alpes (ANR-19-P3IA-0003). CB thanks Marc Ohlmann for its advice.

6 Author's contributions

Everyone conceived the ideas and designed methodology. CB, VM, CM and SD carried out the formal analyses. Everyone carried the investigation. CB developed the software, produced visualizations and wrote the original draft. Everyone validated and reviewed. VM, CM, WT and SD supervised.

7 Code availability

All the R and Python code necessary to reproduce our results is provided at `https://github.com/ChrisBotella/TrophicNetEncoder`. The graph embedding methods except Graph2Vec are available as part of the R package `econetwork` available at CRAN (The Comprehensive R Archive Network).

References

Albouy, C., Velez, L., Coll, M., Colloca, F., Le Loc'h, F., Mouillot, D., and Gravel, D. (2014). From projected species distribution to food-web structure under climate change. *Global change biology*, 20(3):730–741.

Allesina, S., Alonso, D., and Pascual, M. (2008). A general model for food web structure. *science*, 320(5876):658–661.

Bai, Y., Ding, H., Qiao, Y., Marinovic, A., Gu, K., Chen, T., Sun, Y., and Wang, W. (2019). Unsupervised inductive graph-level representation learning via graph-graph proximity. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 1988–1994. International Joint Conferences on Artificial Intelligence Organization.

Bandyopadhyay, S., Aggarwal, M., and Murty, M. N. (2020). Unsupervised graph representation by periphery and hierarchical information maximization. *arXiv preprint arXiv:2006.04696*.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., Ginhoux, F., and Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology*, 37(1):38–44.

Borgwardt, K. M. and Kriegel, H.-P. (2005). Shortest-path kernels on graphs. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 8–pp.

Borrelli, J. J. (2015). Selection against instability: stable subgraphs are most frequent in empirical food webs. *Oikos*, 124(12):1583–1588.

Braga, J., Pollock, L. J., Barros, C., Galiana, N., Montoya, J. M., Gravel, D., Maiorano, L., Montemaggiori, A., Ficetola, G. F., Dray, S., et al. (2019). Spatial analyses of multi-trophic terrestrial vertebrate assemblages in Europe. *Global Ecology and Biogeography*, 28(11):1636–1648.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Camacho, J., Stouffer, D. B., and Amaral, L. A. N. (2007). Quantitative analysis of the local structure of food webs. *Journal of theoretical biology*, 246(2):260–268.

Cirtwill, A. R., Dalla Riva, G. V., Gaiarsa, M. P., Bimler, M. D., Cagua, E. F., Coux, C., and Dehling, D. M. (2018). A review of species role concepts in food webs. *Food Webs*, 16:e00093.

Costa, A., González, A. M. M., Guizien, K., Doglioli, A. M., Gómez, J. M., Petrenko, A. A., and Allesina, S. (2019). Ecological networks: Pursuing the shortest path, however narrow and crooked. *Scientific reports*, 9(1):1–13.

Dunne, J. A., Williams, R. J., and Martinez, N. D. (2002). Network structure and biodiversity loss in food webs: robustness increases with connectance. *Ecology letters*, 5(4):558–567.

Floyd, R. W. (1962). Algorithm 97: shortest path. *Communications of the ACM*, 5(6):345.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1263–1272.

Ivanov, S. and Burnaev, E. (2018). Anonymous walk embeddings. In *International conference on machine learning*, pages 2186–2195.

Kéfi, S., Miele, V., Wieters, E. A., Navarrete, S. A., and Berlow, E. L. (2016). How structured is the entangled bank? The surprisingly simple organization of multiplex ecological networks leads to increased persistence and resilience. *PLoS biology*, 14(8):e1002527.

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *5th International Conference on Learning Representations (ICLR-17)*.

Kortsch, S., Primicerio, R., Aschan, M., Lind, S., Dolgov, A. V., and Planque, B. (2019). Food-web structure varies along environmental gradients in a high-latitude marine ecosystem. *Ecography*, 42(2):295–308.

Lau, M. K., Borrett, S. R., Baiser, B., Gotelli, N. J., and Ellison, A. M. (2017). Ecological network metrics: opportunities for synthesis. *Ecosphere*, 8(8):e01900.

Li, C., Ma, J., Guo, X., and Mei, Q. (2017). Deepcas: An end-to-end predictor of information cascades. In *Proceedings of the 26th international conference on World Wide Web*, pages 577–586.

McCann, K., Hastings, A., and Huxel, G. R. (1998). Weak trophic interactions and the balance of nature. *Nature*, 395(6704):794–798.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*.

Monteiro, A. B. and Faria, L. D. B. (2016). The interplay between population stability and food-web topology predicts the occurrence of motifs in complex food-webs. *Journal of Theoretical Biology*, 409:165–171.

Montoya, J. M., Rodríguez, M. A., and Hawkins, B. A. (2003). Food web complexity and higher-level ecosystem services. *Ecology letters*, 6(7):587–593.

Münkemüller, T., Gallien, L., Pollock, L. J., Barros, C., Carboni, M., Chalmandrier, L., Mazel, F., Mokany, K., Roquet, C., Smyčka, J., et al. (2020). Dos and don'ts when inferring assembly rules from diversity patterns. *Global Ecology and Biogeography*.

Narayanan, A., Chandramohan, M., Venkatesan, R., Chen, L., Liu, Y., and Jaiswal, S. (2017). graph2vec: Learning distributed representations of graphs. In *Proceedings of the 13th International Workshop on Mining and Learning with Graphs (MLG)*.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582.

Ohlmann, M., Miele, V., Dray, S., Chalmandrier, L., O'connor, L., and Thuiller, W. (2019). Diversity indices for ecological networks: a unifying framework using Hill numbers. *Ecology letters*, 22(4):737–747.

O'Connor, L. M., Pollock, L. J., Braga, J., Ficetola, G. F., Maiorano, L., Martinez-Almoyna, C., Montemaggiori, A., Ohlmann, M., and 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.

Pellissier, L., Albouy, C., Bascompte, J., Farwig, N., Graham, C., Loreau, M., Maglianesi, M. A., Melián, C. J., Pitteloud, C., Roslin, T., et al. (2018). Comparing species interaction networks along environmental gradients. *Biological Reviews*, 93(2):785–800.

Pimm, S. L., Lawton, J. H., and Cohen, J. E. (1991). Food web patterns and their consequences. *Nature*, 350(6320):669–674.

Poisot, T., Canard, E., Mouillot, D., Mouquet, N., and Gravel, D. (2012). The dissimilarity of species interaction networks. *Ecology letters*, 15(12):1353–1361.

Roopnarine, P. D., Angielczyk, K. D., Wang, S. C., and Hertog, R. (2007). Trophic network models explain instability of early triassic terrestrial communities. *Proceedings of the Royal Society B: Biological Sciences*, 274(1622):2077–2086.

Schindler, D. E., Carpenter, S. R., Cole, J. J., Kitchell, J. F., and Pace, M. L. (1997). Influence of food web structure on carbon exchange between lakes and the atmosphere. *Science*, 277(5323):248–251.

Shervashidze, N., Vishwanathan, S., Petri, T., Mehlhorn, K., and Borgwardt, K. (2009). Efficient graphlet kernels for large graph comparison. In *Artificial Intelligence and Statistics*, pages 488–495.

Simmons, B. I., Cirtwill, A. R., Baker, N. J., Wauchope, H. S., Dicks, L. V., Stouffer, D. B., and Sutherland, W. J. (2019). Motifs in bipartite ecological networks: uncovering indirect interactions. *Oikos*, 128(2):154–170.

Snyder, W. E., Snyder, G. B., Finke, D. L., and Straub, C. S. (2006). Predator biodiversity strengthens herbivore suppression. *Ecology letters*, 9(7):789–796.

Taheri, A., Gimpel, K., and Berger-Wolf, T. (2018). Learning graph representations with recurrent neural network autoencoders. In *Proceedings of the KDD Deep Learning Day*.

Thébault, E. and Fontaine, C. (2010). Stability of ecological communities and the architecture of mutualistic and trophic networks. *Science*, 329(5993):853–856.

Thompson, R. M. and Townsend, C. R. (2005). Food-web topology varies with spatial scale in a patchy environment. *Ecology*, 86(7):1916–1925.

Torres-Alruiz, M. D. and Rodríguez, D. J. (2013). A topo-dynamical perspective to evaluate indirect interactions in trophic webs: New indexes. *Ecological modelling*, 250:363–369.

Tylianakis, J. M., Laliberté, E., Nielsen, A., and Bascompte, J. (2010). Conservation of species interaction networks. *Biological conservation*, 143(10):2270–2279.

Verma, S. and Zhang, Z.-L. (2017). Hunt for the unique, stable, sparse and fast feature learning on graphs. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30.

Williams, R. J. and Martinez, N. D. (2004). Limits to trophic levels and omnivory in complex food webs: theory and data. *The American Naturalist*, 163(3):458–468.

Wood, S. A., Russell, R., Hanson, D., Williams, R. J., and Dunne, J. A. (2015). Effects of spatial scale of sampling on food web structure. *Ecology and evolution*, 5(17):3769–3782.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2018). How powerful are graph neural networks? In *International Conference on Learning Representations*.

Supporting Information for: An appraisal of graph embeddings for comparing trophic network architectures

Christophe Botella[1,*], Stéphane Dray[2], Catherine Matias[3], Vincent Miele[2], and Wilfried Thuiller[1]

[1]Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LECA, Laboratoire d'Ecologie Alpine, F-38000 Grenoble, France
[2]Université de Lyon, F-69000 Lyon; Université Lyon 1; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Évolutive, F-69622 Villeurbanne, France
[3]Sorbonne Université, Université de Paris, Centre National de la Recherche Scientifique, Laboratoire de Probabilités, Statistique et Modélisation, F-75005 Paris, France
[*]Corresponding author: christophe.botella@gmail.com

A Graph embeddings

A.1 Foodweb metrics in Metrics2Vec

Table 1 describes the 17 network metrics used in Metrics2Vec, relying on the following notation. Consider a directed network $G = (V, E)$ where V is the set of species (nodes) and E the set of directed interactions where each element has the form (u, v) with $u, v \in V^2$. In Table 1, we call G' the *undirected network* derived from G. In other words $G' = (V, E')$ where $\forall v, u \in V^2, (v, u) \in E' \Leftrightarrow ((v, u) \in E$ or $(u, v) \in E)$.

Name	Description	Used in
Density	Ratio of number of edges on number of possible edges for the undirected network: $2\|E'\|/(\|V\|(\|V\|-1))$	Kortsch et al. [2019] Braga et al. [2019] Thompson and Townsend [2005]
Directed density	Ratio of number of edges on number of node pairs: $\|E\|/\|V\|^2$	Kortsch et al. [2019] Wood et al. [2015] Thompson and Townsend [2005]
Average generality	Average number of prey per predator	Kortsch et al. [2019] Braga et al. [2019]
Generality s.d.	Standard deviation of generality	Kortsch et al. [2019] Braga et al. [2019] Wood et al. [2015]
Vulnerability s.d.	Standard deviation of number of predators per prey	Kortsch et al. [2019] Braga et al. [2019] Wood et al. [2015]
Mean trophic level	Average of species Trophic levels	Braga et al. [2019]
Trophic length	Maximum minus minimum trophic levels	
Proportion of omnivores	Proportion of species who consume species at various trophic levels)	Kortsch et al. [2019] Braga et al. [2019]
Level of omnivory	Standard deviation of prey trophic levels	Kortsch et al. [2019]
Top species ratio	Ratio of species with no predators	Kortsch et al. [2019] Braga et al. [2019] Thompson and Townsend [2005]
Basal species ratio	Ratio of species with no prey	Kortsch et al. [2019] Braga et al. [2019] Thompson and Townsend [2005]
Intermediate species ratio	Ratio of species that are neither basal nor top	Kortsch et al. [2019] Braga et al. [2019] Thompson and Townsend [2005]
Modularity	Measure of compartmentalization see Newman [2006]	Kortsch et al. [2019] Tylianakis et al. [2010]
Transitivity /clustering coefficient	Probability that two species linked to a third are also linked together in the undirected network [Wasserman et al., 1994]	Kortsch et al. [2019] Braga et al. [2019] Wood et al. [2015]
Diameter	Length of the longest shortest-path in the (directed) network	
Mean distance / Characteristic path length / Mean Shortest-Path length	Mean length of the directed shortest-paths between all pairs of species	Kortsch et al. [2019] Braga et al. [2019] Wood et al. [2015] Thompson and Townsend [2005]
Assortativity of degrees	Correlation between out and in degrees of nodes pairs [Newman, 2002]	

Table 1: Common network metrics to describe and compare foodwebs architecture.

Moreover, we plot in Figure 1 (resp. in Figure 2) the values of each pairs of metrics for a subsample of 800 simulated networks (resp. Pearson's correlation coefficient between each pairs of metrics over all simulated networks). Figure 2 exhibits 4 groups where each metric is either highly correlated or highly anti-correlated to all others. The first group shows a high connection between modularity and various metrics related to the distribution of degrees. The second group shows a high correlation between all metrics built from the same computation

of trophic levels [Williams and Martinez, 2004]. The third group is composed of statistics of the distribution of distances (diameter, mean_distance), proportions of basal, intermediate and top species and transitivity, which are all linked to nodes pairs distances in the network. The fourth is composed of assortativity alone, which is decoupled from all other metrics. Note that combining a diversity of redundant metrics in the embedding may be useful, as long as they are not fully correlated, because each metric may compensate biases of others in certain cases leading to an overall better separation of a targeted ecological property.

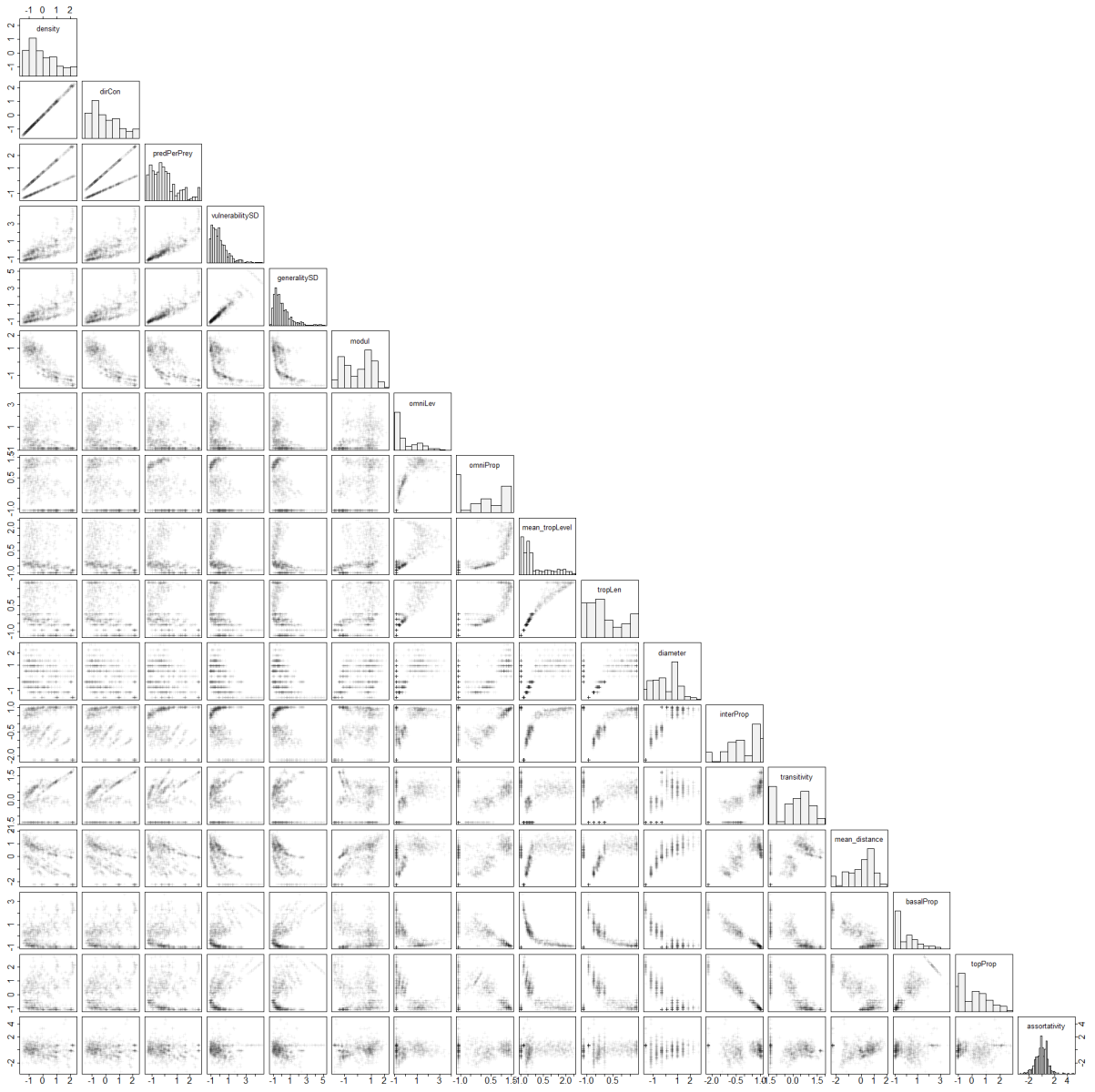

Figure 1: Multi-scatter plots of foodweb metrics used in this study (subsample of 800 simulated networks).

A.2 Triangular motifs in Motifs2Vec

Figure 3 shows the 13 directed triangular motifs whose occurrence proportions are used in Motifs2Vec.

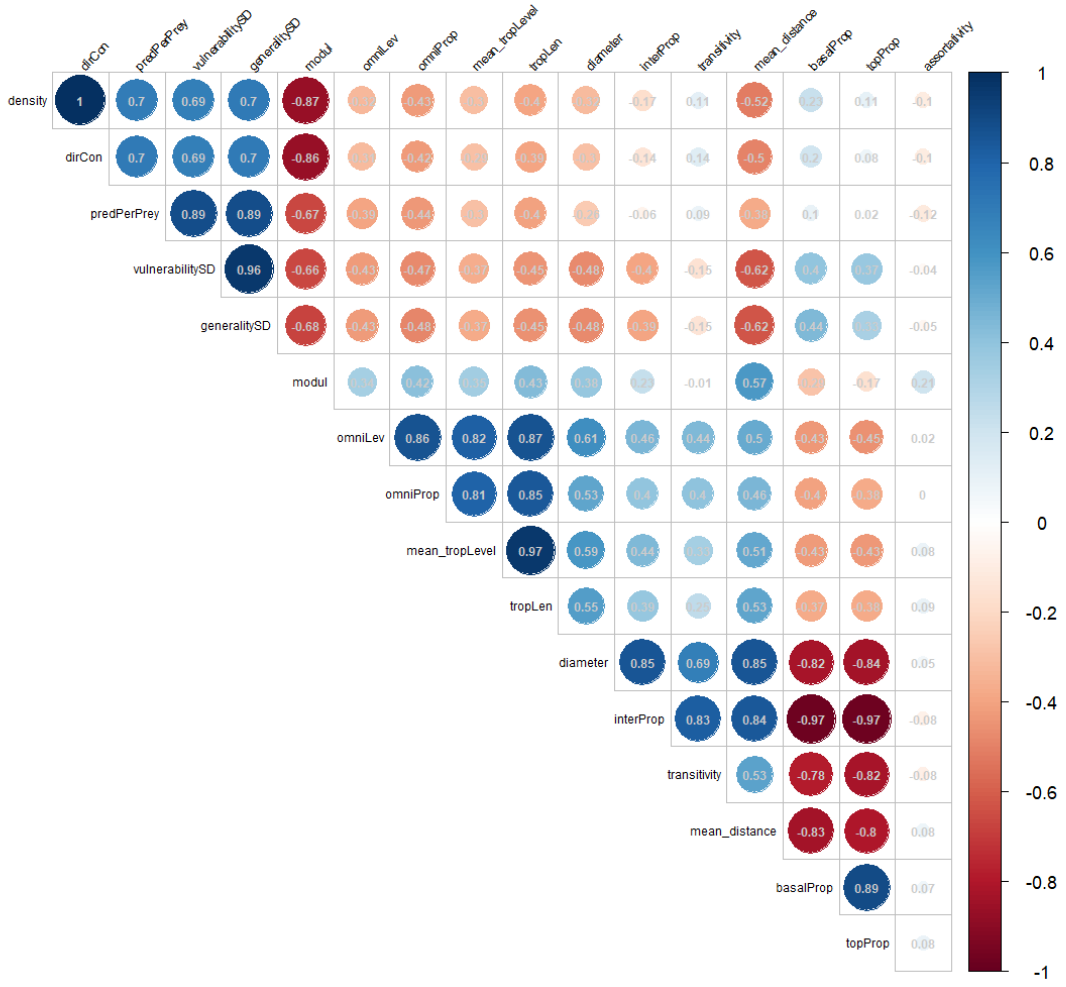

Figure 2: Correlations between foodweb metrics used in this study, computed over all simulated networks.

A.3 Introduction to Graph2Vec

Graph2Vec [Narayanan et al., 2017] is a graph embedding that is based on the decomposition of a network into its *nodes rooted subtrees*. A node rooted subtree represent the local neighborhood around a node and is constructed from an iterative algorithm, as in the Weisfeiler-Lehman graph kernel computation [Shervashidze et al., 2011]. At each iteration, the algorithm concatenates, for each node, its label with those of its neighbors, determining the structure of a (labelled) tree, compresses the character string into a hash-code and relabels the original node with this hash-code. At the end of iteration i, each node has a label (hash-code) that is uniquely associated with a tree of depth i, whose root is the initial node label and the forks of level k have the labels of the neighbors of order k [Shervashidze et al., 2011]. The method then lists for each network all the hash-codes found from iterations 0 up to a user-chosen maximal depth.

Then, Graph2Vec generates an embedding of predetermined dimension where networks having many common hash-codes *and* sharing the absence of many hash-codes (held by other networks from the dataset) are close together. In this second step, the embedding is optimized using the Skipgram model and negative sampling [Mikolov et al., 2013]. It creates at the same time a vector representation for each network of the dataset and for any subtree found across

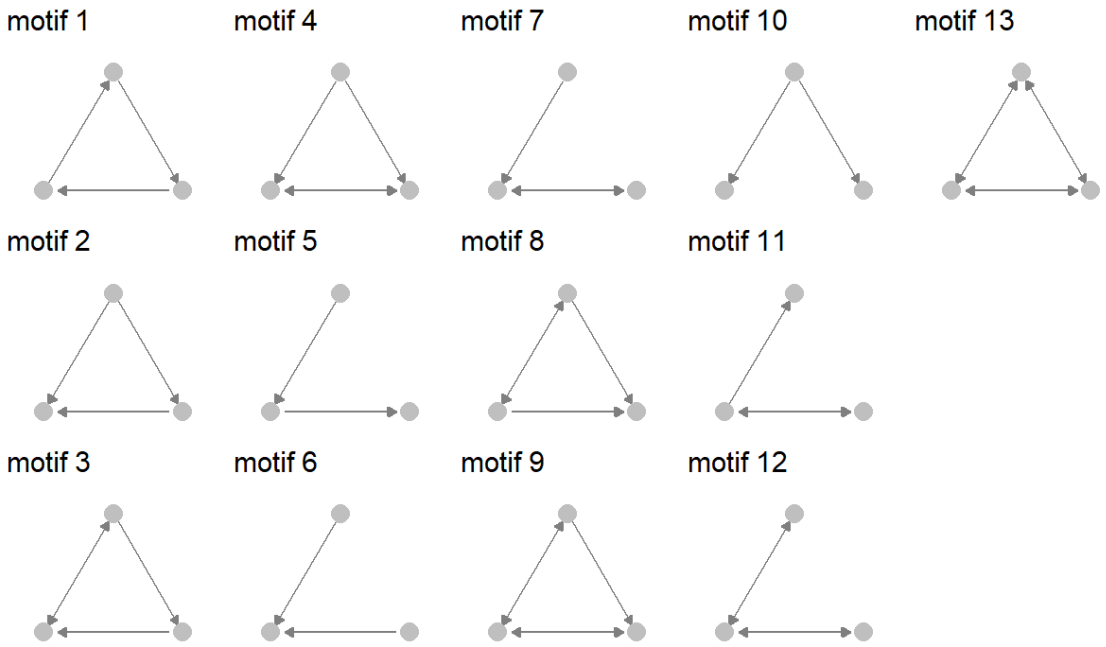

Figure 3: The 13 directed triangular motifs without self-loops used in Motifs2Vec.

all networks. The optimisation of the objective function (Equation (5) in Narayanan et al. [2017]) maximises the proximity between a network vector and the vectors of subtrees that it contains while pushing away the network vector and the vectors of subtrees that it doesn't contain but exist elsewhere in the networks dataset.

A.4 Evaluation and selection of Graph2Vec parameterizations

To simplify the main manuscript, we selected only two versions of Graph2Vec parameterizations. Table 2 reports the predictive accuracy obtained on more diverse parameterizations of Graph2Vec including: various number of iterations (or depth, identified by the prefix "dp" in row names of Table 2), default embedding that relies on the directed network with directions from prey to predators (identified by suffix recto in row names of Table 2) versus its concatenation with the embedding obtained on the transposed networks (no suffix), the integration of trophic groups as node labels (identified by the suffix 'lab' in row names of Table 2) versus no node labels (no suffix).

The predictive accuracy is overall significantly increased when concatenating the recto and verso embeddings compared to the recto embedding alone (Table 2). Also, the predictive accuracy decreased slightly with increasing number of iterations for non-node labelled Graph2Vec versions and for all ecological properties. It is likely that Graph2Vec with higher number of iterations couldn't exhibit consistent patterns of similarities across networks because high order neighborhoods are potentially much more diverse, making each network more dissimilar to the others. No obvious trend appeared among node-labelled Graph2Vec parameterizations, and those were approximately equivalent for all properties. In the main manuscript, we kept only parameterizations with 2 iterations, which were among the best parameterizations in most cases: Graph2Vec_lab_dp2 and Graph2Vec_dp2.

method	maxTrophLen	TrophLens	nModules	omni	generalism	loop	n
Graph2Vec_dp2_recto	50	33	84	67	66	98	83
Graph2Vec_lab_dp2_recto	100	99	100	91	84	100	98
Graph2Vec_dp1_recto	50	32	85	65	66	98	83
Graph2Vec_lab_dp1_recto	99	98	100	91	87	100	99
Graph2Vec_dp4_recto	47	30	83	67	66	98	81
Graph2Vec_lab_dp4_recto	100	99	100	90	83	100	98
Graph2Vec_dp1	70	51	91	83	78	100	93
Graph2Vec_lab_dp1	99	99	100	92	90	100	99
Graph2Vec_dp2	67	50	89	82	76	100	93
Graph2Vec_lab_dp2	100	99	100	92	88	100	99
Graph2Vec_dp4	64	45	88	81	75	99	91
Graph2Vec_lab_dp4	100	99	100	91	86	100	99

Table 2: Predictive accuracy computed for different Graph2Vec configurations and all properties.

B Trophic networks simulation

We describe how we generate, for each network, a group model (a Stochastic Blockmodel, SBM, see Allesina and Pascual [2009]) based on the drawing of six ecological properties, and how we draw a random network from it.

Prior to generate a network, we parameterized its group model. We started by defining the trophic groups of the group model, which are parameterized by three ecological properties. First, we draw n_M uniformly in $\{1, 2\}$ (labelled **nModules**) which is the number of modules in the network. Species that belong to a given module are more likely to interact together than with species from other modules. Second, for each possible module $i \in \{1, n_M\}$, we independently and uniformly draw l_i in $\{1, 2, 3, 4\}$ the trophic length in that module (= number of trophic levels minus one). Each trophic level inside each module represents one trophic group. The list $(l_i)_{i\in[\![1,n_M]\!]}$ is another ecological property, called **TrophLens**, which determines the trophic groups composition. Then, the group model has $n_b = \sum_{i=1}^{n_M}(l_i + 1) \in \{2, ..., 10\}$ trophic groups. **TrophLens** has 20 possible values (one unique module with four possible values for the number of trophic levels, or two modules each one with four possible values). We also introduce **maxTrophLen**, the maximum trophic length across the modules, i.e. **maxTrophLen**=$\max_{i\in[1,n_M]} l_i$. For example, Figure 4 illustrates a potential group model with **nModules**= 2 and **TrophLens**= $(2, 1)$. Trophic groups composing the group model are labelled with the number of the module they belong to and with their trophic level, starting from the basal species and the interactions go from the prey to the predator.

Then, we allowed (directed) interaction probabilities only between specific pairs of groups as follows: the probability P_{pred} of interaction between successive trophic levels in a same module is based on a constant term $P_{base} = 0.7$ and modulated by other variable terms (see Equation (1) below). Three independent ecological properties parameterize the interaction probabilities. Each is drawn as 1 (activated) or 0 (non-activated) with uniform probability. First, $I_o \in \{0, 1\}$ (labelled **omni**) indicates the apparition of omnivory patterns: when activated (I_o=1), the probability of interaction between species from a given trophic group to species from other groups at least two levels higher, inside the same module, was set to $P_{omni} = 0.2$. Secondly, $I_g \in \{0, 1\}$ (labelled **generalism**) increases the predation probability (P_{pred}) by $P_{generalism} = 0.2$. Third, $I_l \in \{0, 1\}$ (labelled **loop**) determines the apparition of an interaction probability $P_{Loop} = 0.15$ between species of a same group (cannibalism was excluded). We also allowed inter-module interactions and only between successive trophic

levels, with a probability $P_{inter} = 0.1$. We apply a corrective term to P_{pred} to compensate the effects of **omni** and **loop** on the expected connectance of the network simulated from the group model. The derivation of this corrective term is shown in the next paragraph and leads to

$$P_{pred} := P_{base} + I_g P_{generalism} - I_o P_{omni} \frac{\sum_{i=1}^{n_M} 1_{l_i>1} \binom{l_i}{2}}{n_b - n_M} - I_l P_{loop} \frac{n(n-n_b)n_b^2}{n^2(n_b - n_M)}. \quad (1)$$

The network size (species richness) was the last parameter of our group model: over the 5000 generated group models, half had $n = 60$ species while the rest had $n = 120$ species to introduce size variability across networks.

Once a group model was parameterized, we draw a trophic network from it. We associated one species to each group and then randomly distributed the remaining species uniformly across the groups. Then, for each pair of species, we independently draw the realisation of an interaction as defined by the group model given their respective groups.

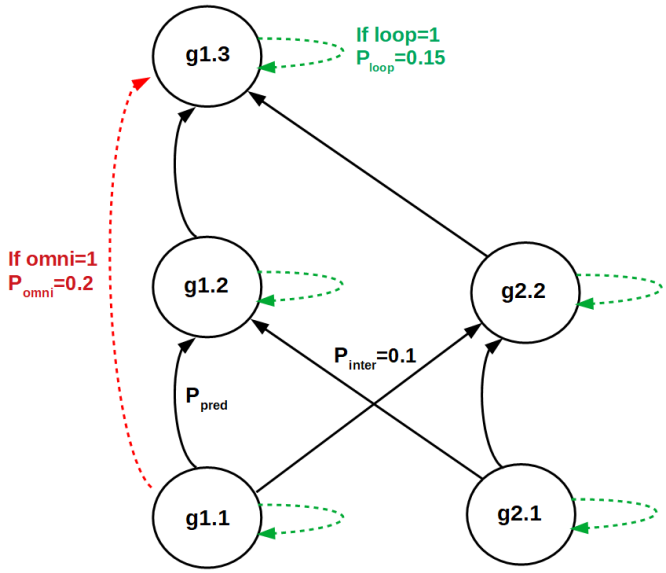

Figure 4: A schematic representation of a group model. It has $n_b = 5$ trophic groups named g1.1 (basal species in module 1), g1.2, g1.3 (top predators in module 1), g2.1 (basal species in module 2), g2.2 (top predators in module 2). Solid black arrows indicate non-null inter-group interaction probabilities. Red and green arrows represent interaction probabilities that would appear if respectively **omni** and **loop** were activated. The group model has two modules (**nModules=2**) composed by 3 and 2 trophic groups respectively (**TrophLens**=(2,1), **maxTrophLen**=2). Each group represents a trophic level inside a module. P_{pred} controls intra-module predation while P_{inter} controls inter-module predation.

Controlling networks connectance through predation link probability. Variables **omni** and **loop** would induce a variation in the expected connectance of the simulated networks if P_{pred} was defined independently of those random properties. We introduced a cor-

rective term in P_{pred} so that the network expected connectance is not affected by **omni** and **loop**. Then, this corrective term insures that connectance of the simulated networks does not vary too much except by the activation of **generalism**, and the effect due to trophic group composition and compartmentalisation.

To justify this corrective term, we firstly establish the expected connectance of simulated networks from an arbibtrary group model setting. Let us denote the following elements of the group model:

- Let $G = (V, E)$ be the random network to be simulated, we denote by n its number of nodes.
- $n_M \in \{0, 1\}$ (**nModules** property) is the number of compartments or Modules in the group model.
- Let $(l_i)_{i \in [\![1, n_M]\!]}$ (**TrophLens** property) be the numbers of trophic levels in each compartment, i.e. the number of groups that they contain. The group model thus contains a total number of $n_b = \sum_{i=1}^{n_M} l_i$ groups.
- Let $\lambda \in V^{n_b}$ be a random partition of nodes V such that every node is affected to one of the groups and the group of a node is drawn according to a uniform multinomial distribution.
- Let $I_o \in \{0, 1\}$ (**omni** property) be the variable indicating a non-null linking probability $P_{omni} = 0.2$ of omnivory links in the group model.
- Let $I_l \in \{0, 1\}$ (**loop** property) be the variable indicating a non-null probability $P_{loop} = 0.15$ of intra-group linking in the group model.
- Let P_{pred} be the simple predation probability between successive trophic groups inside a compartment of the group model.
- Let $I_g \in \{0, 1\}$ (**generalism** property) be the variable indicating the increase of P_{pred} by $P_{generalism} = 0.2$, inducing a higher number of prey per predator in average.

Then, the expected connectance of G may be additively decomposed as:

$$C = C_{\text{pred}} + 1_{n_M=2} C_{\text{inter}} + I_o \, C_{\text{omni}} + I_l \, C_{\text{loop}}.$$

Every term is developed further, and the total expected connectance is thus approximated by (see paragraphs below for details on each contribution):

$$\begin{aligned} C \approx & \frac{1}{n^2}\Big[(P_{pred} + 1_{n_M=2} P_{inter}) \frac{n^2}{n_b^2}(n_b - n_M) + I_o P_{omni} \frac{n^2}{n_b^2} \sum_{i=1}^{n_M} 1_{l_i>2} \binom{l_i - 1}{2} \\ & + I_l P_{loop} \frac{n(n - n_b)}{n_b}\Big] \\ = & \frac{1}{n^2} \frac{n^2}{n_b^2}(n_b - n_M)\Big[P_{pred} + 1_{n_M=2} P_{inter} + I_o P_{omni} \frac{\sum_{i=1}^{n_M} 1_{l_i>2} \binom{l_i - 1}{2}}{n_b - n_M} \\ & + I_l P_{loop} \frac{n(n - n_b) n_b}{n^2 (n_b - n_M)}\Big]. \end{aligned}$$

Thus, as we want to define P_{pred} such that the expected connectance is not affected by **loop** and **omni**, and given that P_{loop} et P_{omni} are constant, it must be decomposed as:

$$P_{pred} := P'_{pred} - I_o P_{omni} \frac{\sum_{i=1}^{n_M} 1_{l_i>2} \binom{l_i - 1}{2}}{n_b - n_M} - I_l P_{loop} \frac{n(n-n_b)n_b}{n^2(n_b - n_M)}.$$

Where P'_{pred} is independent of **loop** and **omni**. We then defined $P'_{pred} := P_{base} + I_g P_{generalism}$, where $P_{base} = 0.7$, to introduce a positive influence of **generalism** on the number of prey per predator. Note that **nModules** and **TrophLens** will impact the expected connectance of the network. We implicitly assume that this impact is realistic and should not be corrected.

Contribution of simple predation C_{pred}.

We compute the approximated expectation of the number of simple predation interaction, i.e. interaction from any group to the group that is just one trophic level above in the same module, divided by the maximum number of interactions in the network n^2. It is computed as follows:

$\mathbb{E}(C_{\text{pred}}) \approx \frac{1}{n^2} \sum_{i=1}^{n_M} \sum_{j=2}^{l_i} P_{pred}\, \mathbb{E}(n_j^i\, n_{j+1}^i) \approx \frac{1}{n^2} P_{pred}\, \frac{n^2}{n_b^2}(n_b - n_M)$

We made, here and in the following paragraphs, the approximate assumption that the number of species n' in a given group follows a binomial distribution, $n' \sim \mathcal{B}(1/n_b, n)$, and that is is independent of the number of species in other groups. Note also that this formula assumes that species are distributed uniformly and independently across groups, whereas in our simulation we first distributed one species per group before distributing the remaining species uniformly. The error brought by our assumption is minor because the probability that there exists a group without any species is always inferior to 0.02 even in the worst case where there are 60 species and the group model has the maximum number of 10 groups. This assumption also applies to the following contributions.

Contribution of omni, C_{omni}.

We compute the approximated expectation of the number of interactions appearing across a non-successive pair of trophic levels inside a same module, divided by the maximum number of interactions in the network n^2. It is computed as follows:

$\mathbb{E}(C_{\text{omni}}) \approx \frac{1}{n^2} \sum_{i=1}^{n_M} 1_{l_i>2} \sum_{j=1}^{l_i-2} \sum_{k=j+2}^{l_i} \mathbb{E}(n_j^i n_k^i) P_{omni}$

$\approx \frac{1}{n^2} \sum_{i=1}^{n_M} 1_{l_i>2} \frac{n^2}{n_b^2} P_{omni} \sum_{j=1}^{l_i-2} \sum_{k=j+2}^{l_i} = \frac{1}{n^2} \frac{n^2}{n_b^2} P_{omni} \sum_{i=1}^{n_M} 1_{l_i>2} \binom{l_i - 1}{2}$

This is because:

$\sum_{j=1}^{l_i-2} \sum_{k=j+2}^{l_i} = \sum_{j=1}^{l_i-2} \sum_{k=j}^{l_i-2} = \sum_{j=1}^{l_i-2} l_i - 1 - j = \sum_{j=1}^{l_i-2} j = \frac{l_i-2+1}{2}(l_i - 2) = \binom{l_i - 1}{2}$

Contribution of loop, C_{loop}.

We computed the expectation of the number of interactions appearing across a pair of species inside a same group, divided by the maximum number of interactions in the network n^2. It is computed as follows:

$\mathbb{E}(C_{\text{loop}}) = \frac{1}{n^2} \sum_{i=1}^{n_M} \sum_{j=1}^{l_i} P_{loop} \mathbb{E}\left((n_j^i)^2 - n_j^i\right) = \frac{1}{n^2} \sum_{i=1}^{n_M} \sum_{j=1}^{l_i} P_{loop} \frac{n}{n_b}(\frac{n}{n_b} - 1)$

$= \frac{1}{n^2} \frac{n(n-n_b)}{n_b} P_{loop}$

Contribution of inter-module interactions, C_{inter}.

We compute the approximated expectation of the number of interactions appearing across a successive trophic levels in distinct modules, divided by the maximum number of interactions in the network n^2. is computed in the following expression which is conditional to '$n_M = 2$', otherwise, the contribution is 0.

$\mathbb{E}(C_{\text{inter}}) = \frac{1}{n^2}(\sum_{j=1}^{l_1-1} P_{inter}\ \mathbb{E}(n_j^1\ n_{j+1}^2) + \sum_{j=1}^{l_2-1} P_{inter}\ \mathbb{E}(n_j^2\ n_{j+1}^1))$
$= \frac{1}{n^2} P_{inter} \frac{n^2}{n_b^2}(n_b - n_M)$

To conclude this section, we show in Figure 5 boxplots of the networks density against 3 quantities: the total number of trophic groups $n_b \in \{2, \ldots, 10\}$ and the binary values of **generalism** and **nModules**.

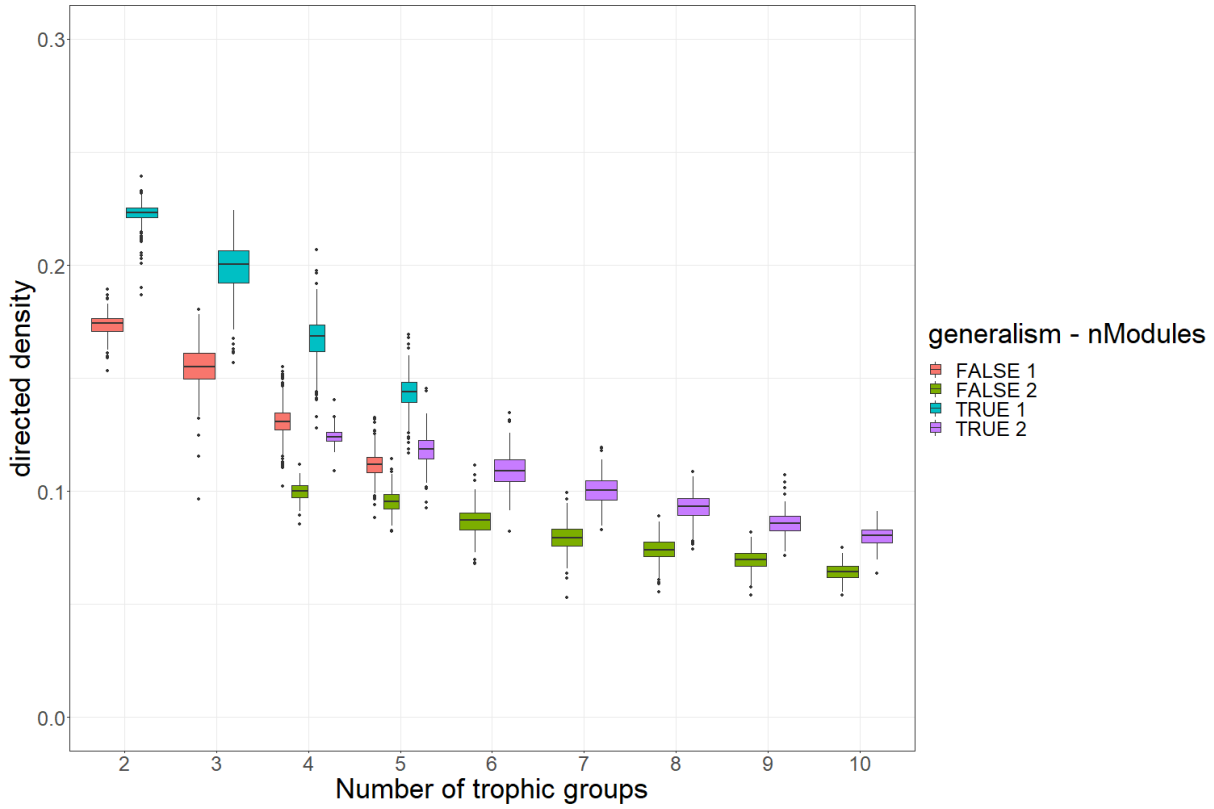

Figure 5: Boxplot of simulated networks density versus number of trophic groups (n_b), **generalism** and **nModules**.

C UMAP plans and ecological properties

Figures 6,7, 8, 9 and 10 hereafter show the UMAP plans of all 7 tested embedding methods where, respectively, the points are colored according to the category of **maxTrophLen**, **nModules**, **omni**, **generalism** and **loop**. Also, the size of the network (60 or 120) is indicated by the shape of the point. In each plan we represented the same random sample of 600 networks taken from the 5000 total simulated random networks.

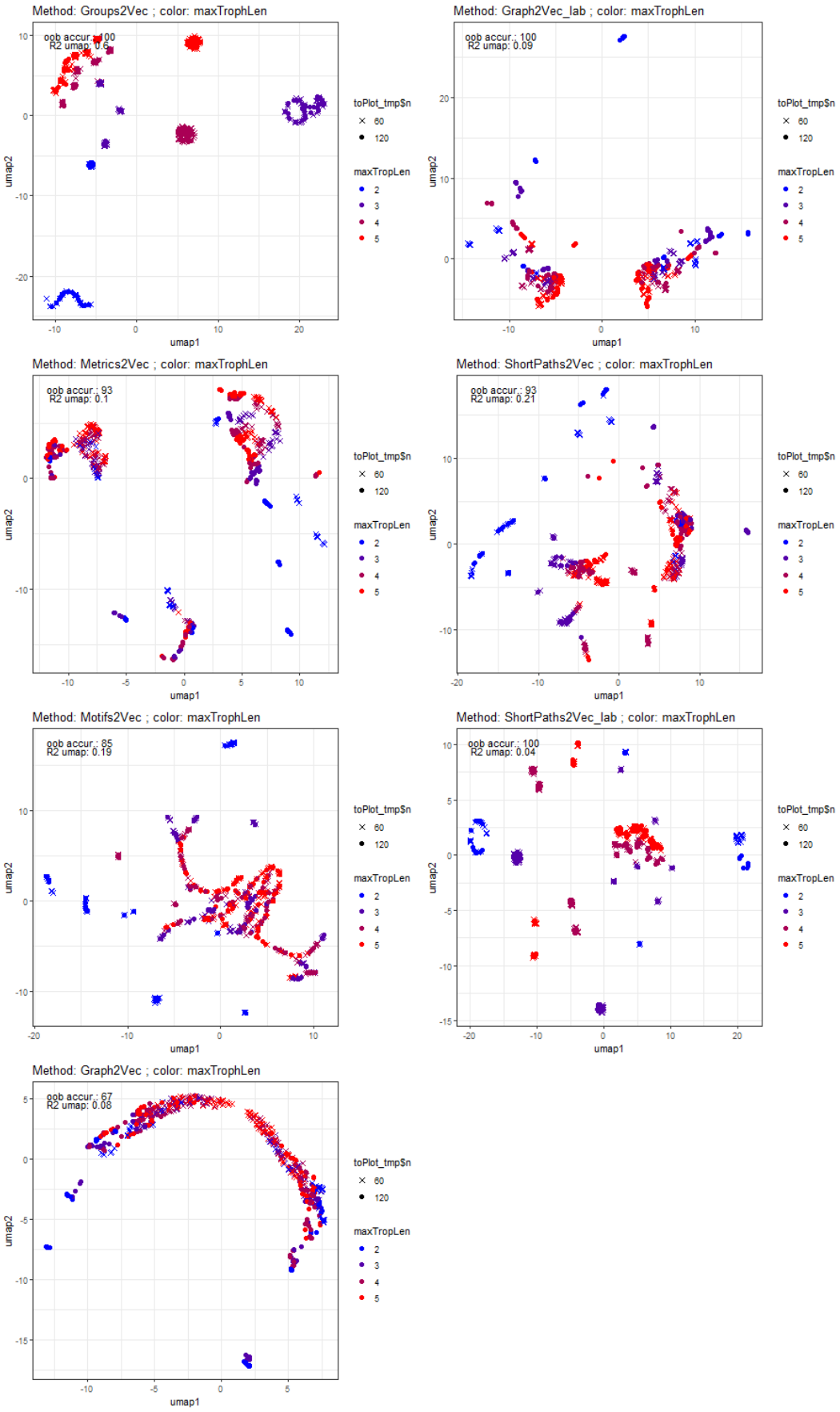

Figure 6: UMAP plans of all tested embeddings colored for **maxTrophLen**.

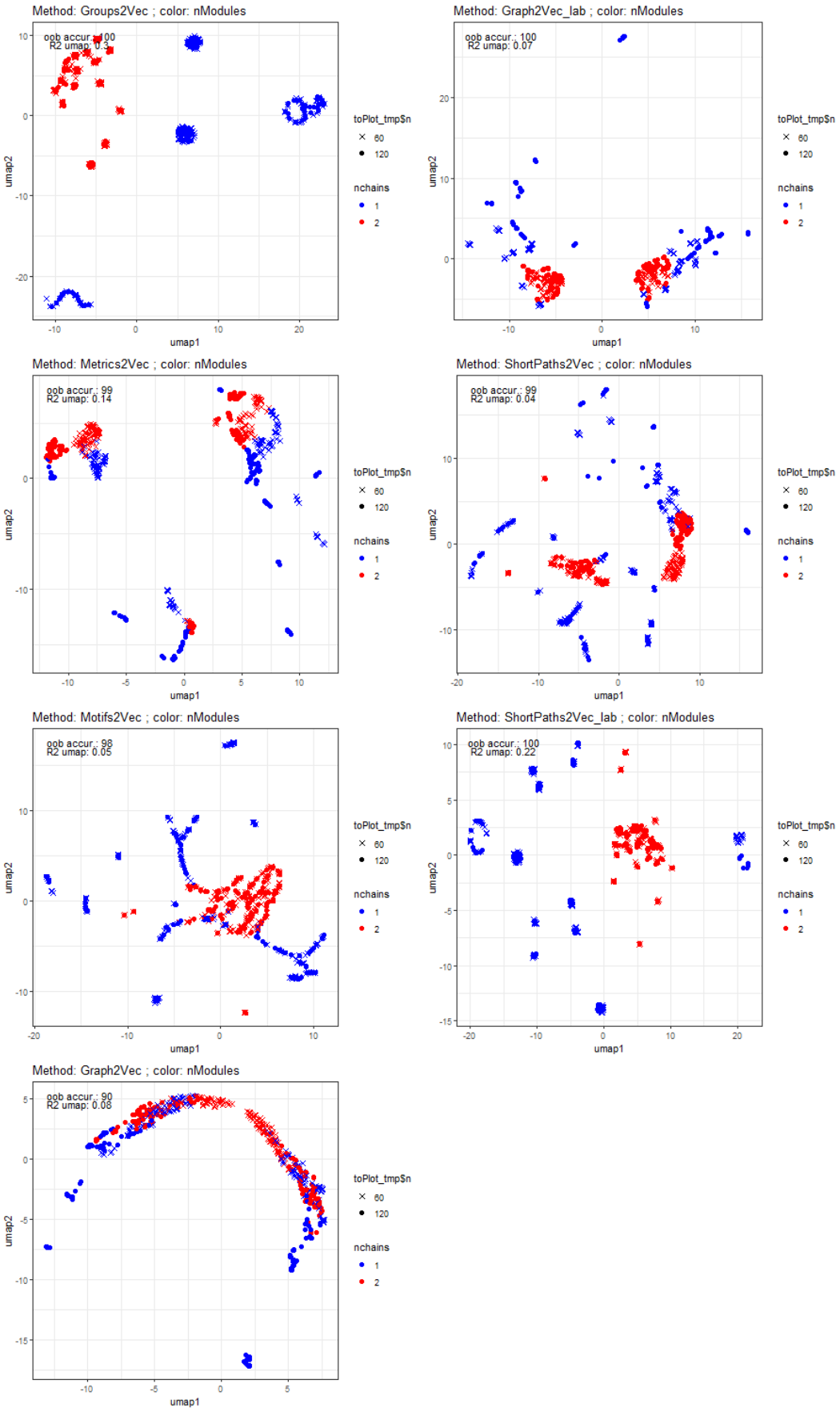

Figure 7: UMAP plans of all tested embeddings colored for **nModules**.

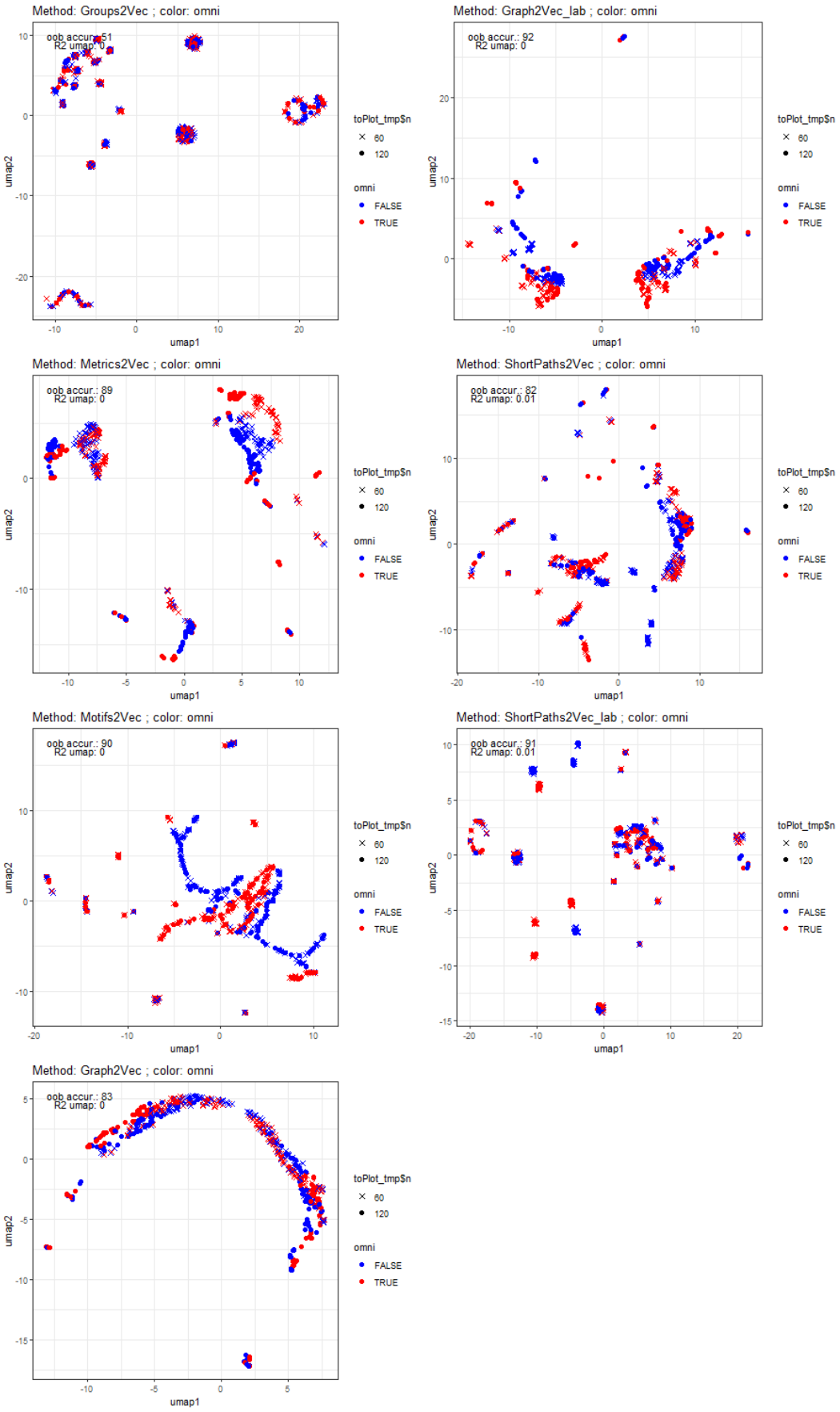

Figure 8: UMAP plans of all tested embeddings colored for **omni**.

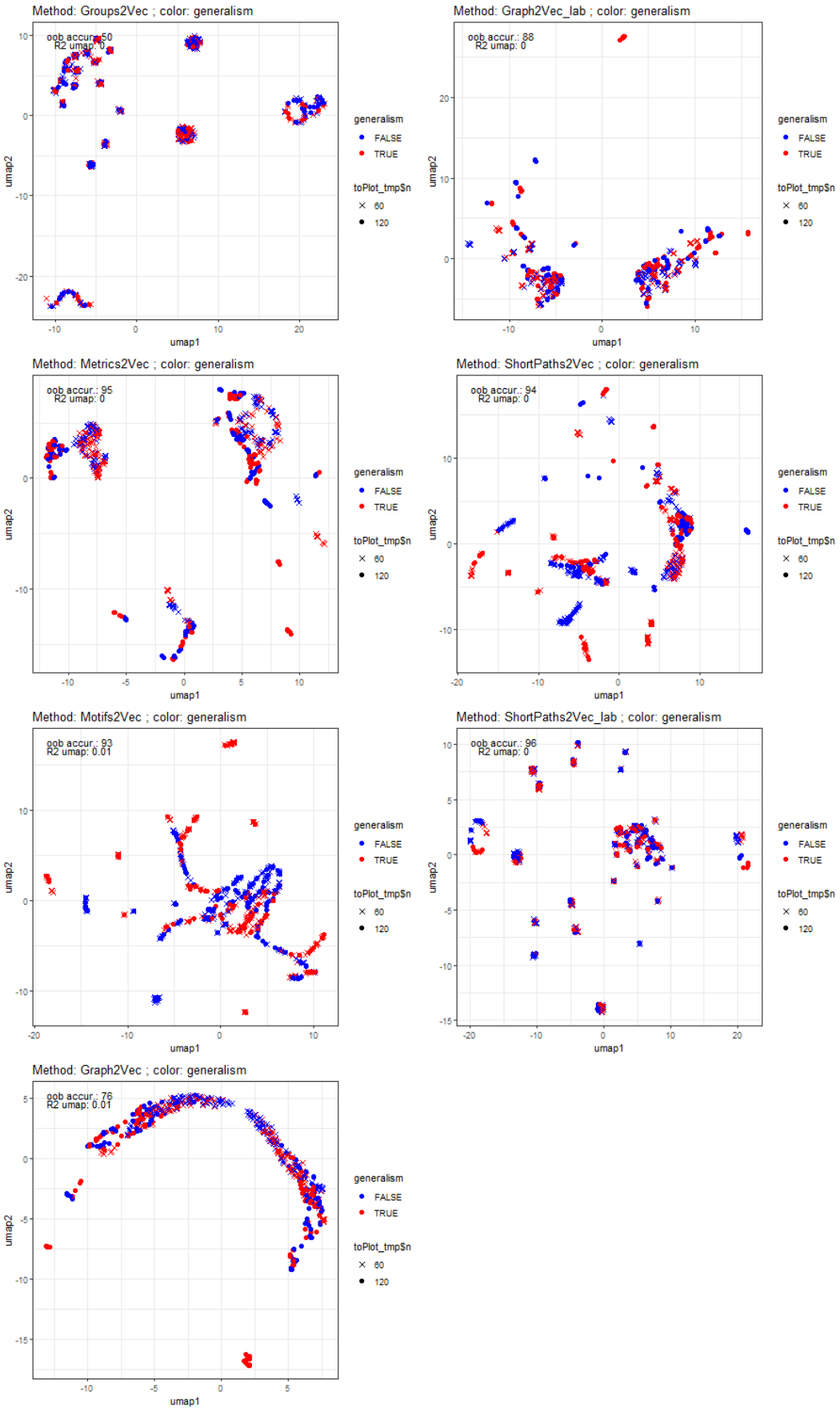

Figure 9: UMAP plans of all tested embeddings colored for **generalism**.

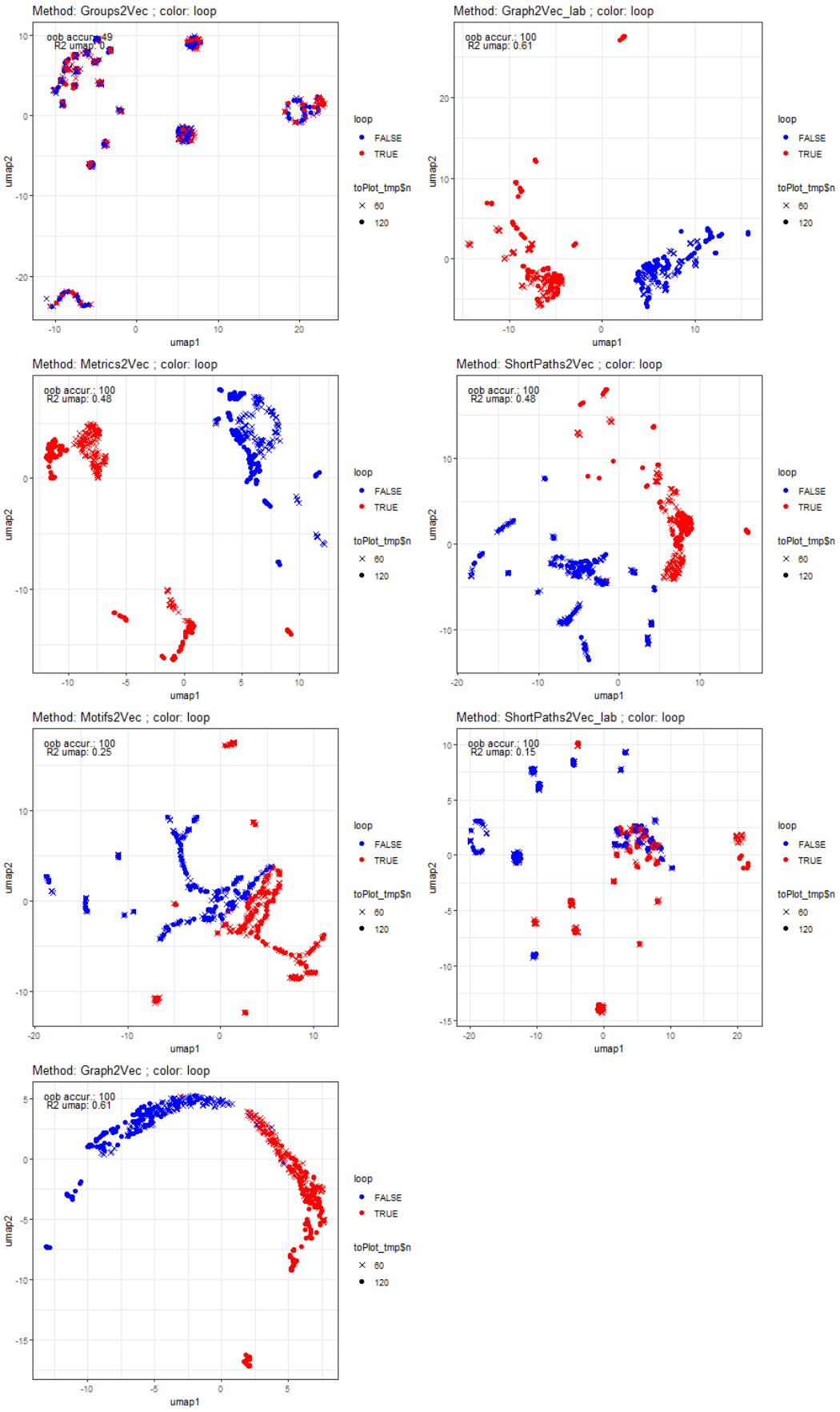

Figure 10: UMAP plans of all tested embeddings colored for **loop**.

D Criterions to measure the quality of the embeddings

D.1 Predictive accuracy

We measured whether the position of a network in an embedding can be related to the category of an ecological property (e.g. the value 2, 3, 4 or 5 for the property **maxTrophLens**). To do so, we relied on Random Forests [Breiman, 2001] trained to predict the category of a network from its position in the embedding. Then, we measured the Out-Of-Bag classification accuracy of this classifier, namely the ratio of well classified networks over the total number of networks (when each tree is trained on all but one network and then used to predict the category of this network). In the following, this measure is simply called predictive accuracy. It quantifies how well some optimized (and non necessarily linear) boundaries in the embedding space may separate the categories of the networks. The predictive accuracy can thus be very good even if networks from a same category are distributed in several disconnected subspaces of the embedding. The Random Forest (R-package **randomForest**) was built with 300 trees and the number of variables available to each tree branch (argument **mtry**) was optimized to minimize the average Out-Of-Bag prediction error. Note also that the Random Forest trained on the embedding obtained from ShortPaths2Vec_lab was learnt after reducing the dimension of this embedding (originally 499) to 60 using Singular Value Decomposition, preserving most of the embedding variability (otherwise the predictive accuracy would have been unfairly reduced compared to other approaches because of overfitting of the random forest).

D.2 Segregation of categories in the embedding

Our second criterion asks, given an ecological property, how well networks from a same category are aggregated in the embedding compared to networks from distinct categories. We measured it with the partial R-squared (R2-ebd herafter) of the nonparametric multivariate analysis of variance [Anderson, 2001] of the embedding, with the grouping of networks given by their ecological category. For a property with K categories, it is given by $R^2_{ebd} = Q^2 \sum_{k,k' \in [\![1,K]\!]^2} \|c_k - c'_k\| / (K^2 \sum_{i,i' \in [\![1,Q]\!]^2} ||e_i - e_j||)$, where Q is the number of networks, $(e_i)_{i \in [\![1,Q]\!]}$ is the embedding vector of network i, the vector c_k is the geometric center of category k, and $\| \cdot \|$ is the Euclidean distance. Thus, $R^2_{ebd} \in [0, 1]$, with $R^2_{ebd} = 0$ if and only if all centers are confounded, and $R^2_{ebd} = 1$ if and only if the points of each category are concentrated at their respective center and at least two centers are distinct. We can say that the highest is the aggregation of categories in the embedding, the closest to 1 is the R2-ebd. A high R2-ebd therefore increases the chances that the user will discover these groups regardless of the clustering or dimension reduction method.

D.3 Measuring segregation of categories in 2D after dimension reduction

We applied UMAP to reduce the dimension of the embeddings to 2, relying on 150 neighbours and the Euclidean distance as hyperparameters of the method. This was done for each of the seven embedding approaches. We used a large number of neighbors to preserve the large scale distances among points and to best reveal the separation of large clusters of networks. Our third criterion, which best approaches the embedding method suitability (in combination with UMAP) for unsupervised analysis, consisted in calculating again the partial R-squared in the two-dimensional UMAP plane obtained from the embedding. We call this metric R2-umap.

D.4 Decrease in segregation of categories due to size variability

This last criterion concerns the robustness of segregation of the ecological property categories to the variability of network sizes. We propose a measure of this phenomenon that is the relative decrease of the partial R-square when the size is variable (i.e. R2-ebd and R2-umap) compared to the corresponding value when the size is constant. To compute that we use the R2-ebd and R2-umap described earlier, which are computed on all networks, i.e. with sizes 60 and 120 together. We also computed the partial R-squared on the networks of size 60 only, called R2-ebd_60 (resp. R2-umap_60) and on networks of size 120 only, called R2-ebd_120 (resp. R2-umap_120). Then, we computed the percent of loss in segregation in the embedding (resp. UMAP plane) due to size variability R2-loss-ebd (resp. R2-loss-umap) with the following formulas

$$R2-\text{loss-ebd} = 100\frac{(\text{R2-ebd}_{60} + \text{R2-ebd}_{120})/2 - \text{R2-ebd}}{(\text{R2-ebd}_{60} + \text{R2-ebd}_{120})/2}.$$

$$R2-\text{loss-umap} = 100\frac{(\text{R2-umap}_{60} + \text{R2-umap}_{120})/2 - \text{R2-umap}}{(\text{R2-umap}_{60} + \text{R2-umap}_{120})/2}.$$

This R2-loss-ebd (resp. R2-loss-umap) is most likely superior to 0 and always inferior to 100. If it equals 0, there is no effect of size variability on the segregation of categories, whereas if it equals 100, then the segregation is totally lost due to network sizes variability. In other words, the higher is the R2-loss-ebd (R2-loss-umap), the less robust is the embedding (resp. UMAP plane obtained from the embedding) to size variability for the segregation of the targeted property. The R2-loss-ebd are given for each method and property in Table 3 while the R2-loss-umap are given in Table 4.

method	maxTrophLen	TrophLens	nModules	omni	generalism	loop
Groups2Vec	0	0	0	26	60	51
Metrics2Vec	10	11	9	9	10	10
Motifs2Vec	0	0	0	0	0	0
Graph2Vec	61	65	64	72	52	23
Graph2Vec_lab	26	21	22	20	43	15
ShortPaths2Vec	9	7	6	3	7	2
ShortPaths2Vec_lab	1	1	1	1	4	2

Table 3: R2-loss on embeddings (R2-loss-ebd): Loss of aggregation of categories (in %) due network sizes variability for each ecological property and each embedding.

method	maxTrophLen	TrophLens	nModules	omni	generalism	loop
Groups2Vec	0	0	0	25	52	40
Metrics2Vec	9	10	4	-3	21	12
Motifs2Vec	0	1	0	-1	3	0
Graph2Vec	13	14	12	55	12	14
Graph2Vec_lab	16	18	15	16	21	10
ShortPaths2Vec	4	7	8	2	18	2
ShortPaths2Vec_lab	3	0	0	0	100	0

Table 4: R2-loss on Umap plans (R2-loss-umap): Loss of aggregation of categories (in %) due network sizes variability for each ecological property and each Umap plan.

E Agreements between embeddings

One complementary question that may be asked is: what is the level of agreement between embedding methods? In other words, what is the degree of similarity between their placement of the simulated networks in their respective embeddings? To answer this question, we use the distance correlation [Székely and Rizzo, 2009] with the Euclidean Distance. This metric varies between 0 and 1. An agreement of 1 means that the Euclidean Distance Matrices (EDMs) of the two embeddings are equals up to a constant factor. In other words, an agreement of 1 means that the two embeddings are equal up to a composition of a rigid transformation (composition of rotation, translation, reflection) and homogeneous dilation (homothety) [Dokmanic et al., 2015]. More generally, this metric is insensitive to those compositions of transformations, which is an important property for our application. For example Graph2Vec may produce unstable embeddings that are equivalent up to those transformations. The distance correlation between all pairs of embeddings of the simulated networks is represented in Figure 11.

We can clearly see from correlations in Figure 11, that are all superior to 0.45, and form the scatter-plots of distances in the lower triangle of the same Figure, that the distances between networks are overall highly correlated between embeddings. It means that each network position relatively to others is somehow consistent from one embedding to another. This may be explained because one can find dimensions that quantify approximately the same topological features across embeddings. For instance, the assortativity column in Metrics2Vec measures the frequency of closed triangles which is highly correlated with several columns of Motifs2Vec.

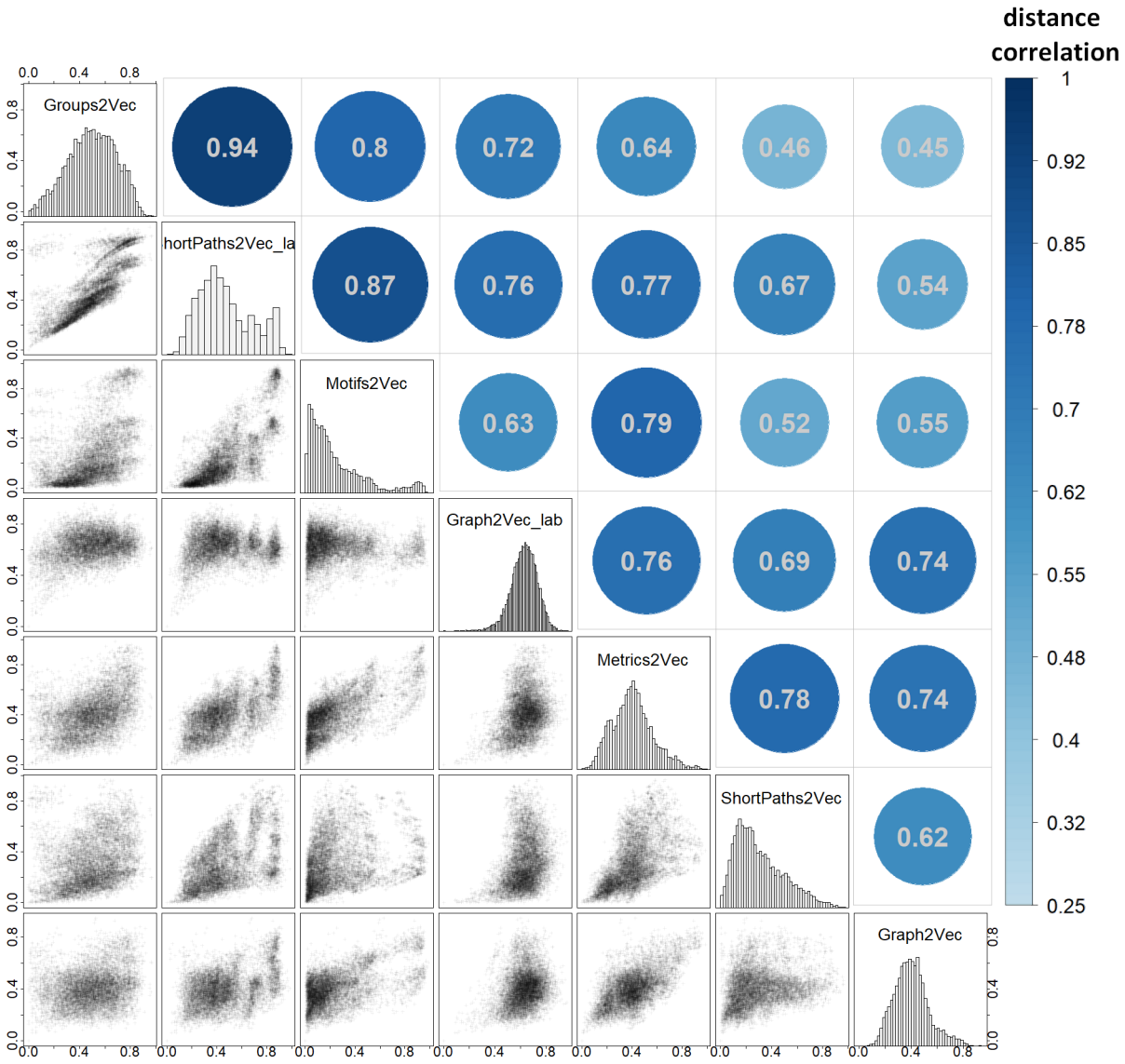

Figure 11: Agreements between embeddings. In the upper triangle, we represent the distance correlation between each pair of embeddings. On the diagonal, we plot the histogram of networks pairs distances per embedding. In the lower triangle, we scatter-plot the networks pairs Euclidean distances for each pair of embeddings: The higher the distance correlation, the more the points are aggregated along a straight line.

References

Allesina, S. and Pascual, M. (2009). Food web models: a plea for groups. *Ecology letters*, 12(7):652–662.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral ecology*, 26(1):32–46.

Braga, J., Pollock, L. J., Barros, C., Galiana, N., Montoya, J. M., Gravel, D., Maiorano, L., Montemaggiori, A., Ficetola, G. F., Dray, S., et al. (2019). Spatial analyses of multi-trophic terrestrial vertebrate assemblages in Europe. *Global Ecology and Biogeography*, 28(11):1636–1648.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Dokmanic, I., Parhizkar, R., Ranieri, J., and Vetterli, M. (2015). Euclidean distance matrices: essential theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 32(6):12–30.

Kortsch, S., Primicerio, R., Aschan, M., Lind, S., Dolgov, A. V., and Planque, B. (2019). Food-web structure varies along environmental gradients in a high-latitude marine ecosystem. *Ecography*, 42(2):295–308.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Narayanan, A., Chandramohan, M., Venkatesan, R., Chen, L., Liu, Y., and Jaiswal, S. (2017). graph2vec: Learning distributed representations of graphs. In *Proceedings of the 13th International Workshop on Mining and Learning with Graphs (MLG)*.

Newman, M. E. (2002). Assortative mixing in networks. *Physical review letters*, 89(20):208701.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582.

Shervashidze, N., Schweitzer, P., Van Leeuwen, E. J., Mehlhorn, K., and Borgwardt, K. M. (2011). Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(77):2539–2561.

Székely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. *The annals of applied statistics*, pages 1236–1265.

Thompson, R. M. and Townsend, C. R. (2005). Food-web topology varies with spatial scale in a patchy environment. *Ecology*, 86(7):1916–1925.

Tylianakis, J. M., Laliberté, E., Nielsen, A., and Bascompte, J. (2010). Conservation of species interaction networks. *Biological conservation*, 143(10):2270–2279.

Wasserman, S., Faust, K., et al. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.

Williams, R. J. and Martinez, N. D. (2004). Limits to trophic levels and omnivory in complex food webs: theory and data. *The American Naturalist*, 163(3):458–468.

Wood, S. A., Russell, R., Hanson, D., Williams, R. J., and Dunne, J. A. (2015). Effects of spatial scale of sampling on food web structure. *Ecology and evolution*, 5(17):3769–3782.