HAL open science

Quantifying the overall effect of biotic interactions on species communities along environmental gradients

Vincent Miele, Catherine Matias, Marc Ohlmann, Giovanni Poggiato, Stéphane Dray, Wilfried Thuiller

To cite this version:

Vincent Miele, Catherine Matias, Marc Ohlmann, Giovanni Poggiato, Stéphane Dray, et al.. Quantifying the overall effect of biotic interactions on species communities along environmental gradients. 2021. hal-03172480v1

HAL Id: hal-03172480

https://hal.science/hal-03172480v1

Preprint submitted on 18 Mar 2021 (v1), last revised 28 Apr 2023 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Quantifying the overall effect of biotic interactions on species communities along environmental gradients

Vincent Miele[1,*], Catherine Matias[2], Marc Ohlmann[3,4], Giovanni Poggiato[4,5], Stéphane Dray[1], and Wilfried Thuiller[4]

[1]Université de Lyon, F-69000 Lyon; Université Lyon 1; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Évolutive, F-69622 Villeurbanne, France

[2]Sorbonne Université, Université de Paris, Centre National de la Recherche Scientifique, Laboratoire de Probabilités, Statistique et Modélisation, F-75005 Paris, France

[3]Univ. Savoie Mont–Blanc, CNRS, LAMA, F-73000 Chambéry, France.

[4]Univ. Grenoble Alpes, CNRS, Univ. Savoie Mont Blanc, LECA, Laboratoire d'Ecologie Alpine, F-38000 Grenoble, France

[5]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, F-38000 Grenoble, France

[*]Corresponding author: vincent.miele@univ-lyon1.fr

Abstract

Separating environmental effects from those of biotic interactions on species distributions has always been a central objective of ecology. Despite years of effort in analysing patterns of species co-occurrences and communities and the developments of sophisticated tools, we are still unable to address this major objective. A key reason is that the wealth of ecological knowledge is not sufficiently harnessed in current statistical models, notably the knowledge on biotic interactions.

Here, we develop ELGRIN, the first model that combines simultaneously knowledge on species interactions (i.e. metanetwork), environmental data and species occurrences to tease apart the relative effects of abiotic factors and overall biotic interactions on species distributions. Instead of focusing on single effects of pair-wise interactions, which have little sense in complex communities, ELGRIN contrasts the overall effects of biotic interactions to those of the environment.

Using simulated and empirical data, we demonstrate the suitability of ELGRIN to address the objectives for various types of interactions like mutualism, competition and trophic interactions.

Data on ecological networks are everyday increasing and we believe the time is ripe to mobilize these data to better understand biodiversity patterns. We believe that ELGRIN will provide the unique opportunity to unravel how biotic interactions truly influence species distributions.

Key-words: biodiversity patterns, C-score, species co-occurrence, metanetwork, Markov random fields, environmental niche.

1 Introduction

Ecologists have always strived to understand the drivers of biodiversity patterns with the particular interest to tease apart the effects of environment and biotic interactions on species distributions and communities (Ricklefs, 2008; Thuiller *et al.*, 2015). Species distributions are influenced by the abiotic environment (e.g. climate or soil properties) because of their own physiological constraints that allow them or not to sustain viable populations in specific environmental configurations. However, the occurrence of a species in a given site is also influenced by other species through all sort of interactions that can be trophic (e.g. a predator needs prey), non-trophic (e.g. plant species need to be pollinated by insects) or competitive (two species with the same requirements might exclude each other) (Guisan *et al.*, 2017; Gravel *et al.*, 2019).

Teasing apart the effects of environmental variations and biotic interactions on species distributions and communities from observed co-occurrence patterns has always been a hot topic in ecology since the earlier debate between Diamond (1975) and Connor & Simberloff (1979), to the recent syntheses on the subject (Blanchet *et al.*, 2020). More than anything, with a few exceptions and despite recent advances like joint species distribution models (Ovaskainen *et al.*, 2017) or elegant null model developments (Peres-Neto *et al.*, 2001; Chalmandrier *et al.*, 2013), the conclusion has been that is almost impossible to retrieve and estimate biotic interactions from observed spatial patterns of species communities (Zurell *et al.*, 2018). This conclusion should thus preclude any attempt to disentangle the relative effects of environment and biotic interactions. A major difficulty of this long-standing issue is that biotic interactions could be of any type (i.e. positive, negative, asymmetric) and that observed patterns average out all these interactions. Observed communities reflect the overall outcome of biotic interactions that is difficult to dissect, especially when analysing pairwise species spatial associations as it is commonly done. Yet, this overall outcome might be worth analysing on its own, for instance to measure the overall strength of biotic interactions in a given community and between communities, how it depends on the co-existing species, and how it varies in space.

Interestingly, so far there have been few attempts to integrate the wealth of existing knowledge to address this fundamental ecological issue (Blanchet *et al.*, 2020; Holt, 2020). Indeed, the spatial analysis of biotic interactions is gaining an increased interest with novel technologies to measure interactions in the field (e.g. camera-traps, gut-content), open databases (i.e. GLOBI, Fungal) and the developments of new statistical tools to analyse them (Tylianakis & Morris, 2017; Pellissier *et al.*, 2018; Ohlmann *et al.*, 2019). The combination of expert knowledge, literature, available databases, and phylogenetic hypotheses has also given rise to large metanetworks that generalise the regional species-pool of community ecology by incorporating the potential interactions between species from different trophic levels along with their functional and phylogenetic characteristics (Maiorano *et al.*, 2020; Morales-Castilla *et al.*, 2015). Despite a single attempt (Staniczenko *et al.*, 2017), information on interaction networks has been poorly integrated to understand and model biodiversity patterns. We believe that the time is ripe to consider network information in the process of modelling species distributions and communities. It implies to integrate both biotic and abiotic information (and their spatial variations) as explanatory factors in statistical models to weight their relative strength.

In this paper, we propose a novel statistical model, called ELGRIN (in reference to Charles Elton and Joseph Grinnell) that can handle the effects of both environmental factors and known ecological interactions (aka a metanetwork) on species distributions. We rely on Markov random fields (MRF, also called Gibbs distribution, e.g., Brémaud, 1999), a family of flexible models that can handle dependencies between variables using a graph. By considering both abiotic and biotic processes in the modelling framework, our approach allows to capture the spatial variation of the relative effects of environmental and biotic dimensions on the composition of ecological communities. It thus provides a convenient way to integrate network ecology in joint species community modeling. More specifically, ELGRIN jointly models the presence and absence of all species in a given area in function of environmental covariates and the topological structure of the known metanetwork. The outcome of the model allows interpreting the relative strength of environment against overall species spatial associations and analysing how this relative strength varies across space (Figure 1). In this paper, we first present the overall modelling framework, then validate its performance using simulations and finally apply the model on vertebrate trophic networks in the European Alps.

2 Material and methods

2.1 Species data and potential interactions

We consider a set of sites or locations indexed by $l \in \{1, \ldots, L\}$, where the occurrence (presence/absence) of N species and a set of environmental variables (stored in the vector W_l) are recorded. For the same set of N species, we assume that we know how they interact, i.e. the metanetwork which can be summarised with a graph $G^\star = (V^\star, E^\star)$ over the set of nodes $V^\star = \{1, \ldots, N\}$ and edges $E^\star$. This graph represents all potential interactions between any pair of species that could occur in a site. This graph represents a *common regional pool of both species and interactions*, which might be obtained, for instance, by aggregating local networks at different locations or from expert knowledge and literature review (e.g., Cirtwill *et al.*, 2019; Maiorano *et al.*, 2020). Note that various types of interactions can be considered here (e.g., trophic, mutualism, competition). Using a mixture of them is technically feasible, $G^\star$ recording the presence of an interaction but not its type. For our model, this graph is undirected with no self-loops (see model specification below). Hereafter, we refer to co-present (or co-absent) species when these species are connected in the metanetwork and jointly present (or absent, respectively) at a given location.

2.2 The statistical model of ELGRIN

Model description For each location $l \in \{1, \ldots, L\}$, we consider a set of random variables $\{X_i^l\}_{i \in V^\star}$ taking values in $\{0, 1\}$ and that represent the presence/absence of species $i \in V^\star$ at location $l \in \{1, \ldots, L\}$. We rely on a *Markov random field* (see for instance Brémaud, 1999) to model the dependencies between species occurrences at location l. These dependencies are encoded through the metanetwork $G^\star$. For each location $l \in \{1, \ldots, L\}$, we thus assume that these random variables are distributed according to a Gibbs distribution specifying the

joint associations between the species occurrence variables $\{X_i^l\}_{i\in V^\star}$, as follows:

$$\mathbb{P}(\{X_i^l\}_{i\in V^\star}) = \frac{1}{Z}\exp\Big(\sum_{i\in V^\star}[a_l + a_i + W_l^\intercal b_i + (W_l^2)^\intercal c_i]X_i^l \tag{1a}$$

$$+\,\beta_{l,co-pres}\sum_{(i,j)\in E^\star}\mathbf{1}\{X_j^l = X_i^l = 1\} \tag{1b}$$

$$+\,\beta_{l,co-abs}\sum_{(i,j)\in E^\star}\mathbf{1}\{X_j^l = X_i^l = 0\}\Big), \tag{1c}$$

where $\mathbf{1}\{E\}$ is the indicator function of event E (either co-absence $X_j^l = X_i^l = 0$ or co-presence $X_j^l = X_i^l = 1$) and Z a normalising constant discussed below. All the model parameters have an ecological interpretation (Table 1). The use of W_l and W_l^2 allows modelling species response to environmental gradient following a bell-shaped relationship, as expected under classical niche theory (Chase & Leibold, 2003).

Sub-equation (1a) is the Grinnellian part of ELGRIN, as it represents some prior probability of species occurrences independently of their interactions. The real-valued parameter a_l controls the expected species richness at location l and can been seen as the richness capacity of the site (Storch & Okie, 2019). Parameters a_i, b_i, c_i capture the response of species i to environment, seen through a vector of environmental covariates W_l. The intercept a_i can be interpreted as the prevalence of species i whereas the vectors b_i, c_i deal with its environmental niche, like in a standard species distribution model (Guisan *et al.*, 2017).

Sub-equations (1b) and (1c) form the Eltonian part of ELGRIN. It considers only interactions $(i,j) \in E^\star$, i.e. the edges of the metanetwork. The β_l represent the overall influence of the interactions (as encoded through $G^\star$) on all species presence/absence at location l. However, this influence may be different for co-presence and co-absence, with parameters $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$ respectively (see Table 2). When a $\beta_{l,co-pres}$ is positive, it represents a positive driving force of co-presence on species distributions. By contrast, when it is negative, it indicates that species co-presence are avoided. The same reasoning holds with $\beta_{l,co-abs}$ for co-absence. Note that we chose the parameters β_l to be specific to location $l \in \{1,\ldots,L\}$ such that the effect of species interactions can vary across space. Finally, Z is a normalising constant that cannot be computed for combinatorial reasons, although the statistical inference procedure will deal with that. All the details of the estimation procedure and parameters identifiability are available in Supporting Information.

Lastly, it is important to note that the metanetwork $G^\star$ cannot be directed in our modelling procedure. Indeed, Markov random fields specify conditional dependencies between random variables $\{X_i^l\}$ in a non-directed way. Our model assumes that these dependencies are given by the interaction network without considering the direction of edges. Consequently, this statistical model of interaction cannot be read in the light of causality. In case of trophic interactions, it consists in assuming that presence/absence of a predator and its prey are intertwined, without specifying top-down or bottom-up control.

ELGRIN is implemented in `C++` for efficiency and is available as part of the `R` package `econetwork` available at `CRAN` (`https://cran.r-project.org/`).

Model interpretation In an hypothetical example where $G^\star$ is an empty graph (no edges, none of the species interact), the random variables $\{X_i^l\}_{i\in V^\star}$ are independent and each species

is present with probability $e^{\alpha_{i,l}}/(1+e^{\alpha_{i,l}}) \in (0,1)$, where $\alpha_{i,l} = a_l + a_i + W_l^\intercal b_i + (W_l^2)^\intercal c_i$. In other words, $\alpha_{i,l}$ is the logit of the probability of presence of species i at location l in the absence of interactions. Assuming that we have included all important environmental covariates, that there is no dispersal limitation, and no model mis-specifications, $\alpha_{i,l}$ is analogous to the fundamental niche parameters of the species (sensu Hutchinson, 1959). It gives the probability of presence of species i at location l when there is no known effect of interactions.

However, when dealing with potential known species interactions (as recorded in $G^\star$), presence/absence information is smoothed across neighbouring nodes in $G^\star$ (see Table 2). More precisely, whenever two species $i, j \in V^\star$ can potentially interact, i.e. whenever $(i,j) \in E^\star$, the presence or absence of one species influences the other so that both variables tend to be equal to 1 (i.e. both species present) when regulated by a positive $\beta_{l,co-pres}$, or equal to 0 (i.e. both species absent) with a positive $\beta_{l,co-abs}$. On the other hand, when $\beta_{l,co-pres}$ (or $\beta_{l,co-abs}$, respectively) is negative, co-presence configurations (or co-absence, respectively) tend to be avoided, meaning that only one of the two species tends to be present. The larger the absolute value of β_l, the stronger the strength of the effects.

2.3 Application to simulated data

Simulation model We used an updated version of the model developed by Münkemüller & Gallien (2015) to simulate communities whose composition is driven simultaneously by biotic and abiotic environmental effects. We considered N species in the species pool and L communities to simulate (i.e. the number of locations). For each community, we associated a single environmental covariate (hereafter called "environment") that we uniformly drew from a range of values between 0 and 100. Each community has the same carrying capacity K (i.e. the exact number of individuals in each location). We defined the environmental niche (or preference) of each species as a Gaussian distribution centered on a given optimum. The environmental niches optima were regularly taken on a grid between 0 and 100, whereas the standard deviations were all equal to a given value σ for simplicity (to define in the set-up procedure). Finally, we considered an undirected network of species interactions, the metanetwork $G^\star$, under two different scenarios: considering these interactions were reciprocal and all negative (competition) or all positive (mutualism), and randomly drawn according to the following rules. Regarding mutualism, we considered that species that facilitate each other tend to have an abiotic niche that is not too close and not too far from each other. The underlying hypothesis is that a species whose niche optimum is really far from the environmental conditions of the location will anyway be maladapted and will not survive. Reciprocally, as we consider only one environmental variable, if two species have the same niche, we considered that there is no reason why they will facilitate each other, they will rather compete. The details of the model as well as the simulation set-up are described in Supporting Information.

2.4 Application to real data

We analyse the newly available Tetra-EU 1.0 database, a species-level trophic meta-web of European tetrapods (Maiorano *et al.*, 2020). This dataset comprises a continental scale, species-level, metanetwork of trophic interactions (i.e. food web) connecting all tetrapods

(mammals, birds, reptiles, amphibians) occurring in Europe. This metanetwork is based on data extracted from scientific literature, including published papers, books, and grey literature (see Maiorano *et al.*, 2020, for a complete description of the data and the reference list used to build the metanetwork). We decided to restrict our analyses on the European Alps that show sharp environmental gradients and varying trophic web distributions (O'Connor *et al.*, 2020). We extracted the species distribution data from Maiorano *et al.* (2013) at a 300 m resolution. We upscaled all species ranges maps to a 10x10 km equal-size area grid and cropped the distribution data on the European Alps. Species were considered present on a given 10x10 km cell if they were present in at least one of the 300 x 300 m cells within it. This yielded species distributions maps for 257 breeding birds, 99 mammals, 36 reptiles, and 30 amphibians over 2138 locations. Environmental covariates were extracted at the same resolution and were selected following previous work on those data (Braga *et al.*, 2019). For climate, we used mean annual temperature, temperature seasonality, temperature annual range, total annual precipitation and coefficient of variation of precipitation that were all extracted from the Worldclim v2 database (`http://www.worldclim.org/bioclim`). Using GlobCover (GlobCover V2.2; `http://due.esrin.esa.int/page_globcover.php`), we extracted the number of habitats present in a given pixel, habitat diversity in a given pixel based on Simpson index and habitat evenness as a measure of habitat complexity. Finally, we added an index of annual net primary productivity (Global Patterns in Net Primary Productivity, v1 (1995), `http://sedac.ciesin.columbia.edu/data/set/hanpp-net-primary-productivity`) and the human footprint index (`http://sedac.ciesin.columbia.edu/data/set/wildareas-v2-human-footprint-geographic`). Since these data were highly correlated, we used a PCA to retain the three leading vectors as environmental covariates (W_l) in ELGRIN.

3 Results

3.1 Simulated communities

When analysing the data simulated under a competition mechanism, we observed that the estimated a_i's were linearly correlated to species frequencies (Figure 2a). However, the estimated a_l's were only slightly correlated with species richness, with a high variability (Figure 2b). This was expected since, in our simulated procedure, we included a carrying capacity in terms of number of individuals but not in terms of species number. In other words, we did not introduce a direct mechanism that would control richness in the simulations. Interestingly, we discovered that the estimated niche parameters were coherent with the species niche as they were introduced in the simulation (Figure 2c). As expected, the quality of estimations is lower at the extrema of the gradient as bell-shaped species response are truncated. Therefore, the estimation of the parameters of the Grinnellian part of our model (subequation (1a)) seems relevant. For the Eltonian part of our model (subequations (1b),(1c)), we noticed that the parameters $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$ were mostly negative (Figures 2d-e). Strikingly, the more negative $\beta_{l,co-pres}$, the less frequent co-presence was as compared to expectation in random assemblages (see Supporting Information). A similar effect was observed for $\beta_{l,co-abs}$

and co-absence. These negative values demonstrates that ELGRIN captures the mark of the impact of competition. They indicated that co-presence and co-absence were avoided, leading to some level of competitive exclusion. This phenomenon was clearly the by-product of the competitive interactions and the carrying capacity in terms of number of individuals (that explicitly induced exclusion). Finally, $\beta_{l,co-pres}$ tended to be much more negative on locations with extreme environmental conditions whereas $\beta_{l,co-abs}$ slightly increased (Figure 2f). Indeed, two interacting species with a niche optimum close to one extreme condition are jointly disadvantaged in the other extreme. Consequently, these species suffered a double penalty (bad environment fit and competition) and were very unlikely to be present, and frequently co-absent.

When focusing on mutualism, we reached similar conclusions than previously about the Eltonian part of our model and the parameters a_i, a_l and niche parameters (Figures 3a–c). Regarding the Eltonian part, we observed a mixture of positive and negative β_l values. A large series of locations had a positive $\beta_{l,co-pres}$ correlating with frequent observed co-presence as compared to expectation (Figure 3d). This observation was due to our simulation framework integrating mutualism that promoted co-presence of interacting species in simulated communities. On the other hand, the negative estimated $\beta_{l,co-abs}$ were linked to more frequent co-absence than expected (Figure 3e). These locations tended to avoid interacting species. For a further analysis, we checked the possible link between environment and β_l values (Figure 3f). We noticed that $\beta_{l,co-pres}$ was negative in extreme environmental conditions whereas $\beta_{l,co-abs}$ increased. Indeed, in a pair of interacting species, it is possible that one species is not adapted to these conditions and the effect of mutualism was not strong enough to make this species present. Since mutualistic interactions are reciprocal, a cascading effect induced that the other species did not benefit any more from mutualism and also disappeared. This was confirmed by the dramatic increase in $\beta_{l,co-abs}$ that we observed in extreme environmental conditions.

3.2 Empirical case study

When we fitted our model to the European vertebrate dataset, the parameters $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$ were highly correlated (see Supporting Information) suggesting joint effects on predator/prey co-presence and co-absence (being in this case two sides of the same coin). In what follows, we therefore mainly dealt with $\beta_{l,co-pres}$.

Parameters a_i and species frequencies were linked by a clear logit function (Figure 4a), showing that the a_i perfectly fitted the empirical species prevalence. The estimated a_l's were linearly correlated with species richness, but with significant variability: indeed, for any given richness value, a_l tends to be greater when β_l was low and negative, but smaller when β_l was high (bluish and reddish points respectively above and below the regression line in Figure 4b). It indicates that the effects of interactions, as measured by the β_l, become partly responsible for species richness. For instance, a richness of 200 species can be due to a moderate richness capacity (a_l close to -1) plus a large co-presence effect (reddish points in Figure 4b). The Eltonian effects are visible here, highlighted by our model when the β_l are high in absolute value.

We also observed a spatial pattern of the β_l estimates, with regions of negative or positive $\beta_{l,co-pres}$ (bluish or reddish colors respectively in Figure 4c). The largest $\beta_{l,co-pres}$ values

were found mainly in the french Alps and in the Eastern zone. Almost all the highest $\beta_{l,co-pres}$ (> 0.05) are in locations with an altitude below 1600 m (Figure 4d, left). In these regions, richness and more importantly connectance tend to be high (see Figure 4e). Since $\beta_{l,co-abs}$ is also positive, co-presence as well as co-absence are favored here: this is the sign of high inter-dependence between preys and predators that were concomitantly present, and sometimes absent. In the opposite, the higher up, the more likely $\beta_{l,co-pres}$ is negative. This is particularly above 1600 m in the central Alps, where almost all the smallest $\beta_{l,co-pres}$ (< -0.5) were estimated (Figure 4d, right).

Locations with negative $\beta_{l,co-pres}$ have a lower richness (Figure 4e). Since $\beta_{l,co-abs}$ tend to be negative as well, co-presence as well as co-absence are disfavored here. Interestingly, these locations also have a different trophic network structure. Relying on the trophic group definition proposed by O'Connor *et al.* (2020), we observed a lower richness in trophic groups in these locations (Figure 4f) as compared to locations with a higher $\beta_{l,co-pres}$, but also a lower diversity in trophic groups. This latter observation suggested the trophic groups were not distributed in a constant manner over the whole region. In summary, where $\beta_{l,co-pres}$ is negative, in particular in higher altitude, the biotic effects captured by ELGRIN are linked to a peculiar structure of the trophic networks observed in these locations. This is confirmed by the analysis of trophic group 4 that gathers about 15 to 20% of the metanetwork species (Figure 4f). These species were more prominent and as well as less connected to the other species when $\beta_{l,co-pres}$ is highly negative (Figure 4g; the reverse holds with highly positive $\beta_{l,co-pres}$). If they are less connected with their potential preys or predators, this means that co-presence is disfavored: this is exactly what negative $\beta_{l,co-pres}$ values in ELGRIN are supposed to model.

4 Discussion

Deciphering the mechanisms explaining spatial patterns of species distributions and communities is likely one of the most active fields of ecological research since the early days of biogeography and community ecology. Still, there was so far no comprehensive statistical approach able to make the best of existing knowledge on biotic interactions, species occurrence and environmental data to measure and quantify the dual effects of environment and biotic interactions on species distributions. Our proposed model that relies on Markov random fields builds on the ability of graph theory to represent known species interactions under a network formalism. This formalism is ideal because it allows within the same model to account for both the effects of the environment and the biotic interactions, which reconciles the Grinnellian vision of species niches (i.e. how species respond to the abiotic environment) with its Eltonnian counterpart (i.e. how species respond to the biotic environment). The mathematical foundations of ELGRIN are strong and its framework is flexible allowing for useful extensions to handle interaction strength, sampling effects and plasticity of interactions (see Supporting Information).

A key element of ELGRIN is the ability to measure the overall effects of biotic interactions on species distributions, which allows to summarise all local pairwise interactions in a single measure (i.e. $\beta_{l,co-abs}$ or $\beta_{l,co-pres}$). This measure can then be mapped, related to spatial layers to understand how the overall effect of biotic interactions vary in space and

in function of the environment or the ecosystem types. Importantly, this measure can also be carefully investigated at a given location in function of the constituent species, trophic groups, specialists vs generalists, connectance and so on. Interestingly, we can thus see our β_l estimates as an extended and more meaningful version of the famous checkerboard score or C-score (Stone & Roberts, 1990), which has been used to quantify local biotic interactions from co-occurrence pattern (e.g., Boulangeat *et al.*, 2012). The main advantage of ELGRIN over the C-score is that instead of inferring biotic interactions from co-occurrences, it quantifies the effects of biotic interactions on species occurrences, while knowing the interactions between species. Our approach is thus not comparable with recent developments on joint species distribution models (JSDMs) that relate species occurrences to environmental conditions, and provides a residual covariance matrix that could be interpreted on the light of missing predictors, mis-specifications and biotic interactions (Ovaskainen *et al.*, 2017; Zurell *et al.*, 2018). This matrix represents covariances between model residuals (the left-over from the environmental effects) and actually provides little information about biotic interactions (Zurell *et al.*, 2018). ELGRIN does not infer any residual covariance and directly accounts for the known interactions through the metanetwork. In JSDM, missing covariates will inevitably lead to spurious estimates of biotic interactions. In ELGRIN, the parameter a_l is supposed to capture most of the unexplained information that is independent of the biotic interactions. This parameter acts as a site random effect in mixed models and is expected to filter out the effects of missing covariates, although some remaining species-specific effects might still percolate into the $beta_l$ estimates.

In the presentation of ELGRIN and in our two case studies, we focused on a single interaction type (e.g. competition, mutualism or trophic interaction). When dealing with a single type of interaction, competition for instance, the modelling is explicit since we clearly understand the effect that one species can have on another species. However, it is more problematic but technically possible to handle a metanetwork composed of different types of interactions. They can have opposite effects such as competition (a species excludes other species) and facilitation (a species facilitates other species) and, since ELGRIN captures an overall impact of these interactions on the distributions at each location, interpreting ELGRIN's results can be tricky in this case. Additionally, it worth noting that since ELGRIN relies on a Markov random field, $G^\star$ is undirected. In other words, when the original metanetwork G encodes asymmetric interactions (e.g. predator-prey), they are then converted in undirected edges that only represent the presence of interactions (whatever their direction). It is thus critical to keep that in mind when interpreting the results of ELGRIN, and when merging different types of interactions together. The same issue happens when hoping to interpret the residual covariance matrix of JSDM through the lens of biotic interactions, since the values of the covariance matrix could reflect any type of interactions between species, that could be asymmetric or symmetric, or both. Finally, ELGRIN does not incorporate spatial dependencies between the locations. Incorporating or not spatial autocorrelation into the understanding of biodiversity patterns is a long standing question in ecology (F. Dormann *et al.*, 2007). Indeed, spatial dependency between locations might bias the statistical estimates since similar locations could be used as replicates. However, this bias might be an issue only if the underlying factors that created this spatial autocorrelation are not included into the environmental covariates. Meanwhile, the integration of the spatial dependency in the model is likely to

complicate the estimation procedure and could dramatically inflate the computing time.

In terms of further perspectives, we might wonder whether this model could be extended for prediction purposes. In principle, it is possible to draw presence/absence data from the model for different values of the environment variables. These different values could allow for predictions in space but also in time. In ELGRIN, known biotic interactions are also introduced in the modelling framework. However, something to keep in mind is that metanetwork will not change and will thus be considered as static and thus representative in space (or in time). If the metanetwork has not been built with that prediction perspective in mind, this might be an issue since we will miss interaction rewiring effects on species distributions. Instead, if the metanetwork is truly a potential metanetwork that tries to incorporate these potential interactions that have been observed yet (i.e. Maiorano *et al.*, 2020), it might be interesting to investigate how biotic interactions might further influence future species distributions in response to environmental changes.

Acknowledgements

Funding was provided by the French National Center for Scientific Research (CNRS) and the French National Research Agency (ANR) grant ANR-18-CE02-0010-01 EcoNet. VM would like to thank LECA laboratory for having hosted him in Chambéry.

References

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

Boulangeat, I., Gravel, D. & Thuiller, W. (2012) Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. *Ecology letters*, **15**, 584–593.

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**, 1636–1648.

Brémaud, P. (1999) *Markov chains: Gibbs fields, Monte Carlo simulation, and Queues*, volume 31. Springer.

Chalmandrier, L., Münkemüller, T., Gallien, L., De Bello, F., Mazel, F., Lavergne, S. & Thuiller, W. (2013) A family of null models to distinguish between environmental filtering and biotic interactions in functional diversity patterns. *Journal of Vegetation Science*, **24**, 853–864.

Chase, J.M. & Leibold, M.A. (2003) *Ecological niches: linking classical and contemporary approaches*. University of Chicago Press.

Cirtwill, A.R., Eklöf, A., Roslin, T., Wootton, K. & Gravel, D. (2019) A quantitative framework for investigating the reliability of empirical network construction. *Methods in Ecology and Evolution*, **10**, 902–911.

Connor, E.F. & Simberloff, D. (1979) The assembly of species communities: chance or competition? *Ecology*, **60**, 1132–1140.

Diamond, J.M. (1975) Assembly of species communities. *Ecology and evolution of communities*, pp. 342–444.

F. Dormann, C., M. McPherson, J., B. Araújo, M., Bivand, R., Bolliger, J., Carl, G., G. Davies, R., Hirzel, A., Jetz, W., Daniel Kissling, W. *et al.* (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, **30**, 609–628.

Gravel, D., Baiser, B., Dunne, J.A., Kopelke, J.P., Martinez, N.D., Nyman, T., Poisot, T., Stouffer, D.B., Tylianakis, J.M., Wood, S.A. & Roslin, T. (2019) Bringing Elton and Grinnell together: a quantitative framework to represent the biogeography of ecological interaction networks. *Ecography*, **42**, 401–415.

Guisan, A., Thuiller, W. & Zimmermann, N.E. (2017) *Habitat Suitability and Distribution Models: With Applications in R*. Ecology, Biodiversity and Conservation. Cambridge University Press.

Holt, R.D. (2020) Some thoughts about the challenge of inferring ecological interactions from spatial data. *Biodiversity Informatics*, **15**, 61–66.

Hutchinson, G.E. (1959) Homage to santa rosalia or why are there so many kinds of animals? *The American Naturalist*, **93**, 145–159.

Maiorano, L., Montemaggiori, A., O'Connor, L., Ficetola, G. & W., T. (2020) TETRA-EU 1.0: A species-level trophic meta-web of European tetrapods. *Global Ecology & Biogeography*.

Maiorano, L., Amori, G., Capula, M., Falcucci, A., Masi, M., Montemaggiori, A., Pottier, J., Psomas, A., Rondinini, C., Russo, D. *et al.* (2013) Threats from climate change to terrestrial vertebrate hotspots in Europe. *PLoS One*, **8**.

Morales-Castilla, I., Matias, M.G., Gravel, D. & Araújo, M.B. (2015) Inferring biotic interactions from proxies. *Trends in ecology & evolution*, **30**, 347–356.

Münkemüller, T. & Gallien, L. (2015) Virtualcom: a simulation model for eco-evolutionary community assembly and invasion. *Methods in Ecology and Evolution*, **6**, 735–743.

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

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T. & Abrego, N. (2017) How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, **20**, 561–576.

O'Connor, L.M.J., Pollock, L.J., Braga, J., Ficetola, G.F., Maiorano, L., Martinez-Almoyna, C., Montemaggiori, A., Ohlmann, M. & Thuiller, W. (2020) Unveiling the food webs of tetrapods across Europe through the prism of the Eltonian niche. *Journal of Biogeography*, **47**, 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**, 785–800.

Peres-Neto, P.R., Olden, J.D. & Jackson, D.A. (2001) Environmentally constrained null models: site suitability as occupancy criterion. *Oikos*, **93**, 110–120.

Ricklefs, R.E. (2008) Disintegration of the ecological community: American society of naturalists Sewall Wright Award Winner Address. *The American Naturalist*, **172**, 741–750.

Staniczenko, P.P., Sivasubramaniam, P., Suttle, K.B. & Pearson, R.G. (2017) Linking macroecology and community ecology: refining predictions of species distributions using biotic interaction networks. *Ecology letters*, **20**, 693–707.

Stone, L. & Roberts, A. (1990) The checkerboard score and species distributions. *Oecologia*, **85**, 74–79.

Storch, D. & Okie, J.G. (2019) The carrying capacity for species richness. *Global Ecology and Biogeography*, **28**, 1519–1532.

Thuiller, W., Pollock, L.J., Gueguen, M. & Münkemüller, T. (2015) From species distributions to meta-communities. *Ecology Letters*, **18**, 1321–1328.

Tylianakis, J.M. & Morris, R.J. (2017) Ecological networks across environmental gradients. *Annual Review of Ecology, Evolution, and Systematics*, **48**, 25–48.

Zurell, D., Pollock, L.J. & Thuiller, W. (2018) Do joint species distribution models reliably detect interspecific interactions from co-occurrence data in homogenous environments? *Ecography*, **41**, 1812–1819.

Variables	Ecological interpretation
$G^\star$	Metanetwork of interactions (undirected)
X_i^l	Presence/absence of species i at location l
W_l	Environmental covariates at location l
Parameters	
a_i	Prevalence of species i
a_l	Richness capacity (or expected number of species) at location l
b_i, c_i	Environmental (abiotic) parameters of species i
$\beta_{l,co-pres}$	Co-presence strength (or avoidance when < 0) at location l
$\beta_{l,co-abs}$	Co-absence strength (or avoidance when < 0) at location l

Table 1: Definition of variables and parameters of the Markov random field model ELGRIN.

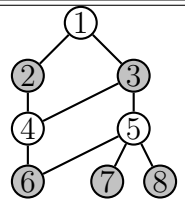
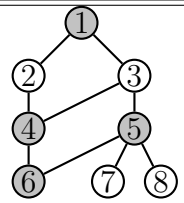
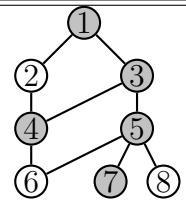
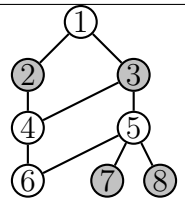
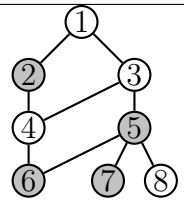
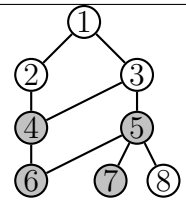
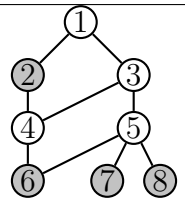
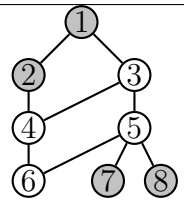
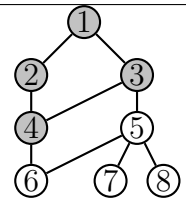
	$\beta_{l,co-pres} \ll 0$ (avoided co-presence)	$\beta_{l,co-pres} = 0$ (random presence)	$\beta_{l,co-pres} \gg 0$ (favored co-presence)
$\beta_{l,co-abs} \ll 0$ (avoided co-absence)			
$\beta_{l,co-abs} = 0$ (random absence)			
$\beta_{l,co-abs} \gg 0$ (favored co-absence)			

Table 2: Simplified view of the different behaviours of the model in function of the parameters $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$. The graph represents the metanetwork containing all potential interactions where species can be either present (gray node) or absent (white node) in a given location l leading to different estimated $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$. When $\beta_{l,co-pres} \ll 0$ or $\beta_{l,co-abs} \ll 0$, interacting species in the metanetwork tend to avoid each other: whenever one is absent, the other tend to be present and conversely. This situation favors a checkerboard pattern on the metanetwork. Reversely, whenever $\beta_{l,co-pres} \gg 0$ (resp. $\beta_{l,co-abs} \gg 0$), there are groups of interacting species that tend to be all present (resp. all absent), inducing sets of gray (resp. white) neighbour nodes in the metanetwork. Whenever $\beta_{l,co-pres} = 0$ or $\beta_{l,co-abs} = 0$, there are sets of interacting species whose states are independent from one another and thus purely random (the proportions of gray and white nodes are governed by the values of the parameters in the Grinnellian part of the model).

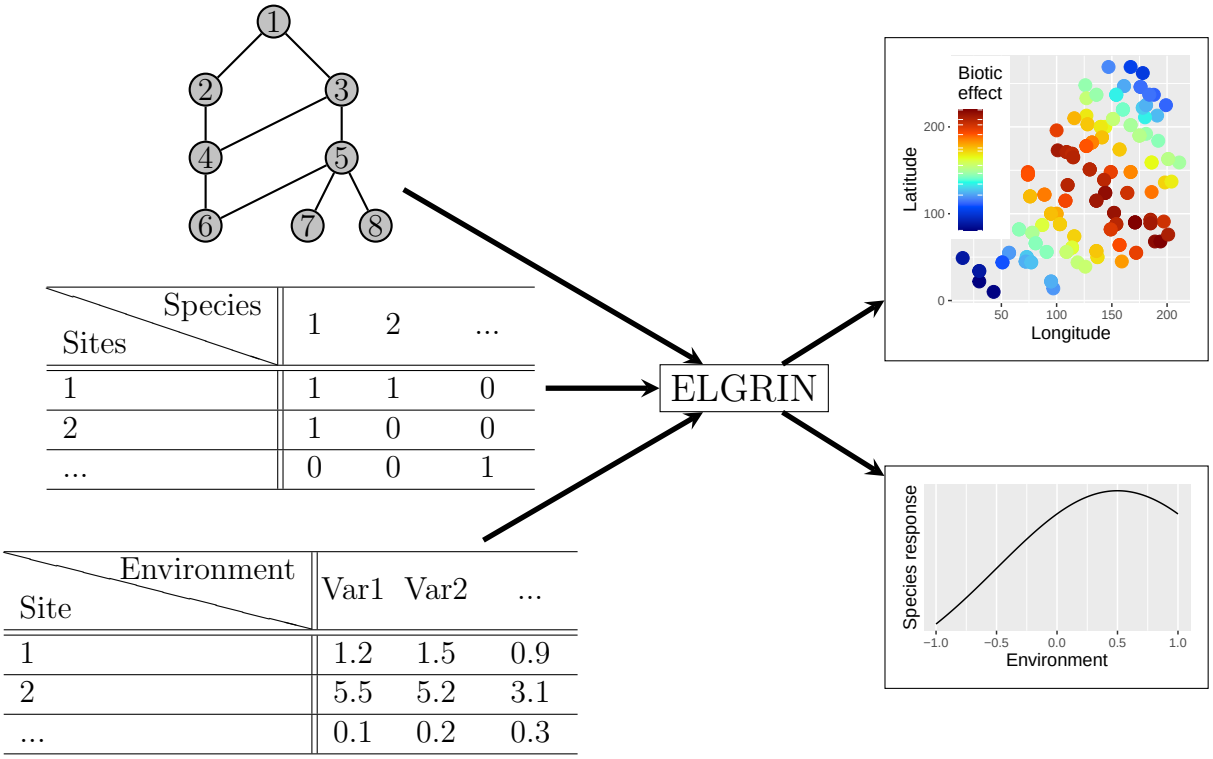


Figure 1: Schematic view of ELGRIN's framework. Given an interaction metanetwork, presence/absence data and environmental covariates for a set of sites, ELGRIN estimates the overall effect of biotic interactions on species distributions at every site, and the environmental response of each species along all sites.

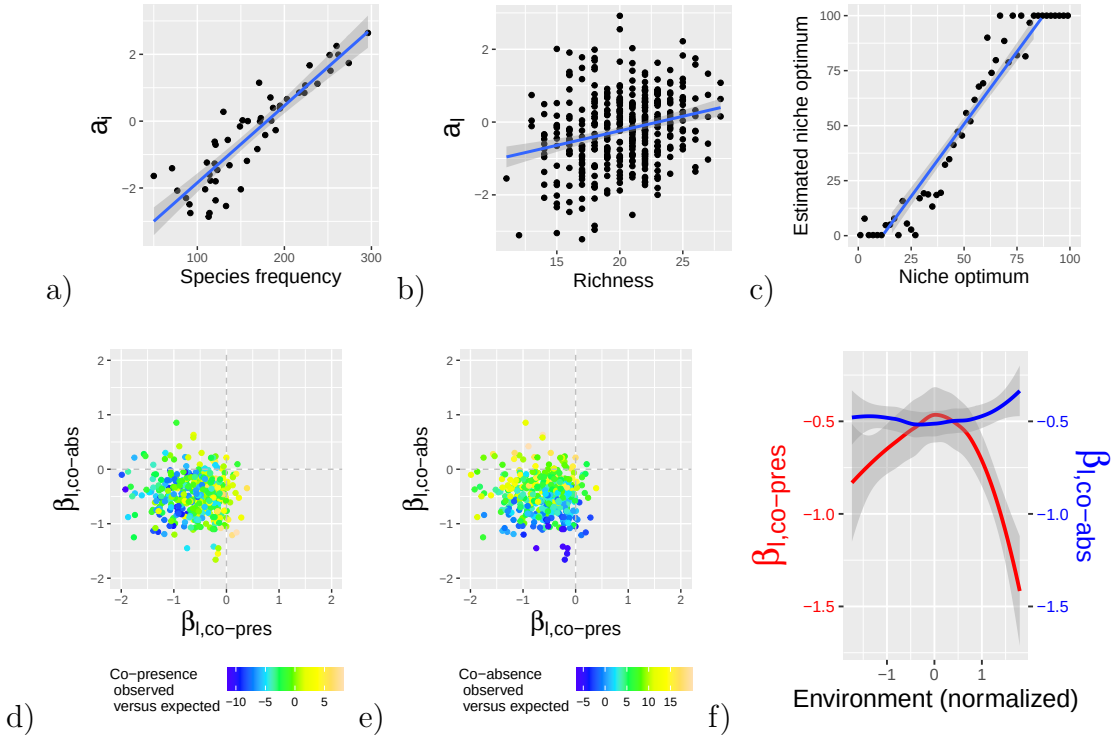


Figure 2: Results of ELGRIN on simulated ecological communities with competition. a) Estimated a_i compared to species frequency. b) Estimated a_l compared to species richness computed at each location. c) Estimated niche optimum (maximum value of the polynomial obtained with the estimated b_i and c_i) versus the optimum chosen for each species (see Material and Methods) d) Scatter plot between $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$ with a color scale showing the difference between the observed and expected number of co-presence of interacting species (see Supporting Information). e) Same as d) but for co-absence. f) Estimated $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$ compared to the environment.

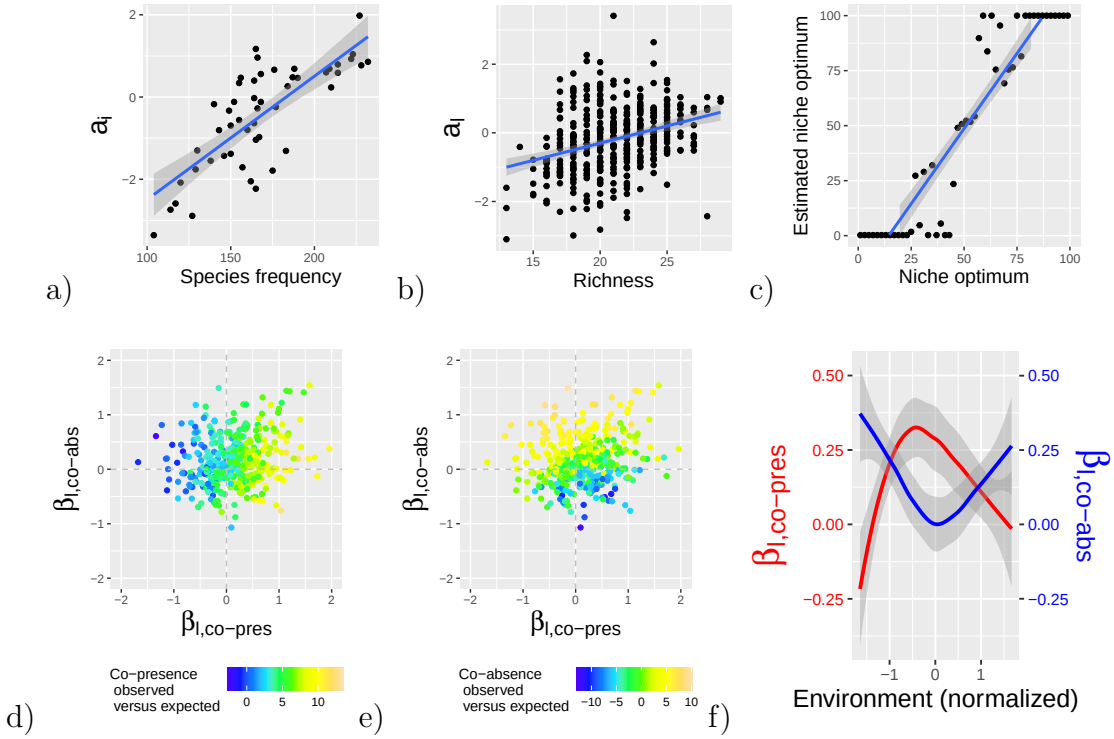


Figure 3: Same as Figure 2 with mutualism.

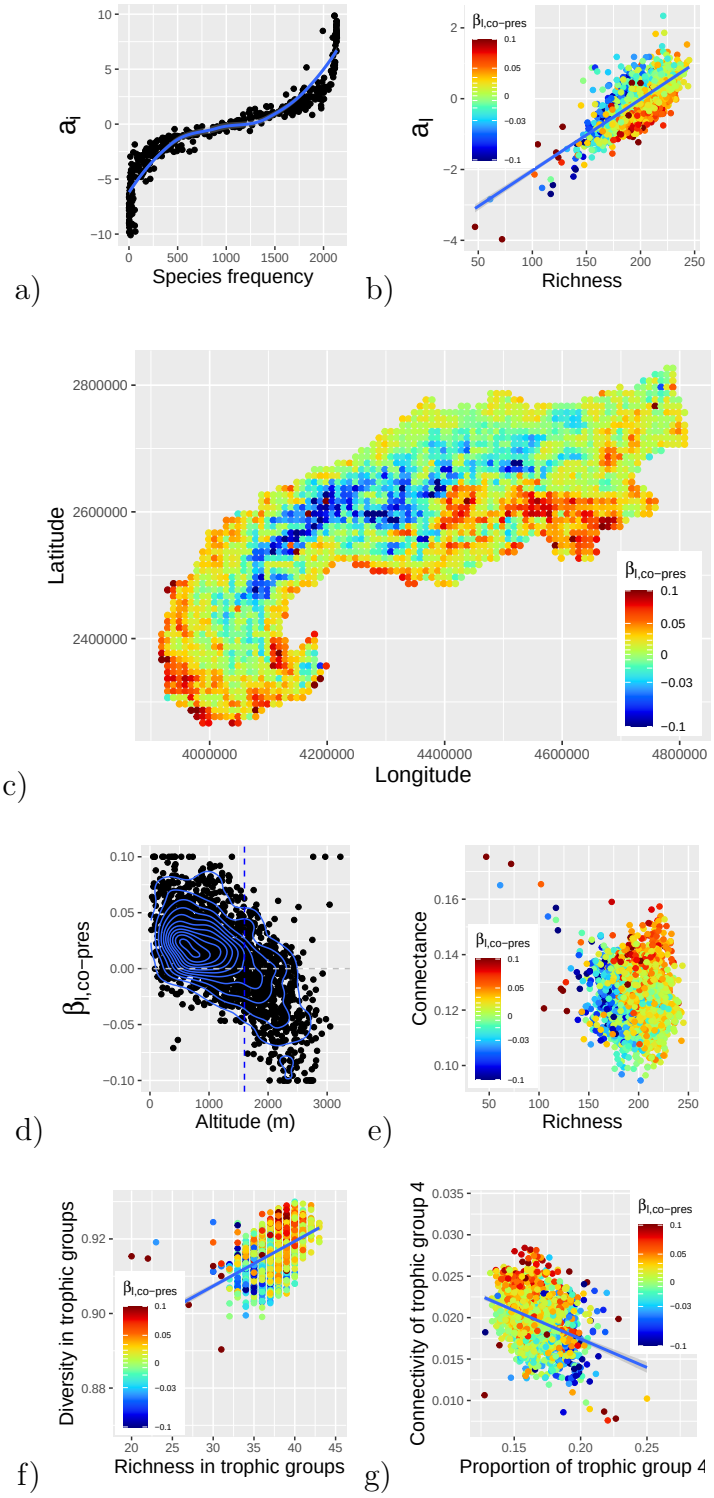

a)
Species frequency
a_i
b)
Richness
a_l
$\beta_{l,co-pres}$
c)
Latitude
Longitude
$\beta_{l,co-pres}$
d)
Altitude (m)
$\beta_{l,co-pres}$
e)
Richness
Connectance
$\beta_{l,co-pres}$
f)
Richness in trophic groups
Diversity in trophic groups
$\beta_{l,co-pres}$
g)
Proportion of trophic group 4
Connectivity of trophic group 4
$\beta_{l,co-pres}$

Figure 4: (Previous page.) Results of ELGRIN on the European tetrapods case study. In the following, the color scale indicates the $\beta_{l,co-pres}$ values. For the sake of representation, β values above 0.1 in absolute value were set to 0.1. a) Estimated a_i compared to species frequency. b) Estimated a_l compared to species richness computed at each location. c) Map of estimated $\beta_{l,co-pres}$ (one dot per location). d) Scatter plot between altitude and $\beta_{l,co-pres}$. e) Scatter plot between richness and connectance at each location. f) Scatter plot between richness and diversity in trophic groups defined in O'Connor *et al.* (2020). g) Scatter plot between the proportion of trophic group 4 defined in O'Connor *et al.* (2020) and its connectivity, as defined by its number of interactions over the total number of possible interactions.

Supporting information for "Quantifying the overall effect of biotic interactions on species communities along environmental gradients", by V. Miele, C. Matias, M. Ohlmann, G. Poggiato, S. Dray and W. Thuiller.

A Model extensions

A.1 Interaction strength

Besides the binary case, it is also possible to handle interaction strengths. An interaction strength can represent a frequency (e.g., the number of visits of a pollinator to a plant), an intensity (e.g., rate of predation, Berlow *et al.*, 2004) or a preference (e.g. modulating trophic links with known affinities of a predator to its preys).

We write $A^\star = (A^\star_{ij})_{i,j\in V^\star}$ the adjacency matrix of the graph $G^\star$. Now, each edge $(i,j) \in E^\star$ is modulated through the weight $A^\star_{ij}$ of the interaction. In this case, sub-equations (1b) and (1c) are replaced by

$$\beta_{l,co-pres} \sum_{(i,j)\in E^\star} A^\star_{ij} \mathbf{1}\{X^l_j = X^l_i = 1\}$$

and

$$\beta_{l,co-abs} \sum_{(i,j)\in E^\star} A^\star_{ij} \mathbf{1}\{X^l_j = X^l_i = 0\},$$

respectively.

A.2 Sampling effects

The random variables X^l_i that indicate the presence of species i at location l might not be exactly observed due to sampling effects. Here, we propose to account for these effects by assuming that each species $i \in V^\star$ is sampled with probability $p_{i,l} \in (0,1)$ at location $l \in \{1,\dots,L\}$. We therefore introduce a new set of random variables $Y^l_i, i \in V^\star, l \in \{1,\dots,L\}$ such that each Y^l_i only depends on X^l_i and is distributed as

$$\begin{aligned}\mathbb{P}(Y^l_i|X^l_i) &= p_{i,l}^{Y^l_i}(1-p_{i,l})^{1-Y^l_i} X^l_i + (1-X^l_i)(1-Y^l_i) \\ &= p_{i,l}^{X^l_i Y^l_i}(1-p_{i,l})^{X^l_i(1-Y^l_i)} \mathbf{1}\{(1-X^l_i)Y^l_i \neq 1\}. \end{aligned} \tag{A.1}$$

Specifically, whenever $X^l_i = 0$ (species i is absent from location l), species i cannot be observed at location l and $Y^l_i = 0$. Now, when $X^l_i = 1$ (species i is present at location l), it is observed ($Y^l_i = 1$) with sampling probability $p_{i,l}$ and unobserved ($Y^l_i = 0$) with probability $1-p_{i,l}$. The parameter $p_{i,l}$ must be given by the user considering three possible cases: species dependent sampling ($p_{i,l} := p_i; i \in V^\star$), location dependent sampling ($p_{i,l} := p_l; 1 \leq l \leq L$) or constant sampling ($p_{i,l} := p$). In this case, the X^l_i become latent variables as we only observe the Y^l_i's. The model turns out to be a hidden Markov random field.

A.3 Plasticity of interactions

Our model is able to assume that interactions are not necessarily induced by the presence/absence variables (we can assume that two species interact in a given location but not in another location). In this case, we consider a sample of observed graphs $G^1, \dots, G^L$ where each $G^l = (V^l, E^l)$ is such that $V^l \subset V^\star$. These graphs represent local interactions that are observed at the different locations $l \in \{1, \dots, L\}$. The main point here is that we assume that these interactions are sampled from the pool of potential interactions encoded in the metanetwork $G^\star$. Let $A^l = (A^l_{i,j})_{i,j \in V^l}$ denote the adjacency matrix of the graph G^l. We assume that any two species that are observed and that can potentially interact (i.e. are linked in the metanetwork $G^\star$) do effectively interact at location l with a probability that depends only on these two species. Namely for any $(i,j) \in E^\star$, conditional on the fact that two species $i, j \in V^\star$ were observed at location l (namely $Y^l_i Y^l_j = 1$), we set

$$A^l_{i,j} | Y^l_i Y^l_j = 1 \sim \mathcal{B}(\epsilon_{ij}), \tag{A.2}$$

and $A^l_{i,j} \equiv 0$ whenever $(i,j) \notin E^\star$ or $Y^l_i = 0$ or $Y^l_j = 0$. This additional parameter $\epsilon = \{\epsilon_{i,j}\}_{i,j \in V^\star}$ allows us to handle interaction plasticity directly in the model.

B Mathematical details on the model

B.1 Identifying the parameters of the Gibbs distribution

We first address the issue of the identifiability of the parameters from the Gibbs distribution. In what follows, we focus on the case of a binary metanetwork $G^\star$. However, our results remain valid in the weighted case, where degrees are replaced by weighted degrees and the cardinality $|E^\star|$ becomes the total sum of the weights.

Let us focus on the model with no covariates ($W_l = 0$) and consider for each location $l \in \{1, \dots, L\}$ the maps $\psi_l = (\{a_i\}_i, a_l, \beta_{l,co-pres}, \beta_{l,co-abs}) \mapsto \mathbb{P}_{\psi_l}$, where

$$\mathbb{P}_{\psi_l}(\{X^l_i\}_{i \in V^\star}) = \frac{1}{Z_{\psi_l}} \exp\Big(\sum_{i \in V^\star} (a_i + a_l) X^l_i + \beta_{l,co-pres} \sum_{(i,j) \in E^\star} 1\{X^l_j = X^l_i = 1\} + \beta_{l,co-abs} \sum_{(i,j) \in E^\star} 1\{X^l_j = X^l_i = 0\} \Big).$$

For any $\psi = (\{a_i\}_{i,l}, \{a_l, \beta_{l,co-pres}, \beta_{l,co-abs}\}_l)$ we also define the global probability distribution $\mathbb{P}_\psi$ as follows

$$\mathbb{P}_\psi(\{X^l_i\}_{i \in V^\star; 1 \le l \le L}) = \prod_{l=1}^{L} \mathbb{P}_{\psi_l}(\{X^l_i\}_{i \in V^\star}).$$

Proposition 1 (Identifying linear combinations of the parameter)**.** *In the model without covariate ($W_l = 0$, for any l), the probability distribution $\mathbb{P}_\psi$ uniquely defines the quantities*

$$\beta_{l,co-pres} + \beta_{l,co-abs}, \tag{B.1}$$

$$\textit{and } a_i + a_l + \beta_{l,co-pres} \deg_{G^\star}(i) \textit{ or equivalently } a_i + a_l - \beta_{l,co-abs} \deg_{G^\star}(i), \tag{B.2}$$

for any $i \in V^\star, l \in \{1, \dots, L\}$, where $\deg_{G^\star}(i)$ is the degree of species i in the metanetwork $G^\star$. Moreover, if there exist 2 species $1 \leq i, j \leq N$ such that $\deg_{G^\star}(i) \neq \deg_{G^\star}(j)$ in $G^\star$, then the probability distribution $\mathbb{P}_\psi$ uniquely defines the additional quantities

$$\beta_{l,co-abs} - \beta_{l',co-abs} \text{ or equivalently } \beta_{l,co-pres} - \beta_{l',co-pres}, \tag{B.3}$$

$$\text{and } a_l - a_{l'}, \tag{B.4}$$

for any $l, l' \in \{1, \dots, L\}$.

Proof. Let us denote $\alpha_{i,l} = a_i + a_l$. As $\mathbb{P}_{\psi_l}$ is a marginal of $\mathbb{P}_\psi$, we start by fixing the location $l \in \{1, \dots, L\}$ and consider the probabilities of specific configurations at this location. We let X^l_{-i} denote the set $\{X^l_j; j \in V^\star, j \neq i\}$. From the knowledge of $\mathbb{P}_\psi$, we obtain for $l \in \{1, ..., L\}$ and $i \in V^\star$ the quantities

$$\begin{aligned}
s^l_0 &:= \log \mathbb{P}_{\psi_l}(\{0, ..., 0\}) = -\log(Z_{\psi_l}) + |E^\star|\beta_{l,co-abs} \\
s^l_1 &:= \log \mathbb{P}_{\psi_l}(\{1, ..., 1\}) = -\log(Z_{\psi_l}) + \sum_i \alpha_{i,l} + |E^\star|\beta_{l,co-pres} \\
s^{i,l}_{10} &:= \log \mathbb{P}_{\psi_l}(\{X^l_i = 1, X^l_{-i} = 0\}) = -\log(Z_{\psi_l}) + \alpha_{i,l} + \beta_{l,co-abs}(|E^\star| - \deg_{G^\star}(i)) \\
s^{i,l}_{01} &:= \log \mathbb{P}_{\psi_l}(\{X^l_i = 0, X^l_{-i} = 1\}) = -\log(Z_{\psi_l}) + \sum_{j \neq i} \alpha_{j,l} + \beta_{l,co-pres}(|E^\star| - \deg_{G^\star}(i)),
\end{aligned}$$

where $|E^\star|$ is the cardinality of the set $E^\star$. It follows

$$\begin{aligned}
r^l_1 &:= s^l_1 - s^l_0 = \sum_i \alpha_{i,l} + |E^\star|(\beta_{l,co-pres} - \beta_{l,co-abs}) \\
r^{i,l}_2 &:= s^{i,l}_{10} - s^l_0 = \alpha_{i,l} - \beta_{l,co-abs} \deg_{G^\star}(i) \\
r^{i,l}_3 &:= s^{i,l}_{01} - s^l_0 = \sum_{j \neq i} \alpha_{j,l} + (\beta_{l,co-pres} - \beta_{l,co-abs})|E^\star| - \beta_{l,co-pres} \deg_{G^\star}(i).
\end{aligned}$$

From these equations, we uniquely obtain

$$\begin{aligned}
t^{i,l}_1 &:= r^l_1 - r^{i,l}_3 = \alpha_{il} + \beta_{l,co-pres} \deg_{G^\star}(i) \\
t^{i,l}_2 &:= r^l_1 - r^{i,l}_2 - r^{i,l}_3 = (\beta_{l,co-abs} + \beta_{l,co-pres}) \deg_{G^\star}(i).
\end{aligned}$$

As a consequence, as soon as there is at least one edge in the metanetwork $G^\star$ (inducing at least on species i with $\deg_{G^\star}(i) \neq 0$) we can obtain the quantities $\beta_{l,co-abs} + \beta_{l,co-pres}$ (recall that $\deg_{G^\star}(i)$ is known) as well as $\alpha_{i,l} + \beta_{l,co-pres} \deg_{G^\star}(i)$ uniquely from the distribution $\mathbb{P}_\psi$. Note also that combining the knowledge of these two quantities, the second is equivalent to knowing $\alpha_{i,l} - \beta_{l,co-abs} \deg_{G^\star}(i)$.

Now, let us recall that $\alpha_{i,l} = a_i + a_l$. For two different locations $l \neq l'$, we have access to

$$t^{i,l}_1 - t^{i,l'}_1 = a_l - a_{l'} + (\beta_{l,co-pres} - \beta_{l',co-pres}) \deg_{G^\star}(i).$$

We now assume that there exist two species $1 \leq i, j \leq N$ such that $\deg_{G^\star}(i) \neq \deg_{G^\star}(j)$ in $G^\star$ and obtain (B.3) as follows

$$\beta_{l,co-pres} - \beta_{l',co-pres} = (t^{i,l}_1 - t^{i,l'}_1 - t^{j,l}_1 + t^{j,l'}_1)[\deg_{G^\star}(i) - \deg_{G^\star}(j)]^{-1}.$$

Combining this with (B.1), it is equivalent to the unique identification of $\beta_{l,co-abs} - \beta_{l',co-abs}$. Finally, going back to $t_1^{i,l} - t_1^{i,l'}$ we uniquely obtain $a_l - a_{l'}$. □

Definition 1 (Equivalence class)**.** *For any parameter* $\psi = (\{a_i\}_i, \{a_l, \beta_{l,co-pres}, \beta_{l,co-abs}\}_l)$*, its equivalence class* $[\psi]$ *is defined as*

$$[\psi] := \{(\{a_i + \gamma \deg_{G^\star}(i) - \delta\}_i, \{a_l + \delta, \beta_{l,co-pres} - \gamma, \beta_{l,co-abs} + \gamma\}_l); \gamma \in \mathbb{R}, \delta \in \mathbb{R}\}.$$

Corollary 1 (Parameter identifiability up to the equivalence class)**.** *In the model without covariate (*$W_l = 0$*, for any* l*) and assuming that there exist 2 species* $1 \leq i, j \leq N$ *such that* $\deg_{G^\star}(i) \neq \deg_{G^\star}(j)$ *in* $G^\star$*, we have that whenever there are two parameter values* $\psi, \tilde{\psi}$ *such that* $\mathbb{P}_\psi = \mathbb{P}_{\tilde{\psi}}$*, then* $\tilde{\psi} \in [\psi]$*. In other words, the equality* $\mathbb{P}_\psi = \mathbb{P}_{\tilde{\psi}}$ *implies that there exist real values* $\gamma, \delta \in \mathbb{R}$ *such that for any* $i \in V^\star$ *and* $l \in \{1, \ldots, L\}$*, we have*

$$\begin{aligned} \tilde{a}_i &= a_i + \gamma \deg_{G^\star}(i) - \delta \\ \tilde{a}_l &= a_l + \delta \\ \tilde{\beta}_{l,co-pres} &= \beta_{l,co-pres} - \gamma \\ \tilde{\beta}_{l,co-abs} &= \beta_{l,co-abs} + \gamma. \end{aligned}$$

Proof. Assume that $\mathbb{P}_\psi = \mathbb{P}_{\tilde{\psi}}$ and define for any location $l \in \{1, \ldots, L\}$ the quantity $\gamma_l := \beta_{l,co-pres} - \tilde{\beta}_{l,co-pres}$. Then we know from Proposition 1 that

$$\begin{aligned} \beta_{l,co-abs} + \beta_{l,co-pres} &= \tilde{\beta}_{l,co-abs} + \tilde{\beta}_{l,co-pres} \\ \tilde{a}_i + \tilde{a}_l + \tilde{\beta}_{l,co-pres} \deg_{G^\star}(i) &= a_i + a_l + \beta_{l,co-pres} \deg_{G^\star}(i). \end{aligned}$$

This induces that

$$\begin{aligned} \gamma_l &= \tilde{\beta}_{l,co-abs} - \beta_{l,co-abs} \\ \text{and} \quad \tilde{a}_i + \tilde{a}_l &= a_i + a_l + \gamma_l \deg_{G^\star}(i). \end{aligned}$$

Let us further prove that γ_l does not depend on l. From Proposition 1 and the additional assumption that at least two species have different degrees in the metanetwork, we have for any locations $l, l' \in \{1, \ldots, L\}$,

$$\beta_{l,co-pres} - \beta_{l',co-pres} = \tilde{\beta}_{l,co-pres} - \tilde{\beta}_{l',co-pres} = \beta_{l,co-pres} - \beta_{l',co-pres} - \gamma_l + \gamma_{l'},$$

which implies that $\gamma_l = \gamma_{l'}$ for any pair of locations. Finally, let us define for any location and any species

$$\delta_l = \tilde{a}_l - a_l \quad \text{and} \quad \delta_i = \tilde{a}_i - a_i.$$

We have established that $\delta_l + \delta_i = \gamma \deg_{G^\star}(i)$. This implies that δ_l is constant through locations and equal to some δ. This concludes the proof. □

Corollary 1 tells us that the model parameter is identifiable up to the equivalence class in Definition 1. Now, we introduce our choice of the representative parameter in this class.

Proposition 2 (Choosing a representative)**.** *In the model without covariate ($W_l = 0$, for any l) and assuming that there exist 2 species $1 \leq i, j \leq N$ such that $\deg_{G^\star}(i) \neq \deg_{G^\star}(j)$ in $G^\star$, for any parameter value $\tilde{\psi}$, it is possible to choose a unique representative $\psi \in [\tilde{\psi}]$ such that the estimated linear regression coefficients of the set of parameters $\{a_i\}_i$ over the degrees $\{\deg_{G^\star}(i)\}_i$ are equal to 0, namely*

$$(\hat{\gamma}, \hat{\delta}) := \inf_{(\gamma,\delta)\in\mathbb{R}^2} \sum_{i\in V^\star} (a_i - \gamma \deg_{G^\star}(i) - \delta)^2$$

satisfies $(\hat{\gamma}, \hat{\delta}) = (0, 0)$.

Proof. Fix a parameter value $\tilde{\psi}$ and consider the linear regression of the set of parameters $\{\tilde{a}_i\}_i$ over the degrees $\{\deg_{G^\star}(i)\}_i$, namely

$$(\tilde{\gamma}, \tilde{\delta}) := \inf_{(\gamma,\delta)\in\mathbb{R}^2} \sum_{i\in V^\star} (\tilde{a}_i - \gamma \deg_{G^\star}(i) - \delta)^2.$$

Then by setting the parameter $\psi = (\{a_i\}_{i,l}, \{a_l, \beta_{l,co-pres}, \beta_{l,co-abs}\}_l)$ as

$$\begin{aligned} a_i &:= \tilde{a}_i - \tilde{\gamma} \deg_{G^\star}(i) - \tilde{\delta}; \\ a_l &:= \tilde{a}_l + \tilde{\delta}; \\ \beta_{l,co-pres} &:= \tilde{\beta}_{l,co-pres} + \tilde{\gamma} \deg_{G^\star}(i) \\ \beta_{l,co-abs} &:= \tilde{\beta}_{l,co-abs} - \tilde{\gamma} \deg_{G^\star}(i) \end{aligned}$$

(for any i, l), we know from Definition 1 that $\psi \in [\tilde{\psi}]$ and also by definition, the estimated values

$$(\hat{\gamma}, \hat{\delta}) := \inf_{(\gamma,\delta)\in\mathbb{R}^2} \sum_{i\in V^\star} (a_i - \gamma \deg_{G^\star}(i) - \delta)^2$$

will now satisfy $(\hat{\gamma}, \hat{\delta}) = (0, 0)$. □

Remark 1. *The choice of the representative parameter given by Proposition 2 is such that the response of species i to the environment does not depend on its degree in the metanetwork and thus on its number of interactions. This is a natural choice to separate the Grinellian part from the Eltonian one in our model. Note that this representative parameter is the one we rely on when interpreting the model. Thus, when we comment the behaviour of the model with respect to different values of its parameter, we always rely on this specific reprensentative.*

In general, when obtaining an estimate of the parameter, we would use a simple linear regression (as described in Proposition 2) to obtain its representative in the equivalence class. In our simulations and applications to datasets, it turns out that because of the initialisation of our algorithm, we never need to perform this additional regression step and the output of the algorithm directly is the representative from Proposition 2.

B.2 Hidden Markov random field and its interpretation

We discuss here the model in its full generality, including possible weights on the metanetwork, sampling effects and plasticity of interactions. We thus have $\mathbf{X} := \{\mathbf{X}^l\}_{1\leq l\leq L} =$

$\{X_i^l\}_{i\in V^\star,1\leq l\leq L}$ (resp. $\mathbf{Y} := \{\mathbf{Y}^l\}_{1\leq l\leq L} = \{Y_i^l\}_{i\in V^\star,1\leq l\leq L}$ and $\mathbf{A} := \{A^l\}_{1\leq l\leq L} = \{A_{i,j}^l\}_{i,j\in V^l,1\leq l\leq L}$) denoting the set of true occurrence variables (resp. observed occurrences and observed interactions). We assume that we observe $(\mathbf{Y}, \mathbf{A})$, while $\mathbf{X}$ are latent random variables.

A Gibbs distribution specifies the joint associations between the species occurrence variables $\{X_i^l\}_{i\in V^\star}$, as follows

$$\mathbb{P}_{\psi_l}(\{X_i^l\}_{i\in V^\star}) = \frac{1}{Z_{\psi_l}} \exp\Big(\sum_{i\in V^\star} [a_l + a_i + W_l^\intercal b_i + (W_l^2)^\intercal c_i] X_i^l + \beta_{l,co-pres} \sum_{(i,j)\in E^\star} A_{ij}^\star \mathbf{1}\{X_j^l = X_i^l = 1\} + \beta_{l,co-abs} \sum_{(i,j)\in E^\star} A_{ij}^\star \mathbf{1}\{X_j^l = X_i^l = 0\}\Big). \tag{B.5}$$

First note that the normalizing constant Z_{ψ_l} is given by

$$Z_{\psi_l} = \sum_{\{x_i\}_{i\in V^\star}\in\{0,1\}^N} \exp\Big(\sum_{i\in V^\star} [a_l + a_i + W_l^\intercal b_i + (W_l^2)^\intercal c_i] x_i + \beta_{l,co-pres} \sum_{(i,j)\in E^\star} A_{ij}^\star 1\{x_i = x_j = 1\} + \beta_{l,co-abs} \sum_{(i,j)\in E^\star} A_{ij}^\star 1\{x_i = x_j = 0\}\Big).$$

In general, this normalising constant Z_{ψ_l} cannot be computed due to the large number of possible configurations appearing in the sum. The statistical inference procedure needs to deal with that.

The model interpretation strongly builds on the *Markov property*, a fundamental characteristic of Markov random fields. Let us denote $\mathcal{N}_i^\star$ the set of species $j \in V^\star$ that are connected to i in the graph $G^\star$ (namely $\{j \in V^\star; A_{ij}^\star \neq 0\}$) and $X_{\mathcal{N}_i^\star}^l$, the set of corresponding random variables X_j^l for $j \in \mathcal{N}_i^\star$. We also recall that X_{-i}^l denotes the set $\{X_j^l; j \in V^\star, j \neq i\}$. Then, under the *Markov property* we have

$$\mathbb{P}_{\psi_l}(X_i^l | X_{-i}^l) = \mathbb{P}_{\psi_l}(X_i^l | X_{\mathcal{N}_i^\star}^l) \propto \exp\Big([a_l + a_i + W_l^\intercal b_i + (W_l^2)^\intercal c_i] X_i^l + \beta_{l,co-pres} \sum_{j\in\mathcal{N}_i^\star} A_{ij}^\star \mathbf{1}\{X_j^l = X_i^l = 1\} + \beta_{l,co-abs} \sum_{j\in\mathcal{N}_i^\star} A_{ij}^\star \mathbf{1}\{X_j^l = X_i^l = 0\}\Big), \tag{B.6}$$

where $\propto$ means proportional (equals up to a normalising constant). More specifically, it means that the conditional occurrence probability of a species i is modulated by the occurrences of the species interacting with i in $G^\star$. In other words, a species presence only depends on abiotic environment and on the species it interacts with. Moreover, the presence/absence variables of any two species are not statistically independent of each other if $G^\star$ is connected (namely, if there exists a path between any two species in $G^\star$). Meanwhile, if $G^\star$ has more than one connected component (i.e. disconnected compartments, Krause *et al.*, 2003), then the presence/absence of species in different components are independent. The Markov property is the cornerstone idea of our model. Indeed, the conditional probabilities of each random variable is specified through (B.6) and is rooted on the idea that the occurrence of a species i at location l depends both on a suitability term, specific to that species and the

local environment, and on the presence/absence of other species with whom it interacts (as encoded in the metanetwork). From this set of conditional probabilities, the Hammersley-Clifford theorem (Besag, 1974) ensures that there exists a proper joint distribution on the random variables $\{X_i^l\}_{i,l}$ and that it is given by equation (B.5).

Now, the observed species occurrence variables $Y_i^l, i \in V^\star, l \in \{1, \dots, L\}$ are distributed such that each Y_i^l only depends on X_i^l (the true occurrence variable) with

$$\begin{aligned}\mathbb{P}(Y_i^l|X_i^l) &= p_{i,l}^{Y_i^l}(1-p_{i,l})^{1-Y_i^l}X_i^l + (1-X_i^l)(1-Y_i^l) \\ &= p_{i,l}^{X_i^l Y_i^l}(1-p_{i,l})^{X_i^l(1-Y_i^l)}\mathbf{1}\{(1-X_i^l)Y_i^l \neq 1\}. \end{aligned} \tag{B.7}$$

In what follows, we choose to impose that the sampling parameters $p_{i,l}$ are set by the user. A consequence of this is that the quantity (B.7) will play no role in the inference procedure. Indeed, it is a constant quantity with respect to the parameter. Finally we set

$$A_{i,j}^l | Y_i^l Y_j^l = 1 \sim \mathcal{B}(\epsilon_{ij}), \tag{B.8}$$

and $A_{i,j}^l \equiv 0$ whenever $(i,j) \notin E^\star$ or $Y_i^l = 0$ or $Y_j^l = 0$.

Building on Equations (B.7) and (B.8), we first obtain the conditional distribution of all observations $(\mathbf{Y}, \mathbf{A})$ given the latent variables $\mathbf{X}$

$$\begin{aligned}\mathbb{P}_\phi(\mathbf{Y}, \mathbf{A}|\mathbf{X}) &= \prod_{l=1}^{L} \mathbb{P}_\phi(A^l|\mathbf{Y}^l)\mathbb{P}(\mathbf{Y}^l|\mathbf{X}^l) \\ &= \prod_{l=1}^{L}\prod_{i\in V^\star}\left[p_{i,l}^{X_i^l Y_i^l}(1-p_{i,l})^{X_i^l(1-Y_i^l)}1\{(1-X_i^l)Y_i^l \neq 1\}\right] \times \prod_{(i,j)\in E^\star} \epsilon_{ij}^{Y_i^l Y_j^l A_{i,j}^l}(1-\epsilon_{ij})^{Y_i^l Y_j^l (1-A_{i,j}^l)}.\end{aligned}$$

Here, the parameter $\epsilon = \{\epsilon_{ij}\}_{i,j\in V^\star}$ drives the distribution of the observation process from the latent one.

Finally, our model is obtained by combining this with Equation (B.5) for the distribution of the latent variables $\mathbf{X}$. Thus the global model is parameterised by $\theta = \{\theta_l\}_{1\leq l\leq L}$ where each $\theta_l = (\psi_l, \epsilon)$. This amounts to the following sets of parameters

$$(\{a_i, b_i, c_i\}_{i\in V^\star}, \{a_l, \beta_{l,co-abs}, \beta_{l,co-pres}\}_{1\leq l\leq L}, \{\epsilon_{ij}\}_{i,j\in V^\star})$$

so there are $3N + 3L + N(N-1)$ parameters when the observed graphs A^l are directed (and $3N+3L+N(N-1)/2$ when the observed graphs A^l are undirected) compared with $N(N-1)L$ observations. However note that in the model inference (see next section), the parameters ϵ_{ij} are pre-estimated (see Equation (C.1)) and do not appear in the main inference algorithm (see Algorithm 1). In what follows, we often use the notation

$$\alpha_{i,l} = a_i + a_l + W_l^\intercal b_i + (W_l^2)^\intercal c_i.$$

A chain graph (Lauritzen, 1996) describing the dependencies among the random variables in this model is given in Figure 5.

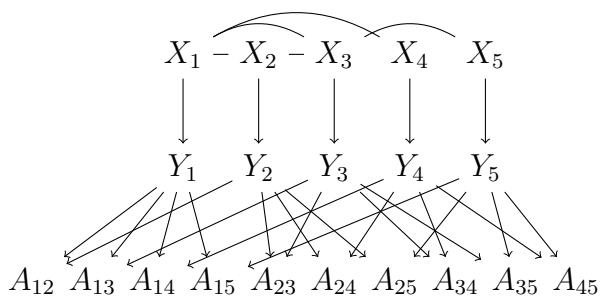


Figure 5: Example of a metanetwork $G^\star$ (relations among the random variables $\{X_i\}_{i\in V^\star}$ with $V^\star = \{1,\dots,5\}$, on the top row) and induced dependency chain graph of all the variables in the model for one observed undirected graph $A = (A_{ij})_{i<j}$ with no self-loops.

C Model inference

We present the inference procedure in the most general case, namely with weighted metanetwork, sampling effects and plasticity of interactions. This means that our inference procedure takes place in the context of a hidden Markov random field model.

C.1 Likelihood

The log-likelihood for observing independent interaction graphs $G^1,\dots,G^L$ at the different locations (and thus species occurrences variables ; indeed it is equivalent to observe $G^1,\dots,G^L$ or $(\mathbf{Y}^1, A^1,\dots,\mathbf{Y}^L, A^L)$) in this model is given by

$$\ell_{n,L}(\theta) = \sum_{l=1}^{L} \log \mathbb{P}_{\theta_l}(G^l),$$

where

$$\mathbb{P}_{\theta_l}(G^l) = \sum_{\{x_i^l\}_{i\in V^\star}\in\{0,1\}^N} \mathbb{P}_{\theta_l}(G^l, \{X_i^l = x_i^l; i\in V^\star\}).$$

As usual in latent variables models, this sum over all possible configurations $\{x_i^l\}_{i\in V^\star} \in \{0,1\}^N$ cannot be computed (unless N is really small). The inference procedure in latent variable models generally relies on the Expectation-Maximisation (EM) algorithm (Dempster *et al.*, 1977). In the context of hidden Markov random fields (HMRF), many difficulties arise that prevent from using this simple strategy.

The complete log-likelihood $\ell^c_{n,L}(\theta)$ contribution of all observations and all latent configurations is given by

$$\begin{aligned}\ell^c_{n,L}(\theta) &:= \log \mathbb{P}_\theta(\mathbf{X}, G^l,\dots,G^L) = \sum_{l=1}^{L} \log \mathbb{P}_{\theta_l}(\mathbf{X}^l, \mathbf{Y}^l, A^l)\\ &= \sum_{l=1}^{L} \log \mathbb{P}_{\psi_l}(\mathbf{X}^l) + \sum_{l=1}^{L}\sum_{i\in V^\star} \log \mathbb{P}(Y_i^l|X_i^l) + \sum_{l=1}^{L}\sum_{i,j\in V^l} \log \mathbb{P}_\phi(A_{i,j}^l|Y_i^l, Y_j^l).\end{aligned}$$

This can be written as

$$\ell^c_{n,L}(\theta) = \sum_{l=1}^{L}\sum_{i\in V^\star}\log(1-\alpha_{i,l}) + \sum_{l=1}^{L}\sum_{i\in V^\star} X_i^l \log\left(\frac{\alpha_{i,l}}{1-\alpha_{i,l}}\right) + \sum_{l=1}^{L} c \sum_{(i,j)\in E^\star} A^\star_{ij} 1\{X_j^l = X_i^l = 1\}$$
$$+ \sum_{l=1}^{L}\beta_{l,co-abs}\sum_{(i,j)\in E^\star} A^\star_{ij}1\{X_j^l = X_i^l = 0\} - \sum_{l=1}^{L}\log(Z_{\psi_l})$$
$$+ \sum_{i\in V^\star}\sum_{l=1}^{L} X_i^l\Big\{Y_i^l\log(p_{i,l}) + (1-Y_i^l)\log(1-p_{i,l})\Big\}$$
$$+ \sum_{i,j\in V^\star}\sum_{l=1}^{L} Y_i^l Y_j^l\Big\{A^l_{i,j}\log\epsilon_{ij} + (1-A^l_{i,j})\log(1-\epsilon_{ij})\Big\}.$$

Here, we restrict our attention to complete datasets $(\mathbf{X}^l, G^l)$ which are compatible, in the sense that whenever $X_i^l = 0$ we also have $Y_i^l = 0$. Otherwise the probability above is 0 and its log is $-\infty$.

C.2 Estimating the frequency of interactions

First, it is important to note that a consequence of the dependence among the $\{X_i^l\}_{i\in V^\star}$ is that the random variables $A^l_{i,j}$ and $A^l_{i'j'}$ are dependent. However, this dependency is entirely carried by the species observations Y_i^l's (which themselves are dependent through the species latent presences X_i^l's). In other words, we have $\mathbb{P}_\phi(A^l|\mathbf{Y}^l, \mathbf{X}^l) = \mathbb{P}_\phi(A^l|\mathbf{Y}^l)$. A consequence is that the parameters ϵ that describe the graph distribution are directly estimated from the data. While the sampling parameters and the random field ones ($\beta_{l,co-abs}, \beta_{l,co-pres}$ and $\alpha_{i,l}$'s) require a sophisticate inference procedure, the ϵ_{ij} parameters are directly estimated by the frequencies

$$\hat{\epsilon}_{ij} = \frac{\sum_{l=1}^{L} A^l_{ij}}{\sum_{l=1}^{L} Y_i^l Y_j^l}. \tag{C.1}$$

Here, the normalising term $\sum_{l=1}^{L} Y_i^l Y_j^l$ is simply the number of simultaneous observations of species i and j across the L different locations, while the numerator counts the number of observed interactions between those species across locations.

C.3 Inference of the random field parameters with simulated field algorithm

Now, we focus on the estimation of random field parameters $\beta_{l,co-abs}, \beta_{l,co-pres}$ and $\alpha_{i,l}$'s. A classical `EM` algorithm would consist in (iteratively) optimising with respect to $\psi = \{\psi_l\}_{1\le l\le L}$ the quantity

$$Q(\psi) = \sum_{l=1}^{L}\mathbb{E}\big(\log\mathbb{P}_{\psi_l}(\mathbf{X}^l, \mathbf{Y}^l)|\psi_l^{(t)}, \mathbf{Y}^l\big) = \sum_{l=1}^{L}\mathbb{E}\big(\log\mathbb{P}_{\psi_l}(\mathbf{X}^l)|\psi_l^{(t)}, \mathbf{Y}^l\big), \tag{C.2}$$

computed with the current value of the parameter $\psi^{(t)} = \{\psi_l^{(t)}\}_{1\leq l\leq L}$. The above quantity has many drawbacks: first it contains the partition functions Z_{ψ_l} that are unknown and cannot be computed. Second, the conditional distribution of $\mathbf{X}^l$ given $\mathbf{Y}^l$ has an intricate dependency structure and thus may not be computed (in fact it is also a Markov random field).

We thus follow the *simulated field algorithm* proposed in Celeux *et al.* (2003). It is based on two different approximations of probability distributions plus a simulation step, as follows. First, the distribution $\mathbb{P}_\psi(\mathbf{X})$ appearing in the complete likelihood is replaced by a mean-field approximation, namely the product distribution

$$\mathbb{P}^1(\mathbf{X}|\psi,\tilde{\mathbf{x}}) = \prod_{l=1}^{L}\prod_{i\in V^\star}\mathbb{P}_{\psi_l}(X_i^l|X_{\mathcal{N}_i^\star}^l = \tilde{x}_{\mathcal{N}_i^\star}^l), \tag{C.3}$$

for some well chosen fixed configuration $\tilde{\mathbf{x}} = (\tilde{x}_i^l)_{1\leq l\leq L, i\in V^\star}$. Second, the conditional distribution $\mathbb{P}_\psi(\mathbf{X}|\mathbf{Y})$ used for integrating the complete log-likelihood in (C.2) is also replaced by a mean-field approximation, that is

$$\mathbb{P}^2(\mathbf{X}|\psi,\tilde{\mathbf{x}},\mathbf{Y}) = \prod_{l=1}^{L}\prod_{i\in V^\star}\mathbb{P}_{\psi_l}(X_i^l|X_{\mathcal{N}_i^\star}^l = \tilde{x}_{\mathcal{N}_i^\star}^l, Y_i^l). \tag{C.4}$$

Note that both distributions (C.3) and (C.4) are probability distributions, contrarily to what happens when relying on pseudo-likelihoods. Third, the choice of the fixed configuration $\tilde{\mathbf{x}}$ relies on a sequential Gibbs sampling from the approximate distribution (C.4). With these three tools at hand, the algorithm consists in iteratively optimising with respect to $\psi = \{\psi_l\}_{1\leq l\leq L}$ the quantity

$$\mathbb{E}^2\big[\log\mathbb{P}^1(\mathbf{X}|\psi,\tilde{\mathbf{x}})\big|\psi^{(t)},\tilde{\mathbf{x}},\mathbf{Y}\big],$$

computed with the current value of the parameter $\psi^{(t)}$ and current simulated field $\tilde{\mathbf{x}}$. Here, $\mathbb{E}^2$ denotes expectation under the probability distribution $\mathbb{P}^2$. This quantity should be compared to the original criterion (C.2).

Let us now fully describe the procedure. For any current parameter value $\psi^{(t)}$ and fixed state value $\tilde{\mathbf{x}}$, we let

$$\tilde{Q}(\psi|\psi^{(t)},\tilde{\mathbf{x}}) = \sum_{l=1}^{L}\sum_{i\in V^\star}\sum_{x\in\{0,1\}}\mathbb{P}_{\psi^{(t)}}(X_i^l = x|X_{\mathcal{N}_i^\star}^l = \tilde{x}_{\mathcal{N}_i^\star}^l, Y_i^l)\log\mathbb{P}_\psi(X_i^l = x|X_{\mathcal{N}_i^\star}^l = \tilde{x}_{\mathcal{N}_i^\star}^l).$$

The algorithm consists in iterating the following two steps at time t,

- `SE`-step: sequentially sample a configuration $\tilde{\mathbf{x}}^{(t)}$ as follows for $1\leq l\leq L$ and $1\leq i\leq n$, sample $(X_i^l)^{(t)}$ according to the conditional distribution

$$x\mapsto\mathbb{P}_{\psi^{(t-1)}}(X_i^l = x|\{X_j^l = (\tilde{x}_j^l)^{(t)}, j\in\mathcal{N}_i^\star, j<i\},\{X_j^l = (\tilde{x}_j^l)^{(t-1)}, j\in\mathcal{N}_i^\star, j>i\}, Y_i^l).$$

This means that we sample the value 0 with probability

$$c\exp\Big(\beta^{(t-1)}_{l,co-abs}\sum_{j\in\mathcal{N}^\star_i}A^\star_{ij}\big[1\{(\tilde{x}^l_j)^{(t)}=0,j<i\}+1\{(\tilde{x}^l_j)^{(t-1)}=0,j>i\}\big]\Big)1\{Y^l_i=0\} \tag{C.5}$$

and we sample the value 1 with probability

$$c\exp\Big(\alpha^{(t-1)}_{i,l}+\beta^{(t-1)}_{l,co-pres}\sum_{j\in\mathcal{N}^\star_i}A^\star_{ij}\big[1\{(\tilde{x}^l_j)^{(t)}=1,j<i\}+1\{(\tilde{x}^l_j)^{(t-1)}=1,j>i\}\big] \\ +Y^l_i\log(p^{(t-1)}_{i,l})+(1-Y^l_i)\log(1-p^{(t-1)}_{i,l})\Big), \tag{C.6}$$

where c is a normalising constant (set such that the 2 probabilities sum to 1).

- M-step: Optimize $\tilde{Q}(\psi|\psi^{(t)},\tilde{\mathbf{x}}^{(t)})$ with respect to $\psi=\{\alpha_{i,l},\beta_{l,co-abs},\beta_{l,co-pres}\}_{i,l}$.

We now express the quantity $\tilde{Q}$ in our model and derive update formulas in our model. First we set

$$\tilde{p}_{i,l,t}(0)=c\exp\Big(\beta^{(t)}_{l,co-abs}\sum_{j\in\mathcal{N}^\star_i}A^\star_{ij}1\{(\tilde{x}^l_j)^{(t)}=0\}\Big)1\{Y^l_i=0\}$$

$$\tilde{p}_{i,l,t}(1)=c\exp\Big(\alpha^{(t)}_{i,l}+\beta^{(t)}_{l,co-pres}\sum_{j\in\mathcal{N}^\star_i}A^\star_{ij}1\{(\tilde{x}^l_j)^{(t)}=1\}+Y^l_i\log(p^{(t)}_{i,l})+(1-Y^l_i)\log(1-p^{(t)}_{i,l})\Big),$$

with the normalising constant c such that $\tilde{p}_{i,l,t}(0)+\tilde{p}_{i,l,t}(1)=1$. Then the vector $(\tilde{p}_{i,l,t}(0),\tilde{p}_{i,l,t}(1))$ is nothing else than the probability distribution $\mathbb{P}_{\psi^{(t)}_l}(X^l_i=\cdot|X^l_{\mathcal{N}^\star_i}=\tilde{x}^l_{\mathcal{N}^\star_i},Y^l_i)$. From this quantity, we obtain

$$\begin{aligned}
&\tilde{Q}(\psi|\psi^{(t)},\tilde{\mathbf{x}})\\
&=\sum_{i\in V^\star}\sum_{l=1}^{L}\Bigg\{\tilde{p}_{i,l,t}(0)\log\left[\frac{\exp\left[\beta_{l,co-abs}\sum_{j\in\mathcal{N}^\star_i}A^\star_{ij}(1-\tilde{x}^l_j)\right]}{\exp\left(\beta_{l,co-abs}\sum_{j\in\mathcal{N}^\star_i}A^\star_{ij}(1-\tilde{x}^l_j)\right)+\exp\left(\alpha_{i,l}+\beta_{l,co-pres}\sum_{j\in\mathcal{N}^\star_i}A^\star_{ij}\tilde{x}^l_j\right)}\right]\\
&\qquad+\tilde{p}_{i,l,t}(1)\log\left[\frac{\exp\left[\alpha_{i,l}+\beta_{l,co-pres}\sum_{j\in\mathcal{N}^\star_i}A^\star_{ij}\tilde{x}^l_j\right]}{\exp\left(\beta_{l,co-abs}\sum_{j\in\mathcal{N}^\star_i}A^\star_{ij}(1-\tilde{x}^l_j)\right)+\exp\left(\alpha_{i,l}+\beta_{l,co-pres}\sum_{j\in\mathcal{N}^\star_i}A^\star_{ij}\tilde{x}^l_j\right)}\right]\Bigg\}\\
&=\sum_{i\in V^\star}\sum_{l=1}^{L}\Bigg\{\tilde{p}_{i,l,t}(0)\Big[\beta_{l,co-abs}\sum_{j\in\mathcal{N}^\star_i}A^\star_{ij}(1-\tilde{x}^l_j)\Big]+\tilde{p}_{i,l,t}(1)\Big[\alpha_{i,l}+\beta_{l,co-pres}\sum_{j\in\mathcal{N}^\star_i}A^\star_{ij}\tilde{x}^l_j\Big]\\
&\qquad-\log\Big[\exp\Big(\beta_{l,co-abs}\sum_{j\in\mathcal{N}^\star_i}A^\star_{ij}(1-\tilde{x}^l_j)\Big)+\exp\Big(\alpha_{i,l}+\beta_{l,co-pres}\sum_{j\in\mathcal{N}^\star_i}A^\star_{ij}\tilde{x}^l_j\Big)\Big]\Bigg\}.
\end{aligned} \tag{C.7}$$

Optimising this quantity with respect to ψ is done numerically. To this aim, we provide below the derivatives of $\tilde{Q}$ wrt ψ.

Let us introduce the following quantities

$$w_i^\star = \sum_{j \in \mathcal{N}_i^\star} A_{ij}^\star,$$
$$w_{i,l}^\star = \sum_{j \in \mathcal{N}_i^\star} A_{ij}^\star X_j^l$$

which are the sum of weights of the neighbours of i in $G^\star$ and the sum of weights of the neighbours of i in $G^\star$ that are present at location l, respectively. Note that we have

$$\sum_{j \in \mathcal{N}_i^\star} A_{ij}^\star (1 - X_j^l) = w_i^\star - w_{i,l}^\star.$$

the sum of weights of the neighbours of i in $G^\star$ that are absent at location l. We also use

$$\mathrm{den}_{i,l}(\beta_{l,co-abs}, \beta_{l,co-pres}, \alpha_{i,l}) = \exp[\beta_{l,co-abs}(w_i^\star - w_{i,l}^\star)] + \exp(\alpha_{i,l} + \beta_{l,co-pres} w_{i,l}^\star).$$

With these quantities at hand and relying on (C.7), we obtain

$$\begin{aligned}\tilde{Q}(\psi|\psi^{(t)}, \tilde{\mathbf{x}}) = &\sum_{i \in V^\star} \sum_{l=1}^{L} \tilde{p}_{i,l,t}(0) \beta_{l,co-abs}(w_i^\star - w_{i,l}^\star) + \tilde{p}_{i,l,t}(1)[\alpha_{i,l} + \beta_{l,co-pres} w_{i,l}^\star] \\ &- \log \mathrm{den}_{i,l}(\beta_{l,co-abs}, \beta_{l,co-pres}, \alpha_{i,l}).\end{aligned}$$

Let us recall that $\alpha_{i,l}$ is a shorthand for the quantity $a_i + a_l + W_l^\intercal b_i + (W_l^2)^\intercal c_i$, so that we finally get, for each $1 \leq l \leq L$ and each $1 \leq i \leq n$, the derivatives

$$\frac{\partial \tilde{Q}}{\partial a_i} = \sum_{l=1}^{L} \Big[\tilde{p}_{i,l,t}(1) - \frac{\exp(\alpha_{i,l} + \beta_{l,co-pres} w_{i,l}^\star)}{\mathrm{den}_{i,l}(\beta_{l,co-abs}, \beta_{l,co-pres}, \alpha_{i,l})}\Big] \tag{C.8}$$
$$\frac{\partial \tilde{Q}}{\partial a_l} = \sum_{i \in V^\star} \Big[\tilde{p}_{i,l,t}(1) - \frac{\exp(\alpha_{i,l} + \beta_{l,co-pres} w_{i,l}^\star)}{\mathrm{den}_{i,l}(\beta_{l,co-abs}, \beta_{l,co-pres}, \alpha_{i,l})}\Big]$$
$$\frac{\partial \tilde{Q}}{\partial b_i} = \sum_{l=1}^{L} W_l^\intercal \Big[\tilde{p}_{i,l,t}(1) - \frac{\exp(\alpha_{i,l} + \beta_{l,co-pres} w_{i,l}^\star)}{\mathrm{den}_{i,l}(\beta_{l,co-abs}, \beta_{l,co-pres}, \alpha_{i,l})}\Big]$$
$$\frac{\partial \tilde{Q}}{\partial c_i} = \sum_{l=1}^{L} (W_l^2)^\intercal \Big[\tilde{p}_{i,l,t}(1) - \frac{\exp(\alpha_{i,l} + \beta_{l,co-pres} w_{i,l}^\star)}{\mathrm{den}_{i,l}(\beta_{l,co-abs}, \beta_{l,co-pres}, \alpha_{i,l})}\Big]$$
$$\frac{\partial \tilde{Q}}{\partial \beta_{l,co-abs}} = \sum_{i \in V^\star} \tilde{p}_{i,l,t}(0)(w_i^\star - w_{i,l}^\star) - \frac{(w_i^\star - w_{i,l}^\star)\exp[\beta_{l,co-abs}(w_i^\star - w_{i,l}^\star)]}{\mathrm{den}_{i,l}(\beta_{l,co-abs}, \beta_{l,co-pres}, \alpha_{i,l})}$$
$$\frac{\partial \tilde{Q}}{\partial \beta_{l,co-pres}} = \sum_{i \in V^\star} \tilde{p}_{i,l,t}(1) w_{i,l}^\star - \frac{w_{i,l}^\star \exp[\alpha_{i,l} + \beta_{l,co-pres} w_{i,l}^\star]}{\mathrm{den}_{i,l}(\beta_{l,co-abs}, \beta_{l,co-pres}, \alpha_{i,l})}. \tag{C.9}$$

The simulated field algorithm is described in Algorithm 1.

Algorithm 1: Simulated field algorithm

Input: Observed presence/absence data $\mathbf{Y}$, adjacency matrix of metanetwork $A^\star$.
Initialization: Choose initial values $\tilde{\mathbf{x}}^{(0)}, \psi^{(0)}$.
Set $t = 1$.
while not converged **do**
  Simulation step:
  for $1 \leq l \leq L$ **do**
    for $1 \leq i \leq n$ **do**
      Sample $(\tilde{x}_i^l)^{(t)}$ from $\{0, 1\}$ relying on the vector of probabilities (C.5) and (C.6).
    end for
  end for
  Compute $\tilde{Q}(\psi|\psi^{(t)}; \tilde{\mathbf{x}})$ from (C.7).
  Maximization step:
  Compute the value $\hat{\psi}$ zeroing the derivatives (C.8)–(C.9).
  Update parameter $\psi^{(t)} = \hat{\psi}$.
  Increment t.
end while

Remark 2. *In the case with no sampling effects (namely $p_{i,l} = 1$), the simulation step is skipped (since $\mathbf{X} = \mathbf{Y}$) and the algorithm reduces to optimizing the quantity*

$$\tilde{Q}_{direct}(\psi|\psi^{(t)}) = \sum_{l=1}^{L} \sum_{i \in V^\star} \sum_{x \in \{0,1\}} \mathbb{P}_{\psi^{(t)}}(X_i^l = x | X_{\mathcal{N}_i^\star}^l) \log \mathbb{P}_{\psi}(X_i^l = x | X_{\mathcal{N}_i^\star}^l).$$

This means that in this specific case, our method roughly consists in a pseudo-likelihood estimation, which is known to be consistent as the number of observations increases (Besag, 1975).

Therefore, the estimation algorithm is more computationally affordable in this case since it consists in a simple iteration of the `M`*-step (i.e. the "maximization step" in Algorithm 1).*

C.4 Additional details on the implementation

The "maximization step" in Algorithm 1 is performed using the vector Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm implemented in the GNU Scientific Library (`https://www.gnu.org/software/gsl/`). We observed that this algorithm was sensitive to the initial value of the parameters. After analyzing synthetic datasets simulated from the model (see Supplementary Figure 6) and estimating the model with various initial values, we validated the following combination of initial parameters:

$$\begin{aligned} a_i = a_l &= \frac{a_0}{2} \\ b_i = c_i &= 0 \\ \beta_{l,co-abs} = \beta_{l,co-pres} &= 0 \end{aligned}$$

with $a_0 = \log(\frac{\bar{Y}}{1-\bar{Y}})$ and $\bar{Y} = \sum_{il} Y_{il}/(nL)$.

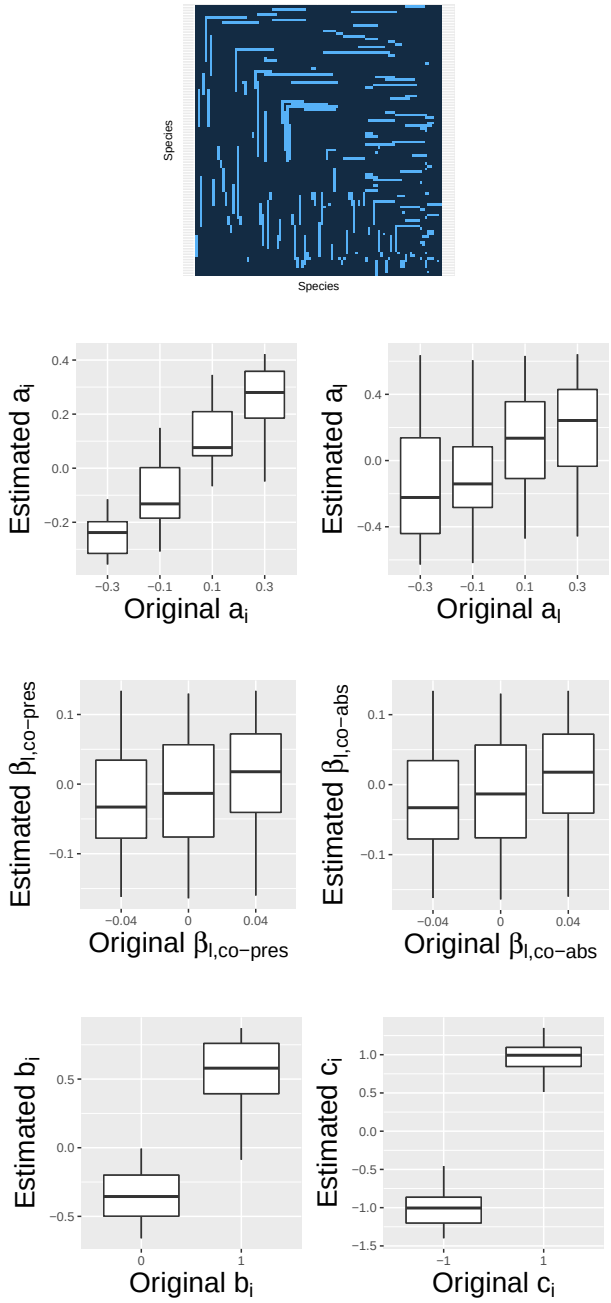


Figure 6: The estimation procedure successfully retrieves the original parameters used to simulate a dataset from the model. Simulation of a community of 100 species in 400 locations. Species are linked in a trophic network generated with the niche model (Williams & Martinez, 2000) with expected connectance of 0.07. Presence/absence data matrix X were simulated with a model with varying a_i, b_i and c_i among species and varying a_l, $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$ among locations. All the parameters were varying at the same time.

D Simulated communities

D.1 Simulation model

The community assembly process was randomly initialized with a set of individuals that were randomly selected in the species pool until the carrying capacity K was reached. At each time step, the probability of an individual from species i to replace a random individual of the community l is $R_{i,l}$. This probability depends on how the environmental conditions at location l are suitable for species i (environmental filter) and on the number of individuals present in community l that interact with species i (competition or mutualism filter). More precisely, we consider the following equation defining the relative importance of environmental and biotic filters respectively:

$$R_{i,l} = \exp\left[\gamma_{env} \log(P_{env,i,l}) \ + \gamma_{metanetwork} \log(P_{metanetwork,i,l})\right]$$

where γ_{env} and $\gamma_{metanetwork}$ are tuning parameters giving weights to abiotic and biotic components, and $P_{env,i,l}$ and $P_{metanetwork,i,l}$ are probabilities of species replacement with different filters. $P_{env,i,l}$ accounts for the environmental filtering and is a rescaled density of the Gaussian niche of species i at the environmental value of location l (the scaling ensures this value ranges in $[0, 1]$). When the environment in community l is suitable to species i, the probability that this species enters this community becomes high.

We then have a term dealing with species interactions, defined as

$$P_{metanetwork,i,l} = \begin{cases} K^{-1} \sum_{j;(i,j)\in E^\star} K_{j,l} & \text{for mutualism,} \\ 1 - K^{-1} \sum_{j;(i,j)\in E^\star} K_{j,l} & \text{for competition,} \end{cases}$$

where $K_{j,l}$ is the number of individuals of species j in community l, such that the total carrying capacity $K = \sum_j K_{j,l}$. In case of mutualism, the larger number of individuals of species connected with i in the metanetwork are present in location l, the higher is the probability of an individual of species i to enter the community. For competition, the opposite effect is induced. The tuning parameters γ_{env} and $\gamma_{metanetwork}$ weight the relative importance of the different filters. This algorithm updates the communities until an equilibrium is reached. To assess the equilibrium state, we calculated the Shannon diversity for each location over time, and checked for convergence. Lastly, we deduced species presence/absence by examining species composition in each location.

D.2 Simulation set-up

Let μ_1 and μ_2 be the niche optima of two species, and σ the standard deviation of their niche. Then, we considered that two species can interact in the mutualistic metanetwork if $\sigma < |\mu_1 - \mu_2| < 2\sigma$. Regarding competition, two species tend to compete if they share the same environmental niche, and thus if $|\mu_1 - \mu_2| < \sigma$. Among all potential species interactions, we randomly sampled 20% of them for both competition and mutualism (see Supporting Information).

We performed simulations with $N = 50$ species and $L = 400$ locations, with a carrying capacity of $K = 40$ individuals. The standard deviations of the Gaussian niche distributions

were set to $\sigma = 20$ for all species. We chose $\gamma_{env} = 1$ and $\gamma_{metanetwork} = 10$ in case of competition and 5 in case of mutualism. We simulated 100 time steps such that the algorithm convergence was achieved in practice. We repeated the whole procedure 10 times and verified that we obtained equivalent qualitative results. Simulations were implemented with R version 3.6.2 and a modified version of the VirtualCom package.

D.3 Adjacency matrices of the metanetworks representing competitive or mutualistic interactions

Figure 7 shows the adjacency matrices simulated under the 2 scenarios (competitive or mutualistic interactions).

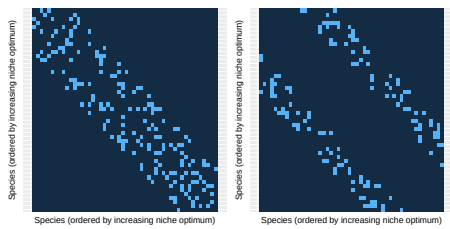


Figure 7: Synthetic adjacency matrices of the metanetworks representing a) competitive (left) or b) mutualistic (right) interactions. Species in rows/columns are ordered according to their increasing species niche optimum.

D.4 Computation of the expected co-presence and co-absence

To demonstrate how ELGRIN captures the effect of species interactions by using parameters $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$, we computed the expected number of co-presences of interacting (in the sense of the metanetwork) species at each location. To do this, we made the hypothesis that interactions do not affect species distributions. In this hypothesis, we considered that co-presence and co-absence were only the byproduct of metanetwork connectance (in other words, interacting species are co-present or co-absent just by chance). More precisely, we computed the expected number of co-presences by multiplying this connectance by the number of possible edges at each location (deduced from the number of present species). Conversely, we computed the expected number of co-absences with the same reasoning but using the connectance of the *complement* metanetwork (i.e. the network composed by the edges that are not in the metanetwork) and the number of absent species.

E Empirical case study : relation between $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$

Figure 8 shows the correlation between the values $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$ estimated through ELGRIN on the European tetrapods case study.

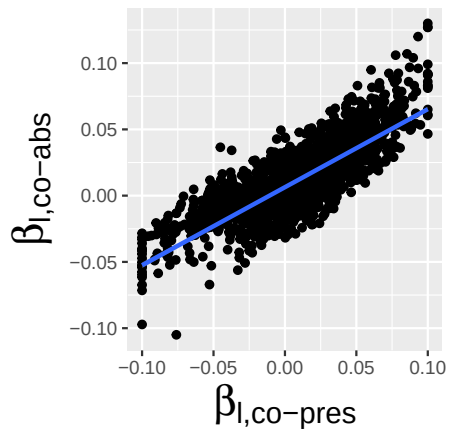


Figure 8: Results of ELGRIN on the European tetrapods case study. The parameters $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$ were highly correlated.

References

Berlow, E.L., Neutel, A.M., Cohen, J.E., De Ruiter, P.C., Ebenman, B., Emmerson, M., Fox, J.W., Jansen, V.A., Jones, J.I., Kokkoris, G.D. *et al.* (2004). Interaction strengths in food webs: issues and opportunities. *Journal of animal ecology* pp. 585–598.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)* 36, 192–225.

Besag, J. (1975). Statistical analysis of non-lattice data. *The statistician* 24, 179–195.

Celeux, G., Forbes, F. & Peyrard, N. (2003). EM procedures using mean field-like approximations for Markov model-based image segmentation. *Pattern Recognition* 36, 131–144.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B* 39, 1–38.

Krause, A.E., Frank, K.A., Mason, D.M., Ulanowicz, R.E. & Taylor, W.W. (2003). Compartments revealed in food-web structure. *Nature* 426, 282–285.

Lauritzen, S.L. (1996). *Graphical models*. vol. 17 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York. Oxford Science Publications.

Williams, R.J. & Martinez, N.D. (2000). Simple rules yield complex food webs. *Nature* 404, 180.