

Uncovering symmetric and asymmetric species associations from community and environmental data

Inferring species association networks [running headline]

KEYWORDS: network inference, representation learning, probabilistic graphical models, species embeddings, latent variable models, species associations networks

AUTHORS:

Sara Si-moussi `sara.si-moussi@univ-grenoble-alpes.fr`
Laboratoire d'Écologie Alpine, CNRS, Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, F-38000, Grenoble

Esther Galbrun `esther.galbrun@uef.fi`
School of Computing, University of Eastern Finland, Kuopio, FI-70211, Finland

Mickaël Hedde `mickael.hedde@inrae.fr`
UMR Eco&Sols, INRAE, IRD, CIRAD, Montpellier SupAgro, Université Montpellier, F-34000, Montpellier

Giovanni Poggiato `giov.poggiato@gmail.com`
Laboratoire d'Écologie Alpine, Univ. Grenoble Alpes, CNRS, Univ. Savoie Mont Blanc, CNRS, F-38000, Grenoble

Matthias Rohr `matthias.rohr@univ-grenoble-alpes.fr`
Laboratoire d'Écologie Alpine, Univ. Grenoble Alpes, CNRS, Univ. Savoie Mont Blanc, CNRS, F-38000, Grenoble

Wilfried Thuiller `wilfried.thuiller@univ-grenoble-alpes.fr`
Laboratoire d'Écologie Alpine, CNRS, Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, F-38000, Grenoble

Correspondence to **Sara Si-Moussi**

Mailing address: LECA, UMR UGA-USMB-CNRS 5553, Université Grenoble Alpes, CS 40700 38058 Grenoble cedex 9, France

E-mail: `sara.si-moussi@univ-grenoble-alpes.fr`

Acknowledgements. We thank Laura Pollock (McGill University) and Tamara Mündenmüller (LECA) for guidance on and access to source code for simulating virtual communities; we thank Philippe Choler (LECA) for discussion and crucial explanations on the ecology of Alpine plant communities. We also thank Li Ping Liu for access to source code on exponential family embeddings.

The research was supported by the Agence Nationale pour la Recherche (ANR) through the MIAI@Univ Grenoble Alpes institute (ANR-19-P3IA-0003) and the GlobNet (ANR-16-CE02-0009), Gambas (ANR-18-CE02-0025) and Forbic (ANR-18-MPGA-0004) projects. Most of the computations presented in this paper were performed using the GRICAD infrastructure¹. SMS was initially supported by a joint PhD fellowship between the French National Institute of Agricultural and Environmental Research (INRAE) and the French Research Institute for digital sciences (Inria) and by the Labex Persyval.

Authorship. SS and WT designed the study. SS conceptualized the inference framework with help from EG. WT and SS designed the evaluation methodology. SS, WT and EG analyzed the results. MH gave additional perspectives to the paper. SS and WT wrote the first version of the paper and all authors contributed critically to editing the manuscript.

Source code. The source code for running the model is available on this GitHub repository link.

¹<https://gricad.univ-grenoble-alpes.fr>

Abstract

Aim There is no much doubt that biotic interactions shape community assembly and ultimately the spatial co-variations between species. There is a hope that the signal of these biotic interactions can be observed and retrieved by investigating the spatial associations between species while accounting for the direct effects of the environment. By definition, biotic interactions can be both symmetric (e.g. competition, mutualism) and asymmetric (e.g. parasitism, predation, hierarchical competition). Yet, most models that attempt to retrieve species associations from co-occurrence or co-abundance data internally assume symmetric relationships between species. Here, we propose and validate a machine-learning framework able to retrieve bidirectional associations by analysing species community and environmental data.

Innovation Our framework (1) models pairwise species associations as directed influences from a source to a target species, parameterized with two species-specific latent embeddings: the effect of the source species on the community, and the response of the target species to the community; and (2) jointly fits these associations within a multi-species conditional generative model with different modes of interactions between environmental drivers and biotic associations.

Using both simulated and empirical data, we demonstrate the ability of our framework to recover known asymmetric and symmetric associations and highlight the properties of the learned association networks. By comparing our approach to other existing models such as joint species distribution models and probabilistic graphical models, we show its superior capacity at retrieving symmetric and asymmetric interactions.

Main conclusions Our framework enables ecologists to obtain a more generalized picture of the spatial associations between species without unrealistic assumptions of symmetry. The framework is intuitive, modular and broadly applicable across various taxonomic groups.

1 Introduction

Understanding the drivers of species distributions and abundances is a long-lasting goal of biogeography [Humboldt et al., 1805]. Niche theory explains the spatial distribution of species by a set of physiological and adaptive properties allowing them to thrive in specific environmental conditions and decline in others [Chase and Leibold, 2003, Pulliam, 2000]. The range of environmental variables (e.g. climate, land cover or soil characteristics) that matches the eco-physiological requirements of a species delimits its Grinnellian niche [Grinnell, 1917]. Habitat suitability models or species distribution models (SDMs) [Guisan et al., 2017] aim to infer this niche by establishing statistical relationships between observed occurrences or abundances of species and the environmental (abiotic) characteristics of the corresponding locations. These models have been particularly useful to predict species in space and time [Thuiller et al., 2019], providing operational tools to conservation biologists [Guisan et al., 2013, Pollock et al., 2020].

Beyond finding suitable habitats, living organisms meet their metabolic demands by feeding on, or acquiring, resources delimited by their Eltonian niche [Elton, 1927]. Through the processes of foraging for food or resources, reproducing and responding to the habitat conditions, species in a community affect each other, directly or via alterations of their surrounding environment (e.g. a large tree provides shade to shade-tolerant under-storey species). Moreover, species with shared resources may exclude one another locally (*competitive exclusion* [Hardin, 1960]) or be different enough in terms of space and resource needs to co-exist (*niche partitioning* [Schoener, 1974]). Conversely, some species facilitate others by modifying the environment in a way that creates habitats or enables access to resources for other species (*engineering and facilitation* [Cuddington et al., 2011]). These biotic interactions can thus be symmetric (e.g. mutualism, competition) in some cases or asymmetric in many other cases (e.g. predator-prey interaction, amensalism, parasitism) [Morales-Castilla et al., 2015]. Although biotic interactions are deemed to take place locally, they are likely driving spatial variation in species abundances [Boulangeat et al., 2012a], and may alter species ranges and leave imprints at large spatial scales [Gotelli, 2002, Wisz et al., 2013], but see [Thuiller et al., 2015].

As a result of species 'Grinnellian' and Eltonian's niches, together with species dispersal abilities, species co-abundances vary in space. These data, measured as *community data*, are usually the corner-stone of analyses aiming to tease apart the relative importance of these processes [Weiher and Keddy, 2001, Thuiller et al., 2013, Ovaskainen et al., 2017]. A natural way to address this objective is to jointly model multiple species distributions against environmental variables, and then, analyse the pairwise co-dependencies between species after controlling for the environmental effects. In theory, these pairwise co-dependencies (i.e. associations) could represent *the net effect* of one species on another, resulting from direct interactions or indirect effects [Ovaskainen et al., 2017]. In practice, due to the intertwined effects of biotic and abiotic processes, they are also the outcome of model mis-specifications and errors, of missing environmental variables and interacting species [Poggiato et al., 2021, Blanchet et al., 2020].

Several statistical frameworks have been proposed to infer these associations, either as their main objective or as a byproduct of the modeling process. These approaches differ in the type of dependencies they can model, in how they accommodate abundance data, and in the way they incorporate environmental covariates. Joint Species Dis-

tribution Models (JSDM, Warton et al. [2015]), the trendy tools at the moment, jointly predict the co-distributions of multiple species. Basically, once environmental covariates are accounted for, the residual correlation matrix is assumed to potentially capture species associations that are unexplained by the modeled covariates Pollock et al. [2014]. Recent implementations incorporate latent factors as a way to account for missing environmental variables and to reduce the parameter space size [Ovaskainen et al., 2017]. Adaptation for abundance data, particularly counts, was achieved through either data transformation techniques or appropriate link functions [Clark et al., 2017, Niku et al., 2019, Ovaskainen and Abrego, 2020, Chiquet et al., 2018, Popovic et al., 2018]. Alternatives mostly rely on Markov Random Fields (MRF) that can be applied to estimate conditional dependencies from a set of co-occurring species [Clark et al., 2018, Harris, 2016], while accounting for the environmental variations. MRF have the statistical property of estimating direct associations between pairs of variables while accounting for all other associations, which makes them highly suitable in Ecology [Clark et al., 2018].

Although these two approaches and others have generated a renewed interest to understand biodiversity patterns from community data, they have also crystallized strong debates on their capacity at revealing true associations that can ultimately be linked to interactions. First, it has been shown that most implementations of JSDMs provide similar predictions and inferences than traditional SDMs since the residual correlation structure does not affect the estimated species-environment relationships [Poggiato et al., 2021]. Second, simulated and empirical case studies have shown the difficulties of these approaches to infer simulated species associations [Zurell et al., 2018, König et al., 2021]. Third, extracting species associations from co-occurrence data proves to be a complex, if not impossible, problem [Blanchet et al., 2020, Cazelles et al., 2016a]. Last, but not the least, since both JSDM and MRF infer a precision matrix from the correlations between prediction residuals, they can only retrieve symmetric associations. This is not a desired properties since most species interactions, hence their induced associations, are likely to be asymmetric [Morales-Castilla et al., 2015].

Still, we believe that analysing community data along environmental gradients can bring useful information to infer species associations. To achieve such a long term goal, we need an approach that can handle both symmetric and asymmetric associations. Doing so requires capturing the way a given species affects the others, but also how the same species is affected by the other species. Interestingly, this duality has long been used in functional ecology to represent how a species respond to the environment through its 'response traits', and how it affects community functioning through its 'effect traits' [Lavorel and Garnier, 2002]. An extension of this response-effect framework has been proposed for trophic interactions by linking response traits at a given trophic level to effect traits at another level [Lavorel et al., 2013, Gravel et al., 2016]. We thus believe that distinguishing how a species respond to a species or a community from how it can affect it in return, which ultimately depend on the intrinsic properties or traits of the species, could provide a more suitable framework to make the best of community data and potentially extract information on species associations.

Outline Here, we propose a framework that builds on this response-effect concept to model species - environment relationships and pairwise symmetric and asymmetric (i.e. bidirectional) associations all-together. To do so, we use machine learning tools to build an efficient dependency network Heckerman et al. [2000] encoding bidirec-

tional species associations from community data. These associations are represented with two sets of embeddings encoding both species effects and responses to other species. Ultimately, the final conditional model of species abundances is built by aggregating both species-environment relationships and the biotic embeddings through different implementations that can best represent mechanistic understanding of the system (e.g. predator-prey interactions, competition-facilitation-amensalism-comensalism).

Through two experiments on simulated datasets and an empirical case study, we illustrate the different implementation of the interplay of biotic associations with environmental covariates. First, we simulate species abundance data with a species community model to evaluate the ability of our framework to recover known associations (both symmetric and asymmetric), where we assume an additive partitioning of environmental and biotic filters, and compare it with state-of-the-art joint species distribution models and Markov-random fields. Second, we evaluate the ability of our model to recover simulated predator-prey associations under different food web topologies, assuming a multiplicative effects of environmental and biotic filters.

Finally, we apply our framework to a well-studied Alpine plant community dataset Choler [2005], Warton et al. [2015] representing an example of hierarchical filtering of assembly rules (environmental effects at regional scale and competitive-facilitate interactions at the community scale). We used this empirical example to illustrate the analysis of structure of the species association networks.

2 The framework

At a high level, our framework models both species–environment responses and species–species associations. It captures how species respond to environmental conditions and to other species (i.e., species response), as well as how they influence the abundance or occurrence of others (i.e., species effect). The resulting graphical model is directional with respect to the environment (which drives species abundance, but not vice versa), and bidirectional with respect to species associations (Fig1a).

More specifically, the incoming edge weights for each species are estimated through a multiple regression that includes both environmental covariates and the abundances of co-occurring species (Fig1b). By incorporating all potential predictors in a single model, the framework quantifies *conditional dependencies* and disentangles environmental effects from biotic ones, allowing a direct assessment of their relative contributions. To represent bidirectional associations, we learn separate embeddings for how a species influences others and how it is influenced by them.

In the following, we define these embeddings and describe the conditional abundance model and its implementation.

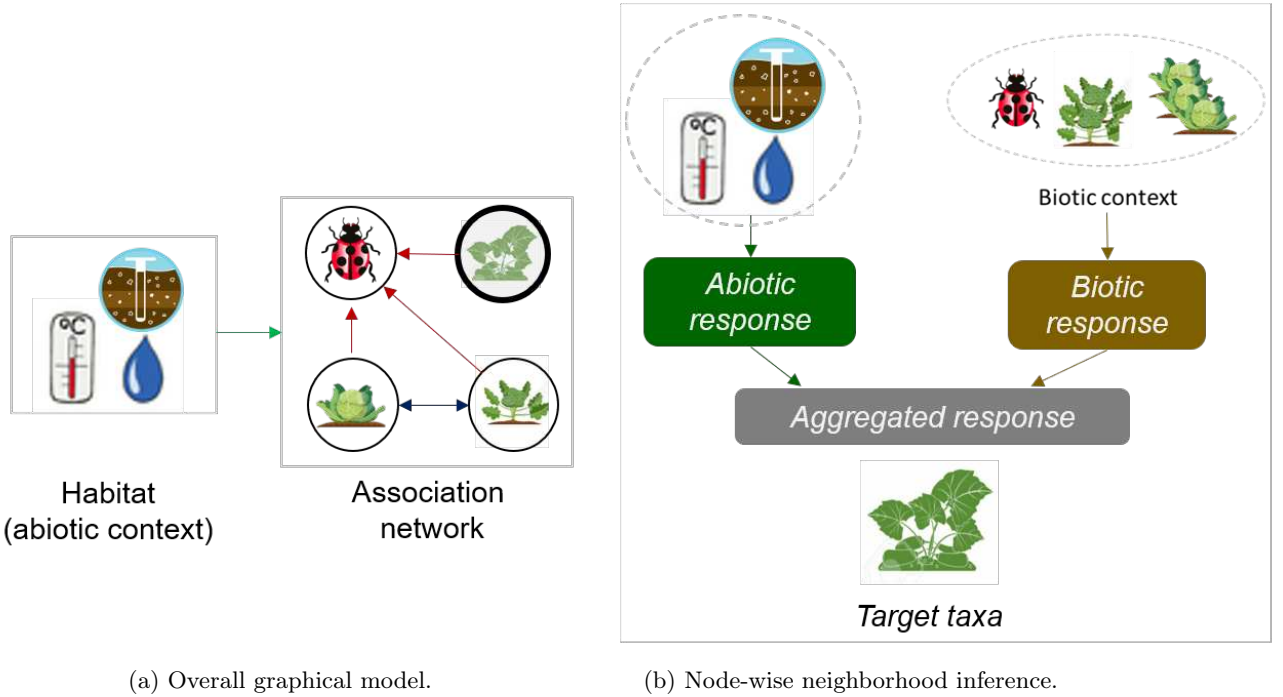


Fig. 1: A graphical illustration of the interplay between the environmental and the biotic filters. (a) Species in a community form a network of associations of different signs: positive (red), negative (blue). Each edge represents an association from a source to a target species. When both species influence each other, the association is bidirectional. All species, along with their associations, respond to the environmental conditions (green). We ignore the reciprocal effect of species on the environment. (b) The abundance of a given species results of its aggregated response to the environment and to the biotic contexts.

Notation We consider a site by species matrix $(\mathcal{K}, \mathcal{S})$, that contains the abundance of species i at site k , denoted y_{ki} . At each site k , the vector \mathbf{x}_k represents the environmental covariates.

2.1 Spatial associations and biotic context

2.1.1 Representing species associations using embeddings

For a given pair of species, a *spatial association* describes the statistical influence of a species on the abundance of another species. The influences can be of different polarity (positive, negative or neutral) and have different intensities (Fig 2b). Several mechanisms can lead to these associations: a direct interaction (e.g. pollination, predation), an indirect interaction through the environment (e.g. resource competition) or a shared correlation to an unmeasured environmental variable or unobserved species [Poggiato et al., 2021].

We note a_{ij} the influence of species j on species i which represents the change in abundance (excess if positive, deficit if negative) of a *target* species i induced by a *source* species j . These values that represent the sign and the strength of the association between all species pairs are stored into an $m \times m$ asymmetric association matrix \mathbf{A} (Fig 2a).

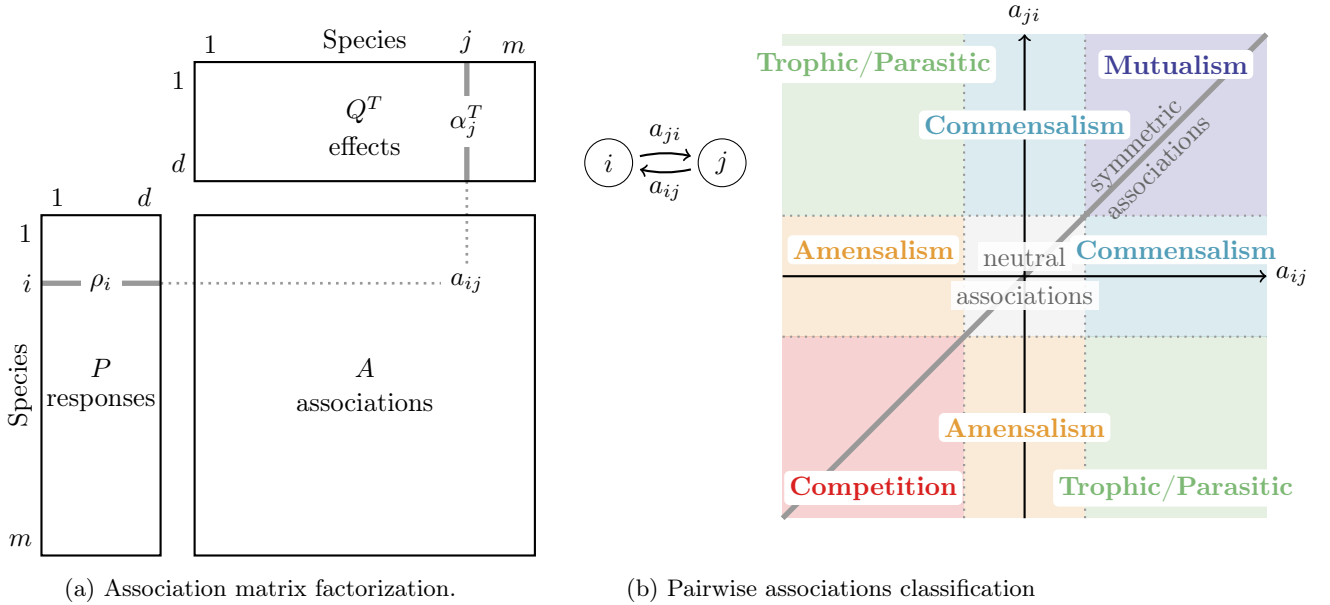


Fig. 2: Association strengths are computed from species response and effects (a). Pairwise association strengths are mapped to potential interaction classes (b). The different quarters of the bi-plot represent the various types of associations between species. The 1:1 line represents symmetric associations.

From a response-effect perspective, any element of \mathbf{A} (e.g. a_{ij}) is the byproduct of the effect of species j and the response of species i . We assume that these parameters represent latent traits or properties of the species that we do not observe, and which we implement as two separate d -dimensional embeddings.

The *effect embedding* of species i , α_i , expresses the type (i.e. traits, properties) of organisms the species allows or impedes when it is present. The *response embedding* of species i , ρ_i , expresses the type (i.e. traits, properties) of biotic context it can tolerate. For instance, trees with dense canopy create shade (effect) that selects only shade-tolerant (response) species and exclude others.

The response and effect embeddings of the different species are collected into two $m \times d$ matrices, respectively denoted as \mathbf{P} and \mathbf{Q} . The association matrix is then written as $\mathbf{A} = \mathbf{P}\mathbf{Q}^T$ (cf. Fig 2a).

2.1.2 Biotic context

The biotic context encodes our assumptions about the potential effects a target species is exposed to at a given site. In the simplest case, without any prior knowledge, it consists of the abundance of other species observed at the same site. Formally, the biotic context of species i at site k , denoted C_{ki} , is defined as follows:

$$C_{ki} = \{j \in \mathcal{S}, j \neq i \text{ and } y_{kj} > 0\}.$$

We obtain the aggregated effect of the biotic context by averaging the effect embeddings of its elements weighted by species' abundances:

$$z_{ki} = \frac{1}{|C_{ki}|} \sum_{j \in C_{ki}} y_{kj} \alpha_j.$$

This formulation allows the presence of opposing effects from different species to balance one another.

The biotic context constrains the structure of the inferred species association network by restricting the set of potential associations *a priori*. For instance, it can be easily adapted for each species according to known interactions. It can also include species from neighboring locations (**spatially-explicit**) up to a chosen radius within which their influence would be considered relevant (e.g. species with low/high mobility). Similarly, we can construct a **temporally-explicit** biotic context from previous observations to account for time-lag and phenological mismatches. (See Supplementary Methods for other biotic context formulation variants).

2.2 A conditional generative model of abundance

2.2.1 Conditional generative model

To disentangle environmental and biotic effects, we represent the response of the species i at site k as an aggregation function f_{agg} of the environmental η_{ki}^A and biotic η_{ki}^B responses (see Eq. (1a)). The environmental response is given by the habitat suitability model(s) h_i involving only the environmental covariates x_k (see Eq. (1b)). The biotic response η_{ki}^B depends on the response embedding ρ_i of the target species and on the biotic context effect z_{ki} resulting in an abundance-weighted sum of pairwise association strengths (see Eq. (1c)). An offset o_i is used to account for variation in exposure or effort.

The response of each species conditional to the environmental conditions and biotic context ($y_{ki} \mid x_k, C_{ki}$), denoted as y_{ki} for clarity, is assumed to follow a distribution \mathcal{F} from the Exponential Family with mean m_{ki} and dispersion ϕ_i parameters. The function g denotes the canonical link function, which relates the aggregated response at site k of species i to the mean (see Eq. (1d)).

The choice of distribution within this family is done according to the data type, for instance, using the normal distribution for biomass, the Bernoulli for presence/absence, or the Negative Binomial for over-dispersed counts.

$$g(m_{ki}) = f_{agg}(\eta_{ki}^A, \eta_{ki}^B) \quad (1a)$$

$$\eta_{ki}^A = h_i(x_k) \quad (1b)$$

$$\eta_{ki}^B = o_i + \rho_i z_{ki} = o_i + \sum_{j \in C_{ki}} (y_{kj} * a_{ij}) \quad (1c)$$

$$y_{ki} \sim \mathcal{F}(m_{ki}, \phi_i) \quad (1d)$$

2.2.2 Aggregation of abiotic and biotic effects

The aggregation function captures the interplay between abiotic and biotic filters (Fig. 3), which can follow different ecological assumptions. In the community assembly rule framework Weiher and Keddy [2001], environmental conditions first define the set of species that can potentially occur, while biotic interactions determine which of these species persist, based on their responses and effects. In some cases, abiotic conditions condition the occurrence of an interaction or share its nature or strength. In others, the biotic context itself can create favorable conditions, for instance, through facilitation.

To reflect these possibilities, we implement and evaluate three aggregation modes: additive, multiplicative, and hierarchical. Although these capture distinct ecological mechanisms, the framework is flexible and can accommodate alternative formulations (Fig. 3).

Additive filters

$$g(m_{ki}) = \eta_{ki}^A + \eta_{ki}^B \quad (2)$$

In the case of additive filters, the biotic context can complement the environmental conditions, and a species may occur if either filter is favorable. For instance, a species might be present even in unsuitable habitat if another facilitator species creates favorable micro-habitat. Conversely, a competitor's presence might exclude a species despite suitable environmental conditions. Here, the biotic and environmental responses are summed to reflect their combined, potentially compensatory influence (Eq.2).

Multiplicative filters

$$m_{ki} = \sigma(\eta_{ki}^A) \times \sigma(\eta_{ki}^B) \quad (3)$$

$$y_{ki} \sim \text{Bernoulli}(m_{ki}) \quad (4)$$

In the case of multiplicative effects, a species can only be present when both abiotic conditions and the biotic context are favorable. This setting is particularly relevant for obligate interactions, such as trophic, host–parasite, or host–symbiont relationships. For example, consider a predator species that requires both suitable abiotic conditions (e.g., temperature) and the presence of at least one prey species. Its occurrence (Eq 4) depends on the

product of these two components (Eq 3), each assessed independently, reflecting the necessity of both filters for survival.

Hierarchical filters The hierarchical filter setting distinguishes between two nested levels of influence: broad-scale environmental filters (e.g., climate) that define the regional species pool, and local-scale habitat features that interact with biotic associations to shape species abundances. This approach mirrors the commonly used assembly rule framework [Thuiller et al., 2013], in which abiotic filters act first, followed by biotic structuring. In practice, we implement this structure using a zero-inflated regression, where the environmental component governs species presence (denoted by the suitability s_{ki}) (Eq 5), and the biotic context influences the abundance y_{ki} conditional on occurrence. In this formulation, a Dirac point mass at zero is used such that when $s_{ki} = 0$, the species abundance is deterministically set to zero (Eq 6).

$$s_{ki} \sim \text{Bernoulli}(\eta_{ki}^A) \quad (5)$$

$$y_{ki} \sim \begin{cases} \mathcal{F}(\eta_{ki}^B; \phi_i), & \text{if } s_{ki} = 0. \\ \delta_0, & \text{otherwise.} \end{cases} \quad (6)$$

Choosing the aggregation function requires knowledge of the ecological community and the expected types of interactions or dependencies that could induce the inferred associations. In contrast, different assumptions can be tested and the best one can be quantified by statistical model selection.

2.3 Inference and model selection

We use the Stochastic Gradient Descent algorithm [Bottou, 2010] to optimize the negative log-likelihood of the observed abundances or occurrences with respect to the parameters, including the response and effect embedding matrices, the parameters of the abiotic response weights, and the species-specific dispersion parameters. Since the biotic context can substantially increase the number of variables and thus the risk of variance inflation due to multicollinearity [Dormann et al., 2013], we introduce elastic net regularization penalties to select meaningful associations for each species.

The proposed model includes a set of hyperparameters that must be selected carefully: the hyperparameters for the habitat suitability models, the embedding dimension, the vector of species offsets, and the regularization coefficient. We implemented two model selection (hyperparameter tuning) strategies. The first relies on information criteria [Konishi and Kitagawa, 2008] to penalize model complexity, such as the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), or its extended version (eBIC). The second strategy uses cross-validation based on predictive performance, using, for instance, the AUC for presence/absence data or Poisson deviance for count data.

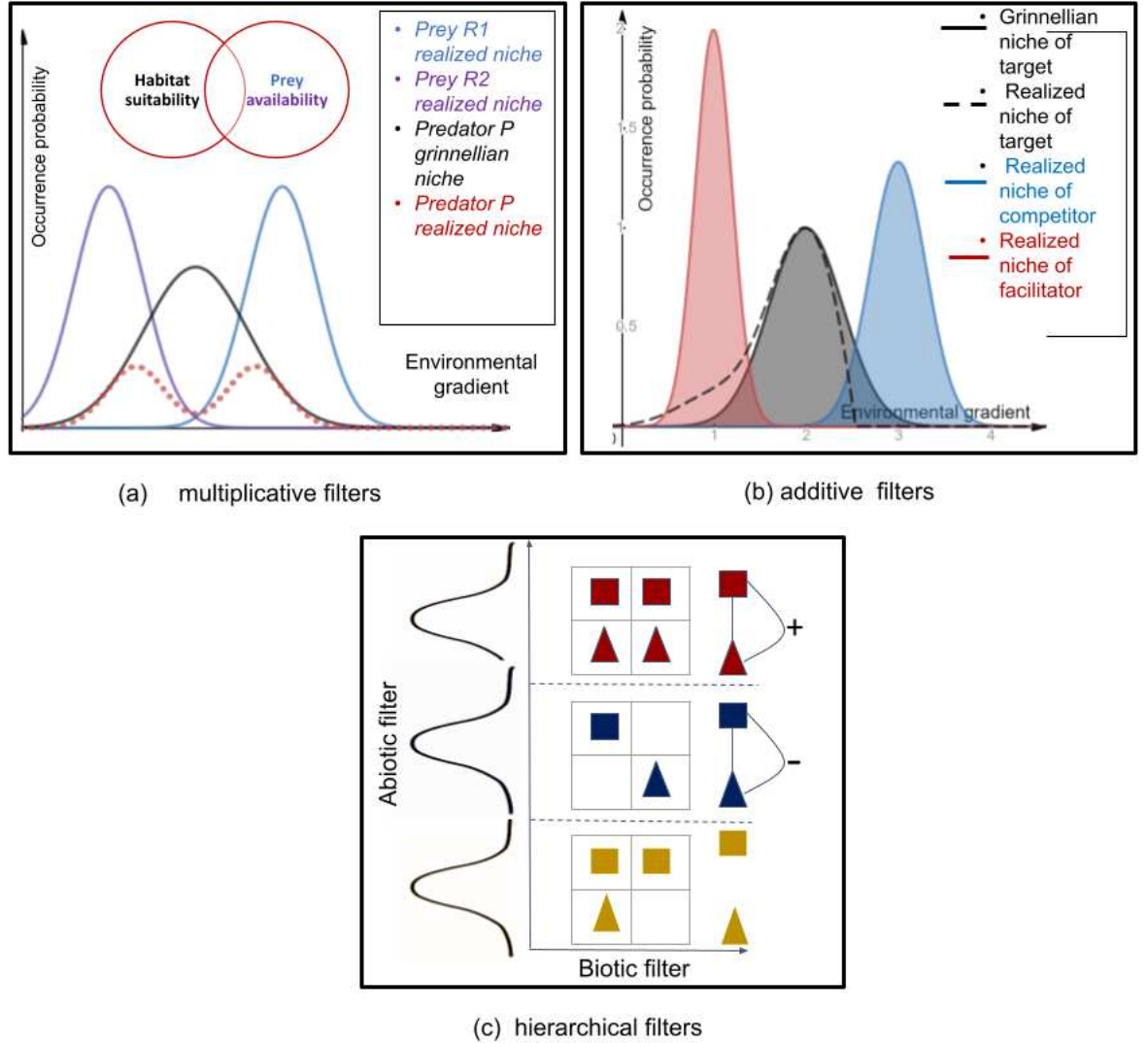


Fig. 3: Examples of scenarios for the aggregation of abiotic and biotic filters. (a) multiplicative filters represent a product of: environmental requirements and biotic resources (realized niche of preys). (b) Additive filters represent the sum of three complementary factors: the effect of the environmental environment and the associations that complement it through either presence of facilitators (resp. competitors) that can extend (resp. restrict) the suitable environmental range. (c) Hierarchical filters show the effect of two nested filters. First, an environmental filter operates at a regional scale, as depicted in the vertical stratification of species into 3 groups (regional pools): red, blue and yellow. Second, a biotic filter that acts locally through the presence of positive or negative associations or lack thereof between species of the regional pool.

2.4 Species association network

The regularization introduces sparsity into the association network by removing links between species that are independent or do not exhibit strong associations [Ohlmann et al., 2018]. Alternatively, the robustness of estimated biotic associations can be assessed through a bootstrap procedure: instead of applying a penalty during inference, the model is fitted to multiple bootstrap samples of the original dataset. Confidence intervals for the mean of each pairwise association are then computed, and associations whose intervals include zero are set to zero. In both the regularized and bootstrap approaches, a threshold can be applied to obtain a discrete version of the association matrix, defined as follows:

$$I_{ij} = \begin{cases} \text{positive} & \text{if } a_{ij} > \epsilon^+, \\ \text{negative} & \text{if } a_{ij} < -\epsilon^-, \\ \text{neutral} & \text{otherwise.} \end{cases}$$

such that ϵ^+ and ϵ^- represent user-defined thresholds on the strength of the positive and negative associations, respectively. The resulting matrix can be seen as a network, where each species is represented by a vertex and a directed edge labeled as positive (resp. negative) from vertex i to vertex j represents a positive (resp. negative) influence of species i on species j .

Based on our embedding definitions, species with similar response embeddings form clusters of rows in the association matrix, referred to as *response groups*, while species with similar effect embeddings form clusters of columns, or *effect groups*. These two sets of groups can be identified simultaneously using a co-clustering algorithm [Govaert and Nadif, 2013]. Their combination reveals blocks in the association matrix corresponding to groups of species that play similar structural roles i.e., are functionally redundant or exchangeable within the network [Gauzens et al., 2015].

3 Test of the framework on simulated species communities

To validate our framework, we conducted two simulation experiments in which community data were generated along an environmental gradient based on species-specific abiotic optima and predefined association matrices.

The first experiment was designed to assess the ability of our model (EA) and competing association inference methods (JSDBMs, MRFs) to recover both symmetric and asymmetric associations under an additive filtering scenario.

The second experiment was designed to test whether our model could recover associations under a multiplicative filtering scenario, where species presence depends on both environmental and biotic context suitability. This setting is not supported by the alternative methods.

3.1 Experiment 1: process-based simulation of community assembly

3.1.1 Community data simulation

We used a process-based stochastic model adapted from *Virtualcom* Münkemüller and Gallien [2015] to simulate the assembly of individuals from a regional species pool into communities, on different locations sampled along an environmental gradient (See Supplementary Methods).

We designed an experiment in which multiple simulations were run on random locations along a single environmental gradient (ranging from 0 to 100), using randomly generated configurations of the prior association matrix. These configurations included:

- Only environmental filtering (Env)
- Only positive associations (Pos)
- Only negative associations (Neg)
- Mixed positive and negative associations (PosNeg)

For each configuration, we varied:

- The species pool size: 10, 20, or 50 species
- The association density: sparse (1/3 of species pairs associated) vs. dense (2/3)
- The association symmetry: symmetric (+/+, -/-) vs. asymmetric associations (+/0, -/0).

Association strengths were fixed at +1 for positive and -1 for negative effects, focusing on association polarity rather than intensity. This factorial design yielded 33 simulation datasets, allowing us to evaluate our framework across a range of conditions and compare its ability to recover symmetric associations against JSDMs and probabilistic graphical models.

3.1.2 Inference

We fitted our model to species count data from the simulated communities using a negative binomial distribution with an exponential link function and an additive aggregation of environmental and biotic filters. Hyperparameters were selected via 10-fold cross-validation, using Poisson deviance as the performance metric.

3.1.3 Evaluation and comparison with JSDMs and graphical models

We also applied five well-established or emerging methods for inferring associations from count data: HMSC [Ovaskainen et al., 2017], EcoCopula [Popovic et al., 2019], EMTree [Momal et al., 2019], MRFcov [Clark et al., 2018], and PLN [Chiquet et al., 2018]. Table 1 summarizes, for each method, the underlying probabilistic model, data requirements, training/inference settings, and any additional post-processing steps.

For all methods, the inferred association matrices were discretized to identify the type of association (positive, negative, or neutral), and compared to the ground-truth using standard multi-class performance metrics: precision, recall, and F1-score. Recall reflects the proportion of true associations of a given type that were correctly identified

(sensitivity), while precision measures the proportion of predicted associations of a given type that were correct (specificity). The F1-score is the harmonic mean of precision and recall, balancing false positives and false negatives.

Framework	Count distribution	Association structure	Graph selection procedure	Learning configuration
Ecological Association Network (EA)	-Negative binomial -Unknown dispersion -Link: log	Dependency network	Cross-validation	Optimizer: adam Maximum number of epochs: 200 Batch size: 16 Early stopping (convergence by monitoring validation loss): - Patience: 5 epochs - Tolerance: 1E-3
EcoCopula Popovic et al. [2018, 2019]	-Negative binomial -Unknown dispersion -Link: log	Copula Gaussian Graphical Model	Graphical lasso	Importance sampling: 1000
EMtree Momal et al. [2019]	Poisson log-normal	Mixture of tree-shaped Gaussian Graphical Models	Edge probability Support: 2/pool_size	Covariance mode: full Number of iterations: 50 Convergence tolerance: 1E-8 Resampling: 5
Hierarchical Modeling of Ecological Communities (HMSC) Ovaskainen et al. [2017]	Poisson	Residual correlation	MAP residual covariance Support level 95%	MCMC: Hamiltonian Monte-Carlo thinning = 10 nChains = 2 Burn-in = 500 nSamples = 5000-7500 Cross-validation: 2 folds
MRFcov Clark et al. [2018]	Gaussian with non-paranormal transformation of count data	Conditional Markov Random Field	Bootstrap (95% CI) Sample proportion: 70% Symetrization function: mean	Bootstrap samples: 500
Poisson Log-Normal network (PLNetwork) Chiquet et al. [2018]	Poisson log-normal	Gaussian Graphical Model	Graphical lasso + stability selection	Covariance mode: full Offset: None

Table 1: Description of the evaluated frameworks and their respective configuration.

3.1.4 Results

The analysis of the relative abundance index (RAI_{ij}) showed good discrimination between positive and negative associations, while neutral associations resulted in more variable RAI_{ij} values (see Supplementary Materials).

Among the six inference methods, some were relatively easy to fit and offered limited control over model selection (e.g., EMTree, EcoCopula, MRFcov), beyond setting the number of iterations and the sampling scheme used for estimating confidence intervals. Execution time varied considerably between methods, largely driven by the model selection procedures employed (e.g., bootstrapping, lasso regularization, cross-validation). In particular, the Bayesian posterior inference in HMSC made it substantially slower than the other approaches.

For each method, we visualized the distribution of inferred association strengths across the different simulation configurations (e.g., environmental filtering only, association types, species pool size; Fig. 4).

All methods except HMSC produced sparse association networks, with low strengths values and were good at discriminating positive and negative associations, while maintaining neutral associations median-centered at zero. Most spurious associations, i.e. neutral pairs with inferred value significantly different from zero, were negative especially in simulations involving only positive associations reflecting the implicit constraint induced by the fixed carrying capacity on the total species count. On the other hand, HMSC produced very dense association matrices despite a large support level for association selection suggesting that some of the inferred associations are indirect associations. There was no difference in inferred strengths neither between symmetric and asymmetric simulations (for all methods).

As species pool size increases and niche overlap becomes more likely, inferred association strengths become more sensitive to niche differences: positive associations weaken with increasing overlap across all methods, while neg-

ative associations show variable responses, with EA and HMSC displaying opposite trends as niche distance increases.

In terms of association type classification, no major performance differences were observed across models for symmetric or dense association structures, the quality of inferred associations depended on species pool size: EA and EcoCopula consistently performed best on positive associations, especially in small pools, while negative associations were challenging for all methods, with EA, MRFcov, and HMSC showing relatively better performance. Detailed results are presented in the Supplementary Materials.

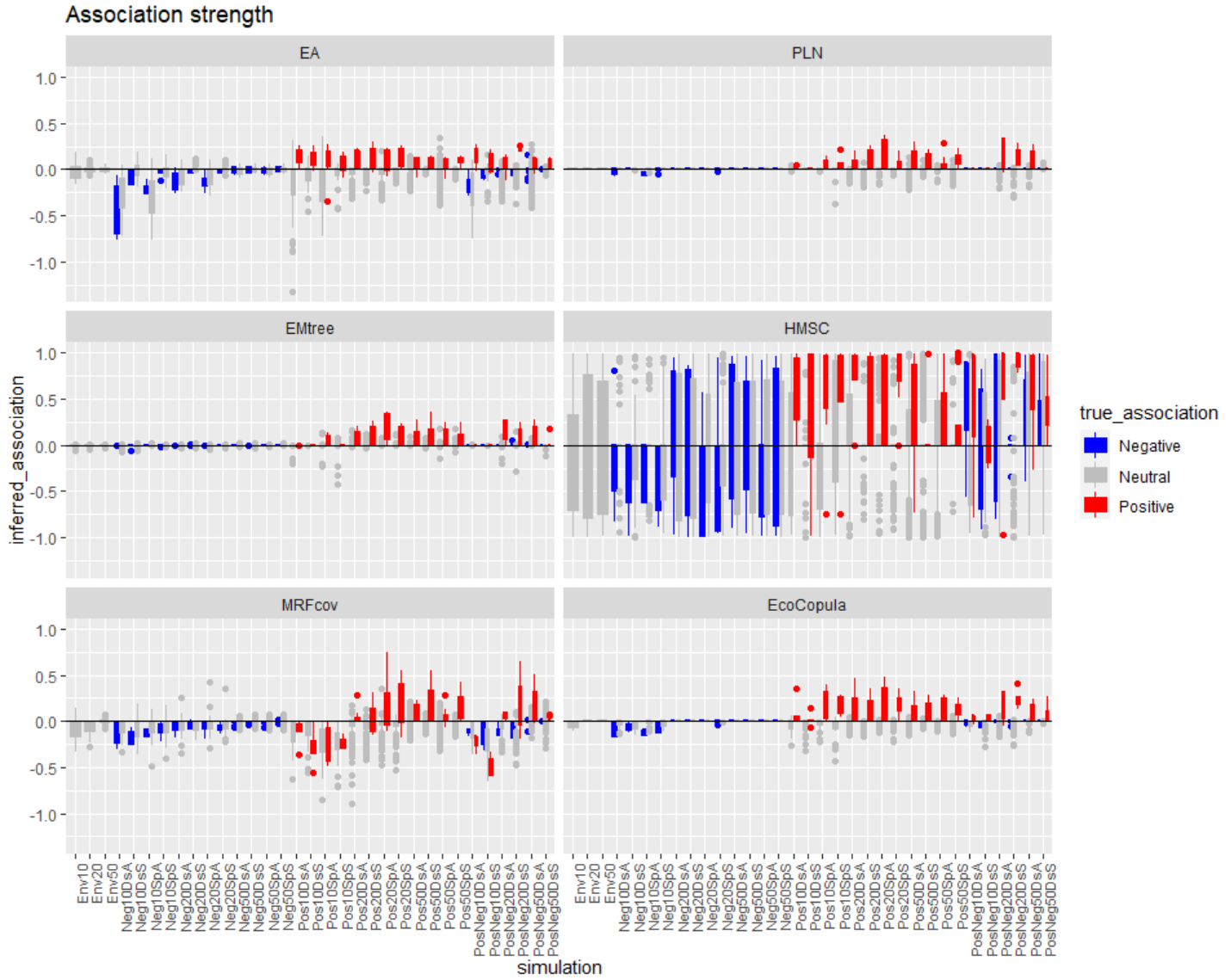


Fig. 4: Distribution of the association strengths inferred by the six methods run for each simulation experiment (in columns, e.g. Neg_S_10_D means simulated **N**egative **S**ymmetric associations for a species pool of **10** species, with **D**ense network of interactions). A data point represents a directed association from a species to another, its color encodes the true type of the association, its coordinates on the y axis represents the fitted association strength by the corresponding model for the simulation configuration x axis. In other words, blue histograms should rather be on the negative side (negative association correctly inferred), red colors on the positive side (positive association correctly inferred) and grey centered around 0 (neutral association)

3.2 Experiment 2: simulation of predator-prey co-occurrences

3.2.1 Community data simulation

In this experiment, we simulated species occurrence data where each species depends jointly on suitable environmental conditions and the presence of at least one prey given by a known predator-prey network (food web). This setup reflects an intersection of abiotic and biotic filters, modeled as a multiplicative response.

Using the `trophic` R package, we generated six food webs with different topologies, each involving the same number of trophic groups ($G = 5$). A trophic group consists of species that share the same prey and are consumed by the same predators, statistically analogous to a latent block structure in a graph.

To each trophic group, we assign $m_G = 5$ species with different abiotic niche optima sampled uniformly along an environmental gradient ranging from 0 to 100. We select 500 sites uniformly in the same gradient.

3.2.2 Inference

We fitted our model to the simulated presence/absence data using a multiplicative filter setting. We used a linear logistic regression with a quadratic term to fit the Gaussian abiotic niche. The simulation model assumes a unidirectional positive dependency of the predators on their preys. Thus, we imposed a non-negative constraint to the embedding vectors to prevent the inference of negative associations and promote sparsity of the association matrix Hoyer [2004]. Consequently, we only inferred two types of associations: positive and neutral. Additionally, we tested whether imposing structure by sharing embeddings between species of the same trophic group improved the ability of the model to retrieve true potential and realized associations.

We used a 10-fold cross-validation to select the combination of embedding dimension and lasso regularization that maximized the accuracy of predicted occurrences.

3.2.3 Evaluation

We evaluated the quality of the recovered associations in terms of accuracy, ROC-AUC, sensitivity, and specificity. As ground truth, we used two reference food webs: (1) the potential food web (metaweb), which includes all possible interactions; and (2) the realized food web, obtained by filtering the metaweb to retain only interactions between species that co-occur at least in one site.

3.2.4 Results

The inferred associations were more faithful to the realized than the potential network (Fig 5). In all cases, incorporating a parameter-sharing constraint within trophic groups allowed to improve the sensitivity with respect to both ground truth networks.

Additional results are available in the Supplementary Materials.

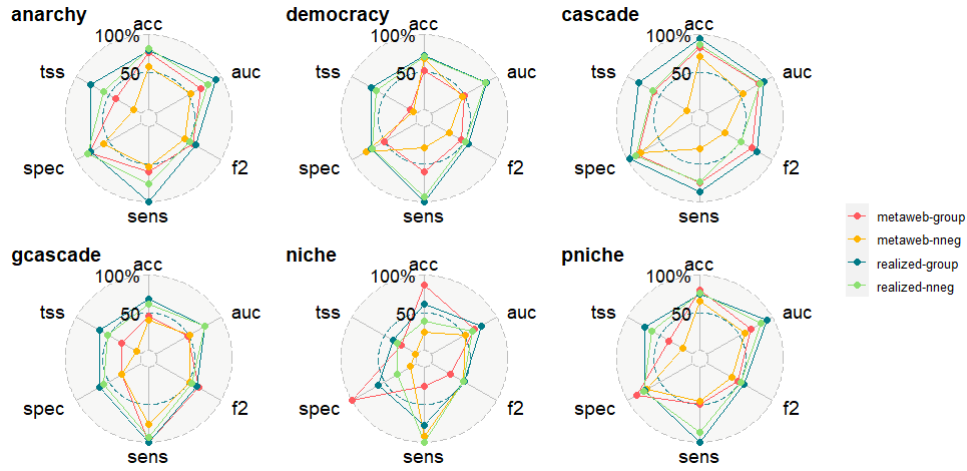


Fig. 5: Network structure inference quality with respect to the potential (metaweb) and the realized networks under two different constraints: non-negative associations and within-group embedding sharing. Performances are reported separately for each food web topology.

4 Empirical case study - Alpine plant associations

4.1 Plant community data

To test our model on a real ecological system, we applied it to an Alpine plant abundance dataset originally published by Choler [2005]. The dataset includes abundance records for 82 plant species surveyed in July 2000 across 75 vegetation plots (each 5×5 m) along a meso-topographical gradient in the French Alps. Environmental and topographic variables were also recorded for each plot.

4.2 Inference and statistical analyses

We fitted our model to this dataset using the hierarchical filtering mode (Fig. 3), assuming that habitat suitability drives species occurrence, while local biotic associations influence species abundance and can lead to local exclusion [Boulangeat et al., 2012b]. Details of data pre-processing and model selection are provided in the Supplementary Materials.

4.3 Results

The application of our approach to the Alpine dataset identified four densely connected modules of different sizes, within which species occupied distinct structural roles in the plant association network. Modules were structured along a gradient of response to the snow melting date (Fig. 6).

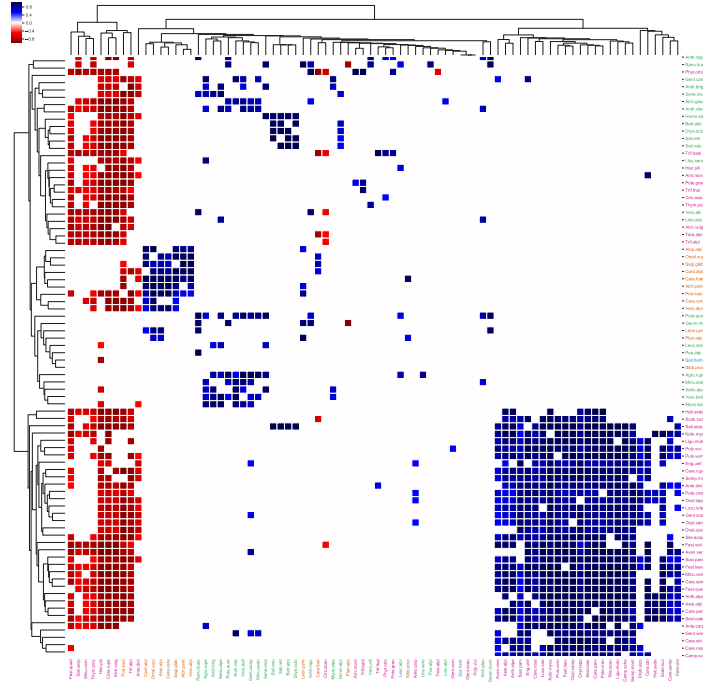
Species from early-melting sites clustered into the same module, characterized by a dominance of positive associations—notably, a largely asymmetric attraction of forbs and grasses toward tall, dominant graminoids such as *Carex* and *Kobresia*. In contrast, forbs and grasses also formed two distinct groups connected by negative associations, indicative of competitive exclusion. Some of these species acted as hubs, linking high-elevation sites to adjacent zones where they also occurred.

The second module encompassed two groups of grasses: tall herbs occurring in favorable conditions, which were primarily structured by negative associations reflecting amensalism and competition; and short herb meadows, ex-

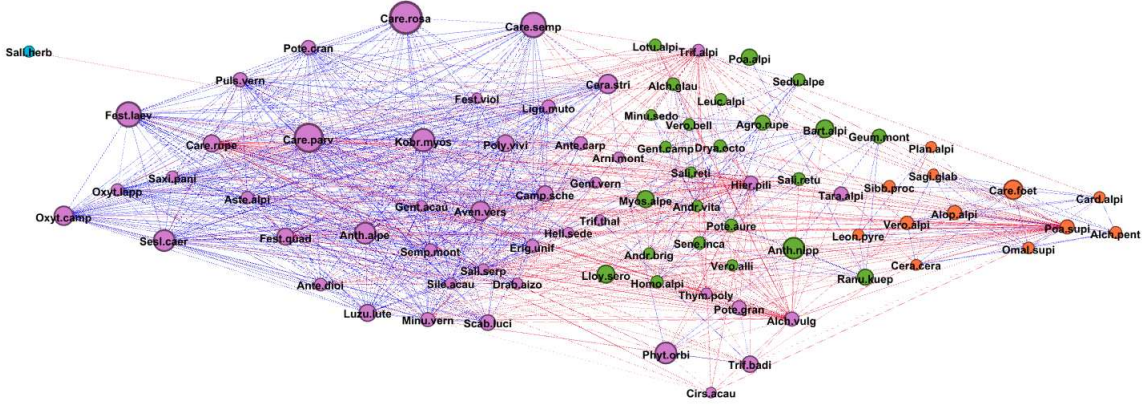
posed to zoogenic disturbance, which exhibited increased abundance when co-occurring with tall herbs, suggesting a facilitative interaction.

The third module represented chionophilous (cold-adapted) vegetation found on late-melting sites. The fourth module encompassed north-facing, isolated communities, dominated by *Salix herbacea*, which showed positive associations with high-altitude communities but remained disconnected from other modules (Fig. 6).

Interestingly, we found a higher proportion of positive associations in communities from stressful environments, such as early-melting sites exposed to wind and erosion due to snowmelt [Choler, 2005]. These associations likely reflect facilitative interactions mediated by graminoids through several mechanisms: graminoids help stabilize the soil [Callaway, 2007, Heilbronn and Walton, 1984], reduce desiccation and frost heaving on stones—thereby supporting seedling survival [Choler et al., 2001]—and create favorable microclimatic conditions that shelter smaller forbs and grasses from wind exposure [Wardle et al., 1998]. In contrast, negative associations were more frequent in species-rich sites, likely driven by competition for limiting resources such as water and nitrogen [Choler et al., 2001].



(a) Inferred plant association matrix. Species in the association matrix are grouped based on a hierarchical co-clustering performed row-wise (yielding response groups) and column-wise (yielding effect groups).



(b) Network of plant associations. Blue (resp. red) edges indicate negative (resp. positive) edge weights. Node colors on the graph represent communities identified by the modularity maximization algorithm Newman [2006] whilst node sizes are scaled according to the plant height. Nodes (except *Salix herbacea*, which represents the vegetation on the northern face of the gradient) are placed from left to right following an ascending order of their response to Snow duration (regression coefficient from the Generalized Linear Model used as a Habitat Suitability Model).

Fig. 6: Plant associations on an Alpine mesotopographic gradient. We highlight the communities (node colors) in figure (b) using colored labels on the matrix (a).

5 Discussion

In this work, we tackled the challenge of inferring interspecific associations from multiple species co-abundances in heterogeneous environments. To do so, we formalized pairwise associations as a function of two sets of latent variables representing the response and the effect of each species with respect to the others. We incorporated these associations into a conditional probabilistic model of abundance that accounts for environmental covariates. We evaluated our approach on both simulated and empirical datasets.

5.1 Disentangling abiotic and biotic drivers

5.1.1 Uncovering positive, negative and neutral associations

Comparatively to other tested frameworks, our method (EA) performed well on both positive and negative associations detection, despite the constraint induced by the embedding-based factorization. On average, discriminating positive and negative associations was within reach of most methods, provided an appropriate pair of thresholds was used to delimit the range of neutral associations. An exception concerned HMSC, which recovered multiple spurious associations that were potentially indirect effects despite the high support level. A more appropriate approach would have been to analyze the inverse covariance matrix. However, inverting the posterior estimate of the covariance matrix suffered from various numerical instabilities.

Due to the upper-limit constraint of the fixed carrying capacity on the total count, all models inferred spurious negative associations between non-interacting species, especially in simulations with positive effects only, as a compensation mechanism. Association strengths were sensitive to niche overlap. For all methods, positive associations were easier to detect between species with overlapping niches. The fact that this pattern was observed for all methods as well as on the pairwise relative abundance indices suggested that the abiotic filter outruled these associations during the community assembly simulation.

Amongst the tested methods, HMSC, PLN/EMtree and EcoCopula first fit the abiotic response then species dependencies are estimated either as random effects (HMSC, PLN) or from the marginal residuals (EcoCopula). The implicit importance given by these inference procedures to the abiotic drivers over the species associations explains the low detection rate of negative associations. In contrast, EA and MRFcov which both rely on an explicit regression over species abundances, do not suffer from the same bias, explaining their superiority in detecting negative effects.

5.1.2 Uncovering prey-predator associations

Several studies discuss the difficulties of recovering biotic interactions from co-occurrences Sander et al. [2017], Barner et al. [2018], Blanchet et al. [2020]. In our experiment, we assumed that trophic interactions induce a dependence of predators on their preys but not vice-versa (directed positive association).

We showed that using an appropriate coupling with the abiotic drivers allows to detect such associations providing that the species pair co-occur. However, the model detected symmetric dependencies when the abiotic niches of the pair overlapped strongly and especially in trophic chains and when the predator did not have other

alternative prey. Moreover, the varying performances in recovering the true network structure for different food web topologies questions the power of the response-effect factorization to represent arbitrary directed acyclic graph (DAG) structures and suggests that a symmetric approach might be more effective, if coupled with knowledge of species trophic levels.

5.1.3 Importance of the abiotic-biotic aggregation function

Species joint responses to abiotic (environment) and biotic (associations) drivers take on different forms, modeled by an aggregation function. Most existing frameworks are limited to linear or additive forms. Linear responses are particularly useful when associations are mediated by the environment (e.g in competition) or can alter it (as in habitat facilitation). In this case, associations compensate the suitability of the environment by either improving micro-habitat conditions or exerting a negative force that counterbalances it.

On the other hand, when associations arise from direct interferences, their detection requires conditioning on co-occurrence, hence on habitat suitability Gravel et al. [2019]. When we fitted an additive architecture to the predator-prey occurrences, the model had a very low detection rate confirming that linear combinations of habitat suitability and biotic effects are not sensitive to such direct associations. These results may be specific to presence/absences and not hold true for abundances.

In general, the choice of an aggregation function depends on the type of interactions expected in the studied system. To guide this choice, several frameworks [Kissling et al., 2012, Boulangeat et al., 2012b, Thuiller et al., 2013] conceptualize the incorporation of eco-evolutionary processes into species distribution models (a.k.a *BI-SDMs* [Dormann et al., 2018]). Besides, theoretical developments extended the theory of island biogeography [MacArthur and Wilson, 2001] to account for trophic interactions [Gravel et al., 2011] and more general interaction networks under environmental constraints [Cazelles et al., 2016a,b].

5.2 From species representations to biotic associations

5.2.1 The meaning of species embeddings

In theory, the effect embedding of a species is equivalent to a factor analysis of all other species abundances (residual abiotic responses if coupled with environmental data) when it is present. The effect embedding is a proxy of the species' influence on the community composition. Combining the effect embeddings of occurring species produces an ordination of the community composition in the embedding space of dimension d : R^d . The species response embedding can be mapped into the same space, we can measure through the dot product the compatibility of the species to the observed community.

Since the community ordination is obtained as a linear combination of present species' effects, species response to the community can be rewritten as a sum of one-to-one responses to each observed species. When the response and effect embeddings are forced to be similar, we recover the same structure used by Latent Variable JSDMs.

Analogously to the species embeddings, Kissling et al. [2012] proposed the concept of *interaction currencies* as surrogates for biotic interactions in distribution models in a similar response-effect framework. Hypothetically, these currencies include resources, biotic variables [Hutchinson, 1957], traits, and other non-consumable

environmental conditions that mediate interactions. Our analysis of embeddings learnt from data in food web simulations showed that they captured both abiotic and biotic species preferences. In the case study on Alpine plants, we found that embeddings were mildly related to functional traits.

5.2.2 Constraining embeddings with prior knowledge

In practice, the embedding dimension is typically significantly smaller than the number of species. While species can have distinct habitat preferences, the biotic role expressed in their interactions and the spatial associations they produce is drawn from a limited number (significantly smaller than the pool size) of behaviors represented by functional groups [Walker, 1992]. A species can belong to one or several functional groups with different proportions. Such information can be mined from online databases or provided by experts [BETSI, 2012, Nguyen et al., 2016, Kattge et al., 2020]. While learning graphical models with large species pools requires large datasets, replacing species with fixed groups Ohlmann et al. [2018] has two advantages: (1) to reduce the parameter space size by sharing embeddings within groups, (2) allowing extrapolation to new settings where different taxa are observed yet from the same modeled groups. Besides, as evidenced by our simulated experiment, using group constraints can improve the ability of inference models to recover potential associations even when species did not co-occur.

5.3 Perspectives

Beyond group constraints, some frameworks [Lo and Marculescu, 2017, Chiquet et al., 2018, Scutari et al., 2019] support white-lists and black-lists, containing authorized and forbidden associations respectively, by penalizing graphs that do not satisfy those constraints. When interaction networks can be described at least partially, the same approach can be used to complete missing edges by harnessing similarities of species interactions. This semi-supervised problem is referred to as *collaborative fitltering* [Fu et al., 2019] and is one of the main applications of dependency networks. Incorporating this link prediction task within a multispecies distribution model would allow to quantify the effect of known and predicted interactions on species distributions.

We motivated throughout our simulation experiments the use of different joint responses for abiotic and biotic drivers depending on the underlying biotic interactions. The fact that interactions require and affect co-occurrences simultaneously are not mutually exclusive [Gravel et al., 2019]. The availability of multi-trophic communities datasets [Derocles et al., 2018] where complex interactions are entangled calls for applications coupling different modes of aggregating abiotic drivers with biotic associations.

6 Conclusion

Biological interactions and other processes induce spatial patterns of co-occurrence and co-abundance. We presented and validated a model of species co-abundances as a function of the habitat and biotic associations. We proposed an asymmetric scheme for modeling associations that is based on learning latent representations of species' responses and effects. Future efforts should be directed towards an incorporation of prior knowledge of

the complete or partial topology of the association networks to guide the inference process. Along with that, a strong theory of how known ecological interactions influence the co-distribution of species is needed to support all these models.

References

- Allison K Barner, Kyle E Coblentz, Sally D Hacker, and Bruce A Menge. Fundamental contradictions among observational and experimental estimates of non-trophic species interactions. *Ecology*, 99(3):557–566, 2018.
- A BETSI. Database for functional traits of soil invertebrates. *French Foundation for Biodiversity Research*, 2:4, 2012.
- F Guillaume Blanchet, Kevin Cazelles, and Dominique Gravel. Co-occurrence is not evidence of ecological interactions. *Ecology Letters*, 2020.
- Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT’2010*, pages 177–186. Springer, 2010.
- I. Boulangeat, D. Gravel, and W. Thuiller. Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. *Ecology Letters*, 15(6):584–593, 2012a.
- Isabelle Boulangeat, Dominique Gravel, and Wilfried Thuiller. Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. *Ecology letters*, 15(6):584–593, 2012b.
- Ragan M Callaway. *Positive interactions and interdependence in plant communities*. Springer, 2007.
- Kévin Cazelles, Miguel B Araújo, Nicolas Mouquet, and Dominique Gravel. A theory for species co-occurrence in interaction networks. *Theoretical Ecology*, 9(1):39–48, 2016a.
- Kévin Cazelles, Nicolas Mouquet, David Mouillot, and Dominique Gravel. On the integration of biotic interaction and environmental constraints at the biogeographical scale. *Ecography*, 39(10):921–931, 2016b.
- Jonathan M Chase and Mathew A Leibold. *Ecological niches: linking classical and contemporary approaches*. University of Chicago Press, 2003.
- Julien Chiquet, Mahendra Mariadassou, and Stéphane Robin. Variational inference for sparse network reconstruction from count data. *arXiv preprint arXiv:1806.03120*, 2018.
- Philippe Choler. Consistent shifts in alpine plant traits along a mesotopographical gradient. *Arctic, Antarctic, and Alpine Research*, 37(4):444–453, 2005.
- Philippe Choler, Richard Michalet, and Ragan M Callaway. Facilitation and competition on gradients in alpine plant communities. *Ecology*, 82(12):3295–3308, 2001.
- James S Clark, Diana Nemergut, Bijan Seyednasrollah, Phillip J Turner, and Stacy Zhang. Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. *Ecological Monographs*, 87(1):34–56, 2017.
- Nicholas J Clark, Konstans Wells, and Oscar Lindberg. Unravelling changing interspecific interactions across environmental gradients using markov random fields. *Ecology*, 99(6):1277–1283, 2018.

- Kim Cuddington, James E Byers, William G Wilson, and Alan Hastings. *Ecosystem engineers: plants to protists*, volume 4. Academic Press, 2011.
- Stephane AP Derocles, David A Bohan, Alex J Dumbrell, James JN Kitson, Francois Massol, Charlie Pauvert, Manuel Plantegenest, Corinne Vacher, and Darren M Evans. Biomonitoring for the 21st century: integrating next-generation sequencing into ecological network analysis. In *Advances in ecological research*, volume 58, pages 1–62. Elsevier, 2018.
- Carsten F Dormann, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J Leitao, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46, 2013.
- Carsten F Dormann, Maria Bobrowski, D Matthias Dehling, David J Harris, Florian Hartig, Heike Lischke, Marco D Moretti, Jörn Pagel, Stefan Pinkert, Matthias Schleuning, et al. Biotic interactions in species distribution modelling: 10 questions to guide interpretation and avoid false conclusions. *Global ecology and biogeography*, 27(9):1004–1016, 2018.
- Charles S Elton. The nature and origin of soil-polygons in spitsbergen. *Quarterly Journal of the Geological Society*, 83(1-5):163–NP, 1927.
- Xiao Fu, Eugene Seo, Justin Clarke, and Rebecca A Hutchinson. Link prediction under imperfect detection: Collaborative filtering for ecological networks. *IEEE Transactions on Knowledge and Data Engineering*, 2019.
- Benoit Gauzens, Elisa Thébault, Gérard Lacroix, and Stéphane Legendre. Trophic groups and modules: two levels of group detection in food webs. *Journal of The Royal Society Interface*, 12(106):20141176, 2015.
- Nicholas J Gotelli. Ecology: Biodiversity in the scales. *Nature*, 419(6907):575, 2002.
- Gérard Govaert and Mohamed Nadif. *Co-clustering: models, algorithms and applications*. John Wiley & Sons, 2013.
- D. Gravel, C. Albouy, and W. Thuiller. The meaning of functional trait structure and diversity to food web dynamics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371:20150268, 2016.
- Dominique Gravel, François Massol, Elsa Canard, David Mouillot, and Nicolas Mouquet. Trophic theory of island biogeography. *Ecology letters*, 14(10):1010–1016, 2011.
- Dominique Gravel, Benjamin Baiser, Jennifer A Dunne, Jens-Peter Kopelke, Neo D Martinez, Tommi Nyman, Timothée Poisot, Daniel B Stouffer, Jason M Tylianakis, Spencer A Wood, et al. Bringing elton and grinnell together: a quantitative framework to represent the biogeography of ecological interaction networks. *Ecography*, 42(3):401–415, 2019.
- Joseph Grinnell. The niche-relationships of the california thrasher. *Auk*, 34(4):427–433, 1917.

- Antoine Guisan, Reid Tingley, John B Baumgartner, Ilona Naujokaitis-Lewis, Patricia R Sutcliffe, Ayesha IT Tulloch, Tracey J Regan, Lluís Brotons, Eve McDonald-Madden, Chrystal Mantyka-Pringle, et al. Predicting species distributions for conservation decisions. *Ecology letters*, 16(12):1424–1435, 2013.
- Antoine Guisan, Wilfried Thuiller, and Niklaus E Zimmermann. *Habitat suitability and distribution models: with applications in R*. Cambridge University Press, 2017.
- Garrett Hardin. The competitive exclusion principle. *science*, 131(3409):1292–1297, 1960.
- David J Harris. Inferring species interactions from co-occurrence data with markov networks. *Ecology*, 97(12): 3308–3314, 2016.
- David Heckerman, David Maxwell Chickering, Christopher Meek, Robert Rounthwaite, and Carl Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1(Oct):49–75, 2000.
- TD Heilbronn and David WH Walton. Plant colonization of actively sorted stone stripes in the subantarctic. *Arctic and Alpine Research*, 16(2):161–172, 1984.
- Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(Nov):1457–1469, 2004.
- Alexander von Humboldt, Aimé Bonpland, et al. *Essai sur la géographie des plantes*. Chez Levrault, Schoell et compagnie, libraires, 1805.
- GE Hutchinson. The multivariate niche. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 22, pages 415–421, 1957.
- Jens Kattge, Gerhard Bönsch, Sandra Díaz, Sandra Lavorel, Iain Colin Prentice, Paul Leadley, Susanne Tautenhahn, Gijbert DA Werner, Tuomas Aakala, Mehdi Abedi, et al. Try plant trait database-enhanced coverage and open access. *Global change biology*, 26(1):119–188, 2020.
- W Daniel Kissling, Carsten F Dormann, Jürgen Groeneveld, Thomas Hickler, Ingolf Kühn, Greg J McInerny, José M Montoya, Christine Römermann, Katja Schiffrers, Frank M Schurr, et al. Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography*, 39(12):2163–2178, 2012.
- C. König, R.W. Wüest, C.H. Graham, D.N. Karger, T. Sattler, N.E. Zimmermann, and D. Zurell. Scale dependency of joint species distribution models challenges interpretation of biotic interactions. *Journal of Biogeography*, 48: 1141–1151, 2021.
- Sadanori Konishi and Genshiro Kitagawa. *Information criteria and statistical modeling*. Springer Science & Business Media, 2008.

- S. Lavorel and E. Garnier. Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the holy grail. *Functional Ecology*, 16(5):545–556, 2002. URL <GotoISI>://WOS:000178119300001.
- Sandra Lavorel, Jonathan Storkey, Richard D. Bardgett, Francesco de Bello, Matty P. Berg, Xavier Le Roux, Marco Moretti, Christian Mulder, Robin J. Pakeman, Sandra Diaz, and Richard Harrington. A novel framework for linking functional diversity of plants with other trophic levels for the quantification of ecosystem services. *Journal of Vegetation Science*, 24(5):942–948, 2013. ISSN 1100-9233.
- Chieh Lo and Radu Marculescu. Mplasso: Inferring microbial association networks using prior microbial knowledge. *PLoS computational biology*, 13(12):e1005915, 2017.
- Robert H MacArthur and Edward O Wilson. *The theory of island biogeography*, volume 1. Princeton university press, 2001.
- Raphaëlle Momal, Stéphane Robin, and Christophe Ambroise. Tree-based inference of species interaction network from abundance data. *arXiv preprint arXiv:1905.02452*, 2019.
- Ignacio Morales-Castilla, Miguel G Matias, Dominique Gravel, and Miguel B Araújo. Inferring biotic interactions from proxies. *Trends in ecology & evolution*, 30(6):347–356, 2015.
- Tamara Münkemüller and Laure Gallien. Virtualcom: a simulation model for eco-evolutionary community assembly and invasion. *Methods in Ecology and Evolution*, 6(6):735–743, 2015.
- Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.
- Nhu H Nguyen, Zewei Song, Scott T Bates, Sara Branco, Leho Tedersoo, Jon Menke, Jonathan S Schilling, and Peter G Kennedy. Funguild: an open annotation tool for parsing fungal community datasets by ecological guild. *Fungal Ecology*, 20:241–248, 2016.
- Jenni Niku, Francis KC Hui, Sara Taskinen, and David I Warton. gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in r. *Methods in Ecology and Evolution*, 10(12):2173–2182, 2019.
- Marc Ohlmann, Florent Mazel, Loïc Chalmandrier, Stéphane Bec, Eric Coissac, Ludovic Gielly, Johan Pansu, Vincent Schilling, Pierre Taberlet, Lucie Zinger, et al. Mapping the imprint of biotic interactions on β -diversity. *Ecology letters*, 21(11):1660–1669, 2018.
- Otso Ovaskainen and Nerea Abrego. *Joint Species Distribution Modelling: With Applications in R*. Cambridge University Press, 2020.
- Otso Ovaskainen, Gleb Tikhonov, Anna Norberg, F Guillaume Blanchet, Leo Duan, David Dunson, Tomas Roslin, and Nerea Abrego. How to make more out of community data? a conceptual framework and its implementation as models and software. *Ecology Letters*, 20(5):561–576, 2017.

- Giovanni Poggiato, Tamara Munkemuller, Daria Bystrova, Julyan Arbel, James S. Clark, and Wilfried Thuiller. On the interpretations of joint modeling in community ecology. *Trends in Ecology & Evolution*, 36(5):391–401, 2021. ISSN 0169-5347. doi: 10.1016/j.tree.2021.01.002.
- Laura J Pollock, Reid Tingley, William K Morris, Nick Golding, Robert B O’Hara, Kirsten M Parris, Peter A Vesk, and Michael A McCarthy. Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (jsdm). *Methods in Ecology and Evolution*, 5(5):397–406, 2014.
- L.J. Pollock, L.M.J. O’Connor, K. Mokany, D.F. Rosauer, M.V. Talluto, and W. Thuiller. Protecting biodiversity (in all its complexity): new models and methods. *Trends in Ecology & Evolution*, 2020.
- Gordana C Popovic, Francis KC Hui, and David I Warton. A general algorithm for covariance modeling of discrete data. *Journal of Multivariate Analysis*, 165:86–100, 2018.
- Gordana C Popovic, David I Warton, Fiona J Thomson, Francis KC Hui, and Angela T Moles. Untangling direct species associations from indirect mediator species effects with graphical models. *Methods in Ecology and Evolution*, 10(9):1571–1583, 2019.
- H Ronald Pulliam. On the relationship between niche and distribution. *Ecology letters*, 3(4):349–361, 2000.
- Elizabeth L Sander, J Timothy Wootton, and Stefano Allesina. Ecological network inference from long-term presence-absence data. *Scientific Reports*, 7(1):1–12, 2017.
- Thomas W Schoener. Resource partitioning in ecological communities. *Science*, 185(4145):27–39, 1974.
- Marco Scutari, Maintainer Marco Scutari, and Hiton-PC MMPC. Package ‘bnlearn’. *Bayesian network structure learning, parameter learning and inference, R package version 4.4*, 1, 2019.
- Wilfried Thuiller, Tamara Münkemüller, Sébastien Lavergne, David Mouillot, Nicolas Mouquet, Katja Schiffrers, and Dominique Gravel. A road map for integrating eco-evolutionary processes into biodiversity models. *Ecology letters*, 16:94–105, 2013.
- Wilfried Thuiller, Laura J Pollock, Maya Gueguen, and Tamara Münkemüller. From species distributions to meta-communities. *Ecology letters*, 18(12):1321–1328, 2015.
- Wilfried Thuiller, Maya Gueguen, Julien Renaud, Dirk N. Karger, and Niklaus E. Zimmermann. Uncertainty in ensembles of global biodiversity scenarios. *Nature Communications*, 10, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-09519-w. URL <GotoISI>://WOS:000462722200003.
- Brian H Walker. Biodiversity and ecological redundancy. *Conservation biology*, 6(1):18–23, 1992.
- DA Wardle, GM Barker, KI Bonner, and KS Nicholson. Can comparative approaches based on plant ecophysiological traits predict the nature of biotic interactions and individual plant species effects in ecosystems? *Journal of ecology*, 86(3):405–420, 1998.

- David I Warton, F Guillaume Blanchet, Robert B O'Hara, Otso Ovaskainen, Sara Taskinen, Steven C Walker, and Francis KC Hui. So many variables: joint modeling in community ecology. *Trends in Ecology & Evolution*, 30(12):766–779, 2015.
- Evan Weiher and Paul Keddy. *Ecological assembly rules: perspectives, advances, retreats*. Cambridge University Press, 2001.
- Mary Susanne Wisz, Julien Pottier, W Daniel Kissling, Loïc Pellissier, Jonathan Lenoir, Christian F Damgaard, Carsten F Dormann, Mads C Forchhammer, John-Arvid Grytnes, Antoine Guisan, et al. The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological reviews*, 88(1):15–30, 2013.
- Damaris Zurell, Laura J. Pollock, and Wilfried Thuiller. Do joint species distribution models reliably detect interspecific interactions from co-occurrence data in homogenous environments? *Ecography*, 41(11):1812–1819, 2018. ISSN 0906-7590. doi: 10.1111/ecog.03315. URL <GotoISI>://WOS:000449886600006.

Uncovering symmetric and asymmetric species associations from community and environmental data

Supplementary Materials

Sara Si-moussi, Esther Galbrun, Mickael Hedde, Giovanni Poggiato, Matthias Rohr, Wilfried Thuiller

Contents

1	Supplements to framework description	2
1.1	Extensions of the biotic context definition	2
1.1.1	Adding conditioning covariates	2
1.1.2	Temporal extension	2
1.1.3	Spatial extension	2
1.1.4	Graph extension	3
2	Supplements to the virtual experiment 1	4
2.1	Simulation how-to	4
2.1.1	Notation	4
2.1.2	Assembly rules	4
2.2	Simulation experiment	5
2.3	Simulation diagnosis	7
2.4	Supplementary results	7
2.4.1	Sensitivity of associations to niche distances	7
2.4.2	Comparative performances on association type inference	8
3	Virtual experiment 2	10
3.1	Simulation how-to	10
3.2	Structural regularities conserved in the embeddings	10
3.3	Supplementary results	10
3.3.1	Network structure inference performances	10
3.3.2	Structural regularities captured by the embeddings	11
4	Supplements to the empirical application	12
4.1	Environmental data preparation	12
4.2	Framework adaptation and training	12
4.3	Embedding dimension and lasso parameter selection	12
4.4	Habitat suitability	13
4.5	Summary association network	13
4.6	Analyzing the functional meaning of plant embeddings	14

1 Supplements to framework description

1.1 Extensions of the biotic context definition

1.1.1 Adding conditioning covariates

In the base model, the estimation of any pairwise association is oblivious to the abiotic or biotic conditions surrounding it. To account for these neighborhood conditions, we extend the base model by allowing the embeddings used to represent the biotic context to vary according to some covariates.

Each site is associated to p conditioning covariates. These covariates are stored alongside an offset in a $n \times (p+1)$ matrix V , such that each of the first p columns of V contains the values of the corresponding covariate for the different sites while the last column is filled with ones. Then, given an embedding dimension d , the covariates are mapped to d dimensions by applying a regression with a weight matrix $W \in \mathbb{R}^{(p+1) \times (d)}$. The resulting conditioning vectors $\beta_k = Wv_k^T$ represent the relative weight associated to a given latent dimension depending on the conditions defined by the covariates.

The extended biotic context is then written as follows, where \odot is the element-wise vector product:

$$z_{ki} = \beta_k \odot \left(\frac{1}{|C_{ki}|} \sum_{j \in C_{ki}} y_{kj} \alpha_j \right) = \frac{1}{|C_{ki}|} \sum_{j \in C_{ki}} y_{kj} \cdot (\beta_k \odot \alpha_j)$$

The biotic associations can be recovered as in the base model, by isolating the pairwise associations in the response variable. However, in this case, the associations we obtain are represented by a three-dimensional tensor instead of a two-dimensional matrix. Each slice along the first dimension of this tensor represents a local association network.

$$a_{kij} = \sum_{l=1}^d (\beta_k \odot \rho_i \odot \alpha_j)_l$$

$$\eta_{ki} = f \left(\sum_{j \in C_{ki}} y_{ki} a_{kij} + o_j \right)$$

By incorporating the association covariates on the latent space, we gain two desirable properties. First, we get a fixed number of parameters that is a factor of the embedding dimension, which is significantly smaller than the number of modeled species. Second, we ensure species with similar latent traits, as captured by the response and effect embeddings, share associations regardless of the surrounding conditions.

1.1.2 Temporal extension

When longitudinal data are available, we denote the abundance of species i at site k at time-point t as $y_{ki}^{(t)}$. Accordingly, the definition of the biotic context for a target species at a given time-point is extended to contain the species, including the target, that were observed in the previous time-point:

$$C_{ki}^{(t)} = \{j \in \mathcal{S}, y_{kj}^{(t-1)} > 0\}$$

$$z_{ki}^{(t)} = \frac{1}{|C_{ki}^{(t)}|} \sum_{j \in C_{ki}^{(t)}} y_{kj}^{(t-1)} \alpha_j$$

1.1.3 Spatial extension

Given a function d that measures the distance between any pair of sites and a radius r , we consider a spatial extension of the base model where the biotic context is defined to contain species that were observed at locations within distance r of the considered site.

$$C_{ki} = \{(j, l) \in \mathcal{S} \times \mathcal{K}, y_{lj} > 0 \text{ and } d(k, l) \leq r\}$$

One can adapt the radius values to each species or group of species. The effect of each contextual element decays with distance to the target location, with a rate τ which controls the decrease in weight per unit of distance.

$$z_{ki} = \sum_{(j, l) \in C_{ki}} y_{lj} \cdot \exp(-\tau d(k, l))$$

1.1.4 Graph extension

So far, we have defined the biotic context based on the local community composition, using the presence or abundance of other species to model pairwise effects on the target species. However, this formulation does not capture higher-order associations among the context species themselves, nor the broader network structure surrounding the target, what we refer to here as the *contextual network*. To address this, we propose leveraging graph embedding techniques to compute latent representations of the contextual network.

2 Supplements to the virtual experiment 1

2.1 Simulation how-to

We used a process-based stochastic model adapted from Virtualcomm (Gallien and Münkemüller 2015) to simulate the assembly of individuals from a regional species pool into communities, on different locations sampled along an environmental gradient. The assembly process is controlled by three filtering mechanisms: the response to the abiotic environment, the outcome of biotic interactions and reproduction. For simplicity, the spatial structure of communities and thus dispersal processes are ignored. In other words, there is no exchange of individuals between neighboring communities. The simulation starts with a given or random initial composition for each community independently. Individuals are replaced through time until an equilibrium state is reached or a user-defined number of iterations is completed. The final communities' composition is returned at the end Fig. 1.

2.1.1 Notation

- We start by sampling n locations uniformly on a single environmental gradient E .
- All locations have the same carrying capacity of K individuals from a common pool of m species $S = \{S_j/j \in [1, m]\}$.
- Each species has its own optimal environmental value $\mu_j \in E$ as well as a niche breadth $\delta_j \in E$.
- Biotic interactions are described by a full interaction matrix $I = (I_{jk})/j, k \in [1, m]^2$; $-1 \leq I_{jk} \leq 1$ where I_{jk} represents the effect of the interaction between the pair (S_j, S_k) on species S_k . We also write: $I = I^+ - I^-$ such that:
 - $I^+ = (I_{jk}^+)/j, k \in [1, m]^2$; $0 \leq I_{jk}^+ \leq 1$ represents the matrix of positive effects (facilitation matrix)
 - $I^- = (I_{jk}^-)/j, k \in [1, m]^2$; $-1 \leq I_{jk}^- \leq 0$ represents the matrix of negative effects (competition matrix)

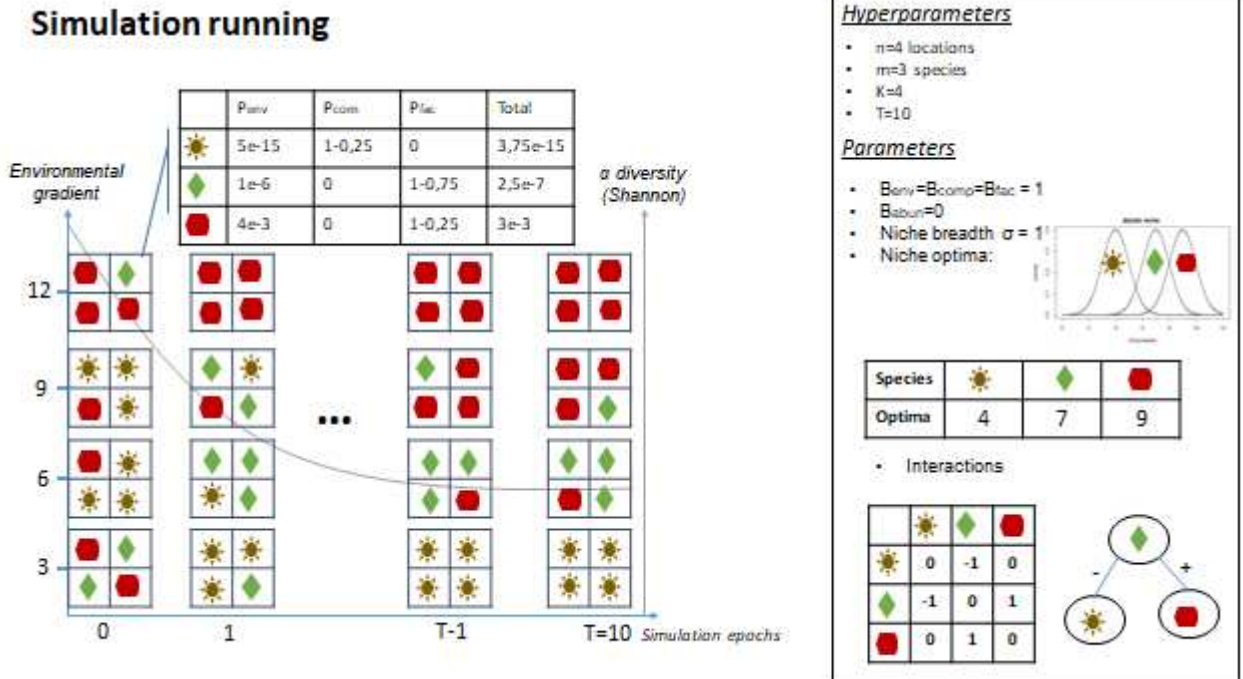


Fig. 1: Simulation procedure.

2.1.2 Assembly rules

At each timestep (epoch), given an actual composition c , the probability that an individual from a given species i to replace any other individual of c is given by the following equation ; such that:

- B_{env} : weights of the abiotic filter.
- B_{comp} : weight of the competition.

- B_{fac} : weight of the facilitation.
- B_{abun} : weight of the reproduction filter, can be interpreted in terms of growth rate.
- $P_{env,i,c}$: the probability of species i to occur under the environmental value E_c is given by the normalized density on E_c of a Gaussian distribution parameterized by its optimum and niche breadth. The closer to its optima, the higher the probability of the species' occurrence.
- $P_{comp,i,c}$: the probability for an individual of species i to join the community given the aggregated effect of its competitors in c .
- $P_{fac,i,c}$: the probability for an individual of species i to join the community given the aggregated effect of its facilitators in c .
- $P_{abund,i,c}$: probability of an individual of species i to join the community as a result of the reproduction of some of the $N_{i,c}$ conspecifics in c .

The unnormalized weights $W_{i,c}$ for each species are then normalized by dividing each one of them by their sum. The result is a vector of probabilities W that sums to 1. Finally, we sample from a multinomial distribution, parameterized with W , K individuals to compose the new community.

2.2 Simulation experiment

We set up an experiment where multiple simulations were run on random locations along a single environmental gradient ranging from 0 to 100 with different randomly selected configurations of the prior association matrix: absence of association (environmental filtering only), positive associations only, negative associations only and a mix of positive and negative associations. In each configuration mode, we varied the pool size, i.e. the number of species (10, 20 or 50), to test how the different models might be affected by the species pool size, the *density* in terms of *number of associated pairs* as a function of the pool size (sparse 1/3 or dense 2/3) and whether the association matrix included asymmetric effects: semi-attraction (e.g. commensalism) or semi-repulsion (e.g. amensalism). Positive (resp. negative) effects were all set to +1 (resp. -1) as we are interested in the polarity of the associations rather than their intensity. The factorial design of this experiment produced 33 simulation datasets Fig 2. These combinations allowed us to test our framework, but also to compare its ability to detect species symmetric associations in respect to other approaches like JSDMs and probabilistic graphical models.

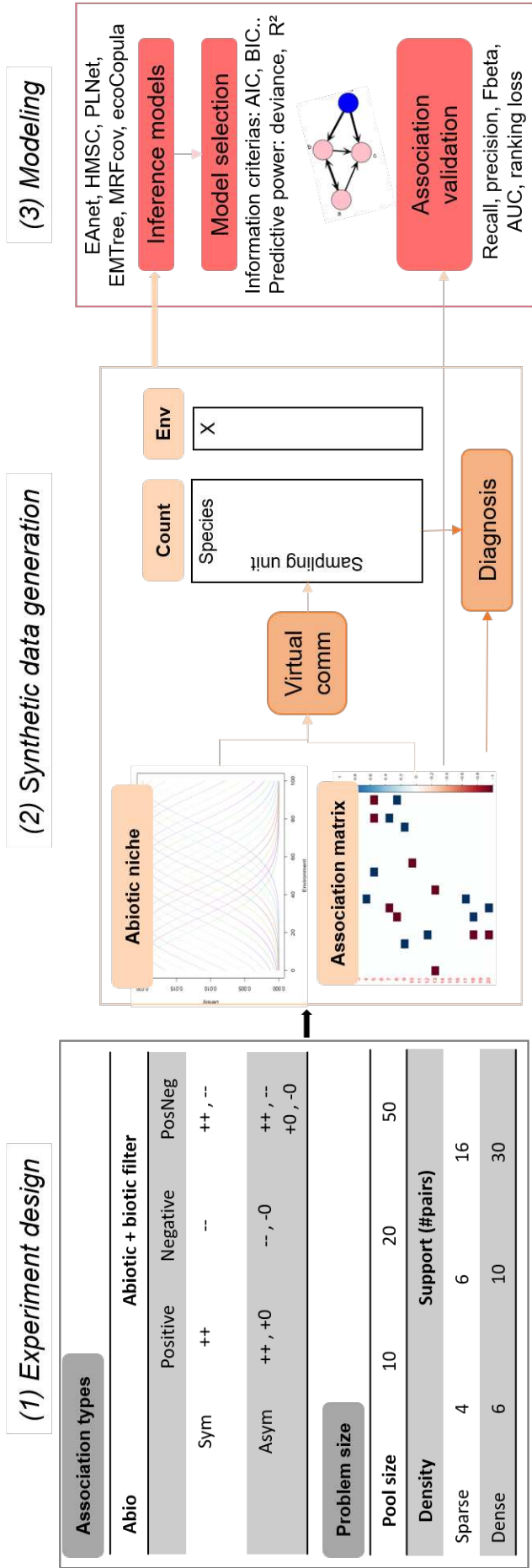


Fig. 2: Description of the simulation experiment 1 with different assembly rules

2.3 Simulation diagnosis

To assess whether the simulated virtual communities reflect the simulation parameters (e.g. two competing species tend not to co-occur), we defined the *relative abundance index* (RAI_{ji}), an asymmetric pairwise index that measures the average change in abundance of the target species i when the source species j is present as compared to its mean abundance irrespective of whether the source species j is present, \bar{y}_i .

$$\Delta_{kji} = \{y_{ki} - \bar{y}_i, \text{ for all } k \in \mathcal{K} \text{ such that } y_{ki} > 0 \text{ and } y_{kj} > 0\}.$$

Then $\text{RAI}_{ji} = \text{avg}(\Delta_{kji})$ across all k sites. The larger the standard deviation $\text{std}(\Delta_{kji})$, the more ambiguous the strength of the effect of species j on species i . If the confidence interval $\text{avg}(\Delta_{ji}) \pm 1.96 \text{std}(\Delta_{ji})$ does not contain zero, then the simulated dependencies unambiguously translate a polarized effect of species j on species i . Otherwise, the simulations led to an equilibrium for which it is not possible to retrieve the parameters.

Before fitting the inference models on the simulated data, we checked using an empirical measure of pairwise association RAI_{ij} whether species dependencies reflected the simulated patterns. Fig 3 depicts the value of the statistic for each directed association a_{ij} . The distribution of RAI_{ij} values showed a good discrimination of positive and negative associations, albeit with different strengths. Neutral associations translated into small RAI_{ij} values, median-centered on zero for simulations with negative associations only. Whereas they spanned a large spectrum of values in simulations with only positive or a mix of positive and negative effects. The proposed indicator is itself a good proxy for association inference, however since it does not account for environmental covariates we use it as a diagnosis tool.

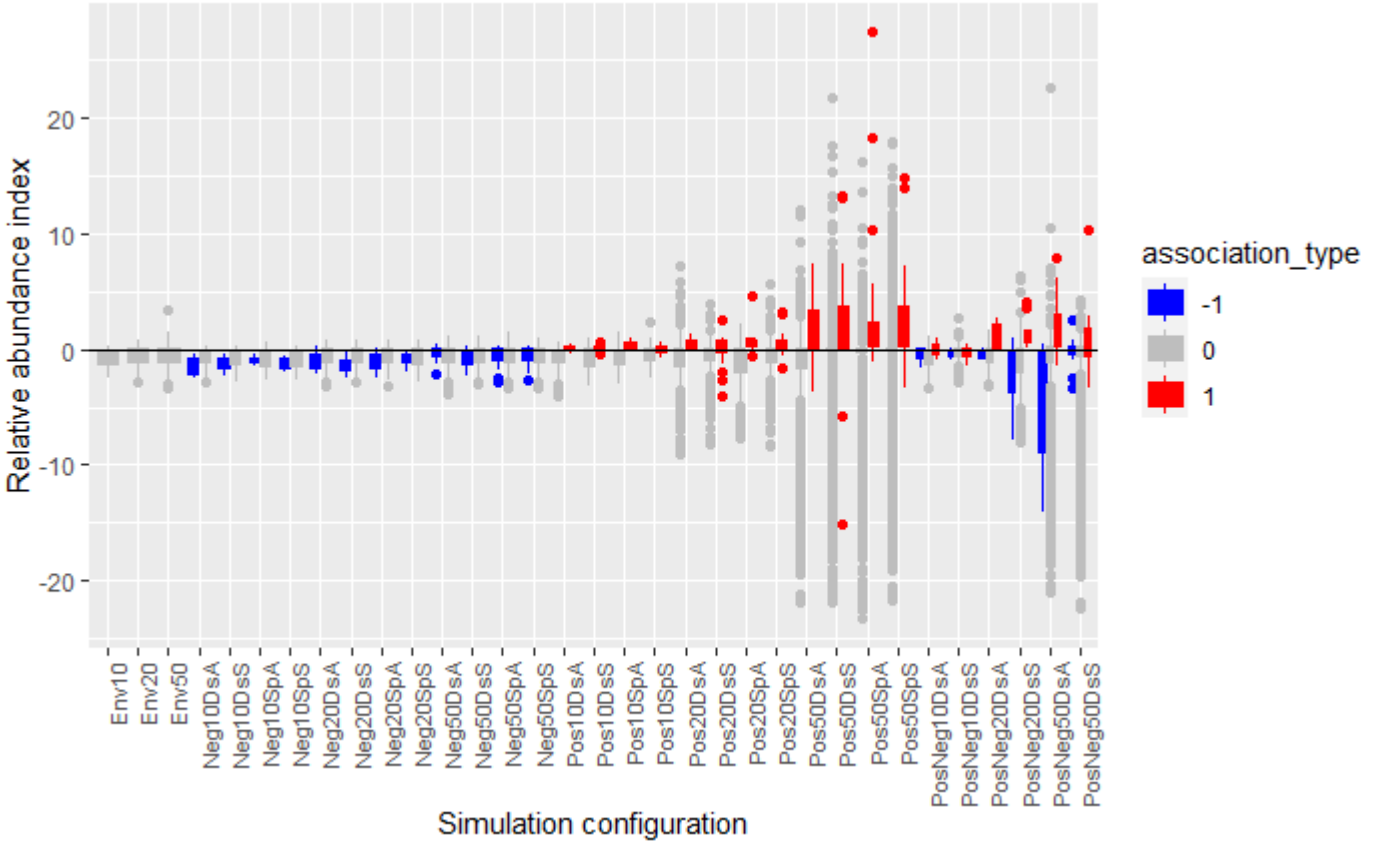


Fig. 3: **Simulation diagnosis.** Distribution of relative abundance indices RAI_{ij} per simulation. Each data point represents a directed association (positive in red, negative in blue and neutral in gray) involving two species from the corresponding simulation. Labels on the x-axis correspond to simulation configurations.

2.4 Supplementary results

2.4.1 Sensitivity of associations to niche distances

On larger pool sizes, the distance between niche optima decreases as the probability of niche overlap increases. Consequently, the estimates of species associations varied with the number of modeled species. Although the strengths appeared to be drawn from a small fixed interval, their value was sensitive to the species niche differences

(Fig. 4). For all tested frameworks, the strength inferred for true positive associations decreased with the amount of niche overlap. For large niche differences (no overlap), MRFcov, EMtree and PLN even reported opposite signs. Conversely, inferred strengths of true negative associations were either invariant to niche difference (MRFcov), very small and close to neutral (EMtree, PLN, ecocopula and EA) or increasing in absolute strength (HMSC). At medium to high niche distance, EA reported an increase in the absolute strength of negative associations. HMSC showed the opposite pattern, suggesting that the negative effects were rather explained by the abiotic covariates.

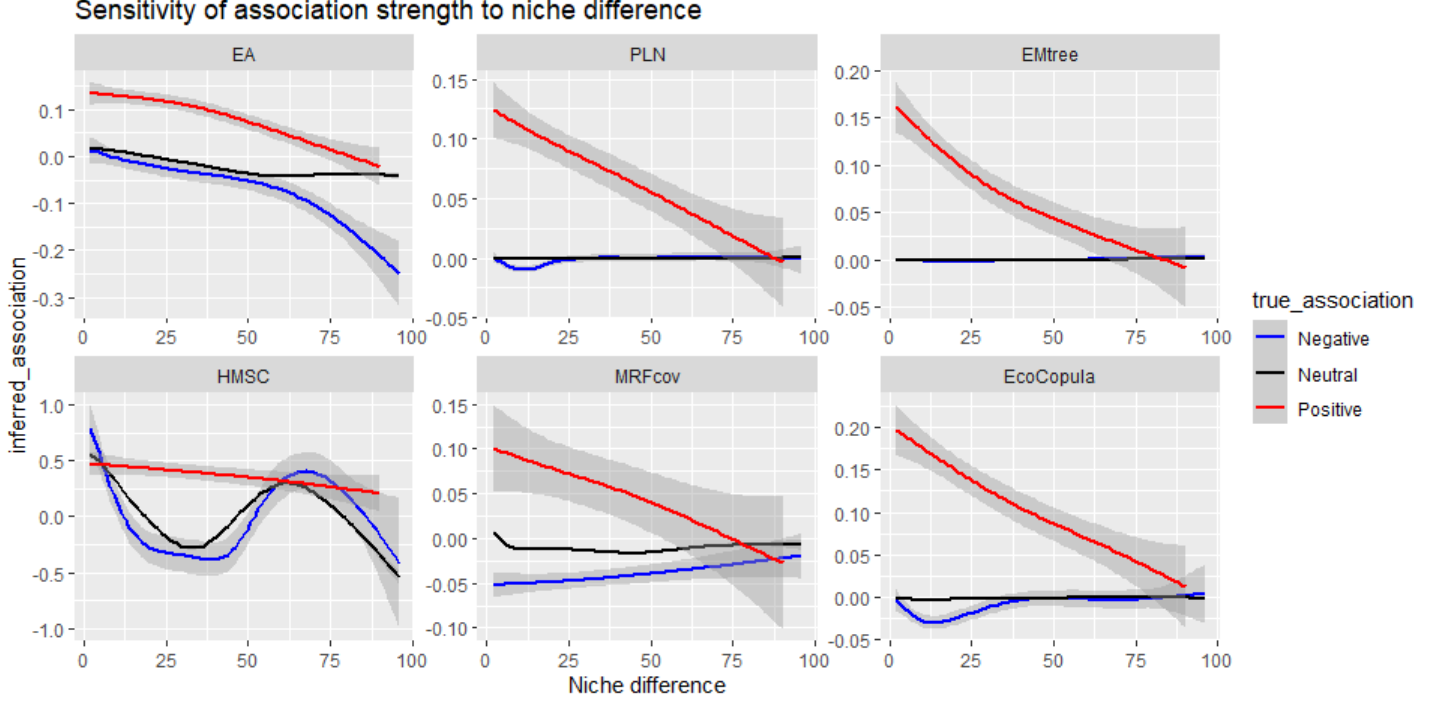


Fig. 4: Sensitivity of the inferred association strength a_{ij} per association type and inference model to the abiotic niche distance measured by the absolute difference between their niche optima $\mu_i - \mu_j$.

2.4.2 Comparative performances on association type inference

We reported the Area Under the Precision-Recall Curve (PR-AUC), the recall and f1-score in Fig 5 for each association type separately. We found no significant difference between the models in respect to their performances in inferring positive vs dense or symmetric/asymmetric datasets. However, the quality of inferred associations varied with the pool size.

On positive associations, EA and Ecocopula outperformed the other methods in all pool sizes. EMtree, PLN and MRFcov reported good performances for 20 and 50 species datasets, but they failed to detect positive associations in 10-species datasets. On negative associations, all models had strong difficulties in retrieving them, and this difficulty was more pronounced for large pool size. EA, MRFcov and HMSC outperformed other methods.

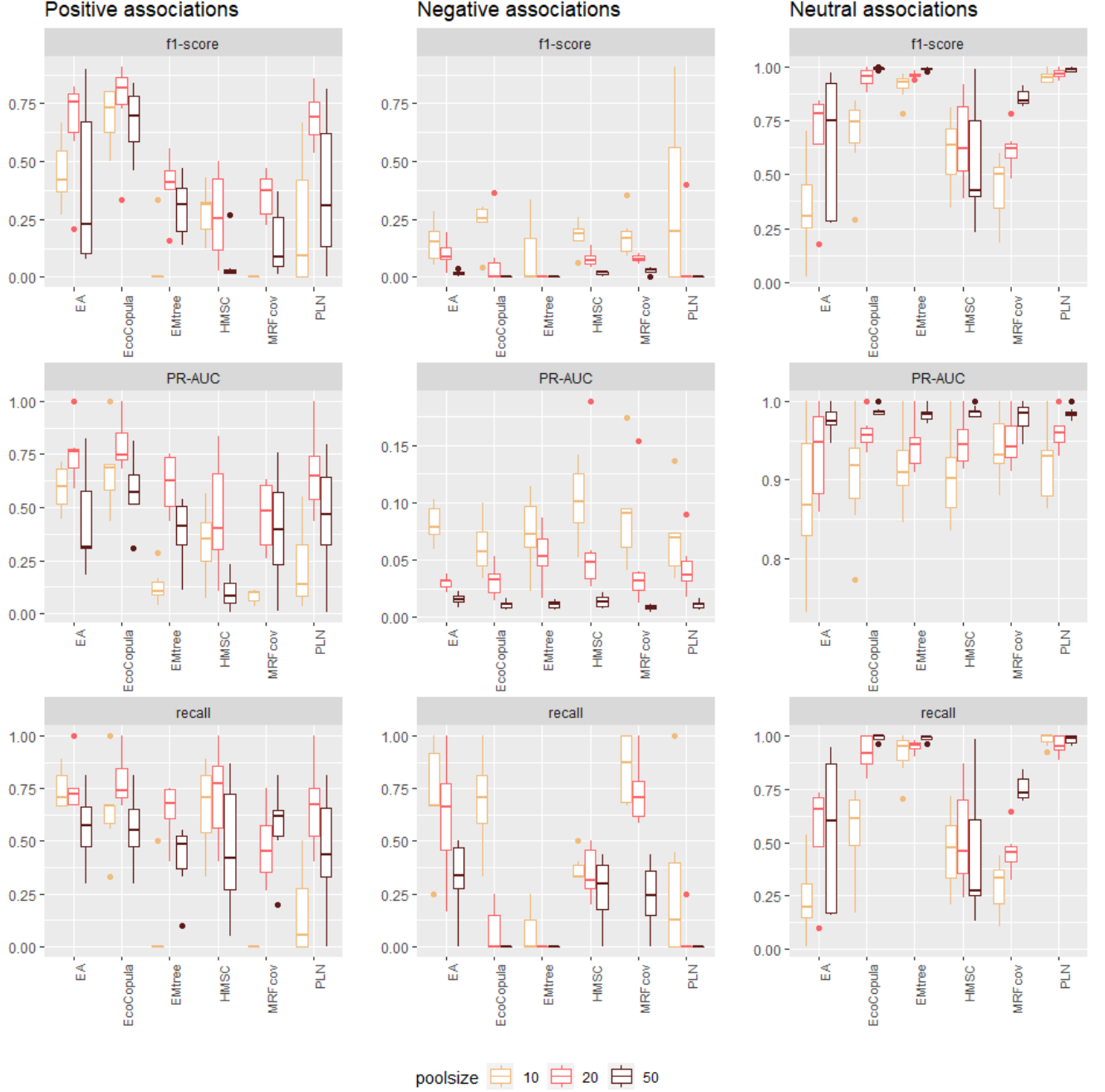


Fig. 5: Inference of true association class per type of association for each model measured by the recall, f1-score and AUC-PR metrics. The higher the value the better performing is the model. AUC-PR is computed on the raw associations while the recall and f1-score are computed on the discretized associations using as a threshold $\epsilon^+ = \epsilon^- = 5E - 2$

3 Virtual experiment 2

In this experiment, we simulated species occurrence data where each species depends jointly on suitable environmental conditions and the presence of at least one prey given by a known predator-prey network (food web). This setup reflects an intersection of abiotic and biotic filters, modeled as a multiplicative response. This multiplicative structure and the asymmetry of predator-prey interactions is not supported by other association inference methods. Therefore, we only evaluated our approach.

3.1 Simulation how-to

Fig 6 illustrates the experimental setup and the procedure used to generate occurrences. Briefly, we assume a bottom-up control so that to be present, a consumer requires the availability of *at least* one resource, in addition to habitat suitability. Basal trophic groups do not depend on any resource, only on the environment. The process resulted in six datasets, each containing 25 species and 500 sites.

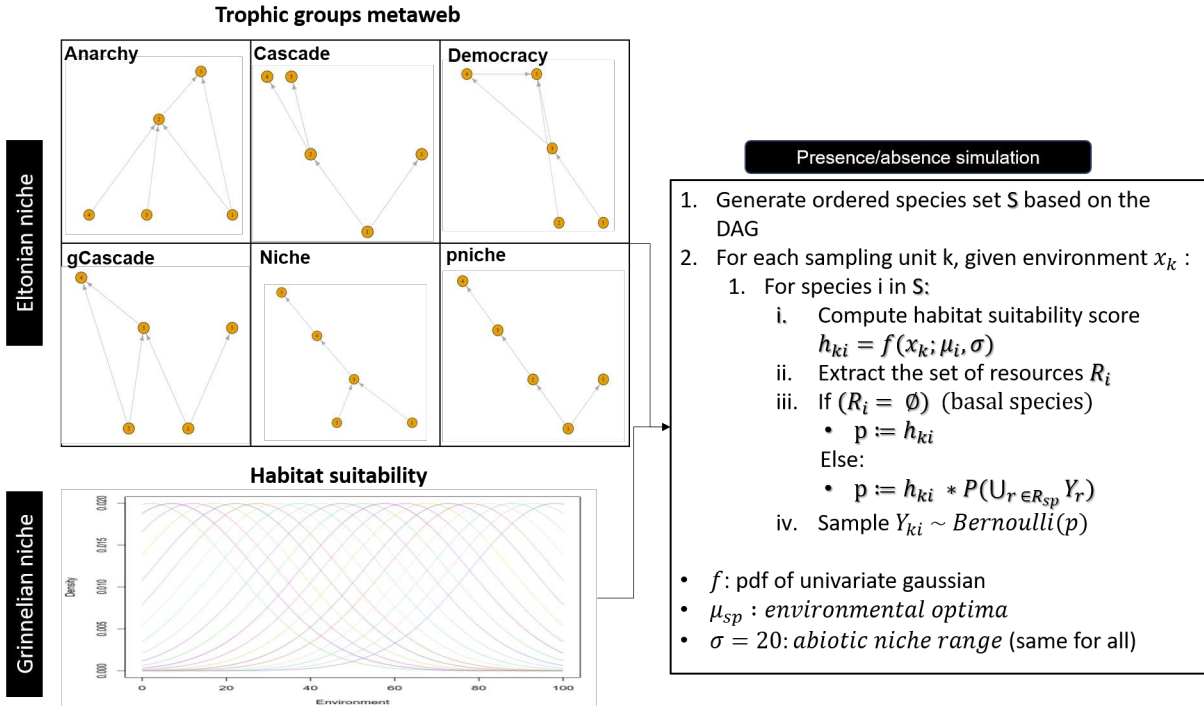


Fig. 6: Simulations of consumer-resource co-occurrences along an environmental gradient given food web topology and true abiotic niches.

3.2 Structural regularities conserved in the embeddings

For each topology, we investigated whether learnt embeddings reflect the underlying clustering of species into trophic groups. Concretely, we performed a Mann-Whitney ranking test [4] checking whether species from the same prior trophic group had more similar embeddings than species from different groups. Finally, we asked to what extent response and effect embeddings captured species abiotic preferences and their biotic requirements. We did that by testing the correlation between embedding similarity and niche overlap measured with environmental optima differences and the proportion of shared preys in both the potential and realized food webs.

3.3 Supplementary results

3.3.1 Network structure inference performances

The inferred associations were more faithful to the realized than the potential network. In all cases, incorporating a parameter-sharing constraint within trophic groups allowed to improve the sensitivity with respect to both ground truth networks.

The main source of error was the confusion of directed associations with symmetric reciprocal associations, introducing spurious associations. This error was particularly frequent between species of inner groups in trophic chains. However, the model managed to break this symmetry for species from groups with a higher diversity of preys (number of distinct groups) of preys.

	Additive abiotic and biotic responses				Intersection of abiotic and biotic requirements			
	Group-level		Species-level		Group-level		Species-level	
	MW	RN	MW	RN	MW	RN	MW	RN
Accuracy (ACC)								
min	0.45	0.80	0.72	0.82	0.45	0.62	0.24	0.38
max	0.83	0.94	0.83	0.94	0.88	0.96	0.72	0.87
median	0.790	0.890	0.765	0.895	0.780	0.735	0.615	0.735
mean (sd)	0.72 \pm 0.15	0.87 \pm 0.06	0.77 \pm 0.04	0.88 \pm 0.05	0.71 \pm 0.18	0.76 \pm 0.12	0.55 \pm 0.19	0.69 \pm 0.18
ROC-AUC (AUC)								
min	0.24	0.07	0.49	0.50	0.50	0.75	0.47	0.62
max	0.80	0.85	0.53	0.74	0.80	0.92	0.58	0.83
median	0.470	0.415	0.505	0.565	0.670	0.855	0.535	0.800
mean (sd)	0.52 \pm 0.22	0.43 \pm 0.27	0.51 \pm 0.01	0.58 \pm 0.09	0.64 \pm 0.12	0.84 \pm 0.07	0.53 \pm 0.04	0.77 \pm 0.08
F2-score (F2)								
min	0.00	0.00	0.00	0.00	0.29	0.50	0.27	0.49
max	0.56	0.65	0.19	0.44	0.68	0.77	0.52	0.55
median	0.120	0.220	0.105	0.220	0.520	0.590	0.415	0.520
mean (sd)	0.19 \pm 0.23	0.25 \pm 0.25	0.10 \pm 0.08	0.19 \pm 0.16	0.52 \pm 0.14	0.61 \pm 0.09	0.40 \pm 0.11	0.52 \pm 0.02
True Skill Statistic (TSS)								
min	-0.06	0.00	-0.01	0.00	0.11	0.37	0.02	0.30
max	0.36	0.77	0.06	0.46	0.60	0.83	0.16	0.63
median	0.000	0.150	0.005	0.120	0.335	0.715	0.085	0.600
mean (sd)	0.08 \pm 0.16	0.24 \pm 0.30	0.02 \pm 0.03	0.15 \pm 0.17	0.34 \pm 0.16	0.68 \pm 0.16	0.09 \pm 0.05	0.54 \pm 0.13

Table 1: Summary of trophic network structure inference performances using two architectures: Additive abiotic and biotic effects, intersection of requirements (reported in the main text), with associations at the species level or the group level (by sharing embeddings within prior trophic groups). Metrics are computed taking in turn the potential **MW** (metaweb) and the realized **RN** network as ground truth. We highlight the best median score for each metric and reference network.

3.3.2 Structural regularities captured by the embeddings

For all topologies, the ranking test showed that both response and effect representations were significantly more similar when species belonged to the same trophic group. Moreover, species embedding similarity correlated positively with both abiotic and biotic niche similarity, with variable significance levels across topologies. Response and effect representations similarity were strongly correlated for two reasons: (1) species with similar responses were likely to have similar effects and vice-versa, (2) both representations were very similar, which explains the symmetry in some of the inferred associations.

	Anarchy	Democracy	Cascade	gCascade	Niche	pNiche
Internal clustering validation						
response	38278(<0.001)	36408(0.0021)	40024(<0.001)	35983(0.0044)	37188(<0.001)	38100(<0.001)
effect	36683(0.0013)	37300(<0.001)	39164(<0.001)	36708(0.0013)	36448(0.002)	38188(<0.001)
Abiotic niche similarity						
response	0.21(<0.001)	0.13(0.0014)	0.13(<0.001)	0.16(<0.001)	0.23(<0.001)	0.16(<0.001)
effect	0.21(<0.001)	0.12(0.0033)	0.18(<0.001)	0.24(<0.001)	0.07(0.065)	0.15(<0.001)
Eltonian niche similarity						
response	0.04(0.29)	0.1(0.015)	0.15(<0.001)	0.09(0.03)	0.12(0.0039)	0.13(0.0011)
effect	0.04(0.31)	0.11(0.0052)	0.2(<0.001)	0.08(0.06)	0.08(0.041)	0.15(<0.001)
Filtered prey similarity						
response	0.14(<0.001)	-0.03(0.39)	0.34(<0.001)	-0.11(0.0052)	0.13(0.0013)	0.03(0.48)
effect	0.13(<0.001)	0.02(0.59)	0.31(<0.001)	-0.09(0.033)	0.05(0.25)	0.08(0.035)

Table 2: Structural regularities captured by the response and effect embeddings, for each food web topology.

4 Supplements to the empirical application

4.1 Environmental data preparation

The plant dataset contained the following set of environmental variables:

slope : the slope inclination in degrees,
snow : the average snowmelt date in Julian days between 1997 and 1999,
physd : the percentage of non vegetated soil due to physical processes,
zoogd : the percentage of non vegetated soil due to marmot activity,
aspect : the relative south aspect, and
form : the microtopographic landform index.

We initially applied a one-hot encoding scheme to the two categorical features (aspect and form) and we scaled the numerical features.

4.2 Framework adaptation and training

We split the observations into a training and a test dataset using a multi-label stratification scheme¹ to ensure that all species were covered and their proportions were preserved in both sets.

For each plant species, we pre-trained a generalized linear model (GLM) with a logit link to relate species occurrences to the environmental variables. We used the learnt weights as initial parameter values in the habitat suitability component of our framework.

We defined the biotic context for a target species as the set of plants observed on the location of interest. We used a negative binomial distribution to fit the plant counts. The embedding vectors were initialized using random samples from a uniform distribution on the $[-0.01, 0.01]$ interval, and subjected to lasso penalties to promote sparsity. Finally, the offset value for each species was set to its average count on occurrence points.

We trained the full model using stochastic gradient descent (with a learning rate of 0.01 and momentum of 0.8) on the training dataset using a subsampling rate of 25% for the negative examples. We monitored the negative log-likelihood of positive examples (presences) on the validation set after each full pass of the training set to assess the convergence of the training. We stopped when the loss stops decreasing or when 200 epochs have elapsed.

4.3 Embedding dimension and lasso parameter selection

The first step in this evaluation was to find appropriate values for the hyperparameters of our model. For a species pool of size m , the embedding dimension d is selected among powers of 2 up to $m/2$, to improve hyperparameter search speed. In our case, with $m = 82$, the embedding dimension is chosen from the set $\{2, 4, 8, 16, 32\}$.

When the value of the lasso penalty parameter λ becomes large, some components of the embedding vectors take extremely small values for all species (below 10^{-5}). These components have no effect on the computed associations. Removing them, shrinks the embeddings to a smaller effective dimension, equal to the number of retained components. In the extreme, very high values of λ lead to effective dimension equal to zero, resulting in a zero association matrix, so that the interaction model is only parameterized by the species offset counts.

For each value of d , we apply the training procedure described previously with increasing values of $\lambda \in \{0.01, 0.015, 0.02, 0.025\}$. We evaluate the resulting models on the test set by computing the effective dimension and the deviance of the predicted counts on positive examples (Fig. 3).

¹Python library `scikit-multilearn`: <http://scikit.ml/>

roles within the modules to create the group-level network. (See Supplementary Results). Finally, we analyzed the resulting patterns in light of existing literature on Alpine plants interactions [2].

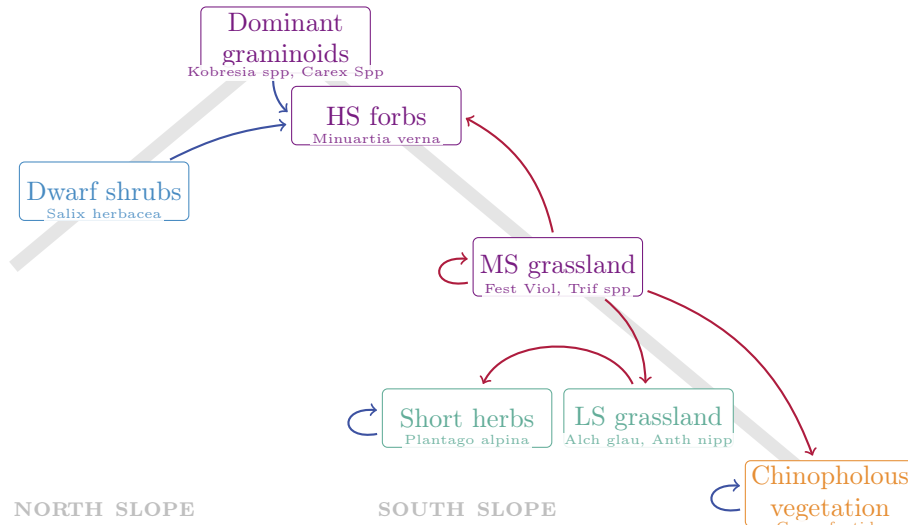


Fig. 8: The summary association network. Structural roles (nodes) are mapped to position in the gradient (Higher-slope HS, mid-slope MS, lower-slope LS) and plant classes (graminoids, grasses/herbs, forbs) and network modules (node colors). Edges go from a source (effect group) to a target (response group). Blue (resp. red) edges represent positive (resp. negative) associations.

4.6 Analyzing the functional meaning of plant embeddings

We investigated the functional determinants of the associations diversity. To do so, we compute the mutual information between the learnt embeddings and the plant traits (reported in [1]). The Mutual Information [6] is an unbounded symmetric and positive score that measures the amount of information contained in one random variable about another. It quantifies the reduction in uncertainty about one random variable given knowledge of another. Zero mutual information indicates independence.

In general, we expect traits related to dispersal capabilities (seed mass, spread) to impact the prevalence of the species, consequently increasing or decreasing the opportunity to affect other species (interaction probability). As a result, we expect such traits to have a higher mutual information with effect embeddings than with response embeddings. Conversely, traits related to nutrient uptake and biomass accumulation potential capture competitive or cooperative abilities of the plant species. Hence, we would expect a high mutual information between these traits and both responses and effects embeddings.

There was a relatively significant contribution of the leaf nitrogen mass and spread to the plants response, whereas leaf angle was found independent (Fig. 9). The Specific Leaf Area contributes significantly to the effect in addition to the Nitrogen mass and on a lesser extent Spread. Height is reported as related to both parameters.



Fig. 9: Mutual information between plant traits and their latent representations. Each bar concerns a specific trait, it represents the stack of mutual information scores from the first to the last (fourth) embedding dimension. The lower (resp. upper) figure shows the results for the response (resp. effect) embeddings.

References

- [1] CHOLER, P. Consistent shifts in alpine plant traits along a mesotopographical gradient. *Arctic, Antarctic, and Alpine Research* 37, 4 (2005), 444–453.
- [2] CHOLER, P., MICHALET, R., AND CALLAWAY, R. M. Facilitation and competition on gradients in alpine plant communities. *Ecology* 82, 12 (2001), 3295–3308.
- [3] GAUZENS, B., THÉBAULT, E., LACROIX, G., AND LEGENDRE, S. Trophic groups and modules: two levels of group detection in food webs. *Journal of The Royal Society Interface* 12, 106 (2015), 20141176.
- [4] MANN, H. B., AND WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.
- [5] NEWMAN, M. E. Modularity and community structure in networks. *Proceedings of the national academy of sciences* 103, 23 (2006), 8577–8582.
- [6] SHANNON, C. E., AND WEAVER, W. A mathematical model of communication. *Urbana, IL: University of Illinois Press* 11 (1949).