Getting More by Asking for Less: Linking Species Interactions to Species Co-Distributions in Metacommunities

Matthieu Barbier*[,1,2,3], Guy Bunin[4], and Mathew A. Leibold[5]

June 4, 2023

* Corresponding author. E-mail: contact@mrcbarbier.org

[1] CIRAD, UMR PHIM, 34090, Montpellier, France

[2] PHIM Plant Health Institute, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, 34090 Montpellier, France

[3] Institut Natura e Teoria en Pirenèus, 09400 Surba, France

[4] Department of Physics, Technion-Israel Institute of Technology, Haifa 3200003, Israel

[5] Biology Department, University of Florida, Gainesville, Florida, USA

Abstract

One of the more difficult challenges in community ecology is inferring species interactions on the basis of patterns in the spatial distribution of organisms. At its core, the problem is that distributional patterns reflect the 'realized niche', the net result of a complex interplay of processes involving dispersal, environmental, and interaction effects. Disentangling these effects can be difficult on at least two distinct levels. From a statistical point of view, splitting a population's variation into contributions from its interaction partners, abiotic environment and spatial proximity requires 'natural experiments' where all three factors somehow vary independently from each other. On a more theoretical level, it is not even clear how to meaningfully separate these processes: for instance, species interactions could depend in many ways on the state of the environment, and these two processes may combine in highly non-additive ways. Here we show that these issues arise almost unescapably, even in a simple theoretical setting designed to minimize them. Using a model of metacommunity dynamics where direct species interactions are assumed to be context-independent, we show that inferring these interactions accurately from cross-species correlations is a major challenge under all but the most restrictive assumptions. However, we also find that it is possible to estimate the statistical moments (mean value and variance) of the species interactions distribution much more robustly, even if the precise values cannot be inferred. Consequently, we argue that study of multi-species spatial patterns can still be informative for theoretical approaches that build on statistical distributions of species parameters to predict macroscopic outcomes of community assembly.

A central issue in community ecology is to identify which processes and mechanisms are most important in determining species presence and abundance across space and time in given communities. While this can be accomplished by carefully designed experimental methods, this is often logistically impractical and indirect inference is used: we attempt to infer parameter values for a **process**-based model via the analysis of naturally-occurring **patterns** of species distributions. While early work focused on the direct analysis of co-distributions (i.e. correlations) in the abundances or occurrences of species, there has, over time, become apparent that this can be misleading and an increasingly sophisticated set of analytical tools that try to do this has emerged.

To give context to this issue, we examine the challenges in using distribution patterns to evaluate species interactions in the critiques and subsequent debate in response to Diamond (1975)'s 'Assembly Rules' that were proposed to explain coexistence patterns in relation to interspecific competition in community and biogeographic data. Diamond (1975)'s assembly rules focused especially on looking at the significance of negative co-distributions in community patterns among potentially competing species (i.e. so called 'checkerboard' patterns of coexistence). While a long-lasting debate ensued on the statistical significance of these patterns (Connor & Simberloff, 1979; Gotelli & McCabe, 2002) and the use of 'null models' (Gotelli & Graves, 1996), few, if any, questioned the basic hypothesis that negative co-distributions were in fact robust indicators of competition until much later.

More recent work has increasingly recognized the confounding roles of environment, disturbance, isolation and dispersal and proposed more sophisticated methods for the study of co-distributions, e.g. (Patterson & Atmar, 1986; Leibold & Mikkelson, 2002; Peres-Neto *et al.*, 2006). These issues have been resurfacing especially in microbial ecology due to a tide of data (Berry & Widder, 2014). Additionally, there has developed a broader focus on other types

of species interactions (e.g. Cazelles et al. 2015). The analysis of species interactions in general, and accounting for environmental, space and dispersal have been merged by the more recent development of more sophisticated methods such as joint species distribution models e.g. (Ovaskainen *et al.*, 2017) and related methods that aim to partition distribution patterns in relation to 'abiotic' (environmental), 'biotic' (involving interactions among species) and 'movement' (involving spatial effects mediated by dispersal effects). While the sophistication of such methods is rapidly developing, especially in addressing computational and statistical issues, this work is increasingly revealing that inferring process from pattern is not trivial (Barner *et al.*, 2018; Thurman *et al.*, 2019; Blanchet *et al.*, 2020; Poggiato *et al.*, 2021).

In parallel with these developments in biogeographic ecology, the same issues arise in work that is more narrowly focused on inferring species interactions from relative abundance patterns in homogeneous communities. Here the data and the inferences ignore the environmental and dispersal dynamics, usually by applying the studies to situations where these are constrained (experimentally) or assumed to be negligible (by the selection of data). The roots of these approaches can be found in early work focusing on indirect effects involving species interactions in multispecies communities (Levine, 1976; Lawlor, 1979) who recognized that pairwise species interactions could be separated into direct components (where species affected each other directly by proximate effects on birth or death rates) or indirect effects (where the effect of a species on another was mediated by direct effects on other mediating species by chains of direct effects). This work highlighted that observable effects of species on each other were most often related to 'net effects' that involved the entire network of possible indirect and direct effects and showed that net interaction between pairs of species involving such indirect effects might have little similarity to the direct interactions that drive them. These authors proposed that such net interactions could be derived in the case of linear models as the coefficients in the inverse of the community matrix that describes direct interactions, suggesting one way forward for analyzing community data to interpreting the resulting distributions (see below).

Schaffer (1981), pointed out that this was not likely to work in the case where qualitative differences in the species composition (presence of absence of individual species in the community) could strongly, and even qualitatively, modify the nature of net interactions among pairs of species. This problem could even involve rare (and thus perhaps hard to detect) species effects. More recently the idea that the inverse of the community matrix could help model net effects has been tentatively invoked as a way forward for interpreting co-distribution patterns (Ovaskainen *et al.*, 2017), but this assumes that the variation is driven by environmental variables that affect the carrying capacity of individual species (and not the interaction coefficients) and that the composition of the community (presence-absence) does not change over the metacommunity. This method is consequently not likely to be sufficiently robust to apply except under very highly controlled or limited conditions.

Nevertheless, the analysis of distribution and co-distribution patterns of species in metacommunities show that while these can be complex, some features can often be related to processes that generate them (Ovaskainen *et al.*, 2017; Ovaskainen & Abrego, 2020; Leibold *et al.*, 2022). For instance, it may be possible to obtain good estimates of the *average* competition strength in a community (Fort, 2018). On the other hand, deriving more exhaustive parameter estimates is still challenging (e.g. (Blanchet *et al.*, 2020; Poggiato *et al.*, 2021; Leibold *et al.*, 2022). Inspired by theory on community assembly in highly diverse biotas under the simplifying assumptions of 'highly disordered' interactions (Barbier *et al.*, 2018), we hypothesized that statistical features (e.g. mean and variance of either direct or net effects) may be a more robust inference target than the precise network of interactions. We further wondered if these statistical properties might also be robust to some of the other concerns outlined above such as the existence of composition change, the amount of dispersal and degree of environmental variation.

Here, we propose to explore limits to inference that stem from the very nature of the dynamical processes, which entangle contributions from the environment and from multiple species in ways that might or might not be possible to disentangle at all. We use process-based simulations of community assembly in disordered communities (Barbier *et al.*, 2018) to generate simulated data under different conditions involving species niches, environmental gradients, dispersal and interaction strength. We then infer parameters from the analysis of resulting patterns to evaluate how well they can be used under these various scenarios.

Given that previous studies have already established that biotic interactions can strongly bias our estimates of species' environmental preferences (Poggiato *et al.*, 2021), we go one step further to show that, even in a setting where we could assume a good estimate of these environmental preferences, we may still be unable to correctly extract species' interactions from their spatial co-occurrence. Yet we find that, even when species interactions cannot be estimated in detail, it remains possible to correctly infer their community-level statistics, i.e. how strong and diverse biotic interactions are overall in the ecological dynamics of the community. This suggests that we may extract more robust information out of spatial patterns by asking for a less detailed description of the underlying processes.

1 Methods

To evaluate how we might infer processes involving species interaction coefficients from patterns in species distributions in a landscape or metacommunity, we considered a highly simplified modeling framework as a limiting case. We assumed that species interacted in a spatially continuous lattice in which local (within cells) interactions could be described by simple Lotka-Volterra equations connected by dispersal from nearby cells. Previous work on inferring species interactions from possible distribution patterns suggested that this could work under similar assumptions, at least under some limiting conditions) (Levine, 1976; Lawlor, 1979). We assume that more complex assumptions (e.g. non-linear interactions, heterogeneous dispersal etc), would make the inferences we are interested in even less likely and our approach should thus be seen as an 'optimistic' evaluation. In essence, we are asking "how well can we **hope** to do in making such inferences with current approaches?"

We intuit that a significant obstacle to inferring the details of species interactions is the covariation between the abundances of all interaction partners, and between each of those and the environmental factors: we cannot discriminate how strongly each of these factors impacts the presence of any given species if our observations do not provide 'natural experiments' where they vary independently (or actual experiments that impose various species compositions in the same environmental conditions (Barbier *et al.*, 2021)). Therefore, we start by focusing on two highly contrasting cases. In one case, we model a scenario that is most likely to be successful in inferring interactions parameters from distributions, because many species compositions are realized for each environmental condition in the absence of local dispersal. We then contrast this with a scenario that includes dispersal that allows environmental tracking by species in the metacommunity.

Within these two scenarios, we analyze a number of simulations that vary in the mean and variance of species' interaction strengths, while remaining in a dynamical regime of 'species sorting' (Leibold *et al.*, 2004) where we expect species abundances to reach a stable equilibrium that reflects how favored they are in their local abiotic and biotic environment. In each simulation, we then attempt to infer interactions in detail, as well as derive their mean and variances.

1.1 Model setting

We consider a metacommunity on a 2D landscape of 64×64 pixels, each represented by a coordinate vector x. At each point, we model a single environmental factor $E(x)$ whose values range in $[-50, 150]$ (Fig. 1). We then simulate Lotka-Volterra dynamics with dispersal from neighboring patches y

$$\frac{dN_i(x,t)}{dt} = N_i(x,t)\left(K_i(x) + A_{ii}N_i(x,t) + \sum_{j\neq i} A_{ij}N_j(x,t)\right) + D\sum_{y\in \mathrm{nei}(x)}\left(N_i(y,t) - N_i(x,t)\right) \tag{1}$$

Furthermore, species are considered extinct when $N_i < N_c = 10^{-3}$: their abundances are set to zero, and only allowed to vary if their net growth rate $d\log N_i/dt$ becomes positive (corresponding to an invasion).

Intra-specific competition is set to $A_{ii} = -1$ for all species, while inter-specific interaction coefficients A_{ij} are independent of the environment and drawn randomly for each species pair (i, j) from a normal distribution with prescribed mean $\langle A\rangle$, standard deviation $\mathrm{std}A$, and symmetry $\mathrm{sym}A = \mathrm{corr}(A_{ij}, A_{ji})$.

The coefficients K_i determine each species' carrying capacities (equilibrium abundance in monoculture) since $A_{ii} = -1$, and thus $N_i = K_i$ at equilibrium in the absence of other species and of dispersal. These carrying capacities are modelled using a unimodal "niche" function of the environmental factor:

$$K_i(x) = e^{-(E(x)-c_i)^2/2w_i^2} - \mu_i \tag{2}$$

with "mortalities" μ_i drawn uniformly between 0 and 0.5, niche centers drawn uniformly $c_i \in [0, 100]$ and widths w_i normal i.i.d. (see parameters in SI). The addition of μ_i ensures that $K_i < 0$ when the environment deviates enough from species' optimum, i.e. species may not grow at all in sites that are too unfavorable.

Previous theoretical and numerical work (Bunin, 2017; Zelnik *et al.*, 2021) has allowed us to choose parameters in this model in order to select a regime of 'species sorting', i.e. species abundances reaching a stable equilibrium in each site based on local environment and interactions: this requires that $\mathrm{std}A$ is small enough to avoid loss of stability leading to complex nonequilibrium dynamics (Bunin, 2017), and D is small enough to avoid significant source-sink dynamics and mass effects, i.e. situations where local abundance are strongly driven by fluxes from neighboring sites (Leibold *et al.*, 2004; Zelnik *et al.*, 2019). These other situations may also be of ecological relevance, but the regime considered here was the most appropriate considering the questions we wish to tackle, as argued in Discussion.

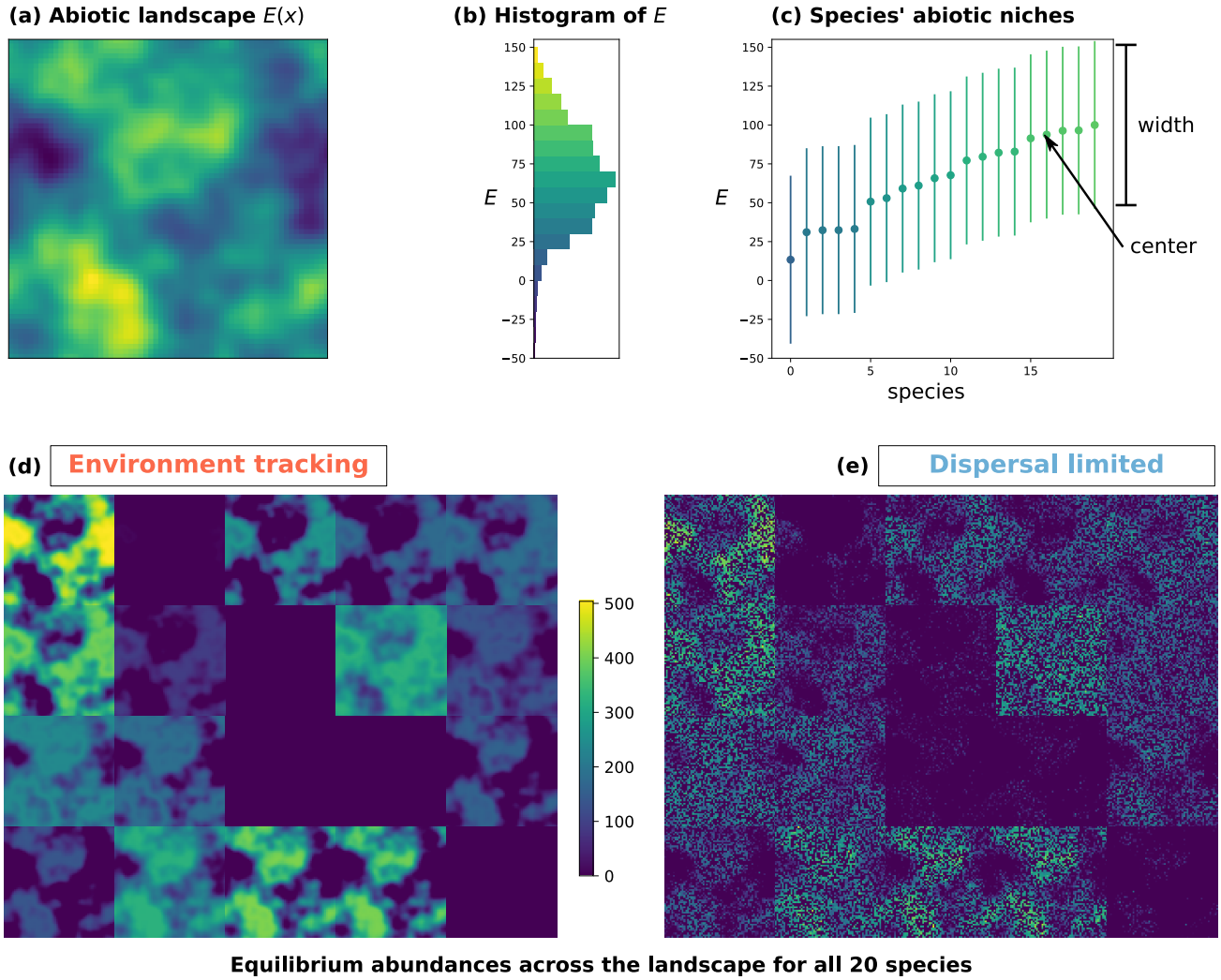

Figure 1: Environmental factor and final species abundances across the landscape in an example simulation run. **(a)** Landscape $E(x)$ with x a two-dimensional coordinate vector (64×64 pixels). **(b)** Histogram $P(E)$ of values of the environmental factor. **(c)** Centers c_i (dots) and width w_i (bars) of abiotic niches for all 20 species. **(d-e)** Equilibrium abundances $N_i(x)$ for all species (rescaled by extinction threshold $N_c = 10^{-3}$ so that values less than 1 indicate local extinction). The two colonization scenarios described in Sec. 1.2 are represented here: **(d)** environment tracking, where every species is initially seeded in every patch (or are allowed small but non-zero dispersal), and **(e)** dispersal limitation, where species are seeded independently in 50% of patches and cannot disperse. We note that several species go fully extinct at the landscape level from competition in the environment tracking scenario **(d)** whereas all species are present in at least part of the landscape in **(e)**. Parameters: $\langle w_i \rangle = 50$, $\langle A_{ij} \rangle = -0.3$, $\mathrm{std}(A_{ij}) = 0.09$, $D = 0$.

1.2 Colonization scenarios and dispersal

We expect that interactions can be inferred more successfully in biodiversity experiments where different species compositions are imposed in the same environmental conditions (Barbier *et al.*, 2021). This suggests testing two distinct scenarios for how species are distributed in the landscape (Fig. 1):

(DL) **"Dispersal limited"** scenario where each species is only seeded (i.e. given positive initial abundance) in half of the patches at random, and cannot disperse between patches.

(ET) **"Environment tracking"** scenario where all species are initially seeded in every cell (or can freely disperse, usually setting $D = 10^{-3}$, see below), and survive or go extinct deterministically because of abiotic and biotic conditions.

We will also vary the dispersal coefficient D, to check whether the DL scenario disappears and ET prevails as soon as $D > 0$ or at some higher value of dispersal.

1.3 Defining direct and net effects

The matrix A_{ij} represents *direct effects*, i.e. the instantaneous impact of species j's abundance N_j on the per capita growth rate of species i at a given time t. By assumption, in the generalized Lotka-Volterra equation (1), these effects are context-independent – they are characteristic of each species pair, and fixed across environmental conditions and across the landscape. This provides an important test case for our ability to infer species interactions, since it entails that we can truly assign a 'ground truth' value to these interactions that we may hope to recover through some inference method.

In the absence of dispersal ($D = 0$), the equilibrium condition for the subset $s(x)$ of species that coexist at location x is given by equation (1) with the content of the parentheses set to 0 (since $dN_i/dt = 0$ but $N_i > 0$)

$$0 = K_i(x) + \sum_{j \in s(x)} A_{ij} N_j(x,t). \tag{3}$$

This linear system of equations can therefore be inverted (Levine, 1976) to yield the equilibrium abundances

$$N_i^*(x,t) = -\sum_{j \in s(x)} \left(A^*(x)\right)^{-1}_{ij} K_j(x) \tag{4}$$

with $A^*(x)$ the submatrix of A restricted to the survivors at site x. As a consequence, we can define

$$V(x) = -\left(A^*(x)\right)^{-1} \tag{5}$$

the matrix of *net effects* at site x, which represents the long-term consequences of interactions: how changing the carrying capacity K_j of species j (making the local abiotic or biotic environment more or less favorable to it) will permanently modify the equilibrium abundance of another species i. Clearly, if species composition $s(x)$ depends on environmental conditions, then so will net effects, even assuming fixed direct effects A_{ij} across the entire landscape. Fig. 2 demonstrates this context-dependence of net effects.

For non-negligible dispersal $D > 0$, there is no simple linear relationship between abundances and carrying capacities. Therefore, net effects at each site depend not only on local species composition, but also on species abundances in other sites as well as the local abundances. A generalization of V near equilibrium can still be made using the inverse of the Jacobian matrix for the full multi-patch dynamics (Gravel *et al.*, 2016). Here we always retain very small values of dispersal D and this issue does not occur.

1.4 Inferring abiotic niches

A frequent objective when modelling species distributions in space is ascertaining the impact of environmental factors, also understood as the 'fundamental niche' of a species (i.e. what range of abiotic conditions it tolerates), often seen as a prior step to deciphering the impact of species interactions. It is however understood that biotic interactions transform this fundamental niche into a 'realized niche' which can bias our perception of species' environmental preferences: for instance, certain species might only occur in extreme environments because other competing species prevent them from occupying the more temperate environments that they would prefer (Poggiato *et al.*, 2021).

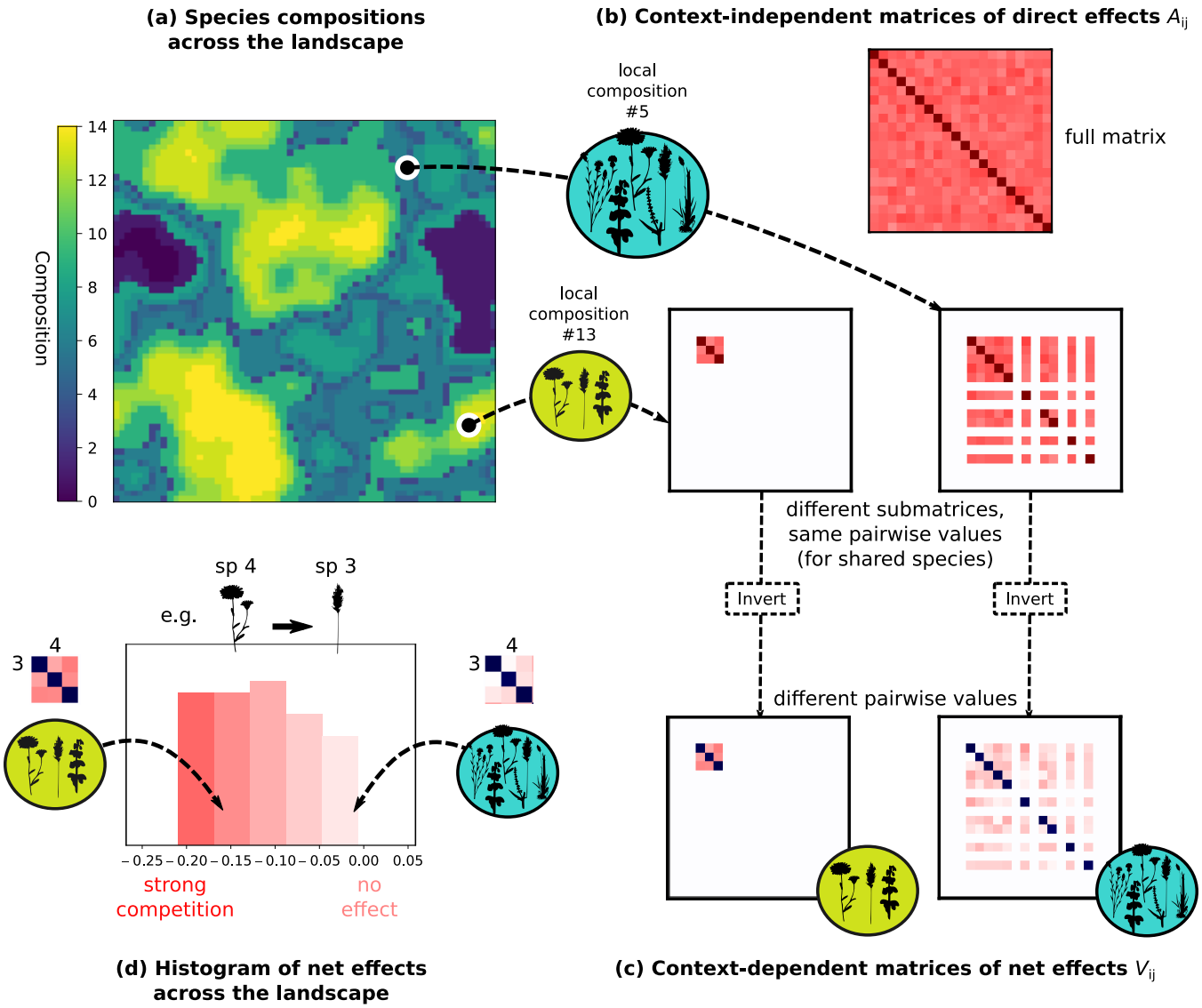

Figure 2: Net effects vary throughout the landscape due to changes in species composition, as illustrated here in one example simulation. While direct effects A_{ij} (instantaneous impact of species j on growth of i) are fixed by assumption in our model (1), net effects V_{ij} (long-term impact of species j on equilibrium abundance of i) are context-dependent and vary due to the presence or absence of other species. **(a)** A number is assigned to each species composition, and mapped through the landscape. Rare compositions are assigned number 0. **(b)** The fixed matrix of direct effects A, and two examples of submatrices restricted to locally surviving species from two different sites. **(c)** Inverting these two submatrices gives local matrices of net effects, $V_{ij}(x)$ at each site x, where individual elements are now different between localities even for pairs that appear in both localities. **(d)** Histogram of local net effects for the most abundant species pair in the metacommunity (species 3 and 4). The net effect $V_{3,4}(x)$ is computed at every site x in the landscape. The mean of these values over the whole landscape gives the spatial average $\overline{V}_{ij}$ referenced in Fig. 3 and 4.

We choose here to explicitly model abiotic niches as carrying capacities K_i that have a Gaussian dependence in the environmental factor E, with species-specific optima and widths (Fig. 1).

Within the dynamical regime and parameter range considered in our simulations, species interactions are indeed distorting the apparent relation between abundance and environment, as illustrated in Fig. S1 and S2, but this distortion remains sufficiently limited that an observer would get a passable estimate of each species' environment preferences, i.e. the center of its fundamental niche and a lower bound on its width, simply by fitting a Gaussian curve to the maximum abundance seen in each environmental condition (details in Appendix).

1.5 Inferring biotic interactions

We infer biotic interaction effects through multilinear regression in two distinct ways. On one hand, we can infer estimates of **net** effects $\widehat{V}_{ij}$ as the multilinear coefficients in

$$N_i(x) = \sum_j \widehat{V}_{ij} K_j(x) \tag{6}$$

On the other hand, following (Xiao *et al.*, 2017; Barbier *et al.*, 2021), we can infer **direct** effects $\widehat{A}_{ij}$ as the multilinear coefficients in

$$N_i(x) = K_i(x) - \sum_{j \neq i} \widehat{A}_{ij} N_j(x) \tag{7}$$

In the first case, we need to know the abiotic niches, i.e. carrying capacities $K_i(x)$ across the landscape. In the second, we can either use known carrying capacities, or infer them as the (site- or environment-dependent) intercept of the relation between the abundance of species i and other species.

Since we noted above that deducing environmental preferences is not the main obstacle in our chosen simulation setting, we hereafter assume that carrying capacities $K_i(x)$ determined by the abiotic factor $E(x)$ are known prior to inferring interactions (e.g. if they are measured in experiments or can independently be estimated from species distributions). We aim to show that even this favorable case presents considerable difficulties, that exist independently from the problem of inferring environmental effects (see Appendix for further discussion).

2 Results

2.1 Direct and net effects

Modern inference methods, e.g. (Ovaskainen *et al.*, 2017), attempt to simultaneously deduce species' interactions, environmental preferences and migration effects from noisy or limited data. Yet significant methodological or conceptual difficulties may arise even with less ambitious goals. Here, we mainly discuss the possibility of estimating species interactions under optimal data conditions: sampling the whole landscape at equilibrium, without any measurement error, and having full knowledge of carrying capacities $K_i(x)$ (i.e. environmental preferences) for each species.

We see in Fig. 3(a,c) that the inference of *direct* effects still depends on the colonization scenario: it is successful with dispersal limitation (DL), but not with environment tracking (ET). In the DL scenario, the inferred matrix $\widehat{A}_{ij}$ is very similar to the ground truth matrix A_{ij}, missing only a few interactions for species that are never present together in the landscape. In the ET scenario, estimates $\widehat{A}_{ij}$ are usually wrong and may even have the wrong sign, though the inference tends to improve here for stronger interactions (as this leads to fewer species coexisting, and thus a simpler inference task).

As for net effects, in our Lotka-Volterra model they are ill-defined at the landscape scale: the value V_{ij} depends on species composition, i.e. the set of surviving species, which varies across the landscape (Fig. 2). Still, we could hope that the *average* value $\langle V_{ij} \rangle$ across the landscape (Fig. 2c) is meaningful. In that case, we expect it should correlate with the apparent net effect, defined as the regression slope of N_i against K_j (putting together values measured across the whole landscape).

We find in Fig. 3(b,d) that this is not the case, which can be explained by the fact that the value of V_{ij} and K_j are actually very correlated across the landscape, so a spatial average of V_{ij} is not representative of local net effects (Fig. 2). We conclude that inferring landscape-scale net effects is an ill-posed problem, as they are not well defined even when direct effects are assumed constant.

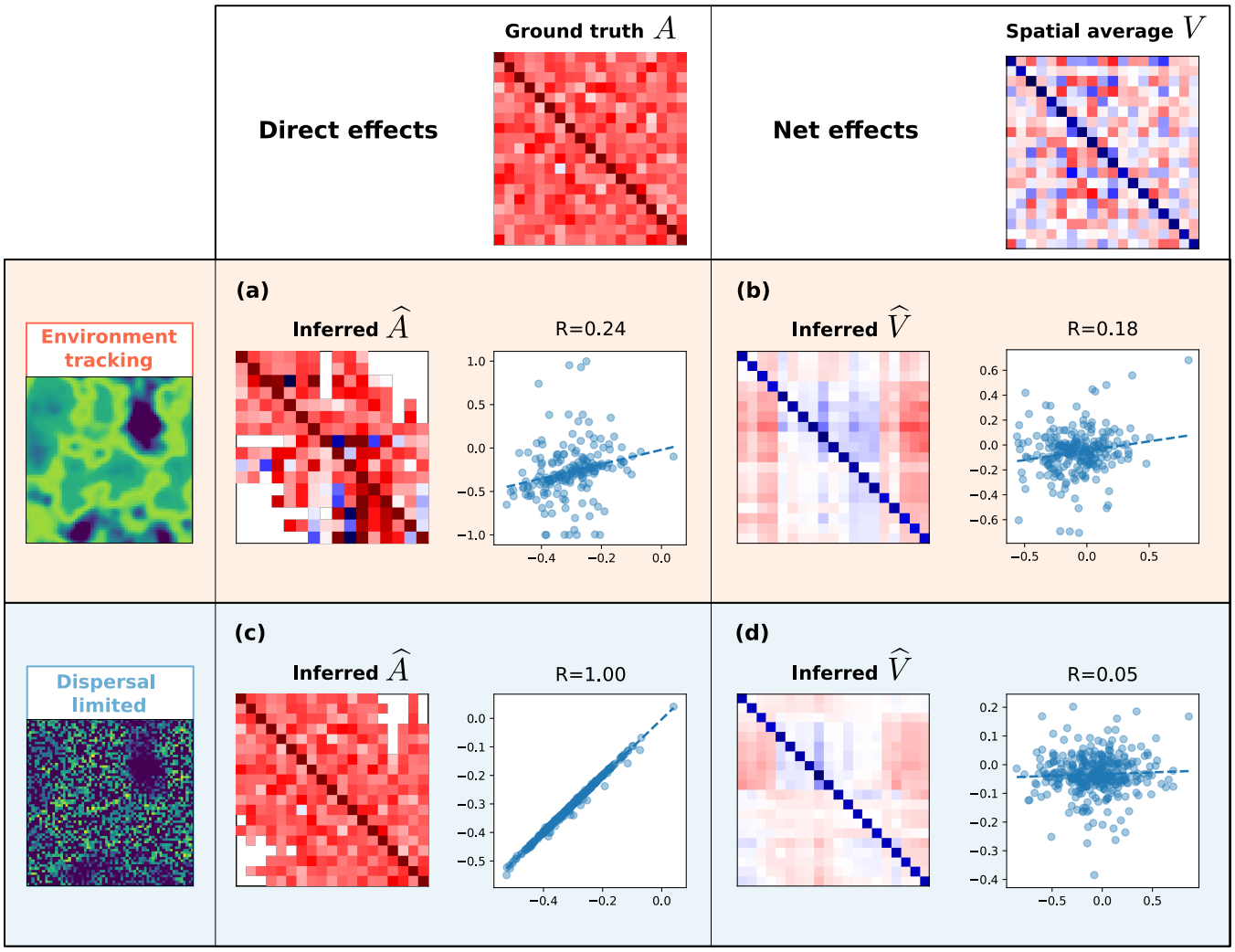

Figure 3: Inferring the full matrix of direct or net effects in the example simulation run under optimal conditions (noiseless data, full knowledge of the abiotic niche of each species, no dispersal, see Methods). Each point corresponds to a pairwise effect of species j on species i. Dashed lines indicate linear regressions. Each row corresponds to a colonization scenario (see Fig. 1 and Sec. 1.2). **(a,c)** Direct effects. On the y-axis, values $\widehat{A}_{ij}$ inferred through hyperplane regression of $N_i(x)$ against $N_j(x)$; on the x-axis, ground truth matrix, A_{ij}. **(b,d)** Net effects V_{ij}. On the y-axis, values $\widehat{V}_{ij}$ inferred through hyperplane regression of $N_i(x)$ against $K_j(x)$. On the x-axis, "ground truth" obtained by spatial average over local matrices, $\overline{V}_{ij}$ (see Methods).

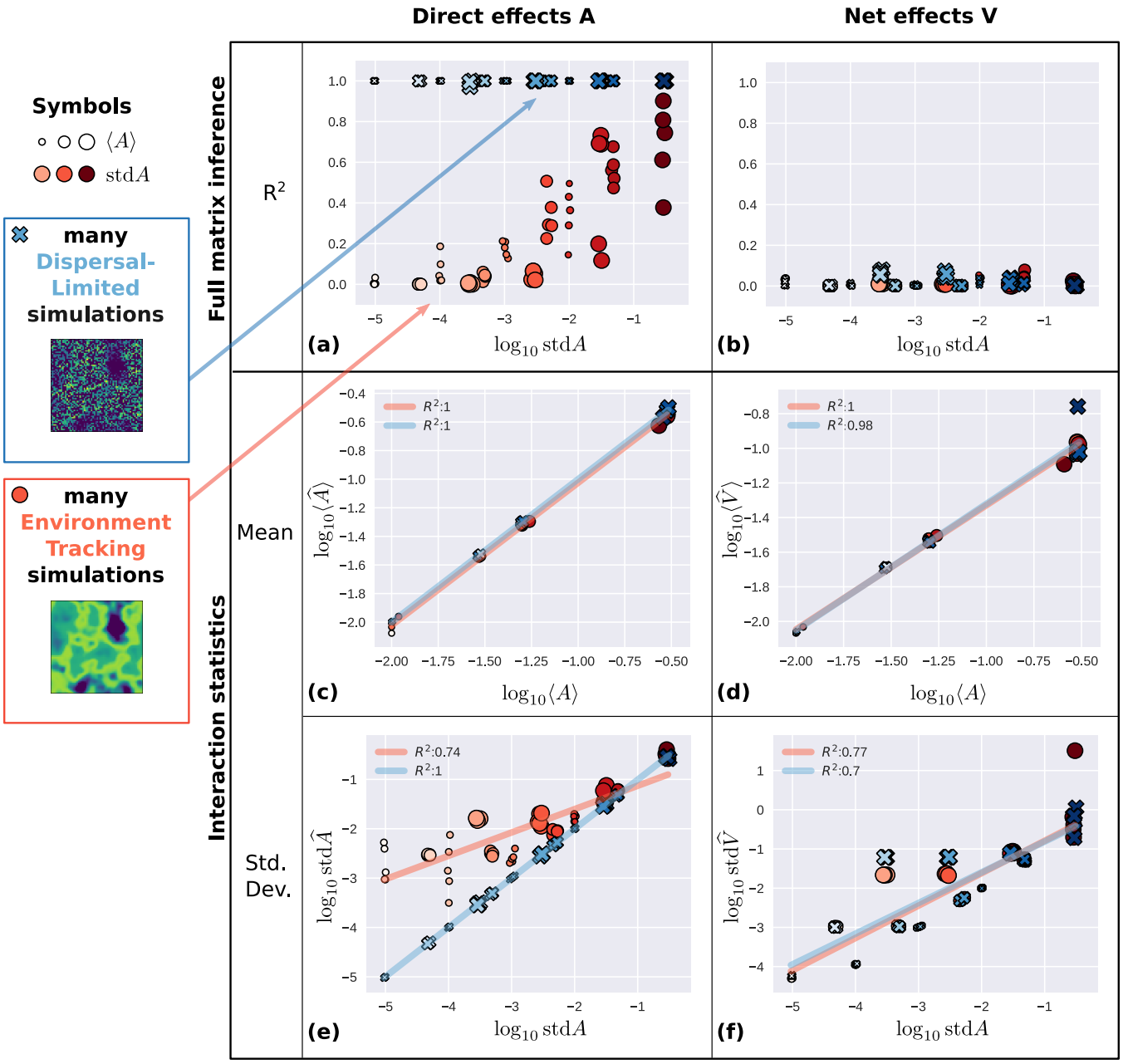

Figure 4: Inferring all interactions or only statistics across many simulation runs with different parameters. Each point is a simulation run, corresponding to one of two colonization scenarios (ET: orange circles, DL: blue crosses) and various values of ground truth statistics $\langle A \rangle \in [0.01, 0.03, 0.05, 0.3]$ (symbol size) and std(A) (symbol color saturation, obtained by multiplying each value of $\langle A \rangle$ by a value in $[0.001, 0.01, 0.1, 1]$). **(a,b)** R^2 of full matrix inference (see e.g. Fig. 3 for one simulation run) is only successful for direct effects under dispersal limitation. However, the statistics of inferred interactions are robustly related to ground truth statistics, for both inferred direct and net effects, $\widehat{A}_{ij}$ and $\widehat{V}_{ij}$. This relationship is very strong for mean interaction **(c,d)**, and weaker for standard deviation **(e,f)**.

2.2 Interaction statistics

While the full matrix inference is unsuccessful in many cases, we see in Fig. 4 that interaction statistics are more robustly estimated. Indeed, there is a very strong correlation between the ground truth mean interaction strength $\langle A \rangle$, and the mean measured over our empirical estimates $\left\langle \widehat{A} \right\rangle$, even when the individual elements $\widehat{A}_{ij}$ are unsuccessfully estimated.

Likewise, there is a strong relationship between $\langle A \rangle$ and average net effects across the landscape, $\langle V \rangle$ (Fig. 4d). These two quantities are not equal even in principle, but a robust relationship suggests that we could use one to infer the other. Standard deviations are also correlated between ground truth and inferred values, though more weakly (Fig. 4e,f). We notice that they are more sensitive to the colonization scenario (DL or ET) and thus we can only roughly deduce the true $\mathrm{std}(A)$ from its empirical estimate (especially at small values, $\mathrm{std}A < 10^{-2}$) without knowing which dispersal scenario we are observing. Finally, symmetry is perfectly estimated in the DL scenario, but entirely incorrect in the ET scenario (see Appendix, Fig. S4).

To summarize, it may be possible to infer the full matrix of direct effects for abundant data, with dispersal limitation or some other phenomenon decoupling species composition from the environment. The full matrix of net effects is not well-defined, and no inference method is successful. Under broader conditions, we can likely only estimate statistical features, most reliably the mean interaction strength.

2.3 Influence of dispersal

We show in Fig. 5 that, as we increase dispersal coefficient D from zero, the transition between dispersal limitation (DL) and environment tracking (ET) scenarios is abrupt in our model, occurring for $D \approx 10^{-3}$.

When dispersal becomes able to overcome the initial absence of a species in a patch, by creating a migrant flux above the extinction threshold N_c and thus allowing species that can invade the patch to reach a nonzero equilibrium, local species composition becomes entirely determined by each site's biotic and abiotic environment (rather than initial conditions and dispersal) and the ability to separate the influence of various species decreases abruptly, as their abundances co-vary much more strongly (which may be seen in Fig. S5 in Appendix).

The abruptness of the transition is due to the fact that our model has a sharp extinction threshold $N_c = 10^{-3}$. Thus, patches where dispersal cannot bring the abundance above the threshold cannot be colonized by the species. Many different species compositions in neighboring patches are preserved until dispersal allows crossing this threshold systematically for all species. However the total number of observed species compositions across the landscape is *not* sharp in D (Fig. S6 in Appendix), suggesting that the ability to correctly infer detailed interactions is not tied to the diversity of compositions over the whole landscape, but rather simply to the existence of diverse compositions in close proximity.

3 Discussion

Many statistical and theoretical methods have been proposed to infer ecological interactions between species from their spatial co-distributions. The best-studied obstacle to this inference is the possible confounding effect of other factors that impact spatial distributions, e.g. the fact that species may appear positively associated because they have similar environmental preferences (Ovaskainen *et al.*, 2017). Recent studies have discussed issues with methods devised to overcome this obstacle, for instance, that interactions may prevent us from correctly understanding species' environmental preferences (Poggiato *et al.*, 2021) Here, we have mainly focused on two further obstacles to the precise inference of species interactions, arising even when we are in the most favorable conditions to address the problems noted previously.

One obstacle is statistical in nature. We find that we can successfully infer direct species interactions only in scenarios where species composition is forced to vary substantially and independently from the environment, for instance due to dispersal limitation or experimental manipulation (as in biodiversity experiments). We cannot do so if the same environmental conditions predictably lead to the same species composition, a situation that we call "environment tracking". The issue is not only that effects of interacting with particular species may be confounded by environmental effects, but they may also be confounded by each other, as the abundances of multiple interaction partners tend to covary positively or negatively based on their similar or dissimilar environmental preferences.

A more conceptual problem is the context-dependence of interactions: there might not exist any constant number that would adequately represent "the effect of species j on species i" across a whole metacommunity, in which case our inference problem is ill-posed from the start. On the one hand, the Lotka-Volterra model used here (1) can be

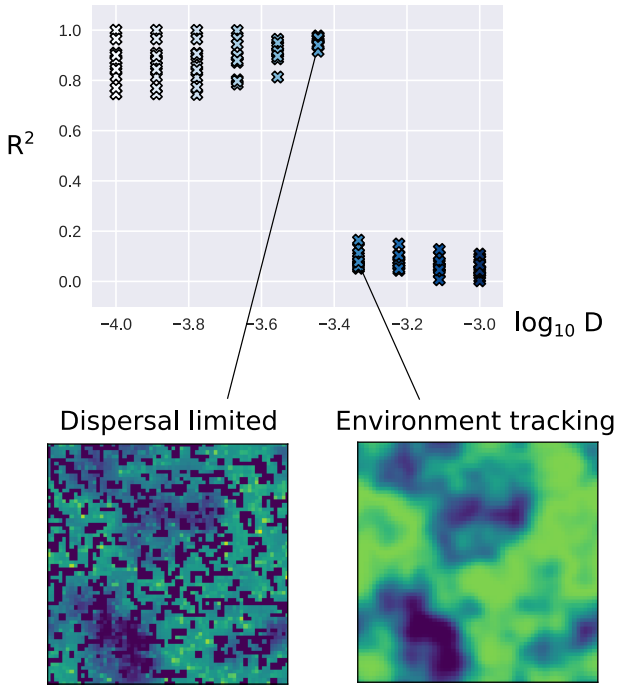

Figure 5: Effect of dispersal on the ability to infer the full matrix of direct species interactions A_{ij}. On the x-axis we represent the dispersal coefficient D in log 10 scale. On the y-axis we show the Pearson R^2 of the fit between true and inferred net effects. Each symbol is a simulation out of sets (all sets share the same landscape, each set has a distinct interaction matrix, and each simulation within a set differs by its value of dispersal). Below, we show the final abundance of one species across the landscape in one simulation set, for values of dispersal right below and above the transition from dispersal limitation to environment tracking. The patchy appearance of the left-hand inset is due to migrant fluxes from neighboring patches being too weak in many cases to overcome the local extinction threshold, as explained in Section 2.3.

thought of as giving a lower bound on the amount of context-dependence we can expect in a plausible ecological setting. The model assumes total context-independence of all direct effects (per-capita instantaneous impacts on growth rates) between species. Yet, the 'net effects' between species, defined to include all indirect impacts arising over time and through intermediates (e.g. indirectly helping a species by directly hindering its competitors), are found to be highly context-dependent as soon as interaction strength is not very small (Zelnik *et al.*, 2022). On the other hand, our choice of looking at long-term abundance patterns, letting species reach an equilibrium, is giving maximal opportunity for such indirect effects to play out – even a rare or slow-growing species has time to exert or mediate significant impacts on others, hence the fact that the matrix of net effects can change drastically when we remove a species, no matter how rare. Thus it may be that observing abundances out of equilibrium, driven by more complex ecological dynamics or external perturbations, and tracking temporal (or spatio-temporal) rather than purely spatial co-distributions, could lessen the interference that might be due to this context dependence and entanglement arising in communities at equilibrium.

Despite these two obstacles, we also found that the community-wide statistical properties, i.e. mean and variance, of direct interaction coefficients could be relatively well inferred from the observed species distribution patterns, even when our detailed estimates of pairwise interactions were entirely incorrect. We did not attempt here to develop novel techniques specifically for the purpose of inferring moments of the distribution of interaction strengths. Instead, we used simple methods to estimate all pairwise interactions, and then computed the moments of these estimates, even when they were individually wrong. It is likely that methods that would be tailored to capture statistical moments directly would be even more robust. But it is rather striking that applying the 'wrong' method still provides reasonable estimates of the moments: it suggests that, even when observational or experimental attempts at measuring interaction strengths (e.g. (Barbier *et al.*, 2021)) yield incorrect numbers, these numbers might still have the right statistics to characterize how important and diverse species interactions are overall in the community-level ecological dynamics.

All our results were obtained in a setting that should be highly favorable to the detection and estimation of species interactions: there are context-independent parameters defining direct interactions, species often coexist, they reach stable abundances that reflect their preferences, dispersal is small and intervenes mainly to allow species to colonize patches where they were not seeded initially (but does not significantly distort equilibrium abundances). Despite all these favorable assumptions, we found that it may not always be possible to precisely infer the details of species interactions. Relaxing some of these assumptions to consider stronger or more context-dependent interactions (e.g. priority effects, environmental modification), more complex dynamics (chaos, transients, external perturbations), stronger spatial fluxes, and observational issues (data limitation, errors and biases), is likely to introduce further difficulties but perhaps also different opportunities for the inference of interactions.

In conclusion, metacommunity ecology provides a more comprehensive conceptual framework than the approaches that set the stage for inferring species interactions from co-distributions (Diamond, 1975; Connor & Simberloff, 1979). Work to date relating metacommunity ecology to species distribution patterns (Morueta-Holme *et al.*, 2016; Ovaskainen *et al.*, 2017; Leibold *et al.*, 2022) is providing some exciting new tools, but the connection between distribution patterns and ecological mechanisms remains elusive. This is, in large part, because correlations between species are highly sensitive, entangling multiple ecological processes and potentially the entire biota, as put forward by Schaffer (1981).

We find that progress might be made by focusing on less detailed and more robust descriptions of distribution patterns. While it may rarely be possible to infer the full set of parameters describing a metacommunity, it might be more feasible to parameterize models inspired by statistical mechanics, e.g. (Gravel *et al.*, 2016; Bunin, 2017), in which only overall statistics of parameters are used to predict a variety of ecological outcomes such as abundance distributions, dynamics and stability. In these so-called "disordered systems" models, outcomes are robust to changing many details of interaction coefficients, as long as species are not organized in a well-ordered structure such as a strict competitive hierarchy. Further refinements of these models have been proposed to capture important large-scale features of the ecological structure of interactions (Barbier *et al.*, 2018). First steps toward parameterizing such models from empirical species co-distributions are being taken in very recent studies (Camacho-Mateu *et al.*, 2023) and we expect that continued progress along these lines will prove timely and useful.

Acknowledgments

We would like to thank Jean-François Arnoldi, Virginie Ravigné, Kara Taylor, Veronica Franz, Jeff Mintz, and Chase Rakowski for discussions during the elaboration of this study and providing useful comments on earlier drafts. M.B. was supported by the CIRAD funding CRESI 2022. G. B. was supported by the Israel Science Foundation (ISF) Grant No. 773/18. M.A.L. was supported by NSF 2025118 and NSF 2224331 grants.

References

Barbier, M., Arnoldi, J.F., Bunin, G. & Loreau, M. (2018). Generic assembly patterns in complex ecological communities. *Proceedings of the National Academy of Sciences*, 115, 2156–2161.

Barbier, M., De Mazancourt, C., Loreau, M. & Bunin, G. (2021). Fingerprints of high-dimensional coexistence in complex ecosystems. *Physical Review X*, 11, 011009.

Barner, A.K., Coblentz, K.E., Hacker, S.D. & Menge, B.A. (2018). Fundamental contradictions among observational and experimental estimates of non-trophic species interactions. *Ecology*, 99, 557–566.

Berry, D. & Widder, S. (2014). Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in microbiology*, 5, 219.

Blanchet, F.G., Cazelles, K. & Gravel, D. (2020). Co-occurrence is not evidence of ecological interactions. *Ecology Letters*, 23, 1050–1063.

Bunin, G. (2017). Ecological communities with lotka-volterra dynamics. *Physical Review E*, 95, 042414.

Camacho-Mateu, J., Lampo, A., Sireci, M., Muñoz, M.Á. & Cuesta, J.A. (2023). Species interactions reproduce abundance correlations patterns in microbial communities. *arXiv preprint arXiv:2305.19154*.

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

Diamond, J.M. (1975). Assembly of species communities. In: *Ecology and evolution of communities* (eds. Cody, M.L. & Diamond, J.M.). Harvard University Press, pp. 342–444.

Fort, H. (2018). Quantitative predictions from competition theory with an incomplete knowledge of model parameters tested against experiments across diverse taxa. *Ecological Modelling*, 368, 104–110.

Gotelli, N.J. & Graves, G.R. (1996). *Null models in ecology.*

Gotelli, N.J. & McCabe, D.J. (2002). Species co-occurrence: a meta-analysis of jm diamond's assembly rules model. *Ecology*, 83, 2091–2096.

Gravel, D., Massol, F. & Leibold, M.A. (2016). Stability and complexity in model meta-ecosystems. *Nature communications*, 7, 1–8.

Lawlor, L.R. (1979). Direct and indirect effects of n-species competition. *Oecologia*, 43, 355–364.

Leibold, M.A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J.M., Hoopes, M.F., Holt, R.D., Shurin, J.B., Law, R., Tilman, D. *et al.* (2004). The metacommunity concept: a framework for multi-scale community ecology. *Ecology letters*, 7, 601–613.

Leibold, M.A. & Mikkelson, G.M. (2002). Coherence, species turnover, and boundary clumping: elements of metacommunity structure. *Oikos*, 97, 237–250.

Leibold, M.A., Rudolph, F.J., Blanchet, F.G., De Meester, L., Gravel, D., Hartig, F., Peres-Neto, P., Shoemaker, L. & Chase, J.M. (2022). The internal structure of metacommunities. *Oikos*, 2022.

Levine, S.H. (1976). Competitive interactions in ecosystems. *The American Naturalist*, 110, 903–910.

Morueta-Holme, N., Blonder, B., Sandel, B., McGill, B.J., Peet, R.K., Ott, J.E., Violle, C., Enquist, B.J., Jørgensen, P.M. & Svenning, J.C. (2016). A network approach for inferring species associations from co-occurrence data. *Ecography*, 39, 1139–1150.

Ovaskainen, O. & Abrego, N. (2020). *Joint species distribution modelling: with applications in R.* Cambridge University Press.

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

Patterson, B.D. & Atmar, W. (1986). Nested subsets and the structure of insular mammalian faunas and archipelagos. *Biological journal of the Linnean society*, 28, 65–82.

Peres-Neto, P.R., Legendre, P., Dray, S. & Borcard, D. (2006). Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology*, 87, 2614–2625.

Poggiato, G., Münkemüller, T., Bystrova, D., Arbel, J., Clark, J.S. & Thuiller, W. (2021). On the interpretations of joint modeling in community ecology. *Trends in ecology & evolution*, 36, 391–401.

Schaffer, W.M. (1981). Ecological abstraction: the consequences of reduced dimensionality in ecological models. *Ecological monographs*, 51, 383–401.

Thurman, L.L., Barner, A.K., Garcia, T.S. & Chestnut, T. (2019). Testing the link between species interactions and species co-occurrence in a trophic network. *Ecography*, 42, 1658–1670.

Xiao, Y., Angulo, M.T., Friedman, J., Waldor, M.K., Weiss, S.T. & Liu, Y.Y. (2017). Mapping the ecological networks of microbial communities. *Nature communications*, 8, 1–12.

Zelnik, Y., Galiana, N., Barbier, M., Loreau, M., Galbraith, E. & Arnoldi, J.F. (2022). How collectively integrated are ecological communities? *bioRxiv*, pp. 2022–12.

Zelnik, Y.R., Arnoldi, J.F. & Loreau, M. (2019). The three regimes of spatial recovery. *Ecology*, 100, e02586.

Zelnik, Y.R., Barbier, M., Shanafelt, D.W., Loreau, M. & Germain, R.M. (2021). Linking intrinsic scales of ecological processes to characteristic scales of biodiversity and functioning patterns. *bioRxiv*, pp. 2021–10.