

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

Matthieu Barbier^{*,1,2,3}, Guy Bunin⁴, and Mathew A. Leibold⁵

February 8, 2024

* 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-conceptual 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 abiotic and biotic environment, and these two processes may combine in highly non-additive ways. Here we show that ~~these issues arise almost inescapably~~ the latter issue arises almost inescapably, even in a simple theoretical setting designed to minimize ~~them~~. Using a model of competitive 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 & Mikkelsen, 2002](#); [Peres-Neto et al., 2006](#)). These issues have been resurfacing especially in microbial ecology

44 due to a tide of data (Berry & Widder, 2014). Additionally, there has developed a broader focus on other types
45 of species interactions (e.g. Cazelles et al. 2015). The analysis of species interactions in general, and accounting
46 for environmental, space and dispersal have been merged by the more recent development of more sophisticated
47 methods such as joint species distribution models e.g. (Ovaskainen et al., 2017) and related methods that aim to
48 partition distribution patterns in relation to ‘abiotic’ (environmental), ‘biotic’ (involving interactions among species)
49 and ‘movement’ (involving spatial effects mediated by dispersal effects). While the sophistication of such methods is
50 rapidly developing, especially in addressing computational and statistical issues, this work is increasingly revealing
51 that inferring process from pattern is not trivial (Barner et al., 2018; Thurman et al., 2019; Blanchet et al., 2020;
52 Poggiato et al., 2021).

53 In parallel with these developments in biogeographic ecology, ~~the same issues arise~~ issues have arisen in work
54 that is more narrowly focused on inferring species interactions from relative abundance patterns in homogeneous
55 communities. ~~Here the data and the inferences ignore the environmental and dispersal dynamics~~ a homogeneous
56 setting (Levine, 1976; Lawlor, 1979). This work ignores confounding effects of environment and dispersal, usually
57 by applying the studies to focusing on situations where these are constrained (experimentally) or assumed to be
58 negligible (by the selection of data). ~~The roots of these approaches can be found in early work focusing on indirect~~
59 ~~effects involving species interactions in multispecies communities~~ (Levine, 1976; Lawlor, 1979) who recognized that
60 pairwise species interactions could be separated into direct components (where species affected each other directly
61 by proximate effects on birth or death rates) or indirect effects (where the effect of a species on another was mediated
62 by direct effects on other mediating species by chains of direct effects). This work highlighted that observable effects
63 of species on each other were most often related to ‘net effects’ that involved the entire network of possible indirect
64 and direct effects and showed that net interaction between pairs of It typically attempts to infer pairwise interactions
65 by comparing abundances across different overlapping sets of species involving such indirect effects might have little
66 similarity to the direct interactions that drive them. These authors proposed that such net interactions could be
67 derived in the case of linear models as the coefficients in the inverse of the community matrix that describes direct
68 interactions, suggesting one way forward for analyzing community data to interpreting the resulting distributions
69 (see below).

70 Schaffer (1981), ~~pointed out that this was not likely to work in the case where qualitative differences in the species~~
71 ~~composition (presence of absence of individual species in the community) could strongly, and even qualitatively, modify~~
72 ~~the nature of net interactions among pairs of species~~. This problem could even involve rare (and thus perhaps hard to
73 detect) species effects. ~~More recently the idea that the inverse of the community matrix could help model net effects~~
74 ~~has been tentatively invoked as a way forward for interpreting~~, see e.g. (Barbier et al., 2021). Another problem
75 remains: how species interactions depend on their biotic context, as it is not obvious that two species interact the
76 same way regardless of which other species are present. We wish to argue here that ecological dynamics inescapably
77 give rise to some context dependence. To clarify, while one may intuitively think of biotic interactions as how one
78 species’ abundance impacts another’s, abundances do not directly cause each other. Rather, species abundances
79 directly influence processes (species dynamics), and these processes in turn control changes in abundances in the
80 short or long term. Even if the direct effects of one species on another’s dynamics were context-independent, its net
81 effects on abundances in the long term can be mediated many other species, along chains of indirect impacts playing
82 out over time, and can thus depend on the whole community’s composition (Schaffer, 1981; Zelnik et al., 2024).
83 We must therefore carefully define direct and net effects and specify which we are trying to infer. It has been
84 proposed that co-distribution patterns across space can be used to deduce a fixed matrix of net effects between
85 species (Ovaskainen et al., 2017), but this assumes that ~~the variation is driven by environmental variables that~~
86 ~~affect the carrying capacity of individual species (and not the interaction coefficients) and that the composition of~~
87 ~~the community (presence-absence) species composition~~ does not change over the metacommunity. This method is
88 consequently not likely to be sufficiently robust to apply except under very highly controlled or limited conditions.

89 Nevertheless, the analysis of distribution and co-distribution patterns of species in metacommunities ~~show that while~~
90 ~~these can be complex, some features can often~~ suggests that, while obtaining exhaustive parameter estimates is
91 still challenging (e.g. (Blanchet et al., 2020; Poggiato et al., 2021; Leibold et al., 2022), some broader features can
92 be related to the processes that generate them (Ovaskainen et al., 2017; Ovaskainen & Abrego, 2020; Leibold
93 et al., 2022). For instance, it may be possible to obtain good estimates of the average competition strength in
94 a community (Fort, 2018). ~~On the other hand, deriving more exhaustive parameter estimates is still challenging~~
95 ~~(A body of theoretical work on so-called “disordered systems” (Barbier et al., 2018) proposes that, when biotic~~
96 ~~interactions are numerous enough and sufficiently independent from each other (unstructured, contrary to e.g.~~
97 ~~(Blanchet et al., 2020; Poggiato et al., 2021; Leibold et al., 2022) a competitive hierarchy), they can be treated as~~
98 random-like, and only their mean and variance matter in determining outcomes of community assembly such as

99 diversity or stability. Inspired by ~~theory on community assembly in highly diverse biotas under the simplifying~~
100 ~~assumptions of ‘highly disordered’ interactions (Barbier *et al.*, 2018)~~ this theory, we hypothesized that statistical fea-
101 tures (e.g. mean and variance of either direct or net effects) may be a more robust inference target than the precise
102 network of interactions. We further wondered if these statistical properties might also be robust to some of the
103 other concerns outlined above such as the existence of composition change, the amount of dispersal and degree of
104 environmental variation.

105 Here, we propose to explore limits to inference that stem from the very nature of the dynamical processes,
106 which entangle contributions from the environment and from multiple species in ways that might or might not
107 be possible to disentangle at all. We use process-based simulations of community assembly in disordered com-
108 munities ~~(Barbier *et al.*, 2018)~~ (as defined above) to generate simulated data under different conditions involving
109 species niches, environmental gradients, dispersal and interaction strength. We consider predominantly competitive
110 interactions, although we believe that many of our results would extend to facilitation, because our simple Lotka-Volterra
111 model can be unstable under strong facilitation. We then infer parameters from the analysis of resulting patterns to
112 evaluate how well they can be used under these various scenarios.

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

120 1 Methods

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

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

139 Within these two scenarios, we analyze a number of simulations that vary in the mean and variance of species’
140 interaction strengths, ~~while remaining in~~. We know from prior literature, e.g. Roy *et al.* (2020); Hu *et al.* (2021),
141 that our simulation model may exhibit complex dynamics, such as chaotic fluctuations, when species interactions
142 are sufficiently strong. To situate ourselves again in the most favorable setting for inference, we choose moderate
143 interactions and dispersal, to remain within a dynamical regime ~~of that corresponds to the~~ ‘species sorting’ ~~(Leibold *et al.*, 2004)~~
144 ~~where we expect species abundances to~~ paradigm of metacommunity theory (Leibold *et al.*, 2004): species abundances
145 reach a stable equilibrium that reflects how favored they are in their local abiotic and biotic environment. In each
146 simulation, we then attempt to infer interactions in detail, as well as derive their mean and variances.

147 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

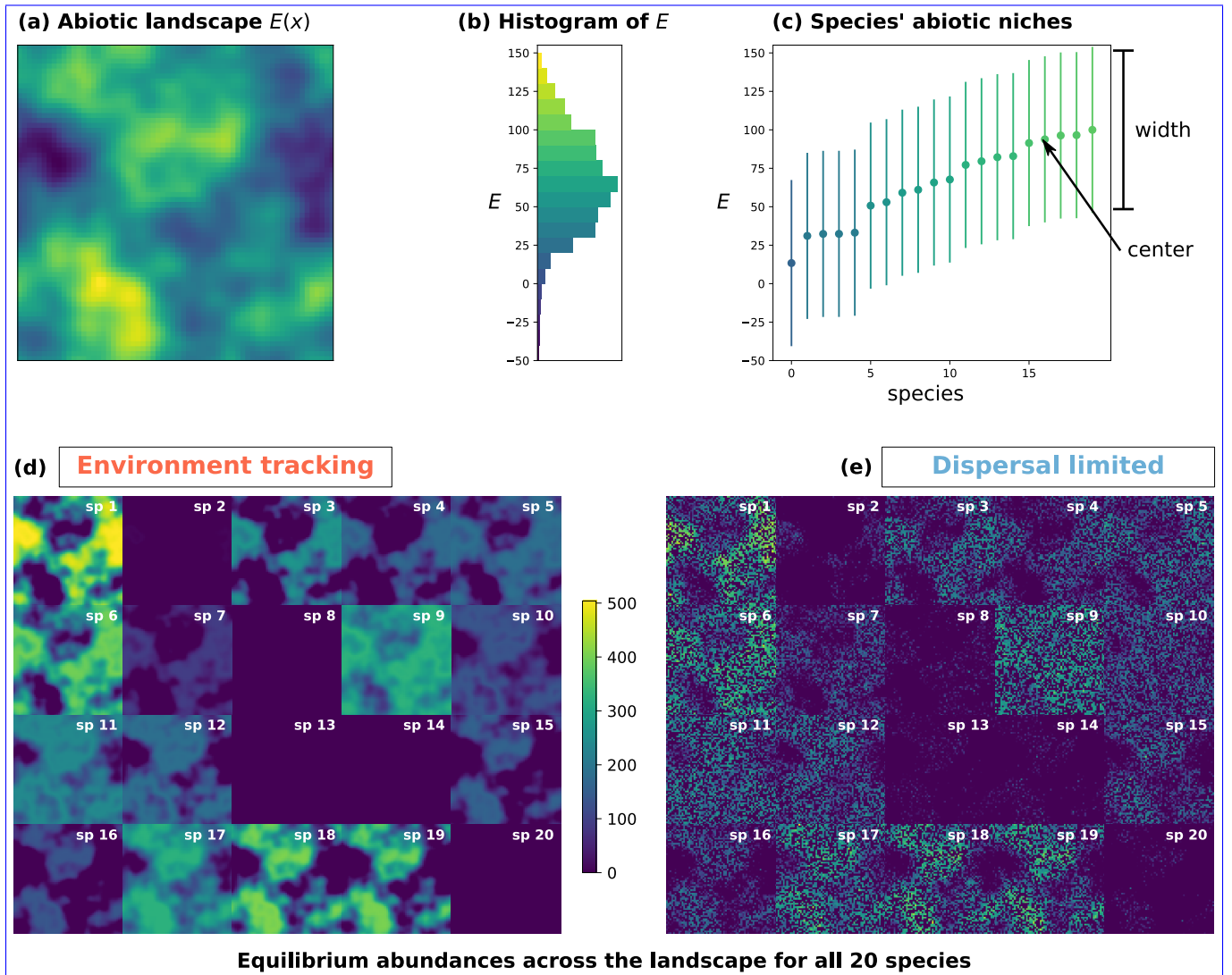


Figure 1: Environmental factor and final species abundances across the landscape in an example simulation run. (a) Landscape $E(x)$ with x a two-dimensional coordinate vector (64×64 pixels called patches). (b) Histogram $P(E)$ of values of the environmental factor. (c) Centers c_i (dots) and width w_i (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-moderate dispersal), and (e) dispersal limitation, where species are seeded independently in 50% of patches at random, and cannot disperse. We note In the latter case, species that would outcompete others in a given environmental condition might be absent by chance in some patches with that condition. Thus, several species go fully extinct at the landscape level from due to competition in the environment tracking scenario (d) whereas, while all species are present in at least part of the landscape in (e). Parameters: $\langle w_i \rangle = 50$, $\langle A_{ij} \rangle = -0.3$, $\text{std}(A_{ij}) = 0.09$, $D = 0$.

Lotka-Volterra dynamics with dispersal from neighboring patches y

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

148 Furthermore, species are considered extinct when $N_i < N_c = 10^{-3}$: ~~their abundances~~, but are allowed to invade
 149 again: their growth rate dN_i/dt is replaced by $\max(0, dN_i/dt)$, which may allow them to regrow above the extinction
 150 threshold if this rate remains positive. At the end of the simulation, the abundances of species under the threshold
 151 are set to zero, ~~and only allowed to vary if their net growth rate $d \log N_i/dt$ becomes positive (corresponding to an~~
 152 ~~invasion) to be ignored in our analyses.~~

153 Intra-specific competition is set to $A_{ii} = -1$ for all species, while inter-specific interaction coefficients A_{ij} are
 154 independent of the environment and drawn randomly for each species pair (i, j) from a normal distribution with
 155 prescribed mean $\langle A \rangle$, standard deviation $\text{std}A$, and symmetry $\text{sym}A = \text{corr}(A_{ij}, A_{ji})$. We typically take a large
 156 negative mean, so that coefficients are predominantly competitive, to avoid the breakdown of the Lotka-Volterra
 157 model with strong facilitation.

158 The coefficients K_i determine each species' carrying capacities (equilibrium abundance in monoculture) since
 159 $A_{ii} = -1$, and thus $N_i = K_i$ at equilibrium in the absence of other species and of dispersal. These carrying
 160 capacities are modelled using a unimodal "niche" function of the environmental factor:

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

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

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

172 1.2 Colonization scenarios and dispersal

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

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

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

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

183 1.3 Defining direct and net effects

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

190 However, these direct effects are not necessarily what we try to infer from empirical metacommunity data, where
 191 we rarely have access to per capita growth rates. Instead, we typically care about how the presence of a species
 192 influences another’s abundance. Since this influence could change over time, we must specify this question further,
 193 e.g. ask how a species affects another’s abundance in the long term, at dynamical equilibrium. The roots of this
 194 question can be found in early work (Levine, 1976; Lawlor, 1979) who recognized that pairwise species interactions
 195 could be separated into direct effects (where species affected each other directly by proximate effects on birth or death
 196 rates) and indirect effects (mediated by chains of direct effects from one species to another). This work highlighted
 197 that observable effects of species on each other were most often related to ‘net effects’, that involved the entire
 198 network of possible indirect and direct effects, and may have little similarity to the direct interactions that drive
 199 them.

200 For mathematical reasons presented briefly here and explained in detail in (Zelnik *et al.*, 2024), Levine and later
 201 authors proposed that net interactions could be derived as the coefficients in the inverse of the matrix of direct
 202 interactions. In the absence of dispersal ($D = 0$), the equilibrium condition for the subset $s(x)$ of species that coexist
 203 at location x is given by equation (1) with the content of the parentheses set to 0 (since $dN_i/dt = 0$ but $N_i > 0$)

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

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

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

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

$$V(x) = - (A^*(x))^{-1} \quad (5)$$

206 the matrix of *net effects* at site x , which represents the long-term consequences of interactions: how permanently
 207 changing the carrying capacity K_j of species j (making the local ~~abiotic or biotic~~ environment more or less favorable
 208 to it) ~~will permanently~~, e.g. via an experimental treatment will modify the equilibrium abundance of another
 209 species i . More broadly, we can define V_{ij} as how any permanent change in the dynamics (per-capita growth rate)
 210 $d \log N_j/dt$ modifies the equilibrium N_i^* .

211 Clearly net and direct effects are not immediately comparable, since A represents the instantaneous impact of
 212 one species’ abundance on another’s dynamics, whereas its inverse V represents the long-term impact of one species’
 213 dynamics on another’s abundance. Yet it can be shown that, properly defined, net effects can be understood
 214 as the sum of all chains of direct effects that connect i and j via any number of intermediate species in the
 215 community Zelnik *et al.* (2024).

216 1.4 Context-dependence of net effects

217 Schaffer (1981) pointed out that the inverse of a matrix depends sensitively on all its elements, and thus, even for
 218 context-independent direct interactions A_{ij} , the specific composition of surviving species can strongly modify the
 219 value or nature of net interactions V_{ij} . For instance, adding a third species may cause two competitors to facilitate
 220 each other indirectly through their competition with a common enemy. While it may seem counter-intuitive that,
 221 say, net effects between two dominant species might change drastically even from adding a rare third species, it is
 222 important to point out that these effects are fully realized only in the long term: the inverse matrix appears naturally
 223 when computing *equilibrium* abundances (4) so that even a rare or slow-growing species has time to exert or mediate
 224 significant impacts on others. Observing abundances out-of-equilibrium could lessen the importance of very indirect
 225 paths, and limit context dependence (Zelnik *et al.*, 2024).

226 Clearly, if species composition $s(x)$ depends on environmental conditions, then so will net effects, even assuming
 227 fixed direct effects A_{ij} across the entire landscape. Fig. 2 demonstrates this context-dependence of net effects.

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

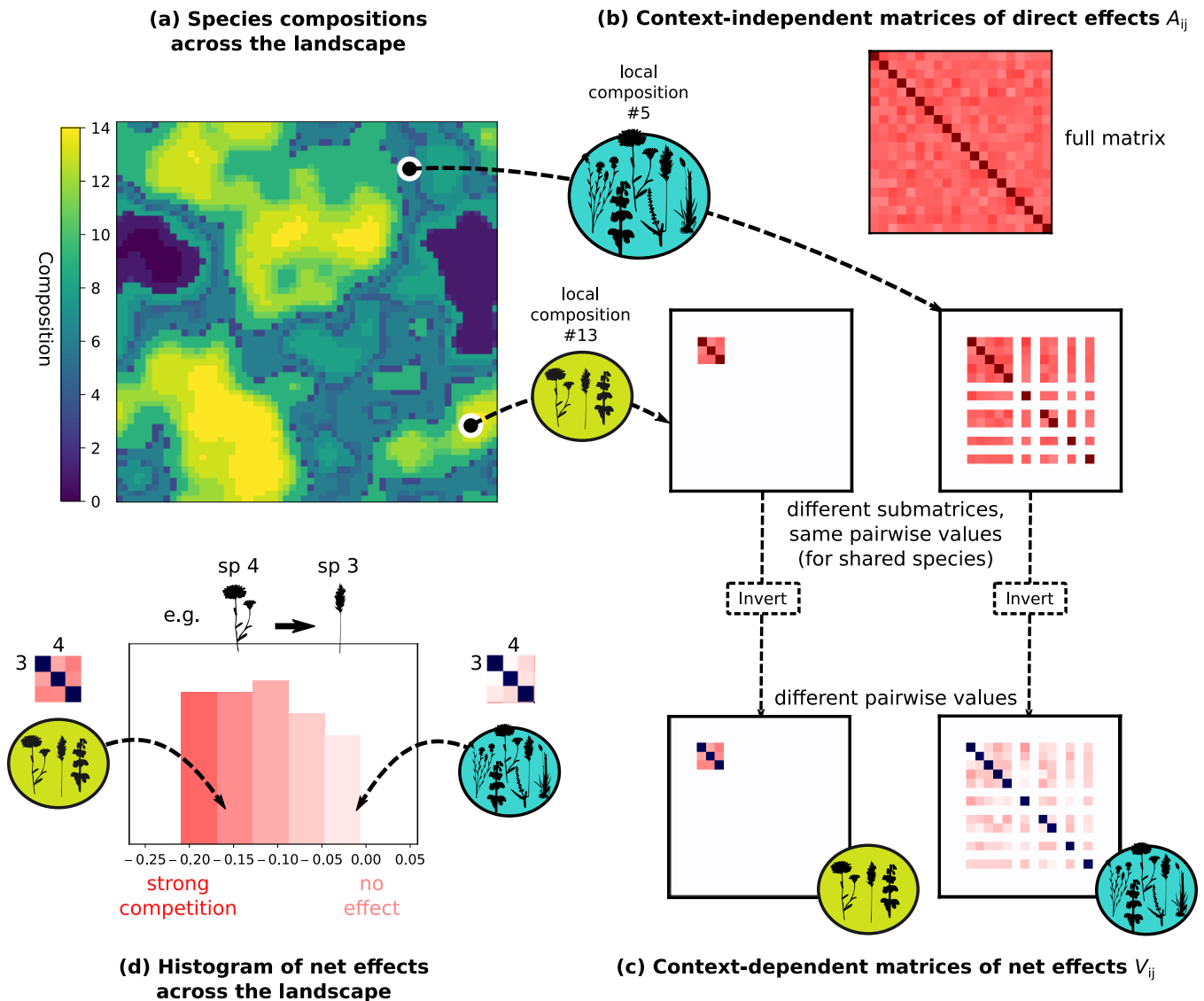


Figure 2: Net effects vary throughout the landscape due to changes in species composition, as illustrated here in one example simulation [with dispersal \(environment tracking in Fig. 1\)](#). While direct effects A_{ij} (instantaneous impact of species j on growth of i) are fixed by assumption in our 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 \bar{V}_{ij} referenced in Fig. 3 and 4.

1.5 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).

~~We choose here to explicitly model~~ Since our focus is rather on the challenges of inferring biotic interactions even under favorable conditions, we do not delve deeply into the problem of simultaneously determining abiotic niches and interactions, which is central in joint Species Distribution Models e.g. Ovaskainen *et al.* (2017).

Nevertheless, using our simulated data, we could simply try to fit the parameters of equation (2), i.e. the true functional form used to generate the data, ignoring species interactions. This amounts to modelling 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~~ As we show in Appendix, 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).

Yet this distortion remains sufficiently limited in our case 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).

Therefore, in the rest of the main text, we entirely bypass the issue of inferring abiotic niche, and investigate inference challenges that remain even assuming that we perfectly know the carrying capacities of every species at every location.

1.6 Inferring biotic interactions

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

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

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

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

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~~ most severe 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.

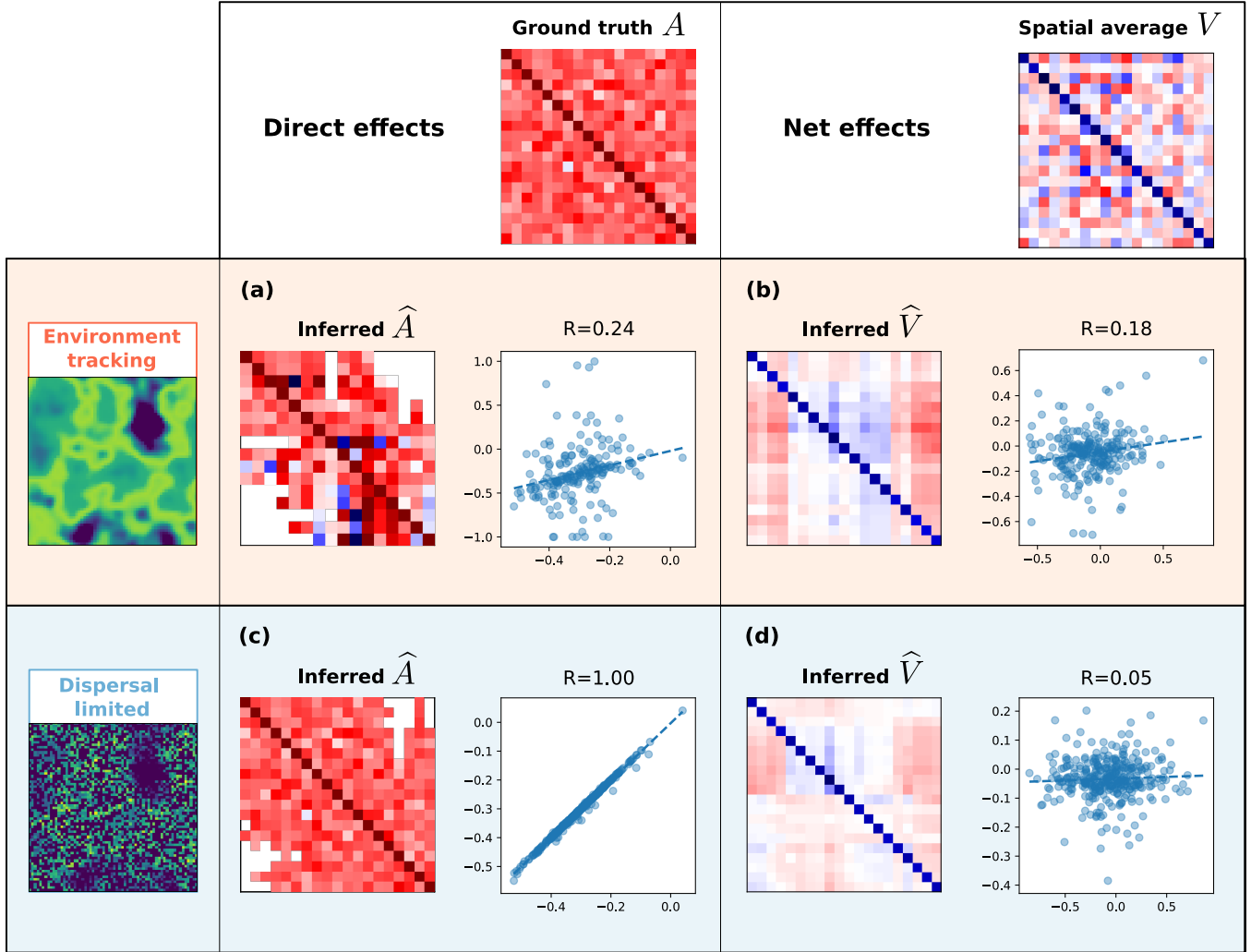


Figure 3: Inferring the full matrix of direct or net effects in the example simulation run under optimal conditions (noiseless data, full knowledge of the abiotic niche of each species, no dispersal, see Methods). Each point corresponds to a pairwise effect of species j on species i . Dashed lines indicate 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 \hat{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 \hat{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, \bar{V}_{ij} (see Methods).

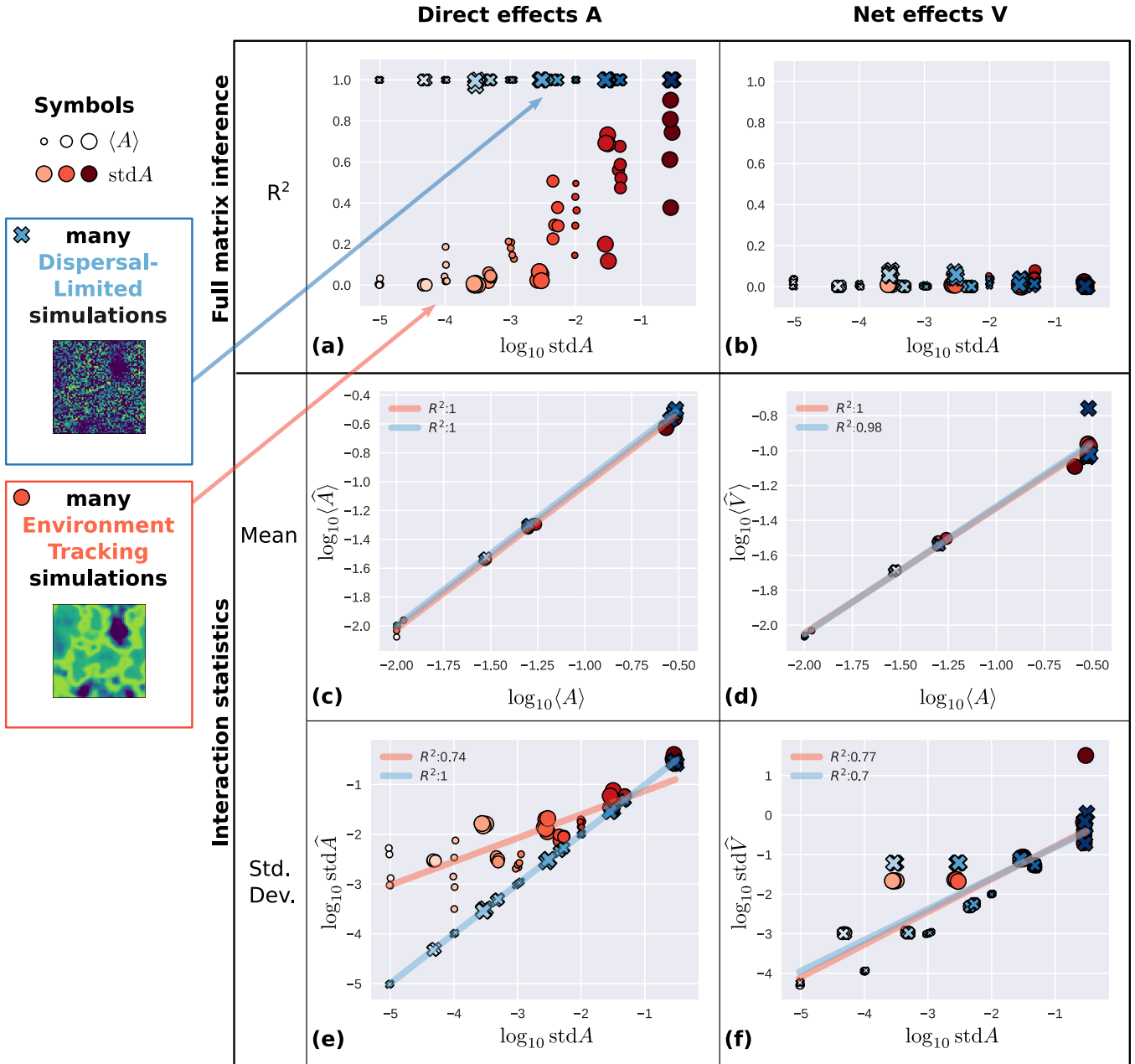


Figure 4: Inferring all interactions or only statistics across many simulation runs with different parameters. Each point is a simulation run, corresponding to one of two colonization scenarios (ET: orange circles, DL: blue crosses) and various values of ground truth statistics $\langle A \rangle \in [0.01, 0.03, 0.05, 0.3]$ (symbol size) and $\text{std}(A)$ (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, \hat{A}_{ij} and \hat{V}_{ij} . This relationship is very strong for mean interaction **(c,d)**, and weaker for standard deviation **(e,f)**.

We break down this issue into the inference of *direct* effects and *net* effects whose definitions we recall here (see Methods for more details). Direct effects correspond to the matrix A_{ij} in (1) which describes how the current abundance of species j influences the instantaneous dynamics (growth or mortality rate) of species i . Net effects are given by the matrix V_{ij} in (5) which describes how a permanent change in species j 's dynamics (e.g. a change of carrying capacity, or removal from the community) would impact the abundance of species i in the long term Zelnik *et al.* (2024). Direct effects are not usually what we try to infer in a biogeographic context since we rarely have access to the population growth rates, but they mediate and explain net effects on abundances.

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 \hat{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 \hat{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.

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 $\langle \hat{A} \rangle$, even when the individual elements \hat{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 $\text{std}(A)$ from its empirical estimate (especially at small values, $\text{std}A < 10^{-2}$) 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

327 the diversity of compositions over the whole landscape, but rather simply to the existence of diverse compositions in
 328 close proximity.

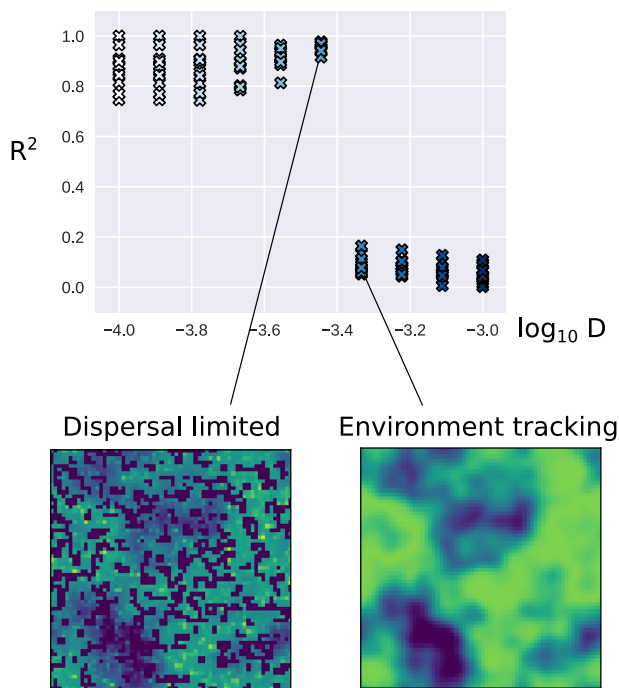


Figure 5: Effect of dispersal on the ability to infer the full matrix of direct species interactions A_{ij} . On the x-axis we represent the dispersal coefficient D in log 10 scale. On the y-axis we show the Pearson R^2 of the fit between true and inferred net effects. Each symbol is a simulation out 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.

329 3 Discussion

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

338 3.1 Statistical issues and non-identifiability

339 One obstacle is statistical in nature, i.e. difficulties in identifying the model due to multi-collinearity. We find
 340 that we can successfully infer direct species interactions only in scenarios where species composition is forced to
 341 vary substantially and independently from the environment, for instance due to dispersal limitation or experimental
 342 manipulation (as in biodiversity experiments). We cannot do so if the same environmental conditions predictably
 343 lead to the same species composition, a situation that we call “environment tracking”.

344 The issue is not only that effects of interacting with particular species may be confounded by environmental effects,
 345 but they may also be confounded by each other, as the abundances of multiple interaction partners tend to covary

346 positively or negatively based on their similar or dissimilar environmental preferences. That latter problem decreases
347 when interactions have more variance (Fig. 4a) or when species are differently influenced by many independent
348 environmental factors (see Appendix, Fig. S7), but this does not suffice here to completely eliminate the problem of
349 model identifiability.

350 All our results were obtained in a setting that should be highly favorable to the detection and estimation of species
351 interactions: there are context-independent parameters defining direct interactions, species often coexist, they reach
352 stable abundances that reflect their preferences, dispersal is small and intervenes mainly to allow species to colonize
353 patches where they were not seeded initially (but does not significantly distort equilibrium abundances). Despite
354 all these favorable assumptions, we found that it may not always be possible to precisely infer the details of species
355 interactions. Relaxing some of these assumptions to consider stronger or more context-dependent interactions (e.g.
356 priority effects, environmental modification), more complex dynamics (chaos, transients, external perturbations),
357 stronger spatial fluxes, and observational issues (data limitation, errors and biases), is likely to introduce further
358 difficulties but perhaps also different opportunities for the inference of interactions. Indeed, we speculate that various
359 obstacles could work against each other: stronger non-linearities may somewhat alleviate model non-identifiability;
360 conversely, having a large number of species and parameters may end up being the main challenge in model fitting
361 and override the importance of details of how each interaction is modelled; finally, other dynamical settings, such as
362 species abundances fluctuating chaotically rather than being at equilibrium, may require entirely distinct methods
363 with different challenges.

364 A more conceptual problem ~~is~~ lies in the context-dependence of interactions: there might not exist any constant
365 number that would adequately represent “the effect of species j on species i ” across a whole metacommunity, in which
366 case our inference problem is ill-posed from the start. On the one hand, the Lotka-Volterra model used here (1) can
367 be thought of as giving a lower bound on the amount of context-dependence we can expect in a plausible ecological
368 setting. The model assumes total context-independence of all direct effects (per-capita instantaneous impacts on
369 growth rates) between species. Yet, the ‘net effects’ between species, defined to include all indirect impacts arising
370 over time and through intermediates (e.g. indirectly helping a species by directly hindering its competitors), are
371 found to be highly context-dependent as soon as interaction strength is not very small (?)(Zelnik *et al.*, 2024). On
372 the other hand, our choice of looking at long-term abundance patterns, letting species reach an equilibrium, is giving
373 maximal opportunity for such indirect effects to play out – even a rare or slow-growing species has time to exert
374 or mediate significant impacts on others, hence the fact that the matrix of net effects can change drastically when
375 we remove a species, no matter how rare. Thus it may be that observing abundances out of equilibrium, driven
376 by more complex ecological dynamics or external perturbations, and tracking temporal (or spatio-temporal) rather
377 than purely spatial co-distributions, could lessen the interference that might be due to this context dependence and
378 entanglement arising in communities at equilibrium.

379 Our work stresses the importance of correctly specifying which concept of biotic interaction one is trying
380 to infer: for instance, estimating context-independent direct effects is sometimes possible even when net effects
381 vary dramatically across the landscape. This is of particular relevance to statistical approaches focusing on the
382 co-distribution of species, e.g. joint Species Distribution Models Ovaskainen *et al.* (2017). The residual covariance
383 between species across the whole landscape can be understood as the spatial aggregation of locally varying net effects
384 which we believe (see Appendix: Estimating interactions from residual species co-variation) is not an appropriate
385 path to deduce direct effects.

386 3.2 Getting more from less

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

398 ~~All our results were obtained in a setting that should be highly favorable to the detection and estimation of species~~
399 ~~interactions: there are~~ From an empirical point of view, our results thus carry encouraging as well as cautionary

notes. Species-level questions may require the precise inference of a given pairwise interaction, and our work comes here as a warning to keep in mind possible barriers to that inference when we do not have grounds to assume ‘natural experiments’ such as permitted here by strong dispersal limitation. On the other hand, we feel that for empiricists interested in estimating the intensity of biotic interactions in a community as a whole, e.g. to know whether community composition is a better indicator of environmental states or internal dynamics, our findings bring some home and a suggestion to turn to methods that strive to estimate community-wide statistics rather than individual pairwise species interactions.

Finally, we must consider the empirical relevance of our study’s assumption that direct species effects 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 ~~and simply additive with each other and with environmental effects.~~ This may seem like a highly restrictive assumption, limiting the value of trying to infer such direct effects. We nevertheless speculate that two facts make this assumption less restrictive than it seems: first, additivity is more likely to hold approximately in direct effects, which occur over a short time, than in long-term net effects; and second, the congruence of many causal factors (species, environmental variables) hopefully means that 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 each and how they interact matter less, and additive effects may be a mechanistically wrong but phenomenologically useful abstraction, at least when focusing on community-wide statistics and outcomes as we are suggesting here.

3.3 Conclusions

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 co-distribution patterns (Morueta-Holme *et al.*, 2016; Ovaskainen *et al.*, 2017; Leibold *et al.*, 2022; Christopher D. Terry *et al.*, 2023) 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.

Code availability

All simulation code is written in Python and available from the following repository: <https://github.com/mrcbarbier/morefromless>

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., Coblenz, 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*.
- Christopher D. Terry, J., Langdon, W. & Rossberg, A.G. (2023). Codistribution as an indicator of whole metacommunity response to environmental change. *Ecography*, p. e06605.
- 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.
- Hu, J., Amor, D.R., Barbier, M., Bunin, G. & Gore, J. (2021). Emergent phases of ecological diversity and dynamics mapped in microcosms. *bioRxiv*.
- 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. & Mikkelsen, G.M. (2002). Coherence, species turnover, and boundary clumping: elements of meta-community 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.

- 489 Ovaskainen, O. & Abrego, N. (2020). *Joint species distribution modelling: with applications in R*. Cambridge
490 University Press.
- 491 Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T. & Abrego, N.
492 (2017). How to make more out of community data? a conceptual framework and its implementation as models
493 and software. *Ecology letters*, 20, 561–576.
- 494 Patterson, B.D. & Atmar, W. (1986). Nested subsets and the structure of insular mammalian faunas and archipelagos.
495 *Biological journal of the Linnean society*, 28, 65–82.
- 496 Peres-Neto, P.R., Legendre, P., Dray, S. & Borcard, D. (2006). Variation partitioning of species data matrices:
497 estimation and comparison of fractions. *Ecology*, 87, 2614–2625.
- 498 Poggiato, G., Münkemüller, T., Bystrova, D., Arbel, J., Clark, J.S. & Thuiller, W. (2021). On the interpretations of
499 joint modeling in community ecology. *Trends in ecology & evolution*, 36, 391–401.
- 500 Roy, F., Barbier, M., Biroli, G. & Bunin, G. (2020). Complex interactions can create persistent fluctuations in
501 high-diversity ecosystems. *PLoS computational biology*, 16, e1007827.
- 502 Schaffer, W.M. (1981). Ecological abstraction: the consequences of reduced dimensionality in ecological models.
503 *Ecological monographs*, 51, 383–401.
- 504 Thurman, L.L., Barner, A.K., Garcia, T.S. & Chestnut, T. (2019). Testing the link between species interactions and
505 species co-occurrence in a trophic network. *Ecography*, 42, 1658–1670.
- 506 Xiao, Y., Angulo, M.T., Friedman, J., Waldor, M.K., Weiss, S.T. & Liu, Y.Y. (2017). Mapping the ecological
507 networks of microbial communities. *Nature communications*, 8, 1–12.
- 508 Zelnik, Y.R., Arnoldi, J.F. & Loreau, M. (2019). The three regimes of spatial recovery. *Ecology*, 100, e02586.
- 509 Zelnik, Y.R., Barbier, M., Shanafelt, D.W., Loreau, M. & Germain, R.M. (2021). Linking intrinsic scales of ecological
510 processes to characteristic scales of biodiversity and functioning patterns. *bioRxiv*, pp. 2021–10.
- 511 Zelnik, Y.R., Galiana, N., Barbier, M., Loreau, M., Galbraith, E. & Arnoldi, J.F. (2024). How collectively integrated
512 are ecological communities? *Ecology Letters*, 27, e14358.

514 A Inferring carrying capacities

515 We compare in Fig. S1 and Fig. S2 the fundamental abiotic niche $K_i(E)$ for each species to the observed distribution
 516 of species abundances, to see whether the former can easily be deduced from the latter or appears distorted due to
 517 biotic interactions.

518 The principle we use here is very simple: since interactions are almost all competitive in most of our simulations
 519 (some of the randomly-drawn values may be positive, representing mutualism, commensalism, etc. but they are
 520 infrequent), we expect the abundance to be at most equal to $K_i(E)$ most of the time. Therefore we devise a very
 521 simple estimate of the fundamental niche by (1) binning observed species abundances into 20 bins based on the value
 522 of the environmental factor, (2) computing a higher bound within each bin (using the 95% percentile to be more
 523 robust to outliers), then (3) ~~performing a Gaussian fit over fitting equation (2) with $\mu_i = 0$, i.e. a Gaussian curve,~~
 524 to these local maxima by linear least squares regression, species by species.

525 (We note that in our simulation model, K_i can be negative, see (2), while the inference approach we use here can
 526 never correctly ascertain how bad an environment is when the species does not grow in it, so we do not attempt to
 527 fit the ‘mortality’ μ_i defined in (2))

528 In Fig. S1, we see that the Dispersal Limitation scenario may help to correctly deduce the fundamental niche
 529 since there are many different species compositions in the same abiotic conditions, and some of these compositions
 530 may be closer to the species being alone, thus we generally recover the overall shape of the niche, though its precise
 531 parameters are slightly biased.

532 In Fig. S2, we observe that in an Environment Tracking scenario, the shape of the realized niche may be quite
 533 different from the fundamental niche, as the (mainly competitive) interactions may restrict species to only a small
 534 part of their potential range. The inference of niche parameters is not as good; still, the environmental gradient is
 535 strong enough here that we can get a rough estimate of abiotic properties through our simple inference method. Given
 536 that the ET scenario is in any case unfavorable to inferring species interactions, we do not discuss the additional
 537 complexity of having wrong estimates of the carrying capacities K_i , and treat them as exactly known when trying
 538 to estimate interactions, but this additional difficulty could be discussed in future work.

539 We show in Fig. S3 how errors made on the inference of each aspect of the niche vary continuously with the
 540 dispersal coefficient, interpolating between the two extreme scenarios shown in the previous two figures.

541 B Inferring symmetry

542 Beyond interaction mean and variance, interaction symmetry has often been proposed theoretically as an important
 543 parameter, see e.g. (Bunin, 2017; Barbier *et al.*, 2018). That was another statistical parameter that we varied and
 544 attempted to infer, see Fig. S4. We find that interaction symmetry is only correctly inferred for direct effects, and
 545 only when the full matrix inference is successful, i.e. in the Dispersal Limited scenario.

546 C Estimating interactions from residual species co-variation

547 Consider patterns of covariation across space:

- 548 • Covariance of carrying capacities “covK”:

$$\text{cov}(K_i, K_j) = \sum_x (K_i(x) - \langle K_i \rangle)(K_j(x) - \langle K_j \rangle) \quad (\text{Eq. S1})$$

- 549 • Covariance of abundances “covN”:

$$\text{cov}(N_i, N_j) = \sum_x (N_i(x) - \langle N_i \rangle)(N_j(x) - \langle N_j \rangle) \quad (\text{Eq. S2})$$

550 To see whether a trace of biotic interactions can be found in residual covariation, we first decompose the covariance
 551 of abundances into two components:

$$\text{cov}N = \text{cov}_{\text{abio}} + \text{cov}_{\text{residual}} \quad (\text{Eq. S3})$$

552 The abiotic component is simply derived from the covariance of carrying capacities

$$\text{cov}_{\text{abio}} = c_i c_j \text{cov}(K_i, K_j) \quad (\text{Eq. S4})$$

553 with unknown scaling coefficients c_i that are inferred by least squares regression.

554 This mimics the way that a simple SDM would predict species covariation (for each species i , a regression between
555 abundance N and the environmental factor). Given (5), we can expect that in the case of weak net effects between
556 species, this formula is approximately correct for our dynamical model with $c_i \approx V_{ii}$.

557 We can then study whether $\text{cov}_{\text{residual}}$ contains information about the full matrix or statistics of direct effects A_{ij}
558 or net effects V_{ij} .

559 All the inferences in the main text were performed assuming that we have full knowledge of abundance $N_i(x)$
560 and carrying capacity $K_i(x)$ for every species i at every point in space x . Here we assume that we have access to
561 less information: only how abundances or carrying capacities covary between species across the landscape, shown in
562 Fig. S5(a,b,c).

563 Removing the covariation between species due to the environment, we can observe how statistics of the coefficients
564 c_i in the regression of abundances to the abiotic niche $N_i \sim c_i K_i(E(x))$, and residual covariation $\text{cov}_{\text{residual}}$ depend
565 on the mean and variance of interactions. We see in Fig. S5 that there is no clear relationship with ground truth
566 parameters. It thus seems impossible, under the studied model setting and approach, to infer even only interaction
567 statistics without having more detailed information on abundances and carrying capacities.

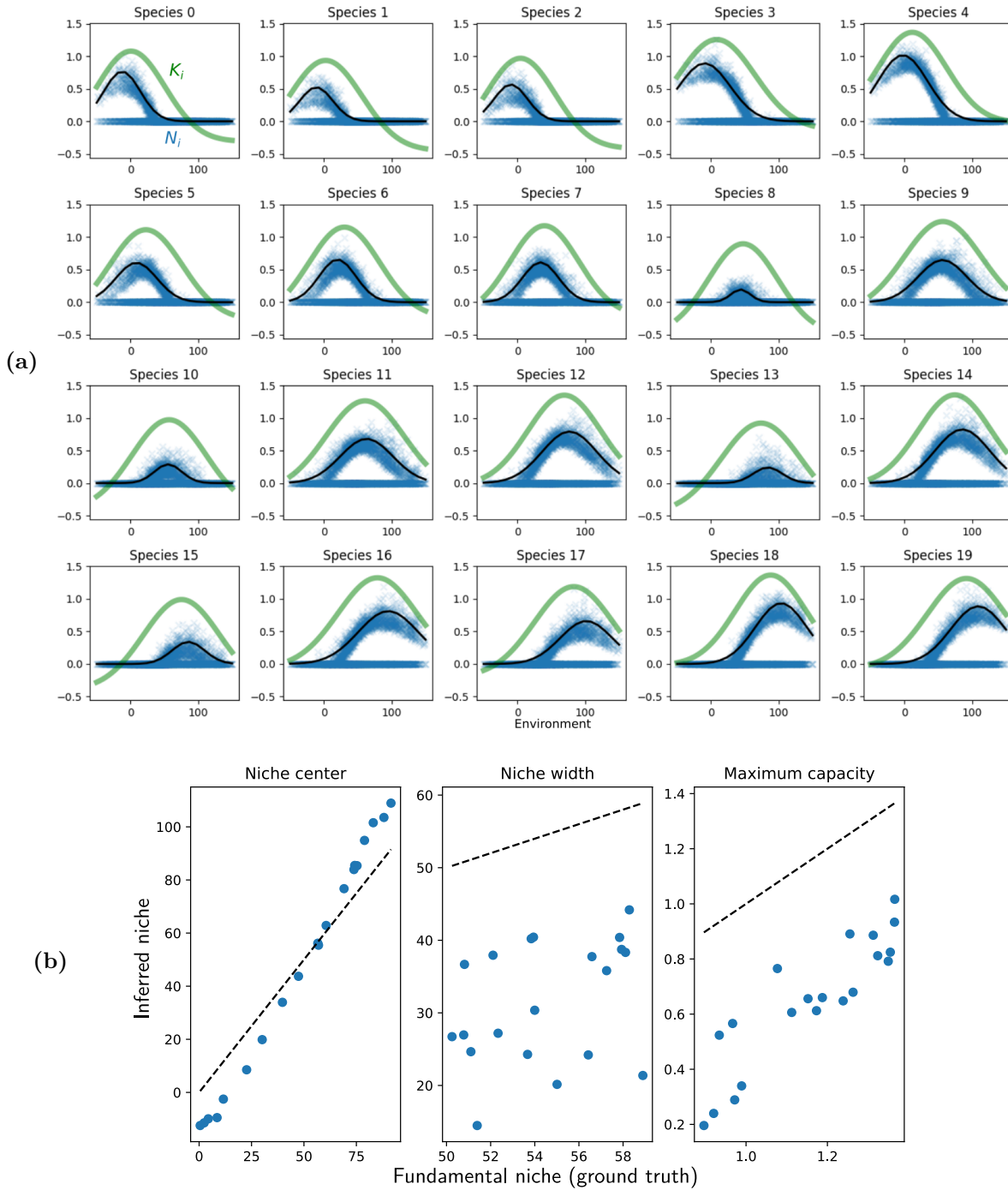


Figure S1: Inferring carrying capacities in one simulation example under conditions of dispersal limitation, see details in Appendix A. **(a)** Fundamental and inferred niches for each species, as a function of values of the environmental factor $E \in [-50, 150]$. Solid green curve: fundamental niche (carrying capacity as a function of environment). Blue symbols: abundances observed throughout the landscape. Solid black curve: niche inferred through our method. **(b)** Comparing the center, width and maximum of the fundamental (x-axis) and inferred (y-axis) niches.

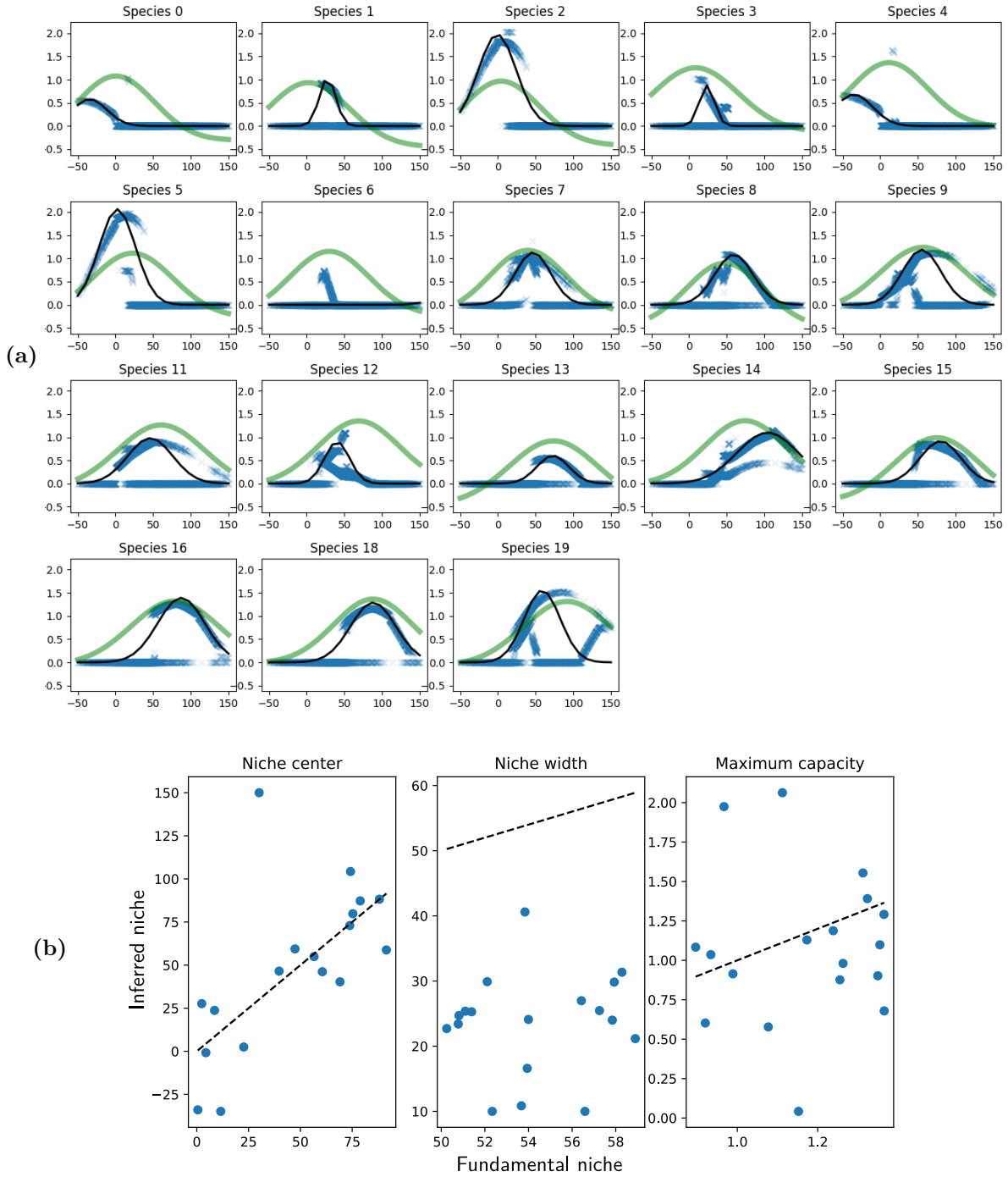


Figure S2: Inferring carrying capacities in one simulation example under conditions of environment tracking, see details in Appendix A and Fig. S1. Some species have gone extinct and are not represented in (a).

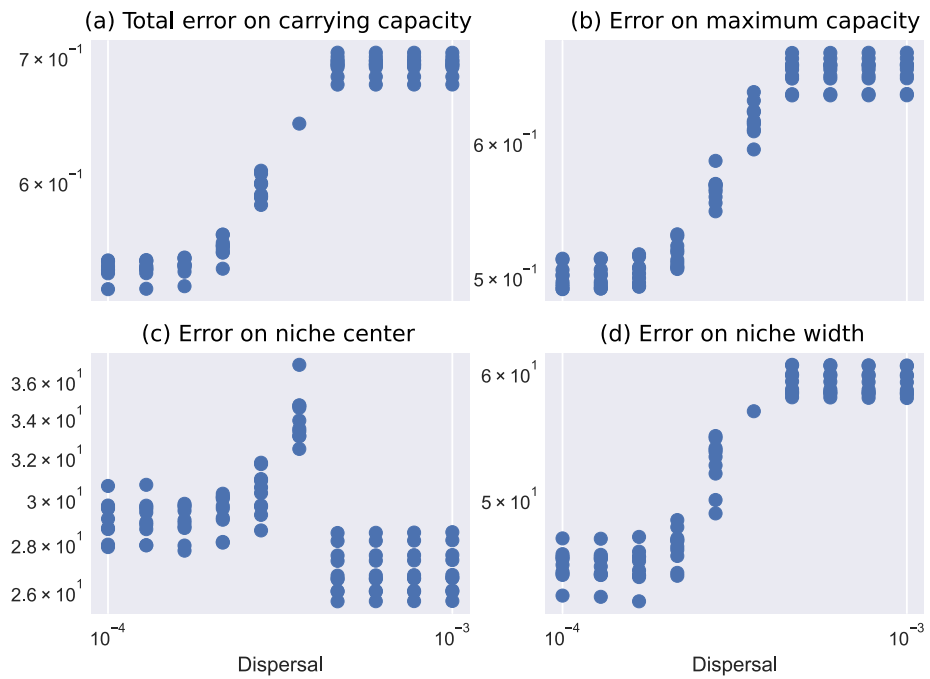


Figure S3: Impact of dispersal coefficient D on error on **(a)** estimates of carrying capacity at each point of the environmental gradient, **(b)** estimates of maximum carrying capacity, **(c)** niche centers, **(d)** niche width. In each case, errors are computed as the square root of the average across species – and average across space in case (a) – of $(\text{groundtruth value} - \text{inferred value})^2$, with the inference approach discussed in Appendix.

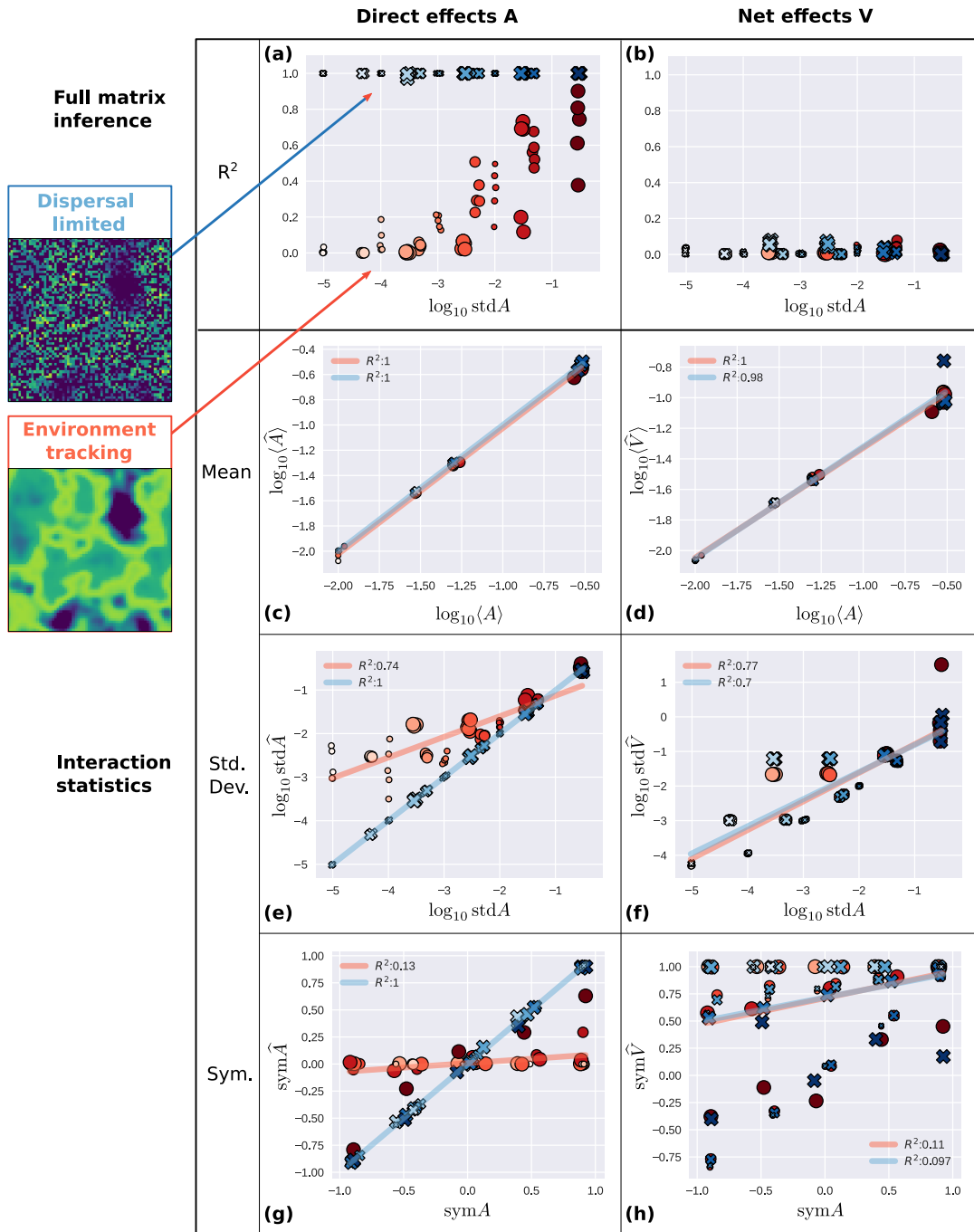


Figure S4: Extension of Fig. 4 including interaction symmetry (g,h) which is only correctly inferred for direct effects, and only when the full matrix inference is successful, i.e. in the DL scenario.

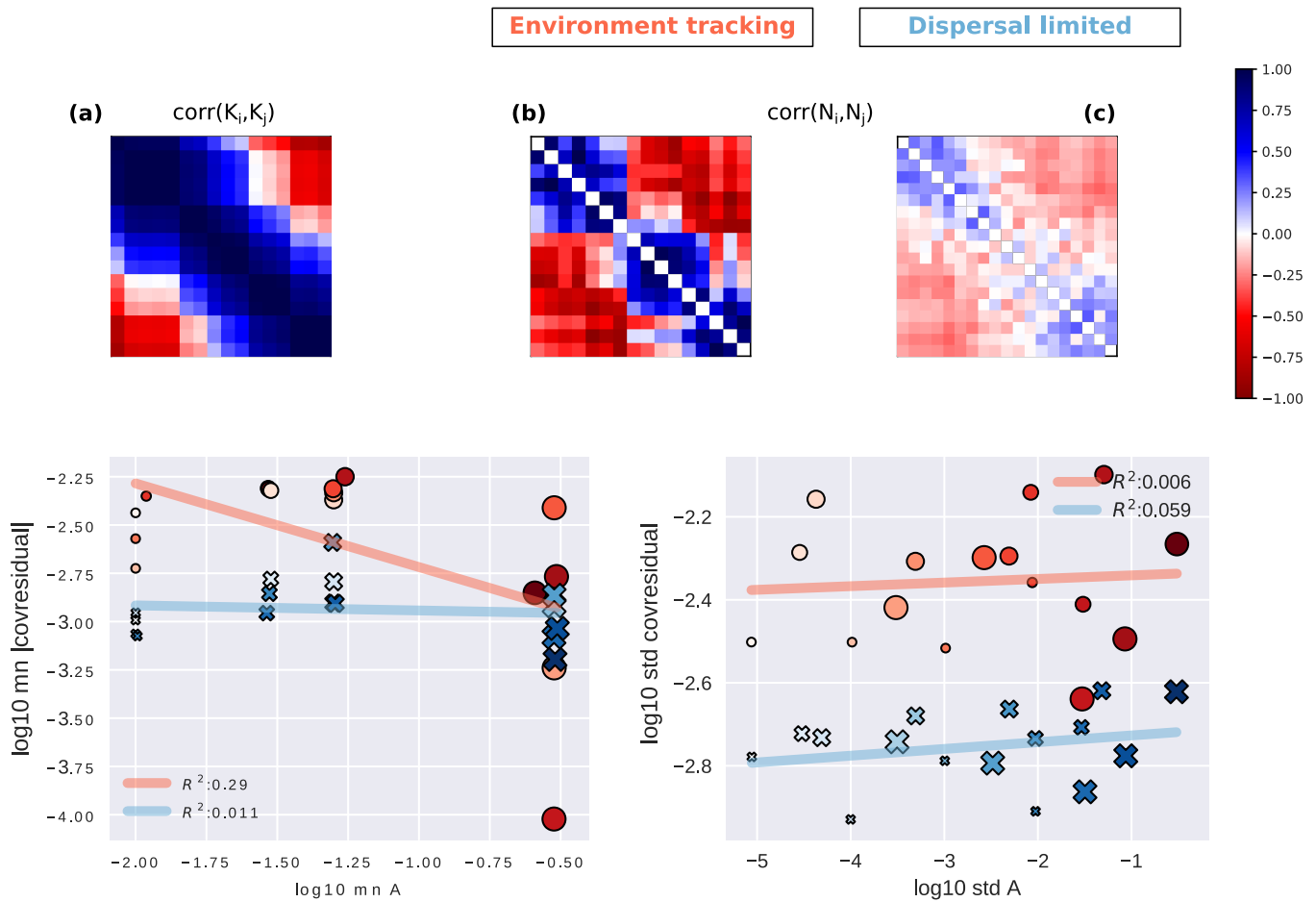


Figure S5: Trying to infer interaction statistics using only the knowledge of covariance or correlation of (a) K and (b-c) N in the two scenarios. From these observations, we compute the residual covariation of species once we control for environment-driven covariation (see Appendix C), and compute its statistics: (d) mean and (e) std. We see no relationship between the statistics of interactions and residual covariation, suggesting that this limited knowledge is insufficient to correctly detect interaction strength

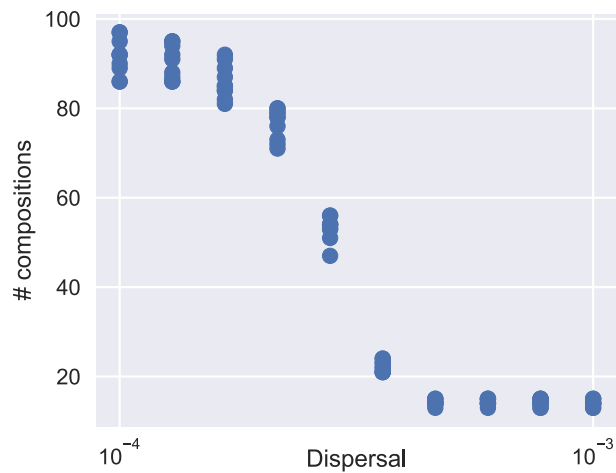


Figure S6: Number of compositions versus dispersal coefficient, showing no abrupt transition contrary to the success of interaction inference in Fig. S6.

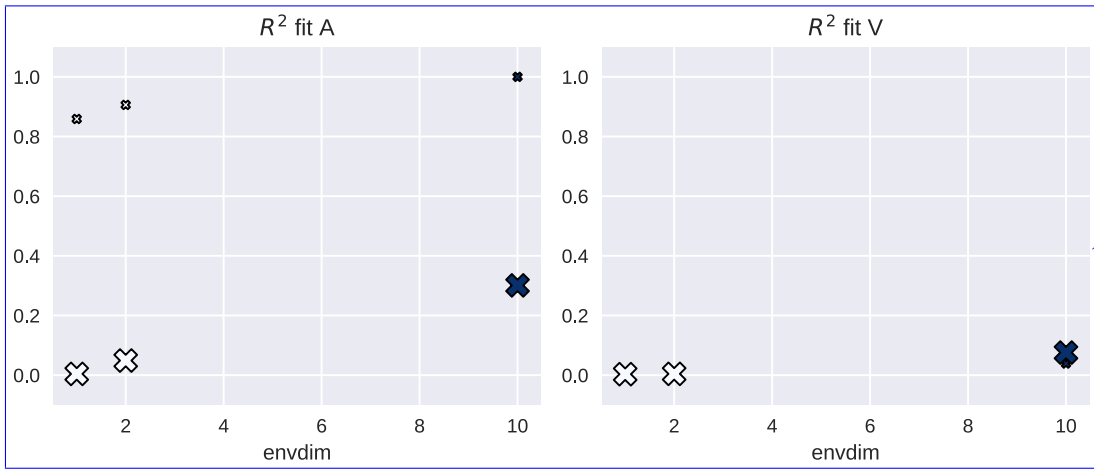


Figure S7: R^2 of inference of direct and net interactions (A and V respectively) depending on the number of independent environmental factors varying through the landscape (1, 2 or 10, shown on the x-axis), and depending on dispersal (small symbols: $D = 0$, large symbols: $D = 0.3$). When there are multiple environmental factors E_j , equation (2) is amended to become $K_j(x) = e^{-\sum_j (E_j(x) - c_{ij})^2 / 2w_{ij}^2} - \mu_j$.