Supporting Information (SI) for "Generic assembly patterns in large ecological communities"

Matthieu Barbier, Jean-François Arnoldi, Guy Bunin and Michel Loreau

The structure of this Supporting Information is as follows:

First, we discuss our assembly perspective on the role of community structure and the emergence of generic patterns in community properties.

Second, we give a systematic description of the numerical experiments, and show which outcomes could or could not be predicted by the reference model.

Third, we carefully describe the reference model and its analytical solution, allowing us to explain what it means to have a "fully disordered system".

Finally, we discuss how to analyze deviations from full disorder, which are important to represent complex communities from which a simple global structure emerges (e.g. hierarchical competition, trophic systems). Previously unpredictable outcomes are found to be correctly predicted by an extended theory, with the addition of supplementary information on this global structure.

Contents

I Introduction

I.1 Assembly

The theoretical study of large communities, whether analytical or numerical, requires some assumptions on the attributes of each species in the community: they can be derived from mechanistic rules, or, barring full knowledge of these rules, sampled from statistical distributions. However, large number of species with randomly sampled attributes are very unlikely to coexist in a stable community [22].

The assembly perspective [19] allows to circumvent this problem: we do not make assumptions about rules or randomness in the *realized* community; instead, we make them in the *pool* of species from which the community emerges through invasions and extinctions. By definition, the community is then a set of species in stable coexistence, and it is generally less random than the pool. Randomness may be more plausible in the species pool, since these candidates for invasion could come from diverse origins, with independent ecological and evolutionary histories.

By computing the stable equilibria of an assembly process, we are effectively asking: what does a community look like when its species have been selected for coexistence by population dynamics (growth, mortality and interactions) alone? Other processes leading to coexistence, such as adaptive or evolutionary dynamics, may induce other community characteristics.

Our control parameters are properties of the species pool. We can talk about generic assembly patterns if pools constructed with different rules and statistical distributions give rise to communities with similar macroscopic properties.

I.2 Model reduction and disorder

Such comparisons allow us to define a protocol of "model reduction": given a model for generating the species pool, we say it is reducible to another model if the latter leads the same quantitative predictions for community properties, with fewer parameters. We can talk about *genericity* if a large and diverse set of models can be reduced to the same model with few parameters (where by "few" we mean a limited number that does not increase when adding more species or more traits to the species). In particular, our analysis hinges on testing whether diverse models inspired by the literature can be reduced to the same minimal *reference model*.

As we explain in Sec. III, a model is *fully disordered* if it can be reduced to our reference model. This reduction has a stark ecological meaning: it signals that each species "sees a fair sample" of the entire community. That is to say, while the "local environment" of each species (its traits and interactions, and those of its neighbors) may be unique, it is statistically representative of the whole community, both in how heterogeneous it is, and in the dynamical feedback it provides. Conceptually, it is similar to the idea of a well-mixed community, except the mixing is not in space but within the interaction web.

This condition is more easily understood from its negation: in less disordered communities, species may for instance occur in guilds with similar interaction patterns, or within a strict hierarchy (such that top predators being only able to feed on mesopredators), instead of each species encountering an unbiased sample of the whole community. Thus, while the drastic reduction to full disorder succeeds for a surprising number of models, it also faces obvious limitations.

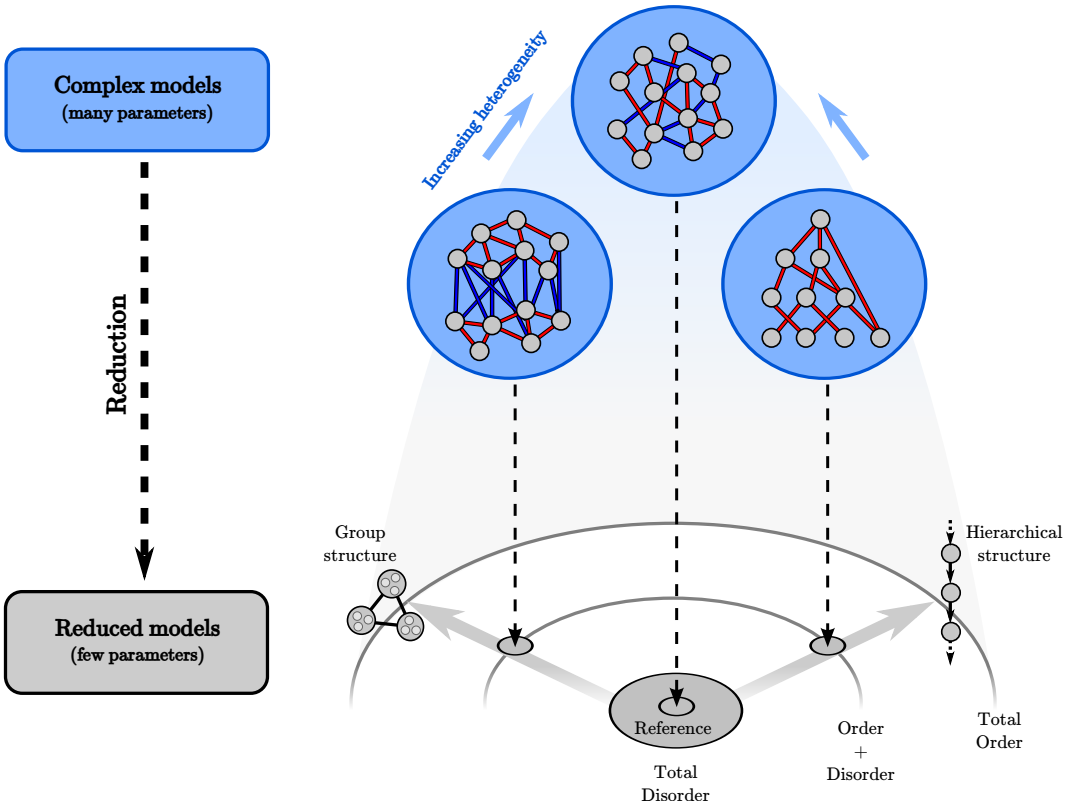

Figure S1: Model reduction. For many complex models (with a large number of parameters), community properties can be predicted from simpler models with relatively few parameters. In the space of these reduced models, the reference (totally disordered) model occupies a central position, from which it is possible to deviate in different directions by adding correlations and structure: for instance, toward hierarchical order, or toward group order. Complexity means that a model is not totally ordered: the more heterogeneity between species, the further away it is from any simple ordered structure – but doing so, it may "move toward the center" i.e. its reduction can become closer to the reference model. For intermediate cases, reduction requires a combination of order and disorder, i.e. one of the extensions of the reference model discussed in Sec. IV.

Yet, even models with more structure can be "mostly disordered", requiring only a few more parameters: rather than tackle the whole complex interaction network, it may be sufficient to consider a minimal extension of our reference model, enriched with simple structural information about hierarchy or group structure, as we show in Sec. IV.

This general framework, which we illustrate in Fig. S1, thus provides a powerful way to combine the points of view of community ecology on heterogeneity and coexistence, and ecosystem theory on large-scale functional structure.

II Numerical experiments

We discuss how the numerical experiments were performed in Sec. II.1, and how their results were compared to the reference model in Sec. II.2.

The full combinatorial space of model features was too vast to be explored systematically. In consequence, two main sets of numerical experiments were performed: a list of specific examples combining various model features among the four categories listed below, described in Table S1, and a systematic exploration of combinations of one interaction type and another model feature, listed in Table S2.

II.1 Simulation models

Simulations were performed by numerically integrating the dynamical equations[1]

$$\frac{d}{dt}B_i = B_i\left(r_i - D_iB_i - f\left(\sum_{j\neq i}^{S} A_{ij}B_j\right)\right) \tag{S1}$$

Note that for simulations, the abundance of species i is expressed by B_i, denoting its absolute biomass as it would appear in data or in realistic simulations. In the main text and in our analysis in Sec. III, we use rescaled abundance variables $N_i = c_iB_i$, where the choice of scaling constants c_i depends on the model as explained in Sec. III.1. For instance, in a competitive system where carrying capacities span many orders of magnitude, we may want to define N_i as the ratio of biomass to carrying capacity, so that all the N_i are on the same scale.

While results for N_i are generic, their translation to absolute values B_i depends on the rescaling choice, and thus on the model. In particular, we find that the distribution $P(N)$ is generally a truncated Gaussian, while the resulting $P(B)$ can be more realistically lognormal or otherwise fat-tailed, as discussed in Sec. III.3.

Each distinct simulation model consisted in a choice of a functional response f, and a set of rules and parameters to generate the coefficients r_i, D_i and A_{ij}. We decomposed the interaction matrix into

$$A_{ij} = G_{ij}a_{ij} \tag{S2}$$

where G_{ij} is the unweighted adjacency matrix of the network structure, and a_{ij} are random variables. Hence, each model can be characterized by the following options which are detailed below in their respective sections II.1.1-4:

List of model features (with additional parameters in parentheses)

1. Functional response:
   (a) Linear

[1] We used the LAPACK integrator DOP853 provided by the SciPy library in Python.

(b) Saturating (B_c)

2. Network structure G_{ij}
   (a) Random (connectivity c)
   (b) Scale-free (p_d)
   (c) Assortativity ($|\mathrm{corr}(z_i, z_j)|$)
   (d) Clustering (p_t)
   (e) Partition (p_b)
   (f) Cascade (p_c)
   (g) Nestedness (p_n)
3. Parameterization
   (a) Independent r_i, D_i and a_{ij}
   (b) Predation (β_{ij}, ϵ)
   (c) Independent carrying capacities (k_i)
   (d) Correlated r_i and a_{ij} (C_{ra})
   (e) Correlated rows in a_{ij} (σ_{row})
   (f) Resource competition (ξ_{ia}, ρ_a, m_i)
   (g) Functional groups (inter- and intra-group parameters)
4. Trait distributions (distribution family, mean, variance, symmetry)
   (a) No rescaling
   (b) Moderate interactions: rescaling $\max|A_{ij}| = 0.1\, D_i$
   (c) Weak interactions: rescaling $\max|\sum_j A_{ij}| = 0.5\, D_i$
   (d) Diffuse interactions: rescaling $A_{ij} \propto 1/d$ (d)

We show in Table S1 which choices and parameter values were used for all simulation results in the main text and SI. Random initial conditions on B_i were drawn for each simulation. The asymptotic equilibrium of the assembly process was attained by allowing an arbitrary number of extinctions and reinvasions: numerically, this was done by setting a threshold for species extinction $B_e = 10^{-15}$, and letting

$$\frac{dB_i}{dt} = \max\left(0, \frac{dB_i}{dt}\right) \qquad \text{if } B_i < B_e. \tag{S3}$$

Thus, species that fell under the extinction threshold would retain the same abundance unless dB_i/dt became positive again, allowing for reinvasions. As Lotka-Volterra dynamics are notably stiff, with rare species dynamics taking very long time to converge, we set a threshold for convergence of

$$\max_i \frac{d}{dt} \log B_i < 10^{-8} \tag{S4}$$

and while this may fail to capture extremely long time dynamics, we generally assumed that aggregated community properties, such as those we measure, would already be very close to their asymptotic value.

II.1.1 Functional response

1a Linear functional response

$$f(z) = z \tag{S5}$$

Linear functional response is the basic frame in which the reference model is defined in Sec. III.

Models in figures

Figure	Name of model in figure	Model features
Fig. 1-3	Resource competition	1a, 2a, 3f, 4a
Fig. 1	"C-R mutualism"	1b, 2a, 3a, 4a
	"Cascade predation"	1a, 2f, 3b, 4b
	"Spatial competition"	1b, 2a, 3c, 4d
Fig. 3	Predation	1a, 2a, 3b, 4a
Figs. 4, S4	Competition	1a, 2a-g, 3a, 4a
	Mutualism	1a, 2a-g, 3a, 4c
	Predation	1a, 2a-g, 3b, 4a
Fig. 5	"Plant-pollinator"	1a, 2e, 3g, 4a
Fig. S3	Competition	1b, 2a-f, 3a, 4a
	Mutualism	1b, 2a-f, 3a, 4c
	Predation	1b, 2a-f, 3b, 4a

Default values

$\langle r_i \rangle$	1
$\mathrm{Var}(r_i)$	0.1
$\langle D_i \rangle$	1
$\mathrm{Var}(D_i)$	0
$\langle a_{ij} \rangle$ or $\langle f_{ij} \rangle$	0.1
$\mathrm{Var}(a_{ij})$ or $\mathrm{Var}(f_{ij})$	0.01
Species number S	100
Connectivity c	0.1
Threshold B_c	5
Symmetry Γ	0.5

Table S1: Model choices and parameter values for results in the main text and SI. *Left:* For each figure and each named model within that figure, we give the corresponding model features as listed above. For Fig. 1, the example models we list here are not exactly those appearing in the source materials (which may not be compatible with the Generalized Lotka-Volterra form we use), but they retain some of their important features: for instance, the simulation model inspired by spatial competition corresponds to the choice of a saturating functional response 1b, random network 2a, independent parameters 3a and diffuse interactions 4d. *Right:* Default parameter values used unless otherwise specified.

1b Saturating functional response This saturation can come from many factors: limited provision or diminishing returns for mutualistic services [14], finite handling time for predators [15] or competition with only close neighbors in space [26].

$$f(z) = \frac{z}{1 + \dfrac{|z|}{AB_c}} \tag{S6}$$

with $A = \langle A_{ij} \rangle$ and B_c an additional parameter giving the population threshold for saturation (see Sec. III.5). This analysis requires the addition of one new parameter, B_c the saturation threshold, from which we recover the basic reference model in the limit $B_c \to \infty$.

II.1.2 Network structure

2a Connectance The overall number of nonzero entries in the interaction matrix is given by $cS(S-1)$ with c the connectance. Simulations for arbitrary connectance agreed with the reference model, and in particular, maintaining the same parameters (μ, σ, γ) while changing connectance would provide identical results (except when this led to very strong interactions at very low connectance, see [3]). For all network structures below, we used fixed connectance $c = 0.1$ by default.

2b Scale-free degree distribution We generated a scale-free network using the Barabasi-Albert algorithm [1], and then rewired the edges randomly with probability $1 - p_d$ to interpolate between scale-free and uniformly random (Erdos-Renyi [11]) degree distributions.

2c Assortativity Correlation between the degrees z_i and z_j of nodes i and j connected by an edge. We could create assortative networks (where hubs are connected together) or disassortative networks (where high-degree nodes are mainly connected to low-degree nodes). We took random networks and

rewired them to tune assortativity, keeping the same number of in- and out- edges for each node, but swapping partners to increase degree correlation or anticorrelation.

2d Clustering We follow the standard interpretation of clustering as a measure of how many triangles there are in a network [28]. We took random networks and rewired them to tune clustering. Our algorithm to do so was the following. First, we compute length-2 paths in the network (by taking the square of the adjacency matrix), then list pairs of nodes that are connected only by a length-2 path, and those that are connected both by a direct interaction and a length-2 path. With probability $|p_t|$, triangles are created by adding direct interactions between pairs of nodes in the first list, or (if $p_t < 0$) destroyed by removing direct interactions in the second list; swapping partners with pairs that appear in neither list allows us to do the same without changing node degrees[2].

2e Partition The network was made more or less bipartite by taking a random network, assigning each species a group index (either 1 or 2), then deleting intra-group links with probability p_b (thus $p_b = 0$ would be a random network, and $p_b = 1$ a perfectly bipartite network).

2f Cascade The classical cascade model [8] has a triangular matrix with only nonzero elements within the triangle. Allowing zeros, we find Directed Acyclic Graphs. We tuned ordering by taking a random network and varying p_n the probability that α_{ij} was set to zero for $i \leq j$. For $p_c = 0$, the matrix was random, while for $p_c = 1$ it was upper-triangular.

2g Nestedness Nestedness is the flipside of directedness: in a fully nested matrix, all elements are contained on one side of the *anti*-diagonal. We tuned nestedness by taking a random network and varying p_n the probability that α_{ij} was set to zero for $i \leq S - j$. For $p_n = 0$, the matrix was random, while for $p_n = 1$ it was fully nested.

II.1.3 Parameterization

3a Independent For competition and mutualism, we could directly draw r_i, D_i, and a_{ij} independently from the weight distributions listed in Sec. II.1.4, with a_{ij} having positive or negative mean respectively.

3b Predation In a predation model, we first drew a matrix β_{ij} of "predation intensity" coefficients as if they were competition coefficients (with average β and other parameters depending on the choice of distribution, see below). We then chose, within each pair, one species to be the predator and one to be the prey. Whenever we had $\beta_{ij} \neq 0$ and $\beta_{ji} = 0$ (e.g. in the nested structure above) we decided that j was the predator, else we randomly selected either i or j. Then, if for instance j was the predator, we set $\alpha_{ij} = \beta_{ij}$ and $\alpha_{ji} = -\epsilon\beta_{ij}$, with ϵ the biomass conversion efficiency of the trophic interaction, which was set here to $\epsilon = 0.1$.

3c Independent carrying capacities In this parameterization discussed in Sec. III.1, we drew three independent sets of parameters: growth rates r_i, carrying capacities k_i and effective interactions C_{ij}, then set

$$D_i = \frac{r_i}{k_i}, \qquad A_{ij} = \frac{r_i}{k_j} C_{ij}. \tag{S7}$$

Note that these independent parameters k_i represent the actual carrying capacities of species, in the same units as B_i, while the K_i appearing below are nondimensional variables that depend on the choice of scaling in Sec. III.1.

[2] All the while, the lists must be maintained dynamically after each operation.

3e Correlated rows While the basic reference model assumes no correlations in the coefficients drawn above, we considered two simple ways of adding correlation structure to their distributions, to test how they cause a deviation from the reference model and how the latter can be extended to account for them. We could introduce row correlation in the interaction matrix: instead of drawing the matrix α directly, we drew a matrix M according to our other rules, then let $\alpha_{ij} = c_i M_{ij}$ with some coefficients c_i drawn independently (e.g. normal).

3d Correlated growth and interactions Similarly, we could introduce correlations between a and r by having $c_i = c(K_i)$ with some function c, for instance $c(K_i) = K_i$. As for strong interactions, such correlations may be minimized by rescaling variables (see Sec. III.2), but we wished here to test how correlations modify results and can be accounted for by analytical extensions.

3f Resource competition This parameterization is explained in detail in Sec. II.1.5. In short, we derived r_i, D_i and a_{ij} from three different sets of parameters: the abundance ρ_a of resource a, the consumption efficiency ξ_{ia} of resource a by species i, and the mortality m_i of species i. Each of these traits were drawn independently from exponential distributions with prescribed mean (see Sec. II.1.5)

3g Functional groups Each species was assigned to one of n groups. We let $D_i = 1$ for all species i. Then, for each group x we drew growth rates with prescribed mean and variance. For each pair of groups x and y, we drew interactions with mean, variance and symmetry. Thus, the total parameters were two vectors (mean and variance of growth rates) and three $n \times n$ matrices (mean, variance and symmetry of interactions).

II.1.4 Trait distributions

Many models called for interaction weights a_{ij}, as well as growth rates r_i and self-interactions D_i, to be drawn directly from probability distributions. In those cases, we always set $D_i = 1$ for simplicity in our simulations (although our approach remains valid otherwise). For specific model parameterizations such as 3c or 3f above, these coefficients were not drawn directly, but computed from other species traits which had to be generated first.

For all these traits, we tested different distributions: normal, uniform, exponential, bimodal (two discrete values) and fat-tailed (power-law with exponent 2). Each of these distributions could be parameterized to set the mean, and most allowed to set the variance, of the generated trait values. Unless otherwise specified in the description of model parameterizations, we used normal distributions with mean and variance selected to obtain the default values in Table S1.

In the case of interaction coefficients a_{ij}, we also wanted to tune their symmetry i.e. the correlation between a_{ij} and a_{ji}. To do so without changing their variance, we can first draw a matrix M of independent coefficients M_{ij} according to the rules selected for a_{ij}, then construct the interaction matrix as follows

$$a_{ij} = \frac{M_{ij} + sM_{ji}}{\sqrt{1+s^2}} \qquad \text{with } s = (1 - \sqrt{1-\Gamma^2})/\Gamma \tag{S8}$$

where we choose the value of auxiliary parameter $\Gamma \in [-1, 1]$ to make interactions more or less symmetrical ($\Gamma = -1$ yields antisymmetrical weights, and $\Gamma = 1$ makes them symmetrical).

4a No rescaling

4b Moderate interactions In most cases, we wished to find stable and diverse equilibria, which are generally difficult to attain with strong individual interactions (such as competitive exclusion causing all species but the strongest competitor to go extinct) in the absence of precise tradeoffs.

Hence, we ensured that interaction magnitudes were always individually smaller than intra-species competition, rescaling the entire matrix so that

$$\max_{ij} |A_{ij}| = 0.1 D_i, \tag{S9}$$

although the sum interaction of a species with all others was generally large ($\left|\sum_j \alpha_{ij}\right| > 1$).

4c Weak interactions Moderate interactions generally ensured the existence of stable multi-species equilibria, except in the case of mutualism with a linear functional response where, to prevent population explosion, we instead rescaled the entire matrix so that

$$\max_i \left| \sum_j A_{ij} \right| = 0.5 D_i \tag{S10}$$

4d Diffuse interactions In diffuse interactions [16], for instance spatial competition between plants, if all space is occupied by some plant, then the more species occupy the same area, the fewer individuals from two given species of plants are going to be in contact with each other (e.g. if each plant has a few neighbors, the more species there are in the system, the less likely it is that a given species is found in the neighborhood of a given plant). To represent this, we used the rescaling

$$A_{ij} \propto \frac{1}{d} \tag{S11}$$

where d is the effective number of species that could figure among interaction partners for a given individual. This parameter depends on S and can be modulated by spatial aggregation, i.e. $d = S$ if species are perfectly well-mixed, and $d = O(1)$ if the same species keep interacting no matter how many other species are introduced.

II.1.5 Example models

While a systematic exploration of combinations of the above features is beyond our purpose, we used some as examples in the main text and to illustrate the sort of ecological settings that can be analyzed within our approach. We now present these specific model choices (whose basic features are listed in Table S1 and further detailed below when necessary) with their ecological motivation.

Cascade predation The combination of predatory interactions and a nested structure is reminiscent of the cascade model [8], whose structure has been and remains in use in many theoretical works on trophic structure, both qualitative and quantitative, alongside more modern proposals such as the allometric niche model [5]. The main difference with these models is that we did not distinguish basal species, which should be the only ones to have positive carrying capacities K_i, meaning that all species here grow on external resources. A realistic model of trophic structure certainly requires this distinction in addition to the triangular interaction matrix, but we wanted here to separately address groups and hierarchical structure as basic model ingredients, as we do in Table S2, and show that each of them causes deviations from the reference model, but can be addressed by its extensions described in Sec. IV.

C-R mutualism Lotka-Volterra dynamics are claimed [14] to fail to represent mutualistic interactions, notably by allowing for boundless population growth. The most common change introduced in the description of mutualistic interactions is a saturating functional response. This represents the simplest case considered in consumer-resource mutualism [13], and these communities are fully predicted by the reference model with the appropriate functional response.

Spatial competition The consequences of spatial competition (e.g. competition for light between plants) have rarely been modelled directly at the population level, instead relying on spatially explicit and even individual-based models [4]. However, a major contribution was the spatial moment perspective of Law and Dieckmann [18], where spatial structure appears as a correction of the first-moment (i.e. spatially averaged) Lotka-Volterra equations, using the second-moment equations (i.e. variance in space).

Here, we approximate this effect further (the validity of these approximations could be tested by comparison to spatially explicit models). First, we must account for antagonistic interactions with at most a finite number n of neighbors: we use a connectivity $c = n/S$ and a saturating functional response with threshold $B_c = n$ (using only one or the other leads to similar results). Second, we must account for the fact that, if individuals are well-mixed spatially, then the more species there are, the fewer individuals from any two species interact together. We do so by scaling interactions as $A_{ij} \propto 1/d$ where d can be constant or proportional to S (see "Diffuse interactions" above). Note that connectivity and interaction scaling allow to control two of the reference parameters, as seen in Fig. 2 in the main text: $\mu \propto n/d$, and $\sigma \propto \sqrt{n}/d$.

Finally, we allow species to have vastly different carrying capacities k_i, and use the parameterization $N_i = B_i/k_i$, $K_i = 1$, $\alpha_{ij} = k_j A_{ij}/D_i$.

Resource competition In this discretized version of a classic ecological model [20, 12], S species compete over R abiotic resources which are steadily resupplied into the ecosystem. It is generally known from competitive exclusion theory that $R > S$ is necessary to at least potentially allow S species to survive – however, these many resources need not all differ by their nature, they could simply be distinguished by spatial or temporal availability within the ecosystem, with some species having an advantage at capturing certain patches rather than others. We define the net growth rate of species i as

$$g_i = \sum_{a=1}^{R} \rho_a(t)\xi_{ia} - M_i \tag{S12}$$

with consumption rate ξ_{ia} for resource a, whose abundance is given by

$$\rho_a(t) = \rho_a - \sum_i \xi_{ia} B_i(t) \tag{S13}$$

where ρ_a is the steady influx of the resource. On the other hand, M_i represents mortality from the energy costs associated with resource acquisition. To avoid favoring specialists or generalists a priori, we make these costs proportional to the total ability of a species to acquire resources:

$$M_i = \rho m_i \sum_a \xi_{ia} \tag{S14}$$

where we factor out the average amount ρ of resource in the system so that m_i is now a dimensionless number representing the intrinsic lack of fitness of species i. Hence, we get the dynamical equations

$$\frac{d}{dt}B_i = g_i B_i = B_i \left(\sum_a \xi_{ia}(\rho_a - \rho m_i) - \sum_j B_j \sum_a \xi_{ia}\xi_{ja} \right) \tag{S15}$$

It is easy to see that they map onto the Lotka-Volterra equations with

$$r_i = \sum_a \xi_{ia}(\rho_a - \rho m_i), \tag{S16}$$

$$D_i = \sum_a \xi_{ia}^2 \tag{S17}$$

$$A_{ij} = \sum_a \xi_{ia}\xi_{ja}. \tag{S18}$$

Let us assume a disordered pool where consumption rates, resource influxes and mortality are all drawn as independent random variables with mean

$$\xi = \langle \xi_{ia} \rangle , \qquad \rho = \langle \rho_a \rangle , \qquad m = \langle m_i \rangle \tag{S19}$$

and standard deviation σ_ξ, σ_ρ and σ_m. Among other things, this implies that all consumers are generalists who can consume any resource but perform better at some. Without loss of generality, we can always set

$$\langle r_i \rangle = R\rho\xi(1-m) = 1 \tag{S20}$$

$$\langle D_i \rangle = R \left\langle \xi^2 \right\rangle = R(\xi^2 + \sigma_\xi^2) = 1. \tag{S21}$$

using a rescaling of time and of B_i (see [3]). The second equation means that there is a tradeoff between the effective intensity ξ and variability σ_ξ of consumption rates.

We thus generated many species pools, using two different distributions of ξ_{ia} (either normal or bimodal) and a normal distribution of ρ_a, and exploring systematically a range of parameters R, σ_ξ (which fixed ξ by the relation above) and σ_ρ. By default, we used $\sigma_\xi^2 = 0.1/R$ and $m = 0.1$; values for the other parameters were explored in Fig. 2 of the main text.

Using the parameterization $N_i = B_i$, $K_i = r_i/D_i$, $\alpha_{ij} = A_{ij}/D_i$, we then computed parameters μ, σ, γ and ζ, as well as the correlation $\langle K_i \alpha_{ij} \rangle_{ij}$. Ignoring the latter led to slightly worse quantitative agreement. We inserted these parameters into the reference model, either with or without that correlation (see Basic correlation structure in Sec. IV), and compared predictions and simulation outcomes for all the community properties listed in Sec. III.4.

Plant-pollinator community This community of $S = 300$ species was divided into two functional groups, one given no intrinsic growth i.e. $K_i = 0$, while the other had K_i drawn with mean 1 and variance ζ. We then drew two types of interactions α_{ij}: mutualistic and competitive, both as exponential distributions with mean -0.01 and 0.01 respectively (larger mutualistic interactions caused population explosions since we did not use a saturating functional response here). The ordering parameter ω decided the probability that mutualistic interactions were exclusively assigned to inter-group links and competitive interactions to intra-group links. For $\omega = 0$, both types of interactions were assigned at random, ignoring the group labels.

We then compared simulation results to the mixture (group structure) model detailed below in Sec. IV: within each group and between groups we measured the average, variance and symmetry of interactions, such as to construct the μ, σ and γ matrices. The fully ordered case $\omega = 1$ had zero variance, i.e. $\sigma = 0$ for each set of interactions, while the intermediate cases $0 < \omega < 1$, where complexity blurred group boundaries, led to nonzero variance and a different mean and symmetry.

We could then insert the μ, σ and γ matrices, as well as the parameters $\langle K \rangle$ and ζ for each group, into the calculations below to compute the abundance distribution and community properties, both per functional group and collectively. This procedure assumed that, even when group boundaries were blurred, we knew which group each species belonged to. In real systems, we would have to infer this clustering of species into groups, but we leave such developments for future work.

II.2 Validation

In this section, we discuss how we compared the results of numerical simulations to the reference model, with results reported in Table S2. We provide in Appendix a description of the algorithms and a sample R code allowing to perform the comparison.

II.2.1 Comparison to the reference model

Computing the reference parameters Simulation models are entirely defined by the functional response f and parameter sets r_i, D_i and A_{ij}. However, we need much less information to parameterize

the reference model. Let us focus here on the case of a linear functional response $f(z) = z$, as the reference model extensions for other choices are discussed in Sec. III.5.

First we must rewrite the equilibrium condition as

$$0 = N_i(K_i - N_i - \sum_{j \neq i}^{S} \alpha_{ij} N_j) \tag{S22}$$

with the following parameters

$$N_i = c_i B_i, \qquad K_i = \frac{c_i r_i}{D_i}, \qquad \alpha_{ij} = \frac{c_i A_{ij}}{D_i} \tag{S23}$$

where we note N_i the effective species abundance, α_{ij} the effective interactions and K_i the effective carrying capacity. The choice of coefficients c_i depends on how the parameter sets r_i, D_i and A_{ij} have been generated (i.e. are they independent parameters, or derived from other underlying traits, causing them to be correlated), see discussion in Sec III.1. In all simulations above, we used $c_i = 1$.

Then, as we explain in the analysis of the reference model (Sec. III), the four reference parameters are defined as

$$\zeta^2 = \overline{K_i^2} - \overline{K_i}^2, \qquad \mu = S\overline{\alpha_{ij}},$$
$$\sigma^2 = S(\overline{\alpha_{ij}^2} - \overline{\alpha_{ij}}^2), \qquad \gamma = \frac{S}{\sigma^2}\left(\overline{\alpha_{ij}\alpha_{ji}} - \overline{\alpha_{ij}}^2\right)$$

where $\overline{X}$ is the average over X including zero elements – thus, the value of these parameters depends not only on the distribution of interaction weights, but also on the network structure specifying which elements are zero or nonzero. With a complete graph structure (i.e. all elements in the matrix α are nonzero), we have direct control over the symmetry parameter $\gamma = \Gamma$ (where the latter is the control parameter described in Sec. II.1.4); else, γ may depend on network structure as well.

These four parameters $\zeta, \mu, \sigma, \gamma$ are the only information required to numerically solve the reference model (see Sec. III and Appendix for the algorithm), which outputs predictions for various community properties.

Comparing simulation outcomes At the end of each simulation, we compute the total biomass

$$T = \sum_i B_i, \tag{S24}$$

total productivity

$$P = \sum_i r_i B_i, \tag{S25}$$

Simpson index

$$D^{-1} = \sum_i \left(\frac{B_i}{T}\right)^2, \tag{S26}$$

and variability in response to demographic[3] noise

$$V = \sum_i \frac{\mathrm{Var}(B_i(t))}{S}. \tag{S27}$$

[3]See discussion in Sec. III.4 for why we consider demographic noise here; this quantity allows to predict variability to environmental noise as well.

Rather than add noise to the simulation and compute the variability with the formula above, it can be evaluated directly [2] (in the linear regime, i.e for small perturbations around an equilibrium) as the trace of the covariance matrix $\mathcal{V}_{ij}$

$$V = \sum_i \mathcal{V}_{ii} \tag{S28}$$

where $\mathcal{V}_{ij}$ is obtained by solving the Lyapunov equation[4]

$$J\mathcal{V} + \mathcal{V}J^T = \mathcal{N} \tag{S29}$$

with $\mathcal{N}$ the diagonal matrix such that $\mathcal{N}_{ii} = N_i$, and J the Jacobian matrix given by

$$J_{ij} = B_i(A_{ij} - D_i\delta_{ij}), \qquad \delta_{ij} = 1 \text{ if } i = j, \text{ 0 otherwise.} \tag{S30}$$

We then compare these results to the analytical predictions computed from the reference model, as given in Sec. III.4.

II.2.2 Results

We now discuss the results presented in Table S2.

Functional response A saturating functional response leads to different results, which can however be captured quantitatively by the expanded analysis presented in Sec. III.5. Since $B_c \to \infty$ is the Lotka-Volterra case, and the $B_c \to 0$ limit is that of independent species, we can generally conjecture that a saturating functional response makes community structure less important than in Lotka-Volterra dynamics, as we indeed see in Fig. S3.

Network structure No network structure caused any significant deviation from the reference model, except for partitioning, which we capture with the mixture model in Sec. IV.1, and directedness (cascade structure) or nestedness which we capture with the model in Sec. IV.2 and Fig. S4.

Parameterization The only deviation from the reference model came from parameterizations that introduced correlations between or within the quantities K_i and α_{ij} defined in (S23). Accurate predictions could however be recovered, using the extended model with an additional parameter representing first-order correlations, see Sec. IV.3.

Trait distributions The interaction matrix could be qualitative (binary edges) meaning that we set $a_{ij} \equiv 1$ in (S2), or quantitative (weighted edges) meaning that we drew a_{ij} according the the rules below. We found that the reference model worked equally well in both cases, meaning that even with qualitative interactions, disorder can prevail simply due to network topology.

As we discuss in Sec. III.2, the analytical inaccuracies coming with strong interactions may be avoided or minimized by rescaling variables when studying specific models (e.g. competition between species with lognormal carrying capacities). Cases where we did not perform any rescaling appear in Table S2 as *strong interactions* if the average interaction was strong $\langle\alpha_{ij}\rangle > 1$, or *fat tails* if only a few such coefficients were found.

III Reference model and community properties

While the discussion given here is meant to be self-contained, the reference model was solved by Bunin [7], and it is introduced and discussed in much greater detail in pedagogical notes available online [3], to which we direct the curious reader.

[4] It was solved numerically using the Bartels-Stewart algorithm, via the linear algebra package of the SciPy library.

	Competition	Predation	Mutualism
Functional response			
Linear (1a)			
Saturating (1b)	Threshold III.5		
Network structure			
Binary/weighted			
Connectance (2a)			
Degree distribution (2b)			
Assortativity (2c)			
Clustering (2d)			
Partition (2e)	Groups IV.1		
Cascade (2f)	Continuous order IV.2		
Nestedness (2g)	Continuous order IV.2		
Parameters			
Interaction weights a			
Normal, uniform			
Exponential, bimodal			
*Strong (mean$\gtrsim 1$)			
*Fat tails			
*Row correlation (3e)	Correlations IV.3		
Growth rates r			
Normal, uniform			
Exponential, bimodal			
*Fat tails			
*Correlations r-a (3d)	Correlations IV.3		
Mechanistic (3c,3f)			

Color	Agreement (error $< 5\%$)
	Reference model
	Reference with few exceptions
	Reference + one parameter
	Reference + more parameters
	Outside of method scope
	No stationary state

Complex examples

Cascade predation	C-R Mutualism	Spatial competition	Resource competition	Plant-pollinator
Continuous order IV.2			Correlations IV.3	Groups IV.1

Table S2: Overview of simulation model properties and their impact on predictability using the reference model (Sec. III) and its variants (Sec. IV). Here, we systematically test pairs of one interaction type and another model feature; more complex combinations were not tried exhaustively, but they appear in the examples listed in the bottom row, described in Sec. II.1.5. *Left:* Predictability for each of the three main interaction types combined with various other traits (in parentheses: corresponding model choices from the list in Sec. II.1). Traits marked by * could, in some models, be eliminated or reduced by a rescaling of the variables, see Sec. II.2.1). Whenever the fully disordered limit must be supplemented with additional information to allow good predictions, we note the required parameter(s) and refer to the relevant extension of the random model (text in the colored cells). *Right:* Agreement between simulations and analytical predictions is judged on a quantitative basis (the error metric defined in Sec. III.4 should be below 5% for all tested parameters). Colors represent whether agreement is found with the basic model (blue), with a simple extension (e.g. for functional response) which adds a single parameter and recovers the reference model in a given limit of this parameter (green), or with a less generic model accounting for robust network structure (purple). Fainter colors represent cases where specific parameter choices can cause quantitative disagreement, without compromising qualitative agreement. Some scenarios, shown in dark grey and black, are outside of the scope of our method altogether, or even incompatible with the premise of equilibrium solutions.

The reference model can be seen as a "maximally generic" model of community assembly, both in terms of its dynamical processes, and parameter structure. Lotka-Volterra dynamics (in the broadest sense) are among the simplest dynamics which allow for species extinction, a key factor in shaping the final assembled community. Another simplifying assumption is that we only consider the equilibrium reached asymptotically at the end of an arbitrarily long sequence of invasions and extinctions, where any species from the external pool has the opportunity of reinvading at any time. We observe in simulations that, while details of community composition may keep changing over long timescales, community-level aggregate quantities of interest are well described by their predicted equilibrium values.

In a certain parameter range which we determine below, the composition of the assembled community is then controlled only by the properties of the global pool of species, instead of depending on the precise assembly process and sequence. Finally, the pool structure is minimally described by one set of parameters characterizing species indivdiually, and another characterizing interactions, and these two parameter distributions are here reduced to their first few moments.

III.1 Dynamics and parameterization

We start from Lotka-Volterra dynamics (see Sec. III.5 for a saturating functional response):

$$\frac{d}{dt}B_i = B_i\left(r_i - D_iB_i - \sum_{j\neq i}^{S} A_{ij}B_j\right) \tag{S31}$$

with B_i the biomass of species i, r_i its intrinsic growth rate, $D_i = r_i$ the density-dependent mortality (also known as intra-species competition), and A_{ij} the bare interactions, expressed as gain or loss of biomass per capita of both species i and j. These three sets of parameters entirely characterize the species pool.

However this expression is not the only possible parameterization. For instance, in competitive communities, it has been shown [23] that a useful parameterization is

$$\frac{d}{dt}B_i = r_iB_i\left(1 - \frac{B_i}{k_i} - \sum_j C_{ij}\frac{B_j}{k_j}\right). \tag{S32}$$

with k_i the real carrying capacity (in units of biomass) of species i and C_{ij} effective interactions. This expression is a better choice if we expect k_i and C_{ij} to be independent properties, while r_i, D_i and A_{ij} are not. In particular, we can choose k_i to be very widely distributed (possibly over many orders of magnitude) to reproduce empirical abundance distributions, and C_{ij} to be more narrowly distributed to permit species coexistence, as we explain below.

Thus, the first step of our model simplification is to choose a parameterization that will lead to maximal independence and minimal heterogeneity of the species attributes, i.e. that will minimize their covariance and variance. The main objective is to avoid the cases listed in Table S2 where our analytical method breaks down, although cases where there are correlations in the coefficients can be tackled by simple extensions of the reference model (see Sec. IV).

Once such a parameterization is found, we can always express the equilibrium condition $dB_i/dt = 0$ as

$$0 = K_i - N_i - \sum_{j\neq i}^{S} \alpha_{ij}N_j \tag{S33}$$

where we note N_i the effective species abundance, α_{ij} the effective interactions and K_i the effective carrying capacity (which can be negative if species i requires others to persist).

The first parameterization S31 corresponds to

$$N_i = B_i, \qquad K_i = \frac{r_i}{D_i}, \qquad \alpha_{ij} = \frac{A_{ij}}{D_i} \tag{S34}$$

where N_i is simply the species biomass, which is the one we will use throughout unless otherwise specified. The second parameterization S32 corresponds to

$$N_i = \frac{B_i}{k_i}, \qquad K_i = 1, \qquad \alpha_{ij} = C_{ij} \tag{S35}$$

where N_i is now the "occupation fraction" of a species' carrying capacity. Let us assume that k_i is spread over orders of magnitude; then, if r_i and A_{ij} are i.i.d., abundant species j may exert a competition pressure $A_{ij}B_j$ on rare species i which is hundreds or thousands of time stronger than the latter's growth r_iB_i, guaranteeing the latter's extinction. By contrast, if C_{ij} are i.i.d., competition pressure from species j will now depend on B_j/k_j, which does not vary as dramatically as B_j, and rare species can coexist with abundant species.

Our results are all computed from (S33); hence, although we will always talk about "abundance", "fitness" and 'interactions", their interpretation can differ depending on the original parameterization.

III.2 Model reduction and minimal parameters

As argued above, for any model with Lotka-Volterra dynamics (i.e. linear functional response), the equilibria of the assembly process are entirely controlled by two properties of the species pool: interactions α_{ij} and carrying capacities K_i. By "model" we thus mean a set of rules and parameters which decide how α_{ij} and K_i are distributed, see e.g. Sec. II for the description of our various simulation models. For instance, the resource competition model derives interactions and carrying capacities from other, more fundamental species characteristics such as their consumption rate for a given resource, itself drawn from some probability distribution.

We can talk of a "model reduction" if there exists a model with fewer parameters whose assembly process will lead to the same quantitative predictions for all macroscopic community properties listed in Sec. III.4, such as diversity, functioning and stability. This reduction may hold for all, or only part, of the parameter space of the original model.

We call a model "fully disordered" when it can be reduced to our reference model, where species fitness and interactions are characterized by only four parameters

$$\begin{aligned} \zeta^2 &= \overline{K_i^2} - \overline{K_i}^2, & \mu &= S\overline{\alpha_{ij}}, \\ \sigma^2 &= S(\overline{\alpha_{ij}^2} - \overline{\alpha_{ij}}^2), & \gamma &= \frac{S}{\sigma^2}\left(\overline{\alpha_{ij}\alpha_{ji}} - \overline{\alpha_{ij}}^2\right) \end{aligned} \tag{S36}$$

(where $\overline{X}$ stands for the average of X including null elements). We can see that $\langle K \rangle$ does not appear here: as part of the parameterization discussed in Sec. III.1, we can always divide abundances by a constant factor to ensure $\langle K \rangle = 1$.

As will become clear in the analysis through the next section, these four parameters emerge naturally if we assume that interactions are disordered ("effectively random" even if they have some structure). Then, the abundance of each species at equilibrium is the result of two distinct random factors: its own carrying capacity and the change in abundance resulting from all interactions.

The result of many direct and indirect interactions is an emergent collective property, which loses sensitivity to the detail of the interactions (as a consequence of the Central Limit Theorem, see Sec. III.3). If the α_{ij} are not widely distributed or strongly correlated in complex ways, then μ, σ and γ suffice to predict their effects.

By contrast, ζ is less universal: this parameter emerges under the assumption that the distribution of carrying capacities K_i is Gaussian, which is not a requirement of our method (the same calculations can be performed for any other choice of distribution, see e.g. Sec. III.5). Still, unless that distribution has fat tails or complex correlations, we have only observed so far that using the full distribution rather than just ζ yields at most small quantitative, rather than qualitative, corrections.

Less disordered models will require more parameters to adequately predict community properties – a natural choice of additional parameter, which is often useful in practice, is the correlation between

carrying capacities and interactions $\langle K_i \alpha_{ij} \rangle$ when it cannot be removed by a choice of parameterization. If these additional parameters remain sufficiently generic (common to many models), we can build classes of "mostly disordered" models that represent usual deviations from the fully disordered limit, such as those represented in Fig. S1 and discussed in Sec. IV.

This model reduction has two main goals. Empirically, if we can find a simpler null model providing the same community-level predictions, this reveals limits to our ability to infer mechanisms from observed patterns. Theoretically, the fully disordered limit carries a stronger message: it means that in a given model, while species might all have different traits, they are in a sense all "sampled from the same distribution". The analysis of the reference model will now allow us to give a precise meaning to this statement.

III.3 Analytical solution

III.3.1 Intuitions

The analysis of the reference model is based on a well-honed technique from statistical physics, the cavity method [6, 24]. Since species abundances N_i are the only variables in the assembly process, all macroscopic properties of the assembled community can be determined from computing the equilibrium distribution of abundances and their correlations to other properties, see Sec. III.4.

The logic of the cavity method is remarkably simple: 1) we compute what happens after a single invasion (from a species randomly drawn from the pool) into an already large community, and 2) we then require that the community properties before and after the invasion be equal. The second part means that we are computing the fixed point of this invasion process, i.e. an equilibrium which cannot be invaded nor destabilized. It is thus different from the single-step invasion often studied in adaptive dynamics (testing whether a given mutant species can invade).

If the system is fully disordered, it is sufficient to single out one invading species to obtain the equations that define the whole community. More precisely, the hypothesis of disorder states that, for any species, its traits (e.g. the strength of its interactions) are sampled without bias from the pool, and the abundances of its interaction partners are sampled without bias from the equilibrium distribution of the entire community. This is a strong, but very successful, hypothesis which conveys an important intuition: no matter how complex the community, if each species samples that complexity fairly, the collective dynamics are in fact surprisingly simple.

In a less disordered system, where different neighborhoods can be distinguished within the community (e.g. functional groups or different positions in a hierarchy), we must instead perform the same computation for an invading species in each possible neighborhood. This idea is explained in Sec. IV.

III.3.2 Calculations

Given the parameters above, we explain in [3] the detailed process of deriving the analytical solution. In short, if we consider a community with the equilibrium abundances N_j^*, then add species 0, its own abundance obeys

$$N_0 = K_0 - \sum_j \alpha_{0j} N_j \tag{S37}$$

where the abundances of other species have now become N_j. Assuming that this change is relatively small, i.e. no species is largely controlled by a single interaction partner, we can take a linear approximation

$$N_j \approx N_j^* + \frac{dN_j}{dN_0} N_0. \tag{S38}$$

We notice from (S37) above (which is also obeyed by species j instead of 0) that adding an interaction partner 0 to the sum is equivalent to changing the carrying capacity K_j by an amount $-\alpha_{j0} N_0$. In

other words

$$\frac{dN_j}{dN_0} = -\alpha_{j0}\frac{dN_j^*}{dK_j}. \tag{S39}$$

This decomposition is helpful since it distinguishes between the effect of the specific interaction α_{j0}, and the general response of species j to any change, which is encapsulated by dN_j^*/dK_j (see [3] for more explanations). We make a further simplifying approximation:

$$\frac{dN_j}{dN_0} \approx -\alpha_{j0}\, v, \qquad v = \left\langle \frac{dN_j^*}{dK_j} \right\rangle. \tag{S40}$$

This coefficient v represents how responsive, on average, the equilibrium abundance of surviving species is to changes in either K_j (defined by the environment) or community composition. It integrates the response not only to the initial perturbation, but also to all subsequent feedbacks as the perturbation propagates through the community[5]. We later compute v explicitly, but for now we keep it as an additional unknown[6]. The fact that we can replace dN_j^*/dK_j by its average value over all species can be justified formally [7], but intuitively conveys the idea that this long-term feedback from the entire community is the same for every species, due to each species occupying a statistically similar position in the community, under the assumption of disorder.

Then we can rewrite (S37) as

$$N_0 = K_0 - \sum_j \alpha_{0j} N_j^* + vN_0 \sum_j \alpha_{0j}\alpha_{j0} \tag{S41}$$

From the definitions in (S36), we notice that the sum on $\alpha_{0j}\alpha_{j0}$ involves γ the reciprocity parameter. Identifying the sample mean with the population mean, and noticing that we are only summing over the $S^* = \phi S$ surviving species in the community, we find

$$\sum_{j=1}^{S^*=\phi S} \alpha_{0j}\alpha_{j0} \approx S^*\overline{\alpha_{ij}\alpha_{ji}} = \phi\gamma\sigma^2 + O(\mu/S) \tag{S42}$$

(where the remainder is small for large communities with distributed interactions, i.e. $\mu \ll S$) and we finally get

$$N_0 = \frac{1}{(1-\phi v\gamma\sigma^2)}\left(K_0 - \sum_j \alpha_{0j}N_j^*\right) \tag{S43}$$

Here, N_0 is a random variable, defined in terms of random variables K_0 and α_{0j} drawn from the species pool distributions, and N_j^* drawn from $P(N)$ the community's equilibrium abundance distribution. In fact, since α_{0j} and N_j^* are independent (N_j^* was the abundance *before* species 0 was added), then the following

$$z_0 = \sum_j^{S^*} \alpha_{0j}N_j^* \tag{S44}$$

is a sum of i.i.d. random variables, and by the Central Limit Theorem it follows an easily computed Gaussian distribution (see [3] for details). Thus, $N_0 = (K_0 - z_0)/(1-\phi v\gamma\sigma^2)$ is fully determined by two independent random variables, K_0 and z_0, for which we can compute

$$\begin{aligned} &\langle K_0\rangle = 1, \qquad \langle z_0\rangle = \phi\mu\langle N\rangle, \qquad \langle K_0 z_0\rangle = \langle K_0\rangle\langle z_0\rangle, \\ &\mathrm{Var}(K_0) = \zeta^2, \qquad \mathrm{Var}(z_0) = \sigma^2\phi\left\langle N^2\right\rangle. \end{aligned} \tag{S45}$$

[5]Hence the total derivative of the *equilibrium* abundance dN_j^*/dK_j meaning that we compute the total change between the equilibria before and after the change in K_j.

[6]Taking "response coefficients" as additional unknowns to help solve the problem is a well-established method in statistical physics, see e.g. [24, 25]. It allows to express a complicated equation on some variables as a simpler pair of coupled equations on these variables and the associated response coefficients.

From this we easily deduce

$$\langle N_0 \rangle = \frac{1 - \phi\mu \langle N \rangle}{1 - \phi v \gamma \sigma^2}, \tag{S46}$$

and

$$\mathrm{Var}(N_0) = \frac{\zeta^2 + \sigma^2 \phi \left\langle N^2 \right\rangle}{\left(1 - \phi v \gamma \sigma^2\right)^2} \tag{S47}$$

We will see that the probability distribution $P_0(N_0)$ of the random variable N_0 is well approximated by a Gaussian[7] with these prescribed mean and variance, however the rest of the calculation would work even if the distribution had a different functional form [3].

If $N_0 < 0$, species 0 could not invade durably. If $N_0 > 0$ however, it could invade[8] and reach abundance N_0. Now, recall that we want to compute the fixed point of the invasion process, when adding a species does not change the abundance distribution. This entails that the distribution $P(N)$ of abundances for (surviving) interaction partners of species 0 is the same as the distribution for species 0 itself *when it survives*, i.e.

$$P(N) = P_0(N) \qquad \text{if } N > 0. \tag{S48}$$

However, since this equality holds only for $N > 0$, moments such as $\langle N \rangle$ are computed only on positive N. Since $N_0 < 0$ indicates extinction, the probability of survival ϕ is simply the integral over $N_0 > 0$. All other moments of $P(N)$ are similarly obtained by integrating $P_0(N_0)$ over survivors:

$$\phi = \int_0^\infty dN_0 \; P_0(N_0), \tag{S49}$$

$$\langle N \rangle = \frac{1}{\phi} \int_0^\infty dN_0 \; P_0(N_0) \; N_0, \tag{S50}$$

$$\left\langle N^2 \right\rangle = \frac{1}{\phi} \int_0^\infty dN_0 \; P_0(N_0) \; N_0^2. \tag{S51}$$

Finally, from (S43) we can compute

$$v = \left\langle \frac{dN_j^*}{dK_j} \right\rangle = \frac{dN_0}{dK_0} = \frac{1}{1 - \phi \sigma^2 \gamma v} \tag{S52}$$

Hence, the equations above can be solved for the four parameters ϕ, $\langle N \rangle$, $\left\langle N^2 \right\rangle$ and v, which appear both on the left-hand side of the equations, and within $P_0(N_0)$ through its mean (S46) and variance (S46).

These equations are transcendental (if we compute the integral, the result involves exponentials and error functions) and therefore the solution cannot be expressed as an explicit formula. However, we explain in Appendix how to find the solution numerically.

III.3.3 Abundance distribution

The theoretical abundance distribution $P(N)$ computed above is Gaussian, which seems to contradict empirical evidence on lognormal Species Abundance Distributions. However, let us recall that in Sec. III.1, we defined N_i as *rescaled* abundances, and in particular, we noted that a natural choice for competitive systems is

$$N_i = \frac{B_i}{k_i} \tag{S53}$$

[7] This is exact if $P(K)$ is itself Gaussian or narrower, and approximate if it has wider tails.

[8] Note an important distinction: traditionally, invasion success is calculated by looking only at the initial growth following the immigration of a species, without considering the ensuing trajectory; here, on the contrary, we check whether species 0 will remain at very long times in the final equilibrium, after all interactions have played out.

with k_i the carrying capacity of species i and B_i its real abundance. If for instance k_i follows a lognormal distribution, as seen in plant monocultures [27], then a Gaussian distribution of N_i means that B_i will also have a fat-tailed, quasi-lognormal distribution. A general explanation for such a distribution of k_i would be a series of independent multiplicative factors such as resource fluxes, conversion efficiencies and reaction rates, whose product would then be lognormally distributed. For the same reason, having k_i follow a multiplicative random walk leads to realistic species abundance distributions [21]. Another important factor is immigration processes in open systems [29].

III.4 Community properties

The analytical solution has provided us with moments of the effective abundance distribution $P(N)$ such as ϕ, $\langle N \rangle$ and $\langle N^2 \rangle$ or σ_N^2 and we can also compute correlations such as $\langle N_i K_i \rangle_i$ or $\langle N_i \alpha_{ij} \rangle_{ij}$ (see [6]). To get back to real community properties, we must first revert the rescaling we performed earlier (see Sec. III.1) from the real biomasses B_i to these effective abundances N_i. In the simplest situation, $B_i = N_i$, $r_i = K_i$ and all community properties of interest are directly defined from the moments and correlations of N_i, K_i and α_{ij} listed above. This was the case for all our simulation models. If however we have a more complex parameterization, there are additional, often straightforward, calculations to be done to come back to the original parameters.

We now detail the predicted community properties, shown in Fig. S2. First, ϕ the fraction of survivors is directly one of the variables we solved for (and does not depend on the choice of parameterization). Then, we get total biomass

$$T = \sum_i B_i = \phi S \langle B \rangle , \tag{S54}$$

total productivity

$$P = \sum_i r_i B_i = \phi S \langle rB \rangle , \tag{S55}$$

(in the figures, we instead display the ratio P/T since P generically grows with T), and Simpson index

$$D^{-1} = I = \sum_i \left(\frac{B_i}{T} \right)^2 = \frac{1 + \sigma_B^2 / \langle B \rangle^2}{S}. \tag{S56}$$

Finally, for stability metrics, variability in response to environmental noise is commonly used. We show in [3] that it is given by

$$V_{\text{env}} = \sum_i \frac{\text{Var}(B_i(t))}{S} = \left\langle \frac{K}{r} B \right\rangle V \tag{S57}$$

where V is given by

$$V \approx \frac{1}{2 - \sigma^2 (\gamma + 1) \frac{\phi S}{S - \mu}}. \tag{S58}$$

For all comparisons we therefore used V, which is a more fundamental quantity: as we explain in [3], it is the common term found when we compute the variability in response to different types of perturbations; in particular, it directly yields variability due to *demographic* noise.

Finally, we also show in Fig. S2 the phase parameter for multistability. Assembly dynamics have two main regimes: one where there is a single global attractor reached asymptotically by any sequence of invasions, and another where multiple attractors exist and the result is history-dependent. The transition from one regime to another as σ increases is signalled by the divergence of the order parameter at a critical value σ_c (see [3, 7] for details and discussion) given by

$$\sigma_c^2 = \frac{1}{\phi (1 + \gamma)^2}. \tag{S59}$$

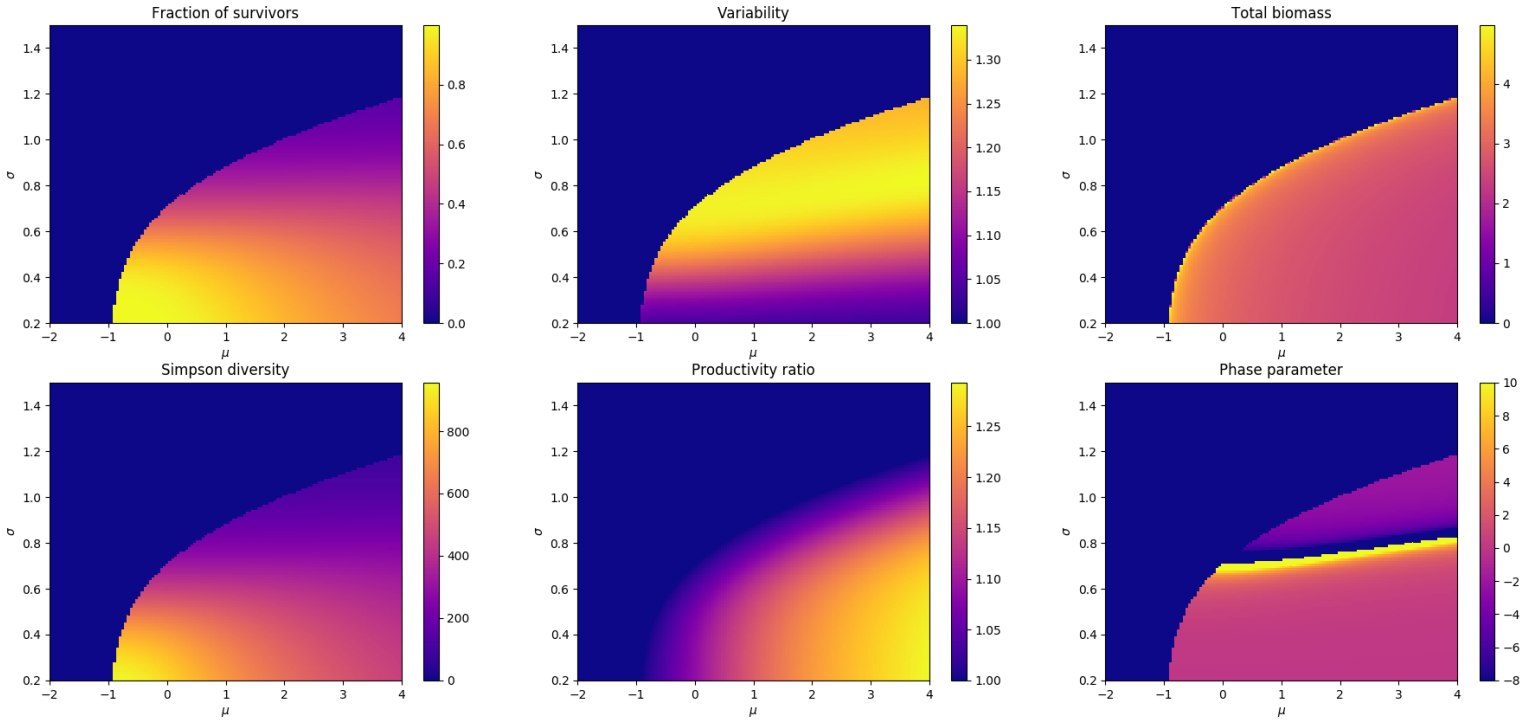

Figure S2: Coexistence, abundance and stability properties for the assembled state of the random Lotka-Volterra model, in the space (μ, σ) for $\zeta = 0.3$ and $\gamma = 1$. The uniform area in the left of each graph signals the parameter region where some interactions are positive and strong enough that abundances diverge. The bottom-right graph showcases the phase parameter: the sharp line where it diverges indicates the transition from the single-equilibrium regime (below the line) where our analytical results are exact for the reference model, to the multistable regime (above the line) where they are approximate.

Our analytical results are exact in the single-equilibrium regime, but even when multiple equilibria arise, our formulas continue to approximately predict the typical features of a community found in one of these equilibria [7]. Likewise, the unique equilibrium, when it exists, is reached in the absence of any noise and at very long times, but community properties do not change significantly when these strict conditions are relaxed, as indeed they are in simulations.

We combined these quantities into a metric of relative error between simulations and reference:

$$\text{Error} = \frac{1}{5} \sum_{x \in \{T,P,\phi,D,V\}} \left(\frac{1}{2} + \frac{1}{2} \left| \frac{x_{\text{simulated}} + x_{\text{predicted}}}{x_{\text{simulated}} - x_{\text{predicted}}} \right| \right)^{-1} \tag{S60}$$

which is symmetrical between under- and over-estimation, has values in $[0, 1]$ and coincides with the straightforward definition of relative error $|\Delta x/x|$ when it is small. We deemed satisfactory agreement when this error was below 5% for all explored parameter values.

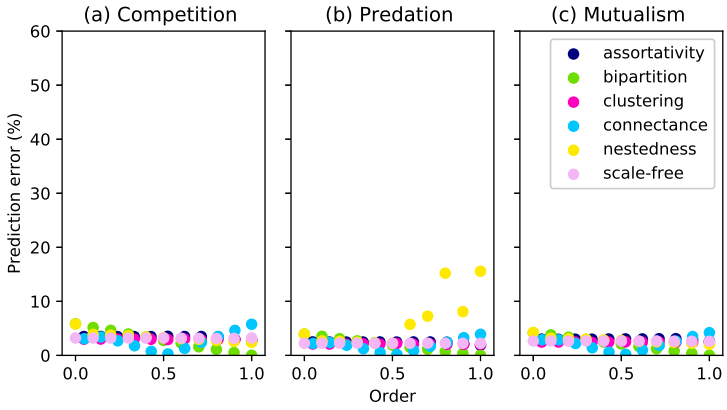

Figure S3: Agreement between simulations and the reference model with a saturating functional response. (a)-(c) For each of the three main interaction types, the relative error (y-axis, between 0 and 60%) of the reference model (with saturating functional response) against simulations, as a function of the degree of structure in the community (x-axis). Each set of symbols indicates a different network structural property: assortativity, partitioning (from complete to bipartite graph), clustering, connectance, nestedness, cascade and scale-free structure. See Sec. II.1 for a description of these structures and their respective control parameters (x-axis), and Fig. 4 in main text for the same comparison with a linear functional response.

III.5 Functional response

The model can be adapted to allow any functional response (and in doing so, we also write out equations allowing for arbitrary distributions of carrying capacities K_i). We will focus on the saturating case:

$$\frac{d}{dt}N_i = \frac{r_i}{K_i}N_i\left(K_i - N_i - \sum_j \frac{\alpha_{ij}N_j}{1+\frac{1}{AN_c}\left|\sum_k \alpha_{ik}N_k\right|}\right) \tag{S61}$$

where we clarify the meaning of the usual saturation half-rate [14] by decomposing it into the average coefficient $A = \langle\alpha_{ij}\rangle$ and the population threshold N_c. Hence, the interaction term saturates when the total population of all partners of a species exceeds N_c. The limit $N_c \to \infty$ recovers the Lotka-Volterra model.

In the context of the cavity method, we isolate species 0 and obtain its equilibrium condition

$$0 = K_0 - N_0 - f\left(\sum_j \alpha_{0j}N_j\right) \tag{S62}$$

where

$$f(z) = \frac{z}{1+\frac{|z|}{AN_c}}. \tag{S63}$$

The same reasoning as above leads us to compute $f(\sum_j \alpha_{0j}N_j)$ at linear order around its value in the absence of species 0:

$$f(\sum_j \alpha_{0j}N_j) \approx f(\sum_j \alpha_{0j}N_j^*) + N_0\frac{d}{dN_0}f(\sum_j \alpha_{0j}N_j). \tag{S64}$$

Let us note

$$z_0 = \sum_k \alpha_{0j} N_j^*. \tag{S65}$$

We have

$$f\left(\sum_j \alpha_{0j} N_j\right) \approx f(z_0) + N_0 f'(z_0) \frac{dz_0}{dN_0} \tag{S66}$$

where, as above,

$$\frac{dz_0}{dN_0} \approx v \sum_j \alpha_{0j}\alpha_{j0}, \qquad \sum_j \alpha_{0j}\alpha_{j0} = \phi\sigma^2\gamma \tag{S67}$$

and finally we find

$$N_0\left(1 - f'(z_0)\phi\sigma^2\gamma v\right) = K_0 - f(z_0). \tag{S68}$$

We can explicitly express N_0 as a function of z_0 and K_0

$$N_0(z_0, K_0) \approx \frac{K_0 - f(z_0)}{1 - f'(z_0)\phi\sigma^2\gamma v} \tag{S69}$$

where (in the case explored here)

$$f(z_0) = \frac{z_0}{1 + |z_0|/AN_c}, \qquad f'(z_0) = \frac{1}{(1 + |z_0|/AN_c)^2}. \tag{S70}$$

As we could expect, $N_0 \to K_0 \pm AN_c$ when $z_0 \to \mp\infty$. Following our previous arguments, z_0 is a Gaussian random variable, with

$$\langle z_0 \rangle = \phi\mu \langle N \rangle, \qquad \mathrm{Var}(z_0) = \phi\sigma^2 \left\langle N^2 \right\rangle, \tag{S71}$$

which we denote

$$P(z_0) = \mathcal{N}(z_0; \mu \langle N \rangle, \sigma^2 \left\langle N^2 \right\rangle). \tag{S72}$$

From there, we can easily solve numerically for the following equations, where N_0 can be replaced by its expression $N_0(z_0, K_0)$ above:

$$\phi = \int_{-\infty}^{\infty} dz_0 dK_0\ \Theta(N_0)\, P(z_0)P(K_0) \tag{S73}$$

$$\langle N \rangle = \frac{1}{\phi}\int_{-\infty}^{\infty} dz_0 dK_0\ \Theta(N_0)\, N_0\, P(z_0)P(K_0) \tag{S74}$$

$$\left\langle N^2 \right\rangle = \frac{1}{\phi}\int_{-\infty}^{\infty} dz_0 dK_0\ \Theta(N_0)\, N_0^2\, P(z_0)P(K_0) \tag{S75}$$

where $\Theta(y)$ stands for the Heaviside step function (1 if $y > 0$, 0 otherwise). Finally, from (S69),

$$v = \left\langle \frac{dN_0}{dK_0} \right\rangle = \frac{1}{\phi}\int dz_0\, \Theta(N_0) \frac{P(z_0)}{1 - f'(z_0)\phi\sigma^2\gamma v} \tag{S76}$$

All right-hand sides are most conveniently rewritten under the general form

$$\frac{1}{\phi}\int_{-\infty}^{\infty} dz_0\ P(z_0)\left(1 - f'(z_0)\phi\sigma^2\gamma v\right)^{-k} \int_0^{\infty} dK_0 K_0^j P(K_0 + f(z_0)) \tag{S77}$$

for $k = 0, 1, 2$ and $j = k$ for the first three equations, $k = 1$ and $j = 0$ for the equation on v.

As we show in Fig. S3, the saturating functional response makes communities *more* predictable by the reference model, since species whose interactions are saturated become indifferent to the state of others and the general community structure and dynamics.

This method can straightforwardly be extended to different functional responses $f(z)$ (all equations where we have not made $f(z_0)$ and $f'(z_0)$ explicit remain valid), and to the case of a species-dependent threshold N_{ci}. Important other variations on Lotka-Volterra dynamics in the literature include: energy reserves and other time-lags between consumption and reproduction (e.g. [10, 17]), and Allee effect or similarly nonlinear density dependence around zero [9]. The former have an effect on dynamics but not on the equilibrium condition studied here, and are therefore irrelevant for our analysis. The latter however can have dramatic consequences on the equilibria, which we will discuss in future work.

IV Reference model extensions

Let us come back to the intuition that "full disorder" means that each species should interact with a statistically unbiased (if not identical) sample of the whole community – i.e. for any species, its interactions and the abundances of its partners can, in effect, be drawn uniformly from the equilibrium distribution of the general community. Then, we can deduce what it means to be less than fully disordered: species can be differentiated statistically by the subset of the community that they interact with directly.

In the context of the cavity method, this means that, when considering an invader, we cannot draw the traits and abundance of its interaction partners at random from the entire equilibrium community; we must consider each possible invasion scenario for the different "neighborhoods" within the community (see examples below). This requires more information as we need the parameters that characterize these neighborhoods in the species pool, from which we can analytically determine the abundances and other community properties within each neighborhood. However, if we have this information, we can apply an extended cavity method and successfully predict simulation results, see the following sections for a few such extensions and Fig. S4 for an example.

In Fig. S1 we show two types of structures that create such neighborhoods in the species pool that persist into the assembled community. Functional groups are a way to summarize strongly clustered "profiles" of species, while hierarchy implies some sort of ordering or ranking between species. While these two types have been extensively studied in simple models, our method works even if these structures (e.g. group boundaries) are blurred rather than exact, and hence, they could be inferred from data in complex communities. However we do not tackle this inference problem here yet, and simply assume that the position of each species within a group or hierarchy in the species pool is known in advance when making analytical predictions.

These two types add structure in the form of correlations in the matrix of species interactions, and possibly also correlations between their interactions and their carrying capacity (e.g. a tradeoff between competitive ability and growth). By contrast, other properties such as a degree distribution simply adds more "local" heterogeneity, i.e. heterogeneity at the level of individual species attributes which does not create distinct neighborhoods[9]. Local heterogeneity should not invalidate the fully disordered approach, and indeed is found to have little to no impact on community properties.

IV.1 Group structure

The equations for the reference model can be extended to any structure comprised of discrete groups, with disordered interactions within and between groups, but different statistics for each set of interactions. Coming back to the equilibrium equation, we can write for species i in group x (which contains

[9]Note that network generated from ensembles such as Barabasi-Albert also come with less local properties such as degree correlations, which could have an impact. We do observe however that they do not, see main text and Fig. S4.

S^x species) as

$$0 = K_i^x - N_i^x - \sum_{j \in G_y} \alpha_{ij}^{xy} N_j^y \tag{S78}$$

where G_y is the set of species in group y. Thus, we now have vector $\boldsymbol{\zeta}$ and matrices $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\gamma}$, defined by

$$\alpha_{ij}^{xy} = \frac{\mu^{xy}}{S} + \frac{\sigma^{xy}}{\sqrt{S}} a_{ij}^{xy} \tag{S79}$$

with

$$\langle a \rangle = 0, \qquad \left\langle a^2 \right\rangle = 1, \qquad \left\langle a_{ij}^{xy} a_{ji}^{yx} \right\rangle = \gamma^{xy}. \tag{S80}$$

The equations to solve are the same as in Sec. III.3, except there are now four equations per group, all coupled: for each group we solve

$$\phi^x = \int_0^\infty dN_0 \; P_0^x(N_0) \tag{S81}$$

$$\langle N \rangle^x = \frac{1}{\phi^x} \int_0^\infty dN_0 \; P_0^x(N_0) \; N_0 \tag{S82}$$

$$\left\langle N^2 \right\rangle^x = \frac{1}{\phi^x} \int_0^\infty dN_0 \; P_0^x(N_0) \; N_0^2 \tag{S83}$$

$$v^x = \frac{1}{u^x} \tag{S84}$$

where $P_0^x(N_0)$ is again a Gaussian distribution for each group x, with the following mean and variance

$$\langle N_0 \rangle^x = \frac{1 - \sum_y \mu^{xy} \varphi^y \langle N \rangle^y}{u^x}, \tag{S85}$$

$$\mathrm{Var}(N_0)^x = \frac{\zeta^2 + \sum_y (\sigma^{xy})^2 \varphi^y \left\langle N^2 \right\rangle^y}{(u^x)^2}, \tag{S86}$$

$$u^x = 1 - \sum_y \varphi^y v^y \gamma^{xy} \sigma^{xy} \sigma^{yx}. \tag{S87}$$

where

$$\varphi^x = \phi^x \frac{S^x}{S}. \tag{S88}$$

IV.2 Hierarchy

We now assume we can characterize a species by its position x on a single axis of possible niches or roles. We show how to solve this in greater detail in [3]. We can take the continuous limit of the group approach, where instead of matrices, $\mu(x, y)$, $\sigma(x, y)$ and $\gamma(x, y)$ are functions.

Note that a strictly nested matrix would have $\mu(x, y) = 0$ for $y > x$, i.e. it can be written using $\Theta(x - y)$ the Heaviside step function. But we find in Fig. S4 that accurate predictions are obtained by replacing the step function by a simple linear dependence:

$$\mu(x, y) = \mu + c_\mu (x - y) \tag{S89}$$

$$\sigma^2(x, y) = \sigma^2 + c_\sigma (x - y) \tag{S90}$$

$$\tag{S91}$$

and fixed γ. When deriving them from a simulation model, the coefficients and slopes can be fitted from the interaction matrix of the species pool, with all slopes zero in the fully disordered case.

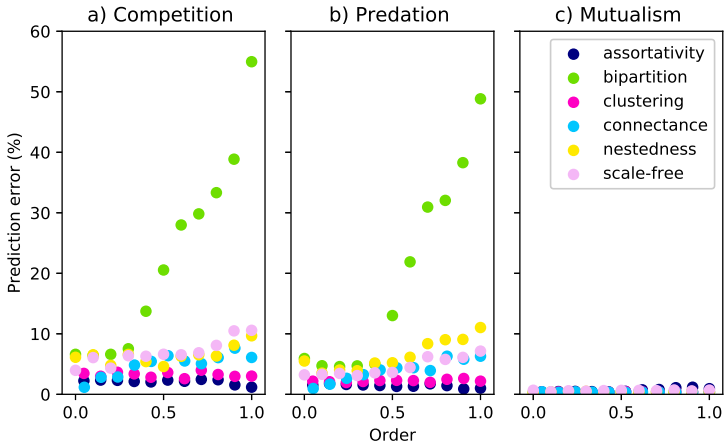

Figure S4: Agreement between simulations and the extended analysis for hierarchical structure. While a cascade structure precluded agreement with the reference model, it is accounted for by the developments in Sec. IV.2. Group structure (e.g. bipartition) however requires a different treatment, discussed in Sec. IV.1. See legend of Fig. S3 for details, and Fig. 4 in main text for comparison with the reference model.

If we add a similar dependence for carrying capacities,

$$\langle K \rangle (x) = K + c_K x \tag{S92}$$

$$\zeta^2(x) = \zeta^2 + c_\zeta x \tag{S93}$$

we can account for tradeoffs such as competition-colonization models: if $c_\mu < 0$ and $c_K > 0$, low rank x corresponds to high competitive ability but low colonization ability.

The solution proceeds similarly to the group approach:

$$\langle N_0 \rangle (x) = \frac{K + c_K x - (\mu + c_\mu x) \langle \varphi N \rangle + c_\mu \langle \varphi N x \rangle}{\tilde{u}(x)}, \tag{S94}$$

$$\mathrm{Var}(N_0)(x) = \frac{\zeta^2 + c_\zeta x + (\sigma^2 + c_\sigma x) \langle \varphi N^2 \rangle - c_\sigma \langle \varphi N^2 x \rangle}{\tilde{u}^2(x)}, \tag{S95}$$

$$\tilde{u}(x) \approx 1 - \gamma \langle \varphi V \rangle \sigma^2. \tag{S96}$$

(in $\tilde{u}$ we should in fact have $\sigma(x,y)\sigma(y,x) = \sigma^2\sqrt{1 - c_\sigma^2 (x-y)^2/\sigma^4}$ but this complicates the analysis, so we must assume that c_σ is small enough). Hence there are two sets of coupled unknowns: for $\Psi = N, N^2, V$ we must compute

$$\langle \varphi \Psi \rangle = \int dx \rho(x) \int_0^\infty dN_0 P_0(N_0, x) \Psi \tag{S97}$$

and (except for V),

$$\langle \varphi \Psi x \rangle = \int dx \rho(x) x \int_0^\infty dN_0 P_0(N_0, x) \Psi. \tag{S98}$$

where $\rho(x)$ represents the probability distribution of niche position x. Finally, we can compute the community averages

$$\langle \Psi \rangle = \frac{\langle \varphi \Psi \rangle}{\langle \varphi \rangle} \tag{S99}$$

which are involved in calculating community properties (see Sec. III.4).

IV.3 Basic correlation structure

Besides group structures and hierarchies, a very simple deviation from the most basic form of the reference model (described in Sec. III) is the presence of first-order correlations within rows in the interaction matrix or between fitness and interaction, with an ecological interpretation given below. Correlations within columns of the interaction matrix can be transformed into row correlations by rescaling the variables x_i differently, see Sec. III.1.

IV.3.1 Fitness-interaction correlations

Sec. III assumed K_0 and α_{0i} to be independent variables. If there is a correlation between them, then

$$\langle K_0 z_0 \rangle = \langle K_0 \rangle \langle z_0 \rangle + S\phi \langle N \rangle C_{K\alpha} \tag{S100}$$

where we now define

$$C_{K\alpha} = \langle K_0 \alpha_{0j} \rangle - \langle K \rangle \langle \alpha \rangle \tag{S101}$$

and as we recall $N_0 = (K_0 - z_0)/(1 - \phi v \gamma \sigma^2)$, it is straightforward to see that

$$\mathrm{Var}(N_0) = \frac{\zeta^2 + \sigma^2 \phi \langle N^2 \rangle - 2S\phi \langle N \rangle C_{K\alpha}}{(1 - \phi v \gamma \sigma^2)^2}. \tag{S102}$$

Hence, such correlations can be accounted for by a simple shift in the variance of $P_0(N_0)$ (and therefore a similar change in the variance of the effective abundances N), at the cost of adding a new parameter, $\langle K_0 \alpha_{0i} \rangle$.

IV.3.2 Row correlation

Let us assume that

$$\alpha_{0i} = \alpha_0 + \frac{\sigma_0}{\sqrt{S}} \beta_{0i} \tag{S103}$$

with

$$\langle \beta_{ij} \rangle = 0, \qquad \mathrm{Var}(\beta_{ij}) = 1, \qquad \langle \alpha_0 \rangle = \mu/S, \qquad \mathrm{Var}(\alpha_0) = \sigma_{\mathrm{row}}^2. \tag{S104}$$

For instance, this allows to capture very skewed degree distributions where a few species serve as hubs and have much stronger average interactions than the others. Then, the equilibrium condition is

$$0 = K_0 - S^* \alpha_0 \langle N \rangle - N_0 - \frac{\sigma_0}{\sqrt{S}} \sum_i^{S^*} \beta_{0i} N_i \tag{S105}$$

where we still assume that $\langle N \rangle$ over partners of 0 does not differ from the community average. Then,

$$N_i \approx N_i^* - \frac{dN_i}{dK_i} \frac{\sigma_i}{\sqrt{S}} \beta_{i0} N_0 \tag{S106}$$

hence

$$\left(1 - \frac{1}{S} \sum_i \sigma_i \sigma_0 \beta_{i0} \beta_{0i} \frac{dN_i}{dK_i}\right) N_0 = K_0 - S^* \alpha_0 \langle N \rangle - \frac{\sigma_0}{\sqrt{S}} \sum_i^{S^*} \beta_{0i} N_i^* \tag{S107}$$

$$\approx (1 - \phi \sigma_0 \sigma \gamma v) N_0 \tag{S108}$$

where

$$v = \left\langle \frac{1}{1 - \phi \sigma_0 \sigma \gamma v} \right\rangle \approx \frac{1}{1 - \phi \sigma^2 \gamma v} \tag{S109}$$

and finally

$$\langle N_0 \rangle = v(\langle K \rangle - \phi\mu \langle N \rangle) \tag{S110}$$

$$\mathrm{Var}(N_0) = v^2 \left(\zeta^2 + S^{*2}\sigma_{\mathrm{row}}^2 \langle N \rangle^2 - 2S^* \langle N \rangle C_{K\alpha} + \sigma^2 \phi \left\langle N^2 \right\rangle \right) \tag{S111}$$

where it turns out that the variance on σ_0 is irrelevant. So in the end we only need two metrics of variance: the usual σ and σ_{row} the standard deviation of $\langle \alpha_{ij} \rangle_i$. Hence row (or column) correlations in the interaction matrix add one more parameter, σ_{row}, compared to fitness-interaction correlations.

References

[1] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[2] Jean-François Arnoldi, Michel Loreau, and Bart Haegeman. Resilience, reactivity and variability: A mathematical comparison of ecological stability measures. *Journal of theoretical biology*, 389:47–59, 2016.

[3] Matthieu Barbier and Jean-François Arnoldi. The cavity method for community ecology. bioRxiv preprint 147728, 2017.

[4] Uta Berger, Cyril Piou, Katja Schiffers, and Volker Grimm. Competition among plants: concepts, individual-based modelling approaches, and a proposal for a future research strategy. *Perspectives in Plant Ecology, Evolution and Systematics*, 9(3):121–135, 2008.

[5] Ulrich Brose, Richard J Williams, and Neo D Martinez. Allometric scaling enhances stability in complex food webs. *Ecology letters*, 9(11):1228–1236, 2006.

[6] Guy Bunin. Interaction patterns and diversity in assembled ecological communities. *arXiv preprint arXiv:1607.04734*, 2016.

[7] Guy Bunin. Ecological communities with lotka-volterra dynamics. *Physical Review E*, 95(4):042414, 2017.

[8] JE Cohen and CM Newman. A stochastic theory of community food webs: I. models and aggregated data. *Proceedings of the Royal Society of London B: Biological Sciences*, 224:421–448, 1985.

[9] Franck Courchamp, Tim Clutton-Brock, and Bryan Grenfell. Inverse density dependence and the Allee effect. *Trends in ecology & evolution*, 14(10):405–410, 1999.

[10] MR Droop. Some thoughts on nutrient limitation in algae. *Journal of Phycology*, 9(3):264–272, 1973.

[11] Paul Erdos and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[12] M Gatto. A general minimum principle for competing populations: Some ecological and evolutionary consequences. *Theoretical Population Biology*, 37(3):369–388, 1990.

[13] J Nathaniel Holland and Donald L DeAngelis. A consumer–resource approach to the density-dependent population dynamics of mutualism. *Ecology*, 91(5):1286–1295, 2010.

[14] J Nathaniel Holland, Donald L DeAngelis, and Judith L Bronstein. Population dynamics and mutualism: functional responses of benefits and costs. *The American Naturalist*, 159(3):231–244, 2002.

[15] Crawford Stanley Holling. The functional response of predators to prey density and its role in mimicry and population regulation. *Memoirs of the Entomological Society of Canada*, 97(S45):5–60, 1965.

[16] Stephen P Hubbell. *The unified neutral theory of species abundance and diversity*. Princeton University Press, Princeton, 2001.

[17] BW Kooi, MP Boer, and SALM Kooijman. Consequences of population models for the dynamics of food chains. *Mathematical biosciences*, 153(2):99–124, 1998.

[18] Richard Law and Ulf Dieckmann. Moment approximations of individual-based models. *IIASA Interim Report*, 1999.

[19] Richard Law and R Daniel Morton. Permanence and the assembly of ecological communities. *Ecology*, 77(3):762–775, 1996.

[20] Robert Mac Arthur. Species packing, and what competition minimizes. *Proceedings of the National Academy of Sciences*, 64(4):1369–1371, 1969.

[21] Ofer Malcai, Ofer Biham, Peter Richmond, and Sorin Solomon. Theoretical analysis and simulations of the generalized lotka-volterra model. *Physical Review E*, 66(3):031102, 2002.

[22] Robert M May. Will a large complex system be stable? *Nature*, 238:413–414, 1972.

[23] Claire Mazancourt, Forest Isbell, Allen Larocque, Frank Berendse, Enrica Luca, James B Grace, Bart Haegeman, H Wayne Polley, Christiane Roscher, Bernhard Schmid, et al. Predicting ecosystem stability from community composition and biodiversity. *Ecology letters*, 16(5):617–625, 2013.

[24] Marc Mézard, Giorgio Parisi, and Miguel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Co Inc, 1987.

[25] Manfred Opper and Sigurd Diederich. Phase transition and 1/f noise in a game dynamical model. *Physical review letters*, 69(10):1616, 1992.

[26] Stephen W Pacala and JA Silander Jr. Neighborhood models of plant population dynamics. i. single-species models of annuals. *The American Naturalist*, 125(3):385–411, 1985.

[27] Bernhard Schmid, Andy Hector, Prasenjit Saha, and Michel Loreau. Biodiversity effects and transgressive overyielding. *Journal of Plant Ecology*, 1(2):95–102, 2008.

[28] M Angeles Serrano and Marián Boguná. Tuning clustering in random networks with arbitrary degree distributions. *Physical Review E*, 72(3):036133, 2005.

[29] William G Wilson and Per Lundberg. Biodiversity and the lotka–volterra theory of species interactions: open systems and the distribution of logarithmic densities. *Proceedings of the Royal Society of London B: Biological Sciences*, 271(1551):1977–1984, 2004.

Appendix: Numerical solution of the reference model

We explain here the basic numerical scheme used to obtain results from the reference model, given the four parameters $\zeta, \mu, \sigma, \gamma$. See below for example R code.

The equations to solve are the following

$$\phi = \int_0^\infty dN_0 \; \mathcal{G}(N_0, \langle N_0 \rangle, \mathrm{Var}(N_0)), \tag{S112}$$

$$\langle N \rangle = \frac{1}{\phi} \int_0^\infty dN_0 \; \mathcal{G}(N_0, \langle N_0 \rangle, \mathrm{Var}(N_0)) \; N_0, \tag{S113}$$

$$\left\langle N^2 \right\rangle = \frac{1}{\phi} \int_0^\infty dN_0 \; \mathcal{G}(N_0, \langle N_0 \rangle, \mathrm{Var}(N_0)) \; N_0^2, \tag{S114}$$

$$v = \frac{1}{1 - \phi\sigma^2\gamma v} \tag{S115}$$

where

$$\mathcal{G}(x, m, v) = \frac{e^{-(x-m)^2/2v}}{\sqrt{2\pi v}} \tag{S116}$$

and

$$\langle N_0 \rangle = \frac{1 - \phi\mu \langle N \rangle}{1 - \phi v \gamma \sigma^2}, \tag{S117}$$

$$\mathrm{Var}(N_0) = \frac{\zeta^2 + \sigma^2 \phi \left\langle N^2 \right\rangle}{1 - \phi v \gamma \sigma^2}. \tag{S118}$$

Let us note

$$w_n(m, v) = \int_0^\infty dN_0 \; \mathcal{G}(x, m, v) x^n \tag{S119}$$

Then

$$w_0(m, v) = \frac{1}{2} \left(1 + \mathrm{erf}(m/\sqrt{2v}) \right) \tag{S120}$$

$$w_1(m, v) = v \frac{e^{-(x-m)^2/2v}}{\sqrt{2\pi v}} + \frac{m}{2} \left(1 + \mathrm{erf}(m/\sqrt{2v}) \right) \tag{S121}$$

$$w_2(m, v) = mv \frac{e^{-(x-m)^2/2v}}{\sqrt{2\pi v}} + \frac{m^2 + v}{2} \left(1 + \mathrm{erf}(m/\sqrt{2v}) \right) \tag{S122}$$

Finally, we use a root solver[10] with ϕ, $\langle N \rangle$, $\left\langle N^2 \right\rangle$ and v as variables, looking for a root of the vector-valued function

$$F(\phi, \langle N \rangle, \left\langle N^2 \right\rangle, v) = \begin{pmatrix} \phi - w_0 \\ \langle N \rangle - w_1 \\ \left\langle N^2 \right\rangle - w_2 \\ v(1 - \phi\sigma^2\gamma v) - 1 \end{pmatrix} \tag{S123}$$

where for each w_n we used $m = \langle N_0 \rangle$ and $v = \mathrm{Var}(N_0)$ which are both explicit functions of $(\phi, \langle N \rangle, \left\langle N^2 \right\rangle, v)$. The same technique generalizes to the extensions of the reference model, except for the fact that w_n cannot always be derived explicitly (and the equation for v may involve an integral), in which case the integrals have to be computed numerically.

R implementation of the comparison scheme

An up-to-date version of this code, plus example files from simulations, are maintained at `http://github.com/mrcbarbier/ecocavity-R`

[10] We used the *hybr* (modified Powell hybrid method) solver from the SciPy package.

```
library(rootSolve)

# This code is provided as part of the Supporting Information for the article
#     "Generic assembly patterns in large ecological communities" (2017)
#     M. Barbier, J.-F. Arnoldi, G. Bunin and M. Loreau.

# The function "prediction_from_matrix(r,A,D)" takes matrices for
# growth rates r, interactions A and self-interactions D and returns
# a list containing the four reference parameters
#    zeta
#    mu
#    sigma
#    gamma
# as well as all predicted quantities
#    biomass: total biomass of the system
#    sdBiomass: standard deviation of individual biomass
#    phi: fraction of survivors
#    survivors: number of survivors
#    simpsonD: Simpson diversity (1/Simpson index)
#    productivity: productivity ratio
#    variability: response to stochastic perturbation
#    press: response to random press perturbation (if infinite, signals multistability)
#
# OPTIONS
#    correlations
#       when set to TRUE (default), the function uses the extension of the reference model
#       that accounts for first-order correlations between the coefficients, and returns
#         cKa: measure of correlation between carrying capacities and interactions
#         sigma_row: measure of row-wise correlations in interactions
#    FR
#       If provided as a list (f,df) consisting of a function f(z) and its derivative f'(z)
#       computes the predictions for the model with functional response f(z)


measure_parameters <- function(r,A,D,PK=TRUE,correlations=TRUE,plotK=FALSE)
{
  K <-r/D
  alpha <- A / rep(D, each = nrow(A))
  S <- length(K)
  if( mean(alpha)>1){
    print("WARNING: Average interactions are strong (A/D>1). Predictions will not be accurate.")
  }
  else if (max(alpha)>1){
    if (max(alpha)>5){
      print("WARNING: Some very strong interactions (A/D>>1). Predictions may not be accurate.")
    }
    else {
      print("WARNING: Some strong interactions (A/D>1). Predictions may not be accurate.")
    }
  }
  avgK <- mean(K)
```

```
  stdK <- sd(K)
  histK <- hist(K,plot=FALSE)
  Kmin <- min(K)-5*stdK
  Kmax <- max(K)+5*stdK
  if(PK){
    PK <- function(x) {
      res <- exp(predict(smooth.spline(x=histK$mids, y=log(histK$density)),x)$y)
      return(res)
    }
  }
  else{
    PK <- function(x){dnorm(x,avgK,stdK)}
  }
  if(plotK){
    ks<-seq(Kmin-.5,Kmax+.5,0.01)
    plot(ks,PK(ks),type="l" )
    lines(ks,dnorm(ks,avgK,stdK),col="red")
    points(histK$mids,histK$density )
  }

  offdiag <- function(a){ a[row(a)!=col(a)] }
  offa <- offdiag(alpha)
  mu <-  (S-1)* mean( offa )
  sig <- sqrt(S-1)*sd(offa )
  gam <- cor(as.vector(offa),as.vector(offdiag(t(alpha))))

  if(correlations){
    corrKA <- (S-1)* mean(offdiag(rep(K, each = nrow(alpha))*(alpha)) ) -  avgK* mu
    sigrow <- sqrt(S-1)*sd(rowMeans(alpha) )
    sig <- sqrt(sig^2 - sigrow^2)
  }
  else
  { corrKA <-0
  sigrow <-0 }

  return( list(S=S,mu=mu,sig=sig,gam=gam, PK=list(avgK=avgK,stdK=stdK,dist=PK,Kmin=Kmin,Kmax=Kmax),co
}



prediction <- function(parameters,correlations=TRUE,FR=FALSE)
{

  PK <- parameters$PK
  if(!(identical(PK$dist,FALSE) & identical(FR,FALSE) )  ){
    #Make integral table for interpolation
    Kmin=PK$Kmin
    Kmax=PK$Kmax
    if(identical(FR,FALSE)){
      #Linear functional response by default
      FR <- list(f=function(x){x},df=function(x){1})
```

```
  }
  Kstep <- (Kmax-Kmin)/100
  KS <- seq(Kmin,Kmax,Kstep)
  Ktable <-  mapply(function(mom){
               mapply(function(z0){
                 integrate( function(x){ FR$f(x)^mom * PK$dist(x+z0) },0, Inf )$value
               }, KS ) }, c(0,1,2) )

  parameters$PK$table <-
    mapply(function(mom){function(K){
      K[K>Kmax]=Kmax
      K[K<Kmin]=Kmin
      approxfun(KS , Ktable[,mom] )(K) }
      }, c(1,2,3)
    )
}

z_parameters <- function(x,S,mu,sig,gam){
  meanz <- mu*x[1]*x[2]
  varz <- sig^2 * x[1] *x[3]
  return(c(meanz,varz) )
}
gaussian_parameters <- function(x,S,mu,sig,gam,PK,correl){
  g <-gam*sig^2 * x[1]
  v <- x[4]
  u <- 1-g*v
  zparam <- z_parameters(x,S,mu,sig,gam)
  mean0 <- PK$avgK - zparam[1]
  corrKA <- correl[1]
  sigrow <- correl[2]
  var0 <- PK$stdK^2 +sigrow^2 * (S-1)*(x[1] *x[2])^2 - 2*x[1]*x[2]*corrKA + zparam[2]
  return(c(mean0/u,var0/u^2,u))
}

erfmom <- function(mom,mean0,var0){
  #If functional response is linear and all distributions are normal
  erf <- function(x) 2 * pnorm(x * sqrt(2)) - 1
  if (var0<=0.001){var0<-0.001}
  xx<-mean0/sqrt(2*var0)
  mom0 <- .5 * (erf(xx)+1 )
  mom1 <- sqrt(var0/2/pi) * exp(-xx^2)
  if (mom==0){
    return(mom0) }
  else if (mom==1){
    return(mean0* mom0 + mom1) }
  else if (mom==2){
    return( (var0+mean0^2)*mom0 + mean0*mom1)}
}

frmom <- function(mom,meanz,varz,g,PK,momden=FALSE){
  if (identical(momden,FALSE) ){ momden <- mom }
```

```
  funcresp <- FR$f
  funcrespd <- FR$df

  Kint <- PK$table[[as.integer(mom+1)]]
  stdz=sqrt(varz)
  res <- integrate(function(z0){
    (1-funcrespd(z0) *g)^(-momden)*Kint(z0) *dnorm(z0,meanz,stdz)  },-Inf,Inf)$value
  return(res)
}

meanfield <-function(S,mu,PK){
  v <- PK$avgK/(1+mu*(1-1/S) )
  return(c(1,v,v^2+PK$stdK^2,1 ) )
}

solve_system <- function(S,mu,sig,gam,PK,correl,maxtrials=100){

  model <- function(x){
    if (identical(FR,FALSE) & identical(PK$dist,FALSE) ){
      gparam <- gaussian_parameters(x,S,mu,sig,gam,PK,correl)
      mean0 <- gparam[1]
      var0 <- gparam[2]
      F1 <- x[1] - erfmom(0,mean0,var0)
      F2 <- x[1] * x[2] - erfmom(1,mean0,var0)
      F3 <- x[1] * x[3] - erfmom(2,mean0,var0)
      F4 <- x[4]*(1-x[4]*gam*sig^2*x[1] ) -1
    }
    else{
      gparam <- gaussian_parameters(x,S,mu,sig,gam,PK,correl)
      mean0 <- gparam[1]
      var0 <- gparam[2]
      zparam <- z_parameters(x,S,mu,sig,gam)
      meanz <- zparam[1]
      varz <- max(0.0001,zparam[2])
      g <- x[1] * sig^2 *gam * x[4]
      g<- min(.9,g)
      F1 <- x[1] - frmom(0,meanz,varz,g,PK)
      F2 <- x[1] * x[2] - frmom(1,meanz,varz,g,PK)
      F3 <- x[1] * x[3] - frmom(2,meanz,varz,g,PK)
      F4 <- x[1] * x[4] - frmom(0,meanz,varz,g,PK,1)
    }
    phimin <-0.001
    cost <- max(0,(phimin-x[1])/phimin ) #max(0,min((0.01-x[1])*10,x[2]^2-x[3]) )
    val <- c(F1 = F1+cost,F2 = F2,F3 = F3, F4=F4)
    #print(c(x,val, cost ))
    return( val)
  }

  ss <- list(root= c( 1,1,1,1),f.root=(100)  )
  failure <- function(s){ max(abs(s$f.root))>0.01  }
```

```
  trials=0
  while (failure(ss) & trials<maxtrials ){
    if (sig>.8 & trials >0){
      x0 <- solve_system(S,mu,sig/1.01,gam,PK,correl,maxtrials=2)
    }
    else{
      if (sig>0.1){
        x0 <- solve_system(S,mu,0.1,0,PK,correl,maxtrials=20)
      }
      else{
        x0 <- meanfield(S,mu,PK)
      }
      }
    try( ss <- multiroot(f = model, start = x0, positive=TRUE) ,silent=FALSE)
    trials <- trials +1
  }
  if (failure(ss)){
    return( c(0,0,0,0) )
  }
  else{
    return(ss$root)
  }
}

compute_results <- function(x,S,mu,sig,gam,PK,correl){
  gparam <- gaussian_parameters(x,S,mu,sig,gam,PK,correl)
  u <- gparam[3]
  phi <- x[1]
  N1 <- x[2]
  N2 <- x[3]
  v <- x[4]
  avgN <- N1*phi
  avgN2 <- N2*phi
  stdN <- sqrt(N2-N1^2)
  simpson <- N2/S/phi/N1^2
  zeta <- PK$stdK/PK$avgK
  if(PK$stdK!=0){
    productivity <- (u*avgN2 + mu*avgN^2 +sig^2/PK$stdK^2
                     *avgN2 * PK$avgK *avgN)/ (1+sig^2/PK$stdK^2*avgN2)
  }
  else{
    productivity <- 1
  }
  press <- 1./( u^2 - phi*sig^2  )
  if(press <0){ press <- Inf }
  variability <- 1/(1- phi* sig^2 * (gam+1)/2 )
  results= c(zeta=zeta,mu=mu,sigma=sig,gamma=gam,
             biomass=avgN*S,sdBiomass=stdN,phi=phi,survivors = S*phi,
             simpsonD = 1/simpson, productivity = productivity/avgN,
             variability=variability,press=press)
```

```
    if(correlations){
      results=c(results,cKa=correl[1],sigma_row=correl[2] )
    }

    return ( results)

  }

  ss <- do.call(solve_system,parameters)
  return(do.call(compute_results, c(list(x=ss),parameters)))

}

prediction_from_matrix <- function(r,A,D=FALSE,PK=TRUE,correlations=TRUE,FR=FALSE){
  if ( identical(D,FALSE) ){
    D <- diag(A)
    diag(A) <- 0
  }
  return(prediction(measure_parameters(r,A,D,PK,correlations),correlations,FR=FR))
}
```