Competition and cooperation in microbial community coalescence

March 27, 2021

1 Abstract

xxx

2 Introduction

Microbial communities are widespread throughout our planet (**?**), from the deep ocean to the human gut, and play a critical role in natural processes ranging from animal development and host health (**??**) to biogeochemical cycles (**?**). These communities are very complex, often harbouring hundreds of species (**?**), making them hard to characterize. Recently, DNA sequencing has allowed a high-resolution mapping of these consortia, opening a niche for ambitious theorists and experimentalists to collaboratively disentangle the complexity of these systems (**????????**). Despite thorough explorations, the mechanisms responsible for the assembly of microbiome communities have only begun to be revealed.

Unlike in the macroscopic world, entire microbial communities often suffer displacements to different regions, where they encounter other communities. The process by which two or more communities that were previously separated join and reassemble into a new community has been termed community coalescence (**?**). This type of event happens repeatedly in microbiomes due to abiotic (wind, tides or river flow), biotic (animal courtship, parent-offspring interactions or leaves falling), and anthropogenic (industrial anaerobic digestion, agriculture, between-human contact) factors (**????**, to name a few). Notwithstanding the frequency and importance of microbial community coalescence, the mechanisms responsible for the community structure and function resulting from coalescence events remain poorly understood (**?**).

Early mathematical models of community-community invasion revealed that when two communities previously separated by a barrier merge due to its removal, asymmetrical dominance of one community over the other one is likely to occur (**??**). As an explanation for this observation, it was argued that, because communities have been assembled through a history of competitive exclusion, they are likely to compete with each other as coordinated entities, rather than as a random collection of species. This result also stems from later theoretical work, where consumer-resource models are used to show that in a microbial community

coalescence event, the winning community will be that which is capable of simultaneously depleting all resources more efficiently. (**??**). Overall, these findings suggest that communities arising from competitive species sorting experiment a certain level of cohesiveness that makes them less vulnerable to invasions by other communities.

Similar results arise from coalescence experiments with methanogenic communities (**?**), where it is reported that during a coalescence event, multiple taxa from the community with the most efficient resource use act as cohesive units and are selected together (ecological co-selection). Further experimental evidence of co-selection in community coalescence has been reported in **?**, where it was shown that the invasion success of a given taxon is determined by its community members. The microbial communities used in these experiments are known to display complex cross-feeding interaction networks, where leaked metabolic by-products of one species act as substrates for others (**???**). This networks can vary in their particular link distribution (the architecture of the flow of metabolites shared across species), but also in their link weights (the fraction of shared by-products, as opposed to the resources kept for private use) (**?**). Several studies point out that the type and strength of the interactions present in microbial communities is a factor that might affect the result of community coalescence (**???**). However, theoretical explorations of community-community invasions up to date have only considered competitive interactions for abiotic resources, leaving out interspecies cross-feeding, an essential feature of this systems.

In this work we develop theory and perform simulations to explore how the presence of syntrophic interactions influences the outcome of community-community encounters. First, we use a consumer resource model that includes metabolic cross-feeding to assemble microbial communities spanning a broad range in the competitive-cooperative spectrum. Second, we propose new metrics of community competition and cooperation, and use them to measure competition and facilitation levels in the assembled communities. Third, we perform coalescence experiments between random community pairs and use our interaction metrics to explain the outcome of the simulations.

3 Theory

Here, we motivate and present the model used for the simulations and explain the assumptions we make. Next, we propose metrics for the levels of competition and mutualism expressed within the framework of the model.

3.1 Model

We use a consumer-resource model adapted from the work of **?**. Consider an environment with m resources present in different concentrations R_j, where $j \in \{1 \dots m\}$. Let now n_α denote the abundance of each bacterial strain α present in the environment, where $\alpha \in \{1 \dots s\}$. Each species is uniquely characterized by its metabolic strategy to harvest resources. Additionally, all species share a core metabolism so that a fraction each consumed resource is released in the form of other resources to the environment. If we now allow the dynamics of this system to unfold, the concentration of each metabolite R_j determines the dynamics of the abundances n_α of each species, which harvest resources according to their metabolic

strategies and secrete by-products through the metabolic network. The change in species abundances translate into changes in the total supply and demand of resources. In turn, resource concentrations R_j are depleted until equilibrium is reached.

The following assumptions are made on the above model: (1) all resources contain the same amount of energy (taken to be 1 for simplicity), (2) a type I functional response, (3) a binary matrix of consumer preferences and (4) a complex environment where all resources are externally supplied in equal amount.

These assumptions correspond to the following equations for the change in consumer biomass and resource abundance with time

$$\frac{dn_\alpha}{dt} = g_\alpha n_\alpha \left((1-l) \sum_j c_{\alpha j} R_j - z_\alpha \right) \tag{1}$$

$$\frac{dR_j}{dt} = \kappa_j - \tau_j^{-1} R_j - \sum_\alpha n_\alpha c_{\alpha j} R_j + l \sum_{\alpha k} n_\alpha D_{kj} c_{\alpha k} R_k \tag{2}$$

In equation **??**, the growth of species α is determined by the resources it harvests, which in turn depends on the resource concentration R_j, and whether or not the species α consumes resource j ($c_{\alpha j} = 1$ or $c_{\alpha j} = 0$, respectively). However, not all the harvested resources contribute to growth. A fraction of them, l, leak back to the environment as metabolic by-products. Additionally, species α retain some of the harvest for its living maintenance cost z_α. For simplicity, we assume this cost to be random:

$$z_\alpha = \chi_0 \sum_j c_{\alpha j} (1 + \epsilon) \tag{3}$$

Here, χ_0 is the average cost of being able to consume a given metabolite, the summation represents the number of resources that species α is able to process, and ϵ is a random fluctuation sampled from a truncated normal distribution (to ensure that $z_\alpha > 0$). The effect of the cost function in equation **??** is to ensure that neither generalists nor specialists are systematically favoured during the community assembly (see section **??** for details).
The remaining resources after subtracting secretions and maintenance are transformed into biomass through g_α, a proportionality constant relating energy to abundance of strain α.

In equation **??**, κ_j and τ_j^{-1} encode the external dynamics of the resources; they are the supply and dilution rate of resource j, respectively. The leakage fraction is l, $(0 < l < 1)$, and D_{jk} encodes the metabolic matrix. Each element of D_{jk} represents the leaked energy of resource j that is transformed into resource k. Note that by definition, D is a row stochastic matrix, i.e., its rows sum to 1. Note that the equations of the model are expressed in index notation; the expressions in matrix form can be found in section **??**

3.2 Competition and facilitation metrics

In this system, competition for resources exists because the metabolic preferences vectors of the species in the community will have, in general, some level of overlap. We define intrinsic competition between a pair of species (α, β) as the level of overlap between their metabolic preferences, independent of their environment's state. We calculate their metabolic overlap

by counting the number of common preferences. Since these are binary, taking the scalar product of the two preference vectors yields the number of common elements between them,

$$(C_i)_{\alpha\beta} = \sum_k c_{\alpha k} c_{\beta k} \tag{4}$$

This intrinsic competition becomes effective competition when we consider the environmental states experienced by the consumers.

The resources available in the environment can either be of abiotic (when they come from the external supply κ) or biotic (if they have been leaked by consumers through l and D) origin. It is necessary to distinguish between the effective competition for these two types of resources because... We calculate the effective competition (henceforth, just 'competition') between a given pair of species on abiotically vs biotically generated resources as follows.

The competition for each abiotically-generated resource is weighted by the normalized resource supply

$$\tilde{\kappa} = \vec{\kappa}/\max(\vec{\kappa}),$$

with the interaction strength given by the non-leaked fraction $1 - l$ (fraction of any resource that is effectively consumed abiotically). Thus, the overall competition C_a between the species pair (α, β) for abiotically-generated resources is

$$(C_a)_{\alpha\beta} = (1 - l) \sum_k \tilde{\kappa}_k c_{\alpha k} c_{\beta k} \tag{5}$$

Now consider the competition C_b between species α and β for biotically-generated resources (which have been leaked to the environment). There will only be competition for the k^{th} biotically-generated resource when both the following two conditions hold: 1) α or β consume resource j and leak a fraction of it in the form of metabolite k; 2) α and β both consume k. Mathematically, this combination of conditions implies:

$$(C_b)_{\alpha\beta}^{k\to j} \propto D_{jk} \left(c_{\alpha j} + c_{\beta j}\right) c_{\alpha k} c_{\beta k} \tag{6}$$

Furthermore, the interaction strength for competition on biotically-generated resources is given by l (the fraction of any resource available because of biotic activity; see section **??** for a detailed explanation of this). Then, the total competition between the species pair (α, β) for biotically-generated resources is the sum of the terms that arise from expression **??** when spanning the full resource set weighted by l:

$$(C_b)_{\alpha\beta} = l \sum_{jk} \tilde{\kappa}_j D_{jk} \left(c_{\alpha j} + c_{\beta j}\right) c_{\alpha k} c_{\beta k} \tag{7}$$

Note that the factor $\tilde{\kappa}_j$ is necessary in both measures of competition (equations **??** and **??**) because all resources ultimately come from externally-supplied resources, and so we need to account for possible differences in the external resource supply rate.

Finally, we can write the total community competition matrix as

$$C_T = C_a + C_b \tag{8}$$

Next, we consider facilitation. Facilitation links form when a species leaks by-products that are used by another species. Therefore, the facilitation of species α to β of metabolite k, due to the consumption of metabolite j, will be non-zero when 1) species α produces k as a by-product when consuming j, and 2) species β consumes k. Analogous to the case of competition for biotically-generated resources, the strength of these interactions is also given by l and the factor $\tilde{\kappa}_j$. Combining these conditions, the total facilitation of species $\alpha \rightarrow \beta$ will be

$$F_{\alpha\beta} = l \sum_{jk} \tilde{\kappa}_j c_{\alpha j} D_{jk} c_{\beta k} \tag{9}$$

Note that facilitation is directional, which implies that the community facilitation matrix F is not symmetric. On the other hand, the community competition matrix C_T is symmetric, since competition is not directional. Averaging the matrices of pairwise metrics **??** and **??** yields a proxy of the community-level competition and facilitation. The matrix diagonals are left out of this averaging because we are only interested in inter-specific competition and facilitation. Additionally, we only consider the upper diagonal elements of the symmetric competition matrix. Then, the community-level competition and facilitation measures are, respectively

$$\mathcal{C} = \langle C_{T_{\alpha\beta}} \rangle_{\alpha>\beta} \quad \& \quad \mathcal{F} = \langle F_{\alpha\beta} \rangle_{\alpha\neq\beta} \tag{10}$$

4 Simulations

Following the explanation of the theory supporting this work, we proceed to the analysis. First, we sample the consumer preferences and metabolic parameters from random distributions. We impose specific constraints to this distributions in order to broaden the range of competition and facilitation levels in the communities. Second, we simulate all instances of community assembly. Next, we perform community coalescence simulations between random pairs of communities. Finally, we analyse the species abundance in the mix and compare it to the original species abundances in each species. We try to explain the differences with the interactions between community members. The parameter values used in the simulations performed here can be found in section **??**.

4.1 Preferential sampling of $c_{\alpha j}$ and D_{jk}

We are interested in generating communities spanning a broad range of cohesion levels, i.e., different competition and facilitation levels. These can be modified by sampling $c_{\alpha j}$ and D_{jk} respectively, with specific constraints.

First, let us consider the problem of modulating the competition level of the community by sampling the metabolic preferences for each strain, encoded in $c_{\alpha j}$. Each species α has a binary vector $\vec{c_\alpha}$ of length m that specifies if resource j is consumed ($c_{\alpha j} = 1$) or not ($c_{\alpha j} = 0$). The metabolic preference matrix C is constructed by sampling its rows $\vec{c_\alpha}$ sequentially. The process of sampling $\vec{c_\alpha}$ is two-fold. We first sample m_α, the number of metabolites that species α consumes, from an exponential distribution. This choice is supported by experimental evidence (**?**). Second, to determine which metabolites are consumed by species α, we sample

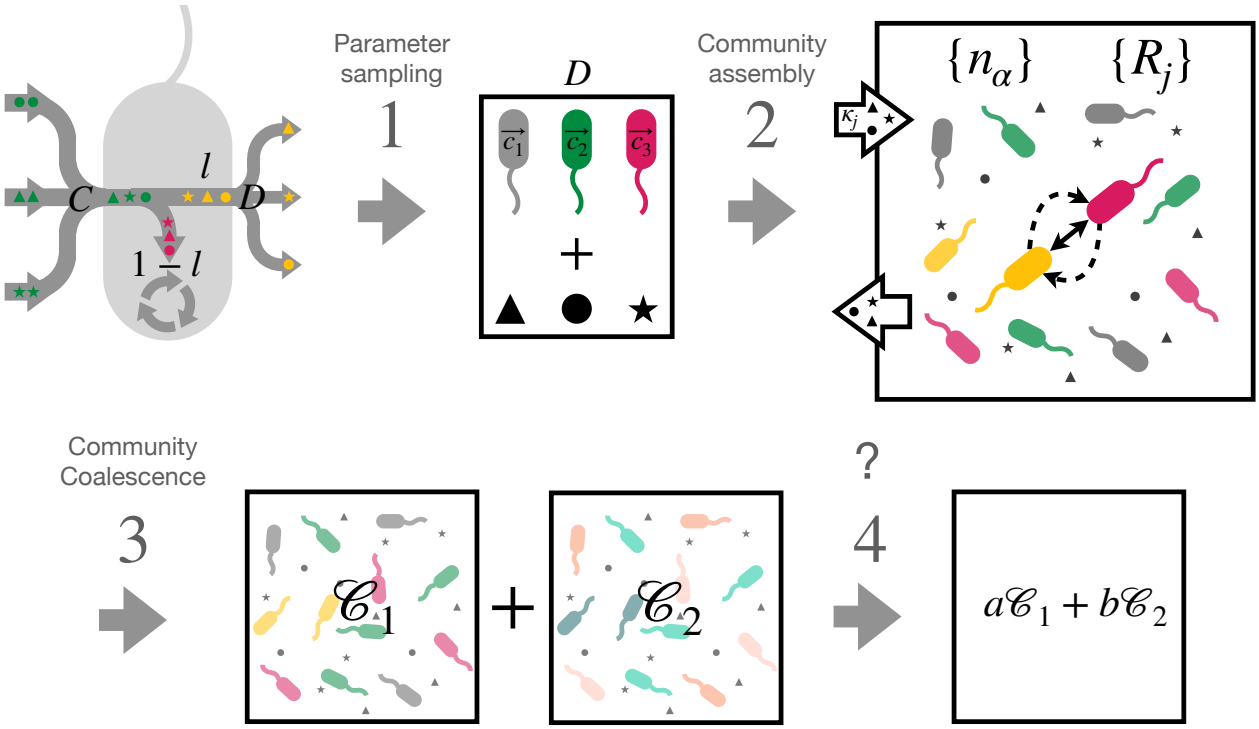

Figure 1: **Workflow scheme.** First, the metabolic preferences of $s = 60$ bacterial strains behaving as explained in section **??** and sketched in the top left corner, and one matrix D, are sampled for each community (parameter sampling). This sampling is done either with taxonomic and metabolic structure, or without it, as specified in section **??**. Second, the dynamics of the sampled system play out in an environment with $m = 60$ resources, according to equations **??** and **??**, until equilibrium of the community is reached (community assembly). Third, these communities are randomly paired up and re-equilibrated together in fresh media (community coalescence). Fourth, the contribution of each community to the final mix is analyzed as a function of the competitive and facilitatory interactions between species (emphasized with black arrows in the top right corner).

m_α metabolites with probability vector $\vec{p_\alpha}$. Note that in this sampling scheme, 'iteration number' and 'species' are equivalent, and denoted by index α.
For species $\alpha = 1$ all metabolites have the same probability of being sampled $1/m$, where m is the total number of resources. This assumption is consequent with the absence of a metabolite hierarchy, since they all carry the same energy. After each iteration, the sampling probability of each metabolite changes according to what has been sampled so far. Let $d_{\alpha j}$, denote the cumulative demand of resource j when the metabolic preferences of species α are sampled. That is, the number of consumers of resource j at iteration α.

$$d_{\alpha j} = \sum_{i=1}^{\alpha} c_{ij} \tag{11}$$

Based on $d_{\alpha j}$, I then compute the probability that species α is assigned resource j as one of

its preferences

$$p_{\alpha j} = (1 - k_c)\frac{1}{m} + k_c \frac{d_{\alpha-1 j}}{\sum_j d_{\alpha-1 j}} \tag{12}$$

where the denominator represents the total number of preferences sampled up until iteration $\alpha - 1$, and acts as a normalization constant.
The strength with which $p_{\alpha j}$ deviates from a uniform distribution is given by the parameter $k_c \in [0, 1)$, the competitiveness factor. When $k_c \to 1$, competition is maximized. Thus, highly demanded metabolites are more likely to be sampled in the next iteration. On the contrary, when $k_c = 0$ the competition is random, since each metabolite is equally likely to be chosen.

The metabolic preferences sampling procedure is implemented in the following algorithm

Algorithm 1: Sampling of metabolic preferences

for $\alpha \in \{1, \ldots, s\}$ **do**
  Sample m_α from an exponential distribution
  Sample vector $\vec{v}$ of m_α integers $\in \{1, \ldots, m\}$ with probability vector $\vec{p}(\alpha)$
  Switch on sampled preferences $\vec{c_\alpha}[\vec{v}] = 1$
  Update $\vec{d_\alpha}$
  Update $\vec{p_\alpha}$ using the new $\vec{d_\alpha}$
end

To illustrate the behaviour of the proposed sampling method, we run this algorithm for the two aforementioned values of k_c in a system with 50 metabolites and 1000 species. Overlaying the cumulative demand vector at each iteration yields representative figures of each case (see figure **??**).

Second, let us consider the problem of modulating the facilitation level of the community by sampling the metabolic cross-feeding topology, encoded in the community metabolic matrix D. Each element of the metabolic matrix, D_{jk}, specifies the fraction of leaked energy from resource j that is released in the form of resource k. Note that by definition $\sum_j D_{jk} = 1$. The matrix row vector $\vec{D_j}$ is sampled from a Dirichlet distribution, with concentration parameter vector $\vec{q_k}$, where

$$q_{jk} = \frac{1}{s}(1 + k_f d_j) \tag{13}$$

Here, s is the sparsity of the metabolic network, ranging from a fully connected network when $s \to 0$ to a sparse one-to-one network when $s \to 1$. The constant k_f is the facilitation factor, which tunes the level of structure that the matrix will have. When $k_f \to 0$ the metabolic matrix has no structure. The concentration parameters is that of a flat Dirichlet distribution, and therefore all metabolites are released with equiprobable fractions. As $k_f \to 1$ the structure of D is fully determined by the resource demands of the community, so that more coveted resources are released at higher fractions. Note that although the two methods for sampling c and D share similarities, they are conceptually different. First, the sampling of c is fully random, in the sense that a vector of probabilities is constructed first, and then preferences are randomly sampled from that distribution. On the other hand, the sampling

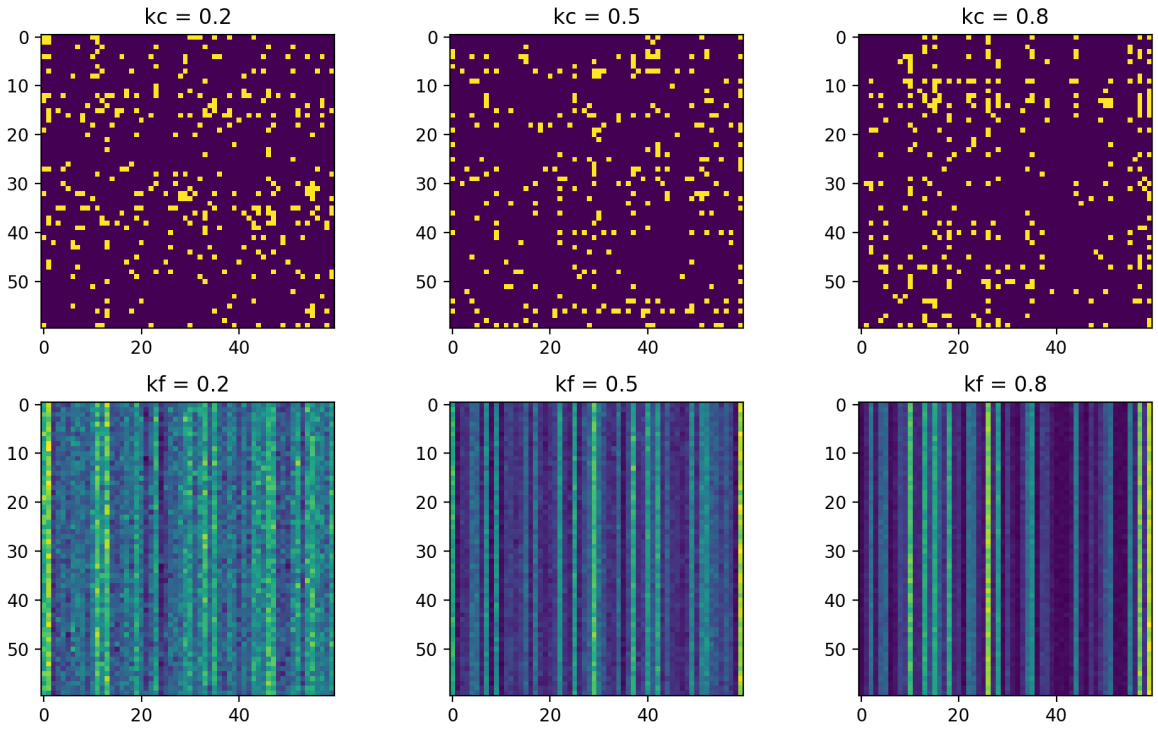

Figure 2: **Preferential structure sampling of C and D** Samples of C (top) and D (bottom) for different values of k_c and k_f respectively in a system with 60 metabolites, 60 species. Note that both constants tune the amount of structure in the matrix, so that when $k_c, k_f \to 0$ there is no structure, and when $k_c, k_f \to 1$ the structure is maximum.

of D has a random term, x_{jk}, and a non-random term that depends only on the vector of demands of the community $\vec{d_s}$. Another difference is that the competitiveness factor is imposed on the preference sampling probability vector, while the facilitation factor is imposed directly on the values of D. Finally, the sampling of D depends on the sampling of c, reflecting that high or low facilitation levels are achieved by tuning the secretion structure of the community to be symmetric to the profile of demands, or independent from it, respectively.

4.2 Sampling $c_{\alpha j}$ and D_{jk} with metabolic and taxonomic structure

Taxonomic and metabolic structure is imposed on the preference matrix and metabolic cross-feeding matrix, respectively, through the partition of the resource space into a set C of n_c resource classes. We define a taxonomic family T as a subset of resource classes, ($T \subseteq C$), characteristic of a species. Therefore, if species α, belongs to taxonomic family T the probability of sampling resource j will be higher if that resource belongs to a resource class that is in the taxonomic family, i.e., $C(j) \in T$, and lower otherwise. The magnitude of this difference is tuned by the taxonomic heterogeneity constant $K_c \in [0, 1]$. Thus, the probability

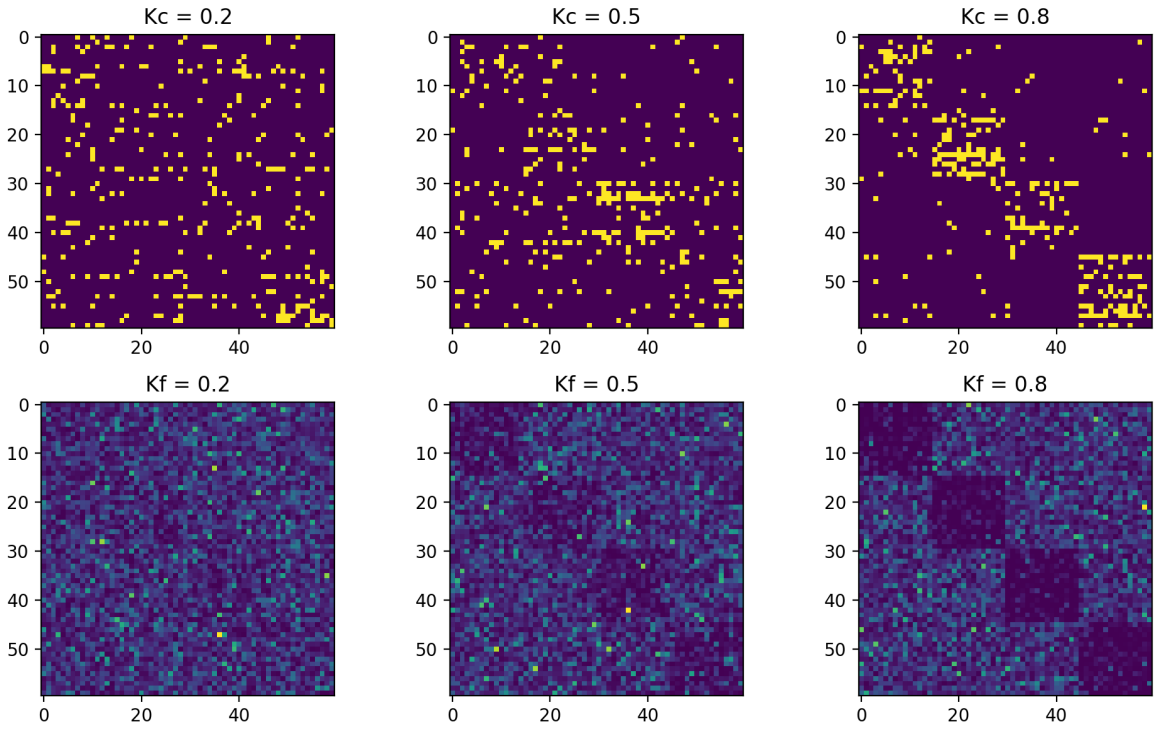

Figure 3: **Sampling with taxonomic and metabolic structure.** Samples of C (top) and D (bottom) for different values of K_c and K_f respectively in a system with 60 metabolites, 60 species and 4 resource classes. Note that both constants tune the amount of structure in the matrix, so that when $K_c, K_f \to 0$ there is no structure, and when $K_c, K_f \to 1$ the structure is maximum.

$p_{\alpha j}$ that species α consumes resource j is given by

$$p_{\alpha j} = \begin{cases} \dfrac{1}{m} + aK_c & \text{if } C(j) \in T \\ \dfrac{1}{m} - bK_c & \text{otherwise} \end{cases} \tag{14}$$

Now we impose that (1) the sum of the elements of $\vec{p_\alpha}$ must equal to 1, and (2) the lower probability must be 0 when $K_c = 1$. These constrains yield the following systems of equations that can be solved for a and b.

$$\begin{cases} n_c \left(\dfrac{1}{m} + aK_c \right) + (m - n_c) \left(\dfrac{1}{m} - bK_c \right) = 1 \\ \frac{1}{m} - b = 0 \end{cases} \tag{15}$$

Finding that

$$a = \frac{1}{n_c} - \frac{1}{m} \qquad \text{and} \qquad b = \frac{1}{m} \tag{16}$$

Finally, the sampling probability of resource j when imposing taxonomic structure, is given by.

$$p_{\alpha j} = \begin{cases} \frac{1}{m}\left(1 + K_c\left(\frac{m}{n_c} - 1\right)\right) & \text{if } C(j) \in T \\ \frac{1}{m}(1 - K_c) & \text{otherwise} \end{cases} \tag{17}$$

Note that K_c controls the amount of taxonomic structure in the community. When $K_c = 1$, the probability of sampling a resource outside the taxonomic family is 0, and all the sampled preferences belong to the subset of classes conforming the taxonomic family. On the opposite end, when $K_c = 0$ the probability of sampling a resource from the taxonomic family is the same as that of sampling a resource outside the taxonomic family.

We impose metabolic structure through a two tier secretion model. The first tier contains by-products that are not in the resource class of the substrate (off-block diagonals of D), and the second one contains the products that belong to the same resource class as the substrates (block diagonals of D). We encode this structure in D by sampling each row from a Dirichlet distribution with concentration parameters q_{jk} that depend on the by-product tier as:

$$D_{jk} = \text{Dir}\left(q_{j1}, q_{j2}, \ldots, q_{jm}\right)_k \tag{18}$$

where

$$q_{jk} = \begin{cases} \frac{1 - K_f}{sM_{C(j)}} & \text{if } C(j) = C(k) \\ \frac{K_f}{sM_{C(j)}} & \text{otherwise} \end{cases} \tag{19}$$

Here, s is the sparsity of the metabolic network, ranging from a fully connected network when $s \to 0$ to a sparse one-to-one network when $s \to 1$. $M_{C(j)}$ is the number of consumers in the class to which resource j belongs. In these simulations $s = 0.05$, and the number of consumers in each class is the same $M_C = m/n_c$

4.3 General case

The most general sampling scheme that we can write is the one that would emerge if we superimpose figures **??** and **??**. Under this sampling scenario, we have a two levels of organization; a coarse structure with inter-guild facilitation, and a fine structure with inter-species facilitation. This sampling allows us to decouple the terms C_b and F, so that their respective effects can be correctly quantified, while keeping the levels of facilitation and competition highly adjustable through the constants k_c and k_f.

The algorithm that we use to construct to sample the metabolic preferences matrix C is analogous to that used in section **??**. The only difference is that the probability $p_{\alpha j}$ that species α is assigned resource j as one of its preferences has a different form, which we derive next.

Our aim is to write a normalized piece-wise function out of equation **??**, where each piece is weighted more or less depending on resource j being taxonomically preferred by species α

or not, respectively. Therefore, we can assume the form of this function to be

$$p_{\alpha j} = \begin{cases} A\left((1-k_c)\frac{1}{m} + k_c\frac{d_{\alpha-1j}}{\sum_j d_{\alpha-1j}}\right)(1+K_c) & \text{if } C(j)\in T \\ \frac{B}{m-n_c}(1-K_c) & \text{otherwise} \end{cases} \tag{20}$$

where A and B are normalization constants that ensure $\sum_j p_{\alpha j} = 1$. Note that K_c modulates the amount of structure in the matrix, ranging from no structure when $K_c = 0$ to maximum taxonomic structure when $K_c = 1$.

In order to obtain expressions for the normalization constants A and B we impose the following constrains on each piece of equation **??**.

$$p^1_\alpha = \sum_{C(j)\in T} p_{\alpha j} = \frac{n_c}{m}(1-K_c) + K_c \tag{21}$$

and

$$p^0_\alpha = \sum_{C(j)\notin T} p_{\alpha j} = \left(1-\frac{n_c}{m}\right)(1-K_c) \tag{22}$$

Note that these two constraints are sensible because they verify two necessary conditions; (1) the total probability sums to one, since $p^1_\alpha + p^0_\alpha = 1$, and (2) probability of sampling metabolites in (outside) the preferred class approaches 1 (0) when K_c increases (decreases). We now solve for A and B by expanding equations **??** and **??** using the definition in expression **??** for $p_{\alpha j}$

$$A\left((1-k_c)\frac{n_c}{m} + k_c\frac{T_c}{T}\right)(1+K_c) = \frac{n_c}{m}(1-K_c) + K_c$$

$$A = \frac{K_c + \frac{n_c}{m}(1-K_c)}{(K_c+1)\left(\frac{n_c}{m}(1-k_c) + \frac{T_c}{T}k_c\right)} \tag{23}$$

and

$$B = 1 - \frac{n_c}{m} \tag{24}$$

where $T = \sum_j d_{\alpha-1j}$, and we have used the following expressions

$$\sum_{C(j)\in T} 1 = n_c \qquad \sum_{C(j)\notin T} 1 = m - n_c \qquad \sum_{C(j)\in T} d_{\alpha-1j} = T_c$$

Thus, the closed form of the sampling probability under the general scenario is

$$p_{\alpha j} = \begin{cases} \left(K_c + \frac{n_c}{m}(1-K_c)\right)\frac{\left((1-k_c)\frac{1}{m} + k_c\frac{d_{\alpha-1j}}{T}\right)}{\left((1-k_c)\frac{n_c}{m} + k_c\frac{T_c}{T}\right)} & \text{if } C(j)\in T \\ \frac{1}{m}(1-K_c) & \text{otherwise} \end{cases} \tag{25}$$

The metabolic matrix D is also constructed with the two levels of structure imposed in C. We sample its elements from a Dirichlet distribution with concentration parameters q_{jk}, that depend on the by-product class and the vector of resource demands as:

$$q_{jk} = \begin{cases} \frac{1}{sM_{C(j)}}(1 + k_f d_j)(1 - K_f) & \text{if } C(j) = C(k) \\ \frac{1}{sM_{C(j)}}(1 + k_f d_j)(1 + K_f) & \text{otherwise} \end{cases} \tag{26}$$

4.4 Community coalescence

We simulate two sets of community assembly instances. First, under the preferential sampling scenario; secondly under the sampling scenario with added metabolic structure. Within each set, we perform 100 simulations at every point of the parameter grid, which encompasses all possible combinations of the following parameter values $s = m = 60$, $l = [0.1, ..., 0.9]$, $k_c = [0, ..., 0.9]$, $k_f = [0, ..., 1]$, $K_c = [0.1, ..., 0.9]$, $K_f = [0.1, ..., 0.9]$. We then use the assembled communities to perform *in silico* community invasion assays.

Here, a community coalescence event is performed by mixing a random pair of communities that have been equilibrated independently and letting the combined system relax to a new equilibrium state. We analyze the species abundance of the mix at equilibrium to try to address the effect of the interactions present in the community in the outcome of the coalescence process.

Previous works in microbial community coalescence (**????**), focus on the cohesion of a community as an important property when predicting coalescence success. Here, we quantify the community cohesion, Θ, by pinning it down to the interactions between the different species in the community. More cohesive communities will be those where competition is minimized and facilitation is maximized, such that

$$\Theta = \mathcal{F} - \mathcal{C} \tag{27}$$

where $\mathcal{F}$ and $\mathcal{C}$ are the measures of competition and facilitation presented in section **??**, equation **??**.

First, we simulate coalescence between pairs of communities assembled under the first sampling scenario (section **??**, figure **??** A). For each value of leakage, we perform $1.5 \cdot 10^4$ simulations where 2 randomly sampled communities are mixed in fresh media (all resources are set to the initial concentrations before assembly). The cohesion values of each community Θ_1 and Θ_2 are recorded before coalescence. An extended system with the two communities (see section **??**) is then created, and its dynamics let to play out (equations **??** and **??**) until a new equilibrium is reached. The species abundance of the mix at equilibrium is analyzed in order to compute the similarity of the outcome community to each of their parent communities. This measure, $S(C_2, C_1) \in [-1, 1]$, specifies the identity of the post-coalescence community in the basis of the parent communities C_1 and C_2 with original richness r_1 and r_2 respectively, and it is calculated as

$$S(C_2, C_1) = \frac{1}{r_1}\sum_{i=1}^{r_1} b_i - \frac{1}{r_2}\sum_{i=r_1}^{r_2} b_i \tag{28}$$

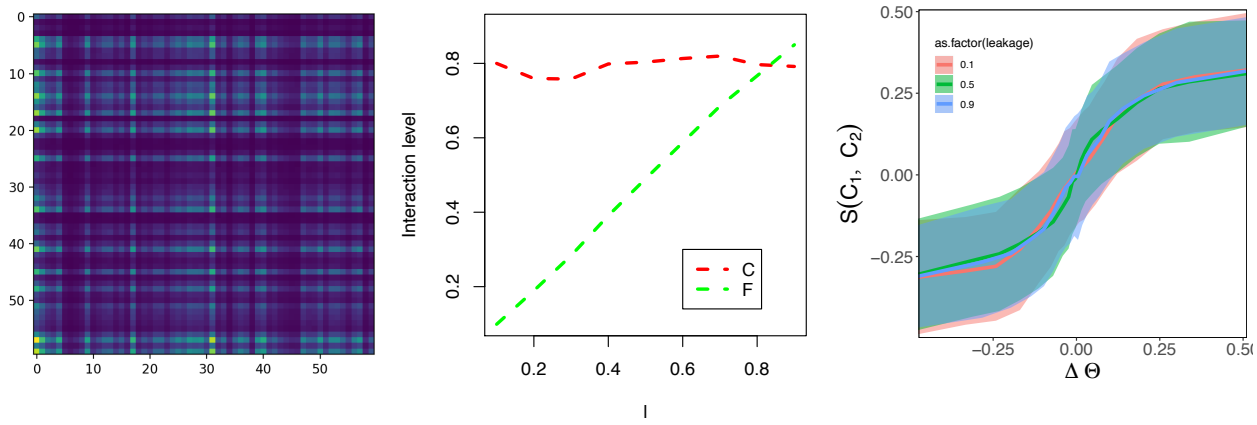

Figure 4: **Results for assumption 1.** Figure A shows one example of the effective matrix of leakage, CD. A matrix element $(CD)_{\alpha k}$ represents the total leakage of metabolite k by species α. Therefore, lighter (darker) horizontal stripes correspond to generalist (specialist) species, and lighter vertical stripes are more demanded metabolites. Figure B shows community-level competition ($\mathcal{C}$) and facilitation ($\mathcal{F}$) averaged across simulations for each levels of leakage. When $l << 1$, abiotic competition dominates over biotic competition, $C_a >> C_b$, but this relation inverts as leakage increases, so that total competition C_T (either for biotically or abiotically-generated resources) remains consistently high across all values of leakage. Facilitation increases linearly with l. Figure C shows that community coalescence success is positively correlated with community-level cohesion. The composition of the post-coalescence community on the basis community 1 ($S = 1$) - community 2 ($S = -1$), $S(C_2, C_1)$ is plotted as a function of the community cohesion difference $\Delta\Theta = \Theta_1 - \Theta_2$ between them. Shown is binned mean (20 bins) over communities with similar $\Delta\Theta$ (solid line) $\pm\sigma$ (shaded).

where b_i represents presence (1) or absence (0) of species i in the post-coalescence community. Note that if $S = 1$, the post-coalescence $C_P = C_1$, and if $S = -1$, then $C_P = C_2$. Also note that this measure is richness independent. This allows us to mix communities with different richness without introducing a bias in the similarity towards the richer community.

Once the community composition of the mix is measured, we plot it against the difference in cohesion of the parent communities $\Delta\Theta$ (see figure **??** C) , finding that the resulting community is similar to its more cohesive parent. In the case of low leakage, facilitation is negligible (see figure **??** B), so in this regime, being more cohesive is equivalent to being less competitive. Therefore, in the low leakage regime, communities that minimize competition succeed in coalescence events. Surprisingly, this trend is consistent even for high values of leakage, where $\mathcal{C} \approx \mathcal{F}$. This suggests that when competition is not negligible, it drives the outcome of community coalescence, overriding any effects that facilitation may have. To uncover the effects of facilitation in the community (which will be expetedly weaker than the effects of competition) we need to switch off competition for a range of leakage values where facilitation is not negligible.

To this end we perform a second set of community coalescence simulations between pairs of communities assembled under the general sampling scenario (section **??**, figure **??** A). This sampling method allows us to substantially lower the term C_b in equation **??**, bringing out a new regime at high l values where $\mathcal{C} << \mathcal{F}$ (see figure **??** B). The simulation pipeline

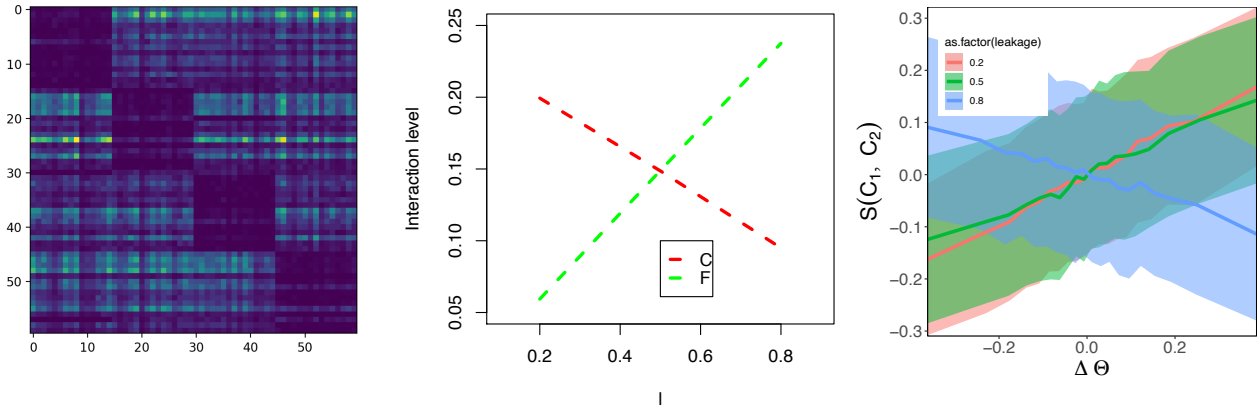

Figure 5: **Results for assumption 2.** Figure A shows an example of the effective leakage matrix with added metabolic structure. Figure B shows community-level competition ($\mathcal{C}$) and facilitation ($\mathcal{F}$) averaged across simulations for different levels of leakage. For low values of leakage, abiotic competition C_a dominates, and for high values of leakage facilitation F is the important term. Biotic competition C_b is negligible due to the imposed class structure. Figure C, shows that success of community coalescence is positively correlated with $\Delta\Theta$ for low values of leakage, when $\mathcal{C} > \mathcal{F}$; but negatively correlated with $\Delta\Theta$ for high values of leakage, when $\mathcal{C} < \mathcal{F}$.

detailed above is now run on these communities. In the low leakage regime, where competition is present, we recover the previous result. In the high leakage regime, competition is negligible, so being more cohesive is equivalent to having higher levels of facilitation. We find that cohesion is negatively correlated with coalescence success, that is, more facilitative communities perform poorly in coalescence events.

5 Discussion

A frequent process by which new microbial networks emerge is through the fusion of two or more communities, an event that has been termed community coalescence (**?**). Numerous theoretical and experimental studies suggest that coalescing communities behave as coherent wholes, and compete against each other like coordinated armies (**???????**). To date, most theory about microbial multi-species invasions focuses on competition, ignoring other interactions arising from metabolic cross-feeding that might change the outcome of coalescence events (**?**).

In the work of **?**, the cohesiveness displayed by coalescing communities in the absence of cooperative interactions resulted in effective resource depletion. This allowed the winning community to engineer an environment more favorable for itself than for the loosing community, which was partially or completely displaced. The latter theoretical prediction has been experimentally verified in methanogenic communities (**?**), which are characterized by a dense metabolic cross-feeding network. However, we cannot help to question this claim: how is it possible that a minimal theoretical setting built exclusively around competition can explain the complex reality of coalescence events in the presence of syntrophy?

In this work we presented a theoretical framework that contemplates the intricate cross-feeding topology displayed by microbial communities and reconciles more realistically theory with observations. The results we obtained in the absence of leakage confirm the previous theoretical prediction that more cohesive communities out-perform in community coalescence. Here, cohesive communities were those were competition was minimum. Because these communities behave analogously to those in **?** in coalescence simulations, it must be that resource use efficiency is a consequence of minimizing of competition.

When leakage was present, we found that reducing competition was still the main force driving the outcome of community coalescence. The difference being that now competition was taking place in another environment; the one engineered by a community that both used and created resources. Ultimately, competition in the biotically-generated resource space exists because bacteria are leaking resources necessary for their own growth. While this might seem disadvantageous, and thus unrealistic at first, leaking essential metabolites is a real phenomenon in many microbial systems, and has been experimentally observed (**??**) and theoretically explored (**??**). Overall, this finding extends the results of previous theoretical studies to accommodate metabolic interdependence; an essential feature of microbial communities, thus constituting a more robust result supporting experimental findings (**?**) and hypotheses (**?**).

The second set of simulations introduced taxonomic classes that confined each consumer to obtain its energy from a subset of resources. We then paired the taxonomic structure with metabolic structure, such that species from a taxonomic class leaked energy in the form of resources that belonged to a different class than their preferred one. These two constraints allowed us to survey a regime of communities with very low competition and high facilitation levels. Performing a second round of coalescence simulations revealed that cohesion was detrimental for coalescence success in cooperative communities, that is, those communities where facilitation was high and competition was negligible were easily displaced by an invading community. This finding is experimentally supported by several studies (**???**) which recognize that strong cooperative links are susceptible to be intercepted by invading species. Nonetheless, recent *in-silico* results of single species invasions on microbial communities have found that cooperative communities are more resistant to invasions than their competitive counterparts (**?**). The contradicting nature of this finding with our own suggests that community invasion ecology cannot simply be extrapolated from our understanding of single species invasions (**?**).

The pairs of coalescing communities in this work were drawn with no richness restrictions. Consequently, the results reported here are independent of the species richness of the mixed communities. Interestingly, several works have pointed to microbial community diversity as an important factor driving resource use efficiency and, therefore, determining community resistance against biotic and abiotic perturbations (**???**). These observations do not necessarily contradict the results reported here. Instead, our findings suggest that community interactions may be a more fundamental mechanism explaining the response of communities to environmental and biotic perturbations, and that biodiversity is rather a consequence of the underlying community interaction network. It is not surprising that experiments have come across biodiversity's influence in community resilience before the effect of community interactions, since the latter is much harder to measure than the former. Understanding biodiversity as an emergent property of the interaction network topology in a microbial com-

munity is a promising line of research (**?**), particularly in the context of climate change (REFERENCE(S)).

Throughout this work, we assumed absence of environmental fluctuations, since the supply rate of all resources was kept fixed. Considering only biotic perturbations allowed us to pinpoint the effect of each interaction separately. While this assumption may be sensible in some cases (**?**), it surely is oversimplifying in others (**?**). When two communities collide, the process will entail, in general, a mixture of biotic and abiotic perturbations. Laying down a theoretical framework to understand the effect of abiotic perturbations, and the interplay between the two, in multi-species microbial invasions is an exciting direction for future research.

The only interactions considered here were facilitation and competition. However, microbial inter-relations are more complex than just a binary classification (**?**), often involving the release of antimicrobial compounds, end-product inhibition, predation, or interactions with spatial dependencies, an aspect that was also omitted here. Finally, stochastic effects were only considered during community assembly, when species were randomly sampled to colonize the initial population. These effects were omitted during community coalescence. Although they have been experimentally shown to be weak (**?**), we cannot ignore that they might had played a significant role in our simulations.

Encounters between microbial communities are becoming increasingly frequent across the globe (**?**). Moreover, *in-vitro/in-vivo* mixing of whole microbial communities is gaining popularity for bio-engineering (**?**), soil restoration (**?**), faecal microbiota transplantation (**??**), and the use of probiotics (**?**). In the absence of robust theory that complements these observations and experiments we present this framework, which ties interactions in microbial communities to the outcome of community coalescence events. Although more work is required so to bridge the gap between theory and experiments, this study constitutes a first step in that direction.

6 Supplementary Material

6.1 Equivalence with other consumer resource models

If we take the model equations **??** and **??** and assume that the dilution all resources is slow, that is, $\tau_j^{-1} << 1$, in the limit when there is no leakage ($l = 0$), we can make the assumption that molecular (resource) dynamics are faster than population dynamics, and therefore that resource concentration R_j at any moment quickly equilibrates to reflect the instantaneous demand thus, $dR/dt \approx 0$. This allows us to separate the time scales between resource and population dynamics, recovering the model presented in **?**

$$\begin{aligned} \frac{dN_\alpha}{dt} &= g_\alpha N_\alpha \Big(\underbrace{\sum_j c_{\alpha j} \frac{\kappa_j}{\sum_\alpha N_\alpha c_{\alpha j}} - z_\alpha}_{\text{Resource surplus } \Delta_\alpha} \Big) \\ R_j &= \frac{\kappa_j}{\sum_\alpha N_\alpha c_{\alpha j}} \end{aligned} \tag{29}$$

The mapping between the notation used in **?** (T), **?** (M) and those used here is provided in the table:

Notation for...	M	Here	T
Species index	i	α	$\vec{\sigma}$
pecies abundance	N_i	N_α	n_σ
Resource a species can harvest	$\vec{c_i}$	$\vec{c_\alpha}$	σ_i
Resource supply	κ_α	κ_j	R_i
Minimal resource requirement	m_i	z_α	$\chi_{\vec{\sigma}}$
Resource weight	w_α	$\vec{1}$	$\vec{1}$
Resource $\rightarrow$ biomass converison factor	g_i	g_α	$(\tau_0 \chi_{\vec{\sigma}})^{-1}$
Resource dilution rate	τ_β^{-1}	≈ 0	0
Leakage factor	l_α	l	0
Metabolic matrix	$D_{\alpha\beta}$	D_{jk}	NA

6.2 Community Assembly

We performed simulations of community assembly for the following parameters: $s = 40$, $m = 40$, $k_c = [0, \ldots, 0.9]$, $k_f = [0, \ldots, 1]$, $l = [0.05, \ldots, 0.95]$. Next, I plot heat maps for each value of l, of the average richness r as a function of k_c and k_f. I find that each value of leakage represents a starkly different regime. In figure **??**A leakage is negligible. Average richness decreases as k_c increases, but does not change with k_f. The leakage fraction is so low that the effect of the level of facilitation the diversity of the community is negligible. Figure **??**B represents an intermediate regime where the leakage level in the community is substantial. In this regime, the maximum average richness is reached around the top-left corner of the heat-map, where competition is lowest and facilitation is highest. Minimum average richness is reached for the opposite conditions. Figure **??**C illustrates communities with very high level of leakage. In this regime, highest diversity is reached by communities with low competition and

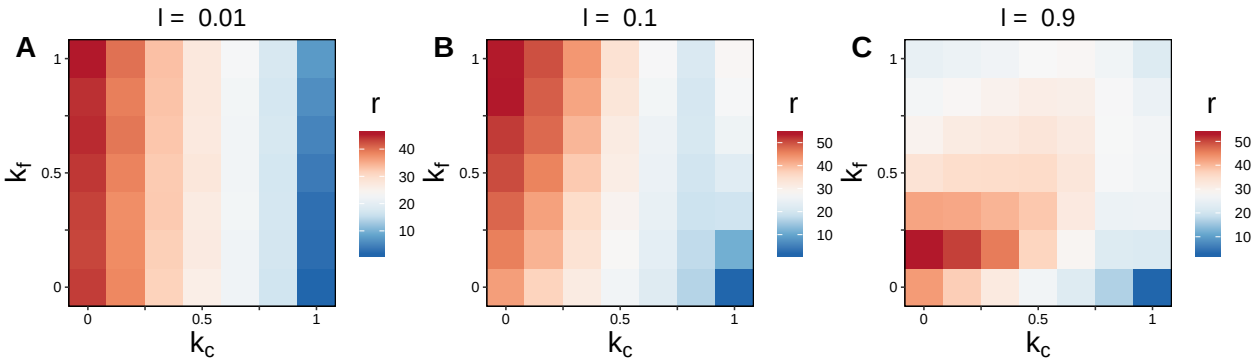

Figure 6: Heat-maps for each value of l of the average richness r as a function of k_c and k_f

facilitation levels. This result is counter intuitive, as one would expect that high facilitation promotes diversity. However, the species in this communities are not efficient at consuming resources, since they leak the majority of what they harvest. Thus, in order to deplete the available resources, they need to perform more cycles of consumption than a species with a lower leakage factor. The optimal interaction topology that ensures efficient resource depletion in several cycles of consumption and leakage is one that minimizes facilitation and competition. To sum up; predictably, the three figures show that biodiversity is favoured by niche separation, since less competitive communities are more diverse. Surprisingly, the benefit of mutualistic interactions changes for each case. When individuals are very selfish (very low l), mutualistic interactions are irrelevant. These become beneficial when individuals are moderately generous. However, they are detrimental, and therefore minimized when individuals are altruistic (very high l).

6.3 Cost function

The cost function used here ensures that neither specialists or generalists are systematically favoured during community assembly, and corresponds to the assumption of approximate neutrality (**????**).

Species α will increment its abundance when its surplus term in equation **??** is greater than 0;

$$\frac{1}{g_\alpha n_\alpha}\frac{dn_\alpha}{dt} = (1-l)\sum_j c_{\alpha j}R_j - z_\alpha > 0 \tag{30}$$

That is, the species that are able to maintain positive growth for the lowest concentration of resources, will be favoured. This is can be achieved by either reducing the cost z_α or increasing the number of resource preferences in equation **??**. However, if one imposes the constrain $z_\alpha = \sum_j c_{\alpha j}$, the condition to maintain positive growth becomes

$$R_j - \chi_0 > 0 \tag{31}$$

Which is independent of the number of preferences of the consumer; that is, species are neutral under this assumption. To break this degeneracy, and place our model in the regime of approximate neutrality, we introduce a small random fluctuation term (ϵ in equation **??**).

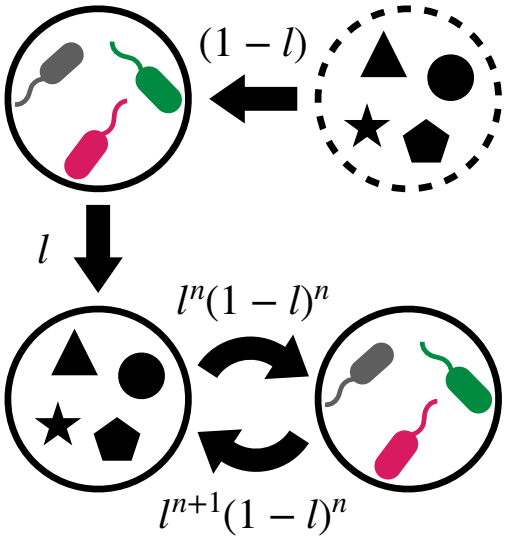

Figure 7: **Mechanism of resource recycling**. The species in this model consume the resources that come through the supply rate κ (abiotic uptake, top arrow) and leak a fraction l back to the environment (left arrow). The leaked resources can be harvested again by the species in the community (biotic uptake, bottom arrows). This can be modeled as an infinite series of consecutive cycles of uptake and leakage where the amount of resources consumed by each strain decreases by a factor of l, after each cycle.

6.4 Resource recycling mechanism

The equilibrium state of the communities in these simulations is a dynamic equilibrium, because in order to maintain the biomass at steady state, the species need to harvest resources from the environment. A portion of those resources is made up by biotic resources, those leaked by the community members. These resources are also competed for in a series of consecutive cycles of consumption and leakage. When a species harvests resource j from the abiotic environment, it intakes a fraction $1-l$ and leaks back a fraction l in the form of other resources (first two arrows in figure **??**). The leaked resources become part of the available substrates, and are competed for in a second cycle of consumption, with intake fraction $l(1-l)$. This process extends up to n cycles, (with $n \to \infty$), as illustrated in figure **??**, so that at cycle n, the fraction of harvested resource is $l^n(1-l)$. The strength of the biotic competition and facilitation links, B, is calculated by summing the fraction of ingested resource over $n \in [1, \infty)$.

$$B = \sum_{n=1}^{\infty} l^n(1-l) = (1-l)\left(\sum_{n=0}^{\infty} l^n - 1\right) = (1-l)\left(\frac{1}{1-l} - 1\right) = l \tag{32}$$

As expected, the strength factor for biotic links is l, the fraction of substrates originally leaked. Note that this result depends on the convergence of the geometric series, which is only true when $l < 1$. In the trivial case $l = 1$, then $B = 0$, and the system would have no competitive, nor facilitative interactions.

6.5 Matrix representations

The implementation of this framework in Python becomes significantly more efficient if the equations are vectorized.
First, the model presented in section **??** can be conveniently expressed in matrix form as follows

$$\begin{aligned}\frac{d\boldsymbol{n}}{dt} &= \boldsymbol{g} \circ \boldsymbol{n}\,(D(\boldsymbol{R})C(\mathbf{1}-\boldsymbol{l}) - \boldsymbol{z}) \\ \frac{d\boldsymbol{R}}{dt} &= \boldsymbol{\kappa} - D(\boldsymbol{R})\boldsymbol{\tau}^{\circ -1} - D(\boldsymbol{R})C^T\boldsymbol{n} + D^T D(\boldsymbol{l}\circ\boldsymbol{R})C^T\boldsymbol{n}\end{aligned} \tag{33}$$

where $\mathbf{1}$ is a column vector of ones of appropriate dimension, and $\circ$ denotes element-wise operation

The equations presented in section **??** can also be vectorized. In the following we use that the factor $\tilde{\kappa}_j = 1$ because in our simulations, all the resources are being supplied in the same amount. For instance, in equation **??**, since metabolic preferences are binary, taking the scalar product of the two preference vectors yields the number of common elements between them.

$$C_a = (1-l)CC^T \tag{34}$$

For equation **??**, only indices j and k can be vectorized, taking the form

$$(C_b)_{\alpha\beta} = l\,(\vec{c_\alpha} + \vec{c_\beta})^T\,D\,(\vec{c_\alpha}\circ\vec{c_\beta}) \tag{35}$$

The vectorization of equation **??** is expressed as

$$F = lCDC^T \tag{36}$$

The coalescence event presented in section **??** is simulated by mixing the species from each community after the assembly process in isolation has been completed. This can be mathematically expressed in matrix form through a system of $s_1 + s_2 + m$ differential equations, where s_1 and s_2 are the number of species that remain in the first and second communities after their assembly, respectively.

$$\frac{d\boldsymbol{n}_e}{dt} = \boldsymbol{g}\circ\boldsymbol{n}_e\,(D(\boldsymbol{R}_e)C_e(\mathbf{1}-\boldsymbol{l}_e) - \boldsymbol{m}_e) \tag{37}$$

$$\frac{d\boldsymbol{R}}{dt} = \boldsymbol{\kappa} - D(\boldsymbol{R})\boldsymbol{\tau}^{\circ -1} - [\mathbb{1}\ \mathbb{1}]\left(D(\boldsymbol{R}_e)C_e^T\boldsymbol{n_e} + D_e^T D(\boldsymbol{l_e}\circ\boldsymbol{R_e})C_e^T\boldsymbol{n}_e\right) \tag{38}$$

Where $[\mathbb{1}\ \mathbb{1}]$ represents the horizontal concatenation of two $m \times m$ identity matrices. The sub-index e stands for *extended*. To form any of the extended vectors in equation **??**, one simply vertically concatenates the vectors from community 1 and community 2. Constructing an extended matrix in equation **??** is done by joining the two matrices alongside their diagonal. For example, constructing the vector of species abundances $\boldsymbol{n}_e$ and the metabolic matrix D_e would be done as

$$\boldsymbol{n}_e = \begin{bmatrix}\boldsymbol{n}^{(1)} \\ \boldsymbol{n}^{(2)}\end{bmatrix} \qquad\qquad D_e = \begin{bmatrix}D^{(1)} & 0 \\ 0 & D^{(2)}\end{bmatrix} \tag{39}$$

where the superscripts indicate belonging to community 1 or 2

6.6 Table of parameter values

Table of parameter values!

bullcrap

The matrix element D_{jk} is constructed by adding two terms. First, the k^{th} element of a random vector following a flat Dirichlet distribution, $\vec{x_j} \sim \mathrm{Dir}(\vec{q})$ where all elements of the parameter vector $\vec{q}$ are equal to one, $q_k = 1$. In this case, the Dirichlet distribution is equivalent to a uniform distribution over the standard $(k-1)$-simplex, where $k \in \{1, \ldots, m\}$, where m is the total number of resources. The Dirichlet distribution has the property that each sampled vector sums to 1, making it a natural way of randomly allocating a fixed total quantity (such as the total secretion flux from a given input) (**?**). Second, a term that depends on the difference in demand between resources j and k, that is, on $d_{sj} - d_{sk}$, where s is the number of species (omited from now on to avoid cluttering the notation). Let Δ_{jk} denote the truncated difference between demands of resource j and k as

$$\Delta_{jk} = \begin{cases} d_k - d_j & \text{if } d_k > d_j \\ 1 & \text{if } d_k = d_j \\ 0 & \text{if } d_k < d_j \end{cases} \tag{40}$$

Then, the fraction of the leaked resource j that is released in the form of resource k is given by

$$D_{jk} = (1 - k_f)x_{jk} + k_f \frac{\Delta_{jk}}{\sum_j \Delta_{jk}} \tag{41}$$