Theoretical ecology without species

Mikhail Tikhonov
Center of Mathematical Sciences and Applications
School of Engineering and Applied Sciences and
Kavli Institute for Bionano Science and Technology, Harvard University, Cambridge, MA 02138, USA

Most of classical theoretical ecology is based on the assumption that organisms in a community can be naturally partitioned into groups of individuals that can be treated as identical. At the same time, mounting experimental evidence from studies of microbial communities raises the intriguing question whether this intuition is an accurate description of the microbial world. This work builds on Mac Arthur's model of competitive coexistence on multiple resources to construct a framework that does not rely on postulated existence of species as fundamental ecological variables. In one parameter regime, effective "species" with a core and accessory genome naturally appear in this model as emergent concepts. However, the same model allows a smooth transition to a highly diverse regime where the species formalism becomes inadequate. An alternative description is proposed based on the dynamical modes of population fluctuations. This approach provides a naturally hierarchical description of community dynamics which is well-defined even when the species description breaks down. The relevance of this framework for understanding the complexity of naturally observed microbial communities is discussed.

Although the basic unit participating in ecological interactions is an individual organism, constructing tractable theoretical models usually requires clustering individuals into discrete groups within which organisms are treated as identical; such classification can be based, for example, on taxonomy, development stage, or phenotype [1]. The resulting tension is a long-standing issue in community ecology: ever since Darwin [2], the difficulty of drawing sharp boundaries partitioning natural diversity into discrete categories [3] and the realization that the individual-level variation can be an important actor in ecological phenomena [4] made the "partitioned community" assumption of well-delimited, uniform groups highly problematic.

The urgency of this issue has been highlighted over the last decade, as advances in sequencing technology prompted a boom in the study of microbial diversity in natural environments [5–11]. In the microbial world, the partitioning problem is intensified by the prevalence of asexual reproduction and horizontal gene transfer [12, 13]. Nevertheless, the currently dominant view in the field is that the partitioned community assumption, while conceptually problematic, is operationally necessary [14–17]. Thus the uncertainty in defining classification criteria is treated merely as a caveat that one should keep in mind when using the familiar species-based perspective (in the broad sense of "species" as a basic unit of classification). Looking at micro-organisms through this lens, we find unprecedented diversity [8, 9, 11], a puzzling prevalence of "rare species" [18, 19], and overall the prospect of understanding these communities as interacting systems of hundreds of poorly characterized types appears rather daunting. But what if, as a thought experiment, we took the difficulty of drawing sharp boundaries very seriously? The species-based intuition is incontestably useful, but might it be forcing onto our data a structure that it does not possess [20, 21]?

Ultimately, of course, it is up to experimental evidence to settle this question, and reports are conflicting [22–26]. However, ideally we should be asking not whether a species-based picture is "adequate", but whether it is superior to alternatives. How could we describe an ecology where species did not exist? As long as species are an operational requirement for describing ecological dynamics, our language will itself be a barrier for exploring what could be a paradigm shift, an excitingly different perspective on the microbial world. Therefore, prior to becoming an experimental matter, it is first a question of constructing a theoretical framework.

In order to construct a setting where alternative analytical approaches could be tested, this work builds on Mac Arthur's model of competitive coexistence on multiple resources [27] to describe communities of organisms defined by a "functional genotype". In one parameter regime, effective "species" with a core and accessory genome emerge in this model as a natural description of surviving genotypes. However, the same model allows a smooth transition to a highly diverse regime where the species formalism becomes inadequate. An alternative description is proposed based on the dynamical modes of population fluctuations. This approach provides a naturally hierarchical description of community dynamics which is well-defined even when the species description breaks down. Within the model described here, the high-diversity regime is more simple dynamically, and it is intriguing to speculate that these results may suggest a more optimistic viewpoint on the complexity of naturally observed microbial communities.

Of course, the issues discussed above have a long history of thought in ecological literature (for a few recent reviews, see [17, 28–31]). Again in the context of microbial ecology, many authors argue for a holistic view of the community [32] and many approaches exist that do not explicitly require the partitioned community assumption, including the very premise of metagenomic characterizations [33, 34]. However, although such approaches

are frequently motivated by similar considerations, their aim is typically not to challenge the partitioned community picture, but to construct complementary, and often coarse-grained viewpoints that may be more appropriate or more convenient to address a particular question. Here, the goal is to investigate a scenario when a partitioned community description would be not merely inconvenient, but incorrect. Although individual-based modeling allows studying such species-less scenarios with simulations, no established theoretical framework currently exists.

I. A MODEL FOR DIVISION OF LABOR: METAGENOME PARTITIONING

To begin, consider the following model for division of labor in large communities. Its mathematical structure will be almost identical to that of Mac Arthur's model of competitive coexistence on multiple resources [27]; however, for the purposes of this work, the interpretation of what constitutes a "species" will need to be modified. To avoid confusion, the model will be defined *de novo*, with notation and interpretation appropriate for this discussion. The exact mapping onto the notations and terminology of Mac Arthur is detailed in the Supplementary Material (SM).

Consider a world with N resources $i \in \{1 \dots N\}$ denoted A, B, etc. These resources can be harvested with "pathways" P_i. An organism is defined by its "functional type", namely the pathways that it carries. Below, the term "functional type" will be preferred to loaded terms "genotype" or "phenotype", since in the context of this model inheritance and evolutionary dynamics will not be considered explicitly. There are $2^N - 1$ possible functional types; they will be denoted using a binary vector of pathway presence/absence: $\vec{\sigma} = \{1, 1, 0, 1, \dots\}$, or by a string listing all resources it can harvest, e.g. "organism $\underline{ABD\dots}$" (the underline distinguishes specialist organisms such as $\underline{A}$ from the resource they feed on, in this case A). Let $n_{\vec{\sigma}}$ be the total number of organisms $\vec{\sigma}$ in the population. The total benefit R_i from a resource i is equally distributed among all organisms carrying the pathway P_i; their number will be denoted T_i:

$$T_i \equiv \sum_{\text{all } \vec{\sigma} \text{ carrying } i} n_{\vec{\sigma}}.$$

For every organism, its individual rate of replication or death is determined by its *resource surplus* $\Delta\varphi$:

$$\Delta\varphi_{\vec{\sigma}} = \sum_i \sigma_i \frac{R_i}{T_i} - \chi_{\vec{\sigma}}. \tag{1}$$

Here the first term is the total benefit harvested by all carried pathways, and the second term represents the maintenance costs of organism $\vec{\sigma}$; these will be discussed shortly. This abstract model might, for example, describe carbon-limited growth of a community of organisms in a well-mixed environment supplied with N different sugars at rates R_i per unit time. P_i is the pathway that allows a microorganism to metabolize a given sugar.

Note that, perhaps counter-intuitively, the starting point for our discussion is not yet a community of unique individuals. Although that is our final destination, to demonstrate continuity with previous ideas the entire argument can be explained as a new perspective on a familiar model, and the generalization to a truly individual-based description will be discussed later.

The resource surplus $\Delta\varphi$ is used to generate biomass. For simplicity, the biomass of an organism can be equated with the number of pathways it carries $|\vec{\sigma}| \equiv \sum_i \sigma_i$. The total biomass change due to resource-dependent growth or death is then given by:

$$|\vec{\sigma}| \frac{dn_{\vec{\sigma}}}{dt} = n_{\vec{\sigma}} \Delta\varphi_{\vec{\sigma}}. \tag{2}$$

If one were to introduce a mutation rate for loss/acquisition of a pathway, this would become a rich dynamical model of "mesoscopic" population genetics for a population of bacteria evolving through horizontal gene transfer. For the purposes of this work, however, we can ignore dynamical questions and focus on the "final equilibrium state", for a given set of costs $\chi_{\vec{\sigma}}$. This is the state to which the community converges in the infinite-time limit, assuming a weak exposure to an external pool of all possible organism types. At such an equilibrium, two conditions must be satisfied: first, all organisms that are present in the community must have zero resource surplus, so that they neither grow nor die. Second, the community must be stable with respect to invasion by organism types that are currently absent: all such organisms must therefore have negative resource surplus. Such an equilibrium always exists and is stable; this is because the dynamics (2), as noted in Ref. [27], can be seen as a gradient-ascent-optimization of the objective function (see SM):

$$F = \sum_i R_i \ln T_i - \sum_{\vec{\sigma}} \chi_{\vec{\sigma}} n_{\vec{\sigma}}.$$

To gain intuition, consider the case of $N = 2$ (Fig. 1). With 2 resources, there are only 3 organism types: $\underline{A}$, $\underline{B}$ and $\underline{AB}$, and the dynamics (2) reduce to the following

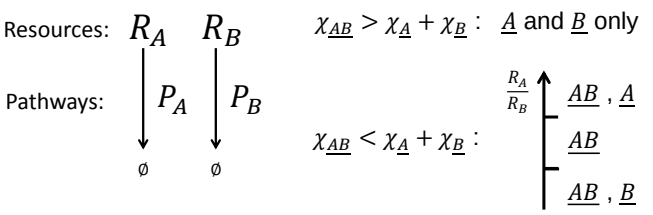

FIG. 1. Equilibria of the metagenome partitioning model for $N = 2$.

three equations:

$$\dot{n}_{\underline{A}} = n_{\underline{A}} \left(\frac{R_A}{n_{\underline{A}} + n_{\underline{AB}}} - \chi_{\underline{A}} \right)$$
$$\dot{n}_{\underline{B}} = n_{\underline{B}} \left(\frac{R_B}{n_{\underline{B}} + n_{\underline{AB}}} - \chi_{\underline{B}} \right)$$
$$\dot{n}_{\underline{AB}} = \frac{1}{2} n_{\underline{AB}} \left(\frac{R_A}{n_{\underline{A}} + n_{\underline{AB}}} + \frac{R_B}{n_{\underline{B}} + n_{\underline{AB}}} - \chi_{\underline{AB}} \right)$$

Analysing these equations, one finds that the final community structure is determined primarily by the cost difference between the specialists and the generalist (see SM). If $\chi_{\underline{AB}} > \chi_{\underline{A}} + \chi_{\underline{B}}$, the generalist $\underline{AB}$ is not competitive and the final state is a community of two specialists $\underline{A}$ and $\underline{B}$. Alternatively, if $\chi_{\underline{AB}} < \chi_{\underline{A}} + \chi_{\underline{B}}$, the dominating type will be $\underline{AB}$, possibly supplemented by either $\underline{A}$ or $\underline{B}$ if there is a sufficient excess of the respective resource. This example demonstrates that the metagenome partitioning model captures the idea of division of labor in a community. Intuitively, some functional capabilities are more compatible that others. If two metabolic enzymes require different conditions for optimal function, or if their enzymatic activity is characterized by an undesirable cross-talk, expressing them in the same organism would require maintaining two separate compartments at an extra cost; for instance, the nitrogen-fixing enzyme nitrogenase is inactivated by oxygen, and as a result, maintaining oxygen respiration and nitrogen fixation functional simultaneously in the same organism would be extremely costly and is never observed. Conversely, an organism can efficiently make use of several enzymes requiring similar specific conditions (e.g. specific levels of pH) by investing into the maintenance of a dedicated compartment only once, but harvesting the benefit from all.

Motivated by these examples, one can complete the model by defining organisms costs $\chi_{\vec{\sigma}}$ based on their enzyme content $\vec{\sigma}$ as follows:

$$\chi_{\vec{\sigma}} = \chi_0 + |\vec{\sigma}| + \sum_{ij} J_{ij} \sigma_i \sigma_j. \tag{3}$$

The cost of an organism is composed of the baseline cost χ_0 of maintaining basic structures (e.g. cell wall, ribosomes, replication machinery), a fixed cost per carried pathway, and a correction that arises from (positive or negative) interactions between pathways, given by an $N \times N$ matrix J_{ij}. For simplicity, J_{ij} is taken to be a constant random matrix of Gaussian elements characterized by mean $\langle J \rangle = \mu$ and variance $\langle (J-\mu)^2 \rangle = J_0^2$. Since one expects cross-talk between enzymes to be detrimental more often than beneficial, μ will be positive. In this simple model, all the complexity of the enzymatic chemistry and properties of the environment is summarized in a random matrix J_{ij} with two parameters. For simplicity, this work will assume all the resources to be supplied in equal abundance: $R_i \equiv R$. Note that resource amount is not a dynamic variable here: the resource flux is always fully consumed by the community. However, as organism $\vec{\sigma}$ multiplies, this increases the total expression T_i of the pathways it carries (for all i such that $\sigma_i = 1$), and therefore the benefit that any one organism can harvest from the corresponding resources R_i/T_i is reduced. In this sense, one can say that, for example, organism $\underline{AC}$ "depletes" resources A and C, because its presence reduces the benefit that other organisms can obtain from these particular resources.

II. THE METAGENOME PARTITIONING MODEL LEADS TO EMERGENT SPECIES WITH A CORE AND ACCESSORY GENOME.

Which of the $2^N - 1$ organism types survive in the final equilibrium state of the community? In traditional evolutionary models, an important role is played by "fitness" of different organisms. In the setting described here, it is natural to define the fitness of an organism as its initial growth rate (at $n_{\vec{\sigma}} = 1$) in a pristine environment, i.e. with no other organisms present. In this situation the total resource harvest collected by $\vec{\sigma}$ is given by $\sum_i \sigma_i R_i = |\vec{\sigma}| R$, and therefore:

$$f_{\vec{\sigma}} = \frac{\Delta\varphi_{\vec{\sigma}}}{|\vec{\sigma}|} = \frac{|\vec{\sigma}| R - \chi_{\vec{\sigma}}}{|\vec{\sigma}|} = R - \frac{\chi_{\vec{\sigma}}}{|\vec{\sigma}|}.$$

We see that the relative fitness of different types is determined by their cost per pathway. Fig. 2A shows the fitness of all organism types for one realization of costs $\chi_{\vec{\sigma}}$ that will be used throughout this work, generated for $N = 20$, $R = 100$, $\chi_0 = 0.9$ and one random realization of the matrix J_{ij}. The matrix J_{ij} was obtained by generating a Gaussian random matrix with $\mu = 0.05$ and $J_0 = 0.05$, and applying a "neighbor bias", shifting elements immediately above and below diagonal $J_{i,i\pm1}$ by a constant $b = -0.15$ (not reflected in Eq. (3)). This favors organism types that express consecutive pathways and was done to improve visual clarity of subsequent figures.

Examining Fig. 2A, one can see that extreme specialists ($|\vec{\sigma}| = 1$) have reduced fitness; this is due to the baseline cost term χ_0 in Eq. 3. As the number of expressed pathways $|\vec{\sigma}|$ is increased, the relative importance of the baseline cost goes down; however, the competitiveness of wide-spectrum generalists ($|\vec{\sigma}| \approx N$) is affected by pathway cross-talk which is, on average, detrimental ($\mu > 0$). The parameters above were chosen to be representative of the regime where the fitness peak $|\vec{\sigma}| = k^*$ is located at an intermediate organism size, in this case $k^* \approx 5$.

For these particular costs $\chi_{\vec{\sigma}}$, the final population state, determined numerically (see SM), is composed of 11 organism types in stable coexistence (Fig. 2B). These types are not necessarily the ones characterized by maximal fitness in an empty environment: as organisms multiply, they modify their environment by depleting resources, altering the fitness landscape for other organism types. For example, note that one of the domi-

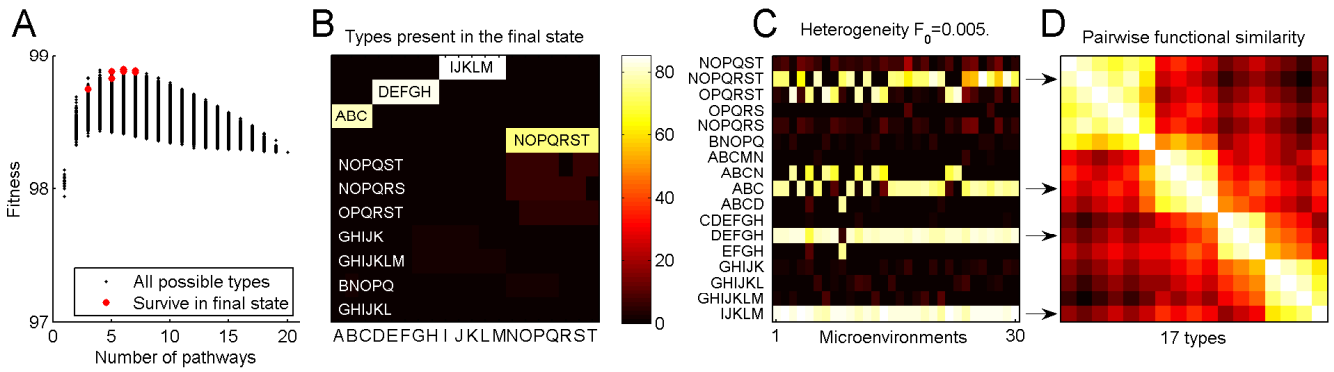

FIG. 2. **The metagenome partitioning model leads to emergent species with a core and accessory genome. A.** Fitness of all $2^N - 1$ organism types, defined as the growth rate in a pristine environment with equally abundant resources $R_i \equiv R = 100$, for one particular random realization of the J_{ij} cost model used throughout the text. The organisms that survive in the final state (red) tend to be selected among high-fitness organisms, but not exclusively (cf. type $\underline{ABC}$), because organisms modify their environment and therefore fitness of other organisms. **B.** The choice of costs in A gives a community with 11 types. Plot shows the pathway content of these types (also provided in labels); color reflects the stable abundance of each type in the population. **C.** A heterogeneous habitat can be thought of as composed of multiple independent micro-environments, each of which may favor different organism types. Here, color reflects the abundance of the specified organism type (rows) in a given micro-environment (columns) for a habitat composed of 30 micro-environments with weak heterogeneity $F_0 = 0.005$. Shown are the 17 types whose biomass exceeds 0.1% of the total. Color scale as in B; rows are ordered to optimize the clustering pattern in the next panel. **D.** A matrix of pairwise Hamming distances between the 17 types of panel C. Organisms form 4 clear groups, or effective "species" with core and accessory genomes.

nant established types (organism $\underline{ABC}$) has a comparatively low fitness when placed in a pristine environment (Fig. 2A); however, it thrives once the other 3 dominant types strongly deplete all resources except A, B and C. In general, the fate of an organism in an ecosystem is determined both by its own fitness and by its interaction with other organisms, in an environment that is modified by their presence. The metagenome partitioning model captures this in a simple setting with few parameters.

How sensitively does the final equilibrium of the community depend on the exact values of $\chi_{\vec{\sigma}}$? In general, any habitat can be thought of as a collection of microenvironments whose characteristics exhibit some degree of heterogeneity. This heterogeneity can be described as a modification of organism costs:

$$\chi'_{\vec{\sigma}} = \chi_{\vec{\sigma}} + H_{\vec{\sigma}}.$$

The magnitude of $H_{\vec{\sigma}}$ compared to the parameters of our model (χ_0, μ, J_0 and the neighbor bias b) characterizes how strongly the tendency of functional traits to associate in an organism is modified across microenvironments. In general, there is no reason to assume $H_{\vec{\sigma}}$ to be small. For example, in *E. Coli*, losing the capacity to produce a metabolically expensive aminoacid such as arginine has a fitness effect that is highly dependent on the environment and varies from strongly deleterious to strongly beneficial when a single environmental characteristic, namely the free concentration of the deficient aminoacid, is modified [35, 36]. Further, even if $H_{\vec{\sigma}}$ is indeed small compared to the relevant parameters of the model, its effect on the final population state can still be significant: intuitively, to have no effect, $H_{\vec{\sigma}}$ should be small compared not to $\chi_{\vec{\sigma}}$ itself, but to differences between $\chi_{\vec{\sigma}}$. Since there are exponentially many organism types, surprisingly small $H_{\vec{\sigma}}$ can modify competition outcome.

Consider first the case when environment heterogeneity is weak. Its effect on organism costs can be translated into a more intuitive effect on fitness; for this, one can set $H_{\vec{\sigma}} = |\vec{\sigma}|F_{\vec{\sigma}}$ and then model $F_{\vec{\sigma}}$ by independently drawing values from a Gaussian distribution of width F_0 for each micro-environment. In other words, environment heterogeneity has the effect of adding a small random contribution to the fitness of all organism types.

Fig. 2C shows the final equilibria for 30 microenvironments at $F_0 = 0.005$ (shown are all the types with significant presence in the habitat, whose biomass combined over all micro-environments exceeds 0.1% of the total). Although F_0 is an order of magnitude smaller than any other parameter in the model, even this small perturbation is often sufficient to change which organisms survive in the final state. If the habitat is sampled on a scale that cannot distinguish between individual micro-environments, all the organism types appearing in Fig. 2C will be observed. Note, however, that these 17 types cluster into just four groups (Fig. 2D), with members of each group exhibiting strong functional similarity to each other. Observing this in an experiment, we would characterize this habitat as harboring 4 "species" of organisms that share a "core genome", differing only in a small number of pathways, the "accessory genome" [37].

Intuitively, this clustering arises for two reasons. Consider a particular organism type $\vec{\sigma}_0$. First, under the J_{ij} cost model, if $\vec{\sigma}_0$ has high fitness (low cost per en-

zyme), i.e. is made primarily of pathways that "go well together", then the fitness of its neighbors in functional space (with one pathway added or removed) will, on average, also be high. Second, the resources not depleted by other organisms shape a niche that favors types with pathway content similar to $\vec{\sigma}_0$. Therefore, if cost modification causes type $\vec{\sigma}_0$ to be displaced, the organism displacing it will likely be one of its high-fitness neighbors that shares largely the same niche. As a result, the metagenome partitioning model contains a regime where species with a core and accessory genome arise as an emergent phenomenon. The expected number M of these effective species can be estimated as the number of high-fitness types ($|\vec{\sigma}| \approx k^*$) that can tile the resource space and therefore coexist at strong abundance in a single micro-environment: $M \approx N/k^*$, which is indeed 4 in this case.

Note that Fig. 2C shows only relatively abundant organisms: the 0.1% biomass threshold eliminates 7 additional types that appear at a low abundance in a single or few microenvironments and thus have a very weak relative presence in the habitat as a whole (Fig. S1). Thus the model described here also naturally includes a previously suggested explanation for the "rare species" phenomenon as a consequence of environment heterogeneity [19]. Note also that setting the number of microenvironments to 30 was an arbitrary choice. If more micro-environments were sampled, new types would occasionally be observed; however, the dominant 4-cluster structure would remain intact (Fig. S2). Importantly, therefore, the interpretation that "the habitat harbors 4 species" remains invariant, while the exact number of types observed in micro-environments can change.

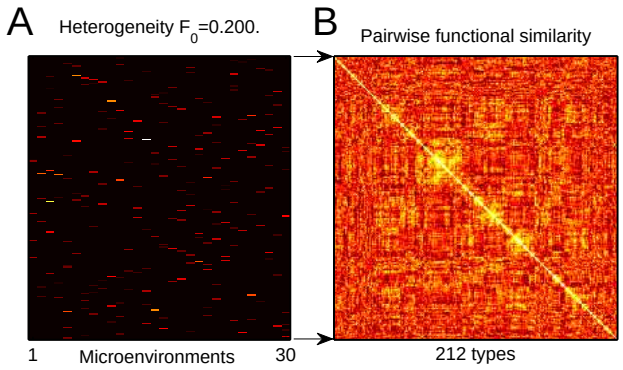

FIG. 3. **The high-diversity regime: species clustering is lost.** Same as Figs. 2CD for a larger heterogeneity parameter $F_0 = 0.2$. **A.** For very large habitat homogeneity, each micro-environment supports different organism types. Axes as in Fig. 2C. Shown are 212 types that exceeded the 0.1% biomass threshold (not labeled to reduce clutter). **B.** The matrix of Hamming distances between 212 types in the habitat no longer possesses a clustered structure.

III. THE MODEL ADMITS A SMOOTH TRANSITION INTO A HIGH-DIVERSITY REGIME WHERE SPECIES CLUSTERING IS LOST.

Consider now the case of large heterogeneity. As F_0 is increased, the clustering of types is progressively reduced. Fig. 3 repeats panels Fig. 2CD with F_0 increased to 0.2. At this heterogeneity parameter, each microenvironment favors its own set of organism types with hardly any overlap between microenvironments; as a result, diversity in the habitat is vastly increased: 30 microenvironemnts now harbor 212 types that exceed the 0.1% total biomass threshold. These types no longer display a clustered structure (Fig. 3B). Whether or not this is an adequate representation of any real habitat harboring a diverse microbial community, this setting allows us to ask a novel type of question: how many species are there in the ecosystem displayed in Fig. 3? Clearly, 212 would be an incorrect answer, just as for the ecosystem displayed in Fig. 2CD the number of effective species was 4 rather than 17. This time, however, no underlying structure of the problem justifies any clustering procedure. The "how many species?" question has no answer in this regime; instead, the "species"-based description is no longer adequate.

How, then, should one characterize this ecosystem? One would like to construct an alternative viewpoint that would be equivalent to counting species when they are well-defined, but would remain applicable even when "species" no longer exist. Note that this is a different task than inventing another metric for characterizing the functional diversity of a community; numerous such metrics have been proposed in the literature [38]. In the problem at hand, we are content with the simplest such metric, namely the number of species; the difficulty lies in, first, defining this quantity in the setting described here, and second, extending it to regimes were species are no longer well-defined.

IV. DYNAMICAL MODES AS AN ALTERNATIVE TO SPECIES

The solution proposed here is to consider the spectrum of dynamical modes in the system.

Consider a community at equilibrium with respect to the deterministic growth/death dynamics (2). Denoting $\vec{n}$ the vector of abundances of all $2^N - 1$ organism types, so that $n_\alpha \equiv n_{\vec{\sigma}_\alpha}$ (using Greek indices to label distinct organism types), one can write these dynamics as

$$\frac{d\vec{n}}{dt} = \vec{g}(\vec{n}), \tag{4}$$

where

$$g_\alpha(\vec{n}) = \frac{1}{|\vec{\sigma}_\alpha|} n_\alpha \Delta\varphi_{\vec{\sigma}_\alpha}.$$

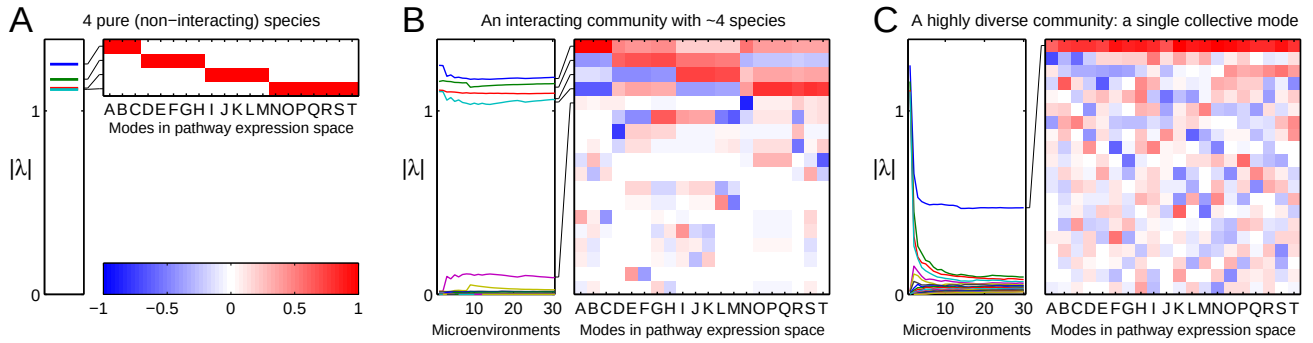

FIG. 4. **The spectrum of dynamical modes quantifies functional diversity. A.** Eigenvalue spectrum (left) and the projection of eigenmodes into the pathway expression space, for a non-interacting community of 4 "pure" species (the 4 dominant types from Fig. 2B). The thin black lines link each mode with the matching eigenvalue in the spectrum. **B.** Left: the eigenvalue spectrum for a community with $F_0 = 0.005$ (cf. Fig. 2) accumulated over K micro-environments, presented as a function of K. Right: the eigenmodes of the entire community ($K = 30$) projected into the pathway space, ordered by decreasing $|\lambda|$. Red, positive components; blue, negative components. The matching eigenvalue is shown for top 5 modes. Note the similarity of the top 4 modes with those of panel A. **C.** Same as B, for the high-diversity regime at large environmental heterogeneity ($F_0 = 0.2$; Fig. 3). Only one mode has a large $|\lambda|$ and corresponds to a collective whole-community response.

This time we will not require that no other type can invade, but only that the abundance of all present organism types had equilibrated and every one of them has zero resource surplus. In other words, we will consider a fixed point of the dynamics (4).

One can now perform the stability analysis around this fixed point, i.e. consider the eigenmodes of the matrix

$$M_{\alpha\beta} = \frac{\partial g_\alpha}{\partial n_\beta}. \tag{5}$$

This matrix characterizes interactions between organism types: $M_{\alpha\beta}$ is the change in growth rate of type α induced by a perturbation in the abundance of type β. Considering separately the types that are present (denote their set S) and those that are absent, one can write:

$$M_{\alpha\beta} = \begin{cases} -\dfrac{n_\alpha}{|\alpha|} \displaystyle\sum_{i\in\alpha\cap\beta} \frac{R_i}{T_i^2} & \text{if } \alpha \in S \\ \dfrac{1}{|\alpha|}\delta_{\alpha\beta}\Delta\varphi_\alpha & \text{if } \alpha \notin S, \end{cases} \tag{6}$$

where $i \in \alpha \cap \beta$ denotes pathways present in both types α and β, and δ is the Kronecker symbol.

One finds that some fluctuation modes are unstable (have positive eigenvalues); these correspond to introducing an organism type that can invade the community ($\alpha \notin S$ and $\Delta\varphi_\alpha > 0$). Once such a type arrives, the community will switch to another temporary equilibrium. Consider these temporary equilibria, restricting dynamics only to the set of organisms that are present:

$$M'_{\alpha\beta} \equiv M_{\alpha\beta}|_{n_\alpha>0, n_\beta>0} = -\frac{n_\alpha}{|\alpha|} \sum_{i\in\alpha\cap\beta} \frac{R_i}{T_i^2}. \tag{7}$$

Within this space, all fluctuation modes are stable (all eigenvalues λ are negative) by the objective function optimization argument (see SM). Borrowing the term introduced in [39], one can call them "ecomodes". Below, it is proposed that the appropriate extension of species in this setting is to consider the most stable ecomodes, i.e. ecomodes with strongly negative $|\lambda|$.

To motivate this, consider first a community composed of just the 4 dominating types of Fig. 2B inhabiting a single microenvironment. These organisms consume non-overlapping sets of resources and do not interact; the interaction matrix $M'_{\alpha\beta}$ is therefore diagonal. For a community composed of "pure" (non-interacting) species, the ecomodes are in one-to-one correspondence with the individual species (Fig. 4A): the eigenvalues $\lambda_i = M'_{\alpha\alpha}$ are given by their cost per enzyme (left panel), and the eigenmodes are set by their pathway content (right panel).

After this simplest example, consider now the entire habitat presented in Fig. 2CD. After summing over several micro-environments, the total community is no longer an equilibrium of any one system. K micro-environments are K independent systems, each with resource abundance R and costs $\chi'_{\vec{\sigma}} \approx \chi_{\vec{\sigma}}$. Consider this as a community of organisms living all together in a habitat with total resources KR. Although this community is no longer at equilibrium, perturbing the abundance of any one type still exerts a well-defined influence on the growth/death rates of other types. Therefore, the interaction matrix $M'_{\alpha\beta}$ can still be defined using Eq. (5), and one can again consider its spectrum (for details, see SM). Fig. 4B present the spectrum computed for the community progressively summed over K micro-environments; from $K = 1$ (a single micro-environment) to $K = 30$ (the entire habitat pooled together). One can see that the spectrum dependence on K quickly stabilizes and is dominated by 4 eigenvalues (Fig. 4B, left). Mapping the eigenmodes into the pathway expression space reveals a complex pattern of interactions (Fig. 4B, right); however,

the structure of the dominant 4 ecomodes bears a clear resemblance to non-interacting case (Fig. 4A) and is determined by the core "genomes" of the dominant organism types. In one of the previous sections, this habitat was characterized as harboring 4 species; however, this interpretation relied on a subjective judgement of Fig. 2D by a human observer. In contrast, the eigenmode spectrum provides an objective description: 4 dominant modes are unambiguously identified.

If the eigenvalue threshold is lowered considerably, a fifth mode becomes apparent. Examining the right panel of Fig. 4B, one can see that this mode corresponds to the competition between types $\underline{NOPQRST}$ and $\underline{OPQRST}$ (compare with Fig. 2C, rows 2 and 3). These two types have similar total abundance in the habitat and only differ by one pathway; in the first approximation, the immediate response of the community to most external perturbations (e.g. fluctuation in resource abundance) will not distinguish between them; however, the difference will become apparent at a longer time scale. Thus, the eigenmode description of the community structure is naturally hierarchical.

Finally, consider the high-diversity regime with 212 types of Fig. 3. By construction, the first micro-environment and the community it harbors is identical to that of Fig. 4B at $K = 1$; therefore, initially the spectrum has the same 4 large eigenvalues. However, as the community is pooled over highly heterogeneous micro-environments, the eigenvalues "crash" (Fig. 4C, left). The spectrum again stabilizes with respect to K, but in a very different regime where it is dominated by a single collective mode (Fig. 4C, right).

V. DYNAMICAL MODES IN AN INDIVIDUAL-BASED DESCRIPTION

The last remaining step of the argument is to show that the eigenmode structure can in fact be defined in a community of unique individuals and does not require assigning them to any groups such as functional types. To see this, one only needs to remark that the exact same procedure allows constructing an interaction matrix between each pair of individuals. First, one must translate the continuous deterministic dynamics (4) into a stochastic dynamics at the level of discrete individuals: each individual has a rate r at which it will either generate a copy of itself or die, so that its expected "abundance" after time dt is $1 \mapsto 1 + r\,dt$. This rate r is determined by the resource surplus experienced by that particular individual. Now, for each individual μ, consider the effect its removal would have on the instantaneous growth rates r of all other individuals ν in the population. This defines the interaction matrix:

$$\mathcal{M}_{\mu\nu} = -\frac{1}{|\mu|}\sum_{i\in\mu\cap\nu}\frac{R_i}{T_i^2}. \tag{8}$$

Unlike Eq. (7), this expression no longer involves any quantities whose definition requires a pre-defined classification (such as n_α, the "abundance of functional type α"), but only a functional characterization of each individual.

For a population of $\mathcal{N}$ individuals, the interaction matrix $\mathcal{M}$ of size $\mathcal{N}\times\mathcal{N}$ will always have $\mathcal{N}$ eigenmodes. Note, however, that if individuals μ and μ' happen to be functionally identical (or very close), then $\mathcal{M}_{\mu\nu} = \mathcal{M}_{\mu'\nu}$ for all ν, and $\mathcal{M}$ has a zero eigenvalue (or, respectively, a very small one). In particular, if individuals were drawn from a limited number K of distinct functional types, as considered here, then $\mathcal{N} - K$ eigenvalues will be exactly zero, and the remaining ones will coincide with the spectrum of $M_{\alpha\beta}$ defined for functional types above. For the community presented on Fig. 4B, the individual-based analysis will show that the community has 4 dominant eigenmodes (the four "species"), 1 smaller eigenvalue reflecting the most important split of one species into two equiabundant competing types, and 12 more (for a total of 17) reflecting the remainder of the internal structure of the 4 clusters. If the level of functional characterization of individuals were more detailed, the remaining eigenvalues would not be zero, but progressively smaller, resolving finer and finer details. The dynamical mode formalism thus provides a naturally hierarchical description for a community of unique individuals. Their grouping into types, when it exists, is established as a result of this analysis; in this framework, it no longer constitutes a fundamental assumption.

VI. DISCUSSION

This work considered the question: can we imagine ecology that is fundamentally *not* a system of interacting species? Constructing a model combining ecology and evolution [40] made it possible to avoid postulating the existence of species as fundamental ecological variables. As demonstrated here, in one parameter regime, species with a core and accessory genome naturally appear in our model as emergent concepts. However, the same model allows a smooth transition to a highly diverse regime where species become an inadequate description. Its dynamics are qualitatively different, exhibiting a single dominating collective mode, as opposed to K interacting modes characteristic of a K-species community. Importantly, both regimes admit a unified, naturally hierarchical description in terms of dynamical modes of community fluctuations that retains all the fine aspects of dynamics and is still meaningful even when the species description breaks down.

The specific example constructed here was intentionally simplified. Species may be well-defined (in terms of clustering of types) without being weakly interacting as in Fig. 2. Further, the high-diversity regime of Fig. 3 was constructed by postulating that any combination of pathways constitutes a possible organism; for any such com-

bination there exists a micro-environment where it is has the highest fitness; and moreover, such an environment is just as likely as any other. In such conditions there is of course no reason for types to cluster into anything that resembles species. The two regimes highlighted here are the opposing extremes (very weak and very strong interaction) and neither is proposed to be an accurate representation of reality. The point, rather, is that a model capable of smoothly interpolating between the two extremes allows us to begin constructing a framework that escapes the confines of the species paradigm. In the interest of clarity of this proof-of-principle argument, one fixed set of parameters of our model was used throughout this paper. Simple arguments demonstrate that the existence of the two regimes highlighted on Figs. 2 and 3 is a generic phenomenon for a wide range of model parameters (see SM). In between these two extremes lie other dynamical regimes not discussed here which may be closer to reality and deserve further investigation.

The dynamics considered in this work were deterministic, and the questions of organism dispersal across micro-environments constituting a habitat was not considered explicitly; these are important simplifying assumptions (see, for example, [41]). Relaxing them can be expected to push away from the neat clustering observed in Fig. 2D, but if so, this only lends more weight to the argument that ecology need not always look like interactions between well-defined species. Another strong simplification was the purely additive dependence of organisms on resources: combinatorial dependence ("and" instead of "or") is absent from the model. This simplification made it possible to capture division of labor in the simplest possible theoretical setting. In real ecological settings, combinatorial effects lead to a proliferation of niches and complexity and cannot be neglected.

The analysis presented here focused on infinite-time final equilibrium states of interacting communities, ignoring all questions about convergence to such equilibria, in particular those associated with the inherently slow sampling of the exponentially large space of possible functional types. Dynamical mechanisms such as inheritance and common descent may provide an independent reason for the partitioned community assumption to hold: even if taxonomic classification for microorganisms is globally problematic, for a given community, it is often appropriate to group together organisms that shared a common ancestor recently as opposed to a million years ago. Virtually all of the literature devoted to the "species problem" justifiably places inheritance and evolutionary dynamics at the center of the discussion. By deemphasizing these much-discussed mechanisms, this work focused on the purely functional aspect: in the community depicted on Fig. 2, types cluster into species not because they share recent ancestors, but because only certain combinations of pathways are competitive. One advantage of this complementary approach is the clear separation between the evolutionary sense of the term "species" and the ecological assumption of a "partitioned community"; as stressed in the introduction, the two are conceptually quite distinct.

Deemphasizing inheritance also lends generality to the argument: for example, it becomes possible to establish a link to the game theory literature. In this field, problems in economics, computer science and biology are modeled as "games" between agents choosing between a limited number of strategies, e.g. "cooperator" / "defector". The species problem as considered in this work could be reformulated as asking whether this limited-strategy description always provides an adequate model for the (in principle, unbounded) spectrum of possible agent behaviors.

More work is required before the dynamical mode description becomes a full-scale theoretical framework capable of challenging the traditional species-based formalism. Nevertheless, the results presented here argue for the possibility that in our quest to understand the microbial communities that shape our environment and our health, a key missing element could be theoretical [42].

VII. ACKNOWLEDGMENTS

I thank Ariel Amir, William Bialek, Simon A. Levin, Anne Pringle, Ned S. Wingreen, and particularly Michael P. Brenner for helpful discussions. This work was supported by the Harvard Center of Mathematical Sciences and Applications and the Simons Foundation.

[1] *Theoretical Ecology: Principles and Applications* eds. May R, McLean A (Oxford University Press, 2007), 272pp.

[2] Ereshefsky M (2010) Darwin's solution to the species problem. Synthese 175:405–425.

[3] Hey J (2006) On the failure of modern species concepts. *Trends Ecol Evol* 21:447–450.

[4] DeAngelis DL, Mooij WM (2005) Individual-based modeling of ecological and evolutionary processes. *Annu Rev Ecol Evol Syst* 36: 147–168.

[5] Gill SR et al. (2006) Metagenomic analysis of the human distal gut microbiome. *Science* 312(5778): 1355–1359.

[6] Turnbaugh PJ et al. (2009) A core gut microbiome in obese and lean twins. *Nature* 457(7228): 480–484.

[7] Caporaso JG et al. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A* 108 Suppl 1: 4516–4522.

[8] Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R (2012) Diversity, stability and resilience of the human gut microbiota. *Nature* 489: 220–230.

[9] Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. *Nature* 486(7402): 207–214.

[10] Human Microbiome Project Consortium (2012) A framework for human microbiome research. *Nature* 486(7402): 215–221.

[11] Gilbert JA, Jansson JK, Knight R (2014) The Earth Microbiome project: successes and aspirations. *BMC Biology* 12: 69.

[12] Staley JT (2006) The bacterial species dilemma and the genomicphylogenetic species concept. *Philos Trans R Soc Lond B Biol Sci* 361(1475): 1899–1909.

[13] Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP (2009) The bacterial species challenge: making sense of genetic and ecological diversity. *Science* 323(5915): 741–746.

[14] Sites JW Jr, Marshall JC (2004) Operational criteria for delimiting species *Annu Rev Ecol Evol Syst* 35: 199–227.

[15] Cohan FM (2006) Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. *Philos Trans R Soc Lond B Biol Sci* 361(1475): 1985–1996.

[16] Cohan FM, Perry EB (2007) A systematics for discovering the fundamental units of bacterial diversity. *Curr Biol* 17(10): R373–386.

[17] Shapiro BJ, Polz MF (2014) Ordering microbial diversity into ecologically and genetically cohesive units. *Trends microbiol*, 22(5): 235–247.

[18] Sogin ML et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A* 103(32): 12115–20.

[19] Lynch MDJ, Neufeld JD (2015). Ecology and exploration of the rare biosphere. *Nature Rev Microbiol* 13: 217–229.

[20] Hendry AP, Vamosi SM, Latham SJ, Heilbuth JC, Day T (2000) Questioning species realities. *Conservation Genetics* 2000, 1(1): 67–76.

[21] Doolitle WF (2012) Population genomics: how bacterial species form and why they don't exist. *Curr Biol* 22(11): R451–3

[22] Acinas SG et al. (2004) Fine-scale phylogenetic architecture of a complex bacterial community *Nature* 430: 551-554.

[23] Hunt DA, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF (2008) Resource positioning and sympatric differentiation among closely related bacterioplankton. *Science* 320: 1081–1085.

[24] Shapiro DJ et al. (2012) Population genomics of early events in the ecological differentiation of bacteria. *Science*, 336: 48–51.

[25] Kashtan N et. al. (2014) Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus. Science* 344(6182): 416–420.

[26] Biller SJ, Berube PM, Lindell D, Chisholm SW (2015) *Prochlorococcus:* the structure and function of collective diversity. *Nature Rev Microbiol* 13: 13-27.

[27] Mac Arthur R (1969) Species packing, and what interspecies competition minimizes. *Proc Natl Acad Sci U S A* 64(4): 1369–71.

[28] De Queiroz K (2007) Species concepts and species delimitation. *Syst Biol* 56: 879–886.

[29] Achtman M, Wagner M (2008) Microbial diversity and the genetic nature of microbial species. *Nature Rev Microbiol* 6: 431–440

[30] Hart MW (2011) The species concept as an emergent property of population biology. *Evolution* 65(3): 613-616.

[31] Cordero OX, Polz MF (2014) Explaining microbial genomic diversity in light of evolutionary ecology. *Nature Rev Microbiol* 12: 263–273.

[32] Konopka A (2009) What is microbial community ecology? *ISME J* 3: 1223–1230.

[33] Tringe SG et al. (2005) Comparative metagenomics of microbial communities. *Science* 308(5721): 554–557.

[34] Vieites JM, Guazzaroni ME, Beloqui A, Golyshin PN, Ferrer M (2009) Metagenomics approaches in systems microbiology. *FEMS Microbiol Rev* 33(1): 236–255.

[35] Mee MT, Collins JJ, Church GM, Wang HH Syntrophic exchange in synthetic microbial communities. *Proc Natl Acad Sci U S A* 111(20): E2149–56.

[36] Pande S et al. (2014) Fitness and stability of obligate cross-feeding interactions that emerge upon gene loss in bacteria. *ISME J* 8: 953–962.

[37] Medini D et al. (2008) Microbiology in the post-genomic era. *Nature Rev Microbiol* 6: 419–430.

[38] Petchey OL Gaston KJ (2006) Functional diversity: back to basics and looking forward. *Ecology Lett* 9: 741–758.

[39] Hekstra DR, Leibler S (2012) Contingency and statistical laws in replicate microbial closed ecosystems. *Cell* 149(5): 11640–73.

[40] Schoener TW (2011) The newest synthesis: understanding the interplay of evolutionary and ecological dynamics. *Science* 331(6016): 426–429

[41] Ronce O (2007) How does it feel to be like a rolling stone? Ten questions about dispersal evolution. *Annu Rev Ecol Evol Syst* 38: 231–253.

[42] Prosser et al. (2007) The role of ecological theory in microbial ecology. *Nature Rev Microbiol* 5: 384–392.

SUPPLEMENTARY MATERIAL

Appendix A: Mathematics of the model

1. Relation to the model of MacArthur

The dynamics (2) can be written as

$$\frac{dn_{\vec{\sigma}}}{dt} = \frac{1}{|\vec{\sigma}|} n_{\vec{\sigma}} \left(\sum_i \sigma_i A_i - \chi_{\vec{\sigma}} \right). \tag{S1}$$

where A_i denotes the "available resources". In the model considered in this work, $A_i = \frac{R_i}{T_i}$. In [27], Mac Arthur considered a model of species competing for renewing resources. In that model, the dynamics of organism populations were identical to (S1), but the availability of resources was given by $A_i = R_i(1 - T_i/r_i)$ (see equations (1)-(3) in [27]), where the extra parameter r_i is the renewal rate (or the "intrinsic rate of natural increase").

The dynamics of the two models, therefore, differ only by the choice of the functional form relating population growth and the corresponding decrease of resource availability. The mapping between the notations of [27] and those used here is provided in the table:

Notation for...	In [27]	Here
Species index	i	$\vec{\sigma}$
Species abundance	x_i	$n_{\vec{\sigma}}$
Resources a species can harvest	a_{ij}	σ_i
Resource carrying capacity	K_j	R_i
Minimal resource requirement	T_i	$\chi_{\vec{\sigma}}$
"Resource weight"	w_i	1
Conversion factor (resources $\mapsto$ biomass)	c_i	$1/\|\vec{\sigma}\|$
Resource renewal rate	r_j	N/A

The main difference between the two models is in the treatment of species. In the model of Mac Arthur, each species i was described by an arbitrary chosen vector of parameters a_{ij} (probability to encounter and consume resource j) and the magnitude of minimal resource requirement (χ in our notation). The space of possibilities is unconstrained, and the types available to form a community are determined by historical contingency; Mac Arthur then asks how many species can co-exist in this way.

In the model considered here, a_{ij} are constrained to be 0 or 1 and are determined by a "functional genome" which also sets the minimal resource requirement of each organism type. The resulting space of possibilities is strongly constrained; in particular, it is exponentially large, but finite. This makes it possible to consider the final equilibrium state determined by ecological interactions only, and to address a very different question: when do the functional types favored by (heterogeneous) environmental conditions cluster to form "species"?

2. Existence and stability of equilibrium

The study of equilibria of the metagenome partitioning model is simplified once the dynamics of the model is reformulated as an optimization problem. This was first done in Ref. [27]; here, because of the difference in the way resource consumption is treated, the objective function being optimized is different, but the argument is similar. Consider the following objective function:

$$F = \sum_i R_i \ln T_i - \sum_{\vec{\sigma}} \chi_{\vec{\sigma}} n_{\vec{\sigma}}.$$

On the domain of interest $\{n_{\vec{\sigma}} \geq 0\}$, this F is bounded from above. To see this, note the inequalities:

$$\sum_i T_i = \sum_{\vec{\sigma}} |\vec{\sigma}| n_{\vec{\sigma}} \leq N \sum_{\vec{\sigma}} n_{\vec{\sigma}}$$

and for $\alpha, \beta > 0$:

$$\alpha \ln x - \beta x \leq \alpha \ln \frac{\alpha}{e\beta}$$

Using these, and setting $\min_{\vec{\sigma}} \chi_{\vec{\sigma}} = \chi^* > 0$, one can write:

$$\begin{aligned} F \leq \sum_i R_i \ln T_i - \chi^* \sum_{\vec{\sigma}} n_{\vec{\sigma}} &\leq \sum_i \left(R_i \ln T_i - \frac{\chi^*}{N} T_i \right) \\ &\leq \sum_i R_i \ln \frac{N R_i}{e \chi^*} \end{aligned}$$

The gradient of F is precisely the "competitive advantage" of organism types:

$$\frac{\partial F}{\partial n_{\vec{\sigma}}} = \sum_i \frac{R_i}{T_i} \sigma_i - \chi_{\vec{\sigma}} = \Delta\varphi_{\vec{\sigma}}$$

(it is helpful to note that $T_i = \sum_{\vec{\sigma}} \sigma_i n_{\vec{\sigma}}$). Therefore, F is always increasing along the trajectories of the model:

$$\frac{df}{dt} = \sum_{\vec{\sigma}} \frac{\partial f}{\partial n_{\vec{\sigma}}} \dot{n}_{\vec{\sigma}} = \sum_{\vec{\sigma}} \frac{n_{\vec{\sigma}}}{|\vec{\sigma}|} (\Delta\varphi_{\vec{\sigma}})^2 \geq 0$$

The population as a whole performs a gradient-ascent optimization of F on the domain $n_{\vec{\sigma}} \geq 0$. Since F is bounded, the final equilibrium always exists and is stable. The argument does not guarantee that such an equilibrium is unique; several stable, non-invadeable equilibria could in principle exist. In practice, for equiabundant resources $R_i \equiv R$ such multi-stability was not observed in simulations: the population always converged to the same state for all the initial conditions sampled. The conditions under which multi-stability could be observed requires further investigation. However, for the "perfect tiling" regime as observed in Fig. 2 the set of dominant types is determined uniquely; this will be proved in "The perfect tiling regime" section below.

3. The maximum number of coexisting types

The traditional question of how many types can coexist for a given set of parameters, although not at the focus of this work, is nevertheless instructive to address. A simple linear algebra argument demonstrates that in the model considered here, this maximum number is N: a stable coexistence is possible only for a number of types that is at most equal to the number of resources. This is because for a given set of K types, the K equilibria conditions $\Delta\varphi_{\vec{\sigma}} = 0$ can be seen as a linear mapping between the N-dimensional vector R_i/T_i and a K-dimensional vector of organism costs $\chi_{\vec{\sigma}}$. In the generic case (i.e. if no special symmetries exist in the cost structure), the existence of such a mapping requires $K \leq N$.

Symmetries in the cost structure can lead to degenerate equilibria circumventing this maximal coexistence condition. Imagine, for example, that all organisms have the exact same cost per pathway $\chi*$. In this maximally degenerate case *any* combination of functional types can coexist, provided that $T_i = R_i/\chi^*$: no division of labor strategy is better than any other.

4. Numerical determination of the community equilibria

To determine the final (non-invadable) equilibrium state of a community, one could imagine choosing a random starting point with a non-vanishing abundance of every possible functional types, and evolving it according to the dynamical equations for time $t \to \infty$. The argument in the previous section guarantees that such evolution would converge to an equilibrium state. Numerically, however, such a procedure is highly memory-intensive: simulating time evolution of 2^N functional types is wasteful, since the final population is guaranteed to contain at most N types with non-zero abundance (see section "The maximum number of coexisting types").

Conveniently, verifying that a configuration is a true final equilibrium is much easier than finding it: one only needs to check that the resource surplus $\Delta\varphi_{\vec{\sigma}}$ is zero for all types that are present and is negative for all those that are absent. This verification is fast and is guaranteed to either confirm that the equilibrium state is correct, or provide a list of types that can invade it. Therefore, a simple heuristic procedure can construct the final equilibrium configuration through an iterated sequence of "guesses", whereby a subset of types is first equilibrated, and then updated by removing types that went extinct and adding those that can invade. This is the approach adopted here.

Specifically, the "initial guess" S_0 is constructed using the "fitness criterion" explained in the main text (low cost per pathway = high fitness): for each pathway i, the 10 most cost-efficient (lowest cost per pathway) functional types ($S_0^{(i)}$) that contained pathway i are determined; the union of these cost-efficient types, all taken at equal abundance of 1 unit, constitutes the "initial guess" $S_0 = \bigcup_i S_0^{(i)}$. After this, the following procedure is iterated: the guessed configuration of types is evolved until satisfactory equilibrium using MatLab's variable-order differential equation solver `ode15s`; a satisfactory equilibrium was defined as a configuration where the vector of derivatives $\dot{n}_{\vec{\sigma}}$ of all simulated types would fall below an arbitrary threshold of 10^{-4}. At this point the algorithm checks if this configuration can be invaded by any type not included in this partial simulation. If not, the configuration is accepted as being within the pre-determined numerical error of the true final equilibrium. If, however, types are found that can invade, they are added to the community at abundance 1, the previously considered types whose abundance fell below 10^{-2} are removed, and the simulation cycle is repeated.

This procedure is guaranteed to converge, because the optimization function F is monotonously increasing at every step. The result of convergence is guaranteed to be a true non-invadable equilibrium because the invadability criterion is checked for all types and is exact.

5. The "perfect tiling" regime

To develop an understanding for the structure of the equilibria of the metagenome partitioning model, consider, first, the situation where the final equilibrium contains a set of dominating organisms that partition resources with no overlap (*cf.* Fig. 4A). This "perfect tiling" regime is an instructive starting point that will help build intuition for the general case.

Call a set of organisms P a "perfect tiling" of the resource space if every resource is consumed by exactly one organism in P. Define the "cost of the tiling" as the total cost of all organisms in P:

$$\chi_P \equiv \sum_{\vec{\sigma}\in P} \chi_{\vec{\sigma}}.$$

The tiling with the lowest cost will be called the optimal tiling P^*. Let S be the set of organisms present in the final (non-invadable) equilibrium state of the community.

Proposition: If S contains a perfect tiling $P \subset S$, its cost is optimal: $\chi_P = \chi_{P^*}$. (And therefore, assuming a non-degenerate cost structure, $P = P^*$).

Proof: Assume the contrary: let $P \neq P^*$ be a perfect tiling that is a subset of the types present in the final non-invadable community equilibrium S. Since all types in P are present at non-zero abundance, their resource surplus $\Delta\varphi$ is zero, and therefore:

$$\sum_{\vec{\sigma}\in P} \Delta\varphi_{\vec{\sigma}} = 0 \quad \Rightarrow \quad \chi_P = \sum_i \frac{R_i}{T_i}$$

By the same token, types in P^* are either present ($\Delta\varphi = 0$) or absent ($\Delta\varphi < 0$), and therefore:

$$\sum_{\vec{\sigma}\in P^*} \Delta\varphi_{\vec{\sigma}} \leq 0 \quad \Rightarrow \quad \chi_{P^*} \geq \sum_i \frac{R_i}{T_i} = \chi_P.$$

Since P^* is the optimal tiling, we conclude that $\chi_P = \chi_{P*}$ as claimed.

To summarize, if the final state contains a perfect tiling, it will necessarily be the one with the lowest total cost. This clarifies how the survival of an organism in the final equilibrium is related to its cost per pathway (its individual "fitness") but also to the interactions with other types: even the organism with the absolute lowest cost per pathway (highest fitness) is not guaranteed to survive.

6. The $N = 2$ case

For $N = 2$, there are only two tilings: $\{\underline{A}, \underline{B}\}$ and $\{\underline{AB}\}$, and any population equilibrium is always in the perfect tiling regime. Applying the argument of the previous section, one concludes that the choice between the two regimes is determined by the difference of the total costs: $\delta\chi = \chi_{\underline{A}} + \chi_{\underline{B}} - \chi_{\underline{AB}}$, as stated in the main text.

7. Stability analysis in the "perfect tiling" regime

The expressions (6) derived in the main text are completely general and are valid for any fixed point of the dynamics (4). For a population in the "perfect tiling" regime, they can be simplified considerably; this analysis provides useful intuition. In this regime, $\alpha \cap \beta = \emptyset$ for any distinct $\alpha, \beta \in S$, and T_i is simply n_α for the unique type α such that $i \in \alpha$. For simplicity, set $R_i \equiv R$ like in the main text; the abundance of the type $\alpha \in S$ is then determined by the condition:

$$\Delta\varphi_\alpha = \sum_{i\in\alpha} \frac{R}{n_\alpha} - \chi_\alpha = |\alpha|\frac{R}{n_\alpha} - \chi_\alpha = 0.$$

Therefore, $R/n_\alpha = \chi_\alpha/|\alpha|$, the cost per pathway.

As a result, the Jacobian $M_{\alpha\beta}$ takes the simple form:

$$M_{\alpha\beta} = \left(\begin{array}{c|c} -\Lambda_\chi & * \\ \hline 0 & \Lambda_{\Delta\varphi} \end{array}\right) \begin{array}{l} \leftarrow \alpha \in S \\ \\ \leftarrow \alpha \notin S \end{array}$$

Here Λ_χ and $\Lambda_{\Delta\varphi}$ are diagonal matrices whose eigenvalues, respectively, are the cost per pathway $\chi_\alpha/|\alpha|$ for types in S, and the resource surplus per pathway $\Delta\varphi_\alpha/|\alpha|$ for types not in S. The asterisk in the upper-right block denotes some non-zero entries that have no effect on the eigenvalues of the Jacobian $M_{\alpha\beta}$. As stated in the main text, for every type $\alpha \notin S$ that can invade the community ($\Delta\varphi_\alpha > 0$), there is an unstable mode with a positive eigenvalue. After restriction to the set of organisms in S, in the "perfect tiling" regime the Jacobian $M_{\alpha\beta}$ becomes purely diagonal $M'_{\alpha\beta} = -\Lambda_\chi$, so that each present type constitutes its own stable eiegenmode, and the eigenvalues are given by the cost-per-pathway ratio of the organisms in the optimal tiling (which are all approximately equal near the fitness peak, see Fig. 2A).

Appendix B: Supplementary figures for Fig. 2

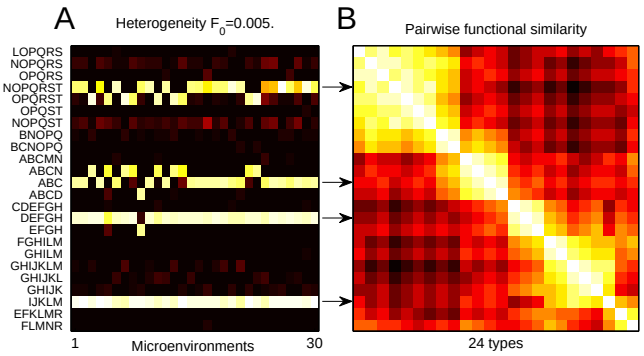

FIG. S1. This figure reproduces Fig. 2CD, except that the 0.1% biomass threshold is not applied. The total number of types observed in 30 microenvironments rises to 24 and includes 7 additional types that appear at a low abundance in a single or few microenvironments and thus have a very weak relative presence in the habitat as a whole ("rare species"). The last two rare types appear to form a new, fifth cluster with an extremely low total abundance in the community.

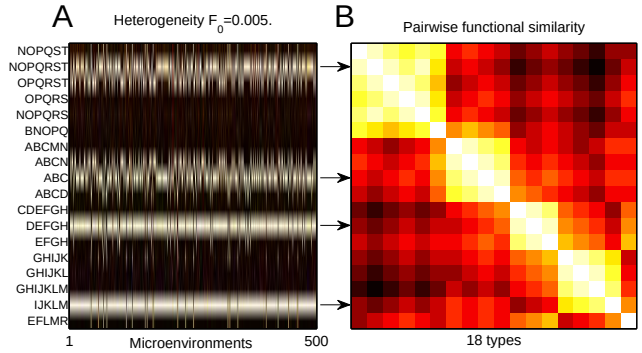

FIG. S2. Sampling more micro-environments increases the total number of all observed types, from 24 in 30 micro-environments to 44 in 500 (not shown). However, the clustered structure of the types crossing the total biomass threshold of 0.1% (presented in this figure) remains essentially intact: the only difference is the addition of an 18th type, which forms its own cluster with a very low relative abundance. By chance, none of the first 30 micro-environments supported this type, and therefore it did not appear on Fig. 2C.

Appendix C: Generality of the clustered (Fig. 2) and not clustered (Fig. 3) regimes

In the interest of clarity of presentation, a single set of parameters was chosen and used throughout the analysis in the main text. However, the two regimes that were discussed one where functional types cluster together (Fig. 2) and one where they do not (Fig. 3), are generic: their existence does not depend on the particular parameter values chosen here.

To see this, consider the following setting: let X be a space of possible organism types, and $\mathcal{F}$ a fitness function $\mathcal{F} : X \mapsto \mathbb{R}$. The only assumption about $\mathcal{F}$ will be that it possesses some continuity with respect to some metric structure on X, so that close types tend to have similar fitness (Fig. S3). Now, just like in the main text, consider a habitat modeled as a collection of heterogeneous micro-environments, the effect of heterogeneity being a random additive contribution of magnitude h to the fitness of each type. However, unlike the main text, imagine that each micro-environment simply selects the highest-fitness type, so that all ecological considerations are eliminated entirely.

If $h \ll h_1$ (see Fig. S3), the heterogeneity does not appreciably modify the fitness profile, and the same type will be selected in all micro-environments. If $h \simeq h_1$, different micro-environments may select different types; however, the continuity of $\mathcal{F}$ will ensure that the selected types cluster around its local maximum or maxima (A and B in Fig. S3). Finally, if heterogeneity is very strong $h \gg h_2$, effectively randomizing the fitness values in each microenvironment, the selected types will no longer exhibit any clustered structure. Therefore, as stated in the main text, the existence of the two regimes depicted in Figs. 2 and 3 depends neither on the parameters nor on the details of the underlying model of ecological interactions, but is a direct consequence of the simplistic manner in which environmental heterogeneity was assumed to manifest itself.

Note, however, that this argument provides no way to interpret either of the regimes in terms of the number of effective species. The non-trivial part of the argument made in this work is not the existence of a regime where

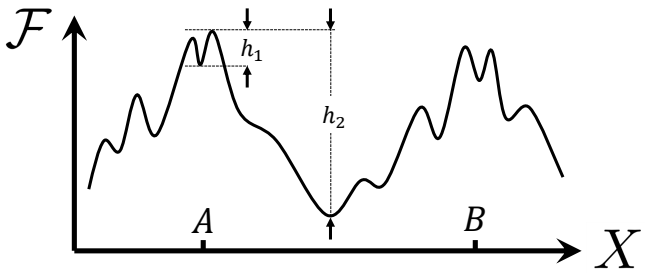

FIG. S3. A sketch of a hypothetical fitness function $\mathcal{F}(X)$, where the multi-dimensional space X is represented as 1-dimensional for simplicity.

types no longer cluster together. Rather, this simple way of tuning the degree of clustering provided a convenient toy-model setup to construct an alternative formalism demonstrating that 17 functional types may represent 17 species, or four, or a collective regime with "no species at all", and the answer to this question is an emergent property of the ecological interactions between individuals.