Editors' Suggestion | Featured in Physics

Diversity, Stability, and Reproducibility in Stochastically Assembled Microbial Ecosystems

Akshit Goyal
The Simons Centre for the Study of Living Machines, NCBS-TIFR, Bengaluru 560 065, India

Sergei Maslov*
Department of Bioengineering and Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA

(Received 3 November 2017; published 13 April 2018)

Microbial ecosystems are remarkably diverse, stable, and usually consist of a mixture of core and peripheral species. Here we propose a conceptual model exhibiting all these emergent properties in quantitative agreement with real ecosystem data, specifically species abundance and prevalence distributions. Resource competition and metabolic commensalism drive the stochastic ecosystem assembly in our model. We demonstrate that even when supplied with just one resource, ecosystems can exhibit high diversity, increasing stability, and partial reproducibility between samples.

DOI: 10.1103/PhysRevLett.120.158102

Natural microbial ecosystems are remarkably diverse, often harboring hundreds to thousands of coexisting species in microscopic volumes [1–3]. How do these ecosystems manage to acquire and maintain such a high diversity? This so-called "paradox of the plankton" [4] is especially surprising given that microbes are capable of rapid exponential growth and fierce competition for nutrients. Indeed, the competitive exclusion principle [2,5] postulates that the number of species in an ecosystem at a steady state cannot exceed the number of available nutrients.

Compounding this puzzle, theoretical studies [6] suggest that highly diverse ecosystems are generally prone to instabilities. This brings up a second question: how do naturally occurring microbial ecosystems manage to remain relatively stable despite their diversity?

Moreover, ecosystems operating under similar environmental conditions could be rather different from each other in terms of species composition [3,7,8]. This apparent lack of reproducibility does not apply equally to different organisms. Some species, classified as "core'" or "keystone," are detected in most individual ecosystems. Other "peripheral" species are only observed in a small fraction of them. Observed species' prevalence distributions (the fraction of similar ecosystems a species is detected in) are often U shaped: their peaks occupied by these core and peripheral species, respectively [7]; often, these are also correlated with species abundances [8]. We are thus presented with a third question: what determines the reproducibility (or lack thereof) of species composition in microbial ecosystems?

Here, we introduce a conceptual model of a stochastically assembling microbial ecosystem, which in spite of its simplicity, addresses and suggests possible solutions to all three of these long-standing puzzles.

To explain the aforementioned high diversity and poor reproducibility, previous models have relied on a number of factors including spatial heterogeneity [5,9], temporal and seasonal variations in resource availability [10,11], thermodynamic constraints [12], microbial "warfare" and cooperation via ecological feedbacks [13], and predation by bacteriophages [14,15]. In contrast to this, our model attributes high diversity to metabolic by-products secreted by microbes due to incomplete resource-to-biomass conversion, which could in turn be used by other species for growth. By its very nature, our model simultaneously exhibits (i) high species diversity, (ii) gradually increasing stability—reached after repeated rearrangements, (iii) a U-shaped prevalence distribution, and (iv) a positive abundance-prevalence correlation.

While our model clearly does not include many of the previously proposed factors known to affect these features, we believe it is a reasonable first order description of some real ecosystems, examples of which include the human oral microbiome [7], methanogenic bioreactors [16], and anaerobic digesters in wastewater treatment plants [3].

Our model describes a dynamic microbial ecosystem in which species attempt to populate the environment externally supplied with a single resource. We assume that species can convert only a fraction of consumed resources into their biomass, while secreting the rest as metabolic by-products. These in turn may serve as nutrient sources for other species in the ecosystem. This allows even one externally supplied resource to support high ecosystem diversity purely via by-product-driven commensal interactions.

New species are constantly introduced to this environment from some external population. Their survival or

extinction is determined by a simple rule dictated by competitive exclusion. Because of the commensal relationship between these species, elimination of just one species may lead to an "extinction avalanche" in which multiple species are lost.

We explore how species diversity in microbial ecosystems is established over time. Moreover, by simulating several instances of ecosystem assembly, we can separate the set of core (high-prevalence) species from those with progressively lower prevalence.

The dynamics in our bioreactorlike environment is fully characterized by the concentrations of individual resources (metabolites) labeled as $C_0, C_1, \ldots$ and the abundances of all resident microbial species labeled as $B_1, B_2, \ldots$. When we initialize the model, each species is assigned a single resource it can grow on and $\beta = 2$ metabolic by-products. All resources are randomly selected from a "universal chemistry" of size $N_{\text{univ}} = 5000$. This choice is inspired by the total number of metabolites in KEGG's metabolic database [17]. However, qualitatively similar results are obtained for much smaller values N_{univ}, for example, the number of carbon sources typically utilized by microbes.

The environment is supplied with a single resource (labeled 0) at a constant flux ϕ_0. After several attempts, the first microbial species (labeled 1) capable of utilizing the resource 0 colonizes the environment. The following equations determine the dynamical behavior of resource concentration C_0 and microbial abundance B_1:

$$\frac{dC_0}{dt} = \phi_0 - \frac{\lambda_1 C_0 B_1}{Y} - \delta C_0, \tag{1}$$

$$\frac{dB_1}{dt} = \lambda_1 C_0 B_1 - \delta B_1. \tag{2}$$

Here, $\lambda_1 C_0$ is the growth rate of the species 1 consuming the resource 0 at a rate $\lambda_1 C_0/Y$, where Y is the yield of the biomass conversion process (microbial concentration per unit resource concentration). The resource affinity λ is assigned by a random draw from a log-normal distribution such that the logarithm of λ has mean 0 and variance 1.

Our model is based on carefully following the flow of resources (e.g., carbon) throughout the ecosystem. Different resources could be interconverted into each other and into the biomass of microbes. Hence it is convenient to measure all microbial abundances in units of the resource concentration. We adopt this change of units for B_i for the rest of this Letter. Microbial yield is given by $Y = (1-\alpha)$, where $(1-\alpha) < 1$ represents the fraction of the consumed resource (e.g., carbon atoms) successfully converted to biomass. The remainder is secreted as two by-products 1 and 2 getting shares $\nu_1\alpha$ and $\nu_2\alpha = (1-\nu_1)\alpha$, respectively.

Another interpretation of these equations would apply if all processes in the ecosystem were energy limited (as opposed to nutrient limited). In this case, it would be convenient to measure the concentrations of both resources and microbes in units of energy density. The factor $(1-\alpha)$ could then be interpreted as an energy conversion efficiency. Because of dissipation, here it would be possible for α (the fraction of the incoming energy flux secreted as by-products) to be smaller than the leftovers from biomass conversion. Barring small corrections, the results of our model would be equally applicable to such energy-limited ecosystems.

We assume that the concentrations of both microbes and resources are diluted at the same rate, δ. It is straightforward to generalize our model to a case where these dilution rates are in fact different (as is often the case in batch-fed bioreactors). Throughout this Letter we are only interested in the steady-state properties of the system, which can be easily derived from Eqs. (1) and (2). At steady state, C_0^* and B_1^* are given by

$$C_0^* = \frac{\delta}{\lambda_1}, \tag{3}$$

$$B_1^* = \frac{(\phi_0 - \frac{\delta^2}{\lambda_1})Y}{\delta} = \frac{\tilde{\phi}_0(1-\alpha)}{\delta}. \tag{4}$$

Here, to simplify our notation, we have introduced the effective flux of a nutrient (adjusted for dilution), which is given by $\tilde{\phi}_0 = \phi_0 - \delta C_0^* = \phi_0 - (\delta^2/\lambda_1)$. Note that (a) at steady state, resource concentration C_0^* depends inversely on λ implying that if two species were to compete for the same resource, the one with a higher λ would drive the resource concentration lower than the other, thus being the only survivor of the two, and (b) unlike the steady-state nutrient concentration, the steady-state species abundance is largely independent of λ. Indeed, λ only enters this equation via the effective resource flux which in the limit of low dilution approximates ϕ_0. Note that using a more general expression for microbial growth, e.g., the Monod equation, does not affect our results [18].

We simulate ecosystem assembly in discrete time steps corresponding to the introduction of a new microbial species into the ecosystem. We assume that these events are sufficiently infrequent for the system to reach steady state between two subsequent immigration attempts. We measure time in the number of attempted species immigrations. For a newly introduced species to survive, it must both have its consumed resource present in the ecosystem, and must also be most competitive on it, i.e., have the highest λ.

In this case, its steady-state abundance is determined by Eq. (4) but with the effective flux $\tilde{\phi}_i$ of its consumed resource. If all by-products are equally partitioned, the average effective flux at trophic layer ℓ is related to the external resource flux via

$$\langle \tilde{\phi}_i \rangle_\ell = \phi_0 \left(\frac{\alpha}{\beta}\right)^\ell - \frac{\delta^2}{\lambda}\left(1 + \frac{\alpha}{\beta}\right)^\ell. \tag{5}$$

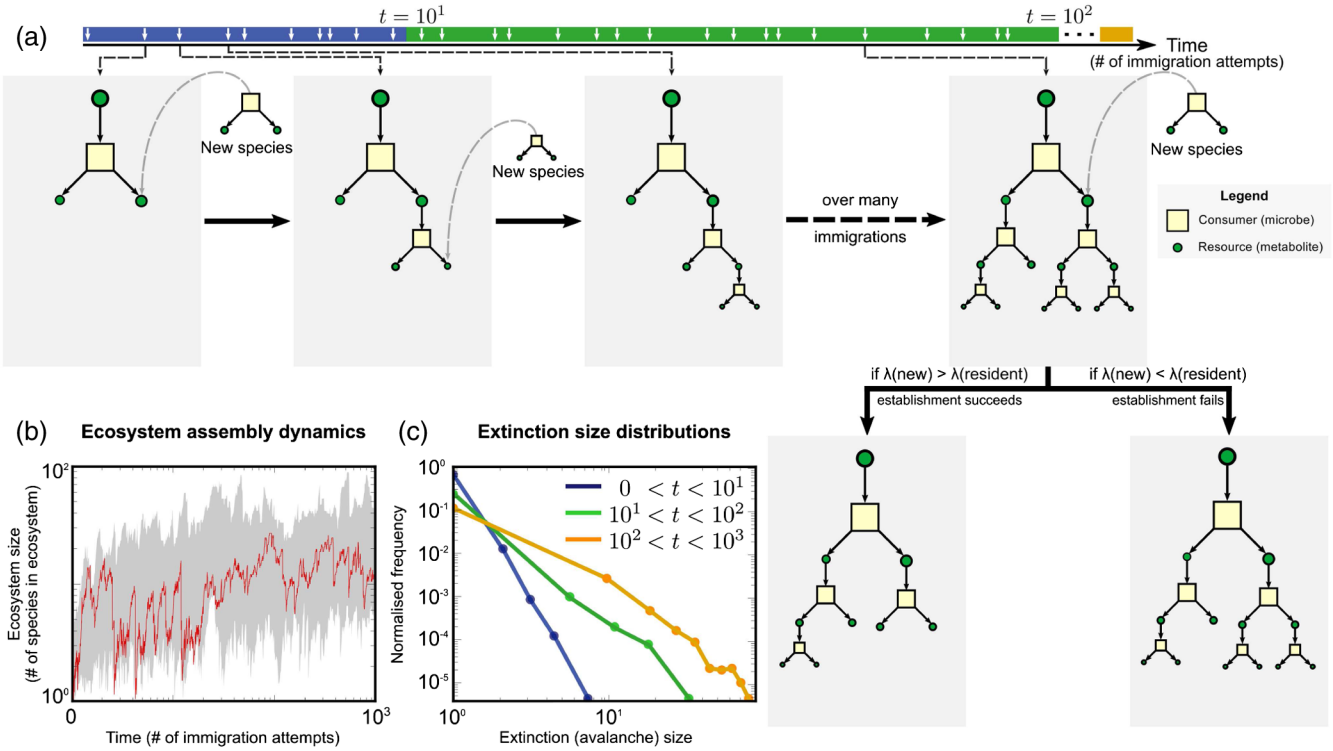

FIG. 1. *Ecosystem assembly in the model.* (a) The diagram illustrates different phases in the assembly dynamics involving species (yellow squares) consuming resources (green circles). Sizes are indicative of the steady-state abundances and concentrations. Initially, only a single externally supplied resource (the largest green circle) is available and consumed by a microbe, which in turn secretes $\beta = 2$ metabolic by-products. New species immigrate into this ecosystem (immigration events marked on the timeline), each using only one resource. Ecosystem establishment is contingent on the following assembly rule: if the resource affinity λ of the new species is higher than any resident species on its chosen resource, the immigrant species survives and the resident goes extinct (along with all its dependents). (b) A sample assembly trajectory (in red) of the ecosystem size (number of species) as a function of time (t, measured in number of immigration attempts) at dilution rate $\delta = 10^{-1}$ days^{-1}. The gray envelope shows ecosystem sizes over 1000 assembly trajectories. (c) Extinction size distributions (number of species that go extinct during a single immigration event) get broader as ecosystem assembly proceeds: $t < 10^1$ (blue); $10^1 < t < 10^2$ (green), and $10^2 < t < 10^3$ (orange).

Note that if a new species competitively displaces another, any species that directly or indirectly depend on the latter for by-products could also go extinct. Very rarely, this extinction might be averted as long as at least one other resident species produces the same by-product and the remaining flux satisfies $\tilde{\phi} > 0$. As ecosystem assembly proceeds, we observe that the distribution of the number of species that go extinct during such events gets broader [see Fig. 1(c)]. Over many steps of ecosystem assembly, as species use and secrete more by-products, the entire ecosystem assumes a treelike structure [see Fig. 1(a) for ecosystem structure and Fig. 1(b) for sample dynamical trajectories].

Our model ecosystems have two interesting emergent features. First, species' steady-state abundances follow a power law. Indeed, at trophic layer ℓ a typical species' abundance is determined by its consumed resource flux as $\langle B^* \rangle_\ell \approx [\langle \tilde{\phi}_i \rangle_\ell (1-\alpha)/\delta]$. Each layer can accommodate β^ℓ species, where β is the number of by-products per species. We can show [18] that the species' steady-state abundance distribution follows

$$\mathcal{N}(B = b) \sim b^{-(1+\frac{\log\beta}{|\log\alpha|+\log(\alpha+\beta)})}. \tag{6}$$

A rank-abundance plot [see Fig. 2(a)] follows a power law with exponent $[|\log\alpha| + \log(\alpha+\beta)]/\log\beta$. For appropriately chosen α, both this expression and our simulations [solid curves in Fig. 2(a)] agree with data from microbial ecosystems sampled from the human tongue [7] and methanogenic bioreactors [16] [open circles in Fig. 2(a)].

Second, the dilution rate δ sets a limit to the number of species in the ecosystem. This happens when the resource flux at the bottom-most layer $\ell_{\max}$ becomes negative [see Eq. (5)]. The number of species in the ecosystem [18] is proportional to $\beta^{\ell_{\max}}$ and given by

$$\mathcal{N}_{\max}(\delta) \sim \delta^{-(\frac{2\log\beta}{|\log\alpha|+\log(\alpha+\beta)})}. \tag{7}$$

For $\beta = 2$ and $\alpha = 0.1$, this expression [black solid line in Fig. 2(b)] approximates our simulated ecosystems [red solid line in Fig. 2(b)]. Note that this expression provides an upper bound to the number of species. It assumes both equal partitioning of all by-products and equal λ's for all species.

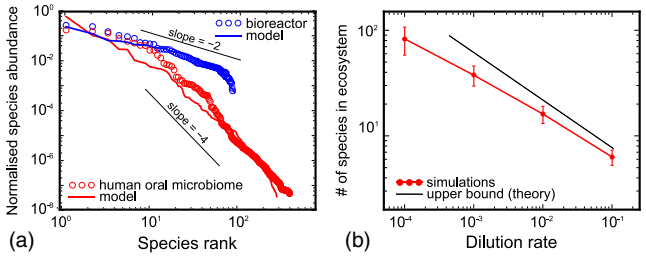

FIG. 2. *Emergent ecological features.* (a) Rank-abundance plot of normalized species abundances in a methanogenic bioreactor [16] (blue circles) and the human oral microbiome [7] (red circles) and, for comparison, simulated ecosystems from our model (corresponding solid lines) with α equal to 0.5 and 0.1, respectively. (b) The dilution rate δ in the chemostat controls the maximal size $\mathcal{N}_{\max}$ of the ecosystem coexisting on a single externally supplied resource. Here, $\alpha = 0.1$ and $\beta = 2$. $\mathcal{N}$ approximately agrees with the expression in Eq. (7).

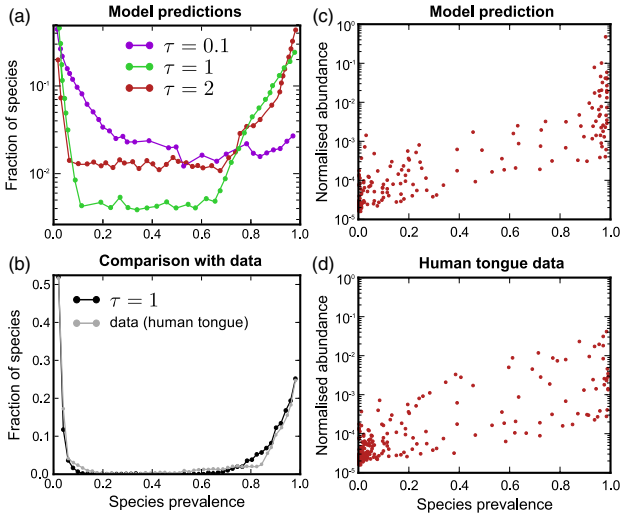

FIG. 3. *Reproducibility from repeated assembly.* (a) Species prevalence distributions from several ecosystems stochastically assembled from a common pool of 1000 species (here $\alpha = 0.1$). Shown are distributions for different times τ in the assembly process (measured in number of immigration attempts by any one species): 0.1 (violet) at which most species have low prevalence, 1 (green) and 2 (red) for which we observe a U-shaped distribution (some core species, most peripheral). (b) The distribution for $\tau = 1$ (black) matches largely with that in longitudinally sampled human oral microbiome (tongue) [7] (gray). (c), (d) Normalized species abundance data correlates positively with species abundance in both (c) simulations and (d) oral microbiome.

We now attempt to understand the reproducibility of species composition in similar ecosystems. For this, we first generate a "species pool" by running one instance of ecosystem assembly until it reaches 1000 species. Even transiently successful colonizers are added to this pool. We then simulate several instances of stochastic ecosystem assembly under identical initial conditions. Species attempt to colonize each ecosystem from the pool randomly (with replacement), with the assumption that each species has the same average immigration rate. The assembly process runs for a fixed time period τ measured by the number of per species immigration attempts.

After collecting several ecosystems for $\tau = \{(1/10), 1, 2\}$, we plot the distribution of species prevalence—the fraction of randomly assembled ecosystems in which a species is present [see Fig. 3(a)]. We observe that the shape of this distribution becomes more pronouncedly "U shaped" with increasing τ. For small values of τ [Fig. 3(a), violet] the distribution is dominated by the species with small prevalence values. This indicates that for such short time of assembly stochastic species colonization dominates. For higher values of τ (shown in green and red), the distribution is "U shaped"; i.e., most species are either core (found in most ecosystem instances) or peripheral (found in a small fraction of them). This can be explained as follows: as assembly proceeds, species from the pool which use high-flux resources with the largest resource affinities establish themselves at the top trophic layers. After some delay, other species from the pool that depend on these can successfully colonize middle trophic layers. However, stochastic colonization continues to dominate in the lowest trophic layers and contributes to the low-prevalence portion of the U shape. Such a U-shaped prevalence distribution is observed in many real microbial communities: such as longitudinal samples of the human oral microbiome [7] [see the gray distribution in Fig. 3(b); the model prediction is shown in black for comparison] and anaerobic digesters in wastewater treatment plants [3].

Real ecosystems are often characterized by a positive correlation between the prevalence and relative abundance of species [see the scatter plot in Fig. 3(d) for oral microbiome and Fig. 2(b) in Ref. [3] for wastewater plants]. This observation is also captured well by our model: high-prevalence species tend to be consumers of resources at higher trophic layers and thus tend to be highly abundant [see model prediction in Fig. 3(c)].

To summarize, we present here a conceptual model of a microbial ecosystem which demonstrates some ecological consequences of metabolic facilitation—all of which are borne well by data from real ecosystems. What distinguishes our model from previous "consumer-resource" [10,20,21] approaches? First, we explicitly model energy conservation in the form of incomplete resource-to-biomass conversion and generate metabolic by-products from what remains. Second, we explicitly handle species abundances to explain why they scale according to a power law. Finally, we generate and explain species reproducibility from many microbial communities by simulating several stochastic assemblies.

Data from microbial ecosystems in different environments corroborate the overarching predictions of this model, namely: human oral microbiome samples [7], soil

communities [8], wastewater treatment plants [3], and methanogenic bioreactors [16]. Interestingly, the oral data we use are from the human tongue, which is believed to be assembled in a specific temporal order; i.e., late-colonizing species depend on the ones that came before them [22]. This is very similar to the mechanism behind our simulations.

Note that our model contains a number of simplifying assumptions that are not realistic. (a) While energy is still strictly conserved in our model through balancing fluxes, we allow microbes to generate by-products with higher energy content (per molecule) out of resources with lower energy. However, the general direction of the metabolic flow in real biochemical pathways is down the energy gradient. We take this into account in model variant A [18] where metabolites are arranged in an energy hierarchy so that by-products always have lower energy content (per molecule) than their parent resource. (b) When generating new species, we assign their by-products randomly from a large set of possible metabolites. In reality, the number of possible metabolic pathways utilizing a given resource is smaller. For example, Ref. [12] lists 5 distinct glucose utilization pathways. We take this into account in model variant B [18] where each resource can be utilized via $\eta = 5$ distinct metabolic pathways, each producing its own set of $\beta = 2$ by-products. (c) Species consuming the same resource are thought to be subject to a "rate-yield trade-off" [10,23,24], which states that microbial species with faster growth rates on a given resource tend to use it less efficiently. Both model variants A and B [18] take this trade-off into account. Simulations from both variants show that our central results remain qualitatively unaffected by all these modifications [18].

Also note that we assume here that each microbe can use only one resource. In reality, microbes can typically use multiple resources for growth. However, such an extension involves several choices. First, one needs to decide if a microbe would consume resources in parallel or sequentially (both cases are observed in real microbes). Second, one may envisage trade-offs between resource affinity per resource and the number of resources. One extreme limit of this trade-off in which the sum of affinities always adds up to the same number has been recently modeled in Ref. [25]. We are currently considering alternative models which incorporate these choices.

A. G. acknowledges support from the Simons Foundation as well as the Infosys Foundation.

*maslov@illinois.edu

[1] C. A. Lozupone, J. I. Stombaugh, J. I. Gordon, J. K. Jansson, and R. Knight, Nature (London) **489**, 220 (2012).
[2] T. P. Curtis, W. T. Sloan, and J. W. Scannell, Proc. Natl. Acad. Sci. U.S.A. **99**, 10494 (2002).
[3] R. Mei, T. Narihiro, M. K. Nobu, K. Kuroda, and W.-T. Liu, Sci. Rep. **6**, 34090 (2016).
[4] G. E. Hutchinson, Am. Nat. **95**, 137 (1961).
[5] S. B. Hsu, S. Hubbell, and P. Waltman, SIAM J. Appl. Math. **32**, 366 (1977).
[6] R. M. May, Nature (London) **238**, 413 (1972).
[7] J. G. Caporaso, C. L. Lauber, E. K. Costello, D. Berg-Lyons, A. Gonzalez, J. Stombaugh, D. Knights, P. Gajer, J. Ravel, N. Fierer, J. I. Gordon, and R. Knight, Genome Biol. **12**, R50 (2011).
[8] A. Barberán, S. T. Bates, E. O. Casamayor, and N. Fierer, ISME J. **6**, 343 (2012).
[9] J. Huisman, P. van Oostveen, and F. J. Weissing, Am. Nat. **154**, 46 (1999).
[10] T. Pfeiffer, S. Schuster, and S. Bonhoeffer, Science **292**, 504 (2001).
[11] R. C. MacLean and I. Gudelj, Nature (London) **441**, 498 (2006).
[12] T. Großkopf and O. S. Soyer, ISME J. **10**, 2725 (2016).
[13] T. L. Czárán, R. F. Hoekstra, and L. Pagie, Proc. Natl. Acad. Sci. U.S.A. **99**, 786 (2002).
[14] T. Thingstad and R. Lignell, Aquatic Microbial Ecology **13**, 19 (1997).
[15] T. F. Thingstad, Limnol. Oceanogr. **45**, 1320 (2000).
[16] M. K. Nobu, T. Narihiro, C. Rinke, Y. Kamagata, S. G. Tringe, T. Woyke, and W.-T. Liu, ISME J. **9**, 1710 (2015).
[17] M. Kanehisa and S. Goto, Nucleic Acids Res. **28**, 27 (2000).
[18] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.120.158102 for derivations and results from model variants, which includes Refs. [10,12,19,22,23].
[19] J. Monod, Annu. Rev. Microbiol. **3**, 371 (1949).
[20] R. MacArthur, Theor. Popul. Biol. **1**, 1 (1970).
[21] J. Rodríguez, J. M. Lema, and R. Kleerebezem, Trends Biotechnol. **26**, 366 (2008).
[22] R. Levy and E. Borenstein, Proc. Natl. Acad. Sci. U.S.A. **110**, 12804 (2013).
[23] R. E. Beardmore, I. Gudelj, D. A. Lipson, and L. D. Hurst, Nature (London) **472**, 342 (2011).
[24] M. Novak, T. Pfeiffer, R. E. Lenski, U. Sauer, and S. Bonhoeffer, Am. Nat. **168**, 242 (2006).
[25] A. Posfai, T. Taillefumier, and N. S. Wingreen, Phys. Rev. Lett. **118**, 028103 (2017).

Supplemental Material
Diversity, stability, and reproducibility in stochastically assembled microbial ecosystems

Akshit Goyal
The Simons Centre for the Study of Living Machines, NCBS-TIFR, Bengaluru 560 065, India.

Sergei Maslov
Department of Bioengineering and Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA.

I. STEADY STATES CONCENTRATIONS VIA THE MONOD EQUATION

Our results for steady state concentrations described in Eqs. (3–4) can be generalized to microbial growth laws described by the Monod equation [1] in which case Eqs. (1–2) become:

$$\frac{dC_0}{dt} = \phi_0 - \frac{\mu_1^{(\max)} C_0}{K_1 + C_0} \cdot \frac{B_1}{Y} - \delta \cdot C_0, \tag{S1}$$

$$\frac{dB_1}{dt} = \frac{\mu_1^{(\max)} C_0}{K_1 + C_0} \cdot B_1 - \delta \cdot B_1, \tag{S2}$$

Here $\mu_1^{(\max)}$ is the maximum growth rate of the microbe 1 realized when its resource is highly abundant, while K_1 (units of concentration) is its half-saturation constant for this resource. The growth law used in the main text is the limiting case of the Monod equation when $C_0 \ll K_1$. In this case growth depends only on the composite parameter $\lambda_1 = \mu_1^{(\max)}/K_1$.

Solving these equations for the steady state one gets:

$$C_0^* = \frac{\delta}{1 - \delta/\mu_1^{(\max)}} \cdot \frac{K_1}{\mu_1^{(\max)}} \quad , \tag{S3}$$

$$B_1^* = \frac{\left(\phi_0 - \delta \cdot C_0^*\right) Y}{\delta} = \frac{\widetilde{\phi_0}(1-\alpha)}{\delta}. \tag{S4}$$

Note that the equation (S4) is exactly the same as the Eq. (4), while Eq. (S3) corrects Eq. (3) for high dilution by setting the "absolute speed limit" $\delta < \mu_1^{(\max)}$ on the dilution rate δ.

Even more generally, any monotonically increasing function $\mu_1(C)$ of the growth rate of the microbe 1 on its only resource concentration would leave Eq. (S4) unchanged, while change the Eq. (S3) to C_0^* such that $\mu_1(C_0^*) = \delta$.

the following rule: a microbe who can survive on a a lower steady state concentration of the resource displaces the one with the higher steady state concentration.

In the course of the evolution via series of competitive displacements, microbes are likely to drive their resource concentrations so low (below corresponding K_i) so that the Monod equations can be safely replaced with their bilinear limit used in the main text.

II. DERIVATION OF SPECIES ABUNDANCE DISTRIBUTIONS

Consider a tree-like ecosystem in a bioreactor as in our model. At steady state, species abundances depend on their effective fluxes corresponding to their consumed resources as in equation (4) in the main text. Consider also the layered arrangement of the ecosystem. Given this, for a species at layer ℓ in the tree, its steady-state abundance b is approximately given by $\frac{\langle\widetilde{\phi_i}\rangle_\ell (1-\alpha)}{\delta}$. In the limit of low dilution δ, we can write b roughly as:

$$\begin{aligned} \log b &\sim \log\phi_0 - \ell \cdot (|\log\alpha| + \log(\alpha+\beta)) + \log(1-\alpha) + \log\delta \\ &\sim \kappa - (|\log\alpha| + \log(\alpha+\beta)) \cdot \ell, \end{aligned} \tag{S5}$$

where κ is a constant. Now, note that that each layer can accommodate β^ℓ species, where β is the number of byproducts per species. To first order, when we ask for the number of species $\mathcal{N}(B > b)$ with abundance greater than

a certain value b, we are asking for species at layer numbers lower than $\ell_{\max}$. Inverting equation (S5), we can write this number as follows:

$$\begin{aligned}\mathcal{N}(B>b) &\sim 1+\beta+\beta^2+...+\beta^{\ell_{\max}} \sim \frac{\beta^{\ell_{\max}+1}}{\beta-1} \\ &= e^{\log\beta\left[\frac{\kappa-\mathrm{b}}{|\log\alpha|+\log(\alpha+\beta)}\right]} \\ &= \kappa' \cdot b^{-\left[\frac{\log\beta}{|\log\alpha|+\log(\alpha+\beta)}\right]},\end{aligned} \tag{S6}$$

where κ' is another constant independent of b. From this cumulative distribution, the normalized species abundance distribution will thus be:

$$\mathcal{N}(B=b) \sim b^{-\left(1+\frac{\log\beta}{|\log\alpha|+\log(\alpha+\beta)}\right)} \tag{S7}$$

III. DERIVATION FOR ECOSYSTEM CAPACITY

We wish to derive the number of species that the ecosystem can accommodate at steady state given a constant dilution rate δ. Note that species cannot survive at steady state unless their steady state abundance is positive. For this to be the case, at the bottom-most layer in the ecosystem $\ell_{\max}$, the resource flux for any consumer species with resource affinity λ must be positive. Using the expression in equation (5) of the main text, this implies that the following relation must hold at $\ell_{\max}$:

$$\begin{aligned}\phi_0\left(\frac{\alpha}{\beta}\right)^{\ell_{\max}} &= \frac{\delta^2}{\lambda}\left(1+\frac{\alpha}{\beta}\right)^{\ell_{\max}} \\ \implies \ell_{\max} &= \frac{\log\phi_0+\log\lambda-2\log\delta}{|\log\alpha|+\log(\alpha+\beta)}.\end{aligned} \tag{S8}$$

Now, the number of species in the ecosystem $\mathcal{N}_{\max}$ is of the order of $\frac{\beta^{\ell_{\max}+1}}{\beta-1} = \beta^{\ell_{\max}} \cdot \frac{\beta}{\beta-1}$:

$$\begin{aligned}\mathcal{N}_{\max}(\delta) &\sim e^{\log\beta\cdot\left(\frac{\log\phi_0+\log\lambda-2\log\delta}{|\log\alpha|+\log(\alpha+\beta)}\right)} \\ &\sim \delta^{-\left(\frac{2\log\beta}{|\log\alpha|+\log(\alpha+\beta)}\right)}.\end{aligned} \tag{S9}$$

IV. DERIVATION FOR RAREFACTION CURVES

Typically surveys of microbial ecosystems also involve measuring 'rarefaction curves', i.e. the number of species observed or detected over the process of sampling several similar ecosystems. Since in our model we performed repeated stochastic assemblies of ecosystems, we can also demonstrate rarefaction curves similar to those observed in the aforementioned surveys. We show below an example of these curves from our model ecosystems in both linear-linear (left) and log-log (right) forms.

We can show that the species' prevalence distributions we discuss in the main text are related to these rarefaction curves. Consider a species pool with N_{pool} species, each species i with an associated prevalence f_i.

Here f_i represents the frequency with which this species is found in a particular ecosystem sample. We wish to derive $N_{\text{obs}}(n)$, the number of species observed or detected after taking n ecosystem samples.

The probability that after n sampling events, a particular species has *not* been detected is $(1-f_i)^n$. Hence, the chance that it is detected at the nth sampling event is $1-(1-f_i)^n$. Summing over all species in the pool, we get the desired expression for $N_{\text{obs}}(n)$:

$$N_{\text{obs}}(n) = \sum_{i=1}^{N_{\text{pool}}} \Big[1-(1-f_i)^n\Big]. \tag{S10}$$

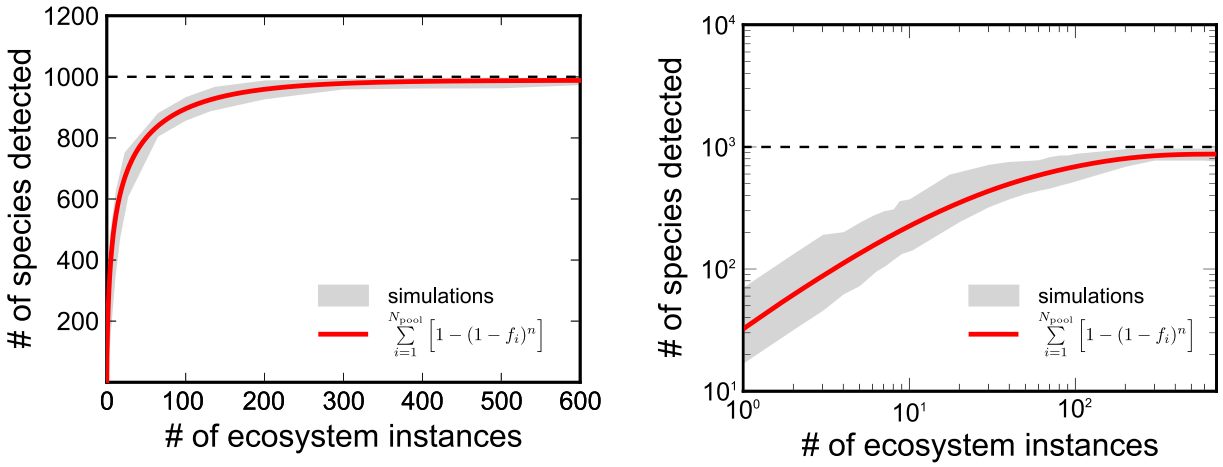

FIG. S1. **Rarefaction curves.** The total number, $N_{\text{obs}}(n)$, of distinct species observed plotted as a function of the number of samples n. Left and right panels show it in linear and logarithmic coordinates respectively. The gray area reflects the variability with respect to the order in which ecosystems are sampled. The red line is the prediction of the Eq. (S10) based on the prevalences f_i of individual species in the pool from our simulations with $\tau = 1$ (see figure 3(A)). The dashed line is set to $N_{\text{pool}} = 1000$, which is the upper bound of $N_{\text{obs}}(n)$.

This expression matches our simulated rarefaction curves quite well (the red solid line indicates the expression using species prevalences and the gray envelope indicates results from several simulated ecosystem samples). In our simulations, $N_{\text{pool}} = 1,000$.

Additionally, note that in case the prevalences f_is are sufficiently dense, we may consider only the prevalence distribution $\mathcal{P}(f)$ instead of this discrete sum. In this case, we get the following integral expression over species prevalences f:

$$N_{\text{obs}}(n) = N_{\text{pool}} \cdot \int_0^1 \mathcal{P}(f) \cdot (1 - e^{-nf}) df. \tag{S11}$$

A practical way to compute it from a known prevalence distribution is to first take a derivative of $N_{\text{obs}}(n)$ with respect to n given by:

$$\frac{dN_{\text{obs}}(n)}{dn} = N_{\text{pool}} \cdot \int_0^1 f \cdot \mathcal{P}(f) \cdot e^{-nf} df, \tag{S12}$$

and then integrate the result over n. Note that $\frac{dN_{\text{obs}}(n)}{dn} \simeq N_{\text{obs}}(n+1) - N_{\text{obs}}(n)$; in other words, the number of new species detected when the number of samples is increased from n to $n+1$. Hence, it stands to reason that it should systematically decrease with n as equation (S12) suggests.

V. MODEL VARIANT WITH HIERARCHY OF NUTRIENTS BASED ON ENERGY OR CARBON CONTENT

In our basic model, for each microbial species that consumes a resource, we pick β byproducts (resulting from resource-to-biomass conversion) randomly from our universe of N_{univ} metabolites. However, one might expect that such secreted byproducts would generally have lower energy or carbon content (per byproduct molecule). To account for this, here we simulate a new variant of the model in which metabolites are arranged hierarchically by their energy content (or equivalently, the number of carbon atoms).

Here we slot all N_{univ} metabolites into five categories (named I, II, III, IV and V) in descending order of each molecule's energy content. We redo our simulations with the following constraint: for each species in the ecosystem, the byproducts it secretes upon consuming a particular metabolite as a resource (this can be from any category, say I) can only have lower energy content than the parent (consumed) metabolite, i.e. it can only be from a category below that of the consumed resource (say only II, III, IV or V). This ensures that no "uphill energy conversion" takes place in the model. We also implemented a plausible biological assumption that the consumption of higher

energy metabolites typically happens with a lower yield (higher α in our model). To account for this we assigned $\alpha = \{0.5, 0.4, 0.3, 0.2, 0.1\}$ to microbes utilizing nutrients from categories I, II, III, IV, V, correspondingly. The central results of our manuscript remain qualitatively (and even quantitatively) unaffected by this modification (see figure S2).

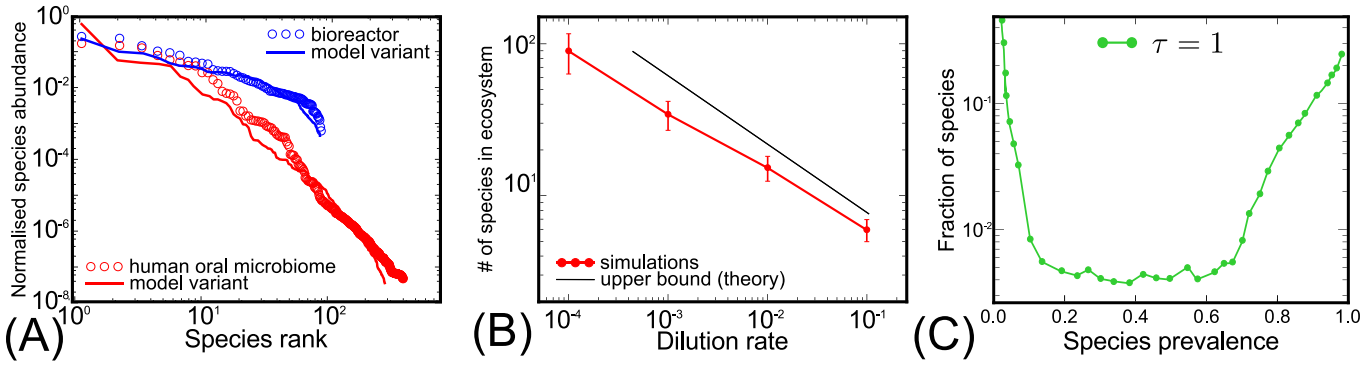

FIG. S2. **Results from simulations of a variant of the model where metabolites are arranged hierarchically by their energy content.** (A) Species abundance-rank distributions for metabolites in 5 categories with values of $\alpha = \{0.5, 0.4, 0.3, 0.2, 0.1\}$ (red) and $\alpha = \{0.9, 0.8, 0.7, 0.6, 0.5\}$ (blue) as in figure 2 in the main text (here $\beta = 2$), and comparison with the data available from the human oral microbiome (red) and methanogenic bioreactors (blue). (B) Scaling of the average number of surviving species in simulated ecosystems and comparison with the theoretical upper bound as in the main text. Here metabolites in 5 categories have values of $\alpha = \{0.5, 0.4, 0.3, 0.2, 0.1\}$ (C) Species prevalence distribution ($\tau = 1$) from simulated ecosystems ($N_{\rm pool} = 1,000$, $\alpha = \{0.5, 0.4, 0.3, 0.2, 0.1\}$, $\beta = 2$). Panels A and B are qualitatively similar to figure 2 for our basic model in the main text. Panel C is qualitatively similar to figure 3(A) for our basic model in the main text.

VI. MODEL VARIANT WITH LIMITED NUMBER OF PATHWAYS PER RESOURCE AND A RATE-YIELD TRADE-OFF

In our basic model, when instantiating species we assign its β byproducts by randomly selecting them from the metabolic universe of N_{univ} metabolites. This is not entirely realistic from the biochemistry standpoint as the number of metabolic pathways utilizing a given resource is finite. For example, Ref. [2] lists 5 different pathways for glucose utilization common to microbes. In general, each of these pathways generates a distinct set of byproducts. To account for this empirical observation we simulated the following variant of our basic model. Each metabolite resource in our model can be utilized via $\eta = 5$ distinct metabolic pathways, each producing its own set of $\beta = 2$ byproducts. Thus when we assign byproducts to organisms we have only 5 distinct choices. That reduces the probability of cascading extinctions in our model. Indeed, if an organism is competitively displaced by another one using the same pathway, the set of byproducts does not change and thus all of the microbes directly or indirectly depending on them for their growth survive.

Furthermore, this version of the basic model also takes into account the rate-yield trade-off reported for microbial species on rather general grounds [3–5]. In a nutshell, the rate-yield trade-off states that a microbial species with a larger value of λ for a given resource would tend to use it less efficiently, i.e. with a lower yield $1 - \alpha$. A classical illustration of this phenomenon is the case of respiration and fermentation glucose utilization pathways: organisms using the former tend to grow slower (smaller λ), but they generate higher yields of ATP (and hence the biomass) than the latter. This trend is reflected in the rules of our model in the following way: the $\eta = 5$ pathways utilizing the same resource have an intrinsic yield hierarchy with $\alpha = \{0.5, 0.4, 0.3, 0.2, 0.1\}$ in pathways arranged in the increasing order of efficiency (yield). The rates λ for an organism is randomly selected from log-normal distribution with standard deviation of $\log \lambda$ equal to $\sigma = 1.5$. The mean values of $\log \lambda$ were determined by their efficiency as $\{2, 1.75, 1.5, 1.25, 1\}$ respectively. This ensures that the least efficient pathway with $\alpha = 0.5$ corresponds to the highest growth rate constant $\lambda = 2$ (on average). In the long run, organisms using these least efficient pathways would tend to dominate the ecosystem [5]. The central results of our manuscript remain qualitatively unaffected by this modification (see figure S3).

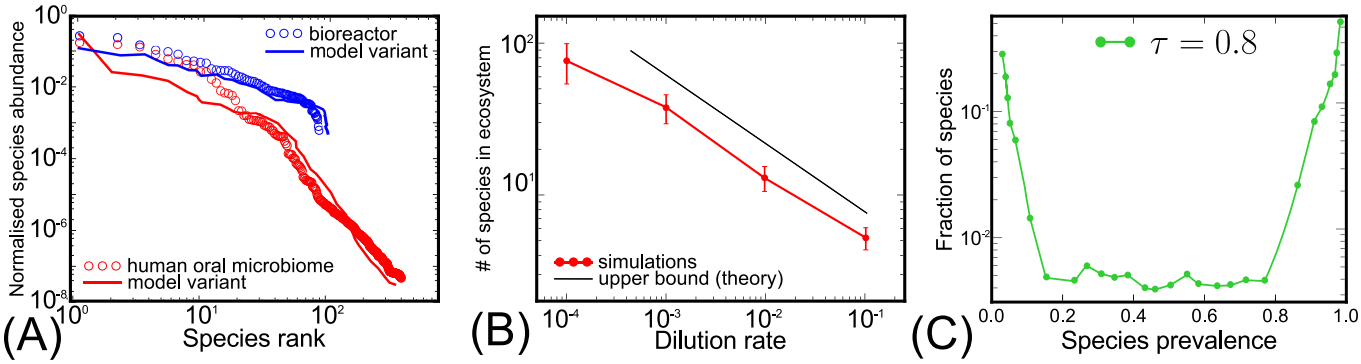

FIG. S3. **Results from a variant of the model with limited number of pathways per resource and a rate-yield tradeoff.** In this variant, every resource can be be utilized via one of $\eta = 5$ distinct pathways. This limits the set of potential byproducts per resource. These 5 pathways are arranged by their efficiency ranging from $\alpha = \{0.5, 0.4, 0.3, 0.2, 0.1\}$ (red in panels A-C) or $\alpha = \{0.9, 0.8, 0.7, 0.6, 0.5\}$ (blue in panel A) (see text for details). (A) Species abundance-rank distributions for metabolites in 5 categories with values of $\alpha = \{0.5, 0.4, 0.3, 0.2, 0.1\}$ (red) and $\alpha = \{0.9, 0.8, 0.7, 0.6, 0.5\}$ (blue) as in figure 2 in the main text (here $\beta = 2$), and comparison with the data available from the human oral microbiome (red) and methanogenic bioreactors (blue). (B) The average number of surviving species as a function of dilution rate. Black line has a power law exponent -0.5. (C) Species prevalence distribution in multiple simulated ecosystems sampled at time $\tau = 0.8$. Here $N_{\rm pool} = 1,000$, $\alpha = [0.1, 0.5]$, $\beta = 2$. Panels A and B are qualitatively similar to figure 2 for our basic model in the main text. Panel C is qualitatively similar to figure 3(A) for our basic model in the main text.

[1] J. Monod, Annual Reviews in Microbiology **3**, 371 (1949).
[2] T. Großkopf and O. S. Soyer, Isme J **10**, 2725 (2016).
[3] T. Pfeiffer, S. Schuster, and S. Bonhoeffer, Science **292**, 504 (2001).
[4] M. Novak, T. Pfeiffer, R. E. Lenski, U. Sauer, and S. Bonhoeffer, Am. Nat. **168**, 242 (2006).
[5] R. E. Beardmore, I. Gudelj, D. A. Lipson, and L. D. Hurst, Nature **472**, nature09905 (2011).