A minimal model for microbial biodiversity can reproduce experimentally observed ecological patterns

Robert Marsland III,* Wenping Cui,† and Pankaj Mehta

Department of Physics, Boston University, Boston, MA

(Dated: May 1, 2019)

Abstract

Surveys of microbial biodiversity such as the Earth Microbiome Project (EMP) and the Human Microbiome Project (HMP) have revealed robust ecological patterns across different environments. A major goal in ecology is to leverage these patterns to identify the ecological processes shaping microbial ecosystems. One promising approach is to use minimal models that can relate mechanistic assumptions at the microbe scale to community-level patterns. Here, we demonstrate the utility of this approach by showing that the Microbial Consumer Resource Model (MiCRM)- a minimal model for microbial communities with resource competition, metabolic crossfeeding and stochastic colonization – can qualitatively reproduce patterns found in survey data including compositional gradients, dissimilarity/overlap correlations, richness/harshness correlations, and nestedness of community composition. By using the MiCRM to generate synthetic data with different environmental and taxonomical structure, we show that large scale patterns in the EMP can be reproduced by considering the energetic cost of surviving in harsh environments and HMP patterns may reflect the importance of environmental filtering in shaping competition. We also show that recently discovered dissimilarity-overlap correlations in the HMP likely arise from communities that share similar environments rather than reflecting universal dynamics. This highlights the promise of minimal models for microbial ecology.

* marsland@bu.edu

† Department of Physics, Boston College, Chestnut Hill, MA

Over the past decade, next-generation sequencing has highlighted the incredible diversity of the microbial ecosystems that fill every corner of our planet. Microbial communities are incredibly complex and and occur in environments ranging from soils to the human body. Large-scale surveys of microbial biodiversity, such as the Earth Microbiome Project (EMP), the Human Microbiome Project (HMP) and the European Metagenomics of the Human Intestinal Tract project (MetaHIT), have revealed a number of robust and reproducible patterns in community composition and function [1–3]. A major challenge for contemporary microbial ecology is to understand and identify the ecological origins of these pattern. This problem is especially difficult because it involves what in the ecology literature has been called the "problem of pattern and scale" [4]: explaining ecological patterns requires connecting processes that occur at very different scales of spatial, temporal, and taxonomical organization.

One potential approach for overcoming the problem of scale is to use mathematical models and large scale simulation to investigate how mechanistic assumptions about environmental and taxonomical structure at the microscopic scale affect the kind of ecological patterns observed at larger scales. A major obstacle in realizing this goal is that any mathematical model that seeks to explain modern microbial sequencing data must deal with the enormous complexity of microbial communities: the numbers of species and consumable molecules in a community can easily reach into the hundreds or thousands [1]. Thus, by necessity any mechanistic model of community assembly will have an extraordinary number of free parameters, presenting a major obstacle for understanding microbial dynamics [5].

One potential strategy for overcoming this difficulty is to exploit the observation that complex systems often have generic behaviors that can be described by sampling parameters from an appropriately chosen random distribution [6, 7]. The most famous example of this is in nuclear physics where the intractably complicated quantum dynamics of the uranium nucleus were successfully modeled using random matrices [6]. Recently, we have adapted these ideas to the microbial setting by formulating a minimal model for microbial population dynamics we term the Microbial Consumer Resource Model (MiCRM) [8–10] (see Figure 1).

The MiCRM builds on the classic framework for resource competition developed by MacArthur and Levins[11]. Like in all consumer resource models, species in the MiCRM are defined by their preferences for resources (Fig. 1d). Species with similar preferences naturally compete with each other, giving rise to competitive exclusion and niche partition-

ing. Crucially, the MiCRM incorporates two additional pieces of biological knowledge that are specific to microbial communities. First, the MiCRM explicitly includes cross-feeding and syntrophy – the consumption of metabolic byproducts of one species by another species [8, 12–14]. This is incorporated into the MiCRM through a stoichiometric metabolic matrix that parameterizes the metabolic transformations of consumed metabolites into secreted byproducts (Fig. 1a,e). Second the MiCRM incorporates stochasticity in dispersal and colonization[15–18]. Due to proximity effects, it is known that new environments are almost always colonized by only a subset of all the species capable of existing in that environment [19]. The MiCRM incorporates stochastic dispersal by seeding new environments through random sampling of a larger regional species pool (Fig. 1b).

Taxonomic and metabolic assumptions are incorporated into the MiCRM through the choice of consumer preferences and metabolic matrices (see Methods and [10] for detailed discussion and implementation details). In the most minimal version of the MiCRM, species have no taxonomic structure (i.e consumer preferences are uncorrelated across species and resources) and metabolism is completely random (i.e. the metabolic matrix has no structure beyond that required by energy and mass conservation). Large-scale surveys such as EMP and HMP often sample communities from very different environmental conditions. For this reason, it is important to be able to incorporate environmental structure and heterogeneity into our models. This is done by choosing which externally supplied resources are present in an environment (Fig. 1c).

Importantly, the MiCRM also allows for the incorporation of additional metabolic and taxonomic structure allowing us to ask how taxonomy and metabolism shape community structure and function. This is implemented in the MiCRM by dividing resources into general resource classes (e.g. sugars, carboxylic acids, lipids, amino acids, etc.) and then using a tiered secretion model where metabolic resources are preferentially secreted into certain resource classes (Methods and [10]). This allows us to incorporate metabolic structure missing in the minimal MiCRM such as the fact that the fermentation of sugars preferentially results in the secretion of carboyxlic acids.

Taxonomic structure can also be easily incorporated into the MiCRM by introducing correlations in species preferences that come from the same family. For example, it is well know that bacteria from the *Enterobacteria* family have a strong preference for fermenting sugars. The MiCRM incorporates such preferences by assigning species to families, with each

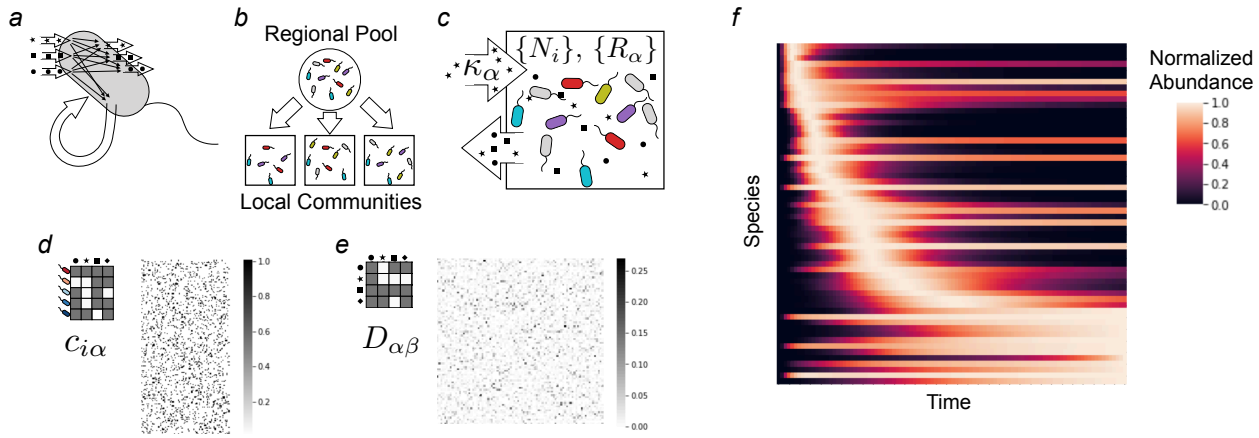

FIG. 1. **A minimal model for investigating microbial biodiversity.** *(a)* The Microbial Consumer Resource Model extends the classic Consumer Resource Model of MacArthur and Levins [11] by incorporating the generic exchange of secondary metabolites observed in microbial communities, as described in the Methods. Each consumed resource type (stars, squares, circles) is transformed into other resource types through metabolic reactions inside each cell, and a fraction of the resulting chemical flux returns to the environment, where it can be consumed by other microbes. *(b)* Communities are initialized by randomly sampling subsets of species from a given regional pool, simulating the effect of stochastic colonization. *(c)* Each community is supplied with a constant influx of specified resource types, and all resources are diluted at a fixed rate. In the simulations reported here, each community is supplied with either one or two resource types, and all other resources are generated as metabolic byproducts. *(d)* Randomly sampled matrix of consumer preferences $c_{i\alpha}$, which determines the composition of the regional pool. *(e)* Randomly sampled metabolic matrix $D_{\alpha\beta}$, which encodes the allowed metabolic transformations and their relative rates. *(f)* Example of colonization dynamics on a single externally supplied resource. The data is normalized by row, so that each white box represents the time at which the species reached its maximum population size.

family preferentially consuming resources from certain resource classes. Importantly, we can control the amount of metabolic and taxonomic structure in the community my modulating just two parameters that control the correlation structures of the consumer preference and metabolic matrix (see Methods and [10]). This addresses the major modeling bottleneck discussed above about how to choose parameters for diverse ecosystems.

In this paper, we use the MiCRM to test simple hypotheses about the mechanistic origins of patterns observed in EMP, HMP and MetaHIT. We find that the MiCRM can qualitatively reproduce observed phenomena with minimum fitting or fine-tuning. We illustrate the utility of the model by identifying ecological mechanisms necessary for reproducing observed patterns and discussing how the MiCRM can be used to identify new patterns and test ecological hypotheses.

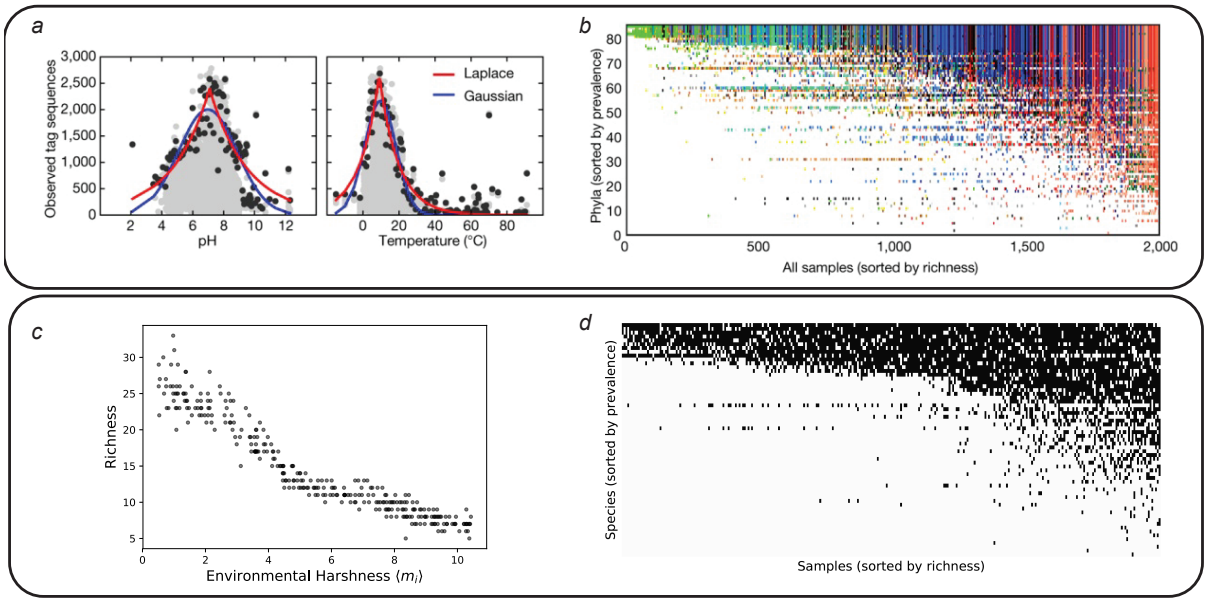

FIG. 2. **Harshness gradient reproduces patterns in global microbial diversity and richness.** *(a)* The number of distinguishable strains observed in each sample of the EMP, plotted vs. pH and temperature. Black dots represent the 99th percentile of all communities at a given pH or temperature. Reproduced from Figure 2 of the initial open-access report on the results of the EMP[1]. *(b)* Presence (colored) or absence (white) of each microbial phylum in a representative set of 2,000 samples from the EMP. Reproduced from Figure 3 of the EMP report [1]. *(c)* The number of species surviving to steady state in simulated communities, plotted vs. environmental harshness. Harsher environments at extreme pH or temperature were simulated by increasing the total amount of resource consumption m_i required for growth (by the same amount for all species). *(d)* Presence/absence of species in simulated communities.

All simulation data and analysis scripts are available at `https://github.com/Emergent-Behaviors-in-Biology/microbiome-patterns`. The model itself is implemented in the freely available Python module Community-Simulator [10] `https://github.com/Emergent-Behaviors-in-Biology/community-simulator`. Since the number of simulations required for comparisons with survey data is necessarily large, our numerical work relies heavily on a novel algorithm implemented in the Community Simulator, which takes advantage of a recently discovered duality between consumer resource models and constrained optimization to quickly and accurately simulate hundreds of communities[20, 21].

RESULTS

Patterns in the Earth Microbiome Project can be explained by energetic costs associated with harsh environments

The Earth Microbiome Project is a systematic attempt to characterize global microbial diversity and function. It consists of over 20,000 samples in 17 environments located on all 7 continents [1]. Recently, a metaanalysis of this data was carried out and several robust patterns were identified. Chief among these was an interesting anti-correlation between richness and environmental harshness reproduced in Figure 2a. Samples near neutral pH or at moderate temperatures ($\sim 15°$C) showed much higher levels of richness than samples from more extreme conditions. Peak richness dropped by a factor of 2 for pHs less than 5 or greater than 9,and temperatures less than 5 °C or greater than 20 °C.

The EMP samples also showed a strongly nested structure: less diverse communities tended to be subsets of the more diverse communities. This is most clearly visible by creating a presence/absence matrix that indicates whether a taxon is present in a sample. Each column in the matrix corresponds to a different sample and each row to a different taxon. When the rows are sorted by taxon prevalence and the columns by richness, as in Figure 2b, one can visually verify that the taxa composing the low-diversity communities are also present in most of the higher-diversity communities.

One possible cause of both these patterns is that microbes require more energy intake to survive in harsher environments[22]. For example, powering chaperones to prevent protein denaturation and running ion pumps to maintain pH homeostasis both require significant amounts of ATP. We hypothesized that increased energy demands could explain the patterns observed in the EMP since they would make it more difficult for obligate crossfeeders to survive on the metabolic byproducts of the primary consumers in a community.

In the MiCRM, the energetic costs of reproduction are encoded in the model parameter m_i, which is the minimal per-capita resource consumption required for net population growth of species i (see Methods for full model equations). The m_i are sampled from a Gaussian distribution with mean 1 and standard deviation 0.01. To vary the harshness of an environment, we added an environment-specific random number m_{env} to the m_i of all species that colonized a given environment. A large m_{env} corresponds to harsh environments

with increased energetic demands whereas small or negative m_{env} corresponds to energetically favorable environments. To mimic the variability in environmental harshness in the EMP, for each community we randomly drew m_{env} uniformly between -0.5 and 10.5.

To selectively test the effect of an energy demand gradient on both alpha (within-sample) and beta (between-sample) diversity, we stochastically colonized 300 simulated communities of 150 species each from a regional pool of 180 species with a chemistry of 90 metabolites. We supplied each community with a constant flux of the same resource type. As discussed above, each of the 300 simulated communities was also assigned a random m_{env} to mimic the effects of environmental harshness on growth rates. The results from this simulation are shown in Figure 2c,d. The same simulation correctly captures both the richness/harshness correlation and the nestedness of the EMP data, suggesting the these large scale patterns may have a simple origin.

We emphasize that these simulations themselves do not confirm the hypothesis that energy gradients are the driver of the observed patterns. Temperature and pH affect microbes in many other ways that are not included in our minimal model. But our simulations do show that accounting for increased energetic costs associated with harsh environments can reproduce the large-scale patterns observed in the EMP even in the absence of any metabolic or taxonomic structure.

Patterns in the HMP can by environmental filtering and competition

The Human Microbiome Project is a large-scale survey of the microbial communities that reside in and on the human body [2]. The HMP was supplemented by the smaller MetaHIT project which focused om sequencing fecal metagenomes from multiple individuals. Initial analysis of the HMP and MetaHIT results on the human microbiome revealed three major patterns, displayed in the top half of Figure 3. First, for a given body site different individuals had very different community compositions. Even at the phylum level, shown in panel *(a)*, the relative abundances of dominant taxa varied dramatically from sample to sample [2]. But samples from different body sites still typically differed more than samples from the same body site, leading to the second pattern, shown in panel (b), of clustering of microbial communities by body site across individuals [2, 3]. Finally, the gradients in the relative abundances of the dominant taxa across a given body site were also visible in

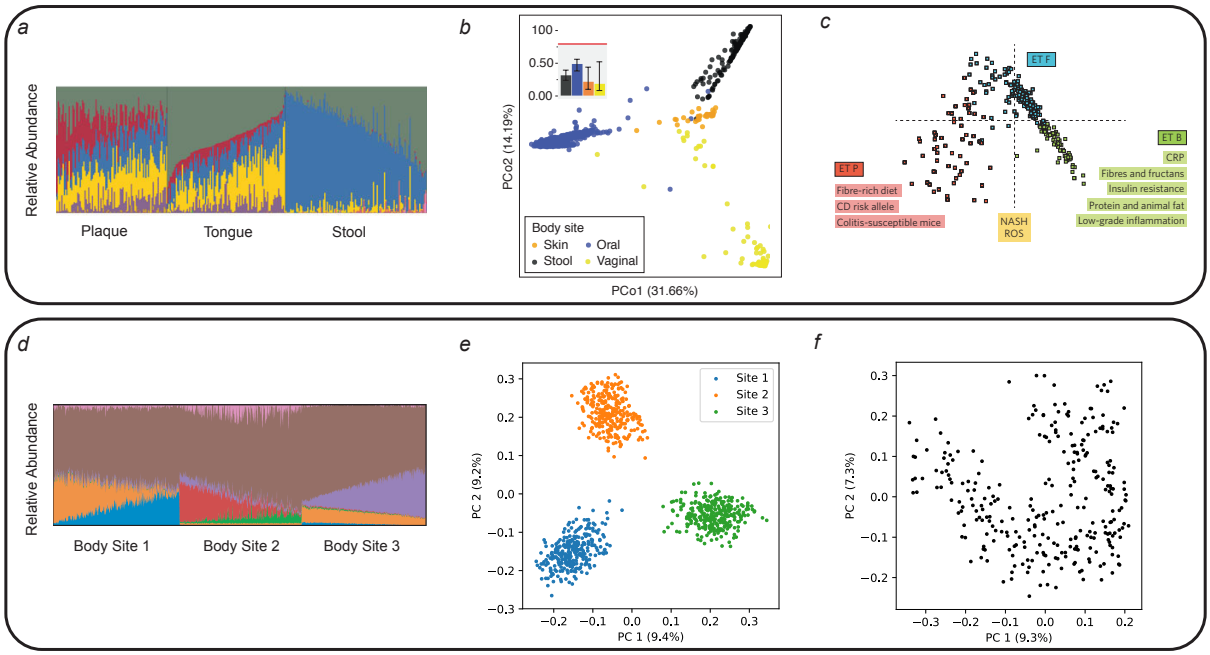

FIG. 3. **Nutrient gradients reproduce patterns in human microbiome survey data.** *(a)* Each column represents one sample from the Human Microbiome Project (HMP). Colored segments represent relative abundances of different phyla in each community. Reproduced from Figure 2 of the initial open-access report on the results of the HMP[2]. *(b)* Principal coordinate analysis (PCoA) of MetaHIT OTU-level community compositions, using the Jensen-Shannon distance metric. Data points are colored by the body site from which the sample was taken. Reproduced with permission from Figure 1 of [23]. *(c)* PCoA of MetaHIT stool samples, with colors showing the three hypothesized enterotypes. Also shown are distinctive features of the gut microenvironment that have been associated with two of the enterotypes. Reproduced with permission from Figure 3 of [23]. *(d)* Each column represents one of 900 simulated samples, each stochastically colonized with 2,500 species from a regional pool of 5,000 species, comprising seven metabolically distinct families. Colored segments represent relative abundances of the seven families. Each of the three "body sites" was supplied with a different pair of resources, with total nutrient supply fixed. Samples are sorted by the relative supply rates of the two resource types. The brown family present in all three environments specializes in the typical byproducts (e.g., carboxylic acids) generated from all the other resource classes. See main text and Methods for simulation details. *(e)* PCoA of species-level compositions of the simulated communities. *(f)* PCoA of simulated samples from Body Site 2.

dimensional reductions of more fine-grained (genus-level) community composition, producing the third pattern shown in panel (c). One-dimensional gradients in composition are known to produce 'u' shaped patterns under standard dimensional reduction techniques [24], and this shape is the main feature visible in the data. There is continuing debate over whether the compositions can be robustly classified into distinct clusters or enterotypes but this is beyond the scope of this this study [23, 25, 26].

As shown in Figure 3c, one factor associated with the compositional gradient is the host's typical diet [23, 26]. Different kinds of externally supplied nutrients, such as fibers and proteins, are thought to encourage growth of different microbial taxa. For this reason, we hypothesized that the patterns in the HMP may arise from heterogeneity in the resources available in different environments. It is clear that reproducing the HMP patterns requires assuming some minimal level of taxonomic and metabolic structure (for example see Fig 3a,c). For this reason, in our simulations we divided resources into six resource classes and species into six families, with each family specializing in one resource class. Each of the three "body sites" was supplied by a unique pair of resources from distinct resource classes (i.e if we label the six resources classes by $A - F$, body site 1 was supplied with a resource from class A and a resource from class B, body site 2 with a resource from class C and a resource from class D, and body site 3 with a resource from class E and a resource from class F). We modeled variability in the availability of resources across individuals at a fixed body site by changing the ratio of the two supplied resources while holding the total supplied energy fixed (see Methods).

To mimic the scale of the actual microbiome data, we generated a regional pool of 5,000 species (approximately the number of OTU's identified in the HMP [2]), and stochastically colonized 300 samples per body site with 2,500 species each. Figure 3d shows the resulting pattern, which mimics the broad range of compositions found in the data. Panel (e) confirms that the compositions from different body sites remain well-separated, and panel (f) shows the expected 'u' shape in the dimensional reduction of compositions from a single body site. This suggests that the patterns found in Fig 3 may reflect the combined effects of environmental filtering and competition between species in the presence of environmental gradients.

Dissimilarity-overlap patterns reflect shared environments not universal dynamics

Another pattern obtained from a more recent analysis of the HMP data is an anti-correlation between overlap and dissimilarity of pairs of communities from a given body site (see Figure 4 and [27] for details). Due to both stochastic colonization and variable environments, there are usually many species in one sample that are not present in the other. Different pairs of samples overlap to different degrees, and this overlap can be measured in

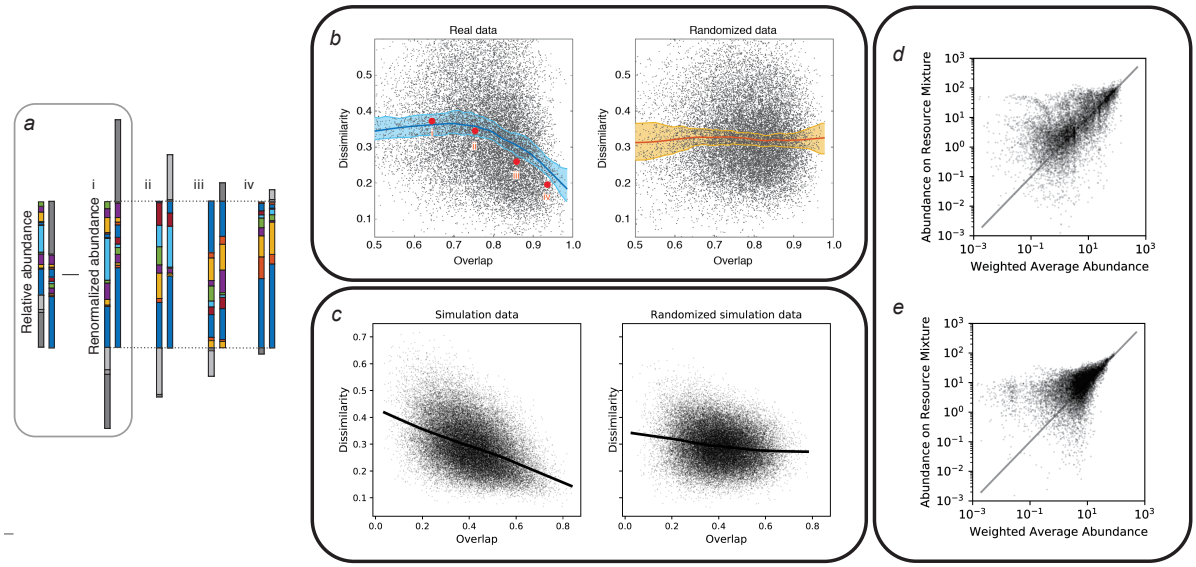

FIG. 4. **Minimal model generates additional predictions about effects of nutrient gradient.** *(a)* The composition of pairs of samples can be compared in two independent ways: "overlap" measures the fraction of each sample comprised by species common to both, and "dissimilarity" measures how different the relative abundance profiles are within this shared pool. The four pairs shown here have increasing overlap and decreasing dissimilarity from left to right, corresponding to the four points indicated in the next panel. *(b)* Dissimilarity and overlap of 17,955 pairs of stool samples from the HMP, analyzed at the genus level. Also shown are the dissimilarity and overlap of randomized data for comparison. Reproduced along with panel *a* with permission from [27]. *(c)* Dissimilarity and overlap of 40,186 sample pairs from the 300 simulated samples of Body Site 2. Also shown are the dissimilarity and overlap of randomized data for comparison. *(d)* Abundance of each species in each simulated sample from Body Site 2, versus the weighted average of the abundance in the two extreme samples along the nutrient gradient. See main text and Methods for details. *(e)* Same as previous panel, but for data generated with unstructured $c_{i\alpha}$ and $D_{\alpha\beta}$.

terms of the ratio of the combined population of the shared species to the total population of the two samples. If one focuses on the subset of species that are shared, one can also compare the relative abundance distributions of the two samples within this shared pool, using standard measures of dissimilarity such as the Jensen-Shannon divergence (see [27] for detailed discussion). These two quantities are not intrinsically related, as can be seen by evaluating them over a randomly generated table of abundances (cf. Figure 4b).

This analysis was initially proposed as a way of distinguishing between "universal" and "host-specific" microbial dynamics. It was argued that if the dynamics of human associated microbial communities were universal, different individuals could be modeled with the same dynamic parameters and this would be reflected by a negative correlations between dissimilarity and overlap across cross-sectional samples. In contrast, for host-specific dynamics

each individual would have their own kinetic parameters and the dissimilarity and overlap would be uncorrelated.

We re-ran the analysis of [27] on our simulated HMP data discussed above and found that the dissimilarity and overlap were negatively correlated at a single body site just as in the real gut microbiome data. (see Figure 4c). However, this correlation was absent if we analyzed pairs of samples from distinct body sites, indicating that this signature likely arises due to the fact the all the communities at a given body site exist in a similar external environment (see `plots/DOC_all.pdf` in the accompanying GitHub repository). Importantly, our simulations show that the negative dissimilarity-overlap correlation observed [27] can be found even in the absence of universal dynamics since environments with different amounts of externally supplied sources generically give rise to communities with different ecological dynamics. Instead, our simulations suggest that the negative correlation between overlap and dissimilarity found in the HMP may reflect the fact that communities at a given body-site experience similar but not identical environments.

"Modular assembly" of microbial communities

Our analysis of our synthetic HMP data also shows a new pattern: the family-level composition of each community along the nutrient gradient is approximately a linear combination of the compositions of the two extreme communities on the gradient (Figure 3d). Quantitatively, if N_i^1 are the population sizes for each family i in the community supplied with a flux κ_1 of resource type 1 alone, and N_i^2 are the population sizes in the community supplied with flux κ_2 of resource type 2 alone, then the community supplied with flux $(1-\alpha)\kappa_1$ of resource type 1 and flux α of resource type 2 has population sizes approximately equal to $(1-\alpha)N_i^1 + \alpha N_i^2$. This "additivity" of communities from different environments can also be seen at the species level, by plotting the actual population sizes N_i^{mix} versus the weighted average prediction, as shown in Figure 4d. The correlation between the actual and predicted population sizes in the simulated HMP data has an R^2 of 0.48.

Since this analysis is performed at the species level, meaningful taxonomic groupings are no longer necessary, so we also checked for additivity in simulations where communities lack any metabolic or taxonomic structure (i.e with unstructured metabolic and consumer preference matrices). Figure 4e shows that the population sizes in the mixed-resource communities

are still well predicted by the weighted average of single-resource communities, with an R^2 of 0.59. This suggests that metabolic and taxonomic structure are not necessary to see this additive pattern and additivity may occur quite generically.

This additive pattern is difficult to test for in field survey data, since there are many additional factors besides diet that vary from sample to sample, many of which may themselves be correlated with diet. Studying this kind of effect requires controlled experiments where the variable of interest can be systematically varied. One recent experiment colonized seawater communities on small beads composed of different kinds of carbohydrates, which served as the sole externally provided carbon source for the community [28]. This simplified scenario reflects the conditions of our minimal model more closely, where only one or two nutrient types were externally supplied. The authors of [28] compared the weighted average of population sizes from communities grown on two different carbon sources, such as agarose and alginate, and the corresponding population sizes from other communities grown on a mixture of the two. They found the same "additivity" effect we observe in our simulations, with even stronger R^2 values of 0.84 and 0.74 for two different carbon source pairs. They termed this property "modular assembly" of microbial communities. Our simulations show that modular assembly may be a generic property of complex microbial communities grown in the presence of multiple, metabolically distinct resources.

DISCUSSION

We have shown that the Microbial Consumer Resource Model introduced in [8] to describe laboratory experiments in synthetic minimal environments can also reproduce a wide range of experimentally observed patterns in survey data such as the HMP and EMP including harshness/richness correlations, nestedness of community composition, compositional gradients, and dissimilarity/overlap correlations. The MiCRM provides a systematic way of exploring the effect of stochastic colonization, resource competition, and metabolic cross-feeding on large-scale observables. By randomly sampling parameters from well-defined probability distributions, we combine a sufficient level of mechanistic detail to make the parameters physically meaningful, while keeping the number of parameters small enough for systematic investigation of the factors that control different patterns.

Our empirical results complement recent theoretical works suggesting that complex

ecosystems may still be well described by random ecoystems [29, 30], suggesting the essential ecology of diverse ecosystems may be amenable to analysis using techniques from statistical mechanics and random matrix theory. For these reasons, the MiCRM is well-suited to serve as a minimal model for understanding microbial ecology.

Our analysis also suggests several hypothesis relating mechanism to large scale patterns in both the EMP and HMP. We have shown that it is possible to reproduce the richness/harshness correlation and the nestedness of the EMP data by assuming that harsh environments pose an additional the energetic cost to organisms. This is true even when communities are grown in otherwise identical environments and lack any taxonomic and metabolic structure. This complements earlier work showing that energy availability is a key driver of community function and structure [9]. Our simulations on the HMP suggest that environmental gradients and resource availability result in significant environmental filtering and naturally explain the clustering of microbial communities by body site. Within similar environments, competition still plays a major role in determining community structure.

Great care has to be taken when interpreting large-scale patterns. For example, the negative correlation between dissimilarity and overlap observed in HMP data in [27] may be indicative of the fact that body-sites across individuals have similar environments rather than a much stronger claim of universal dynamics in the human microbiome. Our work also suggests that many large scale patterns may occur generically across different environmental settings. For example, we have shown that the additivity observed in our synthetic HMP data is also observed in ocean communities grown on synthetic carbon beads [28], suggesting modular assembly may be a generic property of communities grown in environments with metabolically distinct resources.

The analysis presented here shows that it is possible to qualitatively reproduce patterns seen in large-scale surveys such as the EMP and HMP using a simple minimal model. An interesting area of future research is to move beyond qualitative comparisons and ask how minimal models and large scale simulations can be quantitatively compared to large-scale genomic surveys. This problem is especially challenging given the large number of parameters, environments, and experimental designs that must be explained. One potential avenue for doing this is to use statistical methods such as Approximate Bayesian Computation (ABC) [31]. In ABC, the need to exactly calculate complicated likelihood functions is replaced

with the calculation of summary statistics and numerical simulations. In this way, it may be possible to quantitatively relate mechanistic details at the level of microbes to community level patterns observed in large-scale surveys.

METHODS

All synthetic data was generated using the Microbial Consumer Resource Model previously described[9, 10]. We found the fixed points of the dynamics for each community using the Python package Community Simulator [10]:
https://github.com/Emergent-Behaviors-in-Biology/community-simulator.
Principal Coordinate Analysis was performed on the simulated HMP data using the Python package scikit-bio http://scikit-bio.org/. The pairwise distance matrix was generated using standard scipy commands with the Jensen-Shannon distance metric. Dissimilarity-overlap analysis was performed on the simulation data following the procedure described in [27]. All simulation data and analysis scripts are available at https://github.com/Emergent-Behaviors-in-Biology/microbiome-patterns.

A. MiCRM Dynamical Equations

We consider the dynamics of the population sizes N_i of S microbial species and the abundances R_α of M resource types in a well-mixed system, governed by the following set of ordinary differential equations:

$$\frac{dN_i}{dt} = g_i N_i \left[\sum_\alpha w_\alpha (1 - l_\alpha) c_{i\alpha} R_\alpha - m_i \right] \tag{1}$$

$$\frac{dR_\alpha}{dt} = \kappa_\alpha - \tau_R^{-1} R_\alpha - \sum_i c_{i\alpha} N_i R_\alpha + \sum_{i,\beta} D_{\alpha\beta} \frac{w_\beta}{w_\alpha} l_\beta c_{i\beta} N_i R_\beta. \tag{2}$$

For this study, the conversion factors g_i from consumption to population growth were all set to 1, as were the resource qualities w_α and the resource dilution rate τ_R^{-1}. The leakage fractions l_α govern how much of each consumed resource is released into the environment as metabolic byproducts, and was set to 0.8 for all α.

B. Synthetic data for EMP

Synthetic data for the EMP figure was generated using a regional species pool of size $S = 180$ and $M = 90$ potential resources. For the simulations of Figure 2, the elements of the 180×90 consumer preference matrix $c_{i\alpha}$ were sampled from a binary distribution as previously described[9, 10], with $c_0 = 0$, $c_1 = 1$ and $\mu_c = 10$, using only one resource class ($T = 1$) and one consumer family ($F = 1$). The 90×90 metabolic matrix $D_{\alpha\beta}$ was sampled from a Dirichlet distribution as previously described[9, 10], with $d_0 = 0.05$. The m_i were sampled from a Gaussian distribution with mean 1 and standard deviation 0.01. A random number from a uniform distribution between -0.5 and 10.5 was added to all the m_i's from each sample in the EMP simulation. Finally, the κ_α's were all set to 0, except for the single resource externally supplied to each community. The nonzero κ_α was set equal to 200.

C. Synthetic data for HMP

To generate synthetic data for the HMP figure, we assumed a regional species pool of size $S = 5000$ and $M = 300$ resources. For the simulations of Figure 3, a $5,000 \times 300$ consumer preference matrix $c_{i\alpha}$ was sampled from a structured binary distribution as previously described[9, 10], with $T = 6$ classes of 50 resource types each, and $F = 6$ corresponding "families" of 800 species each, with each family specializing in one resource class. The strength of specialization was set to $q = 0.9$, and μ_c, c_0 and c_1 were the same as for the previous simulations. The 300×300 metabolic matrix $D_{\alpha\beta}$ was sampled from a structured Dirichlet distribution as previously described, with $d_0 = 0.3, f_s = 0.45, f_w = 0.45$. The m_i were sampled from a Gaussian distribution with mean 1 and standard deviation 0.01. In each of the three environments, one resource was chosen from each of two resource classes. Each community was supplied a total flux $\kappa = 1000$, with $(1 - \alpha)\kappa$ supplied from the first resource, and $\alpha\kappa$ supplied from the second. The 300 communities in each environment had α values linearly sampled from 0 to 1.

---

[1] Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457 (2017).

[2] Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207 (2012).

[3] Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59 (2010).

[4] Levin, S. A. The problem of pattern and scale in ecology: the robert h. macarthur award lecture. *Ecology* **73**, 1943–1967 (1992).

[5] Hart, S. F. *et al.* Uncovering and resolving challenges of quantitative modeling in a simplified community of interacting cells. *PLoS biology* **17**, e3000135 (2019).

[6] Wigner, E. P. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics (ser. 2)* **62**, 548 (1955).

[7] May, R. Will a Large Complex System be Stable? *Nature* **238**, 413 (1972).

[8] Goldford, J. E. *et al.* Emergent simplicity in microbial community assembly. *Science* **361**, 469 (2018).

[9] Marsland III, R. *et al.* Available energy fluxes drive a transition in the diversity, stability, and functional structure of microbial communities. *PLOS Computational Biology* **15**, e1006793 (2019).

[10] Marsland III, R., Cui, W., Golford, J. & Mehta, P. The Community Simulator: A Python package for microbial ecology arXiv:1904.09367 (2019).

[11] MacArthur, R. Species Packing and Competitive Equilibrium for Many Species. *Theoretical Population Biology* **1**, 1 (1970).

[12] Harcombe, W. R. *et al.* Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Reports* **7**, 1104 (2014).

[13] Zomorrodi, A. R. & Segrè, D. Synthetic ecology of microbes: Mathematical models and applications. *J. Mol. Biol.* **428**, 837 (2016).

[14] Pacheco, A. R., Moel, M. & Segrè, D. Costless metabolic secretions as drivers of interspecies interactions in microbial ecosystems. *Nature Communications* **10**, 103 (2019).

[15] Leibold, M. A. *et al.* The metacommunity concept: a framework for multi-scale community ecology. *Ecology Letters* **7**, 601 (2004).

[16] Vellend, M. *The Theory of Ecological Communities (MPB-57)*, vol. 75 (Princeton University Press, 2016).

[17] HilleRisLambers, J., Adler, P., Harpole, W., Levine, J. & Mayfield, M. Rethinking community assembly through the lens of coexistence theory. *Annual Review of Ecology, Evolution, and Systematics* **43** (2012).

[18] Dini-Andreote, F. & Raaijmakers, J. M. Embracing community ecology in plant microbiome research. *Trends in Plant Science* **23**, 467–469 (2018).

[19] Shurin, J. B. Dispersal limitation, invasion resistance, and the structure of pond zooplankton communities. *Ecology* **81**, 3074 (2000).

[20] Mehta, P., Cui, W., Wang, C.-H. & Marsland III, R. Constrained optimization as ecological dynamics with applications to random quadratic programming in high dimensions. *Physical Review E* **(in press)** (2018).

[21] Marsland III, R., Cui, W. & Mehta, P. The minimum environmental perturbation principle: A new perspective on niche theory arXiv:1901.09673 (2019).

[22] Hoehler, T. M. & Jørgensen, B. B. Microbial life under extreme energy limitation. *Nature Reviews Microbiology* **11**, 83 (2013).

[23] Costea, P. I. *et al.* Enterotypes in the landscape of gut microbial community composition. *Nature Microbiology* **3**, 8 (2018).

[24] Podani, J. & Miklos, I. Resemblance coefficients and the horseshoe effect in principal coordinates analysis. *Ecology* **83**, 3331–3343 (2002).

[25] Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174 (2011).

[26] Gorvitovskaia, A., Holmes, S. P. & Huse, S. M. Interpreting prevotella and bacteroides as biomarkers of diet and lifestyle. *Microbiome* **4**, 15 (2016).

[27] Bashan, A. *et al.* Universality of human microbial dynamics. *Nature* **534**, 259 (2016).

[28] Enke, T. N. *et al.* Modular assembly of polysaccharide-degrading microbial communities in the ocean. *bioRxiv* 387191 (2018).

[29] Barbier, M., Arnoldi, J.-F., Bunin, G. & Loreau, M. Generic assembly patterns in complex ecological communities. *Proceedings of the National Academy of Sciences* **115**, 2156–2161 (2018).

[30] Cui, W., Marsland III, R. & Mehta, P. Diverse communities behave like typical random ecosystems arXiv:1904.02610 (2019).

[31] Csilléry, K., Blum, M. G., Gaggiotti, O. E. & François, O. Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution* **25**, 410–418 (2010).