Emergent predictability in microbial ecosystems

Jacob Moran
Department of Physics, Washington University in St Louis, St Louis, MO, USA
Department of Biophysics, University of Michigan, Ann Arbor, MI, USA

Mikhail Tikhonov*
Department of Physics, Washington University in St Louis, St Louis, MO, USA
(Dated: March 27, 2024)

Microbial ecosystems carry out essential functions for global climate, human health, and industry. These complex communities exhibit a surprising amount of functionally relevant diversity at all levels of taxonomic resolution, presenting a significant challenge for most modeling frameworks. A long-standing hope of theoretical ecology is that some patterns might persist despite community complexity – or perhaps even emerge because of it. A deeper understanding of such "emergent simplicity" could enable new approaches for predicting the behaviors of the complex ecosystems in nature. However, most examples described so far afford limited predictive power, as they focused on reproducibility rather than prediction. Here, we propose an information-theoretic framework for defining, nuancing and quantifying emergent simplicity in empirical data based on the ability of simple models to predict community-level functional properties. Applying this framework to two published datasets, we demonstrate that the majority of properties measured across both experiments exhibit robust evidence of emergent predictability: surprisingly, as community richness increases, simple compositional descriptions become more predictive. We show that this behavior is not typical within the standard modeling frameworks of theoretical ecology, and argue that improving our ability to predict and control natural microbial communities will require a shift of focus: away from complexity of *ecosystems*, and towards prediction complexity of *properties* of ecosystems.

Microbial communities perform key functions in a wide range of contexts, including human health, agriculture, and the global climate. A key goal of microbial ecology has thus been to predict community function from its composition. Recent advances in high-throughput sequencing technologies provide a high taxonomic resolution of the composition of microbial communities, revealing that these communities consistently exhibit high richness (number of coexisting taxa), a surprising diversity at all levels of resolution, high functional redundancy, and a lack of reproducibility in even similar conditions. These characteristics pose significant challenges for making quantitative predictions, which are necessary for applications in industry and medicine.

A long-standing hope of theoretical ecology is that, despite this complexity, some patterns might nevertheless be predictable. The search for emergent regularities in complex ecosystems is as old as ecology itself. For example, classical work uncovered reproducible trophic structures across taxonomically divergent communities [1] or a consistent ratio of predator species to prey species [2, 3]. More recently, moving beyond the patterns persisting despite high ecosystem diversity, it was proposed that some regularities could perhaps arise *because* of it, coining the term "emergent simplicity" [4–8]. This notion is at the foundation of the interface between statistical physics and the so-called large-N ecology, explaining species abundance distributions and other patterns as some parameter becomes large [9–16].

The idea of emergent simplicity is very appealing, suggesting that functional predictions could perhaps be made directly in the complex natural context [17]. However, so far this hope remains out of reach. Most reported instances of "emergent simplicity" fall under the umbrella of functional convergence or self-averaging, focusing on reproducibility rather than prediction [1–8]. Unfortunately, such reproducibility is often approximate or limited to simplified conditions [18], and even when it holds, offers limited predictive power. As an illustration, consider the statement: "if the soil ecosystem is diverse enough, the rate of process P will be $\approx X$ per gram of biomass". This blanket expectation would likely be somewhat naïve, but even if it were true, it would leave key questions unanswered – such as how the process rate would be affected by a perturbation, or what intervention would help increase it. As a result, observing emergent regularities of this kind has generally not led to improved functional predictions.

Understanding and harnessing emergent simplicity could transform our ability to predict and control complex natural communities. However, to realize this potential we require a framework focused on *prediction* rather than reproducibility. Here, we introduce an information-theoretic framework to explicitly quantify the predictive power of simple models in ecological data. This allows us to convert the aspiration of emergent simplicity into a testable quantitative hypothesis. Applying our framework to two published datasets on synthetic ecosystems, we show that, remarkably, the predictive power of simple models indeed improves with community richness, this improvement strongly supported for 4 out of 5 proper-

* tikhonov@wustl.edu

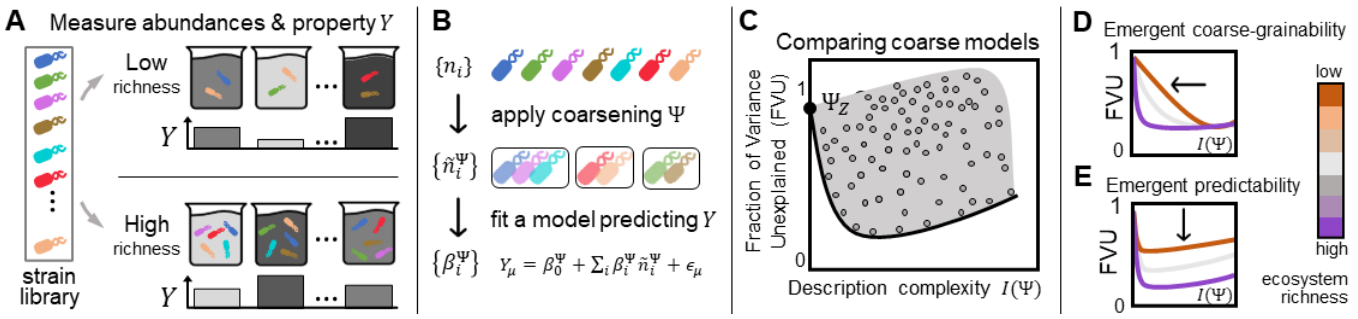

FIG. 1. **A quantitative framework for interrogating emergent simplicity in data.** **A:** A general experimental protocol with community richness (number of strains) as a control parameter. A fixed library of strains is subsampled to assemble communities of low richness (top) or high richness (bottom). Community composition (strain abundances, n_i) and some property of interest Y are measured in each community. **B:** Our approach will be to train models predicting Y based on combined abundances of groups of taxa. Each choice of a *coarsening* Ψ defines a regression model (coefficients $\{\beta_i^\Psi\}$), whose performance we will evaluate. **C:** For each coarsening Ψ, we can evaluate the *description complexity* $I(\Psi)$ (the extent of coarsening; see text) and the out-of-sample prediction error, quantified as the fraction of variance unexplained (FVU) by the corresponding regression model. The cartoon illustrates the expected general shape of the scatter plot in these axes, evaluated for all possible coarsenings Ψ (points). Our focus is on the shape of the Pareto front (thick line), characterizing the best predictive power achievable by models of a given complexity. As community richness increases, the phenomenon of emergent simplicity is identified by the behavior of the Pareto fronts, and can take one of two distinct forms: **D:** emergent coarse-grainability, or **E:** emergent predictability (vertical arrow). We stress that our normalization—focusing on the *fraction* of variance unexplained, equivalent to normalizing all observables to unit variance—ensures that both forms of emergent simplicity defined here are distinct from the simplest law-of-large-numbers intuition (a.k.a. central limit theorem or "self-averaging"); see text.

ties measured across both experiments. These results provide support for emergent simplicity as a valid path towards predictions, one for which community richness is an asset rather than a nuisance.

I. THE THEORETICAL FRAMEWORK

A. A model experimental setup

For concreteness, consider the experimental setup depicted in Fig. 1A. Let $\mathcal{L}$ be a fixed library of S strains, indexed by i. Subsets of these strains are sampled from this library to inoculate N communities, indexed by μ and defined by a binary matrix $b_{i\mu}$ identifying whether strain i is included in community μ. We will assume that community richness $R_\mu \equiv \sum_i b_{i\mu}$ is kept consistent across the dataset; this richness will become our key control parameter for comparisons (a low-richness dataset *vs.* a high-richness dataset). Once the communities are assembled, we measure their microscopic composition (abundance of each strain $n_{i\mu}$) as well as some community-level property of interest Y_μ that we seek to predict (e.g., production of a metabolite). We will not assume that the communities reach any notion of a steady state, merely that some assembly protocol is followed consistently in each trial.

The simplest, and most familiar, notion of emergent simplicity is self-averaging. As community richness increases, one might observe the distribution of Y_μ become increasingly tight. Much previous work suggests that certain properties of high-richness communities may indeed exhibit such "functional convergence." [1–8, 14] However, for the purposes of this work, what we care about is not the total variance of Y_μ, but our ability to explain it from community composition. Our goal is to quantify the *predictive power of simple models* in ecological datasets, and then use empirical data to ask how this predictive power changes with community richness.

B. The candidate 'simple models'

Our approach is motivated by an empirical example, namely wastewater treatment bioreactors that use microbes to digest organic waste into methane. These complex communities assemble semi-naturally and can harbor over 1000 taxa [19]. The rate of digestion (the amount of methane produced) can vary widely [20, 21], but, remarkably, can be explained by surprisingly simple models coarse-graining this vast diversity into just 4 functional classes (acidogens, propionate-degrading acetogens, butyrate-degrading acetogens, and methanogens [22]).

Motivated by this example, throughout this work, our candidate simple models will be indexed by the choice of a *coarsening* Ψ, which maps individual strains into a smaller number of classes: $[1 \ldots S] \xrightarrow{\Psi} [1 \ldots K^\Psi]$ with $K^\Psi \leq S$ (Fig. 1A). Some natural choices of coarsening Ψ could be to group strains based on taxonomy or certain functional traits, but we will allow any choice of Ψ as a valid coarsening.

For any Ψ, we will assess the prediction error of the best model from a specified *model class* $\mathcal{M}$, taking as input the combined abundances $\tilde{n}^{\Psi}_{j\mu} = \sum_{\Psi(i)=j} n_{i\mu}$. This approach echoes other recent attempts to relate community-level properties to coarse-grained compositional features [23, 24]. The model class $\mathcal{M}$ could be, for instance, the class of Generalized Lotka-Volterra models, or some general-purpose ansatz (e.g., training a neural network). In this work, we will pick a particularly simple class of models, namely linear regressions $Y_\mu = \beta^{\Psi}_0 + \sum_i \beta^{\Psi}_i \tilde{n}^{\Psi}_{i\mu} + \epsilon_\mu$ (where ϵ_μ is the residual noise). This choice is motivated by two considerations. First, we will want to apply our framework to data, and as we will see, the limitations of existing datasets make it preferable to pick a model class trainable on fewest data points. Second, it was recently shown that simple regressions can in fact be surprisingly predictive of function [25], motivating us to proceed.

In summary, the candidate "simple models" we will consider are indexed by the choice of a coarsening Ψ. Each Ψ defines a set of coarsened variables ($\tilde{n}^{\Psi}_{i\mu}$), which allow us to train an instance of a linear regression model (coefficients β^{Ψ}_i), whose prediction error we will evaluate.

It is worth noting that a common approach to vary modeling complexity is to keep the variables fixed (e.g. "foxes", "rabbits"), but change the model class (e.g, a non-interacting model; adding quadratic Lotka-Volterra interactions; then higher-order terms; etc.). In contrast, the defining feature of our approach is to fix the model class, and change the variables (vary the extent of taxon grouping into putative functional classes). Our next step is to define quantitative metrics for the performance and complexity of each model thus constructed.

C. Quantifying the performance of simple models

For each candidate "simple model" (coefficients β^{Ψ}_i) indexed by the coarsening Ψ, we will evaluate two features. First, we will compute its out-of-sample *prediction error*. Since we explicitly want to go beyond simple self-averaging, we will normalize away the overall width of the distribution of Y, and report the *fraction* of unexplained variance (FVU), defined as the mean square error of prediction divided by the variance of Y, and confined between 0 and 1.

Second, we will evaluate the *description complexity* $I(\Psi)$, which we define as the mutual information between the label of the coarse-grained class and the identity of the microscopic strain:

$$I(\Psi) = -\sum_{j=1}^{K^{\Psi}} \frac{S^{\Psi}_j}{S} \log \frac{S^{\Psi}_j}{S}.$$

Here S^{Ψ}_j is the number of strains that Ψ assigns to class j; K^{Ψ} is the total number of classes; and S is the total number of strains in the library. Our definition of $I(\Psi)$ implements the intuition that description complexity should grow with the number of classes it distinguishes, but generalizes it to groups of unequal size (for additional discussion, see SI section "Quantifying description complexity"). This metric $I(\Psi)$ reaches its maximum for the fully microscopic description, when each strain is its own class: $K^{\Psi} = S$ and $I(\Psi) = \log S$.

If we were to evaluate description complexity $I(\Psi)$ and the fraction of unexplained variance for all possible coarsenings Ψ, the result would be expected to look like Fig. 1C. Each point in the point cloud corresponds to one choice of Ψ, and describes the predictive power of one candidate simple model.

The left-most point Ψ_z is the 'zero-information' description that groups all strains together. If no compositional information is available (only the total biomass), we expect the fraction of unexplained variance—the ordinate of Ψ_z—to be close to 1. (If Ψ_z is notably below 1, i.e., the chosen property obviously scales with biomass, it might be more meaningful to seek to predict its value per unit biomass; see SI section "Normalization of predictor and response variables".) As the information content of a description increases, better predictions become possible. Models that provide the lowest prediction error for a given complexity constitute the Pareto front (thick line). The Pareto front continues to trend downwards until descriptions become too complex to be fit on limited data (overfitting), and out-of-sample prediction error starts increasing (Fig. 1C). From now on, our focus will be on the properties of this Pareto front.

D. Distinguishing emergent coarse-grainability and emergent predictability

By construction, the point cloud in Fig. 1C, and thus the shape of the Pareto front, explicitly depend on the dataset on which the predictive power is being evaluated. Our goal is to compare the Pareto front constructed for a dataset consisting of high-richness examples *vs.* low-richness examples (Fig. 1A).

We distinguish two behaviors of interest (Fig. 1D). As community richness is increased, we might observe the Pareto front shifting to the left. This would mean that the same prediction accuracy can be achieved by increasingly coarse descriptions ('emergent coarse-grainability'). Separately, we might observe the Pareto front shifting down. This would mean that models of comparable complexity become increasingly predictive ('emergent predictability').

Both phenomena could be collated under the umbrella term 'emergent simplicity,' but are clearly distinct. A given system could exhibit only one of these, or both, or neither. Having built a theoretical framework to distinguish them, we now turn to asking whether these behaviors are expected (based on models), and whether they are observed (based on empirical data).

II. RESULTS

A. Standard ecological models do not generically exhibit either form of emergent simplicity

Before we apply our framework to real data, it is helpful to establish whether observing the behaviors of Fig. 1D would be surprising. For this, we consider two classic models widely used in theoretical ecology: the Generalized Lotka-Volterra model (GLV) with a random interaction matrix, and a Consumer-Resource model (CRM) with a random resource consumption matrix.

To apply our framework, we need to specify the property Y we would seek to predict. Here, we focus on community invadeability by a specified strain. This choice is convenient because it is a relevant and non-trivial community property that can be accessed in both modeling frameworks.

The exact simulation protocol and parameter choices are described in the SI, but our overall approach is similar for both models, and proceeds as follows: **(1) Construct a species pool**, by starting from a random instance of a larger ecosystem and simulating its dynamics to obtain a coexisting set of about 25 species. **(2) Define an "invader,"** drawing a random new species from the same ensemble. **(3) Generate the synthetic dataset.** We simulate the experimental protocol of Fig. 1A, with 1000 samples for each of three tiers of community richness: low richness, 1-5 strains; middle richness, 10-15 strains; and high richness, 20-25 strains. These tiers were chosen to match the empirical dataset of Ref. [26] used below. For each community, we first record the equilibrium abundances of all strains as $n_{i\mu}$, and then simulate the outcome of its invasion by the pathogenic strain, recording the post-invasion abundance of the pathogen as Y_μ. Finally, we **(4) construct the information-prediction Pareto fronts** defined above, using a heuristic algorithm to search the space of possible coarsenings (for details, see SI section 6). The results, averaged over 10 random choices of the invader, are presented in Fig. 2A (GLV model) and Fig. 2B (CRM model).

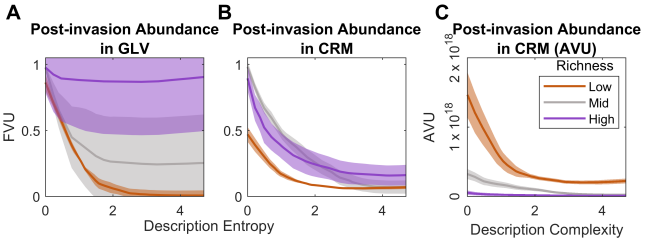

FIG. 2. **Standard ecological models exhibit self-averaging, but not the nuanced forms of emergent simplicity defined here.** As an example, we consider the problem of predicting ecosystem resistance to an invader. **A, B:** Applying our framework to simulated data from two standard ecological models: the Generalized Lotka-Volterra model (panel A) and the Consumer Resource model (panel B), with the observable of interest Y_μ defined as the post-invasion abundance of a designated "invader" strain. Shown are the Pareto fronts in the error-complexity plane, for three tiers of community richness. The Pareto fronts show no signature of either emergent coarse-grainability or emergent predictability *sensu* Fig. 1D. **C:** Same data as in B, replotted with the y axis showing the absolute, rather than fractional, variance of the observable Y. In these axes, the Pareto fronts show a clear shift downwards. This corresponds to the self-averaging behavior (as community richness increases, the distribution of the observable Y becomes more narrow). We conclude that, unlike the simple self-averaging behavior, both forms of emergent simplicity defined here are not a typical feature of standard ecological models.

We make two observations. First, we see that in many cases, models with $I(\Psi) \gtrsim 3$ bits can capture $\gtrsim 75\%$ of variation in these simulated data. (Note that we simulated 1000 samples, and as a result there is no overfitting signature: the Pareto fronts show no uptick at high complexity). Even in the absence of simulated experimental noise, this level of performance was not guaranteed. Thus, Fig. 2 confirms that the model class of linear regressions, while simple, is sufficiently expressive to describe datasets structured by ecological interactions, at least those captured by these standard modeling frameworks.

Second, we see that in both cases, the Pareto fronts shift up and right, not down (Fig. 1E) or left (Fig. 1D). This analysis confirms our intuition that the emergent simplicity as defined here is not a behavior one would generically expect. Contrast this to Fig. 2C, where the same CRM simulation data are replotted using the absolute prediction error rather than relative. Here, the Pareto fronts exhibit a clear shift downwards. This is the phenomenon of 'emergent reproducibility' or functional convergence, where the distribution of Y gets narrower as richness increases. The contrast with Fig. 2A, B highlights that *both* forms of emergent simplicity shown in Fig. 1D, E are distinct from this more familiar scenario, widely reported both empirically [1–8] and in models [14, 24]. In Fig. 1D, E and Fig. 2A, B, the prediction error of Y is always normalized to the overall width of the distribution of Y across examples of the *same* richness tier. Our goal is to quantify forms of emergent simplicity going beyond emergent reproducibility, and our normalization explicitly removes this simpler effect.

To summarize, Fig. 2 demonstrates that standard ecological models do not generically exhibit the nuanced forms of emergent simplicity defined here. Thus, unlike the simple self-averaging, observing either of these phenomena in empirical data would be surprising.

B. Empirical data exhibit emergent predictability

To test the emergent simplicity hypothesis in data, we need a dataset satisfying several conditions. First, it must follow the combinatorial assembly protocol of Fig. 1A, with communities assembled from a fixed pool

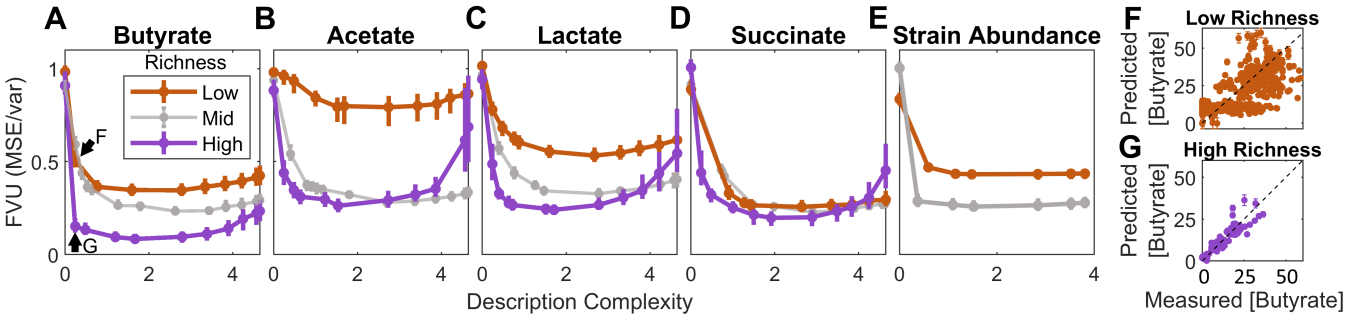

FIG. 3. **Emergent simplicity is observed in both available empirical datasets (Refs. [26] and [27]). A-E:** The Pareto fronts in error-complexity plane, computed for five properties of interest. Lines show the fraction of variance left unexplained by the best model within a given tier of description complexity, when dataset is restricted to communities of low richness (gray), or high richness (black). At high richness, the Pareto front shifts down and to the left, indicating that simpler descriptions become more predictive (compare to Fig. 1C). Each point and error bar shows the median quartiles across 100 random 50-50 splits of the data into training and testing sets. **G,H:** Illustrating the result by plotting predicted versus measured butyrate concentration (the points on the Pareto front indicated by the arrowheads in A). At high richness, a simple compositional feature explains 85% of variance in butyrate production (panel G). In contrast, for low-richness communities, even the most predictive single feature explains only 49% of variance (panel F).

of strains. Surveying the literature, we identified 9 candidate datasets (across both microbial and non-microbial contexts; see Supp Table S1). However, for our purposes it is essential that the assayed communities span a range of richness, so that at least two tiers of community diversity can be defined and compared. Additionally, if N is the number of samples and S the number of strains in the pool, N/S should be sufficiently large (for each richness tier) to allow training the model from our chosen model class. Our choice of linear regressions as model class makes this last condition the most permissive, and identified two datasets whose structure allows our question to be asked. These are the data of Clark *et al.* [26], who assayed communities of 1-25 gut-derived taxa and measured the production of butyrate, acetate, lactate and succinate; and the work of Kehe *et al.* [27], who used a droplet-based platform to screen communities of 1-19 isolates, measuring the abundance of a model plant symbiont *H. Frisingense* as the property of interest. Together, these two datasets provide measurements of five properties we can examine for evidence of emergent simplicity.

As above, for each dataset, we resolve three tiers of community richness (with the high-richness tier, > 20 strains, only available for Clark *et al.* data). Since Clark *et al.* did not directly measure absolute strain abundances in the community (only its total OD and the 16S makeup), we adopted an approach where the relative abundances estimated from the 16S data were first scaled by community OD and then standardized to zero mean and unit variance before training any regression models. For consistency, the same standardization procedure was also applied to Kehe et al data; though omitting this step does not qualitatively change the outcome in either case; see SI fig. S3. We then applied the same heuristic algorithm as above to infer the respective Pareto front. The results are shown in Fig. 3A-E.

Remarkably, we find that four out of five properties exhibit emergent predictability (but not emergent coarse-grainability): simple models become increasingly predictive in communities of higher richness. As a control, this behavior is not observed for measurements of community optical density (OD), or if the values of Y or species abundances are randomly permuted (see SI), indicating that the behavior observed in Fig. 3 is not an artifact of our analysis.

The additional controls we performed include subsampling the data, perturbing the groupings forming the putative Pareto front, and considering alternative metrics of complexity $I(\Psi)$; see SI. For all four properties exhibiting emergent predictability in Fig. 3, the result is robust against all these tests. We conclude that emergent predictability is robustly supported by data.

Furthermore, while we formulated our analysis in abstract terms such as the shape of a Pareto front, we can also confirm the result in more familiar terms by plotting the measured property against the prediction of the simplest models identified by our framework. Two examples are shown in Fig. 3F, G, which correspond to the points on the Pareto fronts labeled with arrowheads in Fig. 3A. At high richness, a simple two-parameter regression explains 85% of variance in butyrate production (Fig. 3G). In contrast, for low-richness communities, no model of comparable complexity can explain more than 49% of variance; the best-performing example is showin in Fig. 3F.

C. Emergent simplicity in functional space

So far we focused on asking how composition predicts functional properties. However, since Clark *et al.* mea-

sured four functional properties simultaneously, we can leverage this fact to also investigate whether these properties are predictive of each other.

In this dataset, the four functional measurements (the concentrations of butyrate, acetate, succinate and lactate produced by the community) define, for each community, a point in what we will call the "functional space". Fig. 4A shows that at high community richness, these points are found along a lower-dimensional subspace. Indeed, for high-richness communities, 73% of variance is explained by PC1, compared to only 41% for low-richness communities (Fig. 4B). This dominant axis of variation is correlated with community pH (Fig. 4C), consistent with Ref. [26] identifying pH-mediated interactions as a key factor structuring these communities.

The emergent low dimensionality in the functional space (Fig. 4A) is distinct and independent from the emergent predictive power of simple compositional features (Fig. 3A-E). Thus, Fig. 4A provides a separate line of evidence indicating that, at least in these synthetic ecosystems, multiple aspects of community structure indeed become simpler and more predictable at high richness.

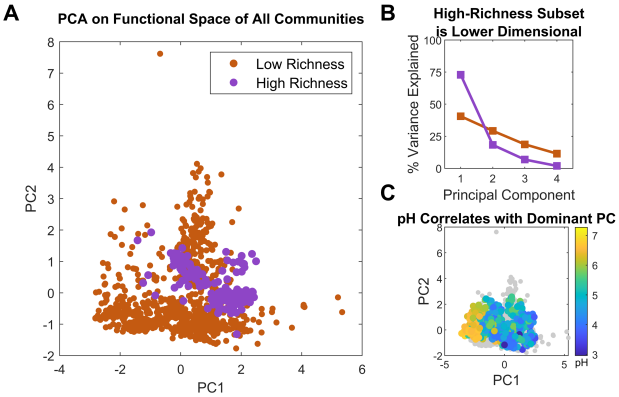

FIG. 4. **Emergent simplicity in functional space. A:** Principle component analysis of the four-dimensional functional measurements of Clark *et al.* (butyrate, acetate, lactate, succinate) indicates that high-richness samples form a lower-dimensional subspace. Principal axes computed for the dataset as a whole; mid-richness communities omitted from plotting for clarity; see SI for full dataset. **B:** PCA spectra computed separately for the two tiers of community richness. For high-richness samples, the dominant component explains 73% of variance, contrasted with 41% for low-richness. **C:** Coloring the PCA plot by measured pH reveals the dominant axis of functional variation is correlated with community pH. Axes same as in A, but including samples of all richness. In gray, communities for which pH was not measured.

D. Mechanistic origin of emergent simplicity is unknown

Why is this occurring? As emphasized above, our normalization procedure for the Pareto fronts ensures that the effect of Fig. 3 cannot be explained by simple self-averaging of the observable of interest.

For specifically butyrate, the results of Clark *et al.* [26] provide some insight into the mechanism. Of the 25 strains in this study, five are capable of producing butyrate. However, in high-richness communities, butyrate production is dominated by one of them, *Anaerostipes caccae* (AC). And indeed, the first point on the high-richness Pareto front in Fig. 3A is the grouping {AC, all other strains}, and provides an excellent prediction for the amount of butyrate produced (Fig. 3G). In contrast, in low-richness mixtures, all five strains contribute to the production, and no single compositional feature is similarly predictive.

Precisely why a single producer becomes dominant is unclear. Clark *et al.* hypothesized that this transition is associated with AC shifting at high community richness to an alternative metabolic pathway, utilizing lactate rather than sugars, a physiological switch mediated by sensing pH and/or the depletion of key resources. Identifying candidate mechanistic explanations for the other properties, and validating these hypotheses, would require new experiments. It is conceivable that the observations of Fig. 3A-D are all a consequence of specific physiological features of individual community members. However, given that we see a common outcome across two independent datasets, it is intriguing to hypothesize that emergent predictability as quantified here might be more general than a mechanistic idiosyncrasy of specific taxa.

III. DISCUSSION

In this work, we developed an information-theoretic framework to nuance and quantify the notion of "emergent simplicity" in ecosystems of increasing community richness (treated as the control parameter). Our framework enabled us to define and distinguish emergent predictability and emergent coarse-grainability, and to contrast both of these phenomena with the emergent reproducibility considered in much previous work. We then applied our framework to data from two separate experiments on synthetic microbial communities, and demonstrated that 4 out of 5 measured properties exhibit robust evidence of emergent predictability, whereby simple compositional descriptions become more predictive at higher richness.

We should note that in general, community richness (number of species) is a poor metric of community diversity, as it is overly sensitive to detection thresholds, and fails to account for the relatedness of community members [28]. This was not an issue for our work, as the synthetic communities considered here were assembled from a small pool of well-delimited taxa. For natural communities, this simple metric of diversity would need to be replaced by one of the better-suited metrics developed in

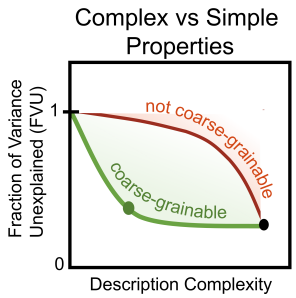

FIG. 5. **Shifting the focus from simple and complex ecosystems to simple and complex *properties* of ecosystems.** As demonstrated in this work, even in a complex (high-richness) community, some properties may be adequately predicted with a low-complexity description; we call such properties coarse-grainable (green Pareto front).

the ecological literature [29–31].

Our framework suggests a shift of viewpoint: instead of talking about simple and complex ecosystems (as quantified, e.g., by community richness), we should be distinguishing simple and complex *properties* of ecosystems. In the language we presented, this distinction is encoded in the shape of the Pareto front, as illustrated in Fig. 5. For properties of interest behaving as in Fig. 3, a comparatively coarse description may provide a "good-enough" prediction power at a fraction of the entropy of the fully microscopic description. We can call such properties coarse-grainable (Fig. 5, green). Conversely, for other properties one might find that useful prediction requires substantial microscopic detail (non-coarse-grainable; Fig. 5, red). Remarkably, in the datasets considered here, all measured properties proved coarse-grainable (compare Fig. 5 and Fig. 3A-E). Further, our key result can be summarized as saying that even as community complexity is increased, the prediction complexity of a given *property* need not do the same, and can even become simpler, as we quantitatively demonstrate.

An empirical confirmation of this behavior is very exciting because learning to harness emergent predictability could enable new methodologies for predicting and controlling complex ecosystems: methodologies for which community richness is an asset rather than a nuisance. Understanding the mechanistic origin of this effect, and developing a theoretical understanding for the conditions of its emergence, are important directions for future work.

All simulations performed in MATLAB (Mathworks, Inc.). The associated code, data and scripts to reproduce all figures in this work are available at `https://github.com/mtikhonov/emergentPredictability`.

ACKNOWLEDGMENTS

We thank S. Allesina, A. Goyal. K. Gowda, C. Holmes, S. Kaplan, S. Kuehn, R. McGee,A. Murugan, O. Venturelli and K. Wood for helpful discussions. This work was supported in part by the National Science Foundation grant PHY-2310746, and the Michigan Pioneer Fellows Program award to JM.

[1] H. Heatwole and R. Levins, Trophic structure stability and faunal change during recolonization, Ecology **53**, 531 (1972).
[2] J. E. Cohen, Ratio of prey to predators in community food webs, Nature **270**, 165 (1977).
[3] F. Briand and J. E. Cohen, Community food webs have scale-invariant structure, Nature **307**, 264 (1984).
[4] J. E. Goldford, N. Lu, D. Bajić, S. Estrela, M. Tikhonov, A. Sanchez-Gorostiaga, D. Segrè, P. Mehta, and A. Sanchez, Emergent simplicity in microbial community assembly, Science **361**, 469 (2018), publisher: American Association for the Advancement of Science.
[5] S. Estrela, J. C. Vila, N. Lu, D. Bajic, M. Rebolleda-Gomex, C.-Y. Chang, J. E. Goldford, A. Sanchez-Gorostiaga, and A. Sanchez, Functional attractors in microbial community assembly, Cell Systems **13**, 29 (2022), publisher: Elsevier.
[6] J. McGrady-Steed, P. M. Harris, and P. J. Morin, Biodiversity regulates ecosystem predictability, Nature **390**, 162 (1997).
[7] S. Naeem and S. Li, Biodiversity enhances ecosystem reliability, Nature **390**, 507 (1997).
[8] F. Carrara, A. Giometto, M. Seymour, A. Rinaldo, and F. Altermatt, Experimental evidence for strong stabilizing forces at high functional diversity of aquatic microbial communities, Ecology **96**, 1340 (2015).
[9] M. Advani, G. Bunin, and P. Mehta, Statistical physics of community ecology: a cavity solution to MacArthur's consumer resource model, Journal of Statistical Mechanics: Theory and Experiment **2018**, 033406 (2018), publisher: IOP Publishing.
[10] R. Marsland, W. Cui, and P. Mehta, A minimal model for microbial biodiversity can reproduce experimentally observed ecological patterns, Scientific reports **10**, 1 (2020), publisher: Nature Publishing Group.
[11] S. Louca, M. F. Polz, F. Mazel, M. B. Albright, J. A. Huber, M. I. O'Connor, M. Ackermann, A. S. Hahn, D. S. Srivastava, S. A. Crowe, and others, Function and functional redundancy in microbial systems, Nature ecology & evolution **2**, 936 (2018), publisher: Nature Publishing Group UK London.
[12] S. Louca, L. W. Parfrey, and M. Doebeli, Decoupling function and taxonomy in the global ocean microbiome, Science **353**, 1272 (2016), publisher: American Association for the Advancement of Science.
[13] S. Louca, S. M. Jacques, A. P. Pires, J. S. Leal, D. S. Srivastava, L. W. Parfrey, V. F. Farjalla, and M. Doebeli, High taxonomic variability despite stable functional structure across microbial communities, Nature ecology & evolution **1**, 1 (2016), publisher: Nature Publishing

Group.
[14] L. Fant, I. Macocco, and J. Grilli, Eco-evolutionary dynamics lead to functionally robust and redundant communities, bioRxiv (2021), publisher: Cold Spring Harbor Laboratory.
[15] J. P. O'Dwyer, Beyond an ecological ideal gas law, Nature Ecology & Evolution **4**, 14 (2020).
[16] J. P. O'Dwyer, S. W. Kembel, and T. J. Sharpton, Backbones of evolutionary history test biodiversity theory for microbes, Proceedings of the National Academy of Sciences **112**, 8356 (2015), publisher: National Acad Sciences.
[17] J. Bergelson, M. Kreitman, D. A. Petrov, A. Sanchez, and M. Tikhonov, Functional biology in its natural context: A search for emergent simplicity, Elife **10**, e67646 (2021), publisher: eLife Sciences Publications Limited.
[18] M. R. Silverstein, J. M. Bhatnagar, and D. Segrè, Metabolic complexity drives divergence in microbial communities, bioRxiv , 2023 (2023).
[19] A. Jeglot, J. Audet, S. R. Sorensen, K. Schnorr, F. Plauborg, and L. Elsgaard, Microbiome structure and function in woodchip bioreactors for nitrate removal in agricultural drainage water, Frontiers in microbiology **12**, 2284 (2021), publisher: Frontiers.
[20] N. E. Korres and A. S. Nizami, Variation in anaerobic digestion: need for process monitoring, in *Bioenergy production by anaerobic digestion* (Routledge, 2013) pp. 194–230.
[21] L. Naik, Z. Gebreegziabher, V. Tumwesige, B. B. Balana, J. Mwirigi, and G. Austin, Factors determining the stability and productivity of small scale anaerobic digesters, Biomass and bioenergy **70**, 51 (2014).
[22] S. Bertacchi, M. Ruusunen, A. Sorsa, A. Sirviö, and P. Branduardi, Mathematical Analysis and Update of ADM1 Model for Biomethane Production by Anaerobic Digestion, Fermentation **7**, 237 (2021), publisher: Multidisciplinary Digital Publishing Institute.
[23] X. Shan, A. Goyal, R. Gregor, and O. X. Cordero, Annotation-free discovery of functional groups in microbial communities, Nature Ecology & Evolution **7**, 716 (2023).
[24] J. Moran and M. Tikhonov, Defining coarse-grainability in a model of structured microbial ecosystems, Mendeley Data 10.17632/gpfw4y46c8. (2022).
[25] A. Skwara, K. Gowda, M. Yousef, J. Diaz-Colunga, A. S. Raman, A. Sanchez, M. Tikhonov, and S. Kuehn, Statistically learning the functional landscape of microbial communities, Nature Ecology & Evolution , 1 (2023).
[26] R. L. Clark, B. M. Connors, D. M. Stevenson, S. E. Hromada, J. J. Hamilton, D. Amador-Noguez, and O. S. Venturelli, enDesign of synthetic human gut microbiome assembly and butyrate production, Nature Communications **12**, 3254 (2021), number: 1 Publisher: Nature Publishing Group.
[27] J. Kehe, A. Kulesa, A. Ortiz, C. M. Ackerman, S. G. Thakku, D. Sellers, S. Kuehn, J. Gore, J. Friedman, and P. C. Blainey, Massively parallel screening of synthetic microbial communities, Proceedings of the National Academy of Sciences **116**, 12804 (2019), publisher: National Acad Sciences.
[28] R. K. Peet, The measurement of species diversity, Annual review of ecology and systematics **5**, 285 (1974).
[29] K. J. Iknayan, M. W. Tingley, B. J. Furnas, and S. R. Beissinger, Detecting diversity: emerging methods to estimate species diversity, Trends in ecology & evolution **29**, 97 (2014).
[30] R. S. Mendes, L. R. Evangelista, S. M. Thomaz, A. A. Agostinho, and L. C. Gomes, A unified index to measure ecological diversity and species rarity, Ecography **31**, 450 (2008).
[31] V. Crupi, Measures of biological diversity: Overview and unified framework, From Assessing to Conserving Biodiversity: Conceptual and Practical Challenges , 123 (2019).
[32] M. Barbier, J.-F. Arnoldi, G. Bunin, and M. Loreau, Generic assembly patterns in complex ecological communities, Proceedings of the National Academy of Sciences **115**, 2156 (2018), publisher: National Acad Sciences.
[33] G. Bunin, Interaction patterns and diversity in assembled ecological communities, arXiv preprint arXiv:1607.04734 (2016).
[34] M. Tikhonov and R. Monasson, Collective phase in resource competition in a highly diverse ecosystem, Physical review letters **118**, 048103 (2017), publisher: APS.
[35] J. J. Faith, P. P. Ahern, V. K. Ridaura, J. Cheng, and J. I. Gordon, Identifying gut microbe–host phenotype relationships using combinatorial communities in gnotobiotic mice, Science translational medicine **6**, 220ra11 (2014).
[36] J. M. Jungers, Y. Yang, C. W. Fernandez, F. Isbell, C. Lehman, D. Wyse, and C. Sheaffer, Diversifying bioenergy crops increases yield and yield stability by reducing weed abundance, Science Advances **7**, eabg8531 (2021).
[37] A. Hector, B. Schmid, C. Beierkuhnlein, M. Caldeira, M. Diemer, P. G. Dimitrakopoulos, J. Finn, H. Freitas, P. Giller, J. Good, *et al.*, Plant diversity and productivity experiments in european grasslands, science **286**, 1123 (1999).
[38] E. Spehn, A. Hector, J. Joshi, M. Scherer-Lorenzen, B. Schmid, E. Bazeley-White, C. Beierkuhnlein, M. Caldeira, M. Diemer, P. Dimitrakopoulos, *et al.*, Ecosystem effects of biodiversity manipulations in european grasslands, Ecological monographs **75**, 37 (2005).
[39] A. Weigelt, E. Marquard, V. M. Temperton, C. Roscher, C. Scherber, P. N. Mwangi, S. Von Felten, N. Buchmann, B. Schmid, E.-D. Schulze, *et al.*, The jena experiment: six years of data from a grassland biodiversity experiment, Ecology **91**, 930 (2010).
[40] C. Roscher, V. M. Temperton, M. Scherer-Lorenzen, M. Schmitz, J. Schumacher, B. Schmid, N. Buchmann, W. W. Weisser, and E.-D. Schulze, Overyielding in experimental grassland communities–irrespective of species pool or spatial scale, Ecology Letters **8**, 419 (2005).
[41] P. B. Reich, J. Knops, D. Tilman, J. Craine, D. Ellsworth, M. Tjoelker, T. Lee, D. Wedin, S. Naeem, D. Bahauddin, *et al.*, Plant diversity enhances ecosystem responses to elevated co2 and nitrogen deposition, Nature **410**, 809 (2001).
[42] P. B. Reich, D. Tilman, S. Naeem, D. S. Ellsworth, J. Knops, J. Craine, D. Wedin, and J. Trost, Species and functional group diversity independently influence biomass accumulation and its response to co2 and n, Proceedings of the National Academy of Sciences **101**, 10101 (2004).
[43] P. B. Reich, Elevated co2 reduces losses of plant diversity caused by nitrogen deposition, Science **326**, 1399 (2009).
[44] J. Van Ruijven and F. Berendse, Positive effects of plant species diversity on productivity in the absence of legumes, Ecology Letters **6**, 170 (2003).

[45] F. I. Isbell and B. J. Wilsey, Increasing native, but not exotic, biodiversity increases aboveground productivity in ungrazed and intensely grazed grasslands, Oecologia **165**, 771 (2011).

[46] H. Krehenwinkel, M. Wolf, J. Y. Lim, A. J. Rominger, W. B. Simison, and R. G. Gillespie, Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding, Scientific reports **7**, 17668 (2017).

[47] M. R. McLaren, A. D. Willis, and B. J. Callahan, Consistent and correctable bias in metagenomic sequencing experiments, Elife **8**, e46923 (2019).

SUPPLEMENTARY INFORMATION

1. Quantifying description complexity, $I(\Psi)$

The most obvious measure of model complexity is the number of model parameters. In the context of this work, this would mean setting $I(\Psi)$ (description complexity) to the number of classes (the cardinality of the partitioning). However, in the main text, we opted for a different metric, defining $I(\Psi)$ as the *entropy* of the partitioning, which weighs classes differently based on the number of strains included. In this section, we comment on this choice.

It is important to emphasize that either choice preserves the phenomenon described in this work (compare main text Fig. 3 with Fig. S1; in either case, the Pareto fronts exhibit the shift indicative of emergent predictability). To explain our choice, we note that the entropy of Ψ is the mutual information between the label of the coarsened class and the identity of the microscopic strain. Consider a scenario where two choices of two-group coarsenings are equally predictive, but one splits strains 50-50, while the other coarsening separates one strain from the rest of the taxa put together. If the latter is equally predictive, our definition identifies it as simpler, and thus preferable (same predictive power, lower complexity).

This property is arguably desirable. If knowing the abundance of just one strain is sufficient to predict function, this scenario makes it easiest to gain insight into the mechanism by which community function is assured. Thus, we would prefer our framework to favor such coarsenings among others with comparable predictive power.

This definition also has some shortcomings. Most notably, any definition relying on counting strains suffers from the fact that such counting is not always straightforward, especially when some strains in the pool are close relatives. For our purposes in this work, this has not been a problem because we applied our framework to data from defined communities constructed from a well-defined set of distinct taxa. (See also the discussion section in the main text, regarding the validity of stratifying communities by "richness".) In the general case, however, the hierarchical nature of relationships between strains makes the choice of ecologically-informed metrics assessing the complexity of an ecosystem description non-trivial, and a key question for future work to address.

Thus, the relative merits of our $I(\Psi)$ compared to the simple cardinality of the partitioning remain to be investigated. However, we believe this question merits further attention, and so used a non-standard $I(\Psi)$ in the main text to motivate this conversation.

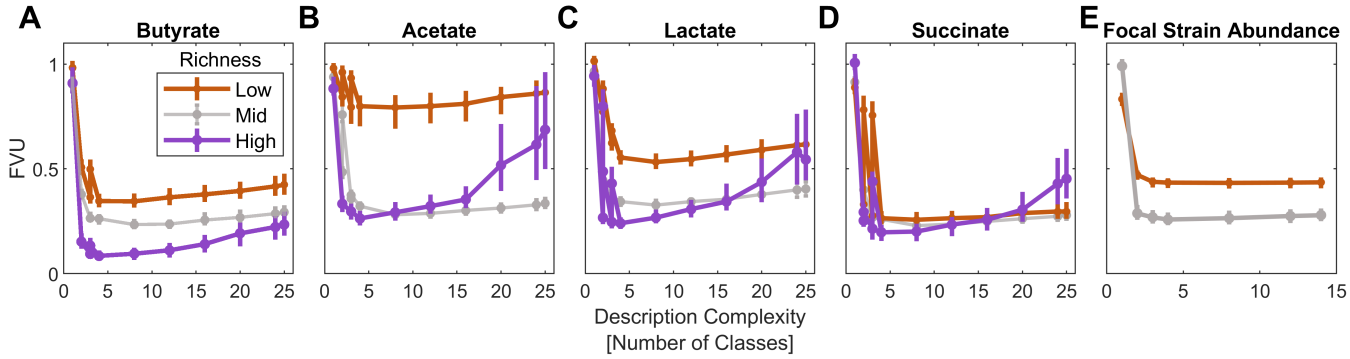

FIG. S1. Replotting result shown in Fig. 3 with description complexity axis $I(\Psi)$ measured in terms of number of classes of a given coarsened description Ψ.

2. Simulations of standard ecological models

We investigate the concepts of emergent simplicity in two standard ecological models: a Generalized Lotka-Volterra (GLV) model and a Consumer Resource (CR) model. The GLV model we consider takes the following form:

$$\dot{n}_i = \frac{r_i}{K_i} n_i \left(K_i - n_i - \sum_{j, j \neq i} a_{ij} n_j \right),$$

which describes the abundance dynamics of each strain i in terms of its growth rate r_i, carrying capacity K_i and interactions with other strains encoded in a_{ij}. For simplicity, we set the growth rates for all strains to be the same

$r_i = 1$ since the key parameters determining equilibrium properties are the K_i and interactions a_{ij} [32]. We then adopt the approach common in large-N theoretical ecology, drawing strain interactions a_{ij} and carrying capacities K_i randomly from specified distributions. Following refs. [32, 33], we use a normal distribution for both K_i and a_{ij}. For the K_i, we set the mean to 1 and $\mathrm{std}(K_i) = 0.3$. For the interactions a_{ij}, the mean and spread of the distribution scale with the number of strains as μ/N and $\sigma/\sqrt{N}$, respectively. We set $\mu = 1$ and $\sigma = 0.8$ (modified from the "simplest" choice $\mu = 1$ and $\sigma = 1$ to facilitate the construction of a coexisting species pool). To obtain a strain library from which to subsample as illustrated in Fig. 1A, we first simulate the dynamics of a larger pool to find a set of coexisting strains (abundance $n_i > 10^{-6}$). With our parameter choices, we found that a random set of 45 strains resulted in a community of 26 coexisting strains, similar to the high richness communities of the empirical datasets we consider in this work.

The CR model we choose to study describes the abundance dynamics of strains competing for a set of resources as follows:

$$\dot{n}_i = n_i \left(\sum_\alpha \sigma_{i\alpha} h_\alpha - \chi_i \right)$$
$$h_\alpha = h_\alpha(T_\alpha) \equiv \frac{b_\alpha}{1 + T_\alpha/K_\alpha}$$
$$T_\alpha \equiv \sum_i n_i \sigma_{i\alpha}.$$

These are the same dynamics postulated in equation 1 of ref. [24] with slightly adjusted notation. Briefly, each strain i is described as a binary phenotype $\sigma_{\alpha i} \in \{0, 1\}$ indicating which resources α each strain can or cannot consume. The growth benefit of each resource h_α is depleted as the total demand T_α for resources increases with increasing abundance of competing strains. The strength of this depletion is set by parameters b_α and K_α. For simplicity, we take $b_\alpha = 1$ for all resources α. Finally, each strain is also described by a growth cost, which we choose to model in the same way as ref. [34], $\chi_i = c + \sum_\alpha \lambda \sigma_{i\alpha} + \epsilon_i$. As in ref. [24], we set the baseline cost $c = 0.1$ and the cost per resource consumed $\lambda = 0.5$. Using the same approach as in studying the GLV model, we draw parameters K_α, ϵ_i and $\sigma_{i\alpha}$ randomly. The "carrying capacities" K_α are Gaussian distributed with mean 10^{10} and standard deviation 10^9. (Note that the scale of abundance is of course arbitrary, and could have been set to 1. The value of 10^{10}, meant to evoke the large typical size of microbial populations, was chosen to visually underscore the difference between the main text Fig. 2C (plotting the *absolute* variance of an invader abundance) and the other panels, such as Fig. 2A, B showing the **fraction** of variance.) Random contributions to each strain's cost ϵ_i are also Gaussian distributed with mean 0 and standard deviation of 0.01. Each strain phenotype $\vec{\sigma}_i$ is a random L-dimensional binary vector where each element represents the ability to consume a resource, and has probability $p = 0.5$ to be 1. Just as we did in the GLV model above, we assemble a community from a larger pool of M phenotypes, choosing $M = 50$ and $L = 40$ such that the number of coexisting strains is similar to the empirical datasets ($\simeq$25).

In both the GLV and CR models, we use the same protocol for our computational experiments, which follows the scheme outlined in Fig. 1A and mimics the laboratory experiments we consider in this work. Using the same distributions used to form the complete community, we randomly draw a focal strain to invade subsets of this community. Specifically, we assemble 1000 communities for each richness tier: low richness with 1 to 5 strains, mid richness with 11 to 15 strains, and high richness with 21 to 25 strains. After a specified amount of simulation time, we note the abundances of all community members, recording them as the "pre-invasion abundances"; these will be used as predictors. We then computationally "invade" the communities with the focal strain, and measure the post-invasion abundance of the focal strain after the same amount of simulation time has elapsed; this abundance is recorded as the response variable we seek to predict. Fig. S2 shows the resulting Pareto fronts inferred by our algorithm (see below for details) for 10 random invading strains. These are the Pareto fronts we average to make Fig. 2 in the main text.

3. Description of empirical datasets

Scanning the literature for empirical studies that assemble communities from a fixed library of strains, we identified 9 candidate datasets for further consideration. These are listed in Table S1 below, cataloguing which datasets satisfy the criteria required for analysis within our framework.

In summary, we determined that the datasets reported in [26] and [27] satisfy all requirements for applying our framework; namely, (1) enough community richness levels are sampled to clearly define at least two distinct richness tiers for comparison, and (2) enough communities are sampled within each richness tier to train and test linear regression models.

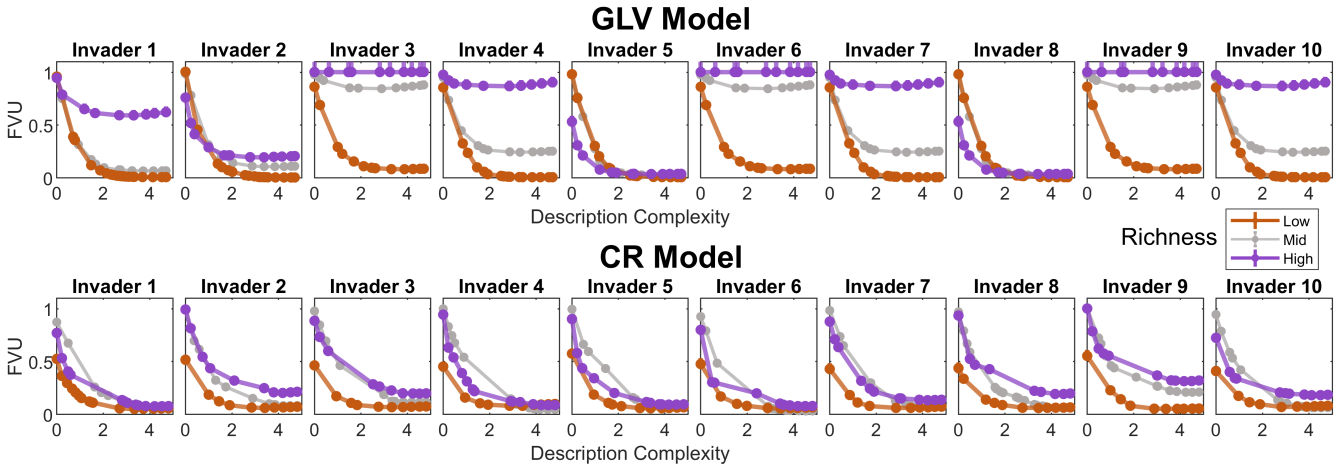

FIG. S2. Plotting the Pareto fronts for predicting post-invasion abundance of randomly drawn strains in a Generalized Lotka-Volterra (GLV) model (top row) and Consumer-Resource (CR) model (bottom row). The panels shown in Fig. 2 are averages over the 10 strains in each ecological model. See text for simulation protocol details.

Dataset	Ecological Context	$\geq$ 2 Richness Tiers (yes/no)	Number of Samples / Number of Strains (by Tier)	Satisfies Inclusion Criteria?
[26]	microbial	yes	870/25 (low), 477/25 (mid), 180/25 (high)	yes
[27]	microbial	yes	19433/14 (low), 346/14 (mid)	yes
[35]	microbial	yes	21/17 (low), 30/17 (mid)	not enough data
[36]	grasslands	yes	350/31 (low), 200/31 (mid), 50/31 (high)	not enough data; not all strains were sampled in low and mid richness samples
[37, 38]	grasslands	yes	345/32 (low), 22/32 (mid)	not enough mid richness samples
[39, 40]	grasslands	yes	200/60 (low), 40/60 (mid), 10/60 (high)	not enough data
[41–43]	grasslands	no	N/A	too few richness levels sampled and not enough data
[44]	grasslands	no	N/A	plots either had all or one species present
[45]	grasslands	no	N/A	low richness only

TABLE S1. Catalogue of candidate published datasets that sample from a fixed library of strains. The last three columns indicate which datasets satisfy the required inclusion criteria for our framework to be applicable.

Clark *et al.* [26] investigated the production of fermentation end-products (butyrate, acetate, lactate, succinate) in synthetic gut communities consisting of 1 to 25 strain isolates. Final strain abundances were determined by first measuring sequenced read counts to compute relative abundances, and then calculating absolute abundances by multiplying by total OD_{600} of each community. These are the data use as inputs to our regression analysis for predicting community function (see below).

The experimental platform of Kehe *et al.* [27] was a microfluidic device containing various geometrical arrangments of microwells for high-throughput, parallel screening of community composition to function mapping. Each microwell array enables the grouping and merging of nanoliter droplets that contain different strain isolates from soil. Unlike Clark *et al.* [26], abundances of input strains were not measured; only the abundance of the focal strain *H. Frisingense* was tracked in arbitrary fluorescence units. Because of this, instead of using abundance as inputs to our framework presented in Fig. 1B, we use the presence ($n_i = 1$) or absence ($n_i = 0$) of strain droplets reported by Kehe etal as our inputs for predicting the final abundance of the focal strain. When we apply a given coarsening Ψ, we sum up the presence of strains that are grouped together for each sample community.

4. Normalization of predictor and response variables

Experimentally, if we think about e.g. soil communities, it would be reasonable to define properties of interest Y as measured in "per gram biomass." In this case the overall biomass is normalized away, and this is the scenario we had in mind when calling Ψ_z the "zero-information" description in Fig. 1C. This is equivalent to trying to predict the property of interest from relative abundances of taxa.

In lab experiments, where an inoculum is provided a fixed amount of resources, normalizing to biomass may not be necessary, since properties of interest can be defined in the absolute sense. (In fact, the resulting biomass could itself be a property of interest.) In this case, Ψ_z is not actually "zero-information," since knowing the overall biomass can be predictive (in which case the fraction of unexplained variance will be smaller than 1).

In our analysis, rather than using the actual abundances of species, we standardized all input variables n to zero mean and unit variance. All coarsening schemes were then applied to these standardized variables. It is important to note that summing standardized abundances is not equivalent to standardizing summed abundances. However, the experimentally available numbers (16S reads) are only proxies for the absolute abundances anyway. The proportionality factor includes the number of 16S copies on the genome, which differs between species, and an PCR amplification bias, which is unknown and often substantial [46, 47]. Furthermore, even if absolute abundances were known, it is not clear if the summation rule implementing the intuition of $\tilde{n}$ as putative "functional classes" should be in units of biomass, in units of metabolic activity, or weighted in some other way.

Thus, although this intuition of $\tilde{n}$ as the combined abundance of a putative functional class makes it easiest to explain our approach, any practical implementation on the available experimental data would implicitly include an unavoidable species-specific factor in front of every abundance value. Given this, we chose to fix these factors by simply standardizing input variables to unit variance, which is generally good practice with regression models, and came with the added benefit of improving performance. That said, we did also confirm that using raw OD-scaled 16S counts without standardization retains our conclusion (Fig. S3).

The standardization step would certainly complicate attempts at mechanistic interpretation of the low-complexity descriptions discovered by our framework (lying along the Pareto front). In general, this strategy—discover a coarse-graining that is predictive, and then gain mechanistic insight by understanding why it works—is one of the promising applications of a framework like ours, or feature selection frameworks more generally. However, in this work, we focus on demonstrating the existence of diversity-aided emergent simplicity. Interpreting the variables is left for future work.

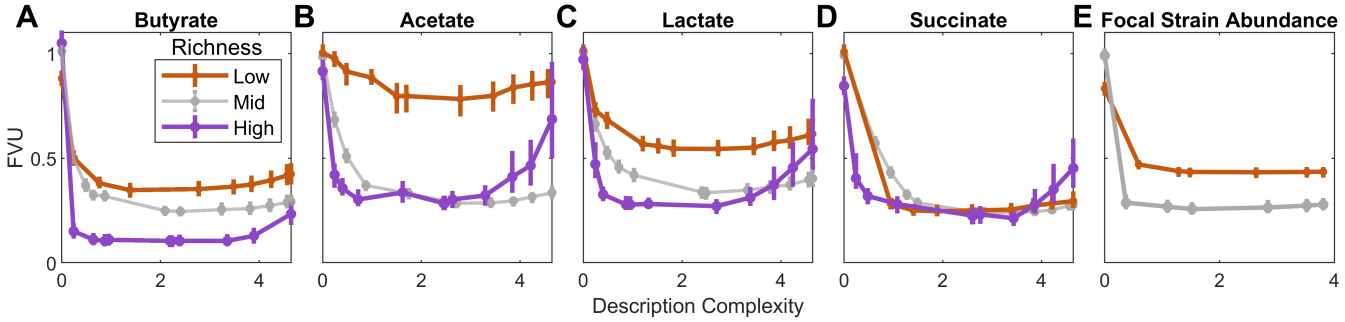

FIG. S3. Reanalyzing result of Fig. 3 with raw input variables (non-standardized) used in regression models. That is, abundance data is not standardized to 0 mean and unit variance.

5. Training and testing of linear regression models

As noted in the main text, we choose to focus on the simple model class of first-order linear regression for reasons stated in section I B. This is the model class we use to apply our framework to all datasets considered in this work – both simulated and empirical. Specifically, we write:

$$\hat{Y}_\mu = \beta_0^\Psi + \sum_i \beta_i^\Psi \tilde{n}_{i\mu}^\Psi,$$

where $\hat{Y}_\mu$ is the model's prediction of property Y measured in the μth community and β_i^Ψ are the regression coefficients of each coarse-grained variable (see main text section I B) to be determined by fitting to a training dataset.

Each regression model corresponding to a coarsened description Ψ is assessed on the same data constructed by randomly splitting datasets into subsets dedicated for either training or testing of regression models. After accounting for replicate communities assayed in each richness tier (i.e., same presence-absence combination $b_{i\mu}$ defined in section I A), 50% of communities are randomly designated for model training/fitting and the other 50% are reserved for testing/validating prediction error. Model fitting was performed using the standard ordinary least-squares method of MATLAB's 'fitlm' algorithm. The predictions of the fitted model where then compared to the measured values of the test set using standard mean-squared error (MSE). The Fraction of Variance Unexplained (FVU) was computed in the standard way of dividing MSE by the variance of Y across the test set samples. This was repeated for 100 random 50-50 splits of the data, taking the median, upper and lower quartiles across these splits as the values for the ordinate axis and for the error bars around it.

6. Deducing the Pareto front of efficient coarsened descriptions

The Pareto front of coarsenings Ψ consist of those with minimal prediction error for a given complexity $I(\Psi)$. Ideally, we would like to know this exactly. However, the brute-force approach of performing an exhaustive search through the space of coarse-grainings is computationally infeasible unless the library of strains is very small. This is because the number of possible ways to partition the strains becomes exponentially large: for S strains, the total number is the Bell number B_S, which is too large to enumerate fully. For example, the 25 taxa used in the Clark *et al.* already results in $B_{25} \approx 5 \times 10^{18}$. This necessitates developing search algorithms to efficiently explore this vast landscape.

With the tractability of a linear regression ansatz, a simple heuristic allows us to construct candidate coarsened descriptions from the microscopic model itself, as we describe below. While these groupings are not guaranteed to be correct or optimal, in practice, we find that they provide good starting points for executing a local search.

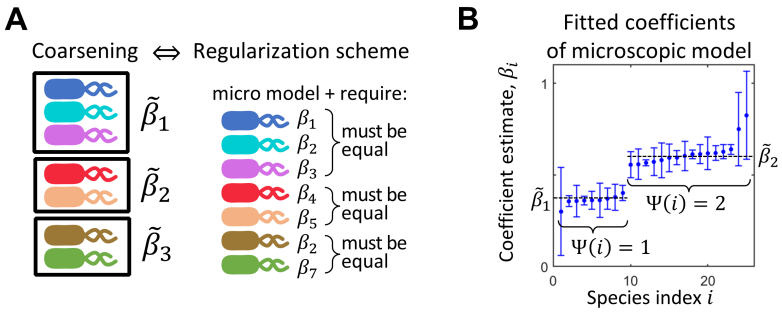

FIG. S4. **A:** In the context of a linear regression model class, the coarsening scheme we develop in this work maps onto a regularization scheme. **B:** This observation suggests a heuristic procedure for constructing candidate coarsenings from the knowledge of the microscopic model (regression with all species kept as separate predictors), by applying a clustering algorithm to the microscopic regression coefficients. In this illustration, there are two clear classes, $\Psi(i) = 1$ and $\Psi(i) = 2$, from which we can form a candidate two-variable coarsened model with coefficients $\tilde{\beta}_\alpha$. Error bars correspond to uncertainty in fit.

Specifically, we note that in the context of linear regression, the coarsening procedure Ψ we consider here (partitioning taxa into groups and summing the abundances of taxa assigned to the same group) can be mapped to a regularization scheme (Figure S4A). This suggests that efficient coarsenings could be constructed by grouping strains with similar regression coefficients (Figure S4B). For this, we implement a generalized K-means clustering algorithm where the cost function accounts for the uncertainty carried by each coefficient estimate: $\text{cost}(\Psi) = \sum_i \frac{\beta_i - \langle \beta_j \rangle_{\alpha_j = \alpha_i}}{\sigma_i}$, where angular brackets denote averaging over all taxa assigned to the same class by the coarsening Ψ. This generalized K-means procedure is used to generate candidate groupings for a range of values of K, and these groupings are used as seeds for local searches optimizing the Pareto front. The specific steps of the Pareto front search algorithm are as follows:

- To obtain accurate estimates initially, the microscopic regression models are fit on all the data before deducing the Pareto front coarsenings via this clustering procedure. This provides the initial guess of the Pareto front by the heuristic depicted in Fig. S4.

- Evaluate the prediction error of all possible single-strain relabelings of strains. That is, for each coarsening Ψ making up the current guess of the Pareto front, generate all neighboring coarsenings that successively re-assign each single strain to all other classes in Ψ.

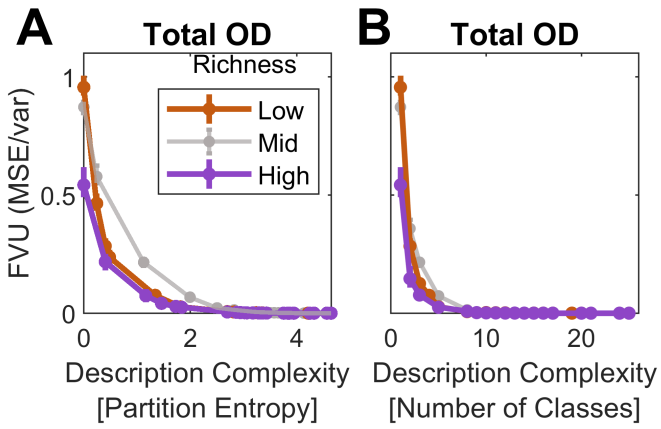

FIG. S5. Using the total OD measured in the Clark *et al.* experiments as a property of interest for prediction, we apply our framework to produce the Pareto fronts with description complexity $I(\Psi)$ measured as partition entropy (**A**) or number of resolved classes (**B**).

- Identify any relabeled coarsenings Ψ that result in a median FVU that fall below the convex hall spanned by the lower quartiles in FVU of the current Pareto front coarsenings and update the current guess of the Pareto front by adding these as the new efficient coarsenings.
- Repeat steps (2) and (3) until no single-strain relabelings of the Pareto front coarsened descriptions fall below the lower FVU quartiles. At this point, the algorithm has converged to at least a local optimum.

7. No emergent simplicity in predicting total OD as property of interest

In addition to measuring fermentation end products, Clark *et al.* measured the total optical density (OD) of their communities to compute strain abundances. Applying our framework to the OD measurements as a control, we find that this property does not exhibit any form of emergent simplicity, indicating that the emergent predictability we observe in Fig. 3 is not an artifact of the distributions of strain abundances (see Fig. S5).

8. Robustness tests

To verify that the observation of emergent predictability is robust, we perform two tests on the low-richness datasets from [26] and [27] to check that the drop in prediction error at higher richness is not due to how the data was collected or analyzed.

First, we perform a down-sampling test, shown in Fig. S6. To test robustness to sample size, we down-sampled the low-richness communities to match the number of unique communities sampled in the high-richness dataset. We performed 5 random subsamplings of the low-richness dataset and computed the Pareto fronts for each down-sample. Comparing these in Fig. S6 to the reference high-richness Pareto fronts (purple dashed line), we find that for 4 out of the 5 properties the effect we observe is not due to a smaller amount of data collected on high-richness communities.

Second, we perform a relabeling test. To check how well our heuristic deduction algorithm does at inferring the Pareto front, which should be at least a local optimum, we evaluate nearby coarse-grained partitions by reassigning individual strains from their original group to all other possible groups. Fig. S7 shows the best of all possible single-move relabelings relative to the low-richness and high-richness Pareto fronts presented in the main text. For all functional properties, performing such a relabeling perturbation either has no effect or decreases prediction power. This confirms that our Pareto front search algorithm does indeed converge.

9. Testing for artifacts

Checking that the emergent predictability observed in Figure 3 is not purely due to chance, we compare the result against Pareto fronts deduced after randomizing the datasets. For designing a stringent randomization test, one seeks

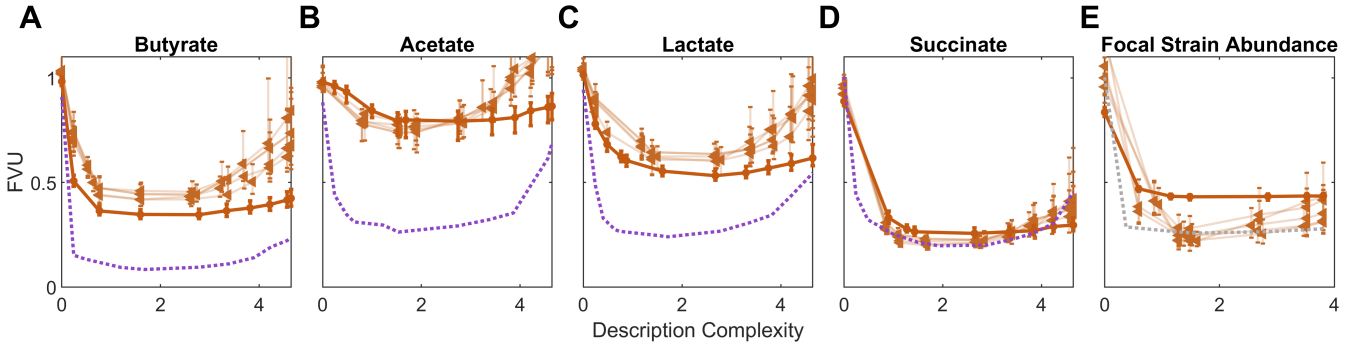

FIG. S6. **Downsampling test: observed emergent predictability is robust to sample size of datasets.** To test if emergent predictability is an effect of differing sample sizes, we downsample the low-richness datasets and re-obtain Pareto fronts. If the Pareto fronts inferred from the downsampled datasets do not drop to encompass the higher-richness Pareto fronts, then the predictability that emerges at higher richness is robust to our downsampling test. Each low-richness dataset was randomly sub-sampled to match the number of unique community member combinations assayed in the high-richness dataset. Plots show 5 random sub-samplings in triangles. Low-richness (orange circles) and high-richness (purple dashed line) Pareto fronts from Fig. 3 are shown for comparison. We find that for the majority of properties, the Pareto fronts obtained after downsampling do not close the gap between low richness and high richness, confirming the shift of the Pareto front is a robust observation.

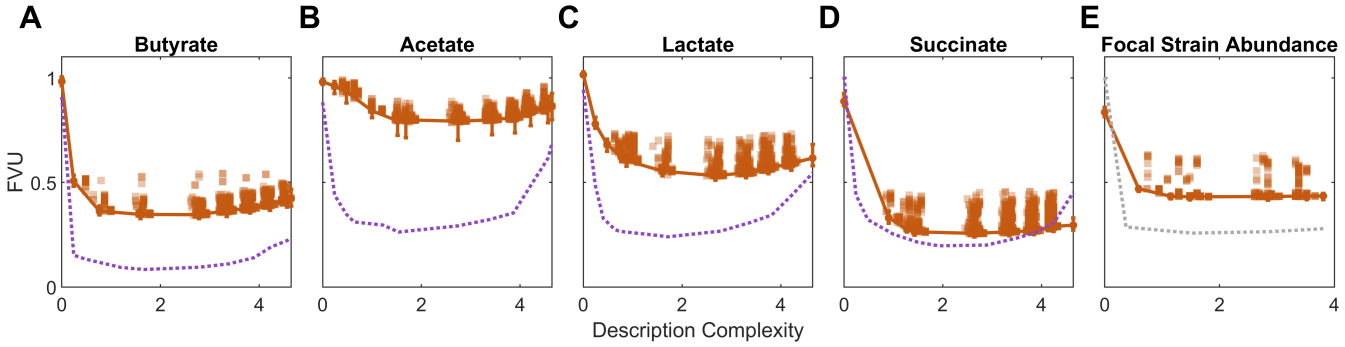

FIG. S7. **Relabeling test: observed emergent predictability is robust to Pareto front detection.** To test that the emergent predictability we observe is not due to our heuristic algorithm failing to detect the Pareto front, we examine the nearest-neighbor coarsened descriptions of the low-richness Pareto front. If any neighboring coarsened descriptions provide more predictive power, possibly dropping to the low prediction error of the high-richness Pareto front, this would then indicate that emergent predictability is not robust to our analysis of the data. Scattering predictive power of perturbed coarsened descriptions making up the low-richness Pareto front, where a perturbation is the relabeling of a single strain from one group to another. Original low-richness and high-richness Pareto fronts are shown for reference. Squares are median FVU of a given relabeled coarsened description Ψ. Only those relabeling perturbations that fall within 20 percent of the minimum FVU are shown for visualization purposes (those with larger error are not shown). We find that for each property, no neighboring coarsened descriptions fall below the Pareto front obtained for the low-richness dataset, confirming the shift of the Pareto front is a robust observation.

to shuffle the data in such a way that breaks the observed effect, while preserving as much of the original structure present in the data. To apply this sentiment to the datasets from [26] and [27], we aim to keep the abundance statistics of each species across communities fixed; so for each strain, we randomly permute its abundance in communities in which the species is initially present. In doing so, we see in Fig. S8 that the emergent predictability in high-richness communities disappears indicated by an increase in prediction error of the high-richness Pareto front obtained for the permuted abundance data. This is similarly broken when we instead permute the measured properties in high-richness communities, shown in Fig. S9. Both tests confirm that the shift of the Pareto front with increasing community richness is not artifact and that the observed emergent predictability is significant.

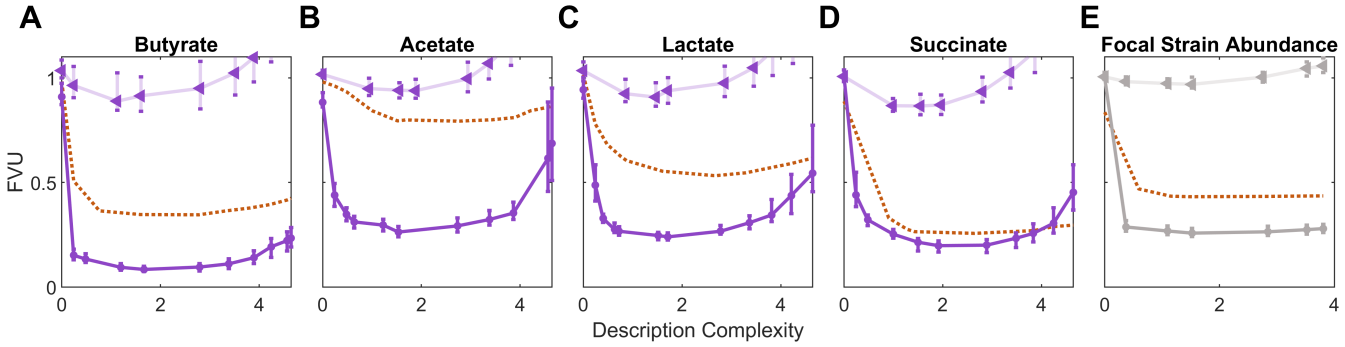

FIG. S8. **Permuting abundance data breaks emergent predictability at high richness.** To test that the observation of emergent predictability is not due to an artifact in the abundance data, we perform the following permutation test to see if it breaks our observation. We randomize the high-richness communities from the empirical datasets by shuffling the abundances of each strain with other instances in which the strains were present (preserving the presence-absence structure of each community). Pareto fronts are then inferred on the permuted data and plotted using triangle markers. The higher-richness Pareto fronts from Fig. 3 are plotted for reference, as well as the low-richness Pareto fronts for comparison (orange dashed lines). We find that the high-richness Pareto fronts obtained from randomized abundance data do worse than the low-richness Pareto fronts for each property of interest, indicating that our permutation test breaks the observed emergent predictability and confirming the shift of the Pareto fronts is not an artifact.

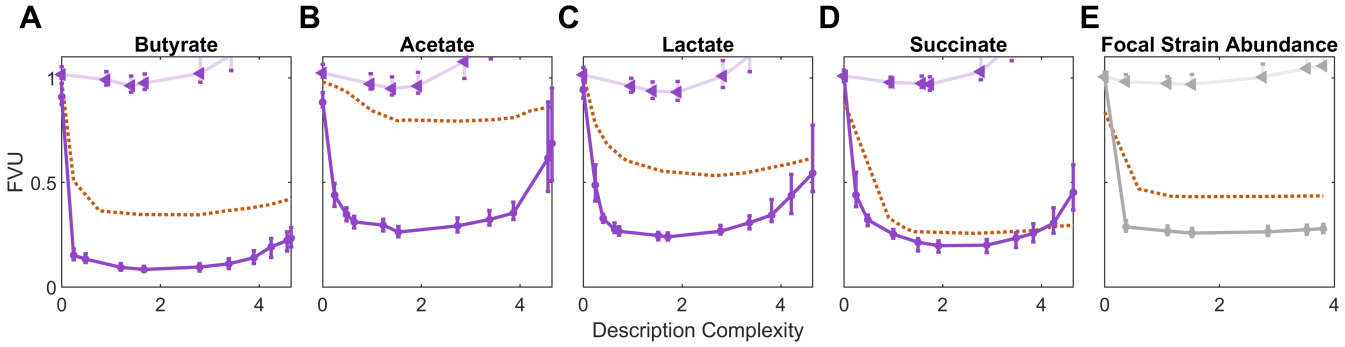

FIG. S9. **Permuting property measurements breaks emergent predictability at high richness.** To test that the observation of emergent predictability is not due to an artifact in the measurements of properties of interest, we perform the following permutation test to see if it breaks our observation. Similar to Fig. S8, we permuted the property data measured in high-ricnhess communities of the emprical datasets. Pareto fronts are then inferred on the permuted data and plotted using triangle markers. The higher-richness Pareto fronts from Fig. 3 are plotted for reference, as well as the low-richness Pareto fronts for comparison (orange dashed lines). We find that the high-richness Pareto fronts obtained from randomized property measurements do worse than the low-richness Pareto fronts for each property of interest, indicating that our permutation test breaks the observed emergent predictability and confirming the shift of the Pareto fronts is not an artifact.