Emergent predictability in microbial ecosystems

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

Lucas C. Graham and Mikhail Tikhonov*
Department of Physics, Washington University in St Louis, St Louis, MO, USA
(Dated: October 8, 2025)

Microbial ecosystems 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, the concept remains partly intuitive with no consistent definition, and most empirical examples described so far afford limited predictive power. Here, we propose an information-theoretic framework for defining and quantifying emergent simplicity in empirical data based on the ability of coarsened descriptions to predict community-level functional properties. Applying this framework to two published datasets, we demonstrate that all five properties measured across both experiments exhibit robust evidence of what we define as "emergent predictability": surprisingly, as community richness increases, simple compositional descriptions become more predictive. We show that standard theoretical models of high-diversity ecosystems fail to recapitulate this behavior. This is in contrast to simple self-averaging, which is well-understood and generic across models. We propose that, counterintuitively, emergent predictability arises when physiological or environmental feedbacks *oppose* statistical self-averaging along some axes of community variation. As a result, these axes of variation become increasingly predictive of community function at high richness. We demonstrate this mechanism in a minimal model, and argue that explaining and leveraging emergent predictability will require integrating large-N theoretical models with a minimal notion of physiology, which the dominant modeling frameworks currently omit.

Microbial communities perform key functions in a wide range of contexts, including human health, agriculture, and the global climate. Predicting community function from its composition is a key goal of microbial ecology. Recent advances in high-throughput sequencing allowed compositional information to be obtained in high resolution, and revealed that natural 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 has been that, despite this complexity, some patterns might nevertheless be predictable [37, 71, 72]. 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 [25] or a consistent ratio of predator species to prey species [8, 12]. More recently, moving beyond the patterns persisting despite high ecosystem diversity, it was proposed that some regularities could perhaps arise *because* of it [3, 20, 22], coining the term "emergent simplicity" [10, 18, 22, 42, 48]. 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 (e.g., the number of species) becomes large [1, 15, 20, 37–39, 51, 52].

The idea of emergent simplicity is very appealing, suggesting that functional predictions could perhaps be made directly in the complex natural context [6, 39]. However, so far this hope remained out of reach. In the absence of a consistent definition, it is unclear precisely what forms of "emergent simplicity" should be seen as surprising or useful. Many reported instances fall under the umbrella of functional convergence, focusing on reproducibility rather than prediction [8, 10, 12, 16, 18, 22, 25, 42, 48, 55, 56]. Such reproducibility can be difficult to distinguish from simple self-averaging or the central limit theorem [27, 41] and 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, as empirical observations of reproducibility are often approximate or limited to simplified conditions [66]. 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 need to make this concept well-defined, measurable, and focused on *prediction* rather than reproducibility. Here, we introduce a theoretical framework to explicitly quantify the predictive power of coarsened compositional descriptions in ecological data. This allows us to convert the aspiration of emergent simplicity into a testable quantitative hypothesis. Applying

* tikhonov@wustl.edu

our framework to two published datasets on lab-assembled ecosystems, we show that, remarkably, for all 5 properties measured across both experiments, the predictive power of coarse descriptions indeed improves with community richness. We further show that standard theoretical models of high-diversity ecosystems fail to recapitulate this behavior. This is in contrast to simple self-averaging, which is well-understood and generic across models. We propose that, counterintuitively, emergent predictability arises when physiological or environmental feedbacks *oppose* statistical self-averaging along some axes of community variation. As a result, these axes of variation can become increasingly predictive of community function at high richness. We demonstrate this mechanism in a minimal model, and argue that explaining and leveraging emergent predictability will require integrating large-N theoretical models with a minimal notion of physiology, which the dominant modeling frameworks currently omit.

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. Our goal is to quantify the *predictive power of simple (coarsened) descriptions* in ecological datasets of this kind, 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 [31]. The rate of digestion (the amount of methane produced) can vary widely [34, 49], but, remarkably, can be largely explained by simple models coarse-graining this vast diversity into just 4 functional classes (acidogens, propionate-degrading acetogens, butyrate-degrading acetogens, and methanogens [7]).

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: $[1 \dots S] \xrightarrow{\Psi} [1 \dots K^\Psi]$ with $K^\Psi \leq S$ (Fig. 1B). Some natural choices of coarsening Ψ might be to group strains based on their taxonomy or key traits, but we will allow any choice of Ψ as valid, and define *description complexity* $I(\Psi)$ as the mutual information between the label of the coarse-grained class and the identity of the microscopic strain:

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

Here S_j^Ψ 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 Supplementary Text, section B 1). This metric $I(\Psi)$ reaches its maximum for the fully microscopic description, when each strain is its own class: $K^\Psi = S$ and $I(\Psi) = \log S$.

For any coarsening Ψ, we will assess the prediction error of the best model from a specified *model class* $\mathcal{M}$, taking as input the combined abundances $\tilde{n}_{j\mu}^\Psi = \sum_{\Psi(i)=j} n_{i\mu}$ and predicting Y. This approach echoes other recent attempts to relate community-level properties to coarse-grained compositional features [47, 65]. The model class $\mathcal{M}$ could be mechanistically motivated (for instance, the class of Generalized Lotka-Volterra models predicting a dynamical property like future abundance of a focal species), or some general-purpose ansatz (e.g., a neural network). In this work, we will pick a particularly simple class of models, namely linear regressions $Y_\mu = \beta_0^\Psi + \sum_i \beta_i^\Psi \tilde{n}_{i\mu}^\Psi + \epsilon_\mu$ (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 ecosystem function [3, 67], 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 ($\{\tilde{n}_{i\mu}^\Psi\}$), which allow us to train an instance of a linear regression model (coefficients β_i^Ψ), 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).

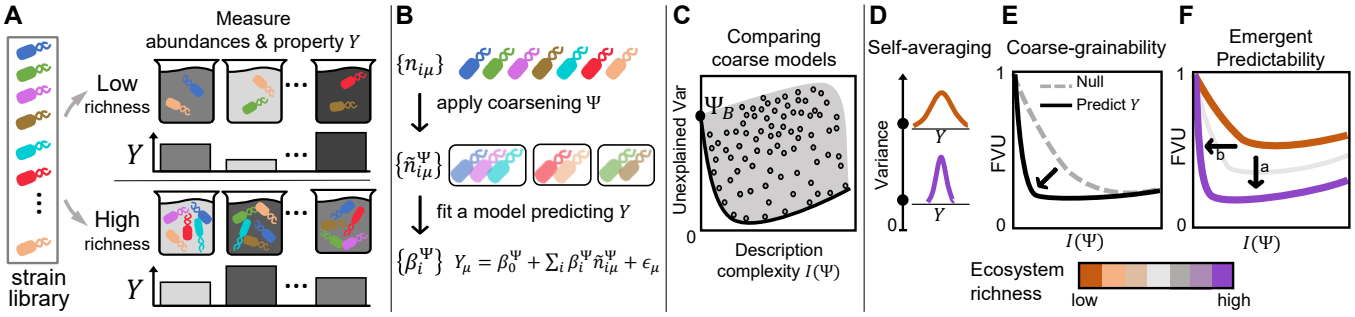

FIG. 1. **A quantitative framework for interrogating emergent simplicity in data. A:** A general experimental protocol with community richness (number of strains) as a control parameter. A fixed library of strains is subsampled to assemble communities of low richness (top) or high richness (bottom). 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 variance unexplained 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. The behavior of the Pareto fronts defines three distinct (but not mutually exclusive) forms of "emergent simplicity": **D:** self-averaging, when higher community richness leads to a tighter distribution of property Y; **E:** coarse-grainability, when coarsened descriptions are surprisingly predictive (compared to a null model), at fixed richness; or **F:** emergent predictability, when the predictive power of coarsened descriptions improves with ecosystem richness.

C. Quantifying the performance of simple models

For each candidate "simple model" (coefficients β_i^Ψ) indexed by the coarsening Ψ, we can quantify its prediction error, computed on held-out data. The well-established bias-variance tradeoff from statistical learning theory indicates that if we could evaluate and plot the prediction error (the variance left unexplained) against the description complexity $I(\Psi)$ 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_B is the least predictive, coarsest description that groups all strains together ($I(\Psi_B) = 0$). If no compositional information is available—only the total biomass—the unexplained variance of Y is close to its total variance. (For properties that scale with biomass, it might be more meaningful to seek to predict their value per unit biomass; see Supplementary Text, section B 2.) 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 behavior of this Pareto front.

D. Distinguishing self-averaging, coarse-grainability and emergent predictability

The intuitive notion of "emergent simplicity" can be codified in three distinct ways. These are important to distinguish, since as we will see, they have very different properties.

The simplest, and most familiar, version of emergent simplicity is self-averaging. As community richness increases, one might observe the distribution of Y_μ become increasingly tight. Much previous work indicates that properties of high-richness communities may indeed exhibit such "functional convergence." [8, 10, 12, 16, 18, 20, 22, 25, 42, 48, 55, 56] We can think of this width as the prediction error of a particularly trivial model which predicts a constant value, such that the unexplained variance is synonymous with the total variance.

In practice, the predictions that interest us are rarely of this "zero-information" kind. Instead, we usually seek to predict properties of a community based on its composition. To define notions of simplicity that go beyond simple self-averaging, it is convenient to normalize away the overall width of the distribution of Y, and think of the *fraction* of unexplained variance (FVU), defined as the mean square error of prediction divided by the variance of Y.

In these normalized units, the Pareto front descends from approximately 1 (no compositional information; worst prediction) to the best prediction attainable with this model class (in the limit of infinite data—neglecting overfitting—this would correspond to the full microscopic description). How quickly it descends encodes the notion of "coarse-grainability". For some properties, we may find that a surprisingly coarse description already enables good predictions; this form of simplicity has also been widely reported [21, 36, 57, 64]. To rigorously define this concept, it is essential to quantify the word "surprisingly," defining a null expectation to which the shape of this curve should be compared. Our framework will enable us to do so, allowing us to examine "coarse-grainability" of ecological properties in empirical data.

Finally, we can ask what happens to the Pareto front when we change ecosystem richness, and compare the fronts con-

structed for a dataset restricted to only high-richness examples *vs.* only low-richness examples, for the same property Y (Fig. 1A). A consistent shift of the Pareto front would indicate that, as richness is increased, models of comparable complexity become increasingly predictive (Fig. 1F, arrow "a"), and/or that same prediction accuracy can be achieved by increasingly coarse descriptions (Fig. 1F, arrow "b"). While the two directions of shift have different verbal interpretations, in general they cannot be formally distinguished, and we will instead quantify "emergent predictability" by measuring the area contained between the two Pareto fronts (see Methods).

The rest of the paper can be summarized in two claims: (1) Self-averaging, coarse-grainability and emergent predictability are all observed in empirical data. However, (2) the three phenomena differ dramatically in their origins, properties and the attention they deserve. Today, the broad usage of the phrase "emergent simplicity" came to encompass any observation of something simple in a complex community context. In this sense, all three phenomena just defined could be collated under this umbrella term. However, as we will show, self-averaging is generic and ubiquitous; on its own, it is neither surprising nor special. Coarse-grainability does not come "for free" (from diversity alone), but can be a straightforward consequence of the fact that some species are more similar than others. While missing from some widely used models that omit such structure, its empirical observation should also not be seen as surprising, unless the observation comes with a quantitative argument that the observed coarse-grainability is stronger than expected. In contrast, emergent predictability is missing from both random and structured theoretical models – and yet, as we will see, it is strongly supported by data. This is the phenomenon that we propose should be seen as "emergent simplicity" *sensu stricto*. Its empirical observation is surprising and demands a theoretical explanation, and we will present evidence supporting a candidate mechanism for its emergence.

II. RESULTS

A. Data selection

To apply our framework, 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 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 identifies two datasets whose structure allows our question to be asked. These are the data of Clark *et al.* [11], 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.* [33], who used a droplet-based platform to screen communities of 1-14 isolates, measuring the abundance of a model plant symbiont *H. frisingense* as the property of interest. For each dataset, we resolve three tiers of community richness: low richness, $1-5$ strains; middle richness, $10-15$ strains; and high richness, $21-25$ strains. Together, these two datasets provide measurements of five properties we can examine for evidence of emergent simplicity. To construct the information-prediction Pareto fronts defined above, we use a Metropolis-Hastings algorithm to search the space of possible coarsenings (for details, see Methods).

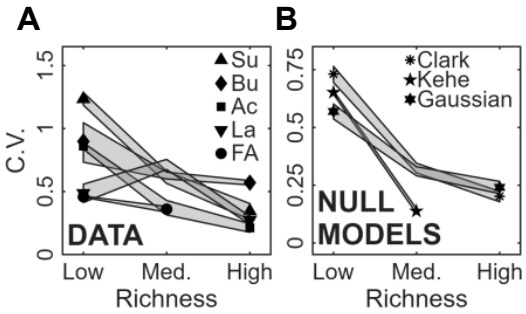

FIG. 2. **Self-averaging is generic. A:** Coefficient of variation (CV) for each of the properties measured in Refs. [11] and [33], computed across communities of a given richness tier. The decreasing trend indicates self-averaging. Legend labels: Bu = Butyrate, Ac = Acetate, La = Lactate, Su = Succinate, FA = Focal species abundance. Shading marks ± 2 SD using standard bootstrapping with 1000 sub-samplings. **B:** CV of synthetic properties generated from a null model where all species contribute proportional to their abundance with a weight drawn from a unit uniform distribution (see text). Three synthetic properties were computed using either shuffled abundance data from Refs. [11, 33] or by drawing random abundances from a Gaussian with mean and variance both set to 1. Null models reproduce the signature decrease of CV with richness, demonstrating that self-averaging is a generic expectation. Shading as in (A).

B. Self-averaging is generic and ubiquitous

For all five properties measured across both datasets, we see that their coefficient of variation decreases with community richness (Fig. 2A). This observation means their value is more reproducible across high-richness communities than it is among low-richness communities, and as such, could be described as a form of "emergent simplicity." However, this same behavior is observed in the simplest of null models that retain no biological structure: self-averaging is preserved even if we randomly reshuffle the measured abundances, or use a purely synthetic property to which each species contributes independently (Fig. 2B; see Methods for details). This is not surprising, since both central-limit averaging and the increasing compositional similarity of high-richness samples (drawing $S > 20$ species out of a pool of only 25) naturally contribute to this effect. Thus, when talking about emergent

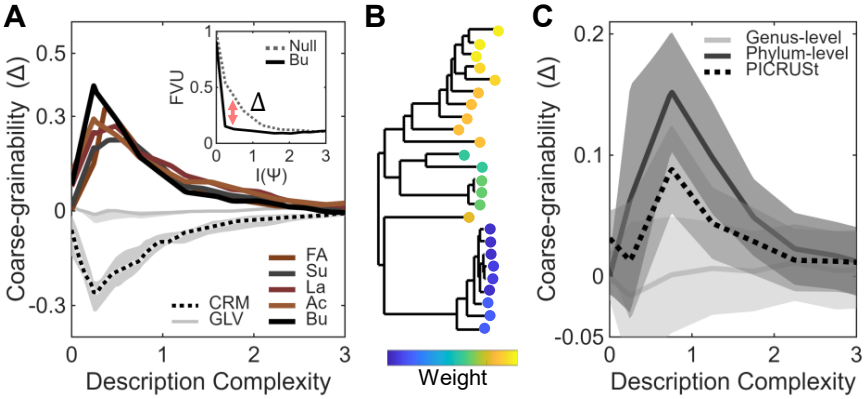

FIG. 3. **Coarse-grainability arises from community structure.** **A:** Difference between the actual Pareto front and null (Δ) as a function of description complexity, shown for empirical community properties measured in Refs. [11] and [33], and contrasted with model-derived properties (community invadability) in two standard ecological models: random-interaction CRM and GLV. $\Delta > 0$ corresponds to coarse-grainability. See inset for the actual Pareto front and null comparison in the case of Butyrate as the property of interest. For CRM and GLV, the shading represents the quartiles in Δ across 10 random invaders. For empirical curves, error bar shading is omitted for clarity but is provided in the Supplementary Text (Fig. S8) Abbreviations as in Fig. 2. **B:** The simplest model capturing the fact that species traits are non-random: species' contributions to a synthetic property (contribution weight) are assumed to be phylogenetically correlated, with conservation depth a tunable parameter (see Methods). The phylogenetic tree corresponds to the 25 species assayed by Clark et al. (reproduced from Ref. [11]). Leaf coloring illustrates the case where contribution weight is deeply conserved. **C:** Δ computed for a synthetic community property generated from the Clark et al. abundance data using phylogenetically correlated weights with conservation depth set to the genus or phylum level (see Methods). As expected, increasing conservation depth parameter leads to greater coarse-grainability. For comparison, the dashed line shows coarse-grainability for a synthetic property generated from central carbon metabolism gene profiles of the same 25 species predicted by PICRUSt2 (see Methods). Lines and shading represent median and quartiles across 10 replicate weight samplings.

simplicity, it is essential that our terms are sufficiently well-defined to avoid any potential confusion with effects of trivial origin. To this end, from now on, we will ignore any absolute changes in the variance, assessing instead the *fraction* of variance explained by various models, and examining the shape and behavior of the Pareto front in these normalized units (Fig. 1E,F). Additionally, our null-model comparisons will employ the same sampling structure as the data, and whenever appropriate, the actual measured abundance tables, so that the compositional similarity of samples is fully retained.

C. Coarse-grainability arises from community structure

We say that a community property Y is coarse-grainable if coarse compositional representations are *surprisingly* predictive of Y, compared to our null expectation. To establish the null expectation, we construct a synthetic property that uses the measured abundance table and preserves some statistics of the microscopic model fit to the actual property of interest, but assumes that species' contributions are random and uncorrelated, eliminating any notion of biological similarity or relatedness between them (for details, see Methods). By construction, the Pareto front for this synthetic property has the same endpoints as the Pareto front of interest (Fig. 3A, inset). We quantify coarse-grainability by measuring the difference between the actual Pareto front and the null comparison, as a function of description complexity; positive values indicate that Y is coarse-grainable. Applying this methodology to the two datasets, we find that all 5 observables across both datasets show a strong signature of coarse-grainability (Fig. 3A). Note that coarse-grainability is defined without reference to varying community richness; the analysis in Fig. 3A is for high-richness communities ($S > 20$; for Kehe et al. $S > 10$).

Coarse-grainability is clearly a useful feature for understanding complex ecosystems. It is therefore worth noting that some of the dominant theoretical frameworks of high-diversity ecology are missing this property. To illustrate this, we consider two classic frameworks widely used to model high-diversity ecosystems: the Generalized Lotka-Volterra model (GLV) with a random interaction matrix, and a Consumer-Resource model (CRM) with a random resource consumption matrix, choosing "community invadability" as the property of interest (specifically, Y is defined as the post-invasion abundance of some specified strain, initially absent from the community). This choice is convenient because it is a relevant and non-trivial community property that can be accessed in both modeling frameworks. The model parameters and simulation protocol are described in the Methods, but the outcome of the analysis is that neither model exhibits coarse-grainability (Fig. 3A, shaded).

The origin of this contrast is easy to pinpoint: real species and interactions are not random. If relatedness of species is taken into account, then even the simplest models can exhibit coarse-grainability. To illustrate this, we construct a simple model in which the only assumption is that closely related species contribute similarly to a community-level property. Using the abundance data from Clark et al., we define a synthetic community property to which each species con-

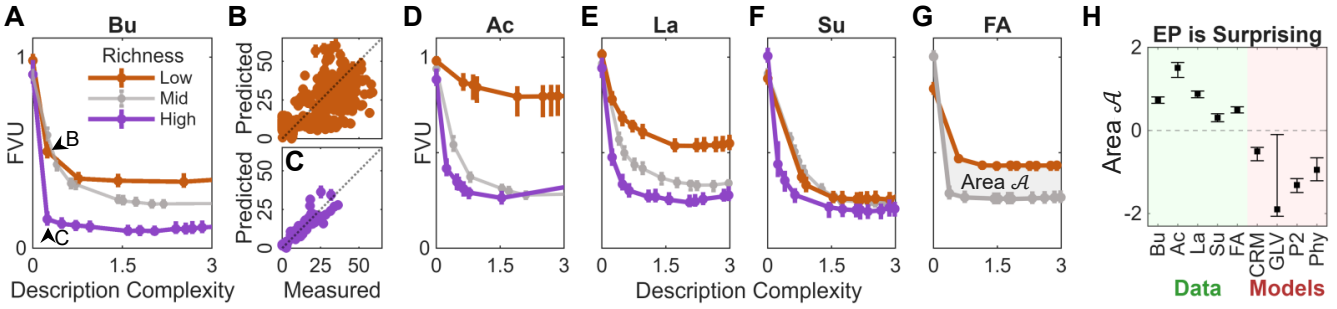

FIG. 4. **Both empirical datasets exhibit emergent predictability. A:** The Pareto front in error-complexity plane, computed for the butyrate measurements of Clark *et al.* [11]. Lines show the fraction of variance left unexplained (FVU) by the best model within a given description complexity, when dataset is restricted to communities of low richness ($S \leq 5$, orange), medium richness (gray) or high richness ($S > 20$, purple). Points indicate the median out-of-sample FVU computed over 100 random 50-50 splits of the data into training and testing sets; error bars are the 1st and 3rd quartiles. The shift of the Pareto front indicates that at higher richness, coarse descriptions become more predictive (*cf.* Fig. 1F). Panels **B-C** illustrate this result by plotting the predicted *vs.* measured concentration for the two models indicated by the arrowheads. At high richness, a highly coarsened model explains 85% of variance in butyrate production (C). In contrast, for low-richness communities, even the most predictive model of comparable complexity explains only 49% of variance (B). **D-G:** Same as (A) for the four other properties measured in Refs. [11] and [33]. The Kehe *et al.* dataset (panel G) only assayed 14 isolates and is missing the high-richness tier ($S > 20$ strains). **H:** To quantify emergent predictability (EP), we measure the signed area $\mathcal{A}$ between the low-richness and highest-available-richness Pareto fronts of previous panels (highlighted in G). The quantification confirms that all measured properties exhibit emergent predictability. This is in contrast to all theoretical models we considered, including those that exhibit coarse-grainability (CRM, random-interaction consumer-resource model; GLV, random-interaction Lotka-Volterra; P2, PICRUSt2 central carbon metabolism; Phy, phylogenetically conserved contributions; other abbreviations as in Fig. 2). Error bars show quartiles of the respective distributions. For empirical data, they represent the uncertainty of Pareto front estimation; for models, they include variability across random model instances (see Methods).

tributes independently, but with weights that are correlated according to phylogenetic relatedness (Fig. 3B shows an example). A single tunable parameter controls the phylogenetic scale at which trait correlations decay, capturing the depth of conservation (see Methods). As expected, increasing the depth of trait conservation leads to greater coarse-grainability (Fig. 3C). Of course, the traits relevant to the measurements of Clark et al. and Kehe et al. are unlikely to have a single relevant scale of conservation depth. As an example of more realistic structure, we also include a coarse-grainability curve for a different synthetic property, constructed from the presence of carbon metabolism genes (KEGG orthologs) predicted by PICRUSt2 [17] for the same set of species; see Methods for details. The shape of this curve confirms that a realistic amount of structure is sufficient to generate coarse-grainability. We note that the scale of the y axis in Fig. 3C depends on arbitrary modeling choices and is not meant to be directly compared to Fig. 3A; our focus is on the coarse-grainability Δ becoming positive at intermediate complexity.

This analysis demonstrates that even minimal models incorporating phylogenetic signal can exhibit coarse-grainability. Some theoretical frameworks intentionally omit this relatedness structure, and thus miss this effect, but in any empirical scenario (or in models using realistic trait matrices, e.g., CRM based on measured resource preferences [24, 28]), at least some degree of coarse-grainability is to be expected.

D. Emergent predictability is surprising and demands explanation

Finally, we turn to emergent predictability, encoded in the behavior of the Pareto front under changes of richness. For each dataset, we resolve three tiers of community richness. We apply the same search algorithm as above to infer the respective Pareto front of each richness tier. The results are shown in Fig. 4A,D-G.

Remarkably, we find that all five properties exhibit emergent predictability. We emphasize that the Pareto front is not an abstract concept: each point corresponds to an explicit model, and we consistently find that coarse models become increasingly predictive in communities of higher richness. As an illustration, Fig. 4B&C plot the measured butyrate concentration against the prediction of the simplest models identified by our framework, labeled with arrowheads in A. At high richness, a coarsening exists whereby a simple two-parameter regression explains 85% of variance in butyrate production (Fig. 4C). In contrast, for low-richness communities, no model of comparable complexity can explain more than 49% of variance; the best-performing example is shown in Fig. 4B.

As a control, this behavior is not observed if the values of Y or species abundances are randomly permuted (see Supplementary Text, Figs. S6, S7), indicating that the behavior observed in Fig. 4 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

Supplementary Text, Figs. S3, S4, S1, respectively. We conclude that emergent predictability is robustly supported by data.

Crucially, however, none of the standard models show this behavior. To see this, we quantify emergent predictability by measuring the area contained between the high-richness and low-richness Pareto fronts (see Methods). Fig. 4H summarizes the contrast between the empirical results and all the theoretical models discussed so far. Each of these models exhibits self-averaging, and multiple models exhibit some degree of coarse-grainability, provided they retain some notion of relatedness between species (PICRUSt; phylogenetically conserved contributions). However, none of these models exhibit emergent predictability. (See also Supplementary Text section C 1.)

E. Mechanistic origin of emergent predictability

At least for the system of Clark *et al.*, the emergent predictability appears to be related to the dominant structuring role of pH, which is one-dimensional. While the data is insufficient to demonstrate this directly (independent measurements of pH were conducted for only a subset of communities), we can obtain strong indirect evidence by leveraging the fact that four functional properties were measured simultaneously. These four measurements (the final concentrations of butyrate, acetate, succinate and lactate produced) define, for each community, a point in what we will call the "functional space". Fig. 5A 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. S11). This dominant axis of variation is strongly correlated with community pH (Fig. 5A, inset), consistent with Ref. [11] identifying pH-mediated interactions as a key factor structuring these communities. This observation likely explains the sudden drop in the Pareto fronts observed in Fig. 4A,D-G. Indeed, any combination of species whose abundance is correlated with pH will explain a substantial fraction of variance in any of the community properties that have a loading on the dominant principal component of Fig. 5. As a result, even a little compositional information can be sufficient to improve predictive power, which is what the Pareto front indicates. Crucially, the analysis of Fig. 4 used no knowledge of pH. Thus, our framework may provide a prior-knowledge-agnostic way for discovering the existence of strong ecosystem-structuring factors.

For specifically Fig. 4A (butyrate), the results of Clark *et al.* [11] provide some additional mechanistic insight. 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. 4A is the grouping {AC, all other strains}, and provides an excellent prediction for the amount of butyrate produced (Fig. 4C). In contrast, in low-richness mixtures, all five strains contribute to the production, and no single compositional feature is similarly predictive.

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. 4A, D-G are all a consequence of some idiosyncrasies of specific taxa. However, the common outcome

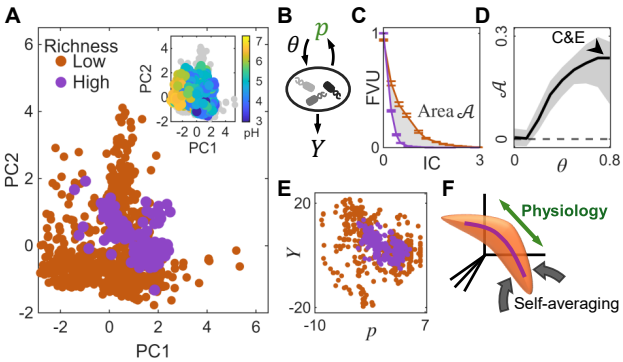

FIG. 5. **Mechanistic origin of emergent predictability.** **A:** Principle component analysis (PCA) of the four-dimensional functional measurements of Clark *et al.* (butyrate, acetate, lactate, succinate) indicates that high-richness samples form a lower-dimensional subspace. Coloring the PCA plot by measured pH (inset) reveals that the dominant axis of functional variation is correlated with community pH. Principal axes computed for the dataset as a whole; mid-richness communities are omitted for clarity (but included in the inset). **B:** A minimal model that can reproduce emergent predictability. The key ingredient is a feedback loop (of strength θ) with a shared environmental variable p. Species differ by their preferred interval of p which allows them to be active; see text for details. Species' contribution strengths to the property of interest Y are random and uncorrelated (no special structure is assumed). **C:** The model exhibits emergent predictability; shown is an example for feedback strength $\theta = 0.8$. Richness tiers match Clark *et al.*, with 800 simulated communities per tier (see Methods). **D:** In the model, emergent predictability is controlled by feedback strength θ. As in Fig. 4, emergent predictability is quantified as the signed area between low- & high-richness Pareto fronts (shaded in panel C). Error bar shading shows quartiles over 10 random model instances. The feedback strength corresponding to the example shown in panels C & E is highlighted. **E:** Scatter plot of the community property Y versus the environmental variable p for the same simulated dataset as used in panel C; colors as in (A). The plot offers an intuitive explanation of the origin of emergent predictability in the model. At higher richness (purple), the point cloud is tighter, but the tightening is anisotropic, allowing p (and, by extension, p-dependent compositional features) to become more predictive of Y. **F:** A cartoon graphical summary of the proposed general mechanism illustrated by our model. Communities are represented as points in an abstract "functional space." As richness increases, communities generally become more functionally similar due to generic forms of self-averaging. However, if physiological or environmental feedbacks oppose self-averaging along some axes of community variation (e.g., the p axis in our model), then at high richness, such axes will become responsible for an increasingly larger fraction of variance in related properties.

observed across two independent datasets suggests that emergent predictability as quantified here may have a more general origin.

F. Candidate general mechanism: ecological feedback that opposes self-averaging

We will now describe and support a candidate general mechanism for emergent predictability to arise. Somewhat counter-intuitively, we propose that emergent predictability arises when some physiological or environmental feedback *opposes* statistical self-averaging along certain axes of community variation. To illustrate, we construct a simple model (Fig. 5B) inspired by Refs. [2, 50, 59]. Let species i be characterized by metabolic activity a_i. The metabolic activity defines its growth, scales its contribution to a collective property of interest Y, but also its influence on a shared environmental variable p (e.g., pH, osmolarity, oxygen content, toxin concentration):

$$\begin{aligned}
\dot{n}_i &= n_i a_i && \text{population growth} \\
\dot{r} &= -\sum_i a_i n_i && \text{nutrient depletion} \\
\dot{p} &= \sum_i a_i n_i c_i && \text{dynamics of } p \\
\dot{Y} &= \sum_i a_i n_i y_i && \text{dynamics of } Y.
\end{aligned}$$

We will assume no special structure in species' contributions to either p or Y: both c_i and y_i will be drawn randomly and independently from Gaussian distributions. However, crucially, we will assume that the activities a_i are impacted by feedback from p. Specifically, each species i has a different interval of p that it prefers, and can only be active within a finite range of a preferred value (denoted ρ_i):

$$a_i = r\, F(|p - \rho_i|).$$

Here, r is nutrient availability, and F has a finite width $\simeq \Delta p$ (we pick F to be a truncated Gaussian; see Methods). As an example, a species acidifying its own environment can poison its own growth in monoculture, even if nutrient is still available [58]. The parameter $\theta = 1/\Delta p$ encodes the feedback strength: large θ corresponds to species highly sensitive to changes in p (a narrow activity range), whereas $\theta = 0$ corresponds to p being irrelevant. This feedback, assumed to be mediated by species physiology, is the key ingredient of our model. For more details, see Methods.

Fig. 5C confirms that this model displays emergent predictability (simulation parameters are listed in Methods). As we expected, the magnitude of the effect is controlled by feedback strength θ (Fig. 5D). Intuitively, this can be understood as follows. In the analytically solvable limit $\theta \to 0$ (no feedback), it is easy to see that p is irrelevant, the species' contributions to Y are subject to simple self-averaging, and there is no emergent predictability (for more discussion, see Supplementary Text, section C 1). At non-zero θ, self-averaging continues to play a role leveling out the quirks of individual species within a tier of similar p preference. However, along the p axis, the p-mediated ecological feedback opposes self-averaging. For instance, a species that is strong driver of p (large c_i) will always drive this environmental variable to the upper limit of its viable range, which is a fixed value set by species physiology ($\simeq \rho_i + \Delta p$) and does not depend on richness. As a result, the axis of variability associated with p preference maintains its relevance while other differences self-average, and so becomes responsible for an increasingly larger fraction of the total variance of the property of interest (illustrated in Fig. 5E, F). A more detailed analysis of this mechanism can be found in the Supplementary Text.

The feedback of the kind we described is expected to be widespread [50], and Fig. 5D demonstrates it is sufficient to endow a system with emergent predictability. Thus, our analysis suggests that emergent predictability is a nontrivial and exciting feature that might well be ubiquitous in natural communities.

III. DISCUSSION

The promise of emergent simplicity is to enable predictions that work because of diversity, not despite it. However, so far, the field was unable to reach a consensus on whether emergent simplicity is real, whether it is surprising, and whether it is useful. Answering the first question requires a quantitative definition; the second requires a null model for comparison. For the third, we need an explicit focus on prediction rather than reproducibility, and a clear understanding of whether emergent simplicity is a biological statement, or a statistical one. Indeed, predictions that are a direct mathematical consequence of the number of species being large, independent of the underlying biology, are unlikely to be sufficiently nontrivial to be useful in practice.

Here, we developed an information-theoretic framework to quantify the notion of emergent simplicity, and applied this framework to data from two separate experiments on microbial communities assembled in the lab. Our approach enabled us to define and distinguish coarse-grainability and emergent predictability, and to contrast both of these phenomena with "emergent reproducibility" (self-averaging) considered in much previous work.

All empirical properties we considered showed a clear signature of coarse-grainability. However, as we discussed, this is not necessarily surprising since for any real ecosystem, at least some degree of coarse-grainability is expected. This does not mean that coarse-grainability is uninteresting. Some empirical reports indicate a truly striking extent of coarse-grainability [31], and it is not clear that this can be fully explained by the relatedness of community members, especially given the extensive, functionally relevant diversity that is known to persist even at the level of strains [23]. A clear quantitative metric of coarse-grainability, as presented here, should help investigate this in future work.

Our main findings concerned emergent predictability, whereby coarser compositional descriptions become more

predictive at higher richness. Our key result is that this phenomenon is robustly supported by empirical data; yet is absent from standard models of theoretical ecology, making it highly non-trivial and surprising.

In four of the five cases (measurements of Clark *et al.*), the phenomenon appeared to be linked to the dominant structuring role of pH in these communities. This observation echoes other recent examples where coarse-grained predictions are enabled by general physiological [18], metabolic [71], or toxicity-mediated constraints [13]. We described a mechanism explaining how structuring factors of this kind may acquire a higher explanatory power at higher diversity. We illustrated it in a minimal model demonstrating how emergent predictability can arise from a physiology-mediated feedback with a collective environmental variable.

Perhaps the most exciting implication of our analysis is that it reconciles the generality of the central limit theorem with the possibility of nontrivial predictions. Indeed, under the mechanism we proposed, the diversity-dependent simplification is ultimately still a form of self-averaging – but one that leads to a stronger collapse onto a lower-dimensional subspace governed by metabolic or physiological constraints (Fig. 5F). Understanding this behavior will require extending the theoretical investigations of large-N ecosystems beyond the Gaussian-interaction regimes [4, 15, 40], and integrating these ongoing efforts with new models capable of capturing the role of such structuring factors.

Our approach has limitations. Our definitions of emergent simplicity required specifying a class of models, and here, due to the limited size of the available datasets, our choice (linear regression) was particularly simple. However, our theoretical framework is more general; with sufficient data, the natural generalization of the Pareto fronts in Figs. 1, 3-5 would be to evaluate the predictive power of machine learning models as a function of the dimension of the compressed (latent-space) representation of community composition, similar to the approach of Ref. [54]. Understanding the extent to which emergent predictability depends on the choice of model class remains a question for future work. We should also 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 [53]. This was not an issue for our work, as the synthetic communities considered here were assembled from a small pool of well-delimited taxa. However, for natural communities, this simple metric of diversity would need to be replaced by one of the better-suited metrics developed in the ecological literature [14, 29, 44].

All simulations performed in MATLAB (Mathworks, Inc.). The associated code, data and scripts to reproduce all figures in this work are publicly available [46].

ACKNOWLEDGMENTS

We thank S. Allesina, A. Goyal. K. Gowda, T. Grand-Pre, C. Holmes, S. Kaplan, S. Kuehn, R. McGee, A. Murugan, O. Venturelli and K. Wood for helpful discussions. MT acknowledges a CAREER award from the National Science Foundation (PHY-2340791). This work was also supported by the National Science Foundation grant PHY-2310746, and the Michigan Pioneer Fellows Program award to JM.

[1] 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(3):033406, 2018. Publisher: IOP Publishing.

[2] D. R. Amor, C. Ratzke, and J. Gore. Transient invaders can induce shifts between alternative stable states of microbial communities. *Science advances*, 6(8):eaay8676, 2020.

[3] S. Arya, A. B. George, and J. P. O'Dwyer. Sparsity of higher-order landscape interactions enables learning and prediction for microbiomes. *Proceedings of the National Academy of Sciences*, 120(48):e2307313120, 2023.

[4] S. Azaele and A. Maritan. Generalized dynamical mean field theory for non-gaussian interactions. *Physical Review Letters*, 133(12):127401, 2024.

[5] 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(9):2156–2161, 2018. Publisher: National Acad Sciences.

[6] 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.

[7] 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(4):237, 2021. Publisher: Multidisciplinary Digital Publishing Institute.

[8] F. Briand and J. E. Cohen. Community food webs have scale-invariant structure. *Nature*, 307(5948):264–267, 1984.

[9] G. Bunin. Interaction patterns and diversity in assembled ecological communities. *arXiv preprint arXiv:1607.04734*, 2016.

[10] 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(5):1340–1350, 2015.

[11] R. L. Clark, B. M. Connors, D. M. Stevenson, S. E. Hromada, J. J. Hamilton, D. Amador-Noguez, and O. S. Venturelli. Design of synthetic human gut microbiome assembly and butyrate production. *Nature Communications*, 12(1):3254, May 2021. Number: 1 Publisher: Nature Publishing Group.

[12] J. E. Cohen. Ratio of prey to predators in community food webs. *Nature*, 270(5633):165–167, 1977.

[13] K. Crocker, K. K. Lee, M. Chakraverti-Wuerthwein, Z. Li, M. Tikhonov, M. Mani, K. Gowda, and S. Kuehn. Environmentally-dependent interactions shape patterns in gene content across natural microbiomes. *Nature Microbiology*, page in print, 2024.

[14] V. Crupi. Measures of biological diversity: Overview and unified framework. *From Assessing to Conserving Biodiversity:*

Conceptual and Practical Challenges, pages 123–136, 2019.

[15] W. Cui, R. Marsland III, and P. Mehta. Diverse communities behave like typical random ecosystems. *Physical Review E*, 104(3):034416, 2021.

[16] M. S. Datta, E. Sliwerska, J. Gore, M. F. Polz, and O. X. Cordero. Microbial interactions lead to rapid micro-scale successions on model marine particles. *Nature communications*, 7(1):11965, 2016.

[17] G. M. Douglas, V. J. Maffei, J. R. Zaneveld, S. N. Yurgel, J. R. Brown, C. M. Taylor, C. Huttenhower, and M. G. Langille. Picrust2 for prediction of metagenome functions. *Nature biotechnology*, 38(6):685–688, 2020.

[18] 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(1):29–42, 2022. Publisher: Elsevier.

[19] 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(220):220ra11–220ra11, 2014.

[20] L. Fant, I. Macocco, and J. Grilli. Eco-evolutionary dynamics lead to functionally robust and redundant communities. *bioRxiv*, 2021. Publisher: Cold Spring Harbor Laboratory.

[21] J. K. Geißert, E. Bohn, R. Mostolizadeh, A. Dräger, I. B. Autenrieth, S. Beier, O. Deusch, A. Renz, M. Eichner, and M. S. Schütz. A computational model of bacterial population dynamics in gastrointestinal yersinia enterocolitica infections in mice. *Biology*, 11(2), 2022.

[22] 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(6401):469–474, 2018. Publisher: American Association for the Advancement of Science.

[23] A. Goyal, L. S. Bittleston, G. E. Leventhal, L. Lu, and O. X. Cordero. Interactions between strains govern the eco-evolutionary dynamics of microbial communities. *eLife*, 11:e74987, Feb. 2022. Publisher: eLife Sciences Publications, Ltd.

[24] M. Gralka, S. Pollak, and O. X. Cordero. Genome content predicts the carbon catabolic preferences of heterotrophic bacteria. *Nature Microbiology*, 8(10):1799–1808, 2023.

[25] H. Heatwole and R. Levins. Trophic structure stability and faunal change during recolonization. *Ecology*, 53(3):531–534, 1972.

[26] 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(5442):1123–1127, 1999.

[27] P.-Y. Ho and K. C. Huang. Challenges in interpreting functional redundancy and quantifying functional selection in microbial communities. *Cell Systems*, 16(8):101350, Aug. 2025.

[28] P.-Y. Ho, T. H. Nguyen, J. M. Sanchez, B. C. DeFelice, and K. C. Huang. Resource competition predicts assembly of gut bacterial communities in vitro. *Nature Microbiology*, pages 1–13, 2024.

[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(2):97–106, 2014.

[30] F. I. Isbell and B. J. Wilsey. Increasing native, but not exotic, biodiversity increases aboveground productivity in ungrazed and intensely grazed grasslands. *Oecologia*, 165:771–781, 2011.

[31] 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(678448):2284, 2021. Publisher: Frontiers.

[32] 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(44):eabg8531, 2021.

[33] 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(26):12804–12809, 2019. Publisher: National Acad Sciences.

[34] N. E. Korres and A. S. Nizami. Variation in anaerobic digestion: need for process monitoring. In *Bioenergy production by anaerobic digestion*, pages 194–230. Routledge, 2013.

[35] 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(1):17668, 2017.

[36] K. K. Lee, S. Liu, K. Crocker, J. Wang, D. R. Huggins, M. Tikhonov, M. Mani, and S. Kuehn. Functional regimes define soil microbiome response to environmental change. *Nature*, Jul 2025.

[37] 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):0015, 2017. Publisher: Nature Publishing Group.

[38] S. Louca, L. W. Parfrey, and M. Doebeli. Decoupling function and taxonomy in the global ocean microbiome. *Science*, 353(6305):1272–1277, 2016. Publisher: American Association for the Advancement of Science.

[39] R. Marsland, W. Cui, and P. Mehta. A minimal model for microbial biodiversity can reproduce experimentally observed ecological patterns. *Scientific reports*, 10(3308):1–17, 2020. Publisher: Nature Publishing Group.

[40] R. Marsland III, W. Cui, J. Goldford, A. Sanchez, K. Korolev, and P. Mehta. Available energy fluxes drive a transition in the diversity, stability, and functional structure of microbial communities. *PLoS computational biology*, 15(2):e1006793, 2019. Publisher: Public Library of Science San Francisco, CA USA.

[41] R. Marsland III, W. Cui, and P. Mehta. A minimal model for microbial biodiversity can reproduce experimentally observed ecological patterns. *Scientific reports*, 10(1):3308, 2020.

[42] J. McGrady-Steed, P. M. Harris, and P. J. Morin. Biodiversity regulates ecosystem predictability. *Nature*, 390(6656):162–165, 1997.

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

[44] 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(4):450–456, 2008.

[45] S. Mirarab, N. Nguyen, and T. Warnow. Sepp: Saté-enabled phylogenetic placement. In *Biocomputing 2012*, pages 247–258. World Scientific, 2012.

[46] J. Moran, L. Graham, and M. Tikhonov. Scripts and data for: Emergent predictability in microbial ecosystems. *Mendeley Data*, V1:10.17632/gxcvv6r5f7.1, 2025.

[47] J. Moran and M. Tikhonov. Defining Coarse-Grainability in a Model of Structured Microbial Ecosystems. *Physical Review X*, 12(2):021038, May 2022. Publisher: American Physical So-

ciety.

[48] S. Naeem and S. Li. Biodiversity enhances ecosystem reliability. *Nature*, 390(6659):507–509, 1997.

[49] 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*, 70:51–57, 2014.

[50] A. V. Narla, T. Hwa, and A. Murugan. Dynamic coexistence driven by physiological transitions in microbial communities. *Proceedings of the National Academy of Sciences*, 122(16):e2405527122, 2025.

[51] J. P. O'Dwyer. Beyond an ecological ideal gas law. *Nature Ecology & Evolution*, 4(1):14–15, 2020.

[52] 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(27):8356–8361, 2015. Publisher: National Acad Sciences.

[53] R. K. Peet. The measurement of species diversity. *Annual review of ecology and systematics*, 5(1):285–307, 1974.

[54] G. Plata, K. Srinivasan, M. Krishnamurthy, L. Herron, and P. Dixit. Designing host-associated microbiomes using the consumer/resource model. *MSystems*, 10(1):e01068–24, 2025.

[55] S. Pollak, M. Gralka, Y. Sato, J. Schwartzman, L. Lu, and O. X. Cordero. Public good exploitation in natural bacterioplankton communities. *Science Advances*, 7(31):eabi4717, July 2021. Publisher: American Association for the Advancement of Science.

[56] S. Pontrelli, R. Szabo, S. Pollak, J. Schwartzman, D. Ledezma-Tejeida, O. X. Cordero, and U. Sauer. Metabolic cross-feeding structures the assembly of polysaccharide degrading communities. *Science advances*, 8(8):eabk3076, 2022.

[57] C. Rao, K. Z. Coyte, W. Bainter, R. S. Geha, C. R. Martin, and S. Rakoff-Nahoum. Multi-kingdom ecological drivers of microbiota assembly in preterm infants. *Nature*, 591(7851):633–638, Mar 2021.

[58] C. Ratzke, J. Denk, and J. Gore. Ecological suicide in microbes. *Nature ecology & evolution*, 2(5):867–872, 2018.

[59] C. Ratzke and J. Gore. Modifying and reacting to the environmental ph can drive bacterial interactions. *PLoS biology*, 16(3):e2004248, 2018.

[60] P. B. Reich. Elevated CO2 reduces losses of plant diversity caused by nitrogen deposition. *Science*, 326(5958):1399–1402, 2009.

[61] 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(6830):809–810, 2001.

[62] 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(27):10101–10106, 2004.

[63] 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(4):419–429, 2005.

[64] M. Shahin, B. Ji, and P. D. Dixit. Embed: Essential microbiome dynamics, a dimensionality reduction approach for longitudinal microbiome studies. *npj Systems Biology and Applications*, 9(1):26, Jun 2023.

[65] X. Shan, A. Goyal, R. Gregor, and O. X. Cordero. Annotation-free discovery of functional groups in microbial communities. *Nature Ecology & Evolution*, 7(5):716–724, 2023.

[66] M. R. Silverstein, J. M. Bhatnagar, and D. Segrè. Metabolic complexity drives divergence in microbial communities. *bioRxiv*, pages 2023–08, 2023.

[67] 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*, pages 1–11, 2023.

[68] 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(1):37–63, 2005.

[69] M. Tikhonov and R. Monasson. Collective phase in resource competition in a highly diverse ecosystem. *Physical review letters*, 118(4):048103, 2017. Publisher: APS.

[70] J. Van Ruijven and F. Berendse. Positive effects of plant species diversity on productivity in the absence of legumes. *Ecology Letters*, 6(3):170–175, 2003.

[71] J. C. Vila, J. Goldford, S. Estrela, D. Bajic, A. Sanchez-Gorostiaga, A. Damian-Serrano, N. Lu, R. Marsland III, M. Rebolleda-Gomez, P. Mehta, et al. Metabolic similarity and the predictability of microbial community assembly. *bioRxiv*, 2023.

[72] C. S. Ward, C.-M. Yung, K. M. Davis, S. K. Blinebry, T. C. Williams, Z. I. Johnson, and D. E. Hunt. Annual community patterns are driven by seasonal switching between closely related marine bacteria. *The ISME journal*, 11(6):1412–1422, 2017.

[73] 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(3):930–931, 2010.

[74] Y. Zhao, O. X. Cordero, and M. Tikhonov. Linear-regression-based algorithms can succeed at identifying microbial functional groups despite the nonlinearity of ecological function. *PLoS computational biology*, 20(11):e1012590, 2024.

MATERIALS AND METHODS

Data selection and processing

Our criteria of data selection, and how they apply to available studies, are presented in Sup. Table S1. These criteria selected two datasets for our analysis, Ref. [11, 33]. The raw data we used is published alongside the original studies, and/or is available on Mendeley Data [46]. Data from Kehe *et al.* were generated using microwell arrays for random community assembly using nanoliter droplets of defined soil isolates, minimal media, and sucrose. Community culturing experiments of Clark *et al.* were performed in an anaerobic chamber with defined isolates assembled in a chemically defined medium to support the growth of diverse gut species. Further experimental details and data preprocessing steps are described in the corresponding publications [11, 33].

For both datasets, we resolve three tiers of community richness (with the high-richness tier, $S > 20$ strains, only available for Clark *et al.* [11]). Since Clark *et al.* did not directly measure absolute strain abundances in a community (only its total OD and the compositional makeup assessed by 16S sequencing), we adopted an approach where the relative abundances estimated from the 16S data were first scaled by community OD (of a given sample), and then standardized to zero mean and unit variance (across samples) 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 Supplementary Text, section B 2 b, and Fig. S2 and for further discussion.

Model-derived synthetic community properties

When using simulated data, our analysis employed models of two kinds: (1) Effective models that adopted a simple linear-contribution ansatz to map community composition directly into a synthetic "community property," and (2) mechanistically-informed dynamical models.

Effective models

These models were employed as a simple setting to establish the *minimal* conditions required for the emergence of each of the flavors of emergent simplicity we defined. Accordingly, in all cases, the synthetic property $\tilde{Y}_\mu$ was defined by a simple linear-contributions model:

$$\tilde{Y}_\mu \equiv \sum_i \tilde{\beta}_i \hat{n}_{i\mu} + \epsilon_\mu. \tag{1}$$

Here, $\hat{n}$ denotes standardized abundances (unless specified otherwise, we used the abundance table measured in Clark *et. al.*). The term ϵ_μ is Gaussian random noise with mean 0; its variance directly sets the FVU of the best predictive model. Finally, the weights $\tilde{\beta}_i$ determine the strength of species' contributions to this synthetic property $\tilde{Y}$. By construction, with sufficient data, the best microscopic model fit to such a synthetic dataset would be the model with coefficients β_i. The different models discussed below differ primarily by the choice of these ground-truth $\tilde{\beta}_i$ (random and uncorrelated; correlated with phylogenetic relationships; or derived from a realistic trait table constructed using PICRUSt2).

a. Null Model for Self-averaging: For Fig. 2B analysis, we construct the simplest null model by drawing $\tilde{\beta}_i$ in Eq. (1) from a unit uniform distribution. For species abundance inputs we use either randomly shuffled abundance data from Clark *et al.* or Kehe *et al.*, or random values drawn from a Gaussian with mean and variance both set to 1; the three choices correspond to the three lines shown in Fig. 2B. Note that in Fig. 2B, our goal is to show that self-averaging can arise with no biological structure whatsoever, which is why we *shuffle* the empirical abundances (or just use Gaussian-random values instead). Specifically for the simulations shown in Fig. 2B, the ϵ_μ term was set to zero.

b. Null Model for Quantifying Coarse-grainability: In order to quantify whether the Pareto front constructed for some property Y descends "faster than expected", we generate a synthetic community property $\tilde{Y}$ whose Pareto front can serve as a null-model comparison. To make the comparison conservative, $\tilde{Y}$ should be constructed based on the same abundance table (no shuffling!). To make the comparison fair, $\tilde{Y}$ must satisfy two constraints: its Pareto front should start at FVU ≈ 1, as is the case for all our empirical properties ("constraint 1"); and the FVU attained by the most predictive model should match that of the reference property Y ("constraint 2").

To this end, the weights $\tilde{\beta}_i$ in Eq. (1) were drawn from a normal distribution with zero mean, and variance matching the spread in the regression coefficients β_i of the "microscopic" model constructed for the reference property Y. Centering $\tilde{\beta}_i$ at zero ensures that the biomass-only description has no predictive power, satisfying constraint 1. Then, the variance of the added Gaussian noise ϵ_μ was adjusted for $\tilde{Y}$ to satisfy constraint 2. (Practically speaking, we adjusted its variance so that the *third* lowest point of the Pareto front of $\tilde{Y}$ matches the third lowest point of the Pareto front of Y. This ensures robustness to the stochasticity inherent to estimating out-of-sample FVU by cross-validation.)

This protocol ensures that the two Pareto fronts (for Y and $\tilde{Y}$) have matching endpoints, and allows us to ask whether the speed of the Pareto front descent observed for Y (its "coarse-grainability") is noteworthy, or is no faster than can be reproduced by the simplest null model described.

c. Phylogenetically Conserved Contributions: In Fig. 3B, C, we sought to demonstrate that even the simplest models implementing the intuition that "some species are more similar than others" can exhibit coarse-grainability. For this, rather than choosing $\tilde{\beta}_i$ in Eq. (1) to be independent and uncorrelated, we drew them in a phylogenetically-informed way. Specifically, we first used the phylogenetic tree of Clark *et al.* species (as presented in the original reference Ref. [11]) to compute the pairwise patristic distance matrix D_{ij}, normalized to set the largest distance to 1. Then, the contribution weights $\tilde{\beta}_i$ were drawn from a multivariate

Gaussian distribution with zero mean and the correlation matrix

$$C_{ij} = \frac{1}{1 + \exp(\Delta_d(D_{ij} - d))}.$$

This sigmoid functional form was chosen to implement the notion of a finite conservation depth: the contributions $\tilde{\beta}_i$ of closely related species are highly correlated, but this correlation falls off sharply beyond a distance threshold d (the steepness parameter is set to a large value $\Delta_d = 50$). The standardized abundance matrix $\hat{n}_{i\mu}$ we used in Eq. (1) was based on Clark *et al.* measured abundances, combined across all richness tiers to have the most data (as a reminder, coarse-grainability is defined without reference to changing richness). The variance of ϵ_μ was adjusted to ensure that the best-performing model on the Pareto front explained 70% of simulated data (matching the typical value attained by the empirical Pareto fronts at high richness).

The distance threshold d is the key parameter varied in the analysis of Fig. 3C. By computing the typical distance between pairs of leaf nodes (Clark *et al.* species) that fall within a given taxonomic level, we can obtain values for the depth parameter d that correspond to genus- or phylum-level conservation of trait correlations. This allowed us to label the curves of Fig. 3C in an interpretable way (e.g., "phylum based conservation"), rather than by the corresponding value of d.

d. Realistic contribution structure (PICRUSt-based model): Finally, we wanted to consider an example with a more realistic structure. To this end, we constructed another model where instead of drawing the weights $\tilde{\beta}_i$ from some distribution, we derive them from an actual matrix of predicted organism traits.

Specifically, we first generated a phylogenetic tree based on the 16S sequences in Clark *et al.* using SEPP [45]. Next, we applied PICRUSt2 [17] to generate a predicted presence/absence table of KEGG orthologs (KOs) belonging to the Carbohydrate Metabolism category (chosen as one general category relevant for the fermentation end-products assayed in the Clark *et al.* study). Any KOs that were present in no species, or in all of them, were discarded. This procedure mapped each species into a 123-dimensional binary "KO profile" $\vec{P}_i$. Then, for each species i we computed the dot product w_i of its KO profile with a fixed vector $\vec{v}$: $w_i = \vec{P}_i \cdot \vec{V}$. For simplicity, we picked $\vec{V} = (1, 1, \ldots 1)$. While this mapping is, of course, still simplistic, it inherits some of the complexity of the PICRUSt-generated trait table, and reflects the fact that relevant organism traits do not typically possess a single scale of conservation depth. Finally, the weights $\tilde{\beta}_i$ were constructed by shifting the distribution of w_i to have zero mean: $\tilde{\beta}_i = w_i - \langle w_i \rangle$. Without this shift, the community biomass alone would already be highly predictive of the property, which would mask the effect we seek to describe. (See important additional discussion in the Supplementary Text Section B 2 a).

Dynamical models

Beyond the effective models described above, our analysis also used three mechanistically-informed dynamical models: two representatives of standard modeling frameworks (GLV and CRM) with "community invadability" as the property of interest, used in Fig. 3&4; and a physiological feedback model used in Fig. 5.

e. Generalized Lotka-Volterra model with random interactions: 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} [5]. 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. [5, 9], 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.

f. Consumer-Resource model with a random consumption matrix: The CRM model describes the abundance dynamics of strains competing for a set of resources. We adopt the following parameterization (identical to Ref. [47] with slightly adjusted notation):

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

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, the relative competitiveness of strains is reflected in their "maintenance cost" χ_i; we follow Ref. [69] to adopt $\chi_i = c + \sum_\alpha \lambda\sigma_{i\alpha} + \delta_i$. As in ref. [47], 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, δ_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} is chosen to mimic the large typical size of microbial populations.) Random contributions to each strain's cost δ_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).

g. Simulated invasion experiments: In both the GLV and CR models, we use the same protocol for our computational experiments, which follows the scheme outlined in Fig. 1A. 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.

h. Physiological feedback model: We take as inspiration the model presented in Refs. [2, 50, 59] of pH-mediated interactions between microbes, and we generalize it to a consumer-resource context. Each species i is characterized by its strength c_i in modifying an environmental variable p and by its preferred environmental variable ρ_i, which in nature would be set by physiological constraints. Species in a community interact by competing for a common nutrient source r and by modifying the environmental variable. The growth rate of species i is expressed as

$$a_i = r\,\max(f(p-\rho_i) - \epsilon, 0),$$

where $f(p-\rho_i) \equiv e^{-\frac{(p-\rho_i)^2}{2\sigma^2}}$ sets the range σ of growth around the preferred environmental pH and ϵ sets a viability threshold outside the preferred environmental variable. The resulting behavior is that as the environmental variable moves further away from a species' preferred state, that species begins to consume and grow less, until eventually it stops consuming and growing entirely. The abundance and environment dynamics are expressed as follows:

$$\begin{aligned}\dot{n}_i &= a_i n_i \\ \dot{r} &= -\sum_i a_i n_i = -\sum_i \dot{n}_i \\ \dot{p} &= \sum_i a_i n_i c_i = \sum_i c_i \dot{n}_i\end{aligned}$$

Note that this is a batch culture model, rather than a steady-state chemostat model. Using a batch culture model allows us to meaningfully tune community richness with all initial species contributing to the dynamics, in contrast to a chemostat version which would only permit at most 2 species surviving at steady state, regardless of their initial number.

Finally, to incorporate some community-level property Y we wish to predict using the final species abundances, species are assigned a weight y_i describing the effect of their metabolism on said property. We chose the dynamics of property Y to be similar to the environmental variable:

$$\dot{Y} = \sum_i a_i n_i y_i = \sum_i y_i \dot{n}_i.$$

The simulations of Fig. 5 used the following parameters: c_i and y_i both drawn independently from a Gaussian with mean 0 and unit variance; ρ_i drawn from a Gaussian with mean 0 and standard deviation 5; and ϵ set to 0.0111, equivalent to species arresting growth when the environmental variable is beyond ± 3 of their preference when feedback strength is 1. Each species was initialized with 0.05 abundance, the initial value of r was fixed at 10. A fixed pool of 25 species was used in each simulation, with richness tiers defined as for all other models and empirical datasets, and 800 unique random communities assembled per richness tier. Once the simulation reaches a steady-state (when the community resource is depleted, or the environmental variable too extreme for any species to grow), the steady state abundances are used as predictors of the community property of interest.

Constructing & manipulating the Pareto fronts

The Metropolis-Hasting algorithm

The Pareto front of coarsenings Ψ consists of those with minimal prediction error for a given complexity $I(\Psi)$. An exhaustive search through the space of all possible coarse-grainings is computationally infeasible unless the library of strains is very small: the total number of ways to partition S strains (the Bell number B_S) grows (super-)exponentially with S, and for $S = 25$ taxa as used in the Clark *et al.* study, $B_{25} \approx 5 \times 10^{18}$. This necessitates developing algorithms to approximate the Pareto front via a heuristic search.

Here, we leverage the approach recently described in [74], adopting a variant of the Metropolis-Hastings algorithm to search the space of possible coarsenings and estimate the Pareto front by retaining the most predictive ones. Briefly, the algorithm initializes a random initial guess for the Pareto

front by drawing a random set of candidate coarsened groupings and evaluating the description complexity and prediction error of each grouping. Then, the algorithm main loop: (1) constructs a new candidate grouping by randomly splitting or merging a randomly selected grouping from the current set, (2) evaluates the information-prediction attributes of the new candidate, and (3) if the new grouping improves on the prediction error of the model currently memorized in the respective complexity bin, the memorized model is overwritten by the new candidate. Otherwise, the replacement is probabilistic with a Boltzmann probability set by the difference in their prediction errors. The main loop is repeated 10^5 times for each of 15 random initializations, and the top-performing groupings are aggregated into the final "best guess" of the Pareto front models. The performance of these best-guess groupings are then cross-validated to obtain a distribution of out-of-sample FVU measurements (see Supplementary Text, section B 4). Finally, we perform the following convergence check: (a) for each coarsening Ψ making up the best Pareto guess, generate all neighboring coarsenings that successively re-assign each single strain to all other classes in Ψ, (b) identify and include any relabeled coarsenings that result in a median FVU that fall below the convex hull spanned by the lower quartiles in FVU of the current Pareto, (c) repeat steps (a) and (b) 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. The output returned by the algorithm consists of a set of coarsenings $\{\Psi\}$, and for each, 100 estimates of out-of-sample FVU, computed by cross-validation over 50-50 splits. The error bars shown in panels 4A,D-G and 5C are the quartiles of these respective distributions.

Additional details on our methodology of constructing and testing the Pareto fronts for robustness and artifacts, and the actual computed Pareto fronts used for all analyses in the main text can be found in the Supplementary Text section B.

Aggregating the Pareto fronts

When we consider synthetic properties where species contributions $\tilde{\beta}_i$ are drawn from a distribution (i.e., the coarse-grainability null model, PICRUSt2 model, and phylogenetic-based model), each random sample of these weights defines its own instance of the synthetic property $\tilde{Y}$, and thus a distinct Pareto front. To summarize this ensemble, we aggregate it into a single Pareto front (with larger error bars), using the following procedure.

As mentioned above (and detailed in Supplementary Text, Sec. B 4), each Pareto front consists of a set of coarsened models identified by the algorithm, and for each such model (a grouping Ψ), its performance is reported as a set of 100 FVU values computed via cross-validation over 100 folds of splitting the data into training and testing sets. When aggregating an ensemble of Pareto fronts, these cross-validation sets are collated together and binned along the axis of description complexity $I(\Psi)$. We then compute the median FVU within each complexity bin; this is the aggregated Pareto front, which is now defined by its vertices at bin center locations, but which we interpolate into a piecewise linear curve. To report an error bar, we similarly aggregate and compute a running average of the magnitude of deviations (relative to this *interpolated* aggregated curve), and report the quartiles of the distribution of these deviations.

Comparing the Pareto fronts

When assessing coarse-grainability (Fig. 3), our task is to compare an actual Pareto front (constructed for some property Y, either empirical or model-derived), and the (ensemble of) null-model comparisons $\tilde{Y}$. To perform this comparison, we first aggregate this ensemble into a single null-model Pareto front as described above. Now, we have two Pareto fronts (with error bars), and our task is to construct a difference curve (also with error bars). To do so, we first interpolate both the Pareto fronts and the error bars, so they are defined at same values of the complexity axis. Then, the difference is trivially computed by subtracting the interpolated curves. As for the error bars, the uncertainties in estimating both curves are independent, and we combine them in quadrature: $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$.

Area between the Pareto fronts

In Fig. 4H, we quantify emergent predictability by computing the area between the high- and low-richness Pareto fronts. In order to compute the uncertainty of its estimation, we compute 100 areas between 100 noisy samples of each of the Pareto fronts (sampling from the 100 FVU values, reported for each model), and record the median and the quartiles of this distribution. Fig. 5D used the same approach, but the areas were additionally aggregated over 10 random instances of the model.

It is important to emphasize that this area-based metric is purely heuristic (note that the area is measured on a log plot, with an arbitrary x-axis cutoff of 3 bits of entropy). This approach is helpful for presentation purposes, affording the simplicity of summarizing several comparisons in a single panel (Fig. 4H). The most direct evidence is to compare the full profiles (as we do for coarse-grainability in Fig. 3); this analysis is presented in Supplementary Text, section B 7.

SUPPLEMENTARY TEXT

Appendix A: Data selection: Other empirical datasets available

Scanning the literature for empirical studies that assemble communities from a fixed library of strains, we identified 9 candidates for further consideration. These are listed in Table S1, and identify the reasons why only the datasets of Refs. [11] and [33] 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.

Appendix B: Constructing the Pareto fronts – Additional details

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. 4 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 being 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.

2. Normalization choices

a. Normalizing the response variable (the community-level property Y)

The question of quantifying the complexity of description, discussed above, is closely tied to the choice of the property of interest, and in particular, its normalization. The easiest way to see this is to observe that with our definition of $I(\Psi)$, the one-group description Ψ_B (all taxa grouped together into a single category) has complexity $I(\Psi_B) = 0$. Technically, this is not a zero-information description: if we think in terms of raw abundances of taxa, summing them is essentially the total community biomass, which can be informative. However, in many situations (say, if studying process rates in natural soil), properties of interest scale with biomass, and the quantity most natural to predict is the value normalized *per unit biomass*. In such cases, our choice of $I(\Psi)$ is particularly appropriate. In contrast, in lab experiments where an inoculum is provided a fixed amount of resources, normalizing to biomass may not be necessary or desired, since properties of interest are meaningfully defined in the absolute sense — in fact, the final biomass of the assembled community could itself be a property of interest.

Fortuitously, in the empirical datasets used here, the one-group coarsening (to which we colloquially refer as the "biomass-only description") explained almost no variance (never exceeding 15% across all richness tiers and measured properties). This simplified our presentation, allowing us to declare that we were primarily focused on scenarios where the one-group coarsening explained no variance, such that we could equate $I(\Psi_B) = 0$ with "having no compositional information." In the interest of conciseness, in the main text it was easiest to frame this as a decision to limit consideration to only a subset of community

properties – those for which this statement is true. However, now is a good time to clarify that properties for which biomass alone explains a significant fraction of the variance are also compatible with our approach / framework. The generalization is straightforward: instead of the fraction of *total* variance explained, one should consider the residual variance after regressing on the biomass. With this approach, the Pareto fronts will start at $\mathrm{FVU} = 1$ *by definition*, which would in fact simplify some downstream analyses. However, defining the y axis of our diagrams in this way would make them appear rather less intuitive on first encounter. To summarize, limiting consideration to properties that are not explainable by biomass is a choice made for ease of presentation, not a limitation of our framework.

One minor consequence of this choice was for the computational invasion experiments (GLV & CRM), where making this analysis comparable with our empirical curves required us to exclude the instances ($\approx 15\%$ of cases) where the combined standardized abundances could explain $>10\%$ of total variance in simulated "community invadability" data at high richness. Had this variance been defined as the residual variance after regressing on the biomass-only prediction, applying such exclusion criteria would not have been necessary. However, the advantage offered by greater simplicity of presentation of our key concepts outweighed this consideration.

b. Normalizing the predictor variables

In our analysis, the input variables supplied as "abundance matrix" n were first standardized to zero mean and unit variance. All coarsening schemes were then applied to these standardized variables.

To understand this choice, it is important to note some limitations of the datasets. The study of Kehe *et al.* [33] did not directly measure the full composition of the final assembled communities: only the abundance of the focal strain *H. Frisingense* was tracked (in arbitrary fluorescence units). The available compositional information is limited to the makeup (binary presence/absence) of the set of droplets combined by the authors' microfluidic platform to form the community initially (each droplet carrying individuals of a given strain). These binary variables are proxies for the *initial* abundances of community members, and in this sense, they still satisfy the criterion of Fig. 1 of "following a consistent assembly protocol from a known compositional state." However, the relationship between this binary n matrix and the actual abundance values is only approximate, and includes unknown proportionality factors.

In the study of Clark *et al.* [11], the composition of final assembled communities was profiled more directly, using 16S sequencing (recording relative abundances) as well as recording the OD_{600} for each community. However, even here, the best estimate of the absolute abundances (multiplying the relative fractions of 16S counts by the recorded community OD) is still only a proxy. The proportionality factor includes the number of 16S copies on the genome, which differs between species, and a PCR amplification bias, which is unknown and often substantial [35, 43]. Because of this, we adopted an approach where the input abundances were standardized to zero mean and unit variance (across samples) before training any regression models. For consistency, the same standardization procedure was also applied to Kehe *et al.* data.

It is important to note that summing standardized abundances is not equivalent to standardizing summed abundances. However, even if absolute abundances were known, it is not clear if the biologically meaningful summation should be best implemented 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 omitting the standardization step retains our conclusion (Fig. S2).

This standardization of abundances 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 predictability. Interpreting the high-performance groupings is left for future work.

3. Convergence of the Pareto search algorithm

We found the convergence of our Metropolis algorithm to be most efficient if the search is performed leveraging all of the data, and evaluates a candidate grouping using its *in-sample* explanatory power. Indeed, if using out-of-sample performance, one must use a very large number of cross-validation folds. Otherwise, the stochasticity of cross-validation (the variability induced by drawing different subsamplings of data) can mask true differences in performance, especially when applied to relatively small datasets like here, so that the decision whether to accept a given Metropolis move can become effectively random, impeding convergence. This large number of cross-validation folds sampled at each evaluation step would increase the runtime by orders of magnitude. Thus, in our approach, during the search phase only, performance of a candidate grouping is assessed using the

in-sample FVU. After convergence, the performance of each grouping identified as belonging to the Pareto front is reassessed using out-of-sample cross-validation. This methodology was previously shown to ensure good convergence on synthetic datasets with known ground truth [74], and performed well in our case as well (see section B 5, "Robustness tests").

4. 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 μ labels communities, $\hat{Y}_\mu$ is the model's prediction of property Y, and β_i^Ψ are the regression coefficients.

To assess the predictive power of a particular coarsened description Ψ, the data is randomly split into training and testing subsets. We note that the experimental design of both studies includes replicate communities (i.e., communities with same presence-absence combination $b_{i\mu}$ defined in section I A). We take care to keep such replicates together, such that 50% of *unique* compositions are randomly designated for model training, and the other 50% are reserved for evaluating the out-of-sample prediction error.

Model fitting was performed using the standard ordinary least-squares method of MATLAB's 'fitlm' algorithm. We then used the test set to compare model predictions to the measured values. The Fraction of Variance Unexplained (FVU) was computed in the standard way of dividing mean-squared error (MSE) of prediction by the variance of Y across the test set samples. This was repeated for 100 random 50-50 splits of the data; the plots report the median value, and the error bars specify the upper and lower quartiles across these splits.

5. Robustness tests

To verify that the observation of emergent predictability is robust, we perform two tests on the low-richness datasets from [11] and [33] 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. S3. To test robustness to sample size, we down-sampled the low-richness communities to match the total number of strain observations across unique community combinations in the high-richness dataset. This ensures each strain is sampled for regression model fitting. We performed 5 random subsamplings of the low-richness dataset and computed the Pareto fronts for each down-sample. We then comparing these curves to the reference high-richness Pareto fronts, to check whether the subsampling might have closed the gap between them. The results are presented in Fig. S3.

For all four observables of Clark *et al.*, subsampling the low-richness data increases stochasticity, but does not significantly shift the Pareto front, confirming that the gap between the high- and low-richness Pareto fronts cannot be attributed to differences in dataset size.

For Kehe *et al.* (Focal strain abundance), we see that the Pareto fronts computed for the downsampled data vary tremendously between subsampling replicates. To understand this, we note that the extent of subsampling is much more dramatic: the 19433 samples of the low-richness tier are downsampled to only 346 (to match the mid-richness tier); compare to 870 samples (low-richness tier) downsampled to 180 (matching the high-richness tier) for the case of Clark et al. This is compounded with the fact that for Kehe et al., the mid-richness tier is the highest available. As a result, for higher-complexity descriptions the gap either disappears, or is dwarfed by the stochasticity. We note, however, that at low-complexity end, the gap persists (see panel Fig. S3F, which plots the distribution of the FVU values along the transect at 0.35 bits), confirming that emergent predictability is supported even after subsampling.

To formally confirm this, we performed Wilcoxon rank-sum tests between the low- and mid-richness FVU distributions at a given description complexity. Panel Fig. S3G plots the p-value of the comparison, and indicates that the high-richness Pareto front remains statistically separable from the ensemble of subsampled low-richness Pareto fronts up to ~ 2 bits.

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. S4 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.

6. Testing for artifacts

Checking that the emergent predictability observed in Figure 4 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 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 to the datasets from [11] and [33], we aim to keep the abundance statistics of each species across communities fixed. For each strain in Ref.[11], we randomly permute its abundance in communities in which the species is initially present. Because Ref.[33] only reports presence-absence of strains, we must instead shuffle by swapping the presence of each strain across a random subset of all possible community samples (i.e., not preserving presence-absence structure) while maintaining the total number of communities in which a given strain is sampled. After doing so, we see that, as expected, the emergent predictability in high-richness communities disappears (Fig. S6), as indicated by an increase in prediction error of the high-richness Pareto front obtained for the permuted abundance data. Emergent predictability is similarly broken when we instead permute the measured properties in high-richness communities, shown in Fig. S7. Both tests confirm that the shift of the Pareto front with increasing community richness is not an artifact of the analysis, and that the observed emergent predictability is significant.

Finally, in addition to measuring fermentation end products, Clark *et al.* measured the total optical density (OD) of their communities to compute strain abundances. For completeness, Fig. S5 presents the result of applying our framework to the OD measurements. We should emphasize, however, that community OD differs from all other community properties used in this work, since technically, in this case the response variable and the predictor variables are *not* technically independent. (Recall that for Clark et al, we use OD measurements for scaling the relative abundances profiled by 16S.) In fact, the only reason this analysis is non-trivial is because, as described above, we standardize the abundance data prior to performing the regressions (without this standardization, the total OD would trivially be a sum of input variables, and be perfectly predictable by *any* coarsened representation). Thus, interpreting Fig. S5 requires extreme caution, and we present it for completeness only. Perhaps the only possible conclusion to draw here is that "not any set of values would be flagged by our framework as exhibiting emergent predictability," reinforcing the point already demonstrated by the permutation tests above.

7. The raw Pareto fronts used in all comparisons

Here, for more direct comparison, we show all the raw Pareto fronts found by our Metropolis search algorithm (see Methods) that were averaged for summary comparisons in the main text. Fig. S8 shows the actual Pareto fronts versus our Gaussian null for quantifying coarse-grainability in Fig. 3. Fig. S9 plots the Pareto fronts for the high- and low-richness communities for models that do not exhibit emergent predictability in Fig. 4. Fig. S10 displays the high- and low-richness Pareto fronts from the physiological feedback model, showing the 10 replicates for 3 feedback strengths that result in varying levels of emergent predictability in Fig. 5C.

Appendix C: Other topics

1. Structured species pool is not sufficient for emergent predictability

In the main text, we explained that in any real-world setting, (at least some amount of) coarse-grainability is always expected to arise simply because species pool has structure: some species are more similar than others.

However, our proposed argument for the origin of emergent predictability may superficially appear similar (particularly when the model-derived results are rephrased in intuitive general terms in the Discussion). Consider a simple scenario with a highly structured pool of species forming two well-defined classes: A and B. Species of the same class are very similar, but A and B are distinct in important ways. Then, the following word-based argument appears plausible: "As the number of species increases, both classes become represented by many members; thus, *within* a class, the quirks of precisely which representatives were selected will average out. In contrast, the difference *between* the two classes will always persist. Thus, as richness increases, one expects the two-group description, A and B, to become increasingly predictive." This intuition appears to suggest that, just like coarse-grainability, emergent predictability can also arise simply as a consequence of species pool structure, contrary to our observations in the main text.

It is therefore highly instructive to convert this word-based argument into a simple quantitative model. Observing why emergent predictability is not, in fact, automatically observed in this case will not only correct a possible misconception, but will also help clarify the mechanism we propose in Fig. 5.

As the simplest model with such a two-class species pool, we will, once again, adopt a simple ansatz where species are

characterized by their contributions β_i to community property Y:

$$Y_\mu = \sum_i \beta_i n_{i\mu}, \text{ where } \beta_i = \begin{cases} \beta_A + \epsilon_i, & i \in A \\ \beta_B + \epsilon_i, & i \in B. \end{cases}$$

In other words, the contribution is defined primarily by the class identity (β_A vs. β_B), while ϵ_i (the "quirks") represent the small deviations of individual strains, and are assumed small. We will take strain $n_{i\mu}$ and ϵ_i to be independent and Gaussian distributed: $n_{i\mu} \sim \mathcal{N}(m_n, \sigma_n^2)$ and $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$, with $\sigma_\epsilon^2 \ll (\beta_A - \beta_B)^2$.

The bimodal distribution of coefficients β_i means that Y will exhibit strong coarse-grainability: the natural two-group description: $n_A = \sum_{i\in A} n_i$ and $n_B = \sum_{i\in B} n_i$ has a much lower error than could be expected by chance. We will now consider whether Y also exhibits emergent predictability. For this, we need to compute the fraction of variance of Y explained by the two-group description, and evaluate its dependence on community richness $S = S_A + S_B$. Here S_A and S_B denote the the number of strains in the two groups. For simplicity, we will assume that $S_A = S_B = S/2$, though this can of course be generalized.

Note that the parameter m_n can itself depend on S – for example, if community is assembled under a common resource limitation, one may expect $m_n \sim 1/S$; similarly, σ_n may or may not scale with S as well. For our argument below, we will only need to consider their *relative* scaling. The ratio σ_n/m_n represents the "noisiness" of each individual species abundance. It would be extremely unusual to see this ratio decrease with S. Indeed, if some mechanism ensured that the abundance of each individual species becomes increasingly predictable with diversity, this would directly lead to an extremely strong form of emergent predictability, whereby not only the macroscopic community-level properties, but even the microscopic-level description is increasingly predictable. No empirical indications support such a strong form. Thus, in general we expect $\sigma_n/m_n \sim S^\alpha$ with $\alpha \geq 0$. Our argument below will not require any stronger assumptions; this general statement will be sufficient.

Computing the variance of the community property Y, we can separate the total variance into two contributions:

$$\mathrm{var}_\mu[Y_\mu] = \underbrace{\mathrm{var}\left[\sum_{i\in A} n_{i\mu}\beta_A + \sum_{i\in B} n_{i\mu}\beta_B\right]}_{\mathrm{var}_{\mathrm{coarse}}} + \underbrace{\mathrm{var}\left[\sum_i n_{i\mu}\epsilon_i\right]}_{\mathrm{var}_{\mathrm{quirks}}},$$

where the first term is precisely the variance explainable by the coarse-grained description, and the second term is the residual contribution from the individual "quirks." To test for emergent predictability in Y, we need to compute the fraction of the explainable variance relative to the total. Simplifying each term we have

$$\begin{aligned} \mathrm{var}_{\mathrm{coarse}} &= (S_A\beta_A^2 + S_B\beta_B^2)\sigma_n^2 = \frac{1}{2}S\sigma_n^2\left(\beta_A^2 + \beta_B^2\right) \\ \mathrm{var}_{\mathrm{quirks}} &= (S_A + S_B)\sigma_\epsilon^2(\sigma_n^2 + m_n^2) = S\sigma_n^2\sigma_\epsilon^2(1 + (\dots)S^{-2\alpha}) \simeq S\sigma_n^2 C\sigma_\epsilon^2 \end{aligned}$$

In the last line, we used the fact that $\alpha \geq 0$, and therefore the $S^{-2\alpha}$ term in the brackets is either negligible, or at most provides a constant correction of order 1. We conclude that in this example, both contributions to the variance scale in the exact same way with S, and the fraction of variance left unexplained by the two-group model remains constant:

$$\frac{\mathrm{var}_{\mathrm{quirks}}}{\mathrm{var}_{\mathrm{coarse}} + \mathrm{var}_{\mathrm{quirks}}} = \frac{2C\sigma_\epsilon^2}{\beta_A^2 + \beta_B^2 + 2C\sigma_\epsilon^2}.$$

Therefore, the observable Y exhibits no emergent predictability, despite being strongly coarse-grainable due to species pool structure. Intuitively, the individual strain "quirks" self-average with richness – but so does the variance explained by the coarse model, scaling in the exact same way. Thus, in the *absolute* sense, the two-class model is an increasingly good one, and this is the phenomenon of self-averaging (unexplained variance decreases). However, the unexplained *fraction* of variance remains constant (no emergent predictability).

Now, imagine if some mechanism were to maintain a finite range of variability along the A–B axis, such that some communities assembled into mostly A-type strains, while others assembled into mostly B-type strains. Then, as the ϵ_i average out, the variance explained by the class representation would indeed become an increasingly dominant fraction. This is what our Fig. 5 model achieves. In a pH-based intuition: the existence of two classes, acid lovers and acid haters, is not enough. But if there is a feedback that drives community pH in either direction based on initial composition, such that the community alternately assembles into a low-pH state dominated by acid-lovers, or a high-pH state dominated by acid haters – this community will display emergent predictability.

In the actual model we use in Fig. 5, note that turning the p-dependent feedback off makes the metabolic activity a_i identical for all species at all times; all species "grow as one" until the resource is depleted. Then, it is trivial to see that the contributions to both p and Y experience pure self-averaging, which affects both quantities in the exact same way. In this limit, the failure to observe emergent predictability has the same origin as in the model just described.

2. PCA on Clark *et al.* functional space

The experiments of Ref. [11] measured the final concentration of four fermentation byproducts in each sample community. We can view these measurements as points in a four-dimensional "functional space." In Fig. 5A, we perform principal component analysis on this functional property data and plot the projections along the first two principal axes for the measurements in low- versus high-richness communities, showing that at higher richness community function is spanned by a lower-dimensional subspace. Fig. S11 quantifies this by plotting the variance explained by each principal component. The plot demonstrates that as richness increases, the first principal axis explains an increasing fraction of total variance, and confirms the numbers quoted in the main text.

SUPPLEMENTARY FIGURES

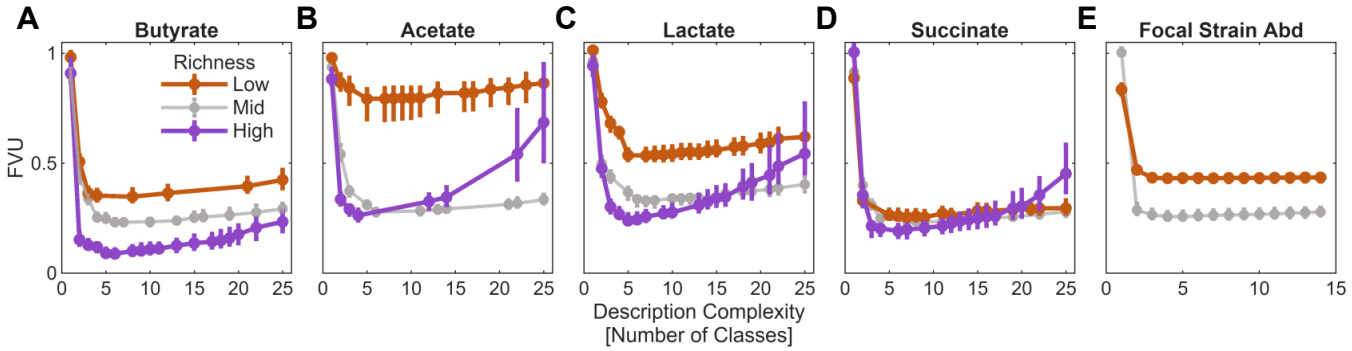

FIG. S1. Redoing the analysis shown in Fig. 4 with description complexity axis $I(\Psi)$ measured in terms of number of classes of a given coarsened description Ψ.

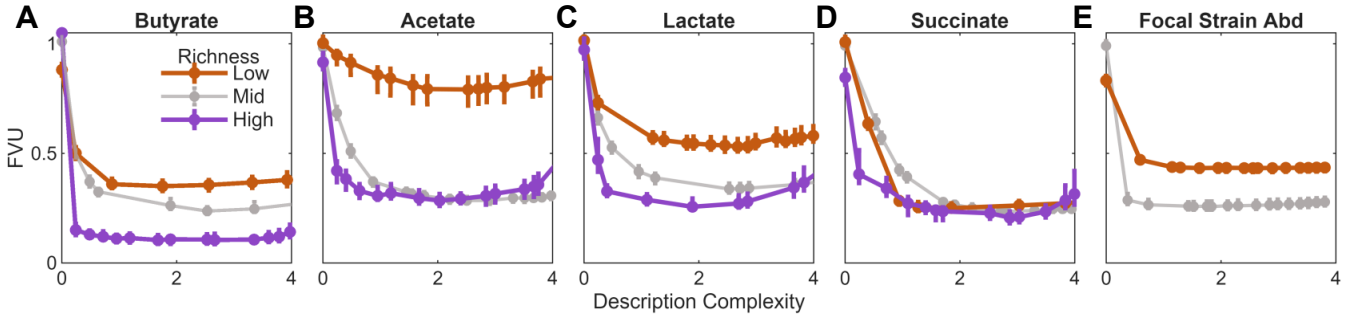

FIG. S2. Reanalyzing result of Fig. 4 with raw input variables (non-standardized) used in regression models. That is, abundance data is not standardized to 0 mean and unit variance.

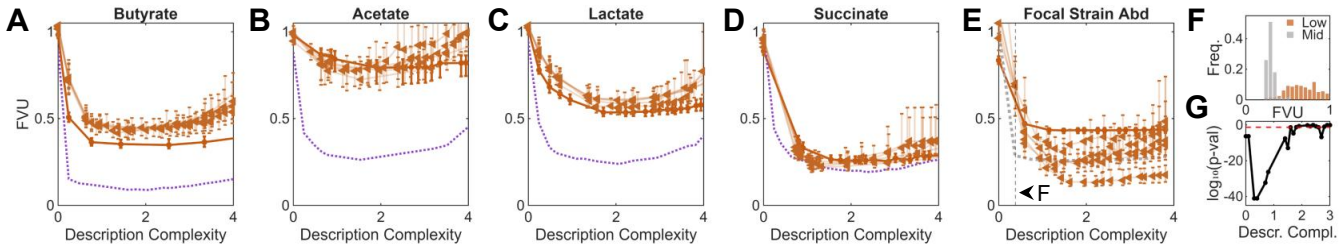

FIG. S3. **Downsampling test: observed emergent predictability is robust to sample size of datasets.** To test if our observation of emergent predictability might be affected by the differnces in the number of samples per richness tier, we randomly subsample the low-richness datasets to match the total number of strain observations across the unique high-richness communities and re-obtain Pareto fronts. Plots show 5 random sub-samplings in triangles. Low-richness (orange circles) and high-richness (purple dashed line) Pareto fronts from Fig. 4 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 that the shift of the Pareto front is a robust observation. The result of this test for Focal Strain Abundance (E) appears weaker than the other properties, but computing p-values from Wilcoxon rank-sum tests on the low- and mid-richness FVU distributions across description complexity (G) shows that the two Pareto fronts are statistically separable up to ~ 2 bits. Panel F shows an example cross-section at the lowest non-zero description complexity grouping of the mid-richness Pareto, comparing the FVU distributions across cross-validation replicates that constitute the mid-richness and subsampled low-richness fronts. To perform the statistical tests, the description complexity axis was binned at a width of 0.25 bits and FVU distributions compared within each bin. Red dashed line in G shows the 0.05 significance threshold.

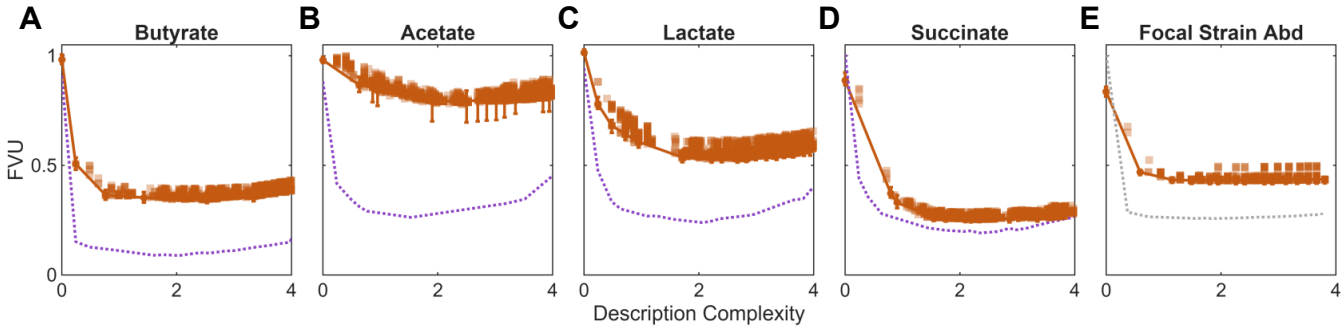

FIG. S4. **Relabeling test: observed emergent predictability is robust to Pareto front detection.** To test that the emergent predictability we observe is not due to our heuristic algorithm failing to detect the Pareto front, we examine the nearest-neighbor coarsened descriptions of the low-richness 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.

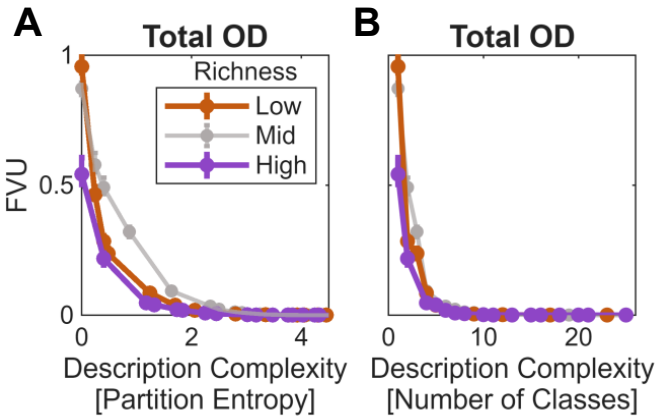

FIG. S5. Using the total OD measured in the Clark *et al.* experiments as a community property for prediction, we apply our framework to produce the Pareto fronts with description complexity $I(\Psi)$ measured as partition entropy (**A**) or number of resolved classes (**B**).

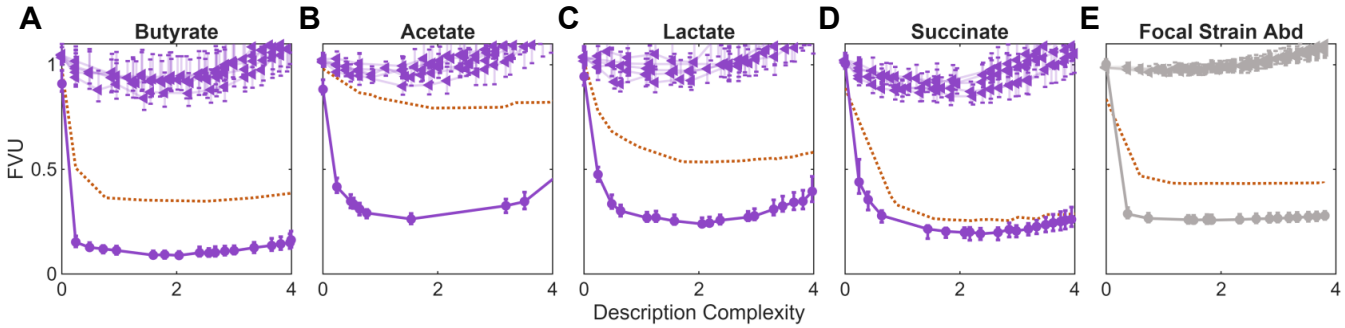

FIG. S6. **Permuting abundance data breaks emergent predictability at high richness.** To test that the observation of emergent predictability is not due to an artifact in the abundance data, we perform the following permutation test to see if it breaks our observation. For Ref. [11] we randomize the high-richness communities by shuffling the abundances of each strain with other instances in which the strains were present (preserving the presence-absence structure of each community). For Ref. [33] we randomize the mid-richness communities by shuffling the presence of each strain while preserving the total number of samples for which a given strain is observed. Pareto fronts are then inferred on the permuted data and plotted using triangle markers; lines with error bars show median and IQR for 5 respective random permutations. The higher-richness Pareto fronts from Fig. 4 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.

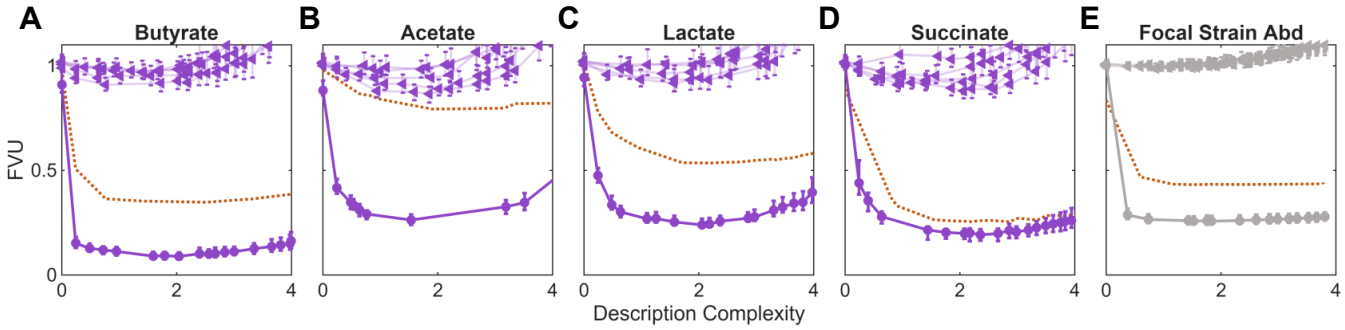

FIG. S7. **Permuting property measurements breaks emergent predictability at high richness.** To test that the observation of emergent predictability is not due to an artifact in the measurements of properties of interest, we perform the following permutation test to see if it breaks our observation. Similar to Fig. S6, 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; lines with error bars show median and IQR for 5 respective random permutations. The higher-richness Pareto fronts from Fig. 4 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.

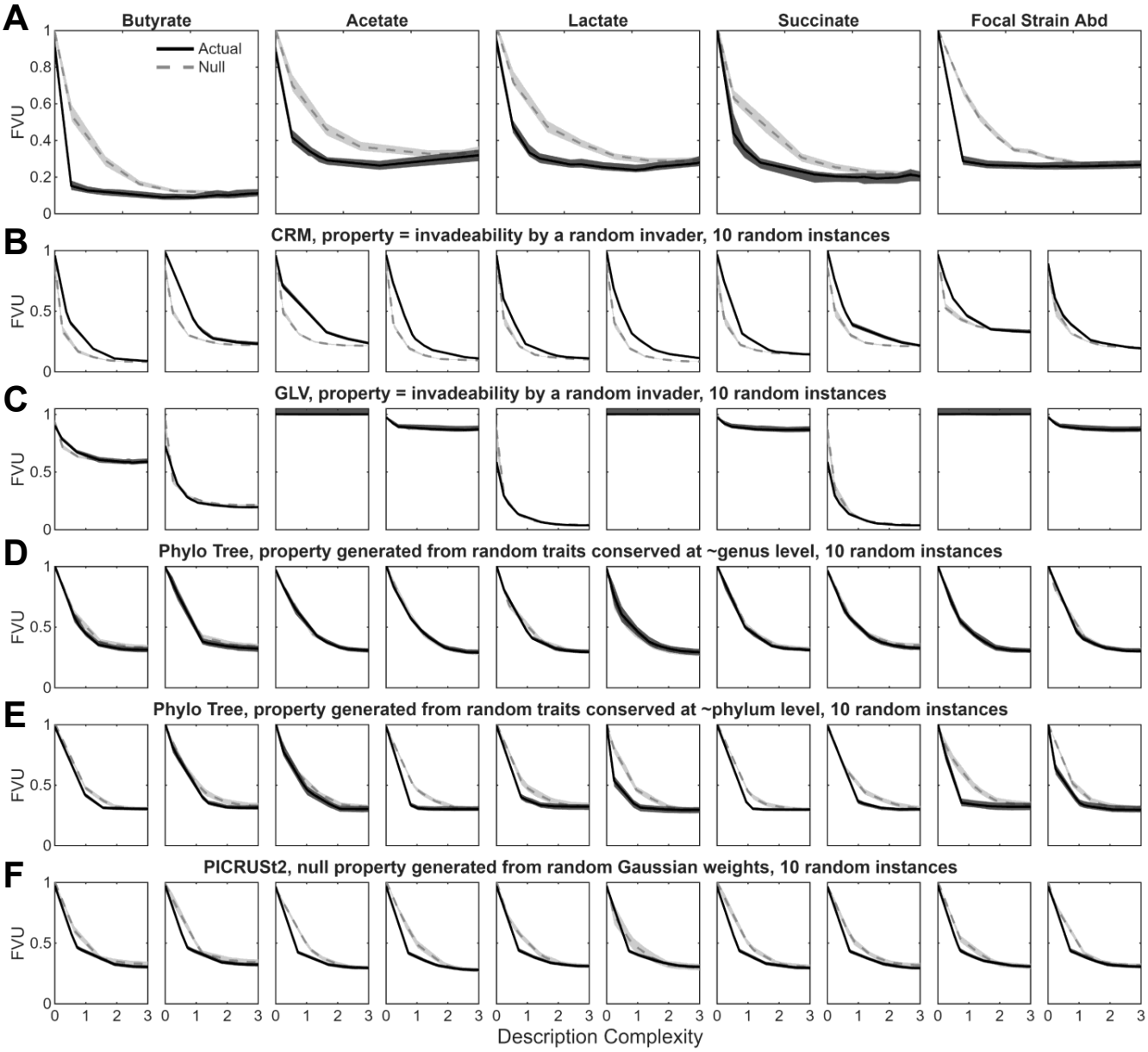

FIG. S8. Plotting the raw Pareto front data for quantifying coarse-grainability in Fig. 3. Each panel shows the median (line) and IQR (shading) across 100 cross-validation replicates for the actual Pareto versus a Gaussian null (see Methods). Each column corresponds to a different random seed for a given model instance. See text for details on simulation protocols and synthetic community property generation. Row **A**: The fermentation products measured in Ref. [11] and focal strain abundance measured in Ref. [33]. Rows **B**,**C**: Predicting post-invasion abundance of randomly drawn strains in a Consumer-Resource model (CRM) and Generalized Lotka-Volterra (GLV) model. Rows **D**,**E**: A synthetic community property produced from a random walk model of an evolved trait on the phylogenetic tree from Ref. [11], with trait conservation depth set to the Genus- (D) or Phylum-level (E). Row **F**: A synthetic community property generated from applying PICRUSt to KEGG orthologs related to carbon metabolism in the species cultured in Ref. [11]. Fig. 3 is based on averaging across the 10 panels for each model as described in the Methods.

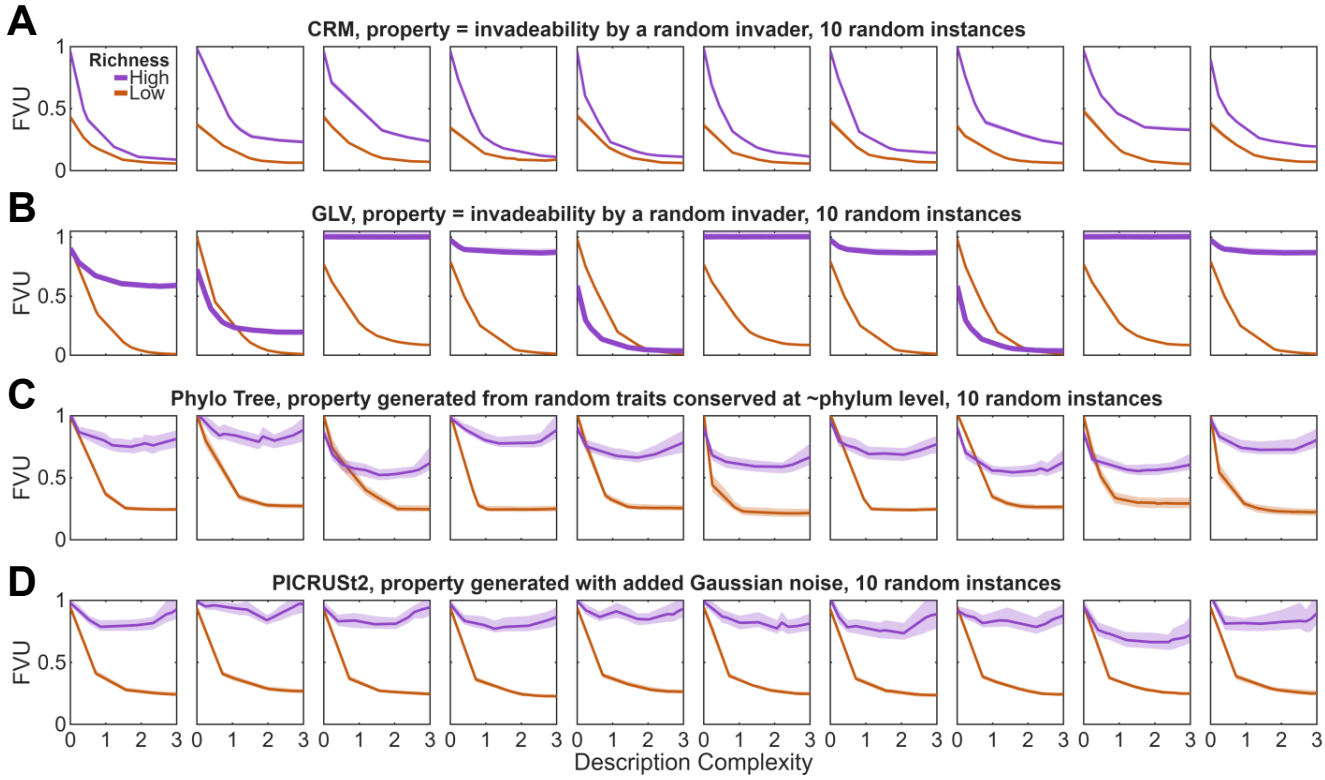

FIG. S9. Plotting the raw Pareto front data for quantifying emergent predictability in Fig. 4. Each panel shows the median (line) and IQR (shading) across 100 cross-validation replicates for the high-richness Pareto versus low-richness. Each column corresponds to a different random seed for a given model instance. See text for details on simulation protocols and synthetic community property generation. Rows **A,B**: Predicting post-invasion abundance of randomly drawn strains in a Consumer-Resource model (CRM) and Generalized Lotka-Volterra (GLV) model. Row **C**: A synthetic community property produced from a random walk model of an evolved trait on the phylogenetic tree from Ref. [11], with trait conservation depth set to Phylum-level (highly structured case). Row **F**: A synthetic community property generated from applying PICRUSt to KEGG orthologs related to carbon metabolism in the species cultured in Ref. [11]. Fig. 4 is based on averaging across the 10 panels for each model as described in the Methods.

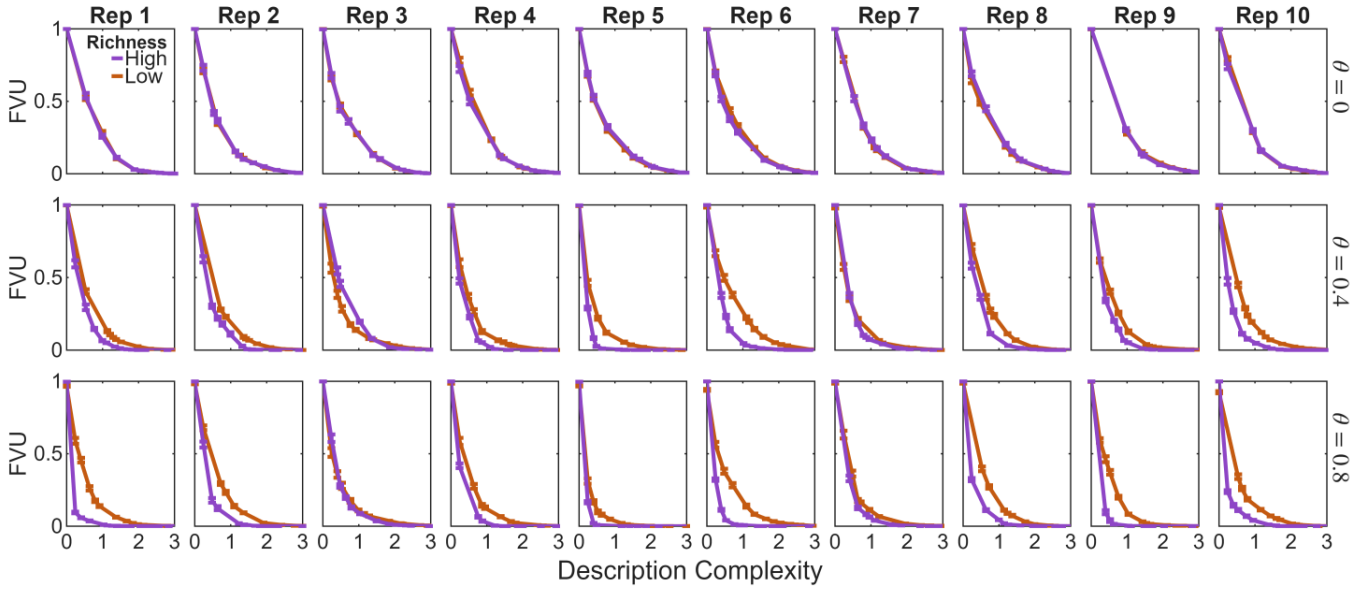

FIG. S10. Plotting the raw Pareto front data for quantifying emergent predictability in the physiological feedback model presented of Fig. 5. Each panel shows the median (line) and IQR (error bars) across 100 cross-validation replicates for the high-richness Pareto versus low-richness. Each column corresponds to a different random seed for a given model instance. See text for details on simulation protocols. Each row corresponds to a different value of feedback strength θ from the environmental variable p. Fig. 5 is based on averaging across the 10 panels for each θ as described in the Methods.

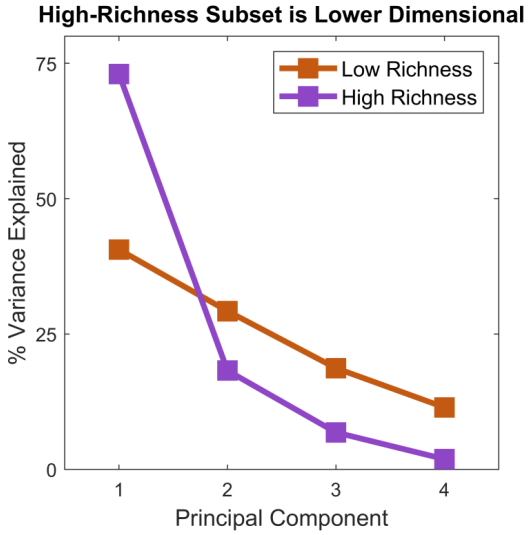

FIG. S11. PCA spectra of the four fermentation byproduct concentrations measured in Ref. [11], computed separately in the two tiers of community richness. For high-richness, the dominant component explains 73% of variance, contrasted with 41% for low-richness.

SUPPLEMENTARY TABLES

Ref.	Context	$\geq$2 richness tiers?	#(samples) / #(strains)	Satisfies criteria?
[11]	microbial	yes	870/25 (low) 477/25 (mid) 180/25 (high)	yes
[33]	microbial	yes	19433/14 (low) 346/14 (mid)	yes
[19]	microbial	yes	21/17 (low) 30/17 (mid)	not enough data
[32]	grasslands	yes	350/31 (low) 200/31 (mid) 50/31 (high)	not enough data; not all strains sampled at low and mid richness
[26, 68]	grasslands	yes	345/32 (low) 22/32 (mid)	not enough data at mid richness
[63, 73]	grasslands	yes	200/60 (low) 40/60 (mid) 10/60 (high)	not enough data
[60–62]	grasslands	no	Not evaluated	not enough data; no richness tiers
[70]	grasslands	yes	Not evaluated	plots either had all or one species
[30]	grasslands	no	Not evaluated	low richness only

TABLE S1. Catalog of published datasets that assemble communities from a fixed library of species/strains.