Spring 5-15-2023

A Framework for Investigating Random Ensembles of Structured Ecosystems and Quantifying Their Emergent Coarse-grainability

Jacob Moran

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Physics

Dissertation Examination Committee:
Mikhail Tikhonov, Chair
Anders Carlsson
Shankar Mukherji
Alexander Seidel
Ralf Wessel
Kevin Wood

A Framework for Investigating Random Ensembles of Structured Ecosystems and Quantifying Their Emergent Coarse-grainability

by
Jacob Thomas Moran

A dissertation presented to
Washington University in St. Louis
in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2023
St. Louis, Missouri

Table of Contents

List of Figures

Acknowledgments

The work in this dissertation would not have been possible without the passionate dedication of Mikhail Tikhonov to mentorship and pioneering science. I am deeply grateful to have had such a caring person as my thesis advisor, and will always appreciate the patience and effort he put towards training me as a scientist. During our time together, Mikhail mentored me on how to *think*, *do*, and *communicate* science. Through our many hours of brainstorming up at the whiteboard together, he provided me the opportunity to think creatively and taught me not how to just *answer* questions, but more importantly, how to *ask* them. For all this and more: I thank you, Mikhail.

Thank you also to Ryan McGee who provided many helpful insights and ideas to the work presented in this dissertation. I greatly appreciate both the biology and learning/information theoretic perspectives he brought to our many discussions. In addition to this, I am grateful for Ryan's friendship through the many times he gave his support and guidance on personal and career matters.

My friends I met at the physics department, and abroad, made my graduate school experience fun and intellectually stimulating. To my dear friends, Daria Kowsari and Stefan Landmann, I owe a special thanks for all the kindness, laughs, and joy you brought to me over the years. I also want to thank my friend Aditya Mahadevan for all the useful and exciting discussions we have had.

To the Wood Lab at University of Michigan, thank you for being so welcoming during my unusual transition for a graduate student. I have greatly enjoyed exchanging scientific discussions over these past months.

I am also very thankful for the many sacrifices my parents, Holly and Joe, have made for me to be on this journey, and for all their love and support along the way. I will not only pay it

forward, but also show my sincere appreciation by using the opportunities you have given me to give back to you in years to come.

Finally, I am deeply thankful to my partner Sydney, who has stood by me to provide love and support through this experience every step of the way. Thank you for sharing your life with me – I am very grateful we set off on this journey together. Over the past nine years, you have not only helped me grow as a scientist, but have also taught me the importance and value of a well-rounded and balanced lifestyle (jelly beans!). It is to you that this dissertation is dedicated.

Jacob Moran

Washington University in St. Louis

May 2023

To my partner, Sydney.

ABSTRACT OF THE DISSERTATION

A Framework for Investigating Random Ensembles of Structured Ecosystems and Quantifying Their Emergent Coarse-grainability

by

Jacob Thomas Moran

Doctor of Philosophy in Physics

Washington University in St. Louis, 2023

Professor Mikhail Tikhonov, Chair

The interface between statistical physics and theoretical ecology has a long history, employing powerful concepts such as ensemble approaches and typicality to study emergent properties of ecosystems. This of course raises the question of what ensembles are useful to describe the typical behaviors of evolution and ecology, but so far, the traditional context of high-diversity ecology has considered ensembles of random, unstructured ecosystems. Although much insight has been gained in this regime, one naturally wonders how representative are random ensembles of real, natural ecosystems that are arguably atypical and highly structured by evolution. Moreover, the question of coarse-graining ecosystems has yet to be addressed because the very ingredient responsible for predictive coarse-grained descriptions – ecosystem structure – is explicitly absent from the current theoretical framework. This dissertation investigates the coarse-grainability of ecosystems within minimal models that intend to capture the atypicality generated by evolution, aiming to establish a conceptual language from which a general theoretical framework can be built.

In the first two chapters, I review the applications of statistical physics in classical models of ecology, moving on to then explore the evolutionary consequences of the atypicality that arises

from evolution. In Chapter 3, I present a model for investigating random, *structured* ecosystems, enabling me to begin studying the emergent coarse-grainability of microbial ecosystems. In particular, I develop the hypothesis that a high strain diversity, despite being nominally more complex, may in fact facilitate coarse-grainability, which is maximized when an ecosystem is assembled in its native environment. Building on this framework in Chapter 4, I provide a more principled approach for defining coarse-grainability by systematically mapping the prediction power versus information content of coarse-grained descriptions of ecosystem composition. Applying this framework to experimental data, I confirm the diversity-enhanced coarse-grainability hypothesis and discuss how this effect cannot be reproduced in standard ecological models parameterized using random ensembles. Finally, I link these results to the theoretical concept of functional attractors of diverse ecosystems.

Chapter 1: Background and Motivation

The sister disciplines, ecology and evolution, have been developing for centuries. With the influential quantitative approaches introduced by pioneers like Lotka and MacArthur [1–3], came a surge of mathematical theory to ecology. This foundational work served as a conduit for importing approaches from statistical physics, starting with the classic papers of May in the 1970s [4]. Prior to this, most studies within mathematical models of ecology were confined to simple ecosystems of low diversity (a few species and resources). However, the natural world displays many ecosystems with a rich diversity of species coexisting. Incorporating this naturalistic complexity into the current modeling framework of the time was the challenge on which May paved progress.

For modeling large, diverse ecosystems, parameterizing the dynamical equations becomes increasingly more complex with the inclusion of more species and environmental factors, especially in any thermodynamic limit (e.g., number of species and resources are taken to infinity with a fixed ratio). This challenge arises in many other contexts involving complex systems, and one influential workaround is to take an ensemble approach from statistical physics, of which Wigner's modeling of heavy atomic nuclei with random matrices is perhaps one of the most notable successes of such an approach [5]. An ensemble approach aims to provide a statistical sense of possible behaviors produced by the dynamical model of interest by sampling parameters from an appropriately chosen random distribution. The key insight then is that patterns that are typical to some ensemble of systems are more generalizable and reproducible than the idiosyncratic details from any one realization. From this, the hope is that any principles one gleans from an ensemble approach are applicable and predictive of specific realizations, assuming a given

realization is representative of what is to be considered typical. However, what is considered typical depends on the choice of ensemble. In the context of ecology, the standard choice is a random, unstructured ensemble of ecosystems [6–13], but natural ecosystems are continuously being structured through evolution [14–25]. Therefore, by processes such as natural selection, evolution generates ecosystems that would be considered *atypical*, raising questions such as: which ecosystem properties or patterns are captured by random, unstructured ensembles, and which are not? How representative are these ensembles of natural populations?

1.1 Insights from Models of Unstructured Ecosystems

Recent applications of statistical physics in large-N ecology (high-diversity context) have primarily focused on predicting emergent statistical trends, such as scaling laws [26] and other global patterns. For example, a generalized version of MacArthur's consumer-resource model, where model parameters are drawn randomly from an unstructured ensemble, can qualitatively reproduce distributions of community composition observed in ocean microbiome data collected across the globe [10]. Work like this serves an important role in highlighting which features presented in empirical observations should be considered surprising due to nontrivial underlying mechanisms versus those that can be reproduced with simple random models.

Other work using similar modeling approaches has sought after understanding the relationship between ecosystem diversity and stability [4,27–29]. For example, a recent theoretical study, also performed in a random consumer-resource model, demonstrates that incorporating metabolic trade-offs in resource consumption leads to highly diverse, yet stable communities with rank-abundance patterns that resemble those observed in real-world ecosystems [29]. Progress such as this contribute to uncovering the principles behind the maintenance of biodiversity in nature; a crucial goal for facing climate change [30–34].

In large part, the progress made in the regime of unstructured ecosystems has been motivated by the question of diversity and coexistence. Despite the rich biodiversity presented by the natural world, intuition from classical theory (i.e., competitive exclusion principle) argues that the diversity of community members should be limited by the amount of energy resources [35], but this is often violated even in simple, well-controlled environments of a lab [21,22]. Recent theory developed in minimal extensions of classic models have proposed plausible mechanisms to reconcile this contrast, showing how diversity may emerge from dynamical oscillations and chaos [36–39], metabolic trade-offs [29,40,41], or cross-feeding [10,42]. And although natural ecosystems are known to be structured in various aspects [14–20,43,44], these models of unstructured populations admittedly do capture empirically observed patterns of diversity as described above.

1.2 A New Direction

Instead of aiming to explain how different eco-evolutionary processes in a given environment enable coexistence of a diverse community, I will take the observation that natural communities are extremely diverse as my starting point for asking when such diversity can be usefully coarse-grained for predicting properties of specific communities. In the case of microbial ecosystems, communities are complex dynamic systems often composed of hundreds of many interacting species. Because of this, expecting the prediction of functional behaviors performed by such heterogeneous many-body systems to be generally infeasible would be understandable. And yet, strikingly, microbial ecosystems appear to be at least partially coarse-grainable, in the sense that some functional observables of interest (e.g., nutrient production) can be predicted by effective models with far fewer variables than the number of interacting lineages. For practical purposes at

least, systems consisting of thousands of taxa, such as industrial bioreactors, are well described by models with just a handful of variables [45,46].

Intuitively, the predominant reasoning for why this is possible is that coarse-grainability directly follows from the hierarchical structure imprinted by evolutionary descent. Taxonomically then, if a community is composed of 100 interacting strains are all close variants of just 10 species that can be further grouped into 2 families, it is intuitively plausible for the community to be approximated by a model with just several variables. However, recent experiments have shown that closely related strains exhibit very different dynamics and interactions in their community context [47], revealing taxonomic structure to likely be inadequate to fully explain coarse-grainability. Ecosystems exhibit other forms of structure at all levels of resolution though, such as the distribution of functional traits across taxa [17,48], and this structure has been shown to matter both for community-level functions [44,49–54] and for theoretical models to capture empirical observations [11]. For example, a random consumer-resource model that incorporates cross-feeding interactions structured by "universal rules of metabolism" can reproduce a wide range of empirical observations [9,11], albeit primarily at the level of broad statistical patterns. Therefore, developing a framework to investigate ensembles of structured ecosystems seems necessary to fully grasp the concept of coarse-grainability in these systems and tap into its potential for practical purposes.

Chapter 2: What Evolves is Atypical

As mentioned in the previous chapter, the mechanisms of evolution generate new structure in ecosystems, constantly pushing populations to atypical corners of fitness distributions. This chapter explores an evolutionary consequence of this atypicality. See Appendix A for a brief demonstration of how evolution generates atypicality in the model I present in this chapter, as well as an example ensemble calculation to demonstrate the amenability of the model to analytical methods from statistical physics.

Expression level is known to be a strong determinant of a protein's rate of evolution. But the converse can also be true: evolutionary dynamics can affect expression levels of proteins. Having implications in both directions fosters the possibility of an "improve it or lose it" feedback loop, where higher expressed systems are more likely to improve and be expressed even higher, while those that are expressed less are eventually lost to drift. Using a minimal model to study this in the context of a changing environment, we demonstrate that one unexpected consequence of such a feedback loop is that a slow switch to a new environment can allow genotypes to reach higher fitness sooner than a direct exposure to it.

The work presented in this chapter has been adapted from the following publication:

> Moran J, Finlay D, Tikhonov M. "Improve it or lose it: Evolvability cost of competition for expression." Physical Review E 103, 062402 (2021).

2.1 Introduction

The rates of protein evolution are affected by a multitude of factors, including protein-protein interactions, stability-based constraints or dispensability [55–64]. However, the strongest single determinate appears to be expression level [65,66]. For instance, substantial evidence suggests that

lower-expressed proteins are less protected from drift, whereas highly expressed proteins are under stronger purifying selection [65–68].

Conversely, the expression level can itself be affected by evolution, especially for proteins or pathways that are dispensable or partially redundant. For example, a protein that is disabled by a deleterious mutation becomes a metabolic burden (or may be directly toxic), favoring a reduction in expression.

Since partial redundancy is believed to be widespread [69], this creates a theoretical possibility of a feedback loop. Consider an organism with several partially substitutable systems or pathways fulfilling a similar function; for example, several metabolic pathways to satisfy its requirement for carbon, or several sensing modalities to respond to environmental cues. In these circumstances, it seems plausible that the systems used more, being under a stronger selection pressure, would be more likely to improve and be used even more. In contrast, the lesser expressed systems could be more likely to deteriorate and be used even less (Figure 2.1).

This process – effectively a “competition for expression” – could be viewed as an extension of Savageau's “use it or lose it” principle, and is conceptually similar to the generalist-to-specialist transition of ecological specialization [70,71], but is rarely discussed in an evolutionary context. One reason, perhaps, is that this intuition appears to predict that highly expressed proteins should evolve faster, the opposite of what is observed empirically [65]. However, as we will show, the consequences of such a feedback interaction are more nuanced.

To do so, we illustrate the feedback loop of Figure 1.1 in a simple minimal model. For highly adapted systems depleted for beneficial mutations, we find that the highest-expressed proteins are still expected to evolve slowest, in full agreement with the empirical observations. In contrast, for

an evolutionary process driven by strong adaptive mutations, e.g. following a strong environmental change, the sign of the correlation between expression and evolutionary rate is predicted to transiently invert. Moreover, at least in our model, the consequences of the "improve it or lose it" feedback include interesting qualitative effects, such as a loss of evolvability caused by an environmental perturbation that is too strong. As an example, we demonstrate how a gradual change to a new environment can lead to a higher rate of fitness gain than direct exposure.

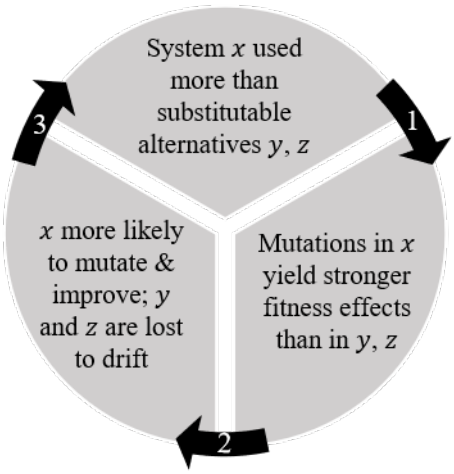

Figure 2.1 | The "improve it or lose it" feedback loop. In this schematic, x, y, and z are partially substitutable systems fulfilling a similar function (e.g., metabolic pathways for alternative sources of carbon). Adaptive mutations in the highest-used system x have stronger fitness effects than y, z (arrow 1). The stronger selection pressure makes system x more likely to mutate and improve (arrow 2). This improvement in x allows the organism to rely on it even more (arrow 3), completing the loop.

2.2 Model and Context

To study the "improve it or lose it" feedback loop, we need an evolutionary model that explicitly includes a notion of usage/expression. For this reason, we adopt the toolbox model from Ref. [72], summarized in Figure 2.2A.

Briefly, we think of a genotype as encoding a set of K systems that can be used at different levels to optimize the fitness of the organism in a given environment. Mathematically, we represent the

K systems as basis vectors $\{\vec{g}_\mu\}$ $(\mu = 1 \ldots K)$ and the environment as a target vector $\vec{E}$ in an abstract L-dimensional space (which can be interpreted as the phenotype space [72]. The fitting problem can be written as,

$$\{a_\mu\} = \mathrm{argmin}_{\{a_\mu \geq 0\}} \left\| \vec{E} - \sum_\mu a_\mu \vec{g}_\mu \right\|, \tag{2.1}$$

where the environment-dependent coefficients $\{a_\mu\}$ can be interpreted as the extent to which the organism relies on a given system $\vec{g}_\mu$ in $\vec{E}$. The quality of fit, which these $\{a_\mu\}$ optimize, can then be interpreted as the fitness of the genotype $G = \{\vec{g}_\mu\}$ in environment $\vec{E}$:

$$F(G, \vec{E}) = -\min_{\{a_\mu \geq 0\}} \left\| \vec{E} - \sum_\mu a_\mu \vec{g}_\mu \right\|. \tag{2.2}$$

In Ref. [72], the coefficients $\{a_\mu\}$ are called "expression level"; however, conceptually, they correspond more closely to the intuitive notion of "usage". Indeed, a larger a_μ in this model corresponds to a system whose deletion would have a stronger fitness effect, rather than one present in a larger copy number (although in practice, the two properties are, of course, correlated [62]). Throughout this work, we refer to $\{a_\mu\}$ as usage coefficients.

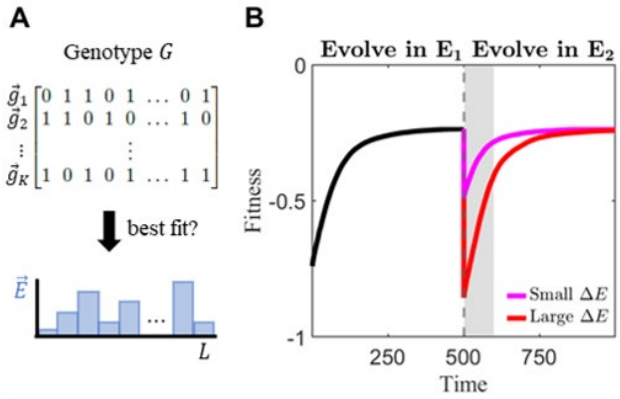

Figure 2.2 | A context to study the "improve it or lose it" feedback loop. (A) In the toolbox model, a genotype is a matrix representing the available "systems" an organism can (linearly) combine to approximate the optimal phenotype required by the environment, $\vec{E}$. The coefficients of the best approximation are interpreted as usage levels a_μ, serving as a proxy for expression. Matrix elements are chosen to be binary (0 or 1) so that mutations in the evolutionary process can be implemented as bit flips. (B) Fitness trajectories of initially random genotypes evolving under $\vec{E}_1$ before switching to $\vec{E}_2$ a distance ΔE away. We choose to study the feedback loop and its consequences during the early-time dynamics after switching (gray region).

For simplicity in simulating evolution within this model, we assume that mutations are rare and selection is strong, so that we need only track the evolutionary trajectory of a single genotype [73]. In each simulation step, we enumerate all beneficial point mutations of the current genotype by performing all single bit-flips of the genotype matrix. We then pick one of these mutations as the first to rise to fixation; in this parameter regime, selection only considers beneficial mutations, and fixation probability is proportional to a mutation's fitness effect. We note that, when evaluating the fitness of mutants, the usage coefficients are optimized for the mutated genotype, and thus typically differ from those of the parent. This corresponds to the assumption that the evolution of $\{a_\mu\}$ occurs on a much faster timescale than evolution of system vectors $\{\vec{g}_\mu\}$ (a separation of timescales; see Section 2.6 for more discussion).

Figure 2.2B shows an example of fitness dynamics of random initial genotypes first exposed to a random environment $\vec{E}_1$ and then to a different random environment $\vec{E}_2$. The feedback loop we

will describe is already present during the early-time dynamics of evolution in $\vec{E}_1$; however, we choose to focus on the time period that follows the environment switch (shaded gray region). This will allow us to use the difference between the two environments, $\Delta E = \left\| \vec{E}_2 - \vec{E}_1 \right\|$ as a natural control parameter (see Section 2.6.3 for parameterization of environment pairs $(\vec{E}_1, \vec{E}_2)$).

In what follows, we use $\vec{E}$ vectors of unit length so that fitness is constrained to $-1 \leq F \leq 0$. We fix $L = 40$ and vary K, and consider genotype matrices with binary values, 0 or 1, initialized randomly with probability $p = 0.5$ of being 1. Since environments are represented by unit vectors with positive components, ΔE is confined to the range $\Delta E \in [0, \sqrt{2}]$. We will show that ΔE controls the strength of the feedback loop, with stronger changes in environment (large ΔE) inducing stronger feedback.

2.3 The toolbox model exhibits the "Improve it or lose it" feedback

Figure 2.3A depicts a representative trajectory of the "improve it or lose it" feedback realized in the toolbox model. The panel shows the dynamics of usage coefficients after a genotype with $K = 5$ systems, pre-adapted to some environment $\vec{E}_1$, was switched to a different environment $\vec{E}_2$, with $\Delta E = 1$ (random environment pairs with a given ΔE were generated as described in the Section 2.6.3). Note that, after each mutation, the usage coefficients are re-optimized according to eq. (2.1) and thus change discontinuously (see Section 2.6.2 and Ref. [72]); however, these steps are typically small, creating an illusion of smooth dynamics. We see that strong adaptive mutations initially concentrate in the two systems with highest usage (frequent redder dots). As they mutate, they also rise in usage, a_μ. In contrast, the lower-used systems decrease in usage, and mutate only rarely, with relatively weak fitness effects (bluer dots).

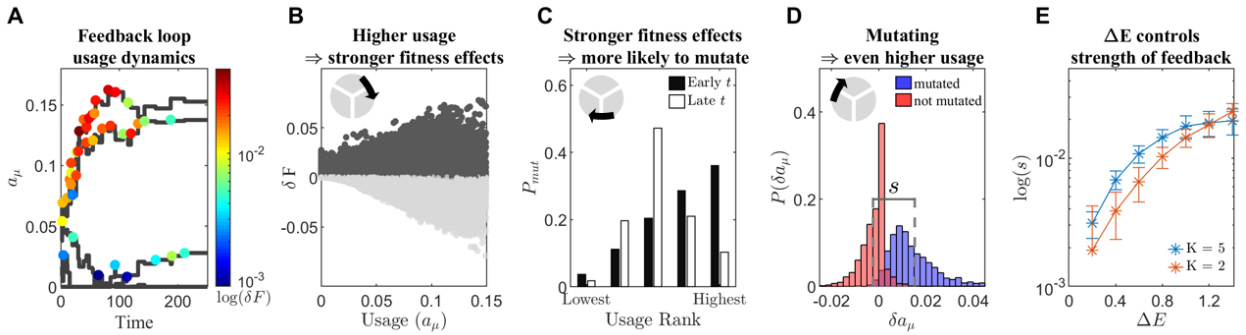

Figure 2.3 | The toolbox model exhibits the feedback loop. (A) An example of evolutionary dynamics of usage coefficients after a genotype adapted to a random environment $\vec{E}_1$ is switched to another random environment $\vec{E}_2$ with $\Delta E \equiv \left|\vec{E}_1 - \vec{E}_2\right| = 1$. Despite similar usage initially, by $t = 100$ only two of $K = 5$ systems remain in use. Dots mark the systems in which a beneficial mutation arose, color indicates fitness effect (red is strongest). In panels (B-D), we examine the statistics of usage dynamics and mutation effects within the first 3 time steps of 20 trajectories in 15 random environment pairs with the same $\Delta E = 1$. Inset pictograms refer to feedback steps as shown in Figure 2.1. (B) Fitness effects of all available mutations in each system versus system usage. Dark and light gray points are beneficial and neutral/deleterious mutations, respectively. Higher-used systems possess stronger fitness effects. (C) Probability of a system to mutate, plotted against its usage rank (ascending order). At early times (black bars), higher used systems are more likely to mutate. As the strong beneficial mutations in highest-used systems are depleted, the probability of mutating shifts towards lower used systems (white bars). (D) Distribution of change in usage of a system that just mutated (blue) or a system that failed to mutate (red) at a particular simulation step. The difference in means of these conditional probability distributions, s, quantifies the strength of the feedback loop. (E) The strength of the feedback loop s is controlled by the magnitude of environmental change ΔE. Error bars represent 1 standard deviation (SD) over 300 replicates.

Although the details of these dynamics are shaped by eq. (2.2) and are of course model-dependent, on a qualitative level the instability driving a subset of usage coefficients up at the expense of others can be directly traced to the feedback loop summarized in Figure 1, as we will now show.

First, agreeing with the intuitive notion of $\{a_\mu\}$ as "usage", systems with higher a_μ tend to harbor stronger fitness effects. To see this in our model, we plot the fitness effects of all available mutations within the first few simulation steps against the usage coefficient of the system where they occur (Figure 2.3B). As expected, both beneficial (dark gray) and deleterious (light gray) mutations are stronger in systems that have a higher usage coefficient a_μ.

As a result, higher-used systems are more likely to mutate, because mutations with a larger fitness effect are more likely to escape drift and fix in the population [74]. The black bars in Figure 2.3C show the early-time probability of each system to mutate, plotted against its usage rank.

Finally, Figure 2.3D shows the distribution of usage changes, defined as the difference in usage δa_μ before and after a simulation step, over the same early time period as described above. Whenever a system mutates, its usage typically increases (Figure 2.3D, blue). In contrast, the systems that did not mutate at that particular timestep typically drop in usage (Figure 2.3D, red). In our model, this also is ultimately a consequence of eq. (2.2), but it is not the model that justifies this behavior. Rather, it is this behavior that justifies using the model, making it appropriate for studying the feedback loop that this behavior induces. In summary, Figure 2.3B-D demonstrates all three arrows from Figure 2.1 at play in our model.

Since a greater separation between the distributions of Figure 2.3D would entail stronger feedback, we can use the difference in the mean of these conditional distributions, denoted as s, as a measure of the feedback strength. Figure 2.3E demonstrates that, as expected, the feedback becomes stronger (increasing $\log s$) as the change in environment becomes more severe.

The rapid evolution of highly used systems (Figure 2.3C) may seem to be at odds with experimental work showing that highly expressed proteins evolve slowest [65]. However, the mechanism described here is fully compatible with the explanations previously proposed for this experimental result. The effect shown in Figure 2.3B (higher used systems have stronger fitness effects) applies to both beneficial and deleterious mutations. For early stages of adaptation driven by beneficial mutations (as considered here), this means the most-used systems will evolve first. However, at later stages, as beneficial mutations are depleted, the same argument dictates that the

most-used systems become the most protected, and evolve slowest. We demonstrate the presence of this effect by replotting the per-system mutation probabilities at a later time (Figure 2.3C, white bars); the probability of mutating begins to shift from higher used to lesser used systems. This result therefore predicts that the negative correlation between expression and evolution rate observed in [65] should transiently invert following a change in environment. If additional factors like interaction and stability constraints on evolutionary rates are considered, our framework predicts that the negative correlation would at least weaken, with the size of the effect controlled by the magnitude of the environmental change. Encouragingly, this transient weakening in negative correlation between expression and evolutionary rate is consistent with recent analysis of evolutionary rates in yeast [75].

2.4 The cost to evolvability

Intuitively, one might expect that the competition for usage mediated by the "improve it or lose it" feedback loop may be detrimental for the organism, since it effectively reduces the number of systems it has available. Implementing this effect in a simple model allows us to make this intuition precise. We will see that, at least in our model, the feedback loop exhibited above reduces the adaptive potential of the genotype, and mitigating its effects can allow for faster adaptation.

For this, we compare the fitness trajectories of genotypes evolving in conditions that exacerbate the feedback and those that weaken it. Specifically, starting from a genotype pre-adapted to $\vec{E}_1$, we compare two ways of adapting it to a new, strongly different environment $\vec{E}_2$: either by exposing it to $\vec{E}_2$ directly (as discussed above), or by changing the environment from $\vec{E}_1$ to $\vec{E}_2$ slowly (on a timescale that is slow compared to mutation fixation). By avoiding large environment jumps, we

expect the gradual switch to weaken the feedback loop. The question we ask is which exposure protocol will ultimately lead to higher fitness in the environment of interest, $\vec{E}_2$.

An example of this comparison is shown in Figure 2.4A. The red curve shows fitness (in the environment of interest $\vec{E}_2$) for genotypes evolving under the slow-exposure protocol, implemented by linearly relaxing the environment vector from $\vec{E}_1$ to $\vec{E}_2$ over a time τ:

$$\vec{E}(t) = \begin{cases} \text{normalize}\left[\vec{E}_2 + \dfrac{\tau - t}{\tau}\left(\vec{E}_1 - \vec{E}_2\right)\right], & \text{if } t < \tau \\ \vec{E}_2, & \text{if } t \geq \tau \end{cases} \tag{2.3}$$

(the environment vector in our model is always normalized to unit length). The τ we use is large relative to the typical time between mutations ($\tau = 100$; compare to Figure 2.3A). The red curve $F_{\text{SE}}(t)$ (slow exposure) is to be compared to the blue curve $F_{\text{DE}}(t)$ (direct exposure), showing fitness of the same initial genotypes evolving directly in $\vec{E}_2$.

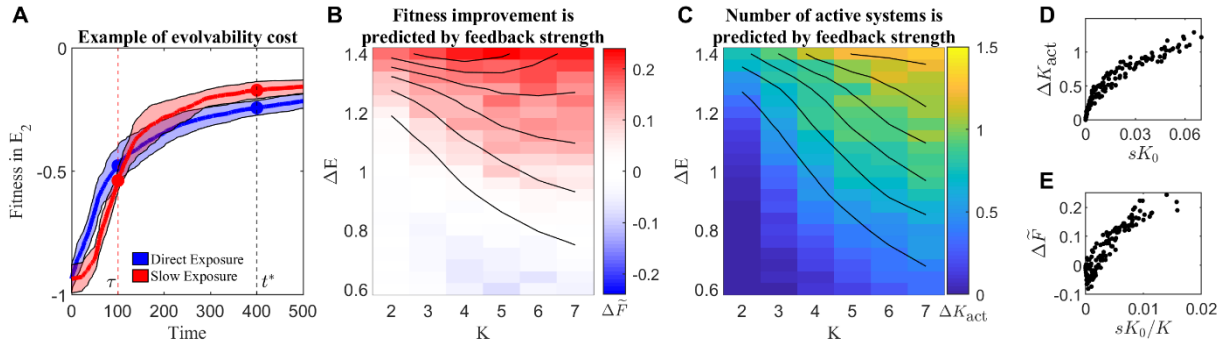

Figure 2.4 | Higher evolvability from slow exposure than direct exposure. (A) Fitness, evaluated in environment $\vec{E}_2$, of genotypes that are either directly exposed (DE) to $\vec{E}_2$ at $t = 0$ (blue trace) or slowly exposed (SE) to $\vec{E}_2$ over a time $\tau = 100$ according to the protocol defined in Eq. (2.3) (red trace). Each trace shows mean ± 1 SD (shading) of 20 replicate trajectories of genotypes with $K = 4$ systems in a random environment pair with $\Delta E = 1.4$. Colored dots highlight that slow exposure leads to higher long-term fitness, despite slower fitness gain initially. The relative improvement in fitness, $\Delta\widetilde{F}$, is measured at an arbitrarily late time point $t^* = 400$ (see Section 2.6.4 and Figure 2.7 for later t^*). (B) Heatmap of the long-term relative fitness improvement, $\Delta\widetilde{F}$. Contour lines show $\Delta\widetilde{F}$ can be predicted by the feedback loop strength s and

the number of initially inactive systems K_0 (see panel E). Here and in the remaining panels, results are averages over 20 trajectories in 15 random environment pairs with varying ΔE. (C) Heatmap of ΔK_{act}, the average difference in number of active systems ($a_\mu > 10^{-3}$) between the SE and DE protocols at $t = \tau$. Contour lines show it is predicted by the product sK_0; see panel D. (D) ΔK_{act}, the increased number of active systems at $t = \tau$, is predicted by sK_0, measured at trajectory start. (E) The long-term fitness improvement $\Delta\widetilde{F}$ at $t = t^*$ is predicted by sK_0/K, measured at trajectory start.

The vertical dashed line at $t = \tau$ marks the timepoint where the "red genotypes" evolving under the slow-switching protocol are finally exposed to $\vec{E}_2$ for the first time. It is therefore not surprising that they are less fit than the "blue genotypes", who have been evolving in $\vec{E}_2$ from the start ($F_{\mathrm{SE}}(\tau) < F_{\mathrm{DE}}(\tau)$; red curve below the blue). However, while more fit, the blue genotypes are manifestly less evolvable: From $t = \tau$ onwards, both red and blue curves document evolution in the same environment $\vec{E}_2$, but the red curve gains fitness much faster, and overtakes the blue.

To quantify the strength of this effect, we consider the relative improvement of fitness provided by the smooth protocol, compared to direct exposure:

$$\Delta\widetilde{F}\,(t^*) \equiv \frac{F_{\mathrm{SE}}(t^*) - F_{\mathrm{DE}}(t^*)}{|F_{\mathrm{DE}}(t^*)|}.$$

While initially negative, in the example of Figure 2.4A this quantity becomes positive at a later time. To demonstrate the robustness of this observation, Figure 2.4B shows $\Delta\widetilde{F}\,(t^*)$ for a range of K and ΔE, computed at an arbitrary late timepoint $t^* = 400$ (see Section 2.6.4 and Figure 2.7 for $\Delta\widetilde{F}$ at a later value of t^*). We see that, at large ΔE, the slow-switching protocol consistently outperforms direct exposure, and more so as K increases. While the scenario of an organism possessing $K = 7$ competing systems fulfilling a similar function is arguably unrealistic, we note that the effect is already present at $K = 2$. (For the purposes of illustration, the example in panel

A used $K = 4$ and a dramatic environment change $\Delta E = 1.4$, when the effect is strongest.) Note that, for simplicity, in Figure 2.4B our slow exposure protocol eq. (2.3) used the same value of the relaxation time $\tau = 100$ for all K and ΔE; optimizing over this parameter and the observation timepoint t^* could of course render the effect stronger.

The origin of this effect is the "improve it or lose it" instability affecting the genotypes undergoing an abrupt environment switch, effectively leaving them with fewer systems. To confirm this, we record the average number $K_{\mathrm{act}}^{\mathrm{DE}}$, $K_{\mathrm{act}}^{\mathrm{SE}}$ of "active" systems (usage $a_\mu > 0.001$) observed at time $t = \tau$ under both protocols. As expected, a slow environment change leaves more systems active; the difference $\Delta K_{\mathrm{act}} \equiv K_{\mathrm{act}}^{\mathrm{SE}} - K_{\mathrm{act}}^{\mathrm{DE}}$ is shown in Figure 2.4C and exhibits a trend similar to Figure 2.4B. Since unused systems harbor weak mutations only (cf. Figure 2.3B), a genotype with few active systems finds itself on a fitness plateau, and its rate of fitness gain is reduced.

Finally, we can quantitatively relate both effects to the strength of the feedback loop as defined above. To start, we focus on the increase in the number of active systems ΔK_{act} in Figure 2.4C. Denote K_0 the number of inactive systems at time $t = 0$ (immediately after the environment switch; usage $a_\mu < 0.001$). This is the number of systems that the slow-exposure protocol could conceivably "rescue". One expects ΔK_{act} to scale with K_0, and if our argument is correct, it should also scale with the strength of the feedback loop s. Indeed, we find ΔK_{act} to be predicted by the product sK_0 (Figure 2.4D). The availability of these additional systems translates into additional adaptive opportunities, and ultimately a higher fitness. In a strongly epistatic model like ours, the exact relationship to the long-term fitness is hard to predict. Nevertheless, it is reasonable to expect the fractional effect on fitness $\Delta\widetilde{F}$ to at least correlate with the fractional effect on the number of active systems $\Delta K_{\mathrm{act}}/K$. If so, then $\Delta\widetilde{F}$ should correlate with sK_0/K, an expectation confirmed in

panel Figure 2.4E. Given the approximate nature of this argument, the correlation observed in Figure 2.4E is in fact surprisingly good. For convenience, the same sK_0 and sK_0/K data, Gaussian-smoothed for visualization purposes, are shown as contour lines superimposed on the heatmaps of Figure 2.4B,C. It is worth emphasizing that our definition of the feedback strength s is computed from the statistics of the first 3 mutations, which only take $t \sim 7 \pm 5$ to occur; and K_0 is similarly measured at the very start of the trajectory. Nevertheless, at least in our model, these early-time properties are predictive of the long-term evolutionary outcome at $t^* = 400$.

2.5 Conclusions and Discussion

In this work, we used a minimal model to explore a possible feedback loop between the usage of a system and its rate of evolution. Within this model, we demonstrated that this feedback loop is particularly pronounced after strong shifts in the selecting environment and can negatively impact evolvability (future fitness gain). In particular, we described a mechanism by which a slow switch to a new environment can allow the genotypes to reach higher fitness sooner than a direct exposure to it. Interestingly, this effect is reminiscent of recent results from the Evolthon crowdsourcing effort, which found that when yeast and *E. coli* populations are slowly exposed to cold temperatures they attain higher fitness than those that undergo a direct exposure [76].

A situation where exposure to a different environment E' can help evolve better fitness in E than a direct exposure to E itself is not, in itself, novel. One well-established scenario for this to occur is the crossing of fitness valleys (or plateaus): much like an enzyme that catalyzes a reaction by stabilizing the reaction transition state, a transient exposure to E' can facilitate reaching a higher fitness peak by enabling prerequisite mutations that would otherwise be unfavorable (or neutral) [77,78]. However, the scenario described here is particularly interesting because the fitness plateau is not an idiosyncratic property of a particular landscape, but emerges through

evolution itself. Fitness landscapes of evolved systems are themselves shaped by evolution [79,80], and at least in our model, the feedback mechanism we described generically induces a fitness plateau following a sufficiently strong environmental change.

To focus on this effect, our proof-of-principle model ignored many other factors contributing to rates of protein evolution. In any realistic scenario, the feedback interaction we described will only be a part of a larger picture. Nevertheless, our analysis predicts that the empirically observed negative correlation between expression and evolution rate would transiently weaken following a change in environment ΔE, and this weakening should be more pronounced for stronger ΔE. We expect this effect to be more evident if other constraints not included in our model are weakened, following, for example, a genome duplication event [75].

It is worth stressing that we considered beneficial mutations only. Clearly, if deleterious mutations were included, our feedback loop would become even stronger: in addition to the effect described, the lesser-used systems would also be less protected from drift [81–83]. This observation could then be seen as the traditional manifestation of the "use it or lose it" principle; in particular, the problem of maintaining redundancy in the face of drift has been extensively discussed [84]. Focusing on beneficial mutations only, and thus explicitly excluding any drift-dependent effects, allows us to highlight a novel aspect. Unlike the discussion of Ref. [84], here, no system is ever fully redundant, and all remain under selection. Nevertheless, some are progressively lost even in the absence of deleterious mutations – simply because the beneficial mutations preferentially target the systems used more, and those that fail to improve become obsolete. This mechanism is clearly analogous to the Red Queen effect [85] (to remain useful, a system must keep improving), except here it applies to an effective competition for expression. In this way, the loss of evolvability described in Figure 2.4 can be seen as a form of a conflict of levels of selection [86]: the

competitive dynamics between lower-level entities (the K "systems" in our model) lead to negative consequences for the organism as a whole -- a decline of phenotypic flexibility and evolvability due to a reduction of the effective K. On a related note, while our model considered K as a fixed parameter, it could easily be extended to allow for system loss and duplication events.

2.6 Technical Details

2.6.1 Interpretation of Model Parameters K and L

In the toolbox model, L is the dimension of the phenotypic trait space. Collectively, the trait dimensions correspond to a list of characteristics that can make a given system suited better or worse in different environmental conditions. This list is potentially infinite! Therefore, although L appears as the dimensionality of some relevant trait space, a better interpretation is that L, with $L \gg K$, effectively sets the magnitude of fitness effects; i.e., mutations in an initially random genotype matrix have fitness effects on the order of $\sim 1/L$.

As stated in the main text, K is the number of systems $\vec{g}_\mu$ encoded in a genotype. Because system vectors are used to fit the conditions a target environment $\vec{E}$ sets on L traits, K can be understood as the dimensionality of phenotype space a genotype can access. In this sense, the impact of increasing K is two-fold. (1) Clearly, from eq. (2.2), a larger K would in general allow for greater fitness since more basis vectors can better fit a target vector, unless K is large enough to already ensure a perfect fit. We therefore focus on the $K \ll L$ regime for the sake of biological plausibility. Possessing more adjustable "knobs" further means that a genotype with larger K has more adaptive opportunities available to gain fitness. (2) By the same token, larger K also means greater flexibility across different environments (multiple L-dimensional targets) due to possessing more available degrees of freedom for tuning. This second effect (phenotypic plasticity) is relevant when

studying performance of the same genotype across multiple environments [72]. However, consequence (1) of larger K is more relevant to the present work's focus on the cost to evolvability that comes with a declining effective K.

Together then, KL reflects the dimensionality of a relevant genotype space in the toolbox model. We stress, however, that the entries in the K-by-L genotype matrix do not correspond to individual loci in a genome, but represent key traits that many loci may contribute towards [87]. For instance, the genotype matrix could represent patterns of trait co-regulation, with each system vector $\vec{g}_\mu$ corresponding to a master regulator of the pathways involved in the L traits. The vector entries then indicate which pathways the master regulatory system does or does not co-regulate. Involving many constituent proteins in general, each individual pathway is of course encoded in more than a single locus. On top of this, additional loci are devoted to the genetic encoding of the master regulators themselves. Therefore, the full genotype space is much larger than KL. In summary, for the purposes of this work, the genotypic space consists of: a subset of loci that we ignore because they do not evolve at all (e.g., membrane synthesis), a subset whose evolution we do not explicitly consider as we assume it occurs on a shorter timescale (master regulators/usage coefficients), and a focal subset that contribute to the key traits of our system vectors $\{\vec{g}_\mu\}$ and evolves on a relevant timescale of interest.

2.6.2 Modeling the Evolutionary Process

For simplicity in simulating the process of evolution, we work in the regime where mutations are rare and selection is strong. In doing so, we only need to track the evolutionary trajectory of a single genotype matrix, representing a clonal population that evolves by sequential mutations that sweep through the entire population as schematized in Figure 2.5. Computationally, we implement a Gillespie-style algorithm, where each loop iteration updates the genotype matrix by mutation

and selection. Having chosen the genotype matrix to be binary, we can implement point mutations as simple bit-flips of one of the matrix elements. Within each step of the algorithm, all point mutations of the current genotype are enumerated. The fitness effect of each mutation relative to the current genotype fitness is computed using eq. (2.2) (reoptimizing the usage coefficients for each mutant). From only the beneficial mutations, one is randomly selected to fix for the next iteration, with the probability of being selected weighted by fitness effect (strong selection, rare mutation regime; [73,74]). Finally, we also update the state of the environment within the Gillespie loop if the environment target vector is dynamic (e.g., eq. (2.3)). To update the environment in a semi-smooth fashion (even if the next mutation has yet to occur), we include an "environment update" event that occurs at a rate comparable to the typical timescale of a fixation event.

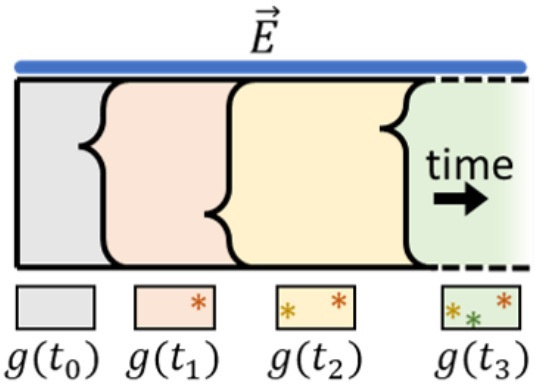

Figure 2.5 | Evolution in strong selection, rare mutation regime.

2.6.3 Constructing Environment Pairs

In this work, we studied evolution of genotypes after switching from one environment to another: An initially random genotype was first computationally evolved in an environment $\vec{E}_1$ until no beneficial mutations remain (highly adapted to $\vec{E}_1$). We then either directly or slowly switch exposure to environment $\vec{E}_2$. Although there are many features of environment pairs that may

matter for an evolving genotype, here, for simplicity, we focus on characterizing each environment pair $(\vec{E}_1, \vec{E}_2)$ by the Euclidean distance between them, $\Delta E = \left\|\vec{E}_2 - \vec{E}_1\right\|$.

To construct a random pair with specified ΔE, we generate two L-dimensional random vectors $(\vec{E}_A, \vec{E}_B)$ from a normal distribution with $\mu = 1, \sigma^2 = 1$ and rotate these vectors towards or away from each other (Figure 2.6A), similar to the approach of Ref. [72]. Specifically, the desired ΔE is attained by rotating the two random vectors away from their arithmetic mean $\vec{E} \equiv \frac{\vec{E}_A + \vec{E}_B}{2}$, according to the following parameterization:

$$\begin{aligned} \vec{E}_1(\delta) &= \text{ normalize}\left[\max\left(\vec{E} + \frac{\delta}{2}(\vec{E}_A - \vec{E}_B), 0\right)\right] \\ \vec{E}_2(\delta) &= \text{ normalize}\left[\max\left(\vec{E} - \frac{\delta}{2}(\vec{E}_A - \vec{E}_B), 0\right)\right] \end{aligned} \tag{2.4}$$

where the "normalize" operation normalizes a vector to unit length and $\max(\ldots,0)$ acts component-wise to ensure that each component is nonnegative. Eq. (2.3) thus parametrically defines a function $\Delta E(\delta)$ that can be inverted for obtaining a random pair of environments $(\vec{E}_1, \vec{E}_2)$ a given ΔE apart. By construction, both vectors $\vec{E}_1$ and $\vec{E}_2$ obtained in this way have unit length and only nonnegative components.

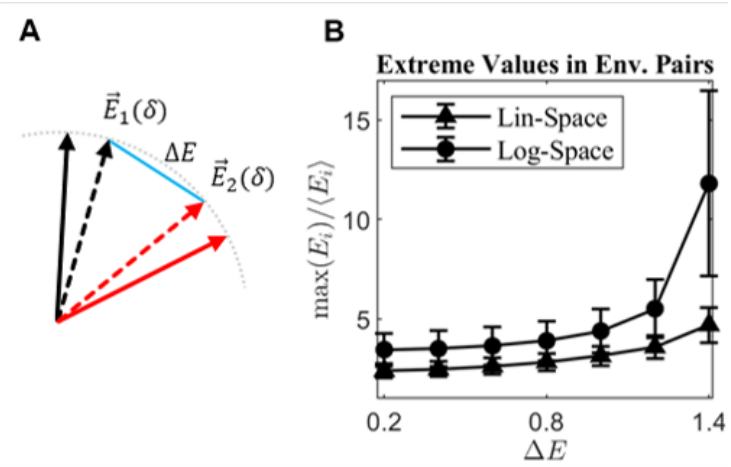

Figure 2.6 | (A) Schematic of environment pair construction: Two L-dimensional vectors are randomly drawn and any negative entries are capped at 0 (red and black solid arrows). The two random vectors are then rotated (parameterized by eq. (2.4) until the desired ΔE is obtained (red and black dashed arrows). (B) Random environment pairs were constructed using either eq. (2.4) or (2.5) over a range of ΔE. For each environment, the maximum component is divided by the average of the components and plotted against ΔE. Error bars correspond to ± 1 standard deviation over 100 replicates. For $\Delta E > 1$, the log-space construction has extreme values that rapidly grow with increasing ΔE, which is precisely the region of interest for this work.

Note that this is slightly different from the precise approach adopted in Ref. [72]. In that work, the rotation was performed in log-space: $\vec{E}_{1,2}(\delta) = \text{normalize}\left[\vec{E}'_{1,2}(\delta)\right]$, where

$$\log \vec{E}'_1(\delta) = \log \vec{E} + \frac{\delta}{2}\left(\log \vec{E}_A - \log \vec{E}_B\right)$$
$$\log \vec{E}'_2(\delta) = \log \vec{E} - \frac{\delta}{2}\left(\log \vec{E}_A - \log \vec{E}_B\right) \tag{2.5}$$

and logarithms are applied component-wise. In Ref. [72], the protocol Eq. (2.5) was adopted as the simplest approach that naturally preserved nonnegativity of vector entries, without the need for explicit truncation. However, for large ΔE, environment pairs constructed in log-space will typically possess extremely large entries (see Figure 2.6B) that focus the majority of selection pressure on a few traits. Since much of our attention in this work concerns the large-ΔE regime

(see, e.g., Figure 2.4B,C above), we opted for the linear-space construction of environment pairs, as defined in eq. (2.4).

2.6.4 Raw versus Gaussian Smoothed sK_0 and sK_0/K

Figure 2.4B&D traces a long-term evolutionary effect – namely, the relative fitness gain $\Delta\widetilde{F}\,(t^*)$ that a slow exposure (SE) protocol provides compared to direct exposure (DE) to a novel environment – to the early-time property of feedback loop strength, s. As described in the main text, the difference in number of active systems ΔK_{act} between SE and DE is predicted by sK_0, where K_0 is the number of inactive systems at $\mathrm{t} = 0$ (the time at which the environment switches from $\vec{E}_1$ to $\vec{E}_2$ for the DE protocol). In turn, we reasoned $\Delta\widetilde{F}\,(t^*) \sim sK_0/K$ for sufficiently late observation time t^*. Figure 2.7A,B provide heatmaps of the raw sK_0 and sK_0/K values, respectively, for each $(K, \Delta E)$ parameter combination (average over 300 trajectories), from which the contour lines used in Figure 2.4B,C were obtained after smoothing with a Gaussian kernel (of width equivalent to 1 heatmap pixel).

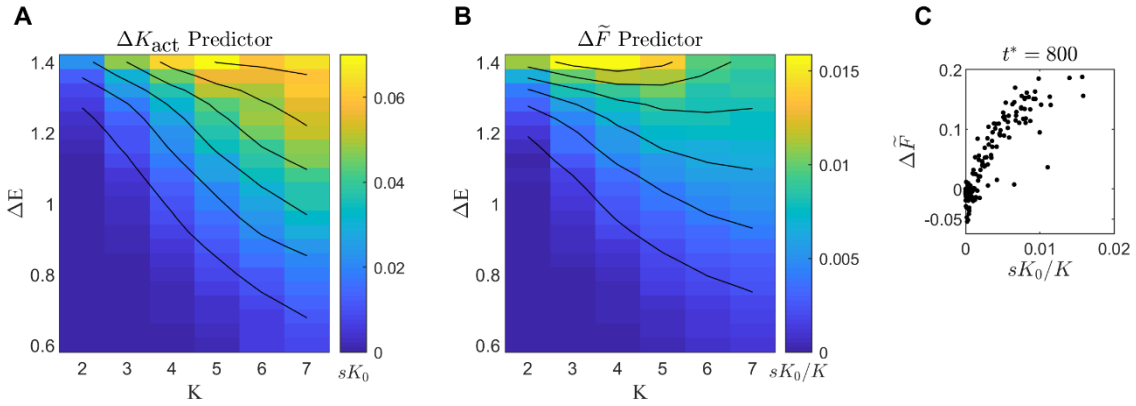

Figure 2.7 | (A-B) Heatmap of feedback strength scaled by the number or fraction of inactive systems during early-times of evolution (first 3 mutations), sK_0 in A and sK_0/K in B, with varying number of systems K and degree of environment change ΔE. The contour lines of Gaussian-smoothed sK_0 and sK_0/K (filter width $\sigma = 1$) overlaid on top as done in Figure 2.4B,C. (C) Relative fitness gain $\Delta\widetilde{F}$, as defined above, measured at $t^* = 800$ (2-times later than the observation time used above) and scattered against the sK_0/K data from (A). Each data point is an average over 300 trajectories.

2.6.5 Dependence of Fitness Improvement on Observation Time

The $\Delta\widetilde{F}\,(t^*)$ reported in panels B and E of Figure 2.4 were measured at $t^* = 400$. Figure 2.7C replots the same results for $t^* = 800$, demonstrating that the observation of Figure 2.4 is not sensitive to the particular choice of the late-time observation point.

Chapter 3: Coarse-grainability in a Model of Structured Ecosystems

Now that we have seen how evolution can generate structure and atypicality and the evolutionary consequences of this, I will next present a framework for using evolutionary dynamics to generate ensembles of random, structured ecosystems.

Despite their complexity, microbial ecosystems appear to be at least partially "coarse-grainable" in that some properties of interest can be adequately described by effective models of dimension much smaller than the number of interacting lineages. This is especially puzzling since recent studies demonstrate that a surprising amount of functionally relevant diversity is present at all levels of resolution, down to strains differing by 100 nucleotides or fewer. Rigorously defining coarse-grainability and understanding the conditions for its emergence is of critical importance for understanding microbial ecosystems. To begin addressing these questions, we propose a minimal model for investigating hierarchically structured ecosystems within the framework of resource competition. We use our model to operationally define coarse-graining quality based on reproducibility of the outcomes of a specified experiment and show that a coarse-graining can be operationally valid despite grouping together functionally diverse strains. Further, we demonstrate that a high diversity of strains (while nominally more complex) may in fact facilitate coarse-grainability, and that, at least within our model, coarse-grainability is maximized when a community is assembled in its "native" environment. Our modeling framework offers a path towards building a theoretical understanding of which ecosystem properties, and in which environmental conditions, might be predictable by coarse-grained models.

The work presented in this chapter has been adapted from the following publication:

Moran J, Tikhonov M. “Defining coarse-grainability in a model of structured microbial ecosystems.” Physical Review X 12, 021038 (2022).

3.1 Introduction

Microbial communities are complex dynamical systems composed of a highly diverse collection of interacting species, and yet they often appear to be at least partially “coarse-grainable”, meaning that some properties of interest can be predicted by effective models of dimension much smaller than the number of interacting lineages. For example, industrial bioreactors consisting of hundreds of species are well described by models with $\lesssim 10$ functional classes [45,46]. What makes this possible? One potential explanation is that coarse-grainability is a direct consequence of the hierarchically structured trait distribution across organisms. If 100 interacting phenotypes are all close variants of only 10 species, which can be further grouped into just two families, it is natural to expect that the diverse community might be approximately described by a 2- or 10-dimensional model. Under this view, effective models are possible because ecosystems are less diverse than a naïve counting of microscopic strains might suggest.

However, recent data reveal this intuition to be too simplistic: a surprising extent of relevant diversity persists at all levels of resolution. Numerous studies have highlighted the role of strain-level variation in shaping the functional repertoire of a microbial population [49–54]. A recent work by Goyal et al. concludes that strains might indeed be “the relevant unit of interaction and dynamics in microbiomes, not merely a descriptive detail” [47]. Surprisingly, however, a greater strain diversity can sometimes enhance predictability instead of undermining it [88]. Equally puzzling, the notion of a bacterial species is undoubtedly useful, despite collapsing together strains that famously may collectively share only 20\% of their genes [89]. Moreover, by some

assessments, the species-level characterization of a community appears to be too *detailed* and can be coarse-grained further [90,91], e.g., to the level of a taxonomic family [9].

Rigorously defining "coarse-grainability" and understanding the conditions for its emergence is of critical importance: harnessing coarse-grainability is our main instrument for understanding, predicting or controlling the behavior of these complex systems. Can an ecosystem be coarse-grainable for some purposes but not others? Or in some environments but not others? Can we ever expect the coarse-grained descriptions derived in the simplified environment of a laboratory to generalize to the complex natural conditions? Addressing this exciting set of general questions is an important challenge at the interface of theoretical microbial ecology and statistical physics.

Here, we introduce a theoretical framework to begin addressing these questions. The novelty of our approach is two-fold. First, we propose a minimal model for investigating structured ecosystems. Much recent work studies the behavior of large microbial ecosystems in the unstructured regime, where the traits of interacting organisms are drawn randomly (e.g., [6,7,10,12,13]). However, real ecosystems assemble from pools of taxa whose trait distributions are highly non-random due to functional constraints, common selection pressures, or common descent. These factors create structure at all levels, from the distribution of genes across strains in microbial pangenomes [14–16] to the distribution of function across taxa [17,43,90,91], with important implications for dynamics, patterns of coexistence, or responses to perturbations [18–20,44]. In natural communities, taxa can often be grouped by identifiable functional roles, often represented by closely related species or strains. As we seek to define and characterize ecosystem coarse-grainability, it seems clear that this structure must play an important role. Our model implements such structure within a consumer-resource framework in a simple, principled way through trait interactions.

The second novelty of our approach is a framework for defining and evaluating a hierarchy of coarse-grained descriptions. The ultimate performance criterion for a coarse-graining scheme would be its ability to serve as a basis for a predictive model, capable of predicting ecosystem dynamics or properties. However, finding the 'most predictive model' is a difficult problem. Here, as a simpler first step, we propose an operational approach which is inspired by the experiments of Ref. [9] and is based on the reproducibility of experimental outcomes. Specifically, we focus on a particular form of coarse-graining in which taxa are grouped together into putative functional groups. Grouping means omitting details, and we say that details are safe to ignore if they do not change the outcome of some specified experiment. Importantly, as we will show, choosing different experiments changes which, or whether, details can be ignored.

Specifically, we will define how ecosystems can be coarse-grainable in the weak sense, where a desired performance of a coarse-graining can be achieved in a given environment, and in the strong sense, where the performance of a given coarse-graining is *maintained* even as environment complexity is increased. We will demonstrate that the same ecosystem can be coarse-grainable under one criterion – even in the strong sense – and not at all coarse-grainable under another. This will reconcile the apparent paradox mentioned above, showing that a coarse-graining can be operationally valid for some purposes, despite grouping together functionally diverse strains. We will explain how strong-sense coarse-grainability arises in the model considered here, and show that this property is context-specific: a coarse-graining that works in the organisms' natural eco-evolutionary context is easily broken if the community is assembled in the non-native environment or if the natural ecological diversity is removed. Finally, we will discuss the extent to which our findings generalize beyond our model.

3.2 An Eco-evolutionary Framework for a Hierarchical Description of the Interacting Phenotypes

In order to study the hierarchy of possible coarse-graining schemes for ecosystems, we need an eco-evolutionary framework that would describe players functionally, by a list of characteristics that can be made longer (more detailed) or shorter (more coarse-grained). In addition, for our purposes we will also want an ability to tune the complexity of the environment, for example, to study the robustness of a coarse-graining between the simplified conditions of a laboratory and the more complex natural environment. In this section, we present our model implementing these two requirements.

3.2.1 The Eco-evolutionary Dynamics

A given environment presents various opportunities that organisms can exploit to gain a competitive advantage. Imagine a world where all such opportunities or "niches" are enumerated with index $i \in \{1 \dots L_\infty\}$. The notation L_∞ highlights that in general, one expects this to be a very large number, corresponding to a complete (and, in practice, unattainable) microscopic description. A strain μ is phenotypically described by enumerating which of these opportunities it exploits, i.e. by a string of numbers of length L_∞ which we will denote $\sigma_{\mu i}$. For simplicity, we will assume $\sigma_{\mu i}$ to be binary ($\sigma_{\mu i} \in \{0,1\}$): strain μ either can or cannot benefit from opportunity i. This will allow us to think of evolution as acting via bit flips $0 \mapsto 1$ and $1 \mapsto 0$, corresponding to the acquisition or loss of the relevant machinery ("trait i") via horizontal gene transfer events or loss-of-function mutations.

We will assume that the fitness benefit from carrying trait i is largest when the opportunity is unexploited, and declines as the competition increases. For a given set of phenotypes present in the community, the ecological dynamics are determined by the feedback between strain abundance

and opportunity exploitation (Figure 3.1A). Briefly, the strain abundances N_μ determine the total exploitation level $T_i \equiv \sum_\mu N_\mu \sigma_{\mu i}$ of opportunity i. The exploitation level determines the fitness benefit $h_i \equiv h_i(T_i)$ from carrying the respective trait; we will choose $h_i(T_i)$ of the form $h_i(T_i) = \frac{b_i}{1+T_i/K_i}$. These h_i, in turn, determine the growth or decline of the strains. Specifically, we postulate the following ecological dynamics:

$$\frac{\dot{N}_\mu}{N_\mu} = \sum_i \sigma_{\mu i} h_i - \chi_\mu \text{ , strain abundance} \tag{3.1a}$$

$$h_i = h_i(T_i) \equiv \frac{b_i}{1 + T_i/K_i} \text{, benefit from } i \tag{3.1b}$$

$$T_i \equiv \sum_\mu N_\mu \sigma_{\mu i} \text{ , exploitation of } i. \tag{3.1c}$$

In these equations, the parameters b_i and K_i describe the environment, with b_i being the fitness benefit of being the first to discover the opportunity i (at zero exploitation $T_i = 0$), and the "carrying capacity" K_i describing how quickly the benefit declines as the exploitation level T_i increases (Figure 3.1B). The quantities χ_μ are interpreted as the "maintenance cost" of being an organism carrying a given set of traits; more on this below.

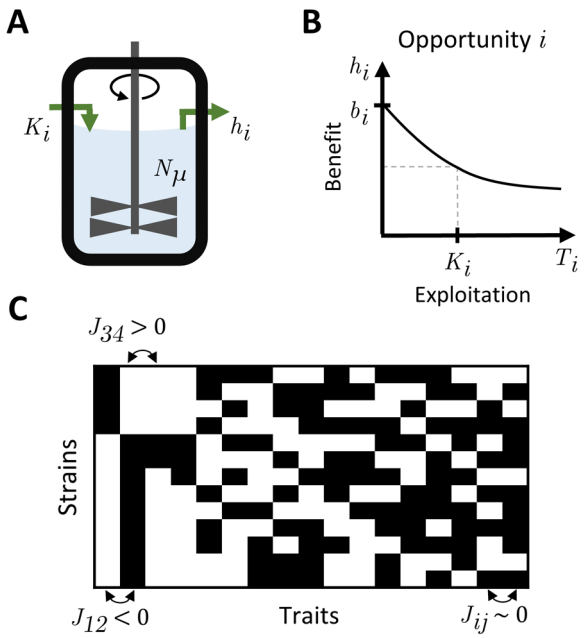

Figure 3.1 | Our eco-evolutionary framework modifies a standard model of resource competition. Organisms engage in ecological competition for limited resources and evolve by gaining or losing traits. Carrying a trait incurs a cost but enables the organism to benefit from the corresponding resource. Here, our novelty is to consider how traits interact with each other. Combinations that interact unfavorably are costly to maintain; as a result, not all phenotypes are competitive. **A:** A metabolic interpretation of our model corresponds to an ecosystem in a chemostat. A set of strains with abundances N_μ compete for a set of substitutable resources indexed by i, e.g., alternative sources of carbon. In this interpretation, K_i correspond to resource supply rates, and h_i are the resource concentrations in the effluent. **B:** For this work, we adopt a more general interpretation where the resources i need not be specifically metabolic. Instead, we think of i as enumerating any *depletable* environmental opportunities that the phenotypes can exploit, which confer a benefit h_i that declines with exploitation level T_i. We parameterize this dependence by the maximum benefit b_i and the carrying capacity K_i (the exploitation level where the benefit is halved); see text. **C:** In our model, phenotypes are binary vectors described by traits they carry. The most competitive phenotypes (rows in the cartoon) are not random, but are shaped by pairwise trait interactions J_{ij}. Strongly synergistic traits ($J_{ij} > 0$) tend to co-occur, while strongly antagonistic traits ($J_{ij} < 0$) are likely not carried together. Such structured phenotypes lead to structured ecosystems, as we investigate.

The dynamics (3.1) is basically the MacArthur model of competition for L_∞ substitutable "resources" [2,3,92]. To these dynamics we add the stochastic arrival of new phenotypes arising through bit flips ("mutations"), as is standard in studies of adaptive dynamics. The combined eco-evolutionary process is simulated using a hybrid discrete-continuous method as described in

Section 3.6.1. As presented so far, our eco-evolutionary model is similar to, e.g., Ref. [93]; our key novelty (trait interactions) will be introduced in the next section. We note, however, that typically the interpretation of resources in models like (3.1) is metabolic [8,10,13,29,94,95]; for example, i might label the different forms of carbon available to a carbon-limited microbial community. Here, we adopt a more general perspective, where i labels any depletable environmental opportunity, which need not be specifically metabolic.

As an example, one way for a strain to survive in chemostat conditions is to develop an ability to adhere to the walls of the device [96]. The wall surface is finite, and provides an example of a non-metabolic limited resource. Similarly, being physically bigger, or carrying a rare toxin could be a useful survival strategy, but in both cases the benefit decreases as the trait becomes widespread in the community. Unlike the forms of carbon, which may be numerous but are certainly countable and finite, the list of exploitable opportunities of this kind could be arbitrarily long ($L_\infty \to \infty$), especially when considering the complexity of natural microbial environments. Note that, by construction, our model allows coexistence of a very large number of phenotypes. In many studies, explaining such coexistence is the aim; here, it is our starting point. Rather than asking how a given environment enables coexistence of a diverse community, we start from the observation that natural communities are extremely diverse, interpret this as evidence for the existence of a very large number of (potentially unknown) limiting factors, and ask whether such diversity of types can be usefully coarse-grained.

Modeling fitness benefits as additive [eq. 3.1a] is certainly a simplification. It is also worth noting that the model (3.1) is special in that it possesses a Lyapunov function [97]; we will return to this point below. Nevertheless, this is a good starting step for our program, namely understanding the

circumstances under which coarse-grained descriptions are adequate. Most crucially, a suitable choice of the cost model χ_μ will allow us to naturally obtain communities with an hierarchical structure of trait distributions across organisms mimicking that of natural biodiversity.

3.2.2 A Simple Cost Model Leads to Hierarchically Structured Communities

Several studies investigated dynamics like (3.1) with costs assigned randomly (e.g., [7,8,10–13]). Here, we seek to build a model where the phenotypes in the community are not random, but are hierarchically structured, reproducing phenomena such as divergent taxa belonging to identifiable functional groups, the fine-scale strain diversity found within a species, or the notion of "core" and "accessory" traits in a bacterial pangenome [98]. For this, consider the following cost structure:

$$\chi_\mu = c + \sum_i \chi_i \sigma_{\mu i} - \sum_{i<j} J_{ij} \sigma_{\mu i} \sigma_{\mu j}. \tag{3.2}$$

The parameter c encodes a baseline cost of essential housekeeping functions (e.g., DNA replication). χ_i is the cost of carrying trait i (e.g., synthesizing the relevant machinery); for most of our discussion, we will set $c = 0.1$, and set all $\chi_i \equiv \chi_0 = 0.5$ for simplicity. The key object for us is the matrix J_{ij}, which encodes interactions between traits and shapes the pool of viable (low-cost) phenotypes (Figure 3.1C). As an example, the enzyme nitrogenase is inactivated by oxygen, so running nitrogen fixation and oxygen respiration in the same cell would require expensive infrastructure for compartmentalizing the two processes from each other; in our model, this would correspond to a strongly negative J_{ij} (carrying both traits is costly). An example for the opposite case of a beneficial interaction (positive J_{ij}) is a branched catabolic pathway, where sharing enzymes to produce common intermediates reduces the cost relative to running the two branches independently. Crucially, in our model, the parameters c, χ_i and J_{ij} are the same for all organisms;

we will refer to them as encoding the "biochemistry" of our eco-evolutionary world.

We now make our key choice. To set J_{ij}, we generate a random matrix of progressively smaller elements, as illustrated in Figure 3.2A. Specifically, we will be drawing the element J_{ij} out of a Gaussian distribution with zero mean and standard deviation $J_0\, f(\max(i,j))$, with a sigmoid-shaped $f(n) = \frac{1}{1+\exp\left(\frac{n-n^*}{\delta}\right)}$ (see Figure 3.2B). Throughout this work, we set $J_0 = 0.2$, $n^* = 10$ and $\delta = 3$. As we will see, this choice for the interaction matrix J implements a hierarchically structured distribution of traits. Intuitively, since high-cost phenotypes are poor competitors, we can think of the interactions J_{ij} as determining the "sensible" trait associations. For strongly interacting traits only some combinations are competitive, resulting in traits that are mutually exclusive ($J_{ij} < 0$) or that frequently co-occur ($J_{ij} > 0$) in low-cost (viable) phenotypes (Figure 3.1C). In contrast, a weakly interacting trait can be gained, lost, or remain polymorphic, as dictated by the environment. An example might be a gene encoding a costly pump that enables the organism to live in otherwise inaccessible (toxin-laden) regions of the habitat. Such a trait is "weakly interacting" if the cost of running the pump does not significantly depend on the genetic background. As we will see, our model will naturally give rise to hierarchically structured sets of phenotypes that share some "core" functions but differ in others to form finer-scale diversity, resembling the notions of "core" and "accessory" traits of a bacterial pangenome [98].

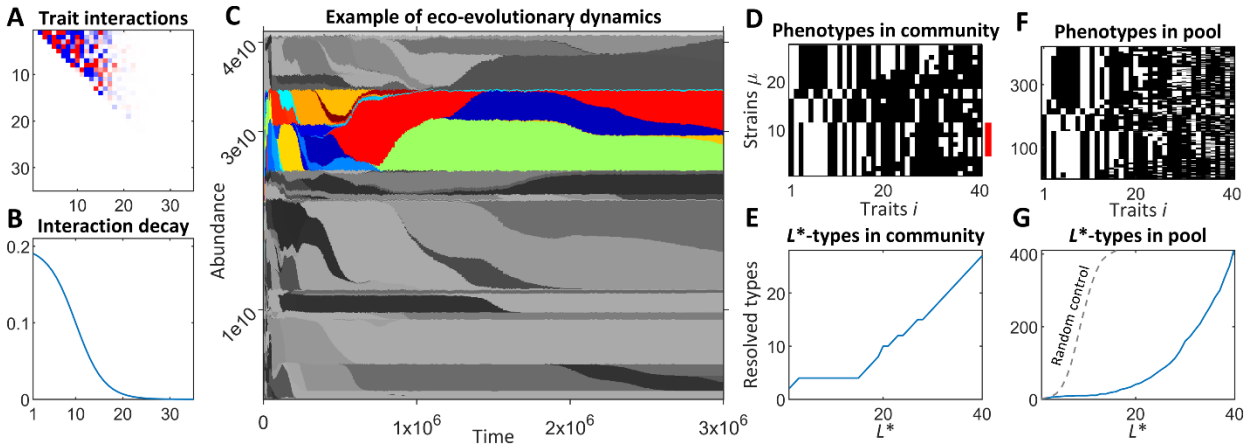

Figure 3.2 | A simple model of trait interactions leads to hierarchically structured ecosystems. **A, B:** In our model, the traits carried by a given phenotype interact with each other to determine its "maintenance cost" (see text). The matrix of pairwise trait interactions J_{ij} is drawn randomly and is the same for all phenotypes, encoding the "biochemical constraints"; panel A shows an example (J_{ij} is triangular with one element per trait pair $i \neq j$). We assume an interaction structure such that a few traits interact strongly while others interact weaker and weaker (panel B). **C:** An example of eco-evolutionary dynamics generated in our model. Shading corresponds to different phenotypes. Although new strains continue to emerge and die out throughout the period shown, they can be grouped into several coarse-grained types of approximately stable abundance (one is highlighted in color). **D:** The phenotypes present at the endpoint of the trajectory shown in C. Each of 27 phenotypes is a row of length $L_\infty = 40$ (white pixels are carried traits). The seven highlighted strains are identical in traits 1-24. We will say that they belong to the same L^*-type", for level of coarse-graining $L^* = 24$. **E:** The number of L^*-types in the community of panel D, shown as a function of L^*. At a coarse-grained level, the community appears to consist of only 4 types (one of these is highlighted in C using color); resolving finer substructure requires $L^* > 15$. **F, G:** Same as D, E for a broader set of strains, pooled over $M_{\text{env}} = 50$ similar environments. The hierarchical structure is maintained (if the trait matrix were randomized, the number of L^*-types would grow exponentially; see the dashed line). Here, we ask: in what sense, if any, could the phenotypic details beyond $L^* \approx 20$-25 be coarse-grained away in this model?

3.2.3 Environment Defines a Strain Pool

To build some intuition about the model defined above, consider Figure 3.2C that shows an example of these eco-evolutionary dynamics for one random biochemistry, and an environment where we set $b_i \equiv b_0 = 1$ for simplicity, and $K_i = K_0 = 10^{10}$ to set the scale of population size as appropriate for bacteria. Grayscale shading corresponds to distinct phenotypes; the community was initialized with a single (randomly drawn) phenotype. The dynamics of Figure 3.2C illustrate that our framework will allow us to define a form of ecosystem stability where all the original

phenotypes may have gone extinct and were replaced by others, and yet at a coarse-grained level the ecosystem structure remains recognizably "the same". Here, starting from about $t \simeq 10^5$, the dynamics resemble a stable coexistence of several coarse-grained "species" (one is highlighted in color), whose overall abundance remains roughly stable even as individual strains continue to emerge and die out. To formalize this observation, we need the notion of coarse-grained "L^*-types", which we will now introduce.

As we continue the simulation, the dynamics converge to an eco-evolutionary equilibrium (a state where the coexisting types are in ecological equilibrium, and no single-bit-flip mutant can invade). In this example, it consists of 27 coexisting phenotypes and is shown in Figure 3.2D. Note that, confirming our expectations, it appears to possess a hierarchical structure. The seven highlighted strains are identical over the first 24 components, and differ only in the "tail" (components 25-40). A coarse-grained description that characterizes organisms only by the first $L^* = 24$ traits would be unable to distinguish these strains; we will say that these strains belong to the same L^*-type with $L^* = 24$. Figure 3.2E plots the number of L^*-types resolved at different levels of coarse-graining L^* (within the community shown in Figure 3.2D). For $L^* = 3\text{-}15$, the number of types remains stable at just 4; the color in Figure 3.2C highlights one of them. Beyond $L^* = 15$, adding more details begins to resolve additional types, up until $L^* = L_\infty$ when the number of L^*-types coincides with the total number of microscopic strains.

Of course, when discussing the diversity of strains one expects to find in a given environment, it is important to remember that no real environment is exactly static, and no real community is in evolutionary equilibrium. To take this into account while keeping the model simple, we will consider not a single equilibrium, but a collection of communities assembled in $M_{\text{env}} = 50$ similar

environments where we randomly perturb the carrying capacity of all opportunities ($K_i = K_0(1 + \epsilon\eta_i)$, with $\epsilon = 0.1$ and η_i are i.i.d. from a standard Gaussian); see Section 3.6. Figure 3.2F shows the set of strains pooled over the 50 ecosystems assembled in this way. This *strain pool* is the central object we will seek to coarse-grain. We stress that its construction explicitly depends on the environment. (Or, more specifically, the particular random set of M_{env} similar environments, but $M_{\text{env}} = 50$ is large enough that the results we present are robust to their exact choice.)

As we see in Figure 3.2F, adding more strains to the pool makes its hierarchical structure even more apparent. Quantitatively, the number of L^*-types (Figure 3.2G) grows much slower than if the traits of each phenotype were randomly permuted (the dashed control curve): microscopically, perturbing the environment favors new strains, but at a coarse-grained level, these new strains are variations of the same few types. This is precisely the behavior that we were aiming to capture in our model. Beyond $L^* \approx 20\text{-}25$, the number of resolved types begins to grow rapidly. Can this diversity be coarse-grained away? Is there a precise sense in which these tail-end traits are "just details"? To answer this question, we must begin by making it quantitative.

3.3 Coarse-graining

3.3.1 Methodology for Defining Coarse-grainability

The L_∞-dimensional description we defined represents the complete list of niches and opportunities present in a natural habitat. Any recreation in the laboratory is simplified, retaining only some of the relevant factors. We will model simplified environments as including resources/opportunities 1 through L (Figure 3.3A). The parameter L represents environment complexity. The other key parameter is the level of coarse-graining detail, L^* (Figure 3.3B). For

each L^*, the identity and combined abundance of L^*-types provides a candidate coarse-grained description of the ecosystem. We seek a quantitative metric for assessing its quality.

Ideally, this assessment would be a comparison of performance of two models: one highly detailed, the other coarse-grained, and our test would evaluate the prediction error for a given property of interest. However, what we built is not a coarse-grained *model*, but a hierarchy of coarse-grained variables. These variables could be used to build any number of models, and identifying the most predictive of these is a highly nontrivial task. Here we sidestep this problem by proposing an operational approach that evaluates a coarse-graining based on the reproducibility of outcomes of a specified experimental protocol.

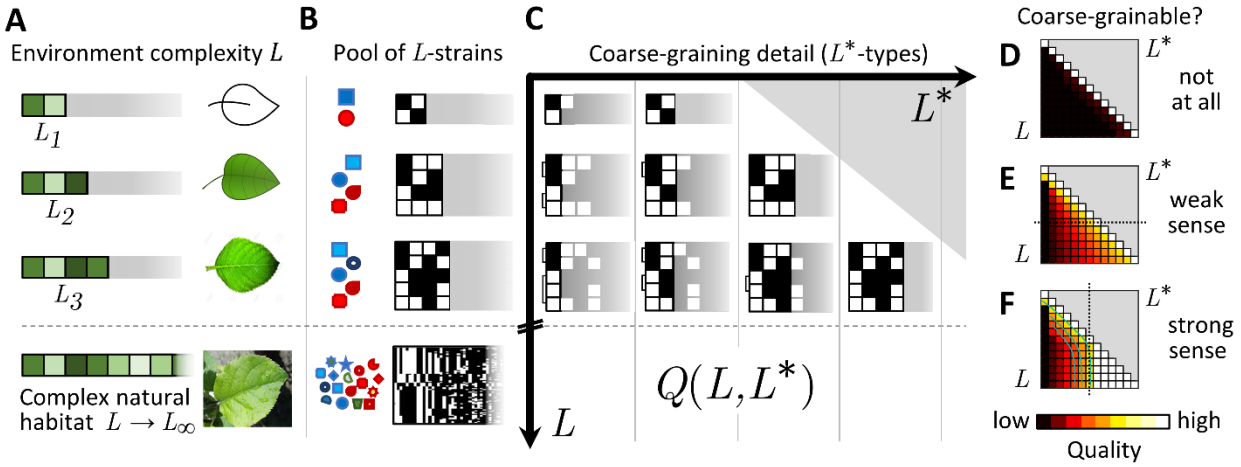

Figure 3.3 | Defining weak and strong coarse-grainability. A: The complex natural habitat is modeled as including a large number L_∞ of exploitable resources or opportunities. In a laboratory, we can consider a sequence of ever-more-detailed approximations including resources $1, \ldots, L$ (with the remaining ones set to zero). **B:** For each environment, the model describes the pool of strains we expect to encounter (the pool of L-strains; see Figure 3.2F). For a given L, the strains are unlikely to carry traits i for resources not provided ($i > L$). As environment complexity L increases, the pool becomes increasingly diverse. **C:** The set of L-strains can be coarse-grained to a varying level of detail $L^* \leq L$. Let $Q(L, L^*)$ be any quantitative metric (to be defined later) scoring the quality of the L^*-coarse-graining in the environment of complexity L. At $L^* = L$, the strain diversity is fully resolved (no coarse-graining). The "coarse-grainability" of the ecosystem is encoded in the behavior of $Q(L, L^*)$ for $L^* < L$. Different metrics Q encode different operational definitions of coarse-grainability. **D:** A non-coarse-grainable ecosystem (*sensu* quality metric Q). The coarse-graining quality remains poor unless the microscopic strain diversity is fully resolved (at $L^* = L$). **E:** Weak-sense coarse-grainability: in any given environment (a fixed L, highlighted),

a desired quality can be achieved with a coarser-than-microscopic description ($L^* < L$). **F:** Strong-sense coarse-grainability: the same coarse-graining (a fixed L^*, highlighted) provides the desired quality even as the environment complexity is increased.

We will describe and contrast two protocols, each of which could be seen as verifying the validity of the coarse-graining, and each yielding its own metric of coarse-graining quality $Q(L, L^*)$; Figure 3.3C. The "diagonal" entries of Q (with $L^* = L$) correspond to an absence of coarse-graining: the description of strains resolves *all* the traits relevant in a given environment. Coarse-grainability is encoded in the behavior of $Q(L, L^*)$ with $L^* < L$ (Figure 3.3D-F). Consider first the behavior of $Q(L, L^*)$ as a function of L^*, with L fixed. If we observe that in a given environment, sufficient quality can be achieved already with $L^* < L$, we will say that the ecosystem is coarse-grainable in the weak sense. For strong-sense coarse-grainability, we ask if the same coarse-grained description continues to perform well even as the environment is made more complex (i.e., instead of fixing L and varying L^*, we fix L^* and vary L). Strong-sense coarse-grainability would be a highly desirable property, but *a priori* it is unclear if it is even theoretically possible.

Crucially, these definitions depend on the choice of the operational criterion for assessing coarse-graining validity (the experiment whose results we require to be reproducible). Below, we will show that the same ecosystem can be coarse-grainable in the strong sense under one criterion, and yet not coarse-grainable at all under another.

3.3.2 Operational Definitions of Coarse-graining Quality $Q(L, L^*)$

In this section, we describe two "experimental" protocols, each of which could be seen as a sensible test of the quality of a coarse-graining. They will establish two alternative criteria for a coarse-graining to be operationally valid, which we will then contrast.

The Reconstitution Test

One possible criterion is the *reconstitution test*. Drawing a random representative for each of the L^*-types in the strain pool, we seed an identical environment with the representatives we chose, allowing them to reach an ecological equilibrium (Figure 3.4B). If the details ignored by the coarse-graining are indeed irrelevant, we would expect such "reconstituted" replicates to all be alike. If the reconstituted communities are found to be highly variable depending on exactly which representative we happened to pick, this will signal that the distinctions we attempted to ignore are, in fact, significant.

Quantitatively, for each L^*-type μ_*, let us denote $n_{\mu_*}^{(\alpha)}$ its final relative abundance (i.e., the fraction of total population size) in the reconstituted replicate α. The coefficient of variation of $n_{\mu_*}^{(\alpha)}$ over α (denoted $\mathrm{CV}_\alpha\left[n_{\mu_*}^{(\alpha)}\right]$) provides a natural measure of variability across replicates. To combine these into a single number, we compute the average such variability over all L^*-types μ_*, weighted by their mean relative abundance across replicates (denoted $\left\langle n_{\mu_*}^{(\alpha)} \right\rangle_\alpha$):

$$Q_{\mathrm{rec}} = \sum_{\mu_*} \left\langle n_{\mu_*}^{(\alpha)} \right\rangle_\alpha \mathrm{CV}_\alpha\left[n_{\mu_*}^{(\alpha)}\right].$$

Since the coefficient of variation is, by definition, $\mathrm{CV}_\alpha\left[n_{\mu_*}^{(\alpha)}\right] = \mathrm{std}_\alpha\left[n_{\mu_*}^{(\alpha)}\right] / \left\langle n_{\mu_*}^{(\alpha)} \right\rangle_\alpha$ our metric simplifies to $Q_{\mathrm{rec}} = \sum_{\mu_*} \mathrm{std}_\alpha\left[n_{\mu_*}^{(\alpha)}\right]$. With this definition, a perfect reconstitution would have $Q_{\mathrm{rec}} = 0$. Conveniently, this is automatically the case if $L^* = L$ (no coarse-graining).

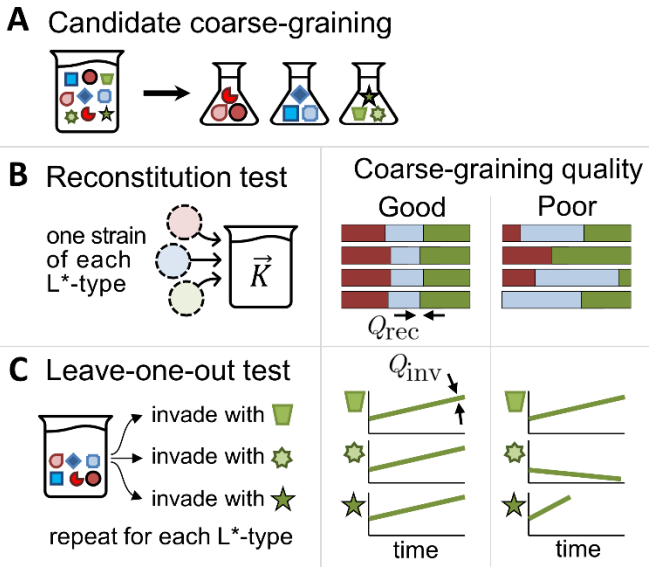

Figure 3.4 | Specific criteria for assessing coarse-graining quality $Q(L, L^*)$. A: In this cartoon, the community is coarse-grained into three operational taxonomic units (OTUs), implemented in our model as L^*-types. **B:** The "reconstitution test". Under this criterion, grouping strains into coarse-grained OTUs is justified if reconstituting a community from a single representative of each OTU yields similar communities regardless of which representatives we pick. As a quantitative measure, we compare the OTU abundances across replicates. **C:** The "leave-one-out test". Under this criterion, grouping strains into coarse-grained OTUs is justified if the strains constituting OTU X (green in this cartoon) all behave similarly when introduced into a community missing X. As a quantitative measure, we compare the invasion rates of the left-out strains.

The Leave-One-Out Test

As we will see, the criterion defined above is extremely stringent and is rarely satisfied. In this section, we introduce a weaker version. Instead of the composition of the entire community, we will explicitly focus on one particular property of interest (below, the invasion rate of a strain). Further, instead of requiring the grouped-together strains to be interchangeable in absolute terms, we will ask that they behave similarly *in the context of the assembled community*.

Specifically, for a given scheme grouping strains into coarse-grained types, consider assembling a community missing a particular coarse-grained type μ_* (the ecological equilibrium reached when combining all the strains in the pool, except those belonging to type μ_*; Figure 3.4C). We will judge the coarse-graining as valid if the different strains constituting the missing type μ_* all behave

similarly when introduced into this community. As one example, we can compare their initial growth rates if introduced into the community at low abundance, called henceforth "invasion rate" (other possible choices include the abundance the strain will reach if established, or the level of niche exploitation h_i in the resulting community; these are shown in the Figure 3.8). If the invasion rates are similar, describing the community as missing the coarse-grained type μ_* would indeed be consistent. If, however, the invasion rates vary strongly, we will conclude that the features our coarse-graining is neglecting are, in fact, important.

Quantitatively, denote the invasion rate of strain μ into a community missing type μ_* as r_{μ,μ_*}. We define

$$Q_{\text{inv}} = \sum_{\mu_*} \bar{n}_{\mu_*} \, \text{std}_{\mu \in \mu_*} r_{\mu,\mu_*},$$

where $\bar{n}_{\mu_*}$ is the relative mean abundance of strains belonging to type μ_* in the pool, and $\text{std}_{\mu \in \mu_*}$ denotes the standard deviation over all strains belonging to μ_* weighted by strain abundance in the pool (i.e., a strain's combined abundance observed across the set of M_{env} environments used to define the pool). Once again, at $L^* = L$ we automatically have $Q_{\text{inv}} = 0$ as this corresponds to the fully microscopic description (each type μ_* is represented by exactly one strain). Note that this averaging convention (weighted by abundance in the pool) is slightly different from that used in the previous section (using average abundance across the assembled replicates). Using the same convention for both Q_{inv} and Q_{rec} would not change our results, but would artificially inflate the latter with noise from low-abundance (rare) strains. (For details, see Section 3.6.3 and Figure 3.9.)

To illustrate the difference between the two criteria, consider the statement that a community consisting of *Tetrahymena thermophila* and *Chlamydomonas reinhardtii* cannot be invaded by

Escherichia coli [99]. What meaning should we ascribe to this statement when phrased in terms of coarse-grained units, rather than specific strains? Under the first criterion, we would require that if we combine any single strain of *T. thermophila*, any strain of *C. reinhardtii*, and any strain of *E. coli*, only the first two would survive. Under the second criterion, we would combine a vial labeled *T. thermophila*, containing the entire diverse ensemble of its strains, with a similarly diverse vial of *C. reinhardtii*, and verify that the resulting community cannot be invaded by any individual strain of *E. coli*.[1]

Note that in our model, the existence of a Lyapunov function [97] means the ecological equilibrium is uniquely determined by the environment and the identity of the competing strains; their initial abundance or the order of their introduction does not matter. While this is a simplification, this property is very useful for our purposes, since any lack of reproducibility between reconstituted communities is then clearly attributable to faulty coarse-graining. In a model where even identical phenotypes could assemble into multiple steady states, distinguishing this variability from the variability due to strain differences would add a layer of complexity to our analysis.

3.4 Results

3.4.1 A Coarse-graining may be Operationally Valid Despite Grouping Functionally Diverse Strains

Throughout this section, we will continue to use an environment with $K_i \equiv K_0$ and $b_i \equiv b_0$ (all L_∞ opportunities are equally lucrative). In practice, when approximating a complex environment in the laboratory, we try to capture the most salient features first. Thus, it would have been perfectly

[1] Although we use this as an example here, we should note that in the original reference [99] this statement was not actually meant as a species-level claim, but indeed referred to the three specific strains used in the experiment.

natural to instead let K_i and/or b_i decline with i; one would expect this to improve coarse-grainability, and this is indeed the case (see Section 3.6.5). The motivation for our choice is two-fold: First, keeping all K_i and b_i the same requires fewer parameters than choosing a particular functional form of decline with i. Second, the regime where no niches are obviously negligible will only make it more striking to find that an ecosystem can be not only coarse-grainable, but coarse-grainable in the strong sense.

Figure 3.5A plots $Q_{\text{inv}}(L, L^*)$ for the leave-one-out test comparing the invasion rates of different strains falling into the same coarse-grained types. We find that any desired coarse-graining quality can be achieved by a sufficient L^*, and is almost unaffected by L. As environment complexity increases and becomes capable of sustaining an ever-growing number of microscopic strains, each L^*-type becomes increasingly diverse. Nevertheless, all the strains in the same L^*-type continue to behave similarly by our invasion-rate-based metric; in other words, under this criterion, the ecosystem is coarse-grainable in the strong sense.

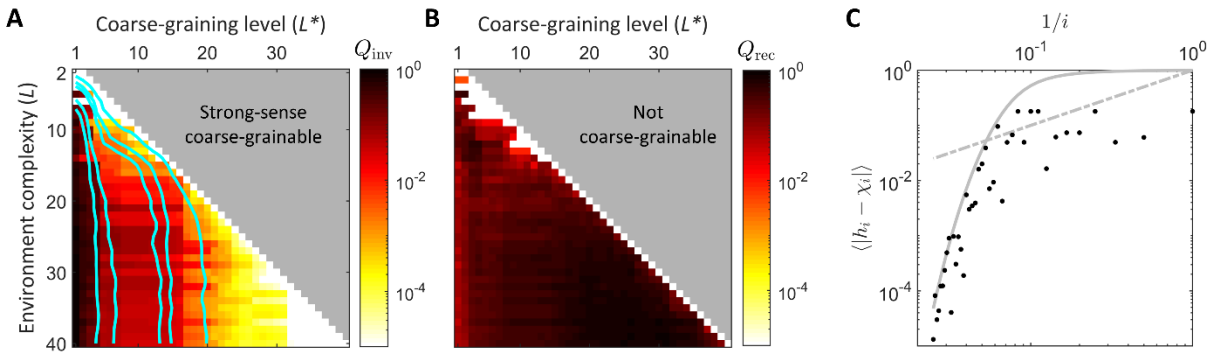

Figure 3.5 | The same ecosystem can be coarse-grainable under one criterion, but not under another. A: If coarse-graining quality is evaluated using the leave-one-out test (assessing reproducibility of strain invasion rates), our ecosystem model is coarse-grainable in the strong sense: the acceptable level of coarse-graining, determined by the desired quality score (isolines of Q), is robust to environment complexity (compare to Figure 3.3F). **B:** In contrast, under the reconstitution test criterion, no amount of coarse-graining is acceptable (compare to Figure 3.3D). This comparison shows that a coarse-graining can be operationally valid for a given purpose (panel A) even when the strains it groups together are functionally diverse (panel B). Both heatmaps represent a single random biochemistry, same in both panels. Isolines in A are averaged over 20

biochemistries to demonstrate robustness (see Section 3.6.2). **C:** Explaining the origin of strong-sense coarse-grainability in our model. The plot shows the scaling with i of $|h_i - \chi_i|$ (computed for $L^* = 30$, $L = 40$, and averaged across communities assembled for the leave-one-out test of panel A). The strong-sense coarse-grainability of panel A is ensured whenever the decay is faster than $1/i$ (dashed gray line). Intuitively, this makes the tail-end traits approximately neutral in the assembled community; see text. We expect this scaling to be controlled by the sigmoidal decay of trait interaction magnitude $|J_{ij}|$, as confirmed here (solid gray line; same as Figure 3.2B but normalized to a maximum of 1 to show the decay of interaction strength rather than their absolute magnitude).

And yet, it would be wrong to conclude that the traits beyond a given L^* are "negligible" in any absolute sense. This is clearly demonstrated by the reconstitution test (Figure 3.5B). If we attempt to reconstruct the community from its members, *every* detail matters: no amount of coarse-graining is acceptable. We will now explain this apparent paradox within our model.

Consider a community at an ecological equilibrium, and let us focus on a particular phenotype σ carrying one of the weakly interacting (tail-end) traits i_0: $\sigma_{i_0} = 1$. What would be the fitness effect of losing this trait? Losing the benefit h_{i_0} from opportunity i_0 is offset by the reduction in maintenance cost; for a weakly interacting trait, the contribution from the term $\sum_j J_{ji_0}\sigma_j\sigma_{i_0}$ is negligible, and the change in cost is simply χ_{i_0}. We conclude that the fitness effect of losing the trait is $\delta f = \chi_{i_0} - h_{i_0}$. At an evolutionary equilibrium, we would therefore have $h_{i_0} = \chi_{i_0}$ (the "functional attractor" state [8]). When this condition is satisfied, we will say that the opportunity or niche i_0 is "equilibrated". If a weakly interacting niche is equilibrated, carrying the respective trait becomes approximately neutral.

Here, our community is not at the evolutionary equilibrium; nevertheless, a sufficiently diverse strain pool will similarly ensure that the opportunities corresponding to the weakly-interacting (tail-end) traits become approximately equilibrated: $h_i \approx \chi_i$. For a simpler model where the

phenotype costs χ_μ are drawn randomly, the mechanism for this can be understood analytically (the "shielded phase" of Ref. [13] }; see also Ref. [100]). Here, the costs are not random, but as long as trait interactions are weak, one expects the behavior to be similar (see SI section S6.2 in Ref. [13]). This expectation is confirmed in simulations. Figure 3.5C shows the observed niche disequilibrium $h_i - \chi_i$ as a function of $1/i$. The plot confirms that the tail-end niches ($1/i \to 0$) are increasingly well-equilibrated ($|h_i - \chi_i|$ decays with i). The strong-sense coarse-grainability of Figure 3.5A is ensured whenever the decay is faster than $1/i$ (dashed gray line). This is because with this scaling, the sum of contributions from the omitted tail-end traits is bounded (see Section 3.6.5). The analytical argument of Ref. [13] leads us to expect the disequilibrium to be controlled by the decaying typical magnitude of interactions $|J_{ij}|$ (solid gray line). If the tail-end niches are equilibrated, carrying the respective traits becomes approximately neutral, and the ability of a strain to invade is entirely determined by its phenotypic profile over non-equilibrated niches, explaining the observations of Figure 3.5A. We conclude that in our model, the strong-sense coarse-grainability is a consequence of the faster-than-$1/i$ decay of interaction strength in Figure 3.2B.

Crucially, however, this approximate neutrality applies only in the environment created by the assembled community, and does not mean that the distinctions are functionally negligible. For instance, consider the (Lotka-Volterra-style) interaction term for a given pair of strains $\mu \neq \nu$:

$$A_{\mu\nu} \equiv \frac{1}{N_\mu}\frac{\partial \dot{N}_\mu}{\partial N_\nu} = \sum_i \sigma_{\mu i}\sigma_{\nu i}\frac{h_i^2}{b_i K_i} = \frac{\sum_i \sigma_{\mu i}\sigma_{\nu i}h_i^2}{b_0 K_0},$$

where we substituted $b_i \equiv b_0$ and $K_i \equiv K_0$ for our environment. Even when tail-end niches are equilibrated with $h_i \approx \chi_i = \chi_0$, we find that each of them contributes equally to the interaction term: no detail is negligible.

This argument directly relates the observed effect to the distinction between a trait that is truly neutral, and one that is effectively neutral in the assembled community only. A truly neutral trait, one incurring almost no cost and bringing almost no benefit, would have $h_i \to 0$ and its contribution to the interaction term $A_{\mu\nu}$ would indeed be small. And indeed, if we repeat our analysis for a scenario where both b_i and χ_i decline with i, we find that neglecting the tail-end traits becomes an adequate coarse-graining also for the reconstitution test (see Figure 3.10).

The conclusion from contrasting Figure 3.5A and Figure 3.5B is worth emphasizing. In the example we constructed, the coarse-grained description is valid *sensu* panel 3.5A. This means that, for instance, we can meaningfully say that "a community assembled of OTU#1 and OTU#2 can be invaded by OTU#4". We can even measure, e.g., the invasion rate, and be assured that it is quantitatively reproducible, with a bounded error bar, across the many strains that constitute OTU#4 at the microscopic level. Despite all this, the *interaction* between the OTUs as coarse-grained units is not actually definable: any specific pair of strains of OTU#1 and OTU#4 may interact differently with each other, as is indeed observed experimentally [47].

Our focus on reproducibility of L^*-type abundances across replicates is inspired by the experiments Ref. [9]. To complete this parallel, we should mention that besides inoculating the same environment with a set of similar inocula, as we did for our reconstitution test (*cf* Figure 3.4B), one could also use the same inoculum to seed a set of similar environments. To implement this in our model, we use the strain pool constructed as described in Section 3.2.3 to inoculate a set of

environments with slight variations in the carrying capacities $K_i \approx K_0$ drawn from a Gaussian distribution of width $\epsilon = 0.1$. This is meant to represent the unavoidable variability present in any experimental replicates of the "same" environment $K_i \approx K_0$, which can affect fitness even when subtle [101]. After assembling the replicate communities, we find that community composition is more reproducible at coarser levels of description (Figure 3.6B,C), consistent with the experimental observations of Goldford et al. [9], and with the interpretation of this pattern as resulting from functional redundancy within coarse-grained types [90,91,102].

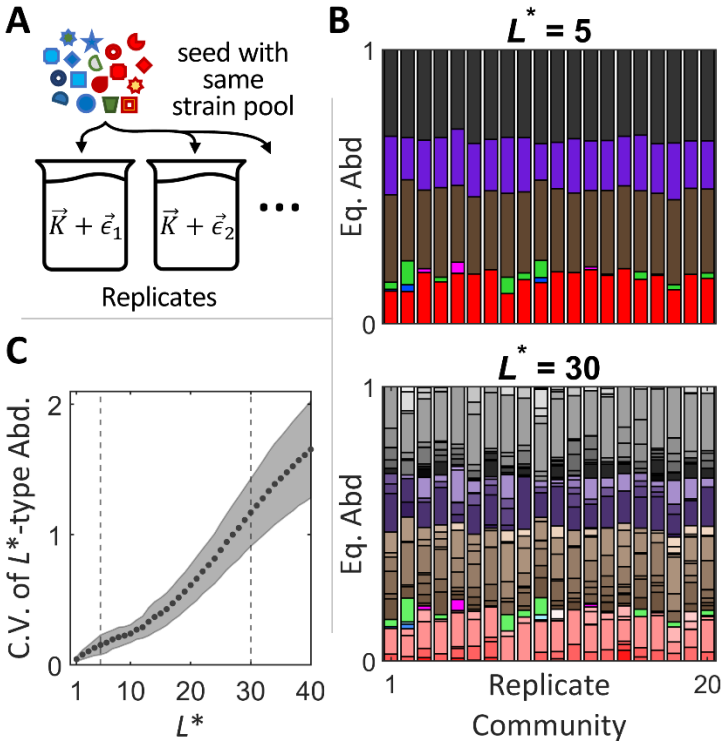

Figure 3.6 | Replicate communities assembled in similar environments are more reproducible at coarser level of description. A: A set of similar environments $\vec{K}+\vec{\epsilon}$ (each carrying capacity modified by 10\% Gaussian noise) is inoculated with the same strain pool, and brought to ecological equilibrium. **B:** Equilibrium relative abundances of coarse-grained L^*-types across 20 replicates, shown for two levels of coarse-graining. A coarser description ($L^* = 5$; 7 resolved types) is more reproducible, consistent with experimental observations [9]. **C:** The variability of coarse-grained descriptions increases with level of detail. Variability is measured as the average coefficient of variation (C.V.) in relative abundance of an L^*-type over 100 replicates, weighted by L^*-type mean relative abundance across replicates. Dashed lines mark $L^* = 5, 30$ shown in B. Datapoints and shading show mean $\pm$ SD over 20 random choices of biochemistry $\{J_{ij}\}$. All simulations performed with $L = 40$.

3.4.2 Using Non-native Strain Pool Reduces Coarse-grainability

The previous section describes a mechanism by which strain diversity can aid coarse-grainability. As we explained, in our model ecosystem the diverse set of strains contained within the coarse-grained units was able to successfully equilibrate the weakly interacting niches, rendering them effectively neutral and leading to the behavior shown in Figure 3.5A. However, for this to occur, the strain pool diversity needs to be derived from a sufficiently similar set of environments, as we will now show.

To see this, we repeat the leave-one-out analysis of Figure 3.5A, except now we inoculate the same test environment of complexity $L = 40$ (using $K_i = K_0$, $b_i = b_0$ as before) with strain pools derived from *other* environments that are increasingly dissimilar to it. Specifically, following the procedure described in Section 3.2.3, we generate strain pools in environments with $K_i = K_0(1 + \epsilon\eta_i)$, where η_i are drawn from the standard normal distribution, and ϵ is the parameter we vary. (The b_i are left at $b_i = b_0$ for simplicity.) The results are presented in Figure 3.7, which shows the performance of different L^*-coarse-grainings under the leave-one-out test.

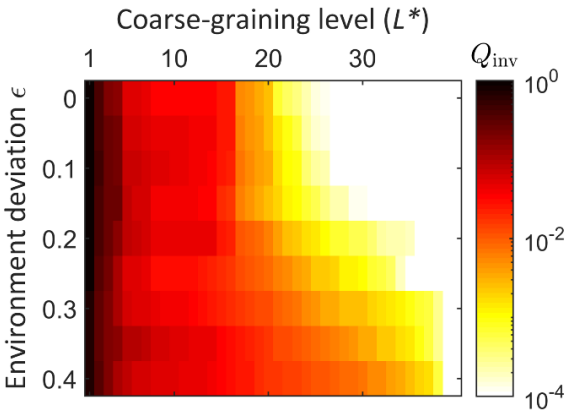

Figure 3.7 | A coarse-graining scheme works best when the environment is populated by the native strain pool. The same test environment as in Figure 3.5A is inoculated with strain pools that evolved in environments increasingly further away (see text). The coarse-graining quality is assessed by leave-one-out experiments, and shown as a function of L^* and environment deviation

ϵ from the test condition. L is fixed at $L = 40$ for comparison with the last row of Figure 3.5A. As the environments for generating strain pools are modified, the traits that were previously negligible can no longer be coarse-grained. The same random biochemistry as in Figure 3.5A was used, and each pixel is averaged over 20 random environments.

At $\epsilon = 0$, this is identical to the protocol of Figure 3.5A. We see that describing phenotypes by 20 traits is sufficient for the invasion rates of grouped-together strains to be consistent within an error bar of $Q < 10^{-2}$. However, as ϵ is increased, and the strain pools we use are derived from increasingly distant environments, the same coarse-graining becomes insufficient. Instead, a substantially higher level of coarse-graining detail L^* is required to maintain the desired quality. In summary, we find that in our model, a coarse-graining scheme works best when the environment is populated by the native strain pool.

3.5 Conclusions and Discussion

The interface of statistical physics and theoretical ecology has a long and highly influential tradition of studying large, random ecosystems, starting from the work of May [4]. The key insight of this approach is that patterns that are typical to some *ensemble* of ecosystems are more likely to be generalizable and reproducible than the details specific to any one realization. However, the choice of the ensemble (and in particular, adding constraints relevant for natural ecosystems) can affect predictions significantly [103–109]. Which predictions of random-interaction models are robust to introducing more realistic structures, and conversely, which phenomena cannot be explained without invoking structural constraints, is an active area of research [110].

Resource competition models – one of the simplest frameworks explicitly linking composition to function – offer a highly promising context to begin addressing these questions, with much recent progress. For example, it was recently shown that cross-feeding interactions structured by shared

"rules of metabolism" (but otherwise random) can reproduce a surprising range of experimental observations [9,11]. This work made it possible to begin disentangling which experimental observations can be seen as evidence for nontrivial underlying mechanisms, and which can be reproduced already in the simplest models.

In this work, we presented a simple framework that allows generating random ecosystems with community structure as a tunable control parameter. Instead of postulating a fixed architecture, such as a number of discrete "families" of phenotypes [11], we use a biologically motivated approach to derive it from functional tradeoffs, parameterized by a matrix of trait-trait interactions J. Simple (few-parameter) choices for J generate communities with complex structures, including hierarchical architectures which, at least superficially, appear to mimic those of natural biodiversity. Perhaps the most immediate benefit from such a framework would be to help develop new ways to quantify the highly multi-dimensional concept of "community structure" across scales, such as, for example, the structure of microbial pangenomes.

In this spirit, here we used this framework to quantify the notion of coarse-grainability. We proposed a way to operationally define the quality of a coarse-grained description based on the reproducibility of outcomes of a specified experiment. We demonstrated that an ecosystem can be coarse-grainable under one criterion, while also not at all coarse-grainable under another.

Specifically, one way to approach the coarse-graining problem is to only group together the individuals that are to a sufficient extent interchangeable. This is the criterion we introduced as a "reconstitution test", and is the criterion implicitly assumed by virtually all compositional models of ecosystem dynamics [111]. However, experimental evidence [47,50–53] suggests that unless we are willing to resolve types differing by as few as 100 bases, this criterion is likely violated in

most practical circumstances. It is certainly violated when grouping strains into functional groups, or taxonomic species or families [89–91,112–114]. One expects, therefore, that explaining the practical successes of such descriptions would require a different definition of what makes a coarse-graining scheme adequate.

We proposed that this can be achieved with only a subtle change to the criterion: namely, by requiring that the grouped strains be approximately interchangeable not in all conditions, but in the conditions created by the assembled community itself. As long as the strains we study remain in a diverse ecological context, and as long as this diversity is derived from a sufficiently similar environment, we find that the coarse-grained description can be consistent in the sense that the strains grouped together possess similar properties of interest (e.g., invasion rate, post-invasion abundance).

In this paper, we focused on a case where the traits were differentiated only by the strength of their interactions, which established a unique hierarchy among them (a clear order in which to include them in the hierarchy of coarse-grained descriptions). In a more general case, the trait cost χ_i, or the trait usefulness in a given environment (b_i, K_i) will set up alternative, potentially conflicting hierarchies. We expect the model to have a rich phenomenology in this regime, which we have not considered here. Another obvious limitation of our analysis is that our model includes only competitive interactions. A simple way to extend our framework would be to include cross-feeding interactions; we leave this extension for future work.

Our analysis introduced a distinction between weak-sense and strong-sense coarse-grainability based on whether the performance of a coarse-graining scheme is robust to increasing the environment complexity. We explained how strong-sense coarse-grainability arises in our model,

linking it to a previously described phenomenon, namely that a sufficiently diverse community may "pin" resource concentrations (here, the exploitation of environmental opportunities) at values that are robust to compositional details [10,13,29,100]. Tracing its origin makes it clear that strong-sense coarse-grainability in our model is only as good as the assumption that the cost of carrying weakly-interacting traits is independent of the phenotypic background. Whether this assumption is ever a good approximation in natural ecosystems remains to be seen. Still, our argument provides an explicit mechanism for how coarse-grainability can not only coexist, but may in fact be facilitated by diversity.

The fact that strong-sense coarse-grainability is at least theoretically possible is intriguing also for the following reason. Throughout this work, we interpreted L as indexing a sequence of ever-more-complex environments (e.g., a minimal medium with 1 carbon source; a mixture of several carbon sources; resuspended homogenized leaf matter; an actual leaf). An alternative perspective, however, is to think of a single environment of interest and take L to be the level of detail at which it is modeled. Any model we could ever consider, however detailed, is necessarily incomplete. Consider the example of the human gut: how important is the exact geometry of the gut epithelium? the effect of peristalsis and flow on small-scale bacterial aggregates? the exact role of the vast diversity of uncharacterized secondary metabolites [115,116]? It seems plausible that the complete list of factors shaping this ecosystem includes many we will never even know about, let alone include in our models. Our analysis raises an intriguing – though at this point, purely speculative – question of whether the tremendous diversity of natural ecosystems might afford our models some unexpected degree of robustness to such unknown details.

In conclusion, there are many reasons to believe that analyzing a species in artificial laboratory environments might be of limited utility for understanding its function or interactions in the natural

environment [117]. Usually, however, the concern is that the laboratory conditions are too simple, and in reality, many more details may matter. Here, we use our model to propose that, at least in some conditions, the opposite can be true: understanding the interaction of two strains in the foreign conditions of the Petri dish may require a much more detailed knowledge of microscopic idiosyncrasies. Removing individual strains of a species from their natural eco-evolutionary context may eliminate the very reasons that make a species-level characterization an adequate coarse-graining of the natural diversity.

3.6 Technical Details

3.6.1 Simulating Eco-evolutionary Dynamics

The eco-evolutionary world in which the dynamics take place is described by the environment (constant in time), the biochemistry (also constant in time), and the state of the ecosystem (dynamically evolving). At any given moment of time, the state of the ecosystem is described by the following information: (1) The identity of each of the phenotypes, described microscopically as vectors of length L_∞; (2) The current abundance (population size) of each of these phenotypes. All simulations are performed in the L_∞-dimensional world of complete microscopic detail; the environments of reduced complexity L are implemented by zeroing out the environmental niches from $L + 1$ onward.

At the level of individual bacteria, for any moment in time, the next "event" to occur will be one of the following: (1) an individual dies; (2) an individual divides, giving rise to an identical sibling; or (3) an individual divides, giving rise to a mutant sibling. Of course, an individual-based simulation is both impractical and unnecessary; instead, we think of these dynamics as a combination of purely ecological updates of phenotype abundances (which can be modeled with

continuous ODEs), and discrete dynamics whereby some strains go extinct, and others are introduced into the population by mutation.

To implement such discrete events, the standard way is to employ a Gillespie scheme [118]. A slight complication here is that when overlaid with ecological dynamics, the rates of such Gillespie events become time-dependent (a mutation favorable right now may cease to be so as ecological dynamics continue). However, this complication is easily resolved using standard methods for implementing a hybrid stochastic-deterministic Gillespie scheme [119]. Briefly, instead of drawing the "time to next event", one must draw a probability threshold, and propagate the continuous dynamics while integrating the rate of an event to occur, up until that accumulated probability crosses the threshold (see [120] for an introduction that is both short and intuitive). As a result, to describe our simulation we just need to define how the state-dependent rates of such events are computed.

To do so, we adapt for our purposes the results of Ref. [121]. From the evolutionary standpoint, the candidate new strain is a mutant that has a chance of escaping drift and become *established* in the population. The probability of becoming established is proportional to the mutation rate, the population size of the parent strain, and the fitness effect δf of the mutation (i.e. the growth rate[2] of the candidate new strain). Once a strain is established, the stochastic effects become negligible and its subsequent dynamics can be modeled deterministically.

The simulation can be summarized with the following pseudocode:

[2] Note that in many evolutionary models, the fitness effect is computed relative to the (ever-increasing) mean fitness in the population. For us here, fitness is not an abstract property that could increase arbitrarily, but is directly defined as a growth rate; and limited resources automatically ensure that the population-mean growth rate is zero.

1. For all single mutants of existing strains, determine the rate at which they would establish in the population, as explicit functions of the abundances of extant strains.
2. Propagate ecological dynamics "for an appropriate length of time" as per standard technique [119].
3. Pick the lucky new strain among the beneficial first mutants; add it to the community at a population size $1/\delta f$; the population of the parent strain is, for consistency, reduced by the same amount.
4. Remove any strain whose abundance is below some predetermined threshold and is declining. (In our simulations, this threshold is set at relative abundance 10^{-6}.)
5. Repeat until the required simulation time has elapsed.

The mutation rate we use is $\mu = 10^{-10}$ per individual per unit time. For context, recall that population size is set by the carrying capacity of environmental niches $N = 10^{10}$ (see Figure 3.2C), so our choice corresponds to $\mu N = 1$. However, as explained below, the mutation rate parameter plays only a minor role in our analysis.

Evolutionary Equilibrium is Not Required

In our analysis of coarse-graining schemes, the community we study is never technically at an evolutionary equilibrium. We explicitly constructed our procedure to avoid making such an unrealistic postulate. Specifically, note that we assemble our strain pool from evolutionary equilibria obtained in *similar* environments, but we then study the interaction of these strains in the original, unperturbed environment, where the condition of evolutionary equilibrium was never imposed.

Of course, we then observe that a sufficiently diverse set of strains derived from sufficiently similar environments assembles into a community that is very close to the evolutionary equilibrium also

in the original environment, and this proximity is largely responsible for the behaviors reported in this work. This, however, is not a caveat, but a feature, as we expect the same to be largely true for real communities as well: the large diversity of strains is quite plausibly sufficient to populate the available niches without requiring *de novo* mutations, relying exclusively on the standing variation.

Mutation Rate is Not a Key Parameter

A corollary of the previous point is that for specifically our purposes here, mutation rate is not a key parameter of our model. Indeed, we only invoke evolution when constructing the strain pool, but each of the combined states is an evolutionary equilibrium, which in this model is guaranteed to be unique. One small caveat is that our evolutionary process is simulated at a finite resolution, and considers first-mutants only. As a result, the trajectory could get stuck in a *locally* non-invadable equilibrium rather than the unique true one, something that would be enhanced by setting the mutation rate too low. In this way, the evolutionary stochasticity (and thus the mutation rate) does technically play a weak role, but we found it to be essentially irrelevant for the parameters used here (specifically, running replicate eco-evolutionary trajectories from random initial phenotypes generated virtually indistinguishable final states).

One artifact that occasionally arises when constructing a strain pool in an environment of complexity L occurs when some obtained phenotypes are identical in all traits $i \leq L$ but are distinguishable in traits past L, which correspond to those resources/opportunities that have been set to zero available benefit ($h_i = 0$ by b_i being set to zero for $i > L$; see Section 3.6.1 above). Note that in our model, it is rare but perfectly reasonable for phenotypes to carry traits that offer no environmental benefit; this occurs whenever the contribution of trait interactions J_{ij} provides a net reduction of maintenance cost. (In other words, surviving strains need not be identically 0 past

$i = L$.) However, what should be true is that an L^*-type with $L = L^*$ can only have one representative: in any environment of complexity $L \leq L^*$, the lowest-cost member in any L^*-type is strictly superior to all others in its L^*-type, and would outcompete them. In practice, the inferior strains are occasionally retained in some replicates due to the evolutionary process only considering first mutants. This has no effect on eco-evolutionary dynamics beyond a slight change to the abundance of the respective type; however, if left uncorrected, it would lead to artifacts in evaluating coarse-graining quality due to the artificially inflated diversity within an L^*-type. To correct for this, we add the following step when generating the strain pool: after collecting together all phenotypes from the ensemble of similar environments, we check for any L^*-types that contain more than one strain at $L = L^*$ and, if so, remove this artifact by retaining only the superior (lowest-cost) phenotype.

3.6.2 Averaging in Figures

All heatmaps shown in figures correspond to a single random biochemistry. The small amount of "graininess" seen in the heatmaps is caused by the quirks of exactly when the L^*-coarse-graining is able to resolve new subtypes. It is shown as is, with no additional averaging or smoothing. In contrast, the isolines overlaid on plots are meant to capture the qualitative trends transcending the quirks of a given biochemistry. They are computed by repeating the same analysis for 20 random biochemistry realizations: the 20 heatmaps are averaged, smoothed with a 1-pixel-wide Gaussian kernel, and the isolines are picked as contour lines of the result.

The only exception to this procedure is Figure 3.7 which investigates the effect of perturbing the environment. In this figure, each pixel is an average over 20 random perturbations (of magnitude specified by epsilon) for a single biochemistry.

3.6.3 Metrics for Coarse-graining Quality

Examples of Other Ecological Properties and their Compatibility with L^*-coarse-graining

As described in the main text, the leave-one-out scheme judges a grouping of strains into coarse-grained OTUs as justified if the strains constituting OTU X all behave similarly when introduced into a community missing X. However, the "similarity of behavior" could itself be assessed by a variety of criteria. In the main text, we focused on comparing the invasion rates of the left-out strains. Some equally interesting alternatives include, for instance, the abundance reached by the invading strain (which might be zero if the strain cannot invade), or the niche occupancy in the resulting community. In each case, we perform the same weighting procedure as used in the main text when defining the leave-one-out test (Q_{inv}).

Invading strain abundance: In this example, we denote n_{μ,μ_*} the relative final abundance reached by strain μ after being introduced in a community missing the L^*-type μ_* (by construction, μ is a representative of μ_*). Similar to the reconstitution test (for abundances), we look at the variability of n_{μ,μ_*} over all $\mu \in \mu_*$, and define the coarse-graining quality Q_{abd} by the weighted average of this quantity over all L^*-types:

$$Q_{\text{abd}} = \sum_{\mu_*} \bar{n}_{\mu_*} \text{CV}_{\mu \in \mu_*}[n_{\mu,\mu_*}].$$

Niche occupancy post-invasion: Here, instead of looking at a compositional property (abundance of a particular coarse-grained type), we ask how similar the effect of the invading strains is on the post-invasion *function* of the community as a whole. The functional properties are encoded in community-wide niche occupancy T_i – or, equivalently, in the niche benefit h_i at equilibrium (the benefit from carrying trait i). Since the environment is kept fixed, T_i and h_i are directly related.

The reason we choose h over T is because in a metabolic interpretation of the resources (community in a chemostat), h_i is directly measurable as resource concentration in the effluent.

Quantitatively, let $h_i^{(\mu,\mu_*)}$ be the h vector of the equilibrium community after strain μ invades the community missing the L^*-type μ_* (again, μ is a representative of μ_*). One might expect computing the average component-wise standard deviation of this vector (across all $\mu \in \mu_*$) to be an appropriate measure of variability, but doing so artificially attenuates any variation by including "equilibrated" niches with $h_i \approx \chi_i$ (see Sections 3.4.1 and 3.6.5) and thus have vanishing variation. We therefore focus only on the variability in the first component h_1^{μ,μ_*}. Similar to the leave-one-out invasion rate analysis, we then compute the weighted average of this quantity over all L^*-types:

$$Q_h(L, L^*) = \sum_{\mu_*} \bar{n}_{\mu_*} \mathrm{std}_{\mu \in \mu_*} h_1^{\mu,\mu_*}.$$

These two new measures of coarse-graining quality, combined with the two considered in the main text, give four metrics $Q(L, L^*)$ encoding four different questions of interest. We chose them to span both compositional and functional properties of a community. How amenable are these four questions to coarse-graining?

The answer is presented in Figure 3.8. The panels are ordered by "degree of coarse-grainability" as defined in Fig Figure 3.8D-F. Our findings are consistent with the recurring observation in the experimental literature that functional properties tend to be more reproducible than compositional ones [9,90,112]: the hierarchy of the four questions presented in Figure 3.8 can be summarized as saying that ecosystems are coarse-grainable in the strong sense when a coarse-graining is evaluated based on functional properties.

Indeed, the invasion rate of a missing strain (panel A) is basically a functional property of the pre-invasion community (the growth rate of a strain is determined by the environmental conditions constructed by the rest of the community). Next comes the niche occupancy (panel B) – a functional property assessed post-invasion. Next is the abundance reached by the invading strain (panel C), a compositional property that we expect to be distinctly less coarse-grainable (indeed, recall how strain-strain interactions are strongly affected by tail-end niches). However, the reconstitution test is still last in the list (panel D), epitomizing the goal of a bottom-up compositional description and requiring the knowledge of *all* microscopic details.

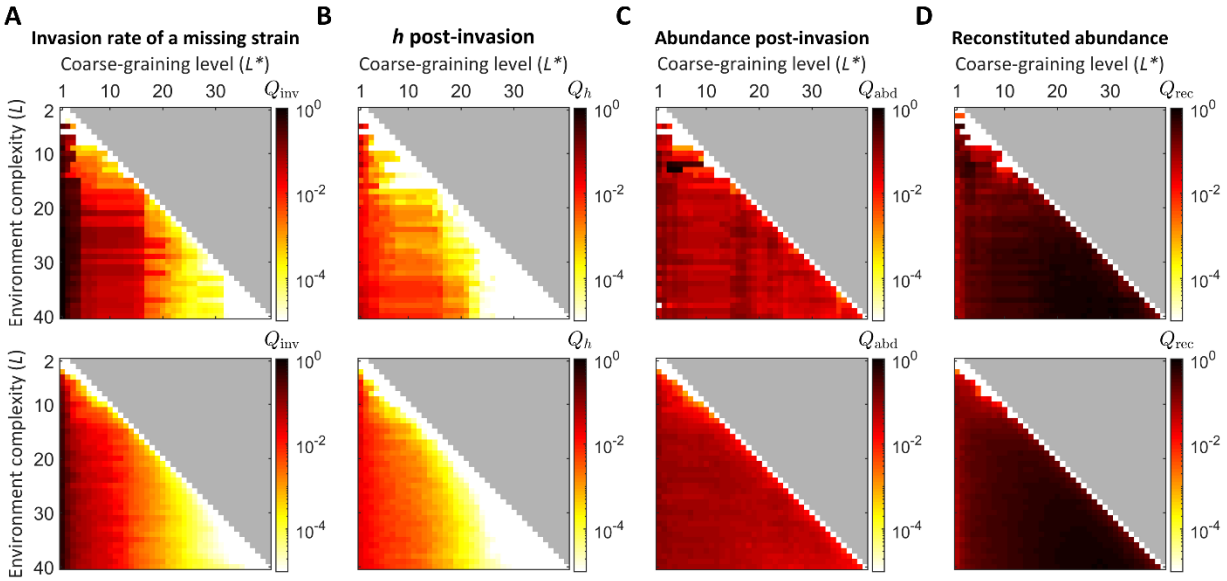

Figure 3.8 | Four questions whose compatibility with the L^*-coarse-graining scheme ranges from excellent to non-existent. Top row of heatmaps show coarse-graining qualities for a single random biochemistry, while those in the bottom row show averages over 20 random biochemistries from which the isolines shown in Figure 3.5 are computed. **A:** Invasion rate of a missing strain; coarse-grained description sufficient. **B:** Niche occupancy post-invasion; coarse-grained description sufficient. **C:** Abundance reached by a strain post-invasion; coarse-graining compatibility is poor. **D:** Reconstitution test (requires functional equivalence of strains); no coarse-graining is possible. Note that the coarse-graining quality metrics shown in panels A-C are all assessed in the framework of the leave-one-out test as defined in the text.

Weighting by Pool Abundance versus Replicate Community Abundance

Recall that the reconstitution scheme (Figure 3.4B) evaluates a grouping of strains into coarse-grained OTUs based on the ability of the coarse-graining to precisely reconstitute replicate communities from representatives of each OTU. If the replicate communities are similar in composition, then we deem the coarse-graining to be an adequate grouping. Specifically, we quantify this in terms of a weighted average coefficient of variation in type relative abundance (n_{μ_*}) at ecological equilibrium across replicates (Q_{rec}). In the main text, we choose to perform the weighting of each type by its mean relative abundance over replicates, rather than by its mean abundance in the pool (as used in the leave-one-out scheme). This choice avoids artificially increasing Q_{rec} with noise from rare, low-abundance types (Figure 3.9). To see this mathematically, consider a low-abundance L^*-type, μ_*^0 that is rarely observed across M replicate communities: say, μ_*^0 has abundance ϵ in only 1 replicate. When M is large, the coefficient of variation is approximately $\text{CV} \approx \frac{\epsilon/\sqrt{M}}{\epsilon/M} \approx \sqrt{M}$. In the case of weighing by replicate-mean abundance, Q_{rec} is simply the sum of standard deviations to which the rare type contributes $\epsilon/\sqrt{M}$, which is very small if ϵ is small compared to M. In the alternative case of weighing by pool-mean abundance, the contribution from the rare type's CV, $\frac{\bar{N}_{\mu_*^0}}{\sum_{\mu_*} N_{\mu_*}}\sqrt{M}$, may or may not be small depending on the relative abundance of the strains that constitute type μ_*^0 in the pool, inflating the noise from its rare appearances.

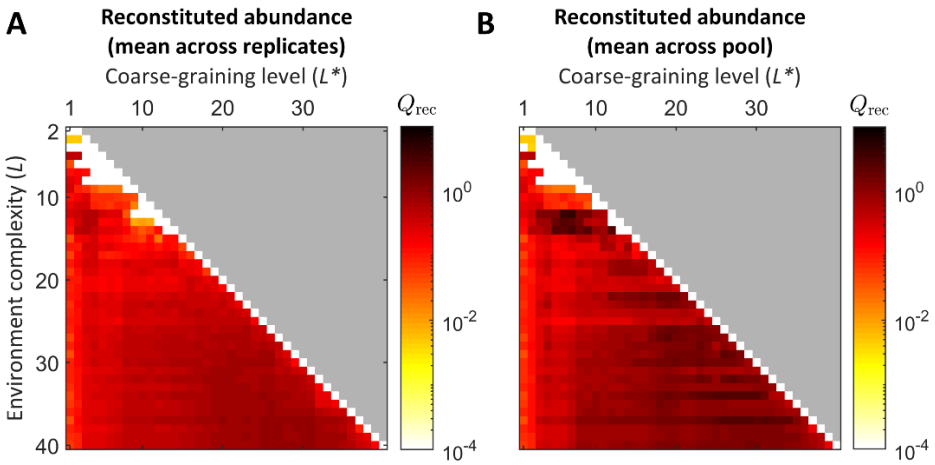

Figure 3.9 | Comparison of weighting L^*-type coefficient of variations (C.V.) in Q_{rec} by a type's mean abundance either as observed across replicate reconstituted communities (**A**) or in constructing the strain pool (**B**). The latter weighting inflates the noise of rare (low abundant) strains causing an increase in Q_{rec}, which indicates larger typical variability within L^*-types. Each panel shows the same biochemistry seed as used in Figure 3.5.

3.6.4 Coarse-graining Truly Neutral Traits

One simple sanity-check of our framework is to make the tail-end niches $i \to L_\infty$ not just weakly interacting, but genuinely close to neutral. A distinction that makes almost no difference in either cost or benefit should surely be negligible for all questions, including the "reconstitution test". To test this, we repeat the analysis for a model where we use the same declining sigmoid-shaped function $f(n)$ to scale down not only the trait interactions (cf. Figure 3.2A,B), but also the trait cost χ_i and benefit b_i. The result is shown in Figure 3.10. As expected, we find that all four example properties and criteria we consider are now coarse-grainable in either the strong or weak sense (this figure is to be compared with Figure 3.8).

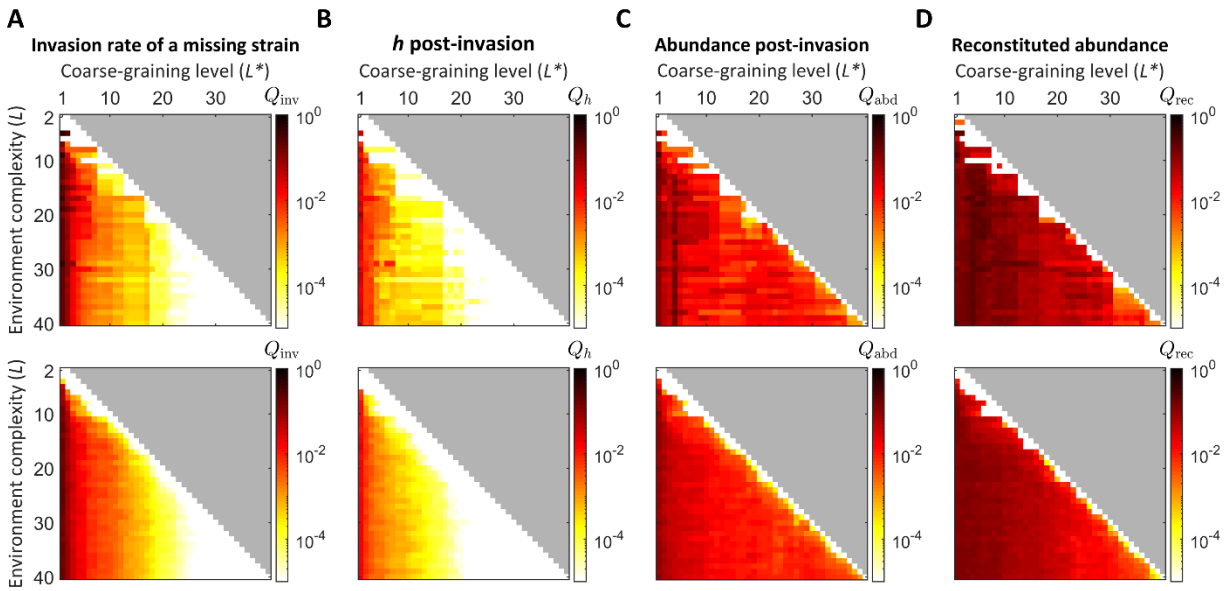

Figure 3.10 | Reproducing Figure 3.8 for a model where tail-end niches are not just weakly interacting, but are also increasingly neutral (bring almost no benefit $b_i \to 0$ and incur almost no cost $\chi_i \to 0$); see text. As expected, phenotypic diversity in the corresponding traits can be adequately coarse-grained away no matter the criteria (compare to Figure 3.8). Again, top row of heatmaps show coarse-graining qualities for a single random biochemistry, while those in the bottom row show averages over 20 random biochemistries.

3.6.5 Effectively Neutral Traits and Asymptotic Scaling of $Q_{\text{inv}}(L, L^*)$

As described in the main text, with respect to the invasion rates of strains in leave-one-out experiments, our ecosystem is coarse-grainable in the strong sense, meaning that a given L^*-coarse-graining maintains a desired quality even as more diversity is resolved with increasing environment complexity L. In other words, the variability of invasion rates between strains within L^*-types can be made arbitrarily small with a sufficient level of coarse-graining (L^*), independent of the environment complexity. This section presents the formal analysis underlying this result.

Recall that $Q_{\text{inv}}(L, L^*)$ measures the typical invasion rate variability between strains within L^*-types so that $Q_{\text{inv}} = 0$ indicates a perfect coarse-graining, which trivially occurs when $L^* = L$. Put mathematically, for invasion rate to be a strong-sense coarse-grainable property, it must satisfy the following criterion: for some sufficient $L^* < L$, there exists some finite (possibly L^*-dependent)

constant $M(L^*)$ such that $Q_{\text{inv}}(L, L^*) < M(L^*)$ for any L. We first analyze only the invasion rate variability between strains within some L^*-type, from which the same analysis applies to any L^*-type.

The invasion rate of strain μ is determined by the pre-invasion niche availabilities h_i set by the community missing type μ_*:

$$r_{\mu,\mu_*} = \sum_{i=1}^{L} \sigma_{\mu i} h_i - \chi_\mu.$$

For a given L^*, all strains within a type μ_* possess identical traits up to $i = L^*$ so that the differences in invasion rates follow from differences in traits $i > L^*$. Let then $r_{.,\mu_*}$ denote the identical contributions to invasion rate from traits $i \leq L^*$ in order to isolate the variable part:

$$r_{\mu,\mu_*} = r_{.,\mu_*} + \sum_{i>L^*}^{L} \sigma_{\mu i}(h_i - \chi_i) + \sum_{j>L^*}^{L} \sum_{i<j} J_{ij} \sigma_{\mu i} \sigma_{\mu j}.$$

For the coarse-grainability criterion to hold, we need to show that each of these sums converge as $L \to \infty$. Dealing first with the trait-trait interaction piece (J_{ij} terms), we define the finite sum $S_j \equiv \sum_{i<j} J_{ij} \sigma_{\mu i}$ and take $L \to \infty$ so that the double sum is converted into a series of finite sums, $\sum_{j>L^*}^{\infty} S_j \sigma_{\mu j}$. In order for this series to converge, we must have $|S_j|$ fall off faster than $1/j$. To determine the scaling of the S_j, note that the terms of each finite sum are Gaussian distributed with mean $\langle J_{ij} \rangle = 0$ and sigmoidally decaying variance $\langle J_{ij}^2 \rangle$ (see Section 3.2.2). Therefore, each finite sum is essentially a random walk of terms that on average sum to $0 \pm \sqrt{j}\sqrt{\langle J_{ij}^2 \rangle}$, where for large j

(as L tends to infinity) the scatter goes like $\pm\sqrt{j}\mathrm{e}^{-j}$ due to the sigmoidal decay of trait interaction strength. With the S_j thus falling off exponentially, we indeed have convergence:

$$\sum_{j>L^*}^{L}\sum_{i<j} J_{ij}\sigma_{\mu i}\sigma_{\mu j} = \sum_{j>L^*}^{L} S_j\sigma_{\mu j} \to M_J(L^*) < \infty \quad \text{as} \quad L \to \infty. \tag{3.3}$$

Similarly, for the non-interaction piece to converge as $L \to \infty$, we must have $|h_i - \chi_i| \sim 1/i^{\alpha}$ with $\alpha > 1$. Figure 3.11 shows the scaling of the typical cost-benefit deviation $\langle|h_i - \chi_i|\rangle$ with$1/i$, where the average is over (left-out) types with the same weighting as used in the main text. Plotted with the simulation data are visual guides that show the scaling of the convergence condition ($1/i$, dashed 1:1 line) and the scaling of the trait interaction sigmoid used to generate J_{ij} (solid gray line). We see that for a sufficient level of coarse-graining ($L^* = 30$, black points) the data indeed falls off faster than $1/i$, verifying that

$$\sum_{i>L^*}^{L} \sigma_{\mu i}(h_i - \chi_i) \to M_h(L^*) < \infty \quad \text{as} \quad L \to \infty. \tag{3.4}$$

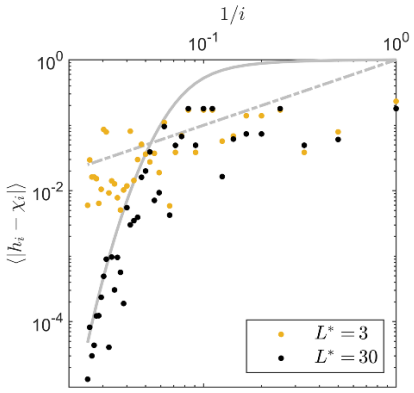

Figure 3.11 | Asymptotic scaling of the typical cost-benefit deviation ($|h_i - \chi_i|$) in the leave-one-out scheme. For weakly interacting traits (large i), the corresponding niche benefits become effectively neutral when sufficient diversity is present in the community missing an L^*-type (black

points; $L^* = 30$), vanishing exponentially (cf. trait interaction decay $|J_{ij}|$ plotted as a solid gray line) such that their sum converges as described in expression (3.4). When L^* is too small (e.g., $L^* = 3$; yellow points), insufficient diversity remains after removing an L^*-type such that the sum will typically no longer converge (terms scaling like dashed $1/i$ line) and coarse-graining quality is poor. Both coarse-grainings are for an environment of complexity $L = 40$ and the same biochemistry as in Figure 3.5.

This result follows from two contributing factors: both (1) weak J_{ij} and (2) phenotypic diversity in the tail-end traits. Ref. [13] has demonstrated that when a community is seeded with a sufficiently diverse strain pool consisting of unstructured, purely random phenotypes ($\sigma_{\mu i}$) the community exhibits a collective phase where strain abundances will adjust to drive fitness benefits to match costs, $h_i \approx \chi_i$ (effectively neutral traits in the so called "S-phase"). To see the presence of this phase in the leave-one-out context, first note that the pool of phenotypes generated within our eco-evo framework are essentially random and unstructured in those traits that are weakly interacting (small J_{ij}) as observed in the example shown in Figure 3.2F. Second, as L^* increases and more strains are resolved, fewer strains are removed in leaving out an L^*-type, increasing the diversity of the remaining community. Figure 3.11 demonstrates the effect of this point: for small L^*, the community to be invaded is not diverse enough for weakly interacting traits to be considered effectively neutral ($h_i \approx \chi_i$) and therefore cannot be ignored ($L^* = 3$, yellow points), unlike the large L^* regime ($L^* = 30$, black points).

Together, (3.3) and (3.4) imply that the typical invasion rate variation between strains is asymptotically bounded by some constant $M(L^*)$ that is independent of L, enabling coarse-grainability in the strong sense.

Chapter 4: Coarse-grainability *in vitro* versus *in silico*

The theory work presented in the previous chapter provided a means for generating random ensembles of *structured* ecosystems, enabling the theoretical investigation into many new exciting questions that were previously inaccessible: e.g., processes shaping microbial pangenomes, or ecosystem coarse-grainability. Continuing with the focus on defining and exploring the question of ecosystem coarse-grainability, in this chapter I generalize the previously introduced coarse-graining procedure to be model-independent and develop a framework for identifying an appropriate level of description for predicting a given ecosystem property of interest. By considering any possible partitioning of taxa as a valid coarse-graining, I can readily apply the framework presented here on experimental data. Using recent data from experiments that tune species richness of microbial communities, I *explicitly* test the hypothesis of the previous chapter that coarse-grained descriptions should be more predictive in highly diverse ecological contexts. My results provide the first empirical evidence of diversity-enhanced coarse-grainability – the hope of "emergent simplicity" sentimentalized by the statistical physics perspective. Completing the loop back to theory, I touch on the difficulties in reproducing this empirical observation in classic, unstructured models of ecology. This work is currently unpublished.

4.1 Introduction

Sequencing-based technologies allow resolving the composition of microbial ecosystems to strain-level detail; however, coarser representations are often found to be more reproducible and more predictive of community-level properties. For example, Goldford *et al.* have demonstrated that both community-level function and composition at the taxonomic level of families are significantly more stable across replicate communities of reconstituted microbiomes than the highly variable

strain- or species-level [9]. Although this suggests that coarse-grained composition could serve as predictors, useful coarse-graining schemes need not be taxonomic or phylogenetic. Indeed, many examples have shown community dynamics and function can be well-captured by coarse-graining diverse sets of taxa into just a handful of functional classes, which typically are not monophyletic [45,46,122]. How do we formalize examples like these? The general principles for selecting an appropriate level of description and identifying predictive groupings for modeling remain elusive.

The notion of "functional guilds" used in classical ecology [123–128] provides hope for a possible path forward. Given the observed functional stability in microbial ecosystems due to shared metabolic capabilities across diverse taxa [90,129–131], it seems plausible that a similar idea could be employed to characterize these systems in simpler terms of a small set of functional roles instead of the overall composition. However, determining the constituents of functional groups has proven difficult in practice, tending to be more of an "art" that often requires expert knowledge of the specific system of interest. The difficulty stems from the lack of clarity in defining relevant traits on which to form functional groups [128,132]. Adding to the challenge, most defined traits contributing to function are hard to measure, especially in a community context. For example, performing metabolomics is costly and the results are not straightforward to parse compared to techniques for measuring composition.

To begin addressing such challenges, I develop a framework based on a particular form of coarse-graining in which taxa are partitioned into putative functional groups, where the combined abundances of taxa in each group serve as a coarse-grained description. This approach leverages the often observed correlations between composition and function; similar to other statistical approaches that have recently shown promising progress [133,134]. The framework presented

here systematically compares all possible coarse-grained descriptions by explicitly quantifying their prediction power and information content. I demonstrate that coarse-grained descriptions can provide similar predictiveness as microscopic, but at a fraction of the entropy budget, enabling the identification of an optimal "good-enough" description for predicting a property of interest. By showing how the selected optimum can decrease in complexity as a function of community diversity, this framework can distinguish between fundamentally different mechanisms underlying the notion "emergent simplicity" in microbial ecology. Applying the framework to experimental data from synthetic microbial communities [135,136] explicitly reveals how ecological diversity can enhance our ability to coarse-grain ecosystems, providing direct evidence that the hope of "emergent simplicity" is likely to be realized in the highly diverse context of natural communities. Finally, I discuss how this empirical observation is an example of the theoretical concept of functional attractors in microbial ecology. Moreover, I argue that this phenomenon is likely difficult to capture in canonical ecological models with parameters drawn from random ensembles, highlighting the importance of both community structure and diversity in determining coarse-grainability.

4.2 A Framework for Quantifying Coarse-grainability

For the sake of concreteness, imagine the following experimental setup: after assembling a set of N communities indexed by $\mu = 1, \dots, N$, measure some functional property Y_μ (e.g., production of methane, or some metabolite), as well as the microscopic composition (abundance of each strain $n_{i\mu}$, where the strains are indexed $i = 1, \dots, P$); see Figure 4.1A.

The implementation of this experimental setup can be performed in a variety of ways. For the purposes of this work, here I will consider the N communities to be assembled from N different

subsets of strains sampled from a fixed library (e.g., isolates stored in the fridge of a lab), inoculating the same medium for each community. This is the setup of the Clark *et al.* experiment [135] on which we will apply our framework in Section 4.3.2 below. The control parameter of this setup is community diversity (sometimes referred to as "species richness"). For example, one can consider low-diversity communities inoculated with only a few strains at a time, or higher-diversity communities inoculated with many strains at a time. An alternative implementation could instead hold diversity fixed and tune the media conditions (environment) as a control parameter. In this case, the N communities are assembled from the same diverse inoculum, but the environment selects strains based on their growth in different media conditions. The general theoretical framework presented below applies to either implementation; the only inputs are the functional property Y_μ and the microscopic composition $n_{i\mu}$ of each community.

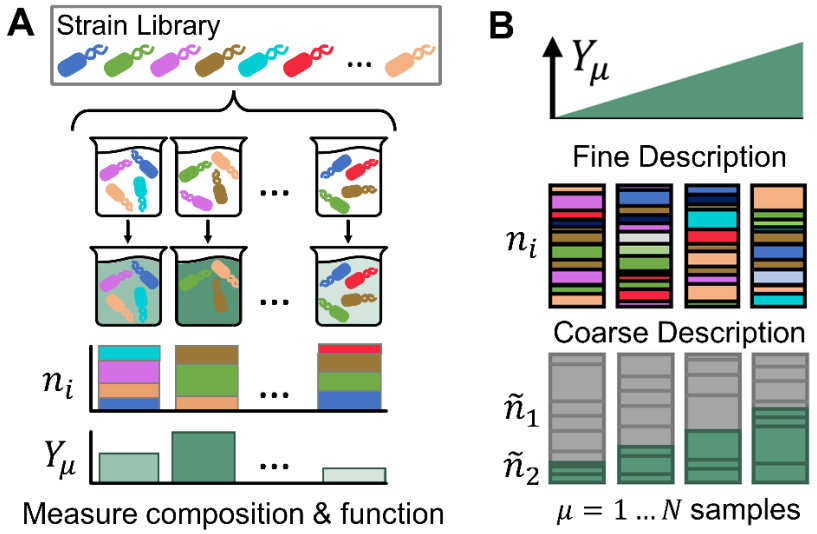

Figure 4.1 | General experimental context and the hope of coarse-grainability. A: Subsets of strains are sampled from a fixed library (e.g., bacterial isolates stored in a lab) to assemble communities that are brought to ecological equilibrium. At steady-state, the microscopic composition of each community is measured as the abundances of each strain n_i (i labeling the strains), as well as a community-level function of interest Y_μ (μ labeling the sample communities). This experimental setup can be implemented in various ways, but in any case, the goal is to learn the relationship between community composition and function. **B:** Example illustration of a predictive coarse description. Fine-scale strain abundances do not obviously correlate with

observed function, but group abundances $\tilde{n}_1, \tilde{n}_2$ (combined abundance of taxa within a functional group) do.

Many examples have established that, even if the microscopic description is available, coarser representations are often still predictive of properties of interest (community-level functions) [45,46]. This is illustrated in Figure 4.1B, where if the communities are ordered by the amount of property Y measured (e.g., how much methane was produced), then what is typically observed is that the microscopic composition is highly variable. But if one knew how to color taxa into just two groups (e.g., those that produce methane in green), then one would notice that the combined abundance of methanogens predicts the amount of methane in each community. This is the hope of coarse-grainability we consider here in this work, which motivates asking: "What is the simplest description sufficient to predict Y?". The framework I will now present converts this question from an intuitive language to well-defined quantitative statements.

4.2.1 Coarse-graining Scheme: Constructing Sets of Variables

To make predictions about properties and behaviors of a system requires two choices: (1) a set of variables that describe the possible states of the system, serving as predictors for the property of interest; (2) a model formulated in terms of these variables. Both are, to an extent, an art, relying on expert knowledge and intuition. Though difficult to implement in practice, there does exist well-developed theoretical guidance for the choice of model. Once the variables are specified, model complexity can be guided by rigorous quantitative principles (e.g., Bayesian model selection). This is the conventional question in modeling ecology, where the variables are fixed (e.g., species abundances) and the focus is on selecting an appropriate complexity of the model. In contrast, the question of coarse-grainability is concerned with the comparison between choices of *variables*. In contrast to model selection, the principles of selecting the level of detail in the

variables remain elusive. Therefore, to make progress in filling this gap, the framework presented here will instead fix a model class (see below) and systematically tune between sets of variables that are detailed versus coarse-grained. In doing this, I will demonstrate how to identify the optimal level of description given a model class, an observable of interest, and ecological conditions.

Here, the particular form of coarse-graining on which I focus is in terms of compositional variables: taxa are grouped together into putative functional groups and the group abundances serves as coarse-grained variables. In general, a microscopic description is set by the resolution of the measurement technique in experiment, but for simplicity, here I will call the microscopic units in the framework "strains". To obtain a coarse-grained description from this, I partition strains into a set of non-overlapping groups (denoting the partitioning as Ψ), and compute the combined abundance of each group (Figure 4.2A): for each α in Ψ, $\tilde{n}_{\alpha\mu} \equiv \sum_{i\in\alpha} n_{i\mu}$, where again μ labels the sample community. To illustrate an example of this procedure, I could describe the composition of each community in the above experimental setup with family-level abundances instead of specifying abundances of individual strains. However, although taxonomy provides one natural hierarchy of coarse-grained description along which to tune, useful coarse-graining schemes need not be taxonomic. For example, Ref. [122] has shown that the dynamics of chitin-degrading communities is well-captured by coarse-graining taxa into three functional classes ("degraders", "exploiters", and "scavengers"), none of which is monophyletic. In general then, any grouping of strains is a valid coarse-graining to consider, and therefore there is a need for some means to compare coarse-grainings. I now show a systematic way of doing this that enables identifying useful coarse-grained descriptions for predicting a property of interest.

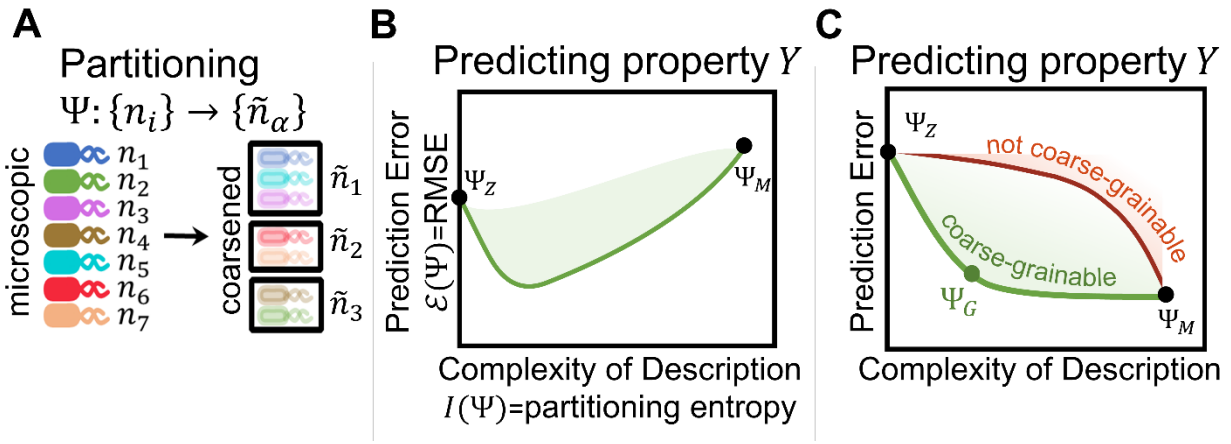

Figure 4.2 | Defining and scoring a (compositional) coarse-graining scheme. A: Here we focus on coarse-graining community composition by partitioning the microscopic taxa into (non-overlapping) putative functional groups. A partitioning Ψ can be any grouping of strains, and maps the set of microscopic variables n_i to a set of coarse-grained variables $\tilde{n}_\alpha$ by summing the abundances of strains i within each group α defined by Ψ (see text). **B:** We evaluate a coarse-graining by quantifying its prediction power for a property of interest Y and information content (the complexity of the description it provides). Specifically, we formulate a model of a specified model class in terms of the coarse-grained variables, and compute the prediction error $\mathcal{E}(\Psi)$ as the standard out-of-sample root-mean-square error (RMSE) of the best trained model; information content $I(\Psi)$ is simply quantified as the entropy of the partitioning (see text). With these, metrics any coarse-graining Ψ can be plotted on a prediction-information plane (light green point cloud). Generically, due to insufficient data for constraining more complex models, this point cloud will curve up as depicted: the microscopic model from Ψ_M overfits and performs worse than the model with zero information based on Ψ_Z (see text). This example of overfitting shows the classic reason for choosing a coarse-grained description. The Pareto front (bold green line) highlights the most efficient coarse-grainings and will serve as the key feature we focus on in this work. **C:** Even when data is sufficient for training the microscopic model to its lowest prediction error possible, it is still imperfect due to model limitation. This enables the possibility for coarse-grainings to be similarly predictive but at a fraction of the entropy budget; identifying an optimal point Ψ_G. When the Pareto front qualitatively behaves in this fashion, we say the ecosystem is coarse-grainable for predicting a given property.

4.2.2 Evaluating a Coarse-graining: Predictive Power & Information Content

The first natural way in which to compare coarse-grainings is by the information they contain or how complex of a description they provide. Explicitly, I quantify this in terms of the information content $I(\Psi)$ of a coarse-graining Ψ, which is the mutual information between the coarse-grained label and the identity of the strain, $I(\Psi) = -\sum_\alpha p_\alpha^{(\Psi)} \log p_\alpha^{(\Psi)}$, where $p_\alpha^{(\Psi)}$ is the number of strains

that Ψ assigns to class α divided by the total number of strains in the library P. Note that this quantity is the entropy of the partitioning Ψ, and encodes the level of detail or complexity of description: resolving every strain is a description with the largest information content.

Another natural aspect by which to compare coarse-grained descriptions is their predictive power. To quantify this, I will specify a *fixed* model class, and access the out-of-sample prediction error $\mathcal{E}(\Psi)$ by computing the standard root-mean-square error (RMSE) of the best trained model for predicting Y in terms of the coarse-grained variables $\tilde{n}_\alpha$. For the purposes of presenting the general framework, the specific choice of model class is not of concern here; the key point is that once I specify a model class, any set of coarse-grained variables has a well-defined prediction error, enabling the selection of an optimal coarse-grained description within the specified class of models. For example, whether I choose a Lotka-Volterra model, or a machine learning algorithm, in any case this framework aims to determine which input variables to use (e.g., species abundances versus family abundances).

4.2.3 Prediction vs. Information Diagram and Coarse-grainability

For each coarse-graining Ψ, evaluating $I(\Psi)$ and $\mathcal{E}(\Psi)$ yields a point in a scatter plot that I will call the prediction-information diagram (Figure 4.2B). To parse the information it contains, first note that if no compositional information is available, the best prediction for Ywill typically be its mean value, resulting in a prediction error of $\mathcal{E}_0 \equiv \mathcal{E}(\Psi_Z) = \mathrm{std}(Y)$; see in Figure 4.2B the point labeled Ψ_Z for zero information (all taxa are combined into just one group). This point and its associated error provide a natural performance ceiling, above which clearly corresponds to overfitting. On the other extreme is the microscopic description Ψ_M (the one with maximal information where all taxa resolved in separate groups). If the amount of data is insufficient to properly train the model on microscopic variables, its out-of-sample prediction error could be

worse than the $\mathcal{E}_0$ ceiling; this is how one would recognize the overfitting regime as depicted in Figure 4.2B. Avoiding overfitting is one valid reason to use simpler (less detailed) descriptions; however, this reason is already well-understood, and in the quest to understand the emergent predictive power of coarse models I will go beyond this benefit. I will thus work in the regime of sufficient data, where $\mathcal{E}(\Psi_M) < \mathcal{E}_0$.

Crucially, even with sufficient data, the microscopic prediction will generally remain imperfect: even when the description is exhaustive, prediction is still limited by model class. In other words, just because the microscopic description is the one that contains all the information about the system, that does not necessarily mean we know how to use this information. Figure 4.2C illustrates a result that follows from this: coarser descriptions may provide similar predictive power at a fraction of the entropy budget (bold green Pareto front of Figure 4.2C), identifying the simplest "good-enough" description Ψ_G for a desired error $\mathcal{E}$. If the coarse-graining Pareto front (optimal $\mathcal{E}$ for a given I) behaves in this way, I will say that our ecosystem is coarse-grainable for predicting property Y. If instead the coarse-graining Pareto front behaved more like the red curve of Figure 4.2C, the ecosystem would be deemed non-coarse-grainable for the given property. Thus, this framework enables the delineation of which observables are coarse-grainable versus not within the scope of a model class. Note that model-class limitation is relevant for any model, as any model omits some underlying complexity.

4.3 Results

So far this work has noted how coarse-grainability encoded in the Prediction-Information diagram depends on the choice of model class and the property of interest, leaving a less intuitive determining factor to explore in detail below: the dependence of coarse-grainability on ecological conditions. The investigation below focuses on the role of ecological context, specifically

community diversity, in shaping the Prediction-Information diagram. First, I will demonstrate the value of the framework presented above by illustrating its capacity for mapping out a potentially rich conceptual space; namely, the capability to distinguish between fundamentally different mechanisms underlying the notion of "emergent simplicity" in microbial ecology. Then, because all the components of the framework are measurable in experiments, I will show how prediction-information diagrams behave for empirical data. Excitingly, by applying the framework to available datasets, I find that a higher-diversity community can be *more* efficiently described with a simpler model, despite its nominally larger complexity. Finally, I discuss the potential sources of this empirical observation based on plausible biological mechanisms for the particular experimental system, as well as propose a more general mechanism based on the theoretical concept of functional attractors.

4.3.1 Nuancing the Notion of "Emergent Simplicity" in Microbial Ecology

To ask how the Pareto front changes as a function of community diversity, I will return to the experimental setup I described above. Again, low-diversity corresponds to randomly sampling subsets of a few strains at a time, while high-diversity means sampling many strains at a time. An important note to make at this point is that in each case strains are sampled from the *same* library. Using the same library to assemble low- and high-diversity communities allows for comparing the prediction-information diagrams of each context on the same set of axes because the microscopic model is the same for both. To see this, imagine using first-order linear regression as the model class and the library consists of 50 strains; then, in each case the microscopic model is 50 parameters (one coefficient for each strain). The question then becomes: is this, or any coarse-grained model, more predictive in the high-diversity context or low-diversity context?

One hope inspired by a statistical physics perspective is that in a higher diverse context, despite being nominally more complex, there might be some form of emergent simplicity possibly due to self-averaging. In the framework presented here, this would correspond to the identified simplest "good-enough" description Ψ_G of Figure 4.2C decreasing in complexity as we increase diversity. Though evidence of emergent simplicity in microbial ecosystems has been reported [9], the term has been used loosely, but my framework allows the notion to be further refined. The prediction-information diagrams can delineate multiple scenarios under which the optimal complexity of description might decrease with community diversity; two are shown in Figure 4.3. Both can be realized in a resource-competition model, and today, both would be conflated under the term "emergent simplicity". However, they correspond to fundamentally distinct mechanisms: model limitation (Figure 4.3A) and the more exciting emergent simplicity in the strict sense Figure 4.3B. In the case of Figure 4.3A, although one can manage to use a simpler and simpler description, this is out of desperation since descriptions perform worse overall. This could correspond to a situation where the specified model class is more and more limited as more species are added because maybe higher order interactions increasingly matter but they are not included in the model. Figure 4.3B shows the opposite trend, where a simpler description can be more predictive than it was in low-diverse context.

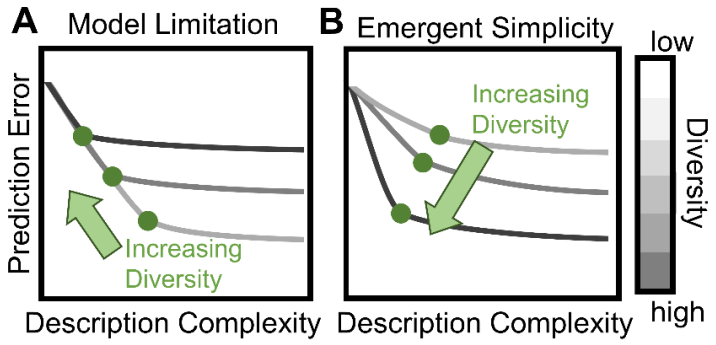

Figure 4.3 | Framework for quantifying coarse-grainability distinguishes mechanisms of "emergent simplicity". Tracking the prediction-information Pareto front of coarse-grainings as a

function of community diversity (number of strains sampled from library in each subset). As diversity increase, the optimal coarse-grained description (green point) may become simpler (naïve sense of "emergent simplicity"; see text) in two different ways. **A:** The optimum performs worse and worse because overall the specified model class becomes more limited at higher diversity. **B:** More excitingly, the optimum not only becomes simpler, but also becomes more predictive in a higher diversity context: emergent simplicity in the strict sense. This framework avoids conflating these fundamentally different scenarios in which optimal descriptions become simpler.

Distinguishing between these mechanisms is essential as they call for qualitatively different methods for improving prediction accuracy: a better microscopic model or a better effective model, respectively. Note again that for this result I have explicitly assumed the regime of sufficient data for model training; considering any data limitation would only make these behaviors more pronounced from symptoms of overfitting. Although this conceptual space can be explored *in silico*, is there any direct experimental evidence for the phenomenon of diversity-enhanced coarse-grainability?

4.3.2 Synthetic Gut Communities Exhibit Diversity-Enhanced Coarse-grainability

The dataset to which I apply this coarse-graining framework comes from experiments published in Clark *et al*. [135], which are essentially the same as the cartoon setup I described above. In their experiments, the library of taxa consisted of 25 species representing the major phyla of the human gut. They assemble synthetic communities in a defined medium using subsets of the 25 species, and then measure composition and function. In their work, the primary focus was on the production of butyrate, a well-characterized metabolite known for its many health benefits in the human gut [137–140]. In addition to this community function, they also measure the concentration of 3 other fermentation products: acetate, lactate, and succinate. Altogether, this provides 4 functional properties from which we can generate prediction-information diagrams and test the hypothesis that ecological diversity can facilitate coarse-grainability.

Figure 4.4 plots the estimated Pareto fronts (solid lines with error bars) of the prediction-information diagrams for each observable measured in communities of low diversity (1-5 species present) and high diversity (>20 species present); see Section 4.5.1 for details on an algorithm for deducing Pareto fronts in the space of coarse-grainings. Prediction error was computed using the standard out-of-sample RMSE of a first-order linear regression ansatz, with each point being normalized by the zero-information prediction $\mathcal{E}_0$ in the respective diversity context. Although this specific choice to use the linear regression class of models may seem too simplistic, the reasons for this choice are two-fold: (1) regression models are widely used for mapping composition to function (in contrast to predicting abundance itself, which is the goal of Lotka-Volterra models) and their performance has been validated in the original reference [135]; (2) the simplicity of a regression ansatz provides the tractability for deducing the coarse-graining Pareto front from the microscopic model itself (see Section 4.5.1).

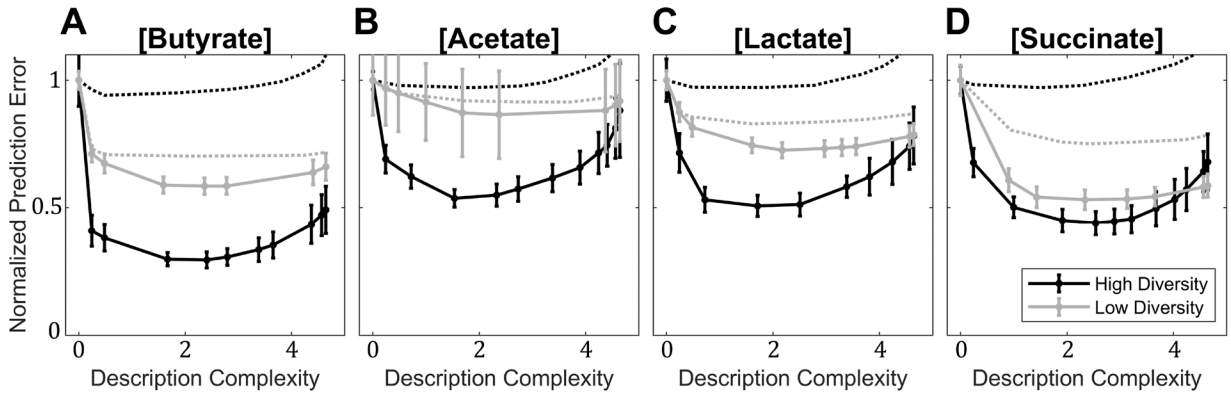

Figure 4.4 | Empirical examples of diversity-enhanced coarse-grainability (emergent simplicity). For each measured function (fermentation products) of the Clark *et al.* dataset, coarse-grained descriptions become more predictive in high diverse communities (>20 species) than they were in the low-diversity context (1-5 species). The inferred Pareto front (solid lines with error bars; see Section 4.5 for technical details) of each panel shifts down and to the left with increasing community diversity (compare to Figure 4.3B). Prediction power of each description was evaluated using a first-order linear regression model class, and RMSE values were normalized by the zero-information prediction error $\mathcal{E}(\Psi_Z)$ to fairly compare the two levels of diversity. Each point shows the median $\pm$ SD across random 50-50 splits of the data into training and testing sets.

Dashed lines show results for the same procedure performed on randomized data (see Section 4.5.4 for technical details).

Comparing low-diversity versus high-diversity in Figure 4.4, we consistently see across observables that the Pareto front shifts down and to the left in the prediction-information plane, indicating that simpler coarse-grained descriptions become more predictive at higher diversity. These results support the hypothesis that ecosystems can be more coarse-grainable at higher diversity, at least for predicting concentrations of fermentation products in communities. To test how surprising these observations are, each panel of Figure 4.4 also plots the Pareto front deduced after randomly permuting the abundance data for each species, shuffling only the samples in which each given species was present (Figure 4.4 dashed lines; see Section 4.5.4 for technical details). Indeed, this stringent randomization test breaks the diversity-enhanced coarse-grainability signature, resulting in the Pareto front shifting in the opposite direction when going from a low diverse context to a higher diverse context. Separately, looking under the hood of the prediction-information diagrams, scatter plots show that measured function indeed correlates better with the deduced coarse-grained variables (see Figure 4.8), further confirming the observations of Figure 4.4. Section 4.5.6 below provides another empirical example of diversity-enhanced coarse-grainability in a separate, independent dataset.

4.4 Conclusions and Discussion

Despite the functionally relevant diversity present at all levels of resolution in microbial ecosystems [47], there exist numerous examples where coarse representations can predict community-level functions without a full microscopic knowledge [45,46]. However, expert knowledge and intuition in each particular instance have been the primary means of identifying useful coarse-grained variables thus far. Quantitative principles for selecting an appropriate level

of description for modeling these systems, as well as a general understanding of the relative roles environment, community structure, and ecological context play in determining coarse-grainability, remain elusive.

In this work, I present a general framework for systematically comparing all possible compositional coarse-grained descriptions by quantifying their information content and their power to predict a functional property of interest. I show how this framework can be used to delineate observables for which an ecosystem can be coarse-grainable or not by way of enabling the identification of an optimal level of description within a given model class. Finally, by nuancing the ways in which the complexity of the selected optimum can decrease with increasing community diversity, I further demonstrate the framework's capacity for exploring emergent coarse-grainability in microbial ecosystems. Applying the framework to experimental data in which community diversity (species richness) is treated as a control parameter, I present empirical evidence for diversity-enhanced coarse-grainability: simpler coarse-grained descriptions become more predictive in a higher diverse context, despite being nominally more complex.

Uncovering this striking phenomenon provides hope for one day understanding microbial ecosystems in their natural ecological contexts, but one naturally wonders why it works. Unraveling the explicit mechanism underlying the empirical observation of emergent coarse-grainability presented here is beyond the reach of the framework in which it was demonstrated. Because I employ the simple model class of linear regression, the mappings from community composition to function are purely correlatory; the predictions a given description produces are not routed in any causal mechanism. However, by leveraging the statistical correlations between composition and function, the efficient coarse-grained descriptions constituting the Pareto fronts of Figure 4.4 have the potential for generating mechanistic hypotheses to probe with further

experiments. In fact, the results of Figure 4.4A for predicting butyrate production are consistent with the mechanistic explanation provided in Clark *et al.* (see Section 4.5.5 for further details): the butyrate-producing species *Anaerostipes caccae* alters its metabolism in a high-diversity context from a low butyrate producing per unit biomass state to high butyrate production per unit biomass depending on environmental pH and resource competition [135]. In some sense, only experiment can tell us the mechanistic story behind each property that exhibits diversity-enhanced coarse-grainability. For the dataset considered here, this thinking would mean that this effect presented by each observable (butyrate, acetate, lactate, and succinate) corresponds to 4 different mechanistic stories, each with their own idiosyncratic details. But perhaps there is a more general rule governing emergent coarse-grainability that encompasses each manifestation of the effect.

The concept of functional attractors from theoretical ecology [8,102] serves as a promising foundation for building a general understanding of emergent coarse-grainability. Though there has yet to be any direct supporting evidence from experiments, the theoretical statement posits that the collective functioning of ecosystems approaches an attractive state in the high diversity limit, regardless of their microscopic composition. I propose that the empirically observed diversity-enhanced coarse-grainability shown here is an example of functional attractors at play outside of just being conceptualized in theoretical models.

Figure 4.5A illustrates the idea of functional convergence in the context of the Clark *et al*. dataset, where the space of functional behaviors is unstructured in low-diversity communities but converges to some structured, low-dimensional manifold at higher diversity. The high-dimensionality of the unstructured low-diversity functional space would require descriptions with high information content for predicting community function. In contrast, if the accessible functional space collapses to a structured "banana" for higher diverse communities (as depicted),

and a compositional feature happens to correlate with its principal coordinate, this would explain the enhanced predictive power provided by coarse-grained composition that emerges.

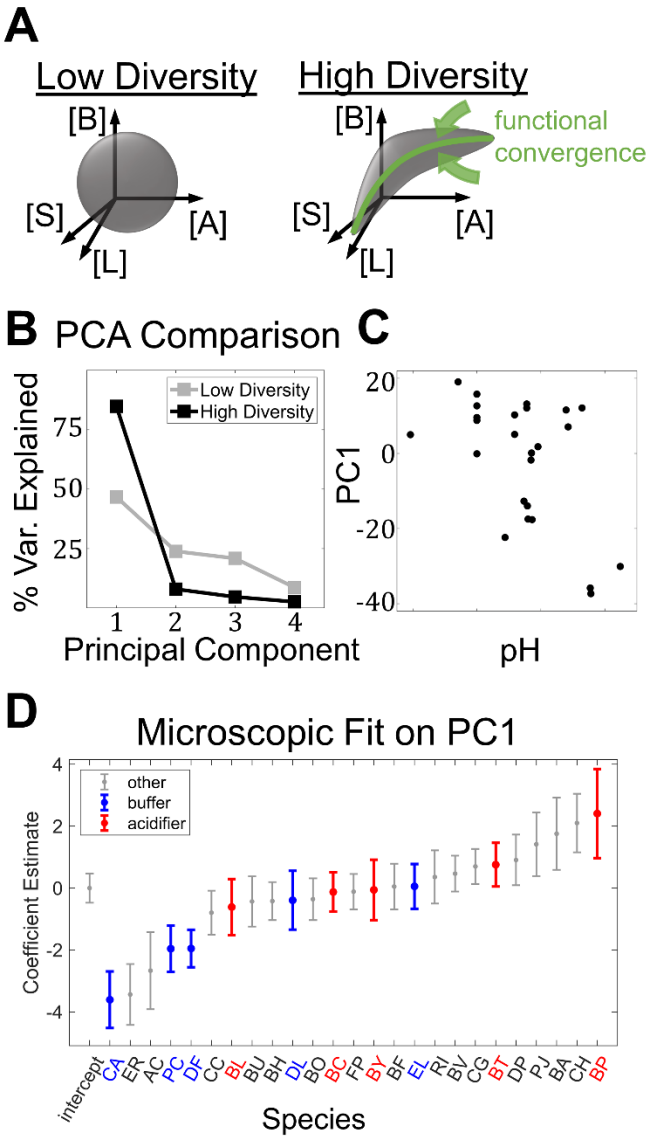

Figure 4.5 | Empirically observed emergent simplicity is consistent with functional attractor mechanism at high diversity. A: This cartoon illustrates the concept of a functional attractor in the 4-dimensional space of fermentation product concentrations measured in Clark *et al.* [135]. The attainable functions spanned by communities of low-diversity resembles an unstructured ball that fills out the full dimensionality of the space (left axes). In contrast, high-diversity communities assemble to some functionally attractive state, converging to a lower-dimensional manifold (or "banana"-fold if you will). If compositional features happen to correlate with the primary functional axis along the "banana", this would explain the observed enhancement in coarse-grainability at high-diversity shown in Figure 4.4. The next few panels exhibit evidence in support of this argument. **B:** Principal component analysis of the functional data (butyrate, acetate, lactate,

succinate concentrations) measured by Clark *et al.* in low- and high-diversity communities confirms that the functional space is substantially lower-dimensional in the high-diversity context, with a single component (PC1) explaining >80% of the variance. **C:** Scattering the projections of fermentation product data along PC1 obtained from PCA on high-diversity communities against measured pH (for communities in which it's reported). The resulting (negative) correlation suggests that environmental pH could be the functional "banana" coordinate determining the location in the 4-dimensional space illustrated in panel A. **D:** To relate compositional features to the suspected functional attractor, we perform a linear regression (as done throughout this work) of PC1 on the microscopic description (individual species abundances). Coloring the fitted coefficients of each species by their roles in shaping environmental pH (as defined by Clark *et al*.) further confirms the link between PC1 and pH, and is contained in compositional features. This result is consistent with the hypothesis that a functional attractor is responsible for the emergent coarse-grainability present in the experimental data analyzed here.

The available data from the Clark *et al.* experiment cannot explicitly determine whether or not the above argument is responsible for the results presented in this work, but Figure 4.5B-D provides some indirect supporting evidence. First, applying standard principal component analysis on the measured functions (butyrate, acetate, lactate, and succinate concentration; see Figure 4.5B) shows that the functional space is much lower dimensional in high-diversity communities than in low-diversity, with ~85% of the functional variance converging along the first principal component (PC1). Second, drawing from the mechanistic insights reported in the original reference [135], in Figure 4.5C I scatter the computed PC1 against pH in those high-diversity communities in which it was measured. The resulting (negative) correlation implicates the contribution of pH in determining the concentration of all 4 fermentation products, nominating pH to be a candidate causal link between these functional readouts. Indeed, the role of environmental pH in shaping metabolic rates of fermentation has been suggested elsewhere [141,142]. Finally, to connect this plausible mechanistic feature to community composition (albeit still in a correlatory manner), I fit the microscopic description (no grouping of species) using a first-order linear regression model to PC1 (using same methods described in 4.5.1). Figure 4.5D shows the fitted coefficients of each species, quantifying their contribution to PC1. Coloring each species by their identified

relationship with environmental pH in Clark *et al.* further supports the connection between PC1 and pH, as well as their correlation with community composition (species abundance). Note that these observations I present in Figure 4.5B-D are only *consistent* with the functional attractor picture, but still do not fully explain the phenomenon of diversity-enhanced coarse-grainability. The argument I thread above is speculative and requires further experimental validation.

As a final point, it's worth further contextualizing the empirical observations presented here by situating them relative to the current scope of theoretical models of ecology. First, the commonly used and influential Lotka-Volterra class of models, which express the dynamics of community composition through a pairwise interaction matrix encoding processes such as predation or cooperation, seems ill-suited for studying questions concerned with coarse-graining ecosystems. Specifically, the dependence of coarse-grainability on diversity is likely difficult to capture in general within the Lotka-Volterra class of models. This essentially follows by definition of these models: any reasonable property of interest (e.g., total biomass or equilibrium abundance) can be trivially related to community composition as a linear combination in terms of the interaction coefficients, and since these are assumed to be context-independent, it is clear that diversity cannot affect coarse-grainability (see Section 4.5.7 for further discussion). Second, the functional attractor concept described above originates from work analyzing the infinite diversity limit of random ecosystems within the setting of a consumer-resource model parameterized by an unstructured ensemble [8]. Though the question of coarse-grainability was not explicitly addressed in the original work, it seems clear that exploring the concept of coarse-graining in ecosystems requires incorporating community structure in some way. The model for studying ensembles of random, *structured* ecosystems presented in Chapter 3 provides a stepping stone in this direction, with the

potential for disentangling the relative roles community structure and diversity play in shaping coarse-grainability.

4.5 Technical Details

4.5.1 Training and Testing of Linear Regression Models

As noted in the main text, any model class can be considered within the framework presented in Section 4.2; one just needs to be specified in order to explicitly compute the prediction error of a given description. The model class I chose for analyzing coarse-grained descriptions of the Clark *et al.* dataset in this work was first-order linear regression for reasons listed in Section 4.3.2. In this specific context, the linear regression models generally take the following form:

$$\hat{Y}_\mu = \tilde{\lambda}_0 + \sum_{\alpha \in \Psi} \tilde{\lambda}_\alpha \tilde{n}_{\alpha\mu},$$

where $\hat{Y}_\mu$ is the model's prediction of property Y measured in the μth community and $\tilde{\lambda}_\alpha$ are the regression coefficients of each coarse-grained group to be determined by fitting to a training dataset.

Each regression model (coarse-grained and microscopic) is trained and tested on the same randomly drawn datasets. After accounting for replicate communities in each diversity bin (see below), 50% of communities are randomly designated for model training/fitting and the other 50% are reserved for testing/validating. Model fitting was performed using the standard ordinary least-squares method of MATLAB's 'fitlm' algorithm. The predictions of the fitted model where then compared to the measured values of the test set using standard root-mean-square error. This was repeated for 100 random 50-50 splits of the data, taking the median and standard deviation across these splits as the values for $\mathcal{E}(\Psi)$ and error bars plotted in Figure 4.4.

4.5.2 Deducing the Pareto Front of Efficient Coarse-grainings

The Pareto front of coarse-grainings in the prediction-error plane correspond to the coarse-grained partitions Ψ with minimal prediction error $\mathcal{E}$ for a given amount of information I. Ideally, we would like to know this exactly. The brute force approach of performing an exhaustive search through the space of coarse-grainings becomes practically infeasible computationally when library of strains becomes large. This is because the number of possible ways to partition the strains becomes exponentially large: for S strains, the total number is the Bell number B_S, which is too large to enumerate fully. For example, the 25 taxa used in the Clark *et al.* already results in $B_{25} \simeq 5 \times 10^{18}$. This necessitates developing search algorithms to efficiently explore this vast landscape.

With the tractability a linear regression ansatz provides, the efficient coarse-grainings making up the Pareto front can be deduced from the microscopic model itself, as I describe below. In general, because the microscopic description corresponds to resolving each taxa into separate groups, the microscopic model will not be available when the number of taxa is very large or data is scarce for model fitting. Fortunately, the Clark *et al.* dataset contains enough sample communities in each diversity bin to constrain the microscopic model of 25 species, permitting the following deduction procedure.

In the context of linear regression, the coarse-graining procedure I consider here (partitioning taxa into groups and combining the corresponding input variable by summing their abundance) can be mapped to a regularization scheme. Figure 4.6A illustrates this interpretation for the simplest case of first-order linear regression, but the mapping straightforwardly extends to higher-order terms.

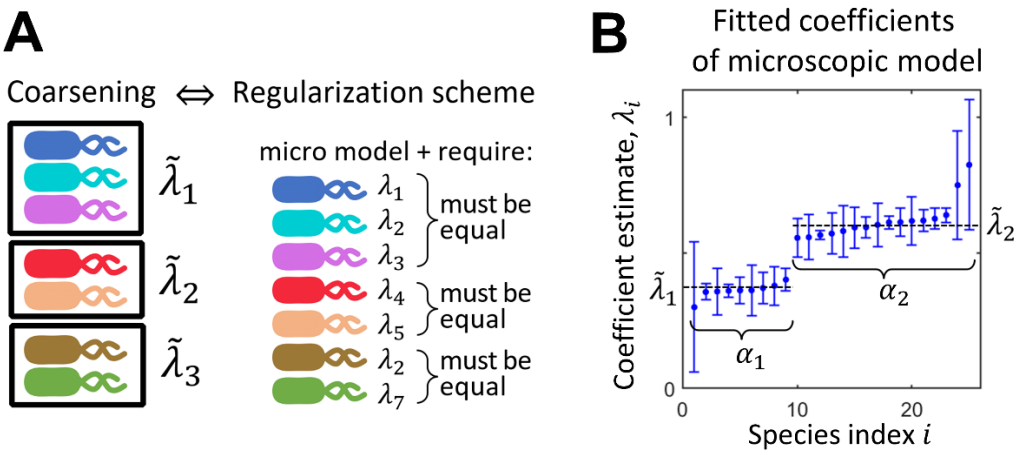

Figure 4.6 | A: In the context of a linear regression model class, the coarse-graining scheme we develop in this work maps onto a regularization scheme. The trivial example depicted in this panel illustrates how grouping strains in a given coarse-graining, such that each strain in a given group is now prescribed the same coefficient $\tilde{\lambda}_\alpha$, is equivalent to regularizing the coefficients of the microscopic model. **B:** Plotting example coefficients λ_i for each species before coarse-graining shows that efficient coarse-grainings can be inferred from the microscopic model (again for first-order linear regression). In this illustration, there are two clear groups to form a two-variable coarse-grained model with coefficients $\tilde{\lambda}_\alpha$. Error bars correspond to uncertainty in fit.

Efficient coarse-grainings then correspond to grouping strains with similar regression coefficients. Let λ_i denote the estimated coefficients of the regression model fit to microscopic variables (individual strains i). For illustration purposes, imagine the values of these coefficients are found to be as depicted in Figure 4.6B. From this coefficient structure we can immediately deduce which two-group coarse-graining would perform best: $\sum_i \lambda_i n_i \approx \tilde{\lambda}_1 \sum_{i\in\alpha_1} n_i + \tilde{\lambda}_2 \sum_{i\in\alpha_2} n_i$. Thus, identifying an efficient coarse-graining reduces to a clustering problem. Specifically, I implement a generalized K-means clustering algorithm where the cost function accounts for the uncertainty carried by each coefficient estimate: $\text{cost}(\Psi) = \sum_i \frac{\lambda_i - \langle\lambda_j\rangle_{\alpha_j=\alpha_i}}{\sigma_i}$, where angular brackets denote averaging over all taxa assigned to the same group. In order to cluster regression coefficients in a sensible fashion, I standardize the measured abundances of each species (input variables) to have 0 mean and unit variance in all regression fitting so that all λ_i are on the same scale. To obtain

accurate estimates initially, the microscopic regression models are fit on all the data before deducing the Pareto front coarse-grainings via this clustering procedure.

4.5.3 Mid-Diversity Communities from Clark *et al.*

To avoid clutter, Figure 4.4 only shows the results of applying the coarse-graining framework to communities of low-diversity (sampling subsets of 1-5 species) and of high-diversity (>20 species) measured in the Clark *et al.* experiment. Their original work also measured communities of intermediate species richness, from which we form a third bin we call mid-diversity (10-15 species). Applying the same methods described above, Figure 4.7 shows the coarse-graining Pareto front for predicting each measured function for all three diversity bins. Despite idiosyncrasies in each case, overall the cross-over from low- to high-diversity qualitatively appears to be continuous with some potentially rich behavior/trends to explore in future work.

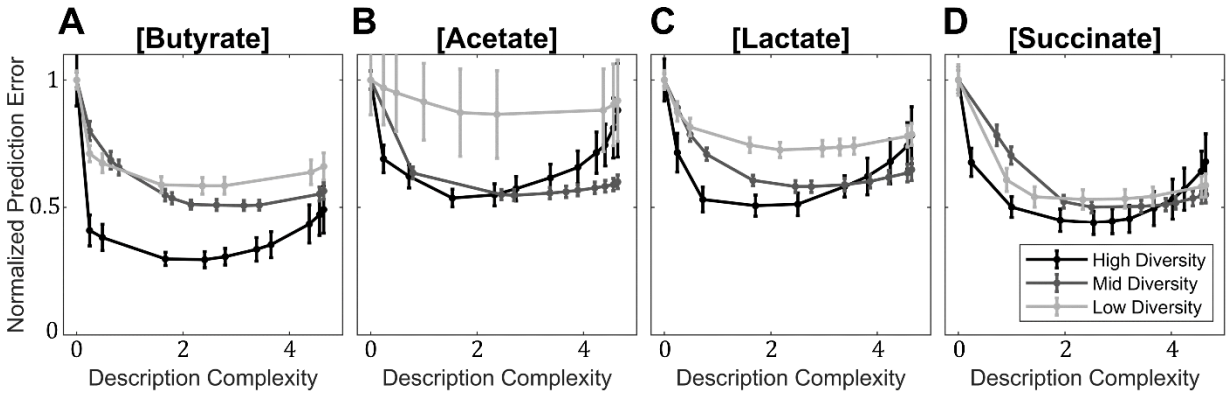

Figure 4.7 | Replotting the inferred Pareto fronts in low- and high-diversity contexts of Figure 4.4 with the addition of those inferred from data measured in mid-diversity communities (10-15 species) in the Clark *et al.* experiments.

4.5.4 Randomization Tests

Checking that the emergent coarse-grainability observed in Figure 4.4 is not purely due to chance, I compare the results to Pareto fronts deduced after randomizing the dataset. 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 sentiment in the Clark *et al.* dataset, I aim to keep the abundance statistics of each species across communities fixed; so for each species, I randomly permute its abundance in communities in which the species is initially present. For this reason, the randomization of low-diversity communities (light gray dashed line in Figure 4.4) only slightly breaks the coarse-grainability observed in the non-randomized Pareto front presumably because in monocultures and cocultures, the dominate structure is contained in the presence-absence of species.

4.5.5 Scattering Measured Function versus Coarse-grained Variables

Further checking the results of Figure 4.4, I convert the information encoded in the prediction-information diagrams into more digestible scatter plots shown in Figure 4.8. Plotting the measured function in low-diversity and high-diversity communities versus the predictive coarse-grained variables (combined species abundance) identified by the respective high-diversity Pareto front, one can see that indeed coarse-grained composition correlates better with function in a higher diverse context. The left column of panels indicate which coarse-grained description is being plotted on the x-axis of the scatter plots. Each selected coarse-graining consists of two groups: the combined abundance of taxa in group 1 are shown in orange, while group 2 is colored purple; each with the same y-axis.

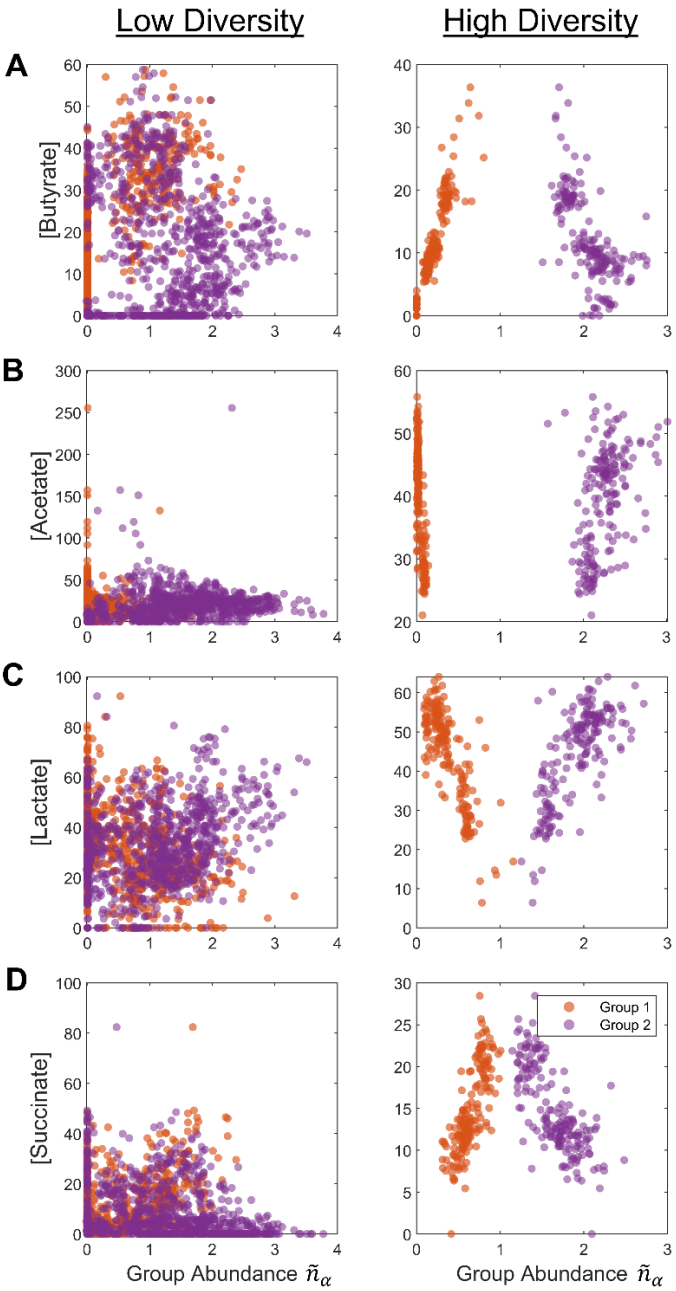

Figure 4.8 | Scatter plots of community function (concentration of fermentation products) versus coarse-grained (group) abundance measured in low-diversity (left column) and high-diversity (right column) communities. As described in Section 4.2.2, species abundances (measured as OD in Clark *et al.*) are summed to give the group abundances that are plotted. Groups 1 and 2 are obtained from the Pareto fronts inferred for predicting each observable in high-diversity, and are therefore consist of different species across panels (see text). The stronger correlation between coarse-grained compositional variables and function at high diversity verifies the results of Figure 4.4.

The highly efficient (very low entropy) coarse-graining for predicting butyrate corresponds to the species *Anaerostipes caccae* (AC) in a group by itself and all others in group 2. As noted in Clark *et al.*, AC is capable of producing butyrate, whose presence or absence strongly determines the concentration of butyrate in high-diversity communities, despite being 1 of 5 butyrate producing species used in their work. Digging into this further, Clark *et al.* performed sought out to decipher the mechanism behind this observation and found two key lines of evidence explaining this behavior: (1) the 4 other butyrate producing species are inhibited by the production of hydrogen sulfide from a non-butyrate producer *Desulfovibrio piger* (DP); (2) AC switches its metabolic behavior depending on environmental pH and availability of energy resources set by other community members, modulating its production of butyrate. Although this reasoning aids in checking the consistency of the framework's output, it's worth noting that the coarse-graining framework independently re-discovered this correlation without incorporating any of the mechanistic knowledge.

Since the other fermentation products (acetate, lactate, and succinate) were not the primary focus of the original work by Clark *et al.*, there is less mechanistic support for the predictive coarse-grainings for these functions. Searching elsewhere for biological justification, I find that my results are at least consistent with observations in the literature. For example, group 1 for predicting acetate concentration consists of just the single butyrate-producing species *Eubacterium rectale* (ER), which is reported to be a net consumer of acetate when producing butyrate [143]. As for succinate, the major bacterial group that employs the succinate route in forming fermentation end products is the phyla Bacteroidetes [143]; consistent with the fact that more than half the constituent species of group 1 in the predictive coarse-grained description for succinate. It's worth re-emphasizing that the groupings identified within the framework developed here are not direct

functional groups given the correlative (not necessarily casual) nature of the analysis; further experimental investigation is required to solidify any mechanistic links.

4.5.6 Diversity-enhanced Coarse-grainability in an Independent Dataset

As far as I am aware, the next best microbial dataset currently available in which species richness (diversity) is varied over a reasonable range (at least an order of magnitude) for defining low- and high-diversity regimes comes from experiments presented in Kehe *et al.* [136]. The community function of interest in this work was the promotion of a known plant symbiont *Herbaspirillum frisingense* measured in terms of the microbes abundance. Synthetic communities were assembled from soil isolates, ranging from 1 to 14 strains sampled at a time (including the symbiont strain).

Unlike the Clark *et al.* dataset, which provides information of species abundance, the only compositional information available from this experiment is reported presence or absence of each species in a given community. The generality of the framework I present allows for this data to still be a valid input: rather than the coarse-grained variables being combined abundances of taxa within a group, they instead indicate when any member of the group is present or if all are absent. Mimicking as best as possible the diversity bins defined for the Clark *et al.* dataset, communities in which 1-5 strains are present is taken as low-diversity, and high-diversity is taken as communities with >10 strains present. Applying the same methods for generating Pareto fronts in the prediction-information plane as outlined above, I observe that this independent dataset also exhibits diversity-enhanced coarse-grainability as shown in Figure 4.9. Although it is (perhaps more) remarkable that this system is more coarse-grainable at higher diversity given the very minimal compositional information, I am unfortunately unable to check if this empirical observation is consistent with the theoretical functional attractor picture because only one function was measured.

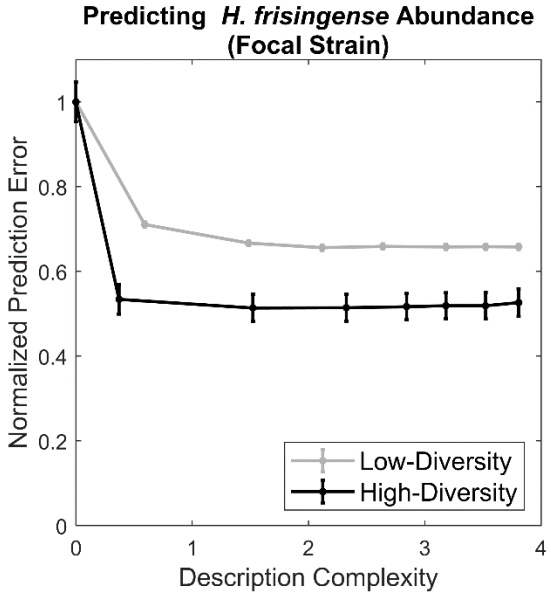

Figure 4.9 | The independent dataset from Kehe *et al.* [136] provides another empirical example of diversity-enhanced coarse-grainability. In this case, predicting the abundance of a focal strain is more coarse-grainable in high diversity communities than in low diversity communities. Pareto fronts are inferred via the same procedure used for analyzing the Clark *et al.* dataset; i.e., points and error bars are median $\pm$ SD across (normalized) RMSE of linear regression models trained on 100 random 50-50 splitting of the data into training and testing sets.

4.5.7 Lotka-Volterra Models and Coarse-grainability

As mentioned in Section 4.4, the Lotka-Volterra class of ecological models aim at describing the dynamics of abundances of taxa. A standard form of this type of model can be expressed as

$$\dot{n}_i = n_i \left(r_i - \sum_j A_{ij} n_j \right),$$

where n_i are again abundances of the microscopic taxa (e.g., strains), r_i is a context-independent growth rate of taxa i, and A_{ij} are context-independent interactions between taxa i and taxa j. I now catalogue obvious observables of interest one might aim to predict about a given community and argue that each cannot exhibit the phenomenon of diversity-enhanced coarse-grainability.

The coarse-grainability of the first several community-level properties are trivially *independent* of community diversity. Each can be seen by inspection. First, consider trying to predict the invasion rate of a strain i (e.g., a pathogen) into an assembled community missing i, denoted as ρ_i. The ability of the focal strain to invade is simply given by the growth rate set by the equilibrium abundances n_j^* of those strains in the assembled community: $\rho_i = r_i - \sum_j A_{ij} n_j^*$. Notice that this being a linear combination of strain abundances means that the coarse-grainability of this property in a regression model class is trivially determined by the choice of A_{ij}. Similarly, obtaining the equilibrium condition of the Lotka-Volterra dynamics for a strain that does not go extinct, the equilibrium abundance of a focal strain in a community is also a trivial linear combination of strain abundances:

$$n_i^* = \frac{r_i - \sum_{j \neq i} A_{ij} n_j^*}{A_{ii}}.$$

Likewise, it then follows that the total biomass of a community at equilibrium $\sum_i n_i^*$ is also a trivial observable in this sense.

One can, however, imagine the following nontrivial observable. Reconsider the idea of attempting to invade an already assembled community with an initially absent strain i, but rather than focusing on its invasion rate, instead consider trying to predict its abundance at a *new* equilibrium post-invasion as a function of the pre-invasion abundances of the old equilibrium. Here, this property is not as straightforward as the above examples because the new equilibrium abundance of any given strain is determined by the abundances of the surviving strains at this new equilibrium, which in general could be an entirely different set of species or entirely different abundances. Exploring this in simulations of a Lotka-Volterra model of the form studied in [109,110], I mimic the

experimental design of Clark *et al.* and assemble communities of low-diversity (1-5 strains) and high-diversity (21-25 strains), with parameters drawn from random ensembles per standard protocols: A_{ij} are normally distributed. Figure 4.10 shows the E-I diagram for predicting the equilibrium abundance of a focal strain after its introduction to pre-assembled communities in each diversity context. This proof-of-principle exploration and the examples catalogued above demonstrates the challenge of capturing emergent coarse-grainability within models of random, unstructured ecosystems.

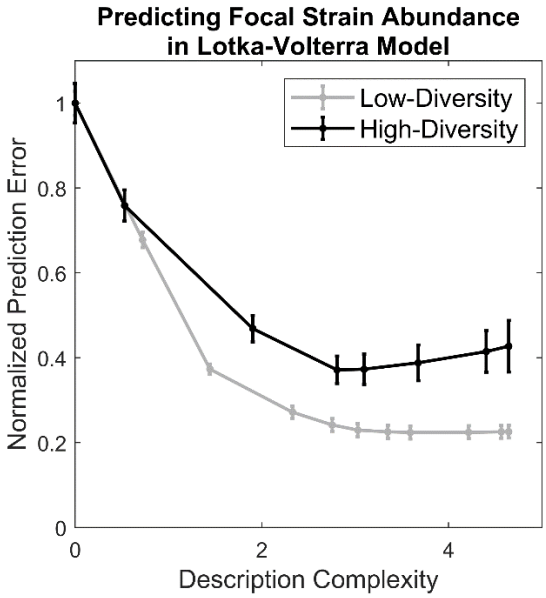

Figure 4.10 | An example of random, unstructured ecosystems within a Lotka-Volterra model are not necessarily any less coarse-grainable for predicting focal strain abundance in high-diversity communities (see text). This example illustrates the difficulty in capturing emergent coarse-grainability in this standard model of ecology.

Chapter 5: Summary and Outlook

The previous chapters contribute to the symbiotic interface shared by statistical physics and theoretical ecology and evolution – biology's yin-and-yang. Part of the rich history of this interface has come from physicists drawing on techniques from other areas, such as disordered systems, to describe eco-evolutionary phenomena with an ensemble approach. This effort has seen a great deal of success, providing interpretation of broad statistical trends observed across ecosystems and distinguishing which patterns can and cannot be captured in simple, null models of unstructured populations. Identifying the failure modes enables the exciting opportunity to begin pushing the scope of current concepts in theoretical physics by generating new frameworks and techniques for tackling these challenges. This dissertation focuses on one such problem area that falls outside the range of conventional statistical physics methods: addressing the observed structures in highly diverse ecosystems and their coarse-grainability.

As discussed throughout the previous chapters, until recently, all models and analytical frameworks of large-N (diverse) ecology hinge on *unstructured* random ensembles to describe ecosystems, omitting the key ingredient responsible for predictive coarse-grained descriptions in empirical examples. Moreover, the renormalization group construction used to explain the emergence of predictive effective models in physics has no known analogue in an ecological context. Although developing a new general framework for the statistical physics of structured, heterogeneous complex systems would fill this gap, this dissertation provides only a step toward this long-term goal. Collectively, the key contributions of the chapters I present above consist of two frameworks: (1) a modeling framework that uses evolution for organically generating ensembles of random, structured ecosystems; (2) a theoretical framework for investigating the emergent coarse-grainability of such ecosystems. I use these developments to theoretically and

empirically demonstrate that simple coarse-grained descriptions counterintuitively emerge in higher diverse communities, and suggest that coarse-grainability is maximized when an ecosystem is assembled in its natural environment.

This body of work highlights the importance of incorporating structure in models of diverse ecosystems, focusing specifically on the observed structure in functional trait distributions across taxa. Ecosystems exhibit many other forms of structure that influence evolution and responses to perturbations. For example, bacterial communities found in nature are often spatially structured, forming biofilms or other spatially organized architectures. Numerous investigations into the evolutionary dynamics of microbes find qualitatively different outcomes when tracking populations in spatially heterogeneous communities grown on a petri dish versus in a well-mixed, homogeneous environment of a test tube [144–146]. In addition to space, microbial communities also self-organize into temporal structures due to differing rates and efficiencies of resource utilization, segmenting time into "temporal niches": each taxa has a boom period then a bust period [21,147,148]. What is a general framework that encapsulates each of these mechanisms of structure? And how do we extend the powerful analytical techniques provided by the statistical physics of disordered systems from unstructured to structured regimes? Answering these questions provides an exciting opportunity to advance both theoretical ecology/evolution and physics.

References

[1] A. J. Lotka, *Analytical Note on Certain Rhythmic Relations in Organic Systems*, Proceedings of the National Academy of Sciences **6**, 410 (1920).

[2] R. MacArthur and R. Levins, *The Limiting Similarity, Convergence, and Divergence of Coexisting Species*, The American Naturalist **101**, 377 (1967).

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

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

[5] E. Wigner, *Characteristic Vectors of Bordered Matrices with Infinite Dimensions Ann. of Math., 62*, (1955).

[6] M. Advani, G. Bunin, and P. Mehta, *Statistical Physics of Community Ecology: A Cavity Solution to MacArthur's Consumer Resource Model*, Journal of Statistical Mechanics: Theory and Experiment **2018**, 033406 (2018).

[7] W. Cui, R. Marsland III, and P. Mehta, *Effect of Resource Dynamics on Species Packing in Diverse Ecosystems*, Physical Review Letters **125**, 048101 (2020).

[8] L. Fant, I. Macocco, and J. Grilli, *Eco-Evolutionary Dynamics Lead to Functionally Robust and Redundant Communities*, BioRxiv (2021).

[9] J. E. Goldford, N. Lu, D. Bajić, S. Estrela, M. Tikhonov, A. Sanchez-Gorostiaga, D. Segrè, P. Mehta, and A. Sanchez, *Emergent Simplicity in Microbial Community Assembly*, Science **361**, 469 (2018).

[10] 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**, e1006793 (2019).

[11] R. Marsland, W. Cui, and P. Mehta, *A Minimal Model for Microbial Biodiversity Can Reproduce Experimentally Observed Ecological Patterns*, Scientific Reports **10**, 1 (2020).

[12] M. Tikhonov, *Community-Level Cohesion without Cooperation*, Elife **5**, e15747 (2016).

[13] M. Tikhonov and R. Monasson, *Collective Phase in Resource Competition in a Highly Diverse Ecosystem*, Physical Review Letters **118**, 048103 (2017).

[14] M. R. Domingo-Sananes and J. O. McInerney, *Mechanisms That Shape Microbial Pangenomes*, Trends in Microbiology **29**, 493 (2021).

[15] O. X. Cordero and M. F. Polz, *Explaining Microbial Genomic Diversity in Light of Evolutionary Ecology*, Nature Reviews Microbiology **12**, 263 (2014).

[16] O. M. Maistrenko et al., *Disentangling the Impact of Environmental and Phylogenetic Constraints on Prokaryotic Within-Species Diversity*, The ISME Journal **14**, 1247 (2020).

[17] H. Wu and E. Moore, *Association Analysis of the General Environmental Conditions and Prokaryotes' Gene Distributions in Various Functional Groups*, Genomics **96**, 27 (2010).

[18] C. O. Webb, D. D. Ackerly, M. A. McPeek, and M. J. Donoghue, *Phylogenies and Community Ecology*, Annual Review of Ecology and Systematics **33**, 475 (2002).

[19] K. Isobe, N. J. Bouskill, E. L. Brodie, E. A. Sudderth, and J. B. H. Mertiny, *Phylogenetic Conservation of Soil Bacterial Responses to Simulated Global Changes*, Philosophical Transactions of the Royal Society B **375**, (2020).

[20] C. A. Serván, J. A. Capitán, Z. R. Miller, and S. Allesina, *Effects of Phylogeny on Coexistence in Model Communities*, BioRxiv (2020).

[21] B. H. Good, M. J. McDonald, J. E. Barrick, R. E. Lenski, and M. M. Desai, *The Dynamics of Molecular Evolution over 60,000 Generations*, Nature **551**, 45 (2017).

[22] D. E. Rozen and R. E. Lenski, *Long-Term Experimental Evolution in Escherichia Coli. VIII. Dynamics of a Balanced Polymorphism*, The American Naturalist **155**, 24 (2000).

[23] D. S. Treves, *Evolution of Acetate Crossfeeding Polymorphisms in Long-Term Populations of Escherichia Coli* (University of Michigan, 1998).

[24] M. A. Kinnersley, W. E. Holben, and F. Rosenzweig, *E Unibus Plurum: Genomic Analysis of an Experimentally Evolved Polymorphism in Escherichia Coli*, PLoS Genetics **5**, e1000713 (2009).

[25] R. F. Rosenzweig, R. Sharp, D. S. Treves, and J. Adams, *Microbial Evolution in a Simple Unstructured Environment: Genetic Differentiation in Escherichia Coli.*, Genetics **137**, 903 (1994).

[26] S. Azaele, S. Suweis, J. Grilli, I. Volkov, J. R. Banavar, and A. Maritan, *Statistical Mechanics of Ecological Systems: Neutral Theory and Beyond*, Rev. Mod. Phys. **88**, 035003 (2016).

[27] K. S. McCann, *The Diversity–Stability Debate*, Nature **405**, 6783 (2000).

[28] C. Ricotta, *From Theoretical Ecology to Statistical Physics and Back: Self-Similar Landscape Metrics as a Synthesis of Ecological Diversity and Geometrical Complexity*, Ecological Modelling **125**, 245 (2000).

[29] A. Posfai, T. Taillefumier, and N. S. Wingreen, *Metabolic Trade-Offs Promote Diversity in a Model Ecosystem*, Physical Review Letters **118**, 028103 (2017).

[30] I. Thompson, B. Mackey, S. McNulty, A. Mosseler, and others, *Forest Resilience, Biodiversity, and Climate Change*, in *A Synthesis of the Biodiversity/Resilience/Stability*

Relationship in Forest Ecosystems. Secretariat of the Convention on Biological Diversity, Montreal. Technical Series, Vol. 43 (2009), pp. 1–67.

[31] S. Lavergne, N. Mouquet, W. Thuiller, and O. Ronce, *Biodiversity and Climate Change: Integrating Evolutionary and Ecological Responses of Species and Communities*, Annual Review of Ecology, Evolution, and Systematics **41**, 321 (2010).

[32] J. K. Jansson and K. S. Hofmockel, *Soil Microbiomes and Climate Change*, Nature Reviews Microbiology **18**, 35 (2020).

[33] R. D. Bardgett, C. Freeman, and N. J. Ostle, *Microbial Contributions to Climate Change through Carbon Cycle Feedbacks*, ISME J **2**, 8 (2008).

[34] M. Zhao, K. Xue, F. Wang, S. Liu, S. Bai, B. Sun, J. Zhou, and Y. Yang, *Microbial Mediation of Biogeochemical Cycles Revealed by Simulation of Global Changes with Soil Transplant and Cropping*, ISME J **8**, 2045 (2014).

[35] G. Hardin, *The Competitive Exclusion Principle*, Science **131**, 1292 (1960).

[36] R. A. Armstrong and R. McGehee, *Competitive Exclusion*, The American Naturalist **115**, 151 (1980).

[37] J. Huisman and F. J. Weissing, *Biodiversity of Plankton by Species Oscillations and Chaos*, Nature **402**, 407 (1999).

[38] M. Pearce, A. Agarwala, and D. S. Fisher, *Stabilization of Extensive Fine-Scale Diversity by Spatio-Temporal Chaos*, BioRxiv 736215 (2019).

[39] A. Mahadevan, M. T. Pearce, and D. S. Fisher, *Spatiotemporal Ecological Chaos Enables Gradual Evolutionary Diversification Without Niches or Tradeoffs*, BioRxiv 2022 (2022).

[40] R. E. Beardmore, I. Gudelj, D. A. Lipson, and L. D. Hurst, *Metabolic Trade-Offs and the Maintenance of the Fittest and the Flattest*, Nature **472**, 342 (2011).

[41] A. Erez, J. G. Lopez, B. G. Weiner, Y. Meir, and N. S. Wingreen, *Nutrient Levels and Trade-Offs Control Diversity in a Serial Dilution Ecosystem*, Elife **9**, e57790 (2020).

[42] S. Estrela, J. Diaz-Colunga, J. C. Vila, A. Sanchez-Gorostiaga, and A. Sanchez, *Diversity Begets Diversity under Microbial Niche Construction*, BioRxiv 2022 (2022).

[43] R. Starke, P. Capek, D. Morais, N. Jehmlich, and P. Baldrian, *Explorative Meta-Analysis of 377 Extant Fungal Genomes Predicted a Total Mycobiome Functionality of 42.4 Million KEGG Functions*, Frontiers in Microbiology **11**, (2020).

[44] A. Eng and E. Borenstein, *Taxa-Function Robustness in Microbial Communities*, Microbiome **6**, (2018).

[45] A. Jeglot, J. Audet, S. R. Sorensen, K. Schnorr, F. Plauborg, and L. Elsgaard, *Microbiome Structure and Function in Woodchip Bioreactors for Nitrate Removal in Agricultural Drainage Water*, Frontiers in Microbiology **12**, 2284 (2021).

[46] S. Bertacchi, M. Ruusunen, A. Sorsa, A. Sirviö, and P. Branduardi, *Mathematical Analysis and Update of ADM1 Model for Biomethane Production by Anaerobic Digestion*, Fermentation **7**, 237 (2021).

[47] 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 (2022).

[48] J. O. McInerney, A. McNally, and M. J. O'Connell, *Why Prokaryotes Have Pangenomes*, Nat Microbiol **2**, 1 (2017).

[49] B. A. Niccum, E. K. Kastman, N. Kfoury, A. Robbat Jr, and B. E. Wolfe, *Strain-Level Diversity Impacts Cheese Rind Microbiome Assembly and Function*, Msystems **5**, e00149 (2020).

[50] T. D. Lieberman, K. B. Flett, I. Yelin, T. R. Martin, A. J. McAdam, G. P. Priebe, and R. Kishony, *Genetic Variation of a Bacterial Pathogen within Individuals with Cystic Fibrosis Provides a Record of Selective Pressures*, Nature Genetics **46**, 82 (2014).

[51] A. Tett et al., *The Prevotella Copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations*, Cell Host & Microbe **26**, 666 (2019).

[52] S. Zhao, T. D. Lieberman, M. Poyet, K. M. Kauffman, S. M. Gibbons, M. Groussin, R. J. Xavier, and E. J. Alm, *Adaptive Evolution within Gut Microbiomes of Healthy People*, Cell Host & Microbe **25**, 656 (2019).

[53] M. A. Lawson, I. J. O'Neill, M. Kujawska, S. G. Javvadi, A. Wijeyesekera, Z. Flegg, L. Chalklen, and L. J. Hall, *Breast Milk-Derived Human Milk Oligosaccharides Promote Bifidobacterium Interactions within a Single Ecosystem*, The ISME Journal **14**, 635 (2020).

[54] M. Roodgar et al., *Longitudinal Linked-Read Sequencing Reveals Ecological and Evolutionary Responses of a Human Gut Microbiome during Antibiotic Treatment*, Genome Research (2021).

[55] K. Jordan, Y. I. Wolf, and E. V. Koonin, *No Simple Dependence between Protein Evolution Rate and the Number of Protein-Protein Interactions: Only the Most Prolific Interactors Tend to Evolve Slowly*, BMC Evol. Biol. **3**, (2003).

[56] T. Friedlander, R. Prizak, N. H. Barton, and G. Tkacik, *Evolution of New Regulatory Functions on Biophysically Realistic Fitness Landscapes*, Nat. Commun **8**, (2017).

[57] G. Reddy and M. M. Desai, *Global Epistasis Emerges from a Generic Model of a Complex Trait*, ELife 10:e64740 (2021).

[58] N. Tokuriki and D. S. Tawfik, *Stability Effects of Mutations and Protein Evolvability*, Current Opinion in Structural Biology **19**, 596 (2009).

[59] J. D. Bloom, S. T. Labthavikul, C. R. Otey, and F. H. Arnold, *Protein Stability Promotes Evolvability*, PNAS **103**, 5869 (2006).

[60] M. Goldsmith and D. S. Tawfik, *Potential Role of Phenotypic Mutations in the Evolution of Protein Expression and Stability*, PNAS **106**, 6197 (2009).

[61] A. E. Hirsh and H. B. Fraser, *Protein Dispensability and Rate of Evolution*, Nature **411**, 1046 (2001).

[62] J. Zhang and X. He, *Significant Impact of Protein Dispensability on the Instantaneous Rate of Protein Evolution*, Mol. Biol. Evol. **22**, 1147 (2005).

[63] W. P. Wall, A. E. Hirsh, H. B. Fraser, J. Kumm, G. Giaever, M. B. Eisen, and M. W. Feldman, *Functional Genomic Analysis of the Rates of Protein Evolution*, PNAS **102**, 5483 (2005).

[64] B-Y. Liao, N. M. Scott, and J. Zhang, *Impacts of Gene Essentiality, Expression Pattern, and Gene Compactness on the Evolutionary Rate of Mammalian Proteins*, Mol. Biol. and Evol. **23**, 2072 (2006).

[65] D. A. Drummond, J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold, *Why Highly Expressed Proteins Evolve Slowly*, Proc. Natl. Acad. Sci. U.S.A. **102**, 14338 (2005).

[66] J. Zhang and J. -R. Yang, *Determinants of the Rate of Protein Sequence Evolution*, Nature Reviews Genetics 409 (2015).

[67] M. A. Savageau, *Design of Molecular Control Mechanisms and the Demand for Gene Expression*, Proc. Natl. Acad. Sci. U.S.A. **74**, 5647 (1977).

[68] U. Gerland and T. Hwa, *Evolutionary Selection between Alternative Modes of Gene Regulation*, Proc. Natl. Acad. Sci. U.S.A. **106**, 8841 (2009).

[69] R. Kafri, M. Springer, and Y. Pilpel, *Genetic Redundancy: New Tricks for Old Genes*, Cell **136**, 389 (2009).

[70] V. S. Cooper and R. E. Lenski, *The Population Genetics of Ecological Specialization in Evolving Escherichia Coli Populations*, Nature **407**, 736 (2000).

[71] T. J. Kawecki, N. H. Barton, and J. D. Fry, *Mutational Collapse of Fitness in Marginal Habitats and the Evolution of Ecological Specialisation*, J. Evol. Biol. **10**, 407 (1997).

[72] M. Tikhonov, S. Kachru, and D. S. Fisher, *A Model for the Interplay between Plastic Tradeoffs and Evolution in Changing Environments*, Proc. Natl. Acad. Sci. U.S.A. **117**, 8934 (2020).

[73] J. H. Gillespie, *A Simple Stochastic Gene Substitution Model*, Theor. Popul. Biol. **23**, 202 (1983).

[74] M. Kimura, *On the Probability of Fixation of Mutant Genes in a Population*, Genetics **47**, 713 (1962).

[75] D. Ascencio, S. Ochoa, L. Delaye, and A. DeLuna, *Increased Rates of Protein Evolution and Asymmetric Deceleration after the Whole-Genome Duplication in Yeasts*, BMC Evo. Bio. **17**, (2017).

[76] S. K. Strauss, D. Schirman, G. Jona, A. N. Brooks, A. M. Kunjapur, A. N. N. Ba, and et al, *Evolthon: A Comunity Endeavor to Evolve Lab Evolution*, PLos Biol **17**, e3000182 (2019).

[77] B. Steinberg and M. Ostermeier, *Environmental Changes Bridge Evolutionary Valleys*, Science Advances **2**, e1500921 (2016).

[78] A. F. Bitbol and D. J. Schwab, *Quantifying the Role of Population Subdivision in Evolution on Rugged Fitness Landscapes*, PLoS Comput Biol **10**, e1003778 (2014).

[79] D. P. Rice, B. H. Good, and M. M. Desai, *The Evolutionarily Stable Distribution of Fitness Effects*, Genetics **200**, 321 (2015).

[80] H. A. Orr, *The Distribution of Fitness Effects Among Beneficial Mutations*, Genetics **163**, 1519 (2003).

[81] F. W. Allendorf, *Protein Polymorphism and the Rate of Loss of Duplicate Gene Expression*, Nature **272**, 76 (1978).

[82] M. Kimura and J. L. King, *Fixation of a Deleterious Allele at One of Two "Duplicate" Loci by Mutation Pressure and Random Drift*, Proc. Natl. Acad. Sci. U.S.A. **76**, 2858 (1979).

[83] T. Ohta, *Time for Spreading of Compensatory Mutations under Gene Duplication*, Genetics **123**, 579 (1989).

[84] M. A. Nowak, M. C. Boerlijst, J. Cooke, and J. M. Smith, *Evolution of Genetic Redundancy*, Nature **388**, 167 (1997).

[85] L. V. Valen, *A New Evolutionary Law*, Evol. Theory **1**, 1 (1973).

[86] S. Okasha, *Evolution and the Levels of Selection* (Oxford University Press, 2006).

[87] G. Martin, *Fisher's Geometrical Model Emerges as a Property of Complex Integrated Phenotypic Networks*, Genetics **197**, 237 (2014).

[88] M. Imhof and C. Schlötterer, *E. Coli Microcosms Indicate a Tight Link between Predictability of Ecosystem Dynamics and Diversity*, PLOS Genetics **2**, e103 (2006).

[89] O. Lukjancenko, T. M. Wassenaar, and D. W. Ussery, *Comparison of 61 Sequenced Escherichia Coli Genomes*, Microbial Ecology **60**, 708 (2010).

[90] S. Louca, L. W. Parfrey, and M. Doebeli, *Decoupling Function and Taxonomy in the Global Ocean Microbiome*, Science **353**, 1272 (2016).

[91] S. Louca, S. M. Jacques, A. P. Pires, J. S. Leal, D. S. Srivastava, L. W. Parfrey, V. F. Farjalla, and M. Doebeli, *High Taxonomic Variability despite Stable Functional Structure across Microbial Communities*, Nature Ecology & Evolution **1**, 1 (2016).

[92] P. Chesson, *MacArthur's Consumer-Resource Model*, Theoretical Population Biology **37**, 26 (1990).

[93] B. H. Good, S. Martis, and O. Hallatschek, *Adaptation Limits Ecological Diversification and Promotes Ecological Tinkering during the Competition for Substitutable Resources*, Proceedings of the National Academy of Sciences **115**, E10407 (2018).

[94] A. Altieri and S. Franz, *Constraint Satisfaction Mechanisms for Marginal Stability and Criticality in Large Ecosystems*, Physical Review E **99**, 010401 (2019).

[95] P.-Y. Ho, T. H. Nguyen, J. M. Sanchez, B. C. DeFelice, and K. C. Huang, *Resource Competition Predicts Assembly of in Vitro Gut Bacterial Communities*.

[96] D. E. Dykhuizen and D. L. Hartl, *Selection in Chemostats*, Microbiological Reviews **47**, 150 (1983).

[97] R. MacArthur, *Species Packing, and What Competition Minimizes*, Proceedings of the National Academy of Sciences **64**, 1369 (1969).

[98] H. Tettelin et al., *Genome Analysis of Multiple Pathogenic Isolates of Streptococcus Agalactiae: Implications for the Microbial "Pan-Genome,"* Proceedings of the National Academy of Sciences **102**, 13950 (2005).

[99] H. Mickalide and S. Kuehn, *Higher-Order Interaction between Species Inhibits Bacterial Invasion of a Phototroph-Predator Microbial Community*, Cell Systems **9**, 521 (2019).

[100] T. Taillefumier, A. Posfai, Y. Meir, and N. S. Wingreen, *Microbial Consortia at Steady Supply*, Elife **6**, e22644 (2017).

[101] G. Kinsler, K. Geiler-Samerotte, and D. A. Petrov, *Fitness Variation across Subtle Environmental Perturbations Reveals Local Modularity and Global Pleiotropy of Adaptation*, Elife **9**, e61271 (2020).

[102] S. Estrela, J. C. C. Vila, N. Lu, D. Bajic, M. Rebolleda-Gomex, C.-Y. Chang, J. E. Goldford, A. Sanchez-Gorostiaga, and A. Sanchez, *Functional Attractors in Microbial Community Assembly*, Cell Systems **13**, 29 (2022).

[103] R. May, *Stability and Complexity in Model Ecosystems. Princeton Univ Press*, Princeton, NJ (1973).

[104] D. Šiljak, *When Is a Complex Ecosystem Stable?*, Mathematical Biosciences **25**, 25 (1975).

[105] A.-M. Neutel, J. A. Heesterbeek, and P. C. De Ruiter, *Stability in Real Food Webs: Weak Links in Long Loops*, Science **296**, 1120 (2002).

[106] N. Rooney, K. McCann, G. Gellner, and J. C. Moore, *Structural Asymmetry and the Stability of Diverse Food Webs*, Nature **442**, 265 (2006).

[107] U. Brose, R. J. Williams, and N. D. Martinez, *Allometric Scaling Enhances Stability in Complex Food Webs*, Ecology Letters **9**, 1228 (2006).

[108] S. Allesina, J. Grilli, G. Barabás, S. Tang, J. Aljadeff, and A. Maritan, *Predicting the Stability of Large Structured Food Webs*, Nature Communications **6**, 1 (2015).

[109] G. Bunin, *Interaction Patterns and Diversity in Assembled Ecological Communities*, ArXiv Preprint ArXiv:1607.04734 (2016).

[110] M. Barbier, J.-F. Arnoldi, G. Bunin, and M. Loreau, *Generic Assembly Patterns in Complex Ecological Communities*, Proceedings of the National Academy of Sciences **115**, 2156 (2018).

[111] D. I. Bolnick, P. Amarasekare, M. S. Araújo, R. Bürger, J. M. Levine, M. Novak, V. H. Rudolf, S. J. Schreiber, M. C. Urban, and D. A. Vasseur, *Why Intraspecific Trait Variation Matters in Community Ecology*, Trends in Ecology & Evolution **26**, 183 (2011).

[112] C. Burke, P. Steinberg, D. Rusch, S. Kjelleberg, and T. Thomas, *Bacterial Community Assembly Based on Functional Genes Rather than Species*, Proceedings of the National Academy of Sciences **108**, 14288 (2011).

[113] B. Segerman, *The Genetic Integrity of Bacterial Species: The Core Genome and the Accessory Genome, Two Different Stories*, Frontiers in Cellular and Infection Microbiology **2**, 116 (2012).

[114] R. G. Beiko, *Microbial Malaise: How Can We Classify the Microbiome?*, Trends in Microbiology **23**, 671 (2015).

[115] P. Vernocchi, F. Del Chierico, and L. Putignani, *Gut Microbiota Profiling: Metabolomics Based Approach to Unravel Compounds Affecting Human Health*, Frontiers in Microbiology **7**, 1144 (2016).

[116] M. R. Wilson, L. Zha, and E. P. Balskus, *Natural Product Discovery from the Human Microbiome*, Journal of Biological Chemistry **292**, 8546 (2017).

[117] 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).

[118] D. T. Gillespie, *Exact Stochastic Simulation of Coupled Chemical Reactions*, The Journal of Physical Chemistry **81**, 2340 (1977).

[119] E. L. Haseltine and J. B. Rawlings, *Approximate Simulation of Coupled Fast and Slow Reactions for Stochastic Chemical Kinetics*, The Journal of Chemical Physics **117**, 6959 (2002).

[120] J. G. Lopez and N. S. Wingreen, *Noisy Metabolism Can Promote Microbial Cross-Feeding*, Elife **11**, e70694 (2022).

[121] M. M. Desai and D. S. Fisher, *Beneficial Mutation–Selection Balance and the Effect of Linkage on Positive Selection*, Genetics **176**, 1759 (2007).

[122] S. Pollak, M. Gralka, Y. Sato, J. Schwartzman, L. Lu, and O. X. Cordero, *Public Good Exploitation in Natural Bacterioplankton Communities*, Science Advances **7**, eabi4717 (2021).

[123] D. Simberloff and T. Dayan, *The Guild Concept and the Structure of Ecological Communities*, Annual Review of Ecology and Systematics **22**, 115 (1991).

[124] J. B. Wilson, *Guilds, Functional Types and Ecological Groups*, Oikos **86**, 507 (1999).

[125] D. M. Benoit, D. A. Jackson, and C. Chu, *Partitioning Fish Communities into Guilds for Ecological Analyses: An Overview of Current Approaches and Future Directions*, Can. J. Fish. Aquat. Sci. **78**, 984 (2021).

[126] S. Díaz and M. Cabido, *Vive La Différence: Plant Functional Diversity Matters to Ecosystem Processes*, Trends in Ecology & Evolution **16**, 646 (2001).

[127] N. Blaum, E. Mosner, M. Schwager, and F. Jeltsch, *How Functional Is Functional? Ecological Groupings in Terrestrial Animal Ecology: Towards an Animal Functional Type Approach*, Biodiversity and Conservation **20**, 2333 (2011).

[128] J. Blondel, *Guilds or Functional Groups: Does It Matter?*, Oikos **100**, 223 (2003).

[129] Human Microbiome Project Consortium, *Structure, Function and Diversity of the Healthy Human Microbiome*, Nature **486**, 207 (2012).

[130] K. Anantharaman et al., *Thousands of Microbial Genomes Shed Light on Interconnected Biogeochemical Processes in an Aquifer System*, Nature Communications **7**, 13219 (2016).

[131] S. Louca et al., *Function and Functional Redundancy in Microbial Systems*, Nature Ecology & Evolution **2**, 936 (2018).

[132] S. A. Levin, *Encyclopedia of Biodiversity.* (Elsevier Science, 2013).

[133] X. Shan, A. Goyal, R. Gregor, and O. X. Cordero, *Annotation-Free Discovery of Functional Groups in Microbial Communities*.

[134] A. S. Raman, J. L. Gehrig, S. Venkatesh, H.-W. Chang, M. C. Hibberd, S. Subramanian, G. Kang, P. O. Bessong, A. A. Lima, and M. N. Kosek, *A Sparse Covarying Unit That Describes Healthy and Impaired Human Gut Microbiota Development*, Science **365**, eaau4735 (2019).

[135] 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*, Nat Commun **12**, 3254 (2021).

[136] J. Kehe, A. Kulesa, A. Ortiz, C. M. Ackerman, S. G. Thakku, D. Sellers, S. Kuehn, J. Gore, J. Friedman, and P. C. Blainey, *Massively Parallel Screening of Synthetic Microbial Communities*, Proceedings of the National Academy of Sciences **116**, 12804 (2019).

[137] Y. Furusawa et al., *Commensal Microbe-Derived Butyrate Induces the Differentiation of Colonic Regulatory T Cells*, Nature **504**, 446 (2013).

[138] Z. Li et al., *Butyrate Reduces Appetite and Activates Brown Adipose Tissue via the Gut-Brain Neural Circuit*, Gut **67**, 1269 (2018).

[139] H. V. Lin et al., *Butyrate and Propionate Protect against Diet-Induced Obesity and Regulate Gut Hormones via Free Fatty Acid Receptor 3-Independent Mechanisms*, PloS One **7**, e35240 (2012).

[140] J. Segain, D. R. De La Blétiere, A. Bourreille, V. Leray, N. Gervois, C. Rosales, L. Ferrier, C. Bonnet, H. Blottiere, and J. Galmiche, *Butyrate Inhibits Inflammatory Responses through NFκB Inhibition: Implications for Crohn's Disease*, Gut **47**, 397 (2000).

[141] Z. E. Ilhan, A. K. Marcus, D.-W. Kang, B. E. Rittmann, and R. Krajmalnik-Brown, *PH-Mediated Microbial and Metabolic Interactions in Fecal Enrichment Cultures*, Msphere **2**, e00047 (2017).

[142] A. W. Walker, S. H. Duncan, E. C. McWilliam Leitch, M. W. Child, and H. J. Flint, *PH and Peptide Supply Can Radically Alter Bacterial Populations and Short-Chain Fatty Acid Ratios within Microbial Communities from the Human Colon*, Applied and Environmental Microbiology **71**, 3692 (2005).

[143] H. J. Flint, S. H. Duncan, and P. Louis, *2 - Gut Microbial Ecology*, in *Designing Functional Foods*, edited by D. J. McClements and E. A. Decker (Woodhead Publishing, 2009), pp. 38–67.

[144] A. Santos-Lopez, C. W. Marshall, M. R. Scribner, D. J. Snyder, and V. S. Cooper, *Evolutionary Pathways to Antibiotic Resistance Are Dependent upon Environmental Structure and Bacterial Lifestyle*, Elife **8**, e47612 (2019).

[145] J. M. Chacón, A. K. Shaw, and W. R. Harcombe, *Increasing Growth Rate Slows Adaptation When Genotypes Compete for Diffusing Resources*, PLoS Computational Biology **16**, e1007585 (2020).

[146] A. Sharma and K. B. Wood, *Spatial Segregation and Cooperation in Radially Expanding Microbial Colonies under Antibiotic Stress*, The ISME Journal **15**, 3019 (2021).

[147] A. Auladell, A. Barberán, R. Logares, E. Garcés, J. M. Gasol, and I. Ferrera, *Seasonal Niche Differentiation among Closely Related Marine Bacteria*, The ISME Journal **16**, 178 (2022).

[148] B. Bloxham, H. Lee, and J. Gore, *Biodiversity Is Enhanced by Sequential Resource Utilization and Environmental Fluctuations via Emergent Temporal Niches*, BioRxiv 2023 (2023).

Appendix A: Appendix of Chapter 2

Here I show supplemental (unpublished) results from the "toolbox model" of evolution I present in Chapter 2.

A.1 Evolved Genotypes versus Random Genotypes

As described in the Chapter 2, when a genotype adapts to an environment in the "toolbox model", its evolution is dominated by an "improve it or lose it" feedback loop: genetic systems of higher usage (expression) mutate, improve and become used even more, while lesser used systems are used less and less becoming obsolete. Through this process, evolution imprints structure on a genotype and its expression, and the properties of evolved genotypes are atypical compared to randomly sampled genotypes of similar fitness.

To show this, I generate two sets of genotypes and contrast their distributions of system usage (i.e., the coefficients of row vectors in the genotype matrix). In order to make a fair comparison, each set of genotypes fall within the same fitness band (see Figure A.1A), but each set is obtained in different ways: one set consists of randomly drawn genotype matrices that happen to fall above a fitness threshold (chosen to be -0.5 based on the parameters of Chapter 2 and the limitations of the naïve sampling method I use here); the other set consists of genotypes that evolved to the specified fitness threshold from an initially random matrix (generally of lower fitness). Plotting the overall usage distribution across all genotypes within the respective sets in Figure A.1B, one can see that the typical genotype from an unstructured ensemble uses the majority of its system vectors $\vec{g}_\mu$; fitting the target environment $\vec{E}$ is a cooperative effort from each system. In contrast, the distribution of the evolved set indicates the corresponding ensemble of structured genotypes are clearly atypical relative to those drawn randomly at similar fitness. Instead, evolved genotypes

consist of many unused systems that have lost out to others in the competition for expression described in Chapter 2.

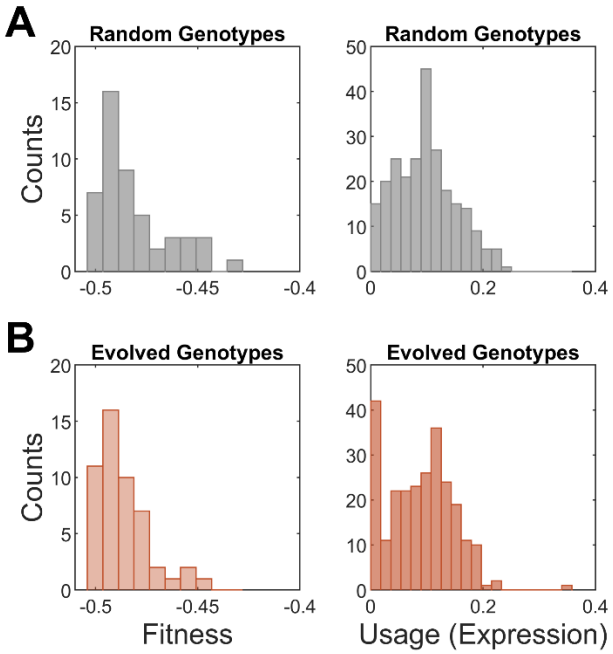

Figure A.1 | Sets of randomly drawn genotypes (A) and evolved genotypes (B) (see text) have distinct expression distributions (right column of panels) despite falling within the same narrow band of fitness (left column of panels). Evolved genotypes are atypical from the point of view of the random genotype usage distribution because many systems are unused (peak at 0) due to the feedback loop mechanism investigated in Chapter 2 leaving evolved genotypes to rely on effectively fewer systems.

The next section analytically calculates the distribution of expression for randomly drawn genotypes using the cavity method from the statistical physics of disordered systems. Although this nicely characterizes the unstructured regime, the applicability of such analytical techniques begins to break down once genotypes evolve. This highlights the need for extensions and generalizations of these techniques to structured ensembles shaped by evolution.

A.2 Cavity Calculation for Toolbox Model

Minimization Conditions

Given an L-dimensional target vector $\vec{E}$ and a random incomplete set of K basis vectors $\vec{e_i}$, find the linear combination of the basis vectors that best fits the target vector subject to the constraint that all coefficients a_i are non-negative. In other words, we need to solve the following constrained minimization problem,

$$\min_{\{a_i \geq 0\}} \left\| \sum_i a_i e_i - \vec{E} \right\|^2 . \tag{A.1}$$

This is equivalent to finding the $\{a_i\}$ such that

$$f \coloneqq \sum_{\mu=1}^{L} (E_\mu - \sum_{j=1}^{K} a_j e_{\mu j})^2 \tag{A.2}$$

is a minimum. This occurs when the $\{a_i\}$ satisfy the following conditions:
either

$$\begin{aligned} a_i &> 0 \\ \frac{\partial f}{\partial a_i} &= \sum_\mu e_{\mu i}(E_\mu - \sum_j a_j e_{\mu j}) = 0 \end{aligned} \tag{A.3}$$

or

$$\begin{aligned} a_i &= 0 \\ \frac{\partial f}{\partial a_i} &= \sum_\mu e_{\mu i}(E_\mu - \sum_j a_j e_{\mu j}) < 0. \end{aligned} \tag{A.4}$$

Consumer Resource Model Mapping

To begin setting up the cavity calculation, it is useful to map the optimization problem onto the dynamics of a consumer resource model (CRM) whose equilibrium solutions correspond to the minimum of our optimization. It turns out that the following CRM equations do the job:

$$\dot{a}_i = a_i \sum_\mu e_{\mu i} R_\mu \tag{A.5}$$

$$\dot{R}_\mu = E_\mu - R_\mu - \sum_j a_j e_{\mu j} \tag{A.6}$$

Notice that at steady state, $\dot{R}_\mu = 0$ implies

$$R_\mu = E_\mu - \sum_j a_j e_{\mu j}$$

Inserting this into the steady state condition for (5), we find the same minimization conditions of the optimization problem from above:

$$0 = a_i \left(\sum_\mu e_{\mu i} E_\mu - \sum_{\mu j} e_{\mu i} e_{\mu j} a_j \right) .$$

Cavity Calculation

Because we are interested in the performance of a randomly drawn genome (i.e., a random binary matrix) in a random target environment, our problem is to solve for the statistical distribution of solutions to the optimization problem (1) given statistical properties of E_μ and $e_{\mu i}$. That is, given the mean and variance of the distribution for E_μ and $e_{\mu i}$, what is the resulting distribution of the a_i and R_μ. To do this, we borrow an approach from statistical physics known as the "Cavity Method" and apply it to the CRM equations (5) and (6). The basic idea behind the method is as follows:

- Assume that we know the equilibrium solutions to CRM equations for a system of size $K-1$ genes and $L-1$-dimensional target for a given realization of parameters
- Now introduce a new gene and component to the target vector so that the system size increases to K by L, which can be seen as a perturbation of order $1/K$
- Assume that the new equilibrium solutions to the larger system respond linearly around the old equilibrium solutions
- Solve for the equilibrium of the newly introduced gene expression, a_0, and residual, R_0
- Since the introduced gene and target component could have been any of the original genes or components, these solutions are self-consistency equations for the entire system. So by solving for the statistical properties of the a_0 and R_0, we are effectively solving for the distribution for any of the $a_{i/0}$ and $R_{\mu/0}$.

For the sake of expanding to linear order in perturbations, we rewrite the CRM equations with an added auxiliary parameter, g_i, whose use will become apparent when we begin taking derivatives:

$$\frac{1}{a_i}\dot{a}_i = \sum_\nu e_{\nu i} R_\nu - g_i, \quad 1 \leq i \leq K \tag{A.7}$$

$$\dot{R}_\mu = E_\mu - R_\mu - \sum_j a_j e_{\mu j}, \quad 1 \leq \mu \leq L. \tag{A.8}$$

To evaluate the scaling of terms and how the large sums will average and fluctuate, it is useful to write the genome entries as

$$e_{\mu i} \equiv \mu_e + \sigma_e d_{\mu i}, \tag{A.9}$$

where we know $\mu_e = p$ and $\sigma_e^2 = p(1-p)$ since the genome is composed of i.i.d. binomial random variables. Moreover, by this definition we must require that $\langle d_{\mu i} d_{\nu j} \rangle = \delta_{\mu\nu}\delta_{ij}$. Equations (7) and (8) now become

$$\frac{1}{a_i}\dot{a}_i = \mu_e \sum_\nu R_\nu + \sigma_e \sum_\nu d_{\nu i} R_\nu - g_i \tag{A.10}$$

$$\dot{R}_\mu = E_\mu - R_\mu - \mu_e \sum_j a_j - \sigma_e \sum_j d_{\mu j} a_j. \tag{A.11}$$

To make the scaling of the first two sums in (10) and (11) a bit more apparent, define

$$\frac{1}{L}\sum_\nu R_\nu \equiv \langle R \rangle$$

$$\frac{1}{K}\sum_j a_j \equiv \langle a \rangle.$$

Inserting these into the CRM equations we obtain

$$\frac{1}{a_i}\dot{a}_i = L\mu_e \langle R \rangle + \sigma_e \sum_\nu d_{\nu i} R_\nu - g_i \tag{A.12}$$

$$\dot{R}_\mu = E_\mu - R_\mu - K\mu_e \langle a \rangle - \sigma_e \sum_j d_{\mu j} a_j, \tag{A.13}$$

where $L\mu_e\langle R\rangle$ and $K\mu_e\langle a\rangle$ are both $\mathcal{O}(1)$ because $\langle R\rangle \sim \frac{1}{L}$ and $\langle a\rangle \sim \frac{1}{K}$, respectively.

Assuming we have solved for the steady state solutions to this K by L system, we now introduce the $i = 0$ gene and $\mu = 0$ component and study how this affects the system:

$$\frac{1}{a_i}\dot{a}_i = L\mu_e\langle R\rangle + \sigma_e \sum_{\nu\neq 0} d_{\nu i}R_\nu - g_i + \sigma_e d_{0i}R_0$$
$$\dot{R}_\mu = E_\mu - R_\mu - K\mu_e\langle a\rangle - \sigma_e \sum_{j\neq 0} d_{\mu j}a_j - \sigma_e d_{\mu 0}a_0.$$

A subtle, but important, point should be mentioned here but won't prove useful until later. As was mentioned in the logic outline of the cavity method, we are interested in the new steady-state solutions to (12) and (13) under this perturbation, some of which will have $a_i = 0$. That said, the sum in the $\dot{R}_\mu = 0$ equation is really over $K^* = \phi K$ terms (i.e., $\sum_{j=1}^{K^*} d_{\mu j}a_j$), where ϕ is the fraction of genes that have nonzero expression ($a_i > 0$) after the perturbation.

Returning to the introduced gene and component, we can think of these additional terms as effective perturbations to the parameters E_μ and g_i,

$$\frac{1}{a_i}\dot{a}_i = \ldots - (g_i + \delta g_i)$$
$$\dot{R}_\mu = E_\mu + \delta E_\mu - \ldots.$$

Given the perturbation is small enough, which is the case for larger and larger K and L since the perturbation is of order 1 / "system size", we assume that the new steady state solutions respond linearly around the old (e.g., denoted by $\bar{a}_i$ and $\bar{a}_{i/0}$, respectively):

$$\bar{a}_i \approx \bar{a}_{i/0} + \sum_\nu \frac{\partial \bar{a}_i}{\partial E_\nu}\delta E_\nu + \sum_j \frac{\partial \bar{a}_i}{\partial g_j}\delta g_j$$
$$\bar{R}_\mu \approx \bar{R}_{\mu/0} + \sum_\nu \frac{\partial \bar{R}_\mu}{\partial E_\nu}\delta E_\nu + \sum_j \frac{\partial \bar{R}_\mu}{\partial g_j}\delta g_j,$$

where $\delta g_i \equiv -\sigma_e d_{0i}\bar{R}_0$ and $\delta E_\mu \equiv -\sigma_e d_{\mu 0}\bar{a}_0$ for the present case. Inserting the linear response expansion into the steady state form of the $i = 0$ and $\mu = 0$ equations of the CRM we obtain

$$0 = \bar{a}_0 \left[L\mu_e\langle R\rangle + \sigma_e \sum_\nu d_{\nu 0}\bar{R}_{\nu/0} - \sigma_e^2 \bar{a}_0 \sum_{\nu,\mu} d_{\nu 0}d_{\mu 0}\frac{\partial \bar{R}_\nu}{\partial E_\nu} \right.$$
$$\left. -\sigma_e^2 \bar{R}_0 \sum_{\nu,j} d_{\nu 0}d_{0j}\frac{\partial \bar{R}_\nu}{\partial g_j} - g_0 + \sigma_e d_{00}\bar{R}_0 \right]$$
$$0 = E_0 - \bar{R}_0 - K\mu_e\langle a\rangle - \sigma_e \sum_j d_{0j}\bar{a}_{j/0}$$
$$+ \sigma_e^2 \bar{a}_0 \sum_{\nu,j} d_{0j}d_{\nu 0}\frac{\partial \bar{a}_j}{\partial E_\nu} + \sigma_e^2 \bar{R}_0 \sum_{i,j} d_{0j}d_{0i}\frac{\partial \bar{a}_j}{\partial g_i} - \sigma_e d_{00}\bar{a}_0$$

Ignoring the $a_0 = 0$ solution for the moment, we solve for a_0 and R_0

$$\bar{a}_0 = \frac{L\mu_e + \sigma_e \sum_\nu d_{\nu 0}\bar{R}_{\nu/0} - \sigma_e^2 \bar{R}_0 \sum_{\nu,j} d_{\nu 0}d_{0j}\frac{\partial \bar{R}_\nu}{\partial g_j} - g_0 + \sigma_e d_{00}\bar{R}_0}{\sigma_e^2 \sum_{\nu,\mu} d_{\nu 0}d_{\mu 0}\frac{\partial \bar{R}_\nu}{\partial E_\nu}} \tag{A.14}$$

$$\bar{R}_0 = \frac{E_0 - K\mu_e\langle a\rangle - \sigma_e \sum_j d_{0j}\bar{a}_{j/0} - \sigma_e d_{00}\bar{a}_0 + \sigma_e^2 \bar{a}_0 \sum_{\nu,j} d_{0j}d_{\nu 0}\frac{\partial \bar{a}_j}{\partial E_\nu}}{1 - \sigma_e^2 \sum_{i,j} d_{0j}d_{0i}\frac{\partial \bar{a}_j}{\partial g_i}} \tag{A.15}$$

This concludes the "setup" of the cavity method. In the sections that follow we will look at the averages and fluctuations of (16) and (17) to obtain the statistical distributions of a_0 and R_0 from which we will solve self-consistent equations for the moments of these distributions numerically.

Distribution of Gene Usage

We ask what are the statistical properties of an ensemble of genomes and targets, each randomly drawn from model specified distributions. At first glance it looks like a nightmare to compute the expected value or variances of (14) and (15) because of the large double sums in both the numerator and denominator. However, these sums can be thought of as random walks that lead to self-averaging to some universal value, simplifying things tremendously. All that is left to do then is to check that the averages and variances of the approximate (14) and (15) have the appropriate scaling that has been found by both numerical and analytical means. To simplify the notation a bit, we define the susceptibilities of our variables to the effective perturbations,

$$\chi_{\mu\nu} \equiv \frac{\bar{R}_\mu}{\partial E_\nu} \qquad \chi_{i\nu} \equiv \frac{\partial \bar{a}_i}{\partial E_\nu}$$
$$\nu_{\mu j} \equiv \frac{\bar{R}_\mu}{\partial g_j} \qquad \nu_{ij} \equiv \frac{\partial \bar{a}_i}{\partial g_j}.$$

Let us first focus on (14), the distribution of gene usage. Denote the double sum in the denominator as

$$S_{\chi^R} \coloneqq \sum_{\nu,\mu} d_{\nu 0} d_{\mu 0} \frac{\partial \bar{R}_\nu}{\partial E_\mu}. \tag{A.16}$$

We can split S_{χ^R} up into sums over different indices and identical indices:

$$S_{\chi^R} = \sum_\mu d_{\mu 0}^2 \chi_{\mu\mu} + \sum_{\nu \neq \mu} d_{\nu 0} d_{\mu 0} \chi_{\nu\mu}$$

Although we have seen in the numerics that the sign of $\frac{\partial \bar{a}_i}{\partial E_\nu}$ is correlated to the value of $e_{\nu i}$, this doesn't matter here because it is safe to assume that these derivatives and $e_{\nu 0}$ (the perturbation) are uncorrelated. This holds for the other derivatives in the other sums as well. Having said that, the average of (16) is

$$\langle S_{\chi^R} \rangle = \frac{K}{\gamma} \chi,$$

where $\chi \equiv \langle \chi_{\mu\mu} \rangle$. From numerics (see Appendix A for a scaling table of all variables) we know that $\chi \sim \mathcal{O}(1)$, which implies that $\langle S_{\chi^R} \rangle \sim \mathcal{O}(K)$. OTOH, the variance of (16) is

$$\mathrm{var}(S_{\chi^R}) = \frac{K}{\gamma} \mathrm{var}(\chi_{\mu\mu}) + (\frac{K}{\gamma})^2 \mathrm{var}(\chi_{\mu\nu}).$$

Here numerics tell us $\mathrm{var}(\chi_{\mu\mu}) \sim \mathcal{O}(\frac{1}{K^2})$ and $\mathrm{var}(\chi_{\mu\nu}) \sim \mathcal{O}(\frac{1}{K})$ so that to leading order in K, $\mathrm{var}(S_{\chi^R}) \sim \mathcal{O}(K)$. These results follow from the random walk nature of these sums over terms that alternate in sign due to the standard normal random variable $d_{\mu i}$. Despite having much fewer terms, the "diagonal" terms coherently add while the "off-diagonal" terms cancel out on average but contribute fluctuations subleading in K, resulting in a sum that self-averages to

$$S_{\chi^R} \approx \sum_\mu d_{\mu 0}^2 \chi_{\mu\mu} \pm \mathcal{O}(\sqrt{K}) \equiv \frac{K}{\gamma} \chi. \tag{A.17}$$

Since the fluctuations in (17) are subleading in K, this contribution can be neglected in the large K limit as the denominator of (14) tends to a the universal value $\sigma_e^2 \frac{K}{\gamma} \chi$.
Now for the double sum in the numerator over mixed indices that we will denote

$$S_{\nu^R} \coloneqq \sum_{\mu,j} d_{\mu 0} d_{0j} \nu_{\mu j}. \tag{A.18}$$

Because the indices are mixed, there are no "diagonal" terms, so the statistics of the $d_{\mu i}$ leave us with (18) having mean 0, $\langle S_{\nu^R} \rangle = 0$. Again, over an ensemble of genome and target realizations, the $d_{\nu i}$ and $\nu_{\mu j}$ we assume to be uncorrelated such that the variance of (18) is simply

$$\mathrm{var}(S_{\nu^R}) = \frac{K^2}{\gamma} \mathrm{var}(\nu_{\mu j}).$$

Because $\mathrm{var}(\nu_{\mu j}) \sim \frac{1}{K^2}$ (see Appendix A), the variance scales as $\mathrm{var}(S_{\nu^R}) \sim \mathcal{O}(1)$. The final piece to determining the scaling of this term in (14) is to account for the statistics of the common $\bar{R}_0$ factor out front. Whether $\bar{R}_0$ and the double sum are correlated random variables or not, we find that

$$\mathrm{var}(\bar{R}_0 S_{\nu^R}) \sim \frac{1}{K}$$

since $\mathrm{var}(\bar{R}_0) \sim \mathcal{O}(1/K)$. Just as we did for S_{χ^R}, we summarize the behavior of the S_{ν^R} term as a mean plus fluctuations:

$$\sigma_e^2 \bar{R}_0 \sum_{\mu,j} d_{\mu 0} d_{0j} \frac{\partial \bar{R}_\nu}{\partial g_j} \approx 0 \pm \mathcal{O}(\frac{1}{\sqrt{K}}). \tag{A.19}$$

Before we continue any further with simplifying (14), we should first check that the scaling of both sides of the equation are consistent. On the left hand side, since we could have picked any gene from the original genome to be the "introduced" gene, we know that a_0 should have the same statistics as any of the other gene expression states. From appendix A, the mean and variance of a_0 scale as $1/K$ and $1/K^2$, respectively. We now turn to the scaling of the numerator of (14) term by term:

- $L\mu_e\langle R\rangle \sim \mathcal{O}(1) \pm 0$
- $\sigma_e \sum_\nu d_{\nu 0} \bar{R}_\nu \sim \mathcal{O}(1) \pm 1$ *
- $\sigma_e^2 \bar{R}_0 \sum_{\mu,j} d_{\mu 0} d_{0j} \frac{\partial \bar{R}_\nu}{\partial g_j} \approx 0 \pm \mathcal{O}(\frac{1}{\sqrt{K}})$
- g_0 is an auxiliary parameter that will be set to zero to solve the self-consistency equations in the next section, but was a parameter of order 1 in simulations.

* the fluctuations could be $\mathcal{O}(1/\sqrt{K})$ if the terms in the sum are uncorrelated, but we know that in order for the scaling of the variance to work out the sum must have correlated terms to give rise to fluctuations up order 1.
So overall, combining the above scaling of the numerator and denominator scaling in (17), the scaling of the right hand side of (14) has an average that is of order $1/K$ and variance of order $1/K^2$ as desired. Now all that is left to do is to formally write down the distribution of a_0. Since (18) has mean 0 and vanishing variance in the limit of large K, we can neglect this term from the numerator of (14). This leaves us with a constant, $L\mu_e\langle R\rangle$, and the single sum and auxiliary parameter, $\sigma_e \sum_\nu d_{\nu 0} \bar{R}_\nu - g_0$, where the sum now includes the $\nu = 0$ term. As was alluded to in the above sections, we are not interested in any single solution set to the optimization problem or CRM equations, but rather the statistics of the solutions, such as the average residual or the spread of gene expression. To do this, we consider the parameters, such as the g_i, being drawn randomly from an arbitrary distribution with mean g and variance σ_g^2 (though it must be well-behaved in some sense – see ref. on cavity method for Lotka-Volterra or Pankaj's paper for discussion). Moreover, the sum $\sum_\nu d_{\nu 0} \bar{R}_\nu$ can be seen as a sum over the steps of a random walk. With this in mind, let

$$g_i \equiv g + \delta g_i$$

such that, by the central limit theorem,

$$-\delta g_0 + \sigma_e \sum_\nu d_{\nu 0} \bar{R}_\nu \tag{A.20}$$

is approximately a Gaussian random variable with mean 0 and variance

$$\sigma_{a_0}^2 = \sigma_g^2 + \sigma_e^2 L q_R,$$

where $q_R \equiv \frac{1}{L} \sum_\nu \bar{R}_\nu^2$. We can then define

$$\sigma_{a_0} z_a \equiv -\delta g_0 + \sigma_e \sum_\nu d_{\nu 0} \bar{R}_\nu, \tag{A.21}$$

where $z_a \sim \mathcal{N}(0,1)$. Finally, including the possible $a_0 = 0$ solution, the distribution for gene usage is a truncated Gaussian:

$$a_0 = \frac{\max\left[0, L\mu_e\langle R\rangle - g + \sigma_{a_0} z_a\right]}{\sigma_e^2 \frac{K}{\gamma} \chi}. \tag{A.22}$$

Distribution of Residuals

We follow the same procedure in finding the gene usage distribution. Now consider equation (15). Define the double sum

$$S_{\chi^a} \coloneqq \sum_{\nu,j} d_{0j} d_{\nu 0} \frac{\partial \bar{a}_j}{\partial E_\nu} = \sum_{\nu,j} d_{0j} d_{\nu 0} \chi_{j\nu}. \tag{A.23}$$

Again, because of mixed indices, $\langle S_{\chi^a} \rangle = 0$. As before, we assume terms in the sum are uncorrelated so that the variance is simply

$$\mathrm{var}(S_{\chi^a}) = \frac{K^2}{\gamma} \mathrm{var}(\chi_{j\nu}) \sim \mathcal{O}(1),$$

where the scaling is informed by the numerics (Appendix A). A slight subtlety arises in evaluating the variance of the term that contains the double sum S_{χ^a} that didn't come up in the previous section. The reason for this subtlety is because, unlike the R_μ, a_0 is not a zero mean random variable. Therefore, the variance can be written as

$$\mathrm{var}(\bar{a}_0 S_{\chi^a}) = \mathrm{var}(\bar{a}_0 \mathrm{var}(S_{\chi^a}) + \langle a_0 \rangle^2 \mathrm{var}(S_{\chi^a}).$$

Then scaling of the correspond term in the numerator is

$$\sigma_e^2 \bar{a}_0 \sum_{\nu,j} d_{0j} d_{\nu 0} \chi_{j\nu} \approx 0 \pm \mathcal{O}(\frac{1}{K}).$$

Next, using the same foregoing assumptions, we compute the average and variance of the sum over like indices from the denominator denoted as

$$S_{\nu^a} \coloneqq \sum_{i,j} d_{0i} d_{0j} \frac{\partial \bar{a}_j}{\partial g_i} = \sum_j d_{0j}^2 \nu_{jj} + \sum_{j \neq i} d_{0i} d_{0j} \nu_{ji}. \tag{A.24}$$

As was mentioned in the subtle remark following (12) and (13), the sum over j is over $K^* = \phi K$ terms. Therefore, the average and variance are found to be

$$\begin{aligned} \langle S_{\nu^a} \rangle &= \phi K \nu \\ \mathrm{var}(S_{\nu^a}) &\approx \phi K \mathrm{var}(\nu_{ii}) + \phi^2 K^2 \mathrm{var}(\nu_{ij}), \end{aligned}$$

where $\nu \equiv \langle \nu_{jj} \rangle$. By the scaling of the response derivatives in Appendix A, the scaling of the average and variance goes like $\langle S_{\nu^a} \rangle \sim \mathcal{O}(1)$ and $\mathrm{var}(S_{\nu^a}) \sim \mathcal{O}(1/K)$. Together, the term corresponding to the sum is on average

$$\begin{aligned} \sigma_e^2 \sum_{i,j} d_{0i} d_{0j} \frac{\partial \bar{a}_j}{\partial g_i} &\approx \sigma_e^2 \sum_j d_{0j}^2 \nu_{jj} \pm \mathcal{O}(\frac{1}{\sqrt{K}}) \\ &\equiv \sigma_e^2 \phi K \nu \pm \mathcal{O}(\frac{1}{\sqrt{K}}). \end{aligned}$$

We see in both double sums, S_{χ^a} and S_{ν^a}, that the fluctuations get arbitrarily small in the large K limit so that we can replace these sums by their average values. Altogether so far this simplifies (15) to

$$\bar{R}_0 = \frac{E_0 - K\mu_e \langle a \rangle - \sigma_e \sum_j d_{0j} \bar{a}_{j/0} - \sigma_e d_{00} \bar{a}_0}{1 - \sigma_e^2 \phi \nu}. \tag{A.25}$$

As was done for (14), we check both sides of (25) to confirm the scaling is consistent. But before we do that, we make one last substitution to make the scaling of the RHS more straightforward. Let

$$E_0 \equiv E + \frac{\delta E_0}{\sqrt{K}},$$

where we assume that $\langle E_0 \rangle \equiv E \sim \mathcal{O}(1)$ and $\mathrm{var}(E_0) \sim \mathcal{O}(1/K)$. Then, since the fitting of the genome to the target leads to $E \simeq K\mu_e \langle a \rangle$, we can reason that

$$E_0 - K\mu_e a \rangle \simeq \frac{\delta E_0}{\sqrt{K}} \sim \mathcal{O}(\frac{1}{\sqrt{K}}).$$

Finally, the scaling of the RHS term by term is as follows:

- the difference, $E_0 - K\mu_e$, has mean 0 and variance $\mathcal{O}(1/K)$
- $\sigma_e \sum_j d_{0j}\bar{a}_j \simeq 0 \pm \mathcal{O}(\frac{1}{\sqrt{K}})$
- the denominator is $\mathcal{O}(1)$.

Therefore, the mean and variance of the RHS is 0 and $\mathcal{O}(1/K)$, respectively, which agrees with the scaling of the residual statistics from Appendix A:

$$\langle R_0 \rangle \sim \frac{1}{\sqrt{K}}, \quad \text{var}(R_0) \sim \frac{1}{K}.$$

All that is left to do now is arrive at the formal expression for the residual distribution. To do this, we make use of the central limit theorem as was done in the previous section for the same reasons. Using the definition of E_0 above,

$$\frac{\delta E_0}{\sqrt{K}} - \sigma_e \sum_j d_{0j}\bar{a}_j$$

is approximately a Gaussian random variable with mean 0 and variance

$$\sigma_{R_0}^2 = \frac{\sigma_E^2}{K} + \sigma_e^2 K q_a, \tag{A.26}$$

where $q_a \equiv \frac{1}{K}\sum_j \bar{a}_j^2$. Define

$$\frac{\delta E_0}{\sqrt{K}} - \sigma_e \sum_j d_{0j}\bar{a}_j \equiv \sigma_{R_0} z_R,$$

where $z_R \sim \mathcal{N}(0,1)$ so that the distribution of residuals is a Gaussian random variable,

$$R_0 = \frac{E - K\mu_e\langle a \rangle + \sigma_{R_0} z_R}{1 - \sigma_e^2 \phi K \nu}. \tag{A.27}$$

Self-Consistency Equations

Although (22) and (27) give us the distributions we sought out for, they are in terms of 7 unknowns: $\{\phi, \langle a \rangle, \langle R \rangle, q_a, q_R, \chi, \nu\}$. Because every gene and residual component is statistically equivalent to gene $i = 0$ and component $\mu = 0$, we can write self-consistent equations by evaluating the moments of (22) and (27) and solving numerically. OTOH, the derivatives are found by directly evaluating the appropriate derivatives of (22) and (27):

$$\chi = \frac{\partial \bar{R}}{\partial E} = \frac{1}{1 - \sigma_e^2 \phi K \nu} \tag{A.28}$$

$$\nu = \frac{\partial \bar{a}}{\partial g} = -\frac{\gamma}{\sigma_e^2 K \chi}. \tag{A.29}$$

Now, to compute the moments it is useful to define new variables

$$\Delta_g \equiv \frac{L\mu_e\langle R \rangle - g}{\sigma_{a_0}}$$

$$\Delta_E \equiv \frac{E - K\mu_e\langle a \rangle}{\sigma_{R_0}},$$

and the functions

$$w_j^a(\Delta) \coloneqq \int_{-\Delta}^{\infty} \frac{\mathrm{d}z}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} (z + \Delta)^j$$

$$w_j^R(\Delta) \coloneqq \int_{-\infty}^{\infty} \frac{\mathrm{d}z}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} (z + \Delta)^j.$$

Now, with these definitions in mind, we note that for random variable $y = \max\left[0, \frac{a}{c} + \frac{b}{c}z\right]$ with z a standard normal random variable, then the moments of y are given by

$$\langle y^j \rangle = \int_{-\Delta}^{\infty} \mathrm{d}z \left[p(z)y(z)^j\right] = \left(\frac{b}{c}\right)^j \int_{-\frac{a}{b}}^{\infty} \frac{\mathrm{d}z}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} (z + \frac{a}{b})^j. \tag{A.30}$$

(30) allows us to concisely write the remaining self-consistency equations to go with (28) and (29):

$$\phi = w_0^a(\Delta_g) \tag{A.31}$$

$$\langle a \rangle = \frac{\sigma_{a_0}\gamma}{\sigma_e^2 K\chi} w_1^a(\Delta_g) \tag{A.32}$$

$$\langle R \rangle = \frac{\sigma_{R_0}}{1 - \sigma_e^2 \phi K\nu} w_1^R(\Delta_E) \tag{A.33}$$

$$q_a = \left(\frac{\sigma_{a_0}\gamma}{\sigma_e^2 K\chi}\right)^2 w_2^a(\Delta_g) \tag{A.34}$$

$$q_R = \left(\frac{\sigma_{R_0}}{1 - \sigma_e^2 \phi K\nu}\right)^2 w_2^R(\Delta_E). \tag{A.35}$$

To summarize, our goal is: given a parameter set $\{E, \sigma_E, g, \sigma_g, \mu_e, \sigma_e, K, L\}$, find the unknowns $\{\phi, \langle a \rangle, \langle R \rangle, q_a, q_R, \chi, \nu\}$ by numerically solving (28), (29), (31) - (35).

Comparing Theory with Simulation

The self-consistency equations were solved using MATLAB's vpasolve function. To compare the distribution of gene usage predicted by the cavity approach against simulated data, a Gaussian pdf was constructed using the mean and variance predicted by (32) and (34) rather than constructing the full truncated distribution of (22). Of course this only tells us approximately how well the Gaussian portion of (22) agrees with data. The agreement of the "δ-peak" that comes from the truncation (due to the nonnegativity constraint on the a_i) was evaluated by comparing the fraction of nonzero a_i, denoted by ϕ, to the fraction of simulated a_i that were nonzero. $1 - \phi$ gives us an idea of the height of the "δ-peak". These were the only two comparisons between the cavity result and simulation, although others might include looking at the distribution of residuals, predicted fitness, etc.

The simulations were performed as follows. For a given set of parameters (note: $g_i = 0 \forall i$ because these are auxiliary parameters, and the $e_{\mu i}$ are binomial random variables such that $\mu_e = p$ and $\sigma_e^2 = p(1-p)$), 50 random genomes were optimized in the least-squares sense to fit a given target vector. The target vector was chosen to be a noisy target centered around 1; i.e., $\vec{E} = (1, 1, \ldots, 1) + \epsilon * \vec{\eta}$, where $\vec{\eta}$ is a random vector drawn from a standard normal distribution and ϵ is a parameter that sets the strength of the Gaussian noise. After obtaining the expression coefficients from the fitting process, the a_i from the 50 trials were collected together, producing the simulated samples of the expression distribution. From this, the relevant statistics were computed. The following figures show the comparison between theory and simulation for parameters $p = 0.35$, $\epsilon = 0.1$, and $K = 20$. The other parameter, γ and thus L, was varied over a range from 0.1 to 0.8. Comparisons were made for other parameters as well and agreement was seen across the board, although if necessary, a more thorough and systematic test could be performed to see for which range of parameters the agreement begins to fail.

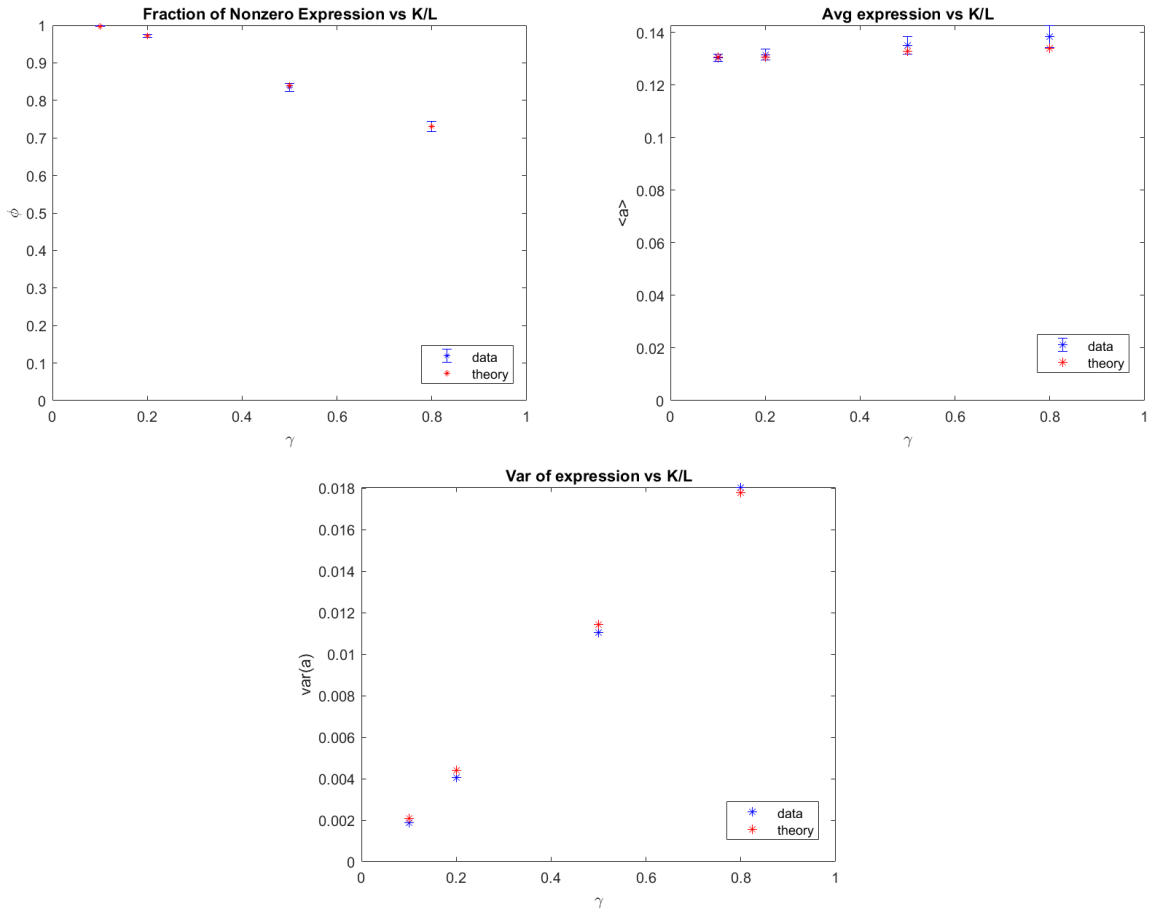

Figure A. 2: Comparing theoretically computed moments of usage/expression distribution with simulated data.

Appendix A: Scaling Table

Some numerical and some analytical arguments for the scaling of variables are presented in what follows, and all results are summarized in a table at the end of this section. First, I present analytical arguments for how gene expression a_i and residual components R_μ scale.

Scaling of Gene Expression

Rewriting random variables and parameters in terms of standard normal variables, let

$$\begin{aligned} a_i &\equiv a + \sigma_a \alpha_i \\ E_\mu &\equiv E + \sigma_E \epsilon_\mu, \end{aligned} \tag{A.36}$$

where the new variables satisfy

$$\begin{aligned} \langle \alpha_i \rangle &= 0 = \langle \epsilon_\mu \rangle \\ \langle \alpha_i \alpha_j \rangle &= \delta_{ij} \\ \langle \epsilon_\mu \epsilon_\nu \rangle &= \delta_{\mu\nu}. \end{aligned}$$

Fitting a random genome, $\mathcal{G} = \{e_{\mu i}\}$, to the target means on average

$$\langle \sum_i a_i e_{\mu i} \simeq E_\mu \rangle. \tag{A.37}$$

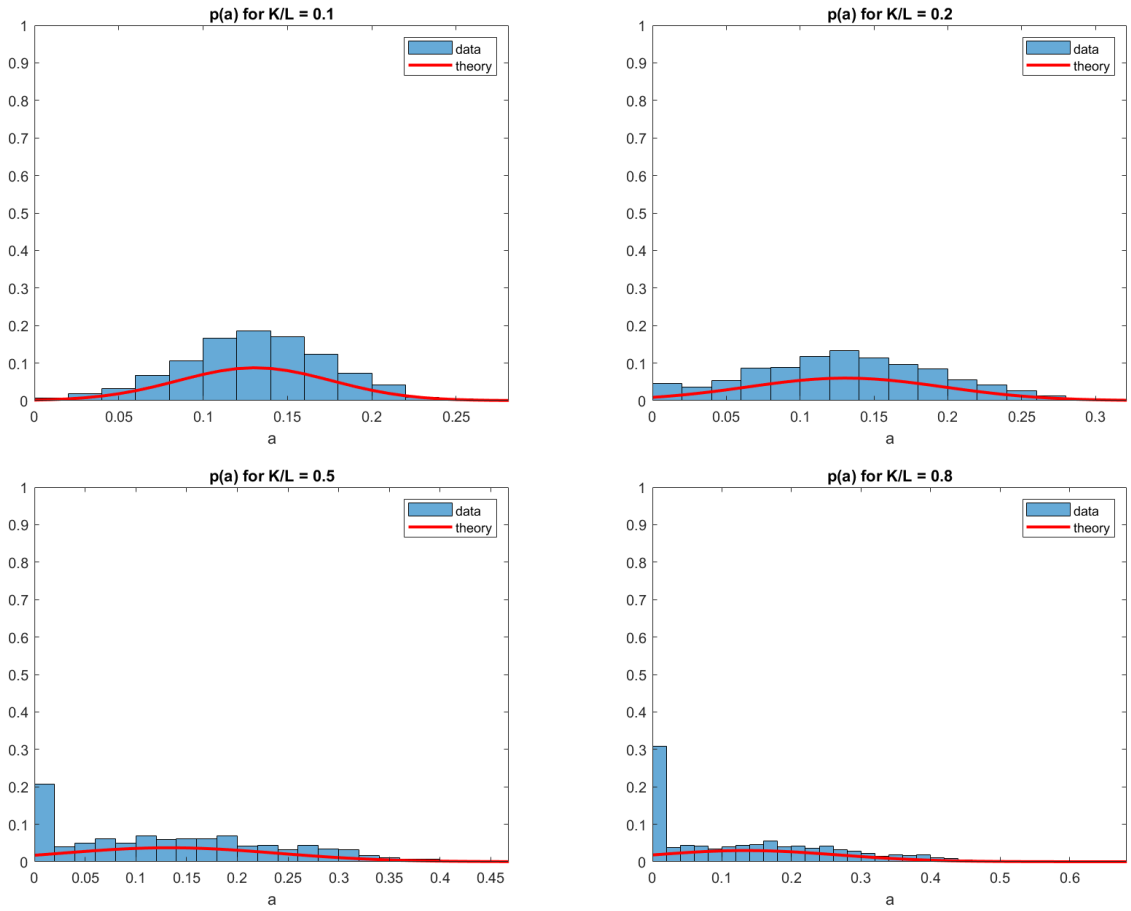

Figure A. 3: Comparing theoretical usage/expression distributions in terms of moments computed from cavity calculation with distributions of simulated data. Panels correspond to varying K/L ratio. Note that the discrepancy due to the peak at 0 expression ($a = 0$) when K/L is large is actually captured by the cavity calculation: the peak corresponds to the fraction with nonzero expression ϕ shown in previous figure.

Inserting (36) into (37) yields

$$\langle a \sum_i e_{\mu i} + \sigma_a \sum_i e_{\mu i}\alpha_i \simeq E + \sigma_e \epsilon_\mu \rangle. \tag{A.38}$$

Obviously the scaling of the random variables a_i will depend on the scaling we choose for the target statistics. In our simulations we chose $E = 1$ and $\sigma_E^2 \sim \mathcal{O}(1/K)$. Equating mean and scatter of (38) allows us to read off the scaling. The mean is simple:

$$a \sum_i e_{\mu i} \simeq E \implies a \sim \mathcal{O}(\frac{1}{pK}).$$

For the scatter, we need to do a bit more work. First, we square both sides to put it in terms of variances, and then separate out the diagonal and off-diagonal statistics:

$$\langle \sigma_a^2 \left(\sum_i e_{\mu i}^2 \alpha_i^2 \sum_{i \neq j} e_{\mu i} e_{\mu j} \alpha_i \alpha_j \right) \simeq \sigma_E^2 \epsilon_\mu^2 \rangle$$

$$\Rightarrow \sigma_a^2 \left(\sum_i \langle e_{\mu i}^2 \rangle \langle \alpha_i^2 \rangle + \sum_{i \neq j} \langle e_{\mu i} \rangle \langle e_{\mu j} \rangle \langle \alpha_i \alpha_j \rangle \right) \simeq \sigma_E^2$$

$$\implies \sigma_a^2 \sum_i \langle e_{\mu i}^2 \rangle \simeq \sigma_E^2.$$

Noting that we choose elements of the genome matrix from a binomial distribution so that $\langle e_{\mu i}^2 \rangle = p$, we conclude that the variance of gene expression scales as $\sigma_a^2 \sim \mathcal{O}(1/pK^2)$, or simply $\sigma_a^2 \sim 1/K^2$.

Scaling of Residuals

Here we make use of both numerics and analytics to find the scaling of the average residual component and their variance. Empirically, we have found that for a fixed K/L ratio, the fitness of a random K by L genome is independent of size of K and L. That is,

$$\left\| \sum_i a_i \vec{c}_i - \vec{E} \right\|^2 \sim \text{const.}, \tag{A.39}$$

where the constant depends on the given K/L ratio. Recall that at equilibrium in the CRM

$$R_\mu = E_\mu - \sum_i e_{\mu i} a_i$$

so that (39) can be rewritten as

$$\left\| \vec{R} \right\|^2 = \sum_\mu R_\mu^2 \simeq L \langle R_\mu^2 \rangle \sim \text{const.},$$

which implies $\langle R^2 \rangle \sim \mathcal{O}(1/K)$. Moreover, because the fitting process minimizes the residual components to roughly $R_\mu \approx 0$, we can assume that the mean $\langle R \rangle$ vanishes faster than the scatter such that $\langle R^2 \rangle$ is a good estimate for σ_R^2. Then the variance also scales as $\sigma_R^2 \sim \mathcal{O}(1/K)$. We have checked numerically the assumption that the mean vanishes faster than the scatter and found that $\langle R \rangle \sim \gamma/K$, where $\gamma \equiv K/L$.

Scaling of Response Coefficients

The following plots present the results of simulations used to determine the scaling of derivatives such as $\frac{\partial a_i}{\partial E_\mu}$. Numerical solutions of both the perturbed and unperturbed CRM dynamics (5) and (6) were found using MATLAB's ode45 solver. Steady-state values of variables were approximated and a threshold of 10^{-9} was used to determine if a steady-state value for a_i was 0 or not. These $a_i = 0$ were removed from the system of equations and the corresponding linear, equilibrium system was solved exactly. If any of the a_i from these solutions are negative, this indicates that our estimate of which a_i to be 0 is too conservative and these negative a_i were set to 0. Finally, the derivatives were computed numerically by taking the first finite difference between the perturbed and unperturbed solutions over the size of the perturbation.

Note, we assume the response to perturbations is in the linear regime so that we can perturb parameters E_μ and g_i separately. Because the variables a_i and R_μ scale with K themselves, we used re-scaled variables α_i and r_μ to work with quantities of order 1. A fixed ratio of $\gamma = 0.2$ was chosen, while K and L varied accordingly. The density of the random binary genomes was selected to be $p = 0.5$, and for simplicity we chose a homogeneous target (i.e., $\sigma_E^2 = 0$). Because all genes and residual components follow the same statistics, we only perturb one component of the target $E_1 + \delta E_1$ and one auxiliary parameter $g_1 + \delta g_1$. To ensure the scaling of the response derivatives were not due to a perturbation that scales with K (as is the case in the cavity method), we used a fixed perturbation of 10^{-3} for all K as well. Finally, we use bootstrapping to estimate the error on the variance of derivatives.

δg perturbation: averages

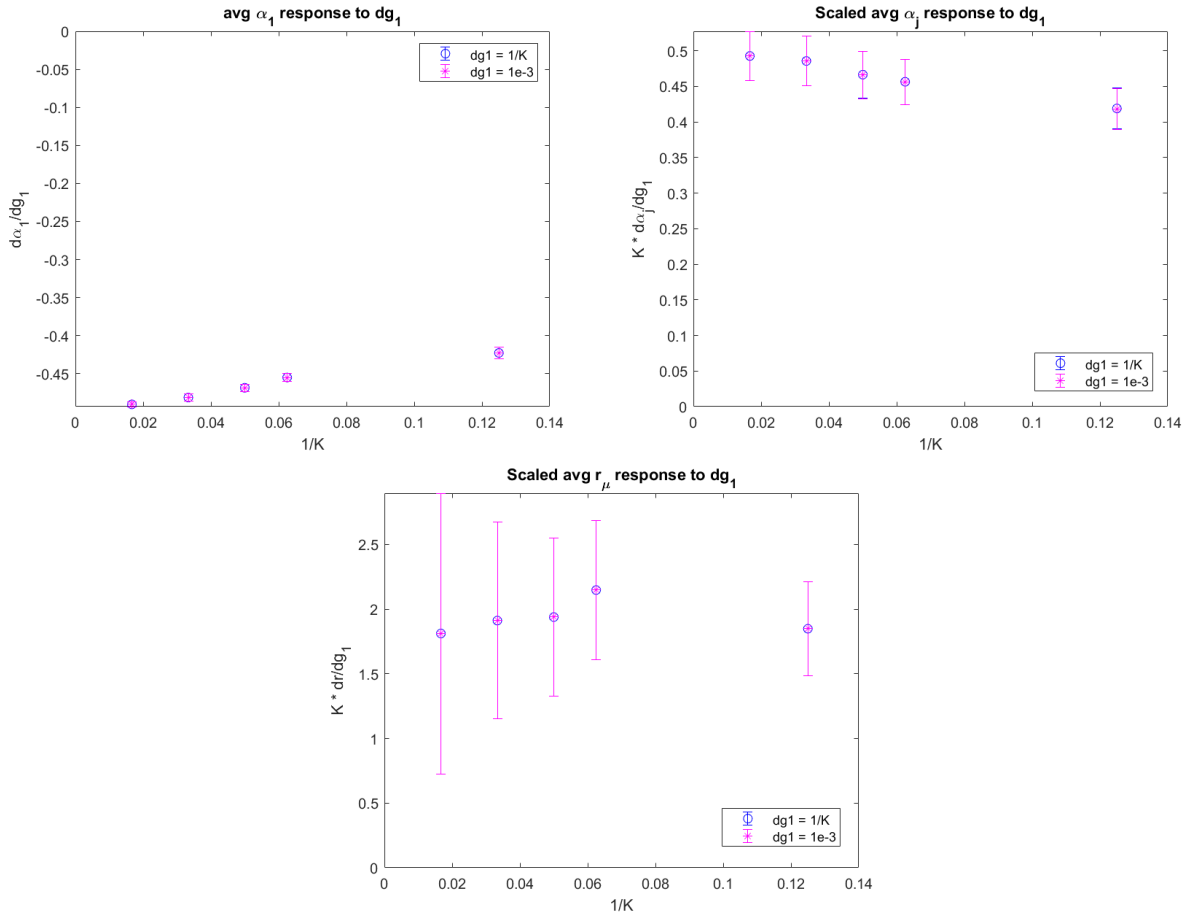

Figure A. 4: Numerical analysis of response variable means scaling with perturbation in cavity method. Each panel shows responses to small (magenta) perturbations and perturbations of size relevant to genome size (adding one gene/system to genome of already K genes/systems; in blue).

δE perturbation: averages

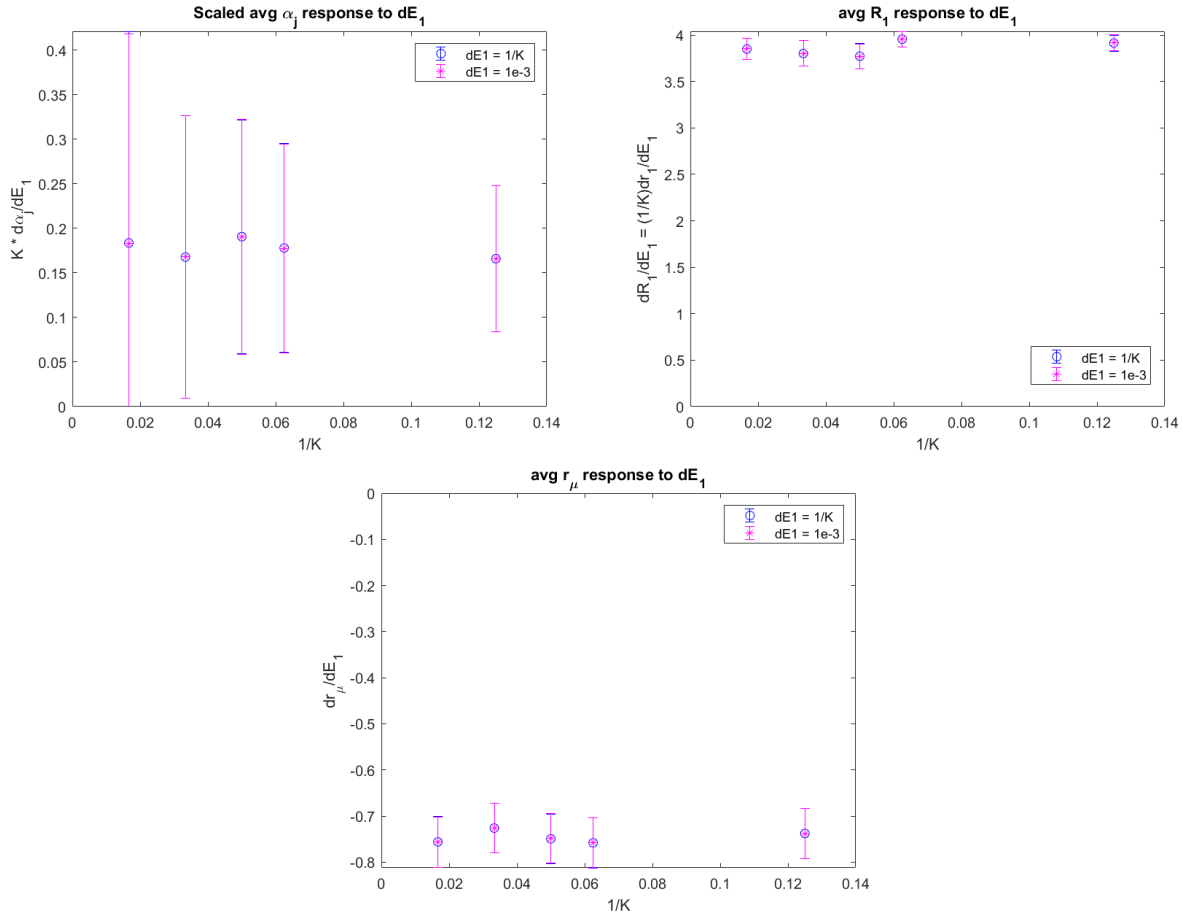

Figure A. 5: Numerical analysis of response variable means scaling with perturbation in cavity method. Each panel shows responses to small (magenta) perturbations and perturbations of size relevant to genome size (adding one gene/system to genome of already K genes/systems; in blue).

δg perturbation: variances

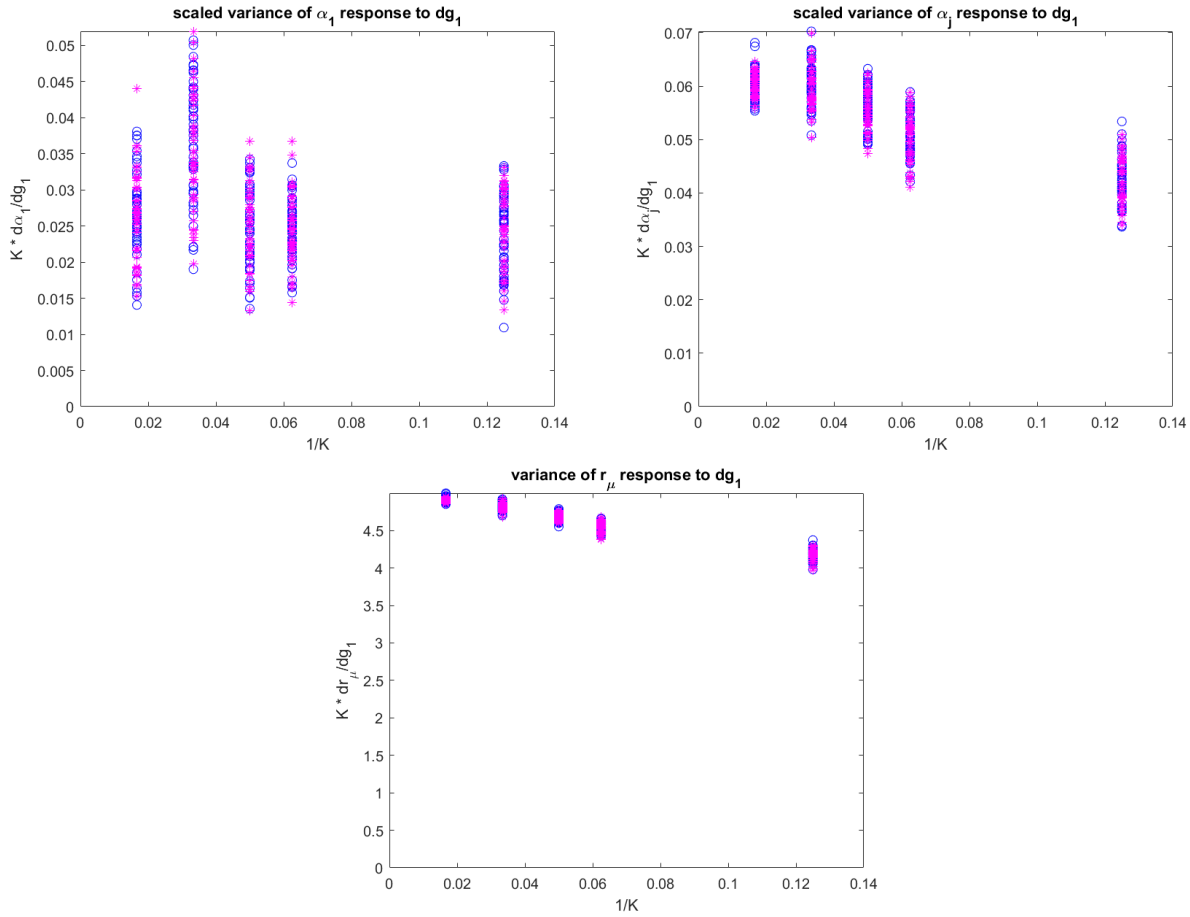

Figure A. 6: Numerical analysis of response variable variance scaling with perturbation in cavity method. Each panel shows responses to small (magenta) perturbations and perturbations of size relevant to genome size (adding one gene/system to genome of already K genes/systems; in blue).

δE perturbation: variances

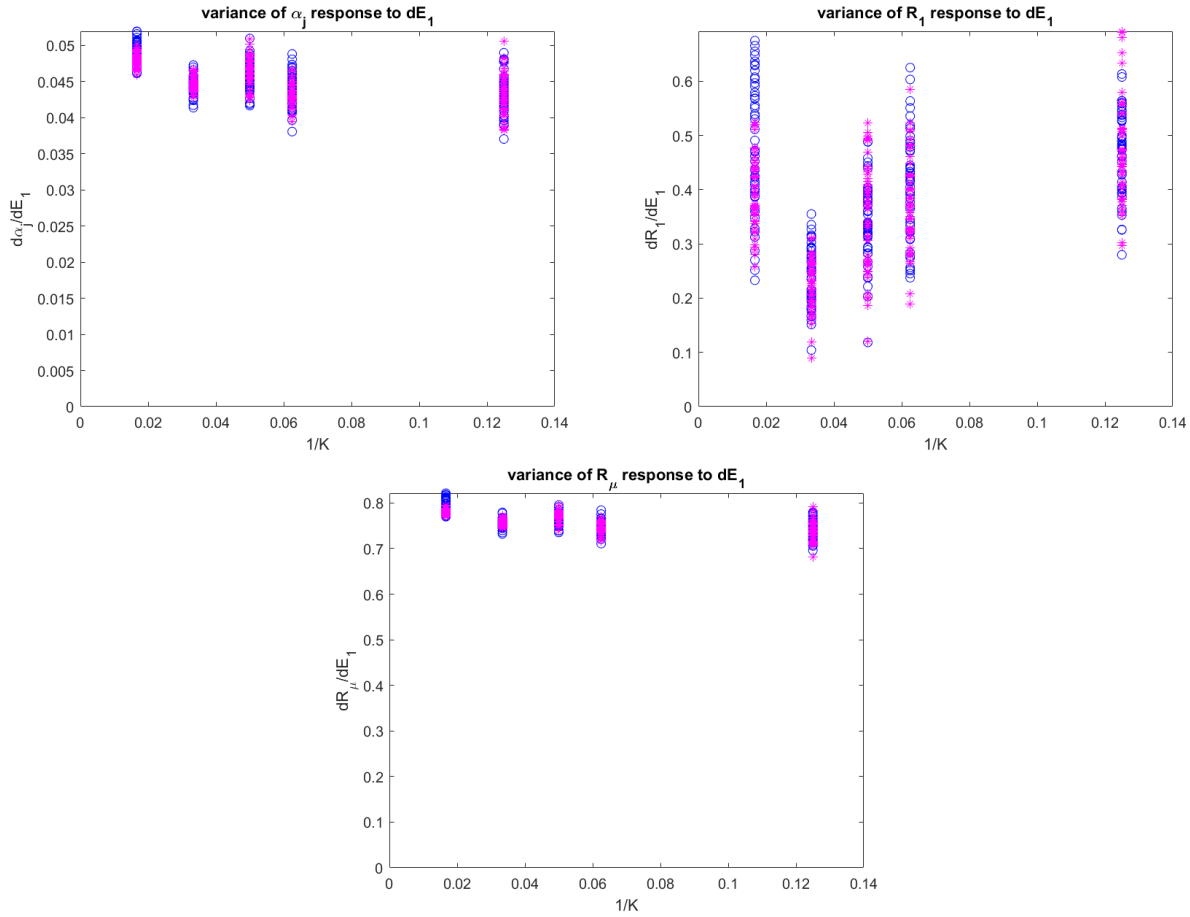

Figure A. 7: Numerical analysis of response variable variance scaling with perturbation in cavity method. Each panel shows responses to small (magenta) perturbations and perturbations of size relevant to genome size (adding one gene/system to genome of already K genes/systems; in blue).

Summary of Scaling Behaviors

variables:

$$\langle R\rangle \sim \frac{\gamma}{K}, \quad \sigma_R^2 \sim \frac{1}{K}$$
$$\langle a\rangle \sim \frac{1}{pK}, \quad \sigma_a^2 \sim \frac{1}{pK^2}.$$

scaled responses:

$$\langle\frac{\partial\bar{\alpha}_j}{\partial E_\mu}\rangle \sim \frac{1}{K}, \quad \langle\frac{\partial\bar{r}_\mu}{\partial E_\nu}\rangle \sim 1, \quad \langle\frac{\partial\bar{r}_\mu}{\partial E_\mu}\rangle \sim K$$
$$\langle\frac{\partial\bar{\alpha}_i}{\partial g_i}\rangle \sim 1, \quad \langle\frac{\partial\bar{\alpha}_j}{\partial g_i}\rangle \sim \frac{1}{K}, \quad \langle\frac{\partial\bar{r}_\mu}{\partial g_i}\rangle \sim \frac{1}{K}.$$

$$\mathrm{var}\left(\frac{\partial\bar{\alpha}_j}{\partial E_\mu}\right)\sim 1,\quad \mathrm{var}\left(\frac{\partial\bar{r}_\mu}{\partial E_\nu}\right)\sim K,\quad \mathrm{var}\left(\frac{\partial\bar{r}_\mu}{\partial E_\mu}\right)\sim 1$$
$$\mathrm{var}\left(\frac{\partial\bar{\alpha}_i}{\partial g_i}\right)\sim \frac{1}{K},\quad \mathrm{var}\left(\frac{\partial\bar{\alpha}_j}{\partial g_i}\right)\sim \frac{1}{K},\quad \mathrm{var}\left(\frac{\partial\bar{r}_\mu}{\partial g_i}\right)\sim 1.$$

inserting back the scaling $a_j \equiv \frac{\alpha_j}{pK}$, $R_\mu \equiv \frac{\gamma}{K} r_\mu$:

$$\langle\frac{\partial\bar{a}_j}{\partial E_\mu}\rangle\sim \frac{1}{K^2},\quad \langle\frac{\partial\bar{R}_\mu}{\partial E_\nu}\rangle\sim \frac{1}{K},\quad \langle\frac{\partial\bar{R}_\mu}{\partial E_\mu}\rangle\sim 1$$
$$\langle\frac{\partial\bar{a}_i}{\partial g_i}\rangle\sim \frac{1}{K},\quad \langle\frac{\partial\bar{a}_j}{\partial g_i}\rangle\sim \frac{1}{K^2},\quad \langle\frac{\partial\bar{R}_\mu}{\partial g_i}\rangle\sim \frac{1}{K^2}.$$

$$\mathrm{var}\left(\frac{\partial\bar{a}_j}{\partial E_\mu}\right)\sim \frac{1}{K^2},\quad \mathrm{var}\left(\frac{\partial\bar{R}_\mu}{\partial E_\nu}\right)\sim \frac{1}{K},\quad \mathrm{var}\left(\frac{\partial\bar{R}_\mu}{\partial E_\mu}\right)\sim \frac{1}{K^2}$$
$$\mathrm{var}\left(\frac{\partial\bar{a}_i}{\partial g_i}\right)\sim \frac{1}{K^3},\quad \mathrm{var}\left(\frac{\partial\bar{a}_j}{\partial g_i}\right)\sim \frac{1}{K^3},\quad \mathrm{var}\left(\frac{\partial\bar{R}_\mu}{\partial g_i}\right)\sim \frac{1}{K^2}.$$