THE UNIVERSITY OF CHICAGO

THE COMMUNITY FUNCTION LANDSCAPE: A RATIONAL FRAMEWORK FOR TRANSLATIONAL MICROBIOME ENGINEERING

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

INTERDISCIPLINARY SCIENTIST TRAINING PROGRAM:
ECOLOGY AND EVOLUTION

BY
MAHMOUD YOUSEF

CHICAGO, ILLINOIS
AUGUST 2025

© 2025 by Mahmoud Yousef

All Rights Reserved

For everyone - family, friends, mentors, colleagues - who made me who I am today

Alhamdulillah: praise God for everything

TABLE OF CONTENTS

LIST OF FIGURES viii

LIST OF TABLES x

ACKNOWLEDGMENTS xi

ABSTRACT xiii

1 INTRODUCTION 1
- 1.1 Complex Functions of Microbiomes 1
- 1.2 Challenges in Synthetic Community Design 2
- 1.3 Thesis Aims 4
  - 1.3.1 Aim 1: Establish the Community-Function Landscape as a Framework for Understanding and Predicting Community Function 5
  - 1.3.2 Aim 2: Engineer Microbial Communities that Remediate Organic Pollutant While Resisting Toxicity 6
  - 1.3.3 Aim 3: Engineer a Core Ecologically-Interacting Community that Clears a Human Gut Pathogen 7

2 DEVELOPING FUNCTIONAL LANDSCAPES OF MICROBIAL COMMUNITY FUNCTION 8
- 2.1 Introduction 8
- 2.2 Results 10
  - 2.2.1 Formulating the Approach to Fitting Community-Function Landscapes 10
  - 2.2.2 Community-Function Landscapes are Learnable Via Regression and Enable Predictions 11
  - 2.2.3 Epistasis is Rare in Many Functional Landscapes 14
  - 2.2.4 Ecological Models Indicate When Landscapes Can Be Learned 18
  - 2.2.5 Rugged Landscapes are More Difficult to Infer 19
- 2.3 Discussion 23
- 2.4 Methods 24
  - 2.4.1 Datasets 24
  - 2.4.2 Regressions 26
  - 2.4.3 Calculating Epistasis 26
  - 2.4.4 Simulations of the generalized Lotka-Volterra model 27
  - 2.4.5 Simulations of the consumer-resource model with random cross-feeding networks 29
  - 2.4.6 Inferrability of Rugged Landscapes 30
  - 2.4.7 Data and Code Availability 31
- 2.5 Supplementary Discussion 32
  - 2.5.1 Low Ruggedness in Landscapes is not an Artifact of LASSO Regularization 32

2.5.2 Understanding the Impact of Undersampling on Regression Fits . . . 34
2.5.3 Supplementary Figures . . . 37

3 EPISTASIS DRIVES TOXICITY RESISTANCE IN BIOREMEDIATION AND ENABLES RATIONAL DESIGN . . . 41
3.1 Introduction . . . 41
3.2 Results . . . 43
3.2.1 A Library of Isolates that Enable Community Design for BPA Degradation . . . 43
3.2.2 Low-Dimensional Variation in Community Function . . . 46
3.2.3 Inferring the Functional Landscape of BPA Degradation Despite High Epistasis . . . 49
3.2.4 Epistasis is the Primary Driver of Toxicity Resistance . . . 52
3.2.5 Functional Landscape Enables Generation of Unseen Communities . . . 55
3.2.6 Enabling Design for Bioremediation in Soils . . . 57
3.3 Discussion . . . 60
3.4 Methods . . . 62
3.4.1 Isolation of Strains . . . 62
3.4.2 BPA Concentration Assay . . . 64
3.4.3 Synthetic Community Design . . . 64
3.4.4 Imputation to predict AUCs across concentrations of BPA . . . 66
3.4.5 Regressions . . . 66
3.4.6 PCA Analysis . . . 68
3.4.7 BPA Extraction and Community Inoculation in Soils . . . 68
3.4.8 Data and Code Availability . . . 69
3.5 Supplementary Discussion . . . 69
3.5.1 Measuring CO_2 Production from BPA Degradation . . . 69
3.5.2 Stability Analysis Shows Covariance Between Features Impedes LASSO Regularization . . . 71
3.5.3 Regression Models are Statistically Significant . . . 74
3.5.4 Second-Order Models Outperform Models Truncated at First- and Third- Order Terms . . . 76
3.5.5 Fitting a Low Rank Regressor . . . 77
3.5.6 Supplementary Figures . . . 81

4 RATIONAL DESIGN OF A SYNTHETIC COMMUNITY THAT SUPPRESSES A MULTI-DRUG RESISTANT HUMAN GUT PATHOGEN . . . 89
4.1 Introduction . . . 89
4.2 Results . . . 90
4.2.1 Designing Communities for the Clearance of *K. pneumoniae* . . . 90
4.2.2 A RandomForest Regressor Infers the Landscape of *K. pneumoniae* Suppression and Enables Community Design . . . 93
4.2.3 Community Structure Reveals 15-member Group Essential ('SynCom15') for *K. pneumoniae* Suppression . . . 94

4.2.4 SynCom15 Clears *K. pneumoniae* in Different *In-Vitro* Environments 98
4.2.5 SynCom15 Clears *K. pneumoniae* in Clinically Relevant Mouse Model of Infection . . . 99
4.3 Discussion . . . 104
4.4 Methods . . . 105
4.4.1 Creation and whole genome sequencing of strain bank . . . 105
4.4.2 Initial design strategy for bacterial communities . . . 109
4.4.3 Experimental workflow for *K. pneumoniae* clearance assay . . . 110
4.4.4 Training and validation of RandomForest (RF) model . . . 111
4.4.5 PC analysis of in-silico consortia . . . 112
4.4.6 Computing Similarity of Strain Projections onto PC46 . . . 115
4.4.7 Characterization of Mice Used for All Experiments Spanning Figures 4.3 and 4.4 . . . 116
4.4.8 Creating GF and antibiotic (Ab)-SPF cecal extract media . . . 116
4.4.9 Preparation of Mice Stool Samples for Fecal Microbiota Transplant (FMT) . . . 117
4.4.10 SPF Mouse Model of *K. pneumoniae* Infection . . . 117
4.4.11 Determining Engraftment of SynCom15 Strains in SPF Mice . . . 118
4.4.12 Data and Code Availability . . . 118
4.5 Supplementary Discussion . . . 119
4.5.1 RandomForest Regressor Outperforms Linear Regressions . . . 119
4.5.2 Supplementary Figures . . . 120

5 CONCLUSION . . . 123
5.1 Future Directions . . . 124

REFERENCES . . . 128

LIST OF FIGURES

1.1 The Community Function Landscape Approach to Predicting Community Function 4

2.1 Community function is predictable from species presence/absence in empirical datasets. . . . 13
2.2 Empirical community-function landscapes are not rugged . . . 16
2.3 Ecological models indicate both optimism and caution for inferring community-function landscapes. . . . 20
2.4 Simulations Show Rugged Landscapes are Harder to Infer . . . 22
2.5 Ridge Regularization Does Not Lead to Rampant Epistasis in Inferred Landscapes 33
2.6 Regressions Tolerate a Moderate Level of Undersampling . . . 36
S.2.1 Residuals obtained from second-order regressions . . . 37
S.2.2 Model residuals obtained from second-order regressions as a function of community richness . . . 38
S.2.3 Model predictions from second-order regression fits with all data in-sample . . . 39
S.2.4 Model predictions from second-order regression fits with all data in-sample and experimental replicates included . . . 40

3.1 Isolates enable construction of synthetic bacterial communities that degrade BPA 45
3.2 BPA Degradation Varies in a Low Dimensional Manner Across Concentrations . 47
3.3 Inferring Community-Function Landscapes Shows Prominent Epistasis . . . 51
3.4 Epistasis Increases Gradually as Toxicity Increases . . . 54
3.5 Community Function Landscape Successfully Guides the Design of New, Functional Communities . . . 56
3.6 Designing for BPA Remediation in Soils . . . 58
3.7 Standard Curve of BPA Concentration Assay . . . 65
3.8 BPA Extraction Efficiency from Soil Slurries . . . 70
3.9 CO_2 production by BPA Degraders during BPA degradation . . . 72
3.10 LASSO Regularizations are Unstable with BPA Data . . . 73
3.11 Linear Regression Models . . . 75
3.12 Low-Rank Regressor Fits to BPA AUC Data . . . 80
S.3.1 Depiction of the BPA Degradation Landscape Across the Five Initial Concentrations 81
S.3.2 Random Communities of Soil Isolates Do Not Degrade BPA . . . 82
S.3.3 BPA Degradation of the isolated 16 strains from 30 ppm BPA . . . 83
S.3.4 Comparing Degradation AUC Across the 5 Initial Concentrations . . . 84
S.3.5 SoftImpute Optimization and Further Validation . . . 85
S.3.6 Analyzing Regression Coefficients to Understand the Functional Landscape . . . 86
S.3.7 Smooth Variation Between Landscapes Enable Cross-Predictions . . . 88

4.1 Creating a Generative Model for the Suppression of *K. pneumoniae* . . . 92
4.2 Statistical Analysis Reveals Blocks of Strains Correlated with *K. pneumoniae* Clearance . . . 96
4.3 SynCom15 Clears *K. pneumoniae* in Multiple *in-vitro* Environments . . . 98

4.4 SynCom15 Sustainably Clears *K. pneumoniae* in a Clinically Relevant Mouse Model of Infection 101
4.5 Algorithm for Creative Diverse Communities Via UMAP 109
4.6 RandomForest Regressor Outperforms Linear Regressions 120
S.4.1 No Individual Strain Comprising SynCom15 Suppresses *K. pneumoniae* On Its Own 121
S.4.2 RandomForest Regressor Feature Importance Scores 122

LIST OF TABLES

2.1 Calculated F_1 Statistics for Six Inferred Landscapes 15
2.2 Information about the six datasets used in this chapter 26
2.3 Calculated F_1 Statistics for Six Inferred Landscapes with Ridge Regularized Regressions 32

3.1 Information about the seven soil samples obtained in this chapter 62
3.2 Information about the 16 strains isolated 63
3.3 L_2 Penalty Values for Five Regressors on BPA Degradation 68

4.1 RandomForest Improves over Linear Additive Model 95
4.2 Strain identification of the 46 strains comprising the restricted strain bank . . . 107
4.3 Regression of principal components against predicted *K. pneumoniae* abundance 113
4.4 Comparing Linear Regressions Performance to RandomForest 119

ACKNOWLEDGMENTS

I've always heard that it takes a village to raise a PhD, and after experiencing it myself, I assert with complete confidence that the sentiment rings true.

I would like to thank my advisors - Dr. Seppe Kuehn and Dr. Arjun Raman - for their time and effort in advising me and mentoring me throughout the entire duration of my PhD. Both are incredible mentors, and heavily invested in my growth as a scientist much more than progressing the research within this thesis. I also am grateful for my thesis committee members - Dr. Tim Wootton who not only chaired my committee but also ensured I was connected to the Department of Ecology and Evolution despite the mess caused by COVID-19, and Dr. John Novembre whose statistical expertise guided nearly every aspect of this work.

The work presented in this thesis is the result of many joint efforts with lab members and other collaborators. I am deeply indebted to Abigail Skwara and Drs. Karna Gowda and Mikhail Tikhonov for their incredible contributions to the data in Chapter 2, Kiseok Lee and Jonathan Tang for their expertise with soil experiments that enabled the results in Chapter 3 to come together, and Drs. Rita Oliveira and Robert Chen for pioneering *in vivo* experiments that tied togehter Chapter 4. I also thank the many conversations with Kuehn and Raman lab members for their advice throughout this entire project.

The collaborative training environment at the University of Chicago MSTP is something I will always cherish as being an essential part of my training. I thank the MSTP for their constant support and guidance in helping me navigate the program. I thank the Pritzker School of Medicine for opening their doors for clinical training and the network of physicians that guided my mentorship by these incredible doctors. I thank the Department of Ecology and Evolution for welcoming me to their vibrant community of scholars and peers.

My family - mom, dad, sisters - have had my side throughout my entire endeavor. They have witnessed my ups and downs, and were always ready to be by my side whenever anything

went wrong, even if it be driving for an hour to deliver my laptop charger that I forgot at their place when visiting them. I love you all so much, and thank you for everything.

To my loving wife: Esraa. We spent most of my PhD in a long-distance marraige waiting for your green card. I don't know anyone who could stand all of that stress and pressure just for me, but you're a real trooper. Thank you.

And most importantly of all, my baby daughter Layal, born in January of this year to light up my world when the workload was at its most difficult. The only one who I can say that, no matter how difficult my day has been, never failed to put a smile on my face. I love you. I can't wait to watch you grow up.

ABSTRACT

The metabolic activity of microbial communities is essential for environmental and host health, influencing processes from bioremediation to host immune health. Given this importance, the rational design of microbiomes with targeted functional properties is an important objective. Designing microbial consortia with targeted functions is challenging due to complex community interactions and environmental heterogeneity. In this thesis, we introduce the Community Function Landscape - a statistical framework that maps the composition of a community to its function. Validation on six datasets of microbial community function reveal that the landscapes capture the average functional impact of each strain, enabling accurate predictions of community function. We apply the functional landscape to two highly complex microbial functions: the degradation of the toxic organic compound BPA, and the suppression of the human gut pathogen *Klebsiela pneumoniae*. In both cases, the functional landscape revealed statistical interactions ('epistasis') between strains that guided the design of consortia with optimal functional performance. Overall, the results of this thesis establish the Community Function Landscape as a framework for the rational design of microbial communities for application in contexts relevant for human and environmental health.

CHAPTER 1

INTRODUCTION

1.1 Complex Functions of Microbiomes

Microbes pervade nearly every environment on the globe [1, 2]. Large communities of microbes - microbiomes - are responsible for many of the geochemical processes that enable life on Earth, including nitrogen [3, 4] and carbon cycling [5]. Many microbes also exhibit functional properties that directly impact human health. Although some single strains are capable of these health-related functions [6], many of these functions emerge from communities composed of diverse taxa, such as metabolite production [7], pathogen suppression [8, 9], and detoxification [10, 11], (Fig. 1.1A). It is well-recognized that many of these functions arise from collective properties of diverse microbiomes, rather than summing together the activities of individual members [12–14]. In other words, 'the whole is greater than the sum of the parts'.

Utilizing microbial communities to optimize their functional properties for human benefit has been a long-standing endeavor. Indeed, humans have manipulated environmental conditions for millennia to recruit microbial communities with desirable properties, from food fermentation [15] to wastewater treatment [16]. Recently, efforts to rationally design consortia with defined composition and targeted functions have seen a rise in interest [17, 18]. However, such pursuits prove to be difficult, since microbial community function is affected by many factors, including cellular gene expression [1, 19], ecological interactions [12, 20, 21], intra-specific variation [22], and environmental influences [23–25]. All of these different 'parts' within microbial communities interact in dynamic and unintuitive manners to give rise to collective functional properties not found in the activity of individual constituents. Therefore, linking community composition to community function remains a difficult challenge, yet an essential one to overcome.

1.2 Challenges in Synthetic Community Design

The main challenge of rational community design is elucidating the linkage between community composition (its structure) and function [20]. This linkage would unearth 'design principles' - benchmarks to meet when designing communities in order to achieve the desired functional property [17, 18, 24]. Understanding and navigating this linkage would then facilitate the rational synthetic design of microbial communities. This philosophy drives much of the interest in microbial engineering.

One technique of unveiling design principles attempts to learn from functional natural communities. This method, known as the "top-down approach," enriches undefined microbiomes from their environmental contexts and inoculates them in synthetic conditions to perform their function [26, 27]. While this approach seems promising in fields such as bioremediation [28, 29], a significant drawback lies in the design space's restriction to the composition of the extracted community [18]. Any analysis of design principles are limited to the mechanisms of the extracted community, and thus building consortia to optimize a function is restricted to the limits of the extracted community [28, 30]. In addition, attempting to analyze such communities typically results in incomplete information on metabolic pathways, where the biochemical reactions (and the enzymes involved) of intermediary steps are often unknown [31, 32], preventing a complete understanding of how such a community performs its function. Furthermore, these extracted communities often exhibit random behavior with unreliable functional output [33, 34]. Any resulting analysis is therefore restricted to only a "local" optimization since the metabolic space for optimization is limited to the space occupied by the enriched community [18], versus a more "global" optimization, attempting to generalize the metabolic design space beyond a single working example.

Instead of learning from undefined microbiomes, communities can be defined by building a network of known metabolic and ecological linkages. In this "bottom-up" technique, species are combined into consortia based on genomic and metabolic properties thought to produce

the desired function cooperatively [18, 20, 27]. Such an approach suffers from a formidable challenge: collective properties of microbial communities cannot be easily deduced from single strain data [29, 35, 36], typically due to interaction effects between strains [12, 21, 37]. While interactions between pairs of strains are widespread [27, 35], larger groups of strains exhibit altered activity in the presence of an additional strain due to interactions between that strain and the rest of the group ("higher-order interactions") [12, 13, 38, 39]. Predicting these interaction effects requires knowledge of what functions specific taxa can perform from genetic data [18]. However, many genes have unknown functions, and predicting a strain's role in a particular pathway (i.e., asking whether a strain performs a specific biochemical reaction) by analyzing gene content is limited [26, 27, 40]. Therefore, we cannot accurately predict how large consortia perform their functions.

Despite the many challenges above, many efforts have cleverly employed a variety of techniques in synthetic design efforts to optimize functions. Efforts to combine top-down and bottom-up approaches yielded keystone species that were essential for identifying necessary interactions that enabled targeted design [28]. Systems biology approaches attempt to gather large-scale data to enable near-perfect descriptions of microbial systems, in order to quickly scale design efforts to find optimal communities [30]. Other endeavors relied on ecological modeling to simplify bottom-up approaches [37, 41]. Generalized Lotka-Volterra models have been used to predict interactions between pairs of strains that enabled rational design of complex and diverse communities [7]. Consumer-Resource Models also served to accurately predict nitrogen-fixation dynamics in co-cultures from single-strain genotypic and phenotypic data alone [42]. Neural networks have even been deployed in order to more accurately model complex interactions within microbiomes [41, 43]. However, all of these approaches required genotypic or phenotypic data of single strains, and/or data on interactions between pairs of strains, in order to enable predictions - information that is not always available [30]. Furthermore, the translation from *in-vitro* community functions to *in-vivo*

performance remains a deep mystery to solve [24].

1.3 Thesis Aims

Rationally building microbial communities with optimal desirable function remains a difficult challenge. While many statistical approaches have successfully created functional communities [31, 44, 45], developing a 'generative' model that reliably predicts the function of unseen communities remains challenging [46]. Results indicated that the ideal approach for learning design principles is to learn from already functional communities [7].

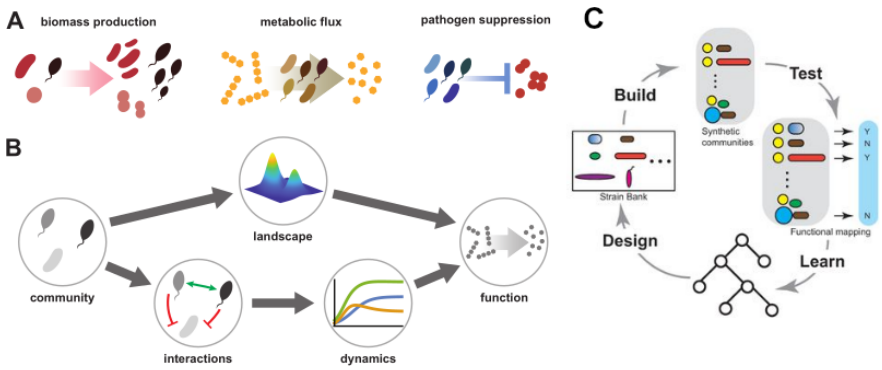


Figure 1.1: **The Community Function Landscape Approach to Predicting Community Function** (**A**) Examples of microbial community functions including (left to right): production of biomass, conversion of substrate to product, and suppression of a pathogen. (**B**) Contrasting the statistical landscape view (top) of predicting community function with the dynamical view (bottom). In the dynamical view, species abundance dynamics are predicted via an ecological model, which integrates knowledge or measurements of interactions between populations. In contrast, the statistical landscape approach neglects dynamics and measures community function for a set of consortia, allowing functions for all possible community combinations to be inferred statistically. (**C**) A standard Design-Build-Test-Learn (DBTL) framework. Communities are designed (D) and built (B) from a strain bank, tested (T) for desired function, and a statistical model mapping community composition with function is learned (L). New communities are then designed based on the learned model and the process is iterated.

In this work, we introduce the community function landscape as a statistical tool by which a quantifiable microbial property can be described as a function of the composition of a community. By leveraging approaches developed in the field of genetics, we can build

a simple statistical model which considers only strain presence and absence as its input - thus sidestepping the many challenges of rational design specified earlier. The community function landscape learns from a set of communities in order to predict other communities, thus allowing for community effects to be learned directly [35]. We also explicitly avoid the need to measure dynamical system properties, thus simplifying the statistical approach (Fig. 1.1B). The landscape approach can be applied to the 'Design-Build-Test-Learn' cycle of microbiome engineering [7, 26](Fig. 1.1C), where learning community effects enable the generation of highly functional communities.

In this thesis, we will address each of the following three aims:

1.3.1 Aim 1: Establish the Community-Function Landscape as a Framework for Understanding and Predicting Community Function

Microbial consortia exhibit complex functional properties in contexts ranging from soils to bioreactors to human hosts. Understanding how community composition determines function is a major goal of microbial ecology. Here we address this challenge using the concept of community-function landscapes – analogs to fitness landscapes – that capture how changes in community composition alter collective function. Using datasets that represent a broad set of community functions, from production/degradation of specific compounds to biomass generation, we show that statistically-inferred landscapes quantitatively predict community functions from knowledge of species presence or absence. Crucially, community-function landscapes allow prediction without explicit knowledge of abundance dynamics or interactions between species, and can be accurately trained using measurements from a small subset of all possible community compositions. The success of our approach arises from the fact that empirical community-function landscapes appear to be not rugged, meaning that they largely lack high-order epistatic contributions that would be difficult to fit with limited data. Finally, we show that this observation holds across a wide class of ecological models,

suggesting community-function landscapes can be efficiently inferred across a broad range of ecological regimes. Our results open the door to the rational design of consortia without detailed knowledge of abundance dynamics or interactions.

1.3.2 Aim 2: Engineer Microbial Communities that Remediate Organic Pollutant While Resisting Toxicity

Bisphenol-A (BPA) is a carcinogenic and endocrine-disrupting pollutant that shows poor promise as a bioremediation target due to bacterial susceptibility to toxicity, slow rates, and incomplete BPA metabolism. Although communities of microbes may overcome these challenges by having strains complement each other, a lack of understanding of mechanism of BPA degradation and toxicity resistance have any practical efforts difficult. Here, we apply the community-function landscape approach to understand the mechanism of toxicity resistance and guide the design of functional consortia. Leveraging the community function landscape with 70 synthetic communities inoculated at five initial concentrations with BPA, we discover that the variation between the activity of the communities is low-dimensional with BPA toxicity. Inspection of the coefficients reveal that although epistasis dominates the landscapes, it rises gradually with toxicity, and that strong epistatic interactions are the key to toxicity resistance. Further, we utilize landscapes to rationally design communities with predictable BPA degradation dynamics *in-vitro*. Remarkably, designed synthetic communities are able to remediate BPA in contaminated soils. Our results demonstrate that toxicity can drive epistatic interactions in community-function landscapes and that these landscapes can guide microbial engineering for translational use.

1.3.3 Aim 3: Engineer a Core Ecologically-Interacting Community that Clears a Human Gut Pathogen

The Human Gut Microbiome is a large ecosystem with many moving parts in terms of interactions, metabolites, and dynamics that all affect its functional properties. Engineering consortia for specific functional properties within the human gut advantageous for human health remains an increasingly important endeavor, although elucidating design principles is immensely challenging. From a diverse set of human gut commensal strains and clearance of multi-drug resistant *Klebsiella pneumoniae* as the functional target of interest, we used a community function landscape to describe pathogen suppression in a manner agnostic to mechanism of action, bacterial interactions, or compositions of natural microbiomes, using a highly nonlinear regression formalism. Statistical inference performed on the landscape identified 15 strains that were key for community function. Combining these strains into a community ('SynCom15') suppressed *K. pneumoniae* across unrelated *in-vitro* environments as well as in clinically relevant *in-vivo* models of infection. Analysis revealed that members of SynCom15 interacted to enable engraftment into mice as well as rapid *K. pneumoniae* clearance. These results show that the community function landscape can be successfully employed to describe a complex community function involving high-order interactions, and can be used to identify 'super communities' for optimal functionality.

CHAPTER 2

DEVELOPING FUNCTIONAL LANDSCAPES OF MICROBIAL COMMUNITY FUNCTION

2.1 Introduction

Biology is a science of the organization of systems, relying on the study of components to understand changes within a system. On the level of proteins, protein function relies heavily on interactions between amino acids and the resulting protein folding. On the level of organisms, genotypic variation drives changes in phenotype and behavior. In both cases, interactions between constituent parts give rise to functional system properties, meaning that the whole of a system is much greater than the sum of its parts.

One of the most powerful conceptual frameworks for understanding the role of components within a system is through a landscape, where the height of the landscape encodes a scalar-valued function or fitness of the system, and position on the landscape corresponds to a particular configuration of components. A landscape then serves as a mapping between the configuration of components within a system to a scalar output description of the state of the system [47]. These landscapes enable exploration of how changing the composition of parts of a system affect the system's state. For example, fitness landscapes have been widely used in genetics to understand the effect of mutations on gene fitness [48], interactions between genes and their effect on phenotype [49, 50], and evolutionary trajectories [51–53].

Importantly, these fitness landscapes sidestep the need to incorporate information on dynamic processes affecting protein fitness. Rather, this approach attempts to understand the average impact a residue or mutation exerts on a protein's fitness compared to the *background average* of that fitness [52, 54]. Notably, these landscapes only consider the presence or absence of residues or mutations, as well as their co-occurence and discordance to estimate average impact on protein activity (a phenomenon known as background epistasis) [54,

55]. For example, protein fitness landscapes based on linear regressions do not account for complex interactions between residues that promote specific folding states to enable protein function, but rather look at the co-occurrence and discordance of residues that *on average* contribute to specific folding states [50, 52]. Despite this dramatic simplification, this approach to fitness landscapes has shown to be highly successful at predicting protein fitness.

Microbial communities as a system resemble protein function. The constituents of microbial communities (single strains) are analogous to residues or mutations in a fitness landscape, and the measured community function is analogous to protein fitness. The functions of microbial communities are also governed by complicated dynamical interactions between strains [13, 14, 21], similarly to interactions between protein residues and mutations. These interactions can potentially be simplified as understanding the 'average functional impact' of a strain's presence or absence in a community [39] through a Community Function Landscape.

This analogy inspires a landscape approach to understanding quantifiable microbial community function (Fig. 1.1B), given the successes of protein fitness landscapes. By utilizing six available datasets of community composition and a measured community function, we fit a regression-based functional landscape and show that the landscape accurately describes community function in all six cases. Inspection of these landscapes show that epistasis is rare in microbial function landscapes, indicating that strain presence/absence dominates the landscapes over interaction terms. These results highlight that learning the properties of communities can be accomplished without a detailed understanding of the interactions between taxa or their abundance dynamics.

2.2 Results

2.2.1 Formulating the Approach to Fitting Community-Function Landscapes

We first began by defining the statistical approach to fitting community-function landscapes using data on microbial community functions. For each experiment with microbial communities, a defined pool of N total species was used to construct a set of s communities. Each community was incubated for a set amount of time, and then a functional property of interest was measured. There are a total of $2^N - 1$ possible combinations of species, and frequently scaling to all possible combinations is infeasible for most properties of interest. We therefore are interested in defining a functional landscape approach that can fit to a small set $s << s^N - 1$ of communities while also providing a global picture of the landscape.

We formulate the problem as a linear regression of the following form:

$$y = \beta_0 + \sum_i \beta_i x_i + \sum_{i,j} \beta_{ij} x_i x_j + ... \tag{2.1}$$

where y is the measured functional property of interest of a community, and x_i represents the presence/absence of strain i in that community. β_i represents the additive effect of strain i in a community, and β_{ij} represents the interaction effects between strains i and j, which is a measure of of impact beyond the additive effects of both strains i and j individually. This is analogous to pairwise epistasis in protein fitness landscapes, and in this context we call β_{ij} epistatic interactions between strain i and j. Ellipses in Eqn. 2.1 denote higher-order epistasis terms, such as three-way epistatic terms (β_{ijk} denoting interactions between strains i, j, and k).

For denoting a strain's presence and absence in a community, we adopt the convention that $x_i = 1$ if species i is present and $x_i = -1$ if species i is absent. This convention (corresponding to the Fourier expansion of the landscape) simplifies interpretability of regression coefficients (see discussion in References [56, 57]). In brief, using $x_i = \pm 1$ allows β_i to be

understood as the average impact of adding or removing strain i as compared to all theoretical community compositions. Similarly, β_{ij} is understood to be the average background epistatic interaction between strains i and j [39, 56, 58]. This allows us to directly analyze model coefficients and make statements of the global functional landscape without measuring all $2^N - 1$ possible communities.

2.2.2 Community-Function Landscapes are Learnable Via Regression and Enable Predictions

After defining the approach to fitting community-function landscapes, we sought to apply the formalism to a number of available datasets of microbial community function. We compiled six datasets where synthetic bacterial communities were constructed from a defined pool of strains. These datasets measured a variety of community functions, including metabolite production [7, 59] and degradation [21], growth activity on a single carbon source [60], total community growth (see Section 2.4.1), and abundance of a single target strain within a community [61]. All six datasets fit our experimental conditions outlined previously - synthetic communities built from a specified pool of strains are inoculated for a set amount of time, and a functional property is measured in such a way that it can be represented as a scalar quantity. Details about the species pools and community constructions for each dataset is found in Tbl. 2.2.

We fit a regression of the form Eqn. 2.1 to each of the six datasets. Since most of the datasets utilized a pool of communities $s << 2^N - 1$, we employed LASSO regularization during inference [62], using cross-validation to optimize the regularization hyperparameter (see Section 2.4.2).

We investigated truncating the model at successively higher orders of epistatic terms in order to determine what degree of model complexity is needed to accurately predict community function. To assess out-of-sample generalization error of the regressions, we

employed an additional leave-one-out cross validation scheme, where each data point was iteratively left out and the model was fit to the remaining data communities. This allows for out-of-sample predictions for each distinct community. The quality of the fits are shown in Fig. 2.1 for increasingly complex models, with shaded bars showing in-sample R^2 and white bars showing out-of-sample R^2 (see Section 2.4.2). Scatter plots visualize the quality of second-order out-of-sample model predictions.

Remarkably, across all six datasets studied here, models of first or second order provide high-quality predictions ($R^2 \sim 0.8$ for second-order models). In most cases, additive models alone (e.g., $y = \beta_0 + \sum_i \beta_i x_i$) already have strong predictive power ($R^2 > 0.5$), although the addition of second-order terms yielded an increased quality of fit for all datasets. We note that residuals of observed versus predicted values demonstrate similar patterns of heteroskedasticity across datasets, where communities with higher values of the function relative to the mean tend to be underestimated, and those with lower values of the function tend to be overestimated (Fig. 2.1, Fig. S.2.1). It is possible that these patterns are a consequence of bias induced by regularization [63]. We note that errors in our predictions do not correlate with community richness (Fig. S.2.2) indicating that our models generalize well to diverse communities.

For the dataset by Clark et al. (Fig. 2.1A), the authors of the original study predicted butyrate production using a complex model that parameterized interactions and abundance dynamics. The quality of our statistical predictions using species presence/absence are similar to those obtained using a complex dynamical model [7], suggesting that detailed dissections of community dynamics are not always necessary to make reasonably good predictions of community function.

For all six datasets, we employed the leave-one-out cross-validation approach to the models fit to second-order terms. Impressively, we find that the regressions perform very well in predicting the unseen communities (Fig. 2.1). The success across these datasets shows the

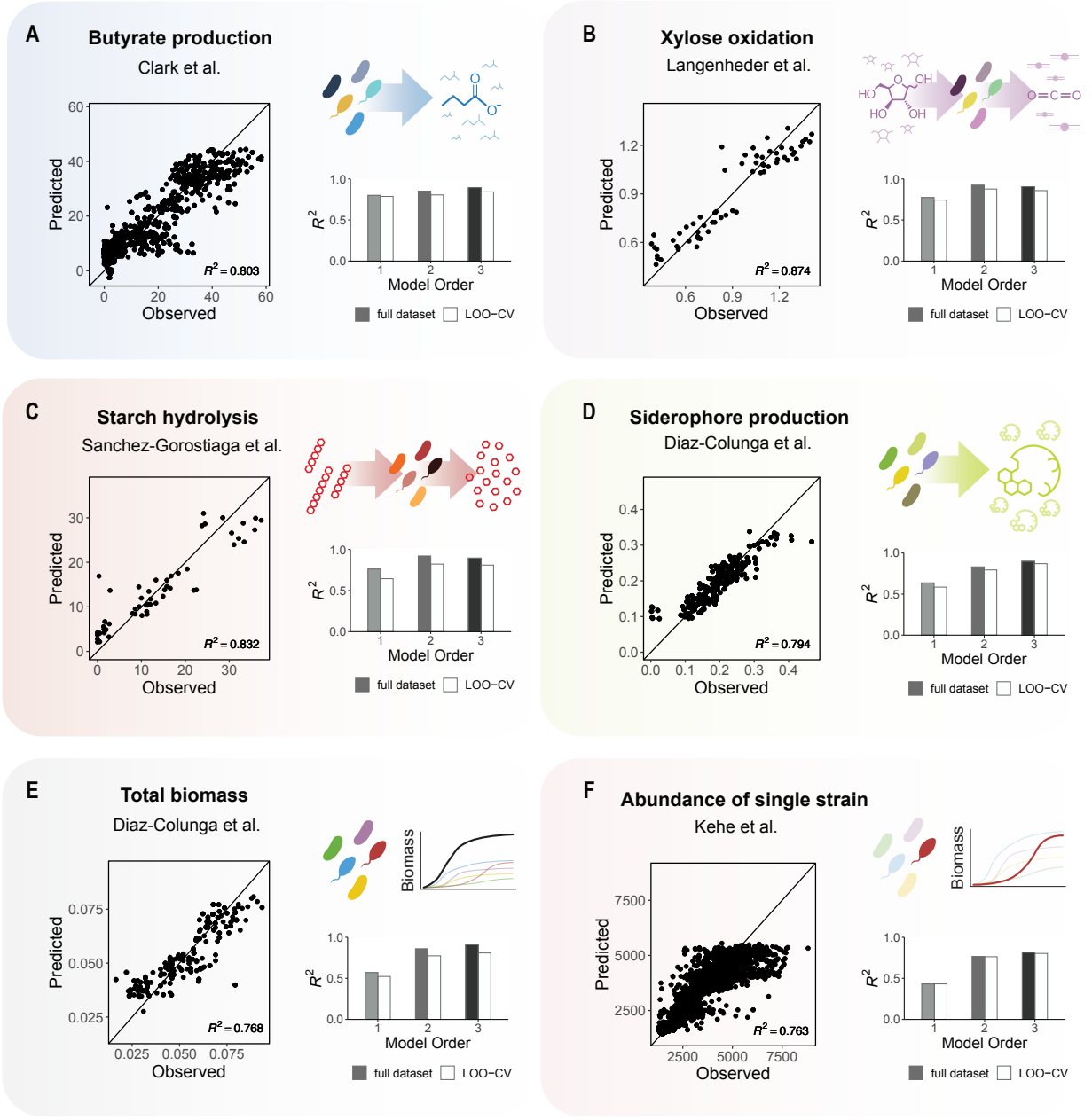


Figure 2.1: **Community function is predictable from species presence/absence in empirical datasets** For each dataset, regularized linear regressions were performed using models truncated at the first, second, and third order (Eqn. 2.1). Bar plots show the quality of fit (R^2) for each of these models, either using all experimental data or using a systematic leave-one-out cross validation approach (labeled "LOO-CV"). Scatter plots show out-of-sample prediction values for the second-order regression fits obtained via the leave-one-out cross validation procedure. Analyses are shown for datasets by (A) Clark et al. [7], (B) Langenheder et al, [60], (C) Sanchez-Gorostiaga et al. [21], (D) Diaz-Colunga et al. [59], (E) data from the Sanchez lab, and (F) Kehe et al. [61]

power of functional landscapes in serving as predictive tools of community function when relying on strain presence/absence alone.

2.2.3 Epistasis is Rare in Many Functional Landscapes

We demonstrated that regressions based on species presence/absence can predict microbial community functions with surprising accuracy. In particular, simple models, containing only additive and pairwise epistatic terms, explained the vast majority of the variation in the data. This framework represents a strategy for approximating the empirical community-function landscapes for these datasets. We wanted to gain intuition for why these regressions appeared to be so successful. To do this, we sought to quantify the ruggedness of community-function landscapes.

In an evolutionary context, the ruggedness of a fitness landscape dictates the number of local fitness optima and has important implications for the predictability of evolution [51]. In a community context, very rugged landscapes are expected to be much harder to approximate globally using regression methods such as those used here, simply because ruggedness arises from a substantial number of strong high-order epistatic terms. High-order terms are challenging to learn statistically due to the explosive combinatorial increase in the number of terms as model order increases. Thus, we expect that rugged landscapes will be difficult to learn statistically, while non-rugged ones will be straightforward to approximate using low-order models.

We quantified the ruggedness of our landscapes via the F_1 statistic. The F_1 statistic is a metric that quantifies the relative contribution of additive versus epistatic terms to the total variation captured by the landscape [57], and is quantified as such:

$$F_1 = \frac{\sum \beta_i^2}{(\sum \beta_i^2 + \sum \beta_{i,j}^2) + ...} \tag{2.2}$$

where the ellipses denote higher-order epistatic terms, if present. Note that if $\beta_{i,j} = 0$ for

all i, j (as well as all higher-order epistatic terms), then $F_1 = 1$ indicating a purely additive landscape, but as $\sum \gamma_{i,j}^2 + ... >> \sum \beta_i^2$, $F_1 \to 0$, indicating a purely rugged/epistatic landscape.

Calculating the F_1 statistic for all six regressions truncated at third-order terms shows that the landscapes are all mostly dominated by additive terms ($F_1 > 0.7$ in all cases, Tbl. 2.1). This result indicates that epistasis is rare in these landscapes, and as such the underlying landscapes are relatively smooth.

Dataset	**Calculated F_1 Statistic**
Butyrate Production	0.733
Growth on Xylose	0.934
Starch Hydrolysis	0.911
Siderophore Production	0.763
Total Biomass	0.706
Abundance of Single Strain	0.710

Table 2.1: **Calculated F_1 Statistics for Six Inferred Landscapes** truncated at third order terms

The F_1 statistic can be restructured to quantify the extent of higher-order interaction coefficients of a specific order: F_2 denotes the extent of second-order epistatic terms in the landscape, F_3 denotes the extent of third-order epistatic terms, etc. The unique mathematical properties utilizing a Fourier expansion basis for our regressions (i.e. $x_i = \pm 1$) enable each F statistic to be exactly the amplitudes of a Fourier amplitude spectrum of a given order (see Section 2.4.3). It can be shown that each amplitude is a reflection of the total variation of the landscape that is captured by terms of that order [56, 58]. Thus, we sought to compute a Fourier amplitude spectrum for each dataset to more precisely understand the extent of epistasis in these datasets.

We started with the dataset from Langenheder et al. [60], in which growth on xylose was measured for all $2^6 - 1 = 63$ species combinations that can be formed from a 6-species pool. This combinatorially complete dataset allowed us to compute the coefficients of the exact full-order empirical landscape, which is a model of the form Eqn. 2.1 that includes epistatic terms

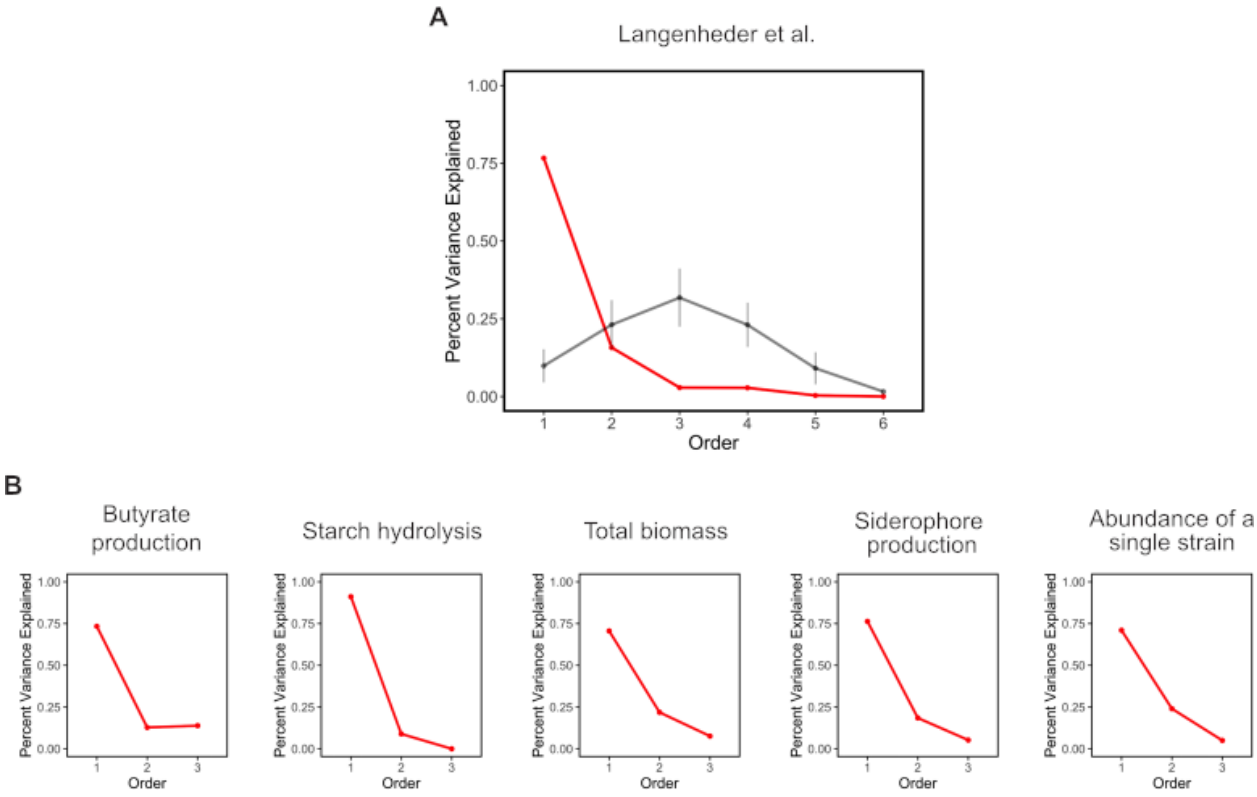


Figure 2.2: **Empirical community-function landscapes are not rugged** (**A**) Fourier amplitude spectrum obtained from the combinatorially complete landscape from Langenheder et al. [60]. Normalized amplitude values correspond to the fraction of total landscape variance that is captured by terms at each order. The empirical amplitude spectrum (red) demonstrates that landscape variance is primarily explained by first-order terms, with higher-order terms explaining decreasing fractions of the variance. This denotes a low degree of ruggedness in the empirical landscape. Black traces show spectra obtained from randomized landscapes (points and error bars indicate the mean and standard deviation, respectively, across 100 randomizations). Unlike the empirical landscapes, the randomized versions are rugged: first-order terms explain a relatively small fraction of total variance. (**B**) For each additional dataset that is not combinatorially complete, the fitted coefficients of the regularized third-order linear regression are used to infer the normalized amplitude spectrum at first through third order. As in panel **A**, first-order terms explain more variance in the landscape than terms at second and third orders, indicating a lack of ruggedness.

of all possible orders (i.e., up to sixth order). Computing the Fourier amplitude spectrum, shown in Fig. 2.2A (red line), indicates that $\sim 78\%$ (F_1) and $\sim 16\%$ (F_2) of the variance in the landscape is explained by first and second-order terms, respectively, leaving $\sim 6\%$ of variance remaining for higher-order terms. In other words, this exact community-function landscape displays a low degree of ruggedness, as it is dominated by low-order terms, and is largely free of consequential higher-order terms.

To assess the statistical significance of this result, we performed a randomization test by computationally shuffling the assignments between function measurements and community compositions. For 100 such randomizations, we inferred the new landscape and computed the Fourier amplitude spectra. The results are plotted in Fig. 2.2A (black line). For the randomized landscapes, we found that terms of third-order were most important, and additive terms alone captured only 10% of the variance; the peak at third order arises from the fact that there are combinatorially more terms possible at this order than any other. We concluded that a lack of ruggedness in the true landscape is not spurious but rather a distinctly non-random structural feature of this landscape.

We performed similar analyses for the remaining five datasets to estimate the relative importance of coefficients of different orders. Although these datasets are not combinatorially complete, and therefore do not permit the inference of the exact amplitude spectra, we inferred a truncated amplitude spectra via third-order regression. We found that in each case, the dominant coefficients are additive, with pairwise coefficients typically the next most-important (Fig. 2.2B). The large fraction of variance explained by additive and pairwise coefficients across these cases again indicates a low degree of ruggedness, and are in concordance with the calculated F_1 statistics (Tbl. 2.1).

We conclude that, across a diverse range of microbial community functions, empirical landscapes possess a low degree of ruggedness, corresponding to the dominance of low-order (i.e., additive and pairwise epistatic) terms, and a notable absence of higher-order

epistatic terms. This enables low-order statistical models to faithfully parameterize empirical landscapes and thereby accurately predict community functions.

2.2.4 Ecological Models Indicate When Landscapes Can Be Learned

We have demonstrated multiple empirical examples where the functional landscape of a microbial community proves to be non-rugged, allowing low-order statistical models to predict the function of interest. To understand the expected generality of this observation beyond the six datasets considered here, we turned to ecological models, generating large ensembles of random communities for which the functional landscapes can be evaluated *in-silico*. Our goal was not to fit models to the empirical examples above, but to probe the conditions under which a lack of ruggedness is expected to be rare or common, and to identify the ecological scenarios under which low-order statistical inference is expected to fail.

We first sought to interrogate synthetically-generated community-function landscapes using the generalized Lotka-Volterra (gLV) model, with total abundance (biomass) serving as the function of interest. The gLV model has many variants that differ in assumptions regarding the structure of randomly-generated species interactions, but as recently argued by Barbier *et al.*, a large class of such variants can be reduced to a four-parameter "reference model" [64]. We adopted this reference model for our analyses (see Section 2.4.4), focusing on a sweep of the two key parameters describing interaction strength (μ) and interaction variability (σ).

For each combination of μ and σ, we performed 10 trials of generating a random pool of $N = 10$ species (a pool small enough to evaluate the combinatorially complete set of all possible communities), and computed the exact community-function landscape (mimicking our procedure for the Langenheder *et al.* data in Fig. 2.2A). The fraction of variance explained by the first two orders of terms, equivalent to the predictive power (R^2) of the second-order approximation of the landscape, is shown in Fig. 2.3A. These values are averaged over the

10 trials at each point in the μ-σ plane. Note that as interaction variability σ becomes large, the Lotka-Volterra model becomes unstable, causing species abundances to diverge [65]. In our simulations, any points in μ-σ plane that encountered divergences in more than 5 trials are indicated in gray (Fig. 2.3A). Additional details about these simulations are described in Section 2.4.4.

We found that across the entire range of parameters for which the dynamics of the model are stable, second-order regression provides excellent fits ($R^2 \gtrsim 0.9$) to gLV community-function landscapes (Fig. 2.3A; the non-gray region is all yellow). In the reference model, positive μ corresponds to interactions that are competitive on average, while larger values of σ correspond to a greater degree of ecological heterogeneity. High quality fits across a broad sweep of the μ-σ parameter space provides strong evidence that our findings from six empirical datasets applies to a wide range of ecological scenarios.

2.2.5 Rugged Landscapes are More Difficult to Infer

Despite the evident simplicity of community-function landscapes in both empirical datasets and in the gLV model, it is possible to conceive of scenarios for which a low-order approximation will be a poor one. A simple case in point is the example of a nutrient X_1 that is broken down through a chain of L reactions following a linear pathway $X_1 \xrightarrow{n_1} X_2 \xrightarrow{n_2} \ldots \xrightarrow{n_L} X_{L+1}$, where each reaction is performed by a specialized community member n_i. If X_1 is the only nutrient supplied to the consortium, and the function of interest is the concentration of the end product X_{L+1}, then the concentration of X_{L+1} will be non-zero if and only if all species $n_1, \ldots n_L$ are present. Mathematically, if presence or absence were denoted as $x_i = 0$ or 1, respectively, such a landscape would be described by a single Lth-order term: $y = \beta_{1\ldots L}\, x_1 x_2 \ldots x_L$. Under the convention used here, where presence or absence is denoted by $x_i = 1$ or -1, respectively, this corresponds to interactions of all orders being equally strong. In either case, it is clear that no low-order approximation could adequately capture

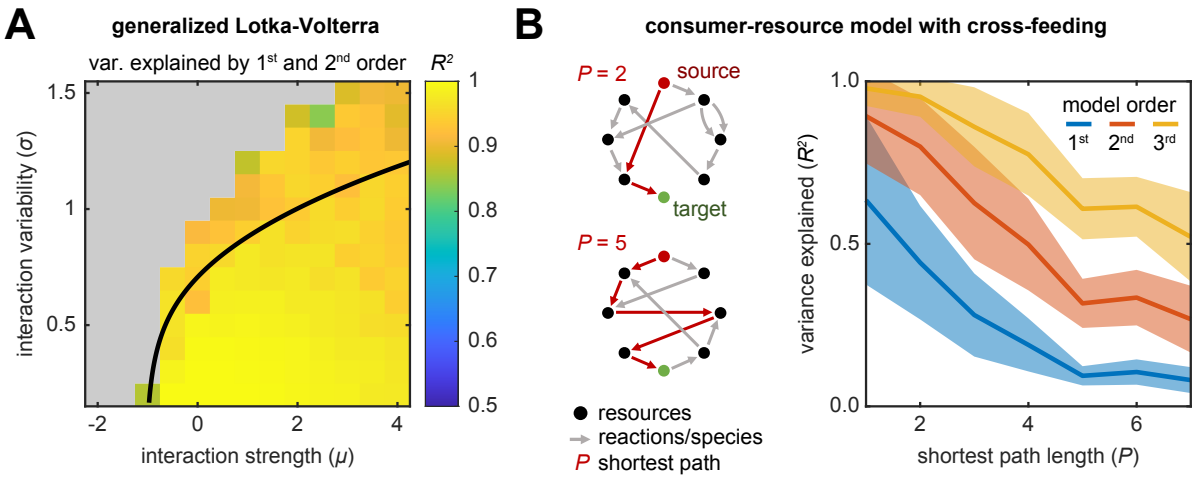


Figure 2.3: **Ecological models indicate both optimism and caution for inferring community-function landscapes.** (**A**) The generalized Lotka-Volterra (gLV) model was used to generate combinatorially-complete synthetic community-function landscapes, taking total abundance as the function of interest. These landscapes were then fit to Eqn. 2.1 exactly, with the variance explained by combined first and second order terms shown (R^2). Landscapes were generated using an $N = 10$ species pool, and two parameters controlling the structure of randomly-drawn interactions were varied: an interaction-strength parameter μ, and an interaction variability parameter σ. 10 landscape trials were generated over a grid of points in the μ-σ plane, and R^2 values averaged over these trials are shown. Note that as interaction variability σ becomes large, the model becomes unstable, causing species abundances to diverge. In our simulations, any points in μ-σ plane that encountered divergences in more than 5 landscape trials are indicated in grey. The black curve shows an analytical prediction (computed in the $N \to \infty$ limit [65]) for the stability boundary of the gLV model, beyond which species abundances tend to diverge. (**B**) A consumer-resource model (CRM) was used to generate synthetic community-function landscapes with randomly-generated cross-feeding networks (schematic). One resource was supplied externally ("source"), while the function of the community was defined to be the concentration of a different resource ("target"). These combinatorially-complete landscapes were fit exactly, with the variance explained by models truncated at first, second, and third-order shown (R^2). Cross-feeding networks, comprising $N = 10$ species and $L = 8$ resources, were generated in order to vary the length of the shortest path (P) connecting the source resource to the target. The variance explained by first through third-order models are shown as a function of P.

this landscape.

To build on the intuition from this thought experiment, we used a consumer-resource model (CRM) to generate synthetic landscapes with $N = 10$ species competing for $L = 8$ nutrients. Only one resource was supplied externally ("source", Fig. 2.3B, red dot), while the function of the community was defined to be the concentration of one of the other resources ("target", Fig. 2.3B, green dot). In each trial, we constructed a random cross-feeding network for a pool of $N = 10$ species, with each species capable of converting one randomly chosen resource into another. We then computed the complete functional landscape, as above for the gLV, by simulating all possible subsets of this 10-species pool. Based on the intuition above, we expected the predictive power of low-order approximations to correlate with the length of the shortest path (P) connecting the source resource to the target (see schematic, Fig. 2.3B). As in the thought experiment, cross-feeding networks with long paths from source to target (i.e., large P) are expected to have rugged landscapes that would be poorly characterized by a low-order model. Simulation details are described in Section 2.4.5.

The predictive power (R^2) of low-order models fit to these synthetic CRM landscapes is shown in Fig. 2.3B as a function of the shortest chain length P. We observed that, while successful in cross-feeding networks with small P, low-order models were increasingly challenged as P increases, with R^2 dropping below 0.5 for second-order models beyond $P = 4$. Increasing model order from additive-only to third-order substantially improved predictive power, particularly at large P, consistent with the idea that increasing P increases the prevalence of higher-order terms. These results illustrate that low-order approximations of community-function landscapes may fail to make accurate predictions in ecological scenarios with long chains of trophic dependencies or other situations when function is strongly contingent on the simultaneous presence of multiple species.

To understand the generality of failure modes of inferring community function landscapes, we explored the inferrability of landscapes with various degrees of ruggedness. To this effect,

we considered a system of $N = 10$ strains, and generated 60 *in-silico* communities from these ten strains. We generated a family of regressions up to second-order terms and controlled the true F_1 statistic of the model to be between 0 and 1, in increments of 0.1, to represent landscapes of varying ruggedness. Each landscape included a noise term η to represent imperfect fits (see Section 2.4.6). For each level of the F_1 statistic, 100 trials of data collection were simulated, and the regressions were fit to the data with LASSO regularization in the same method described previously.

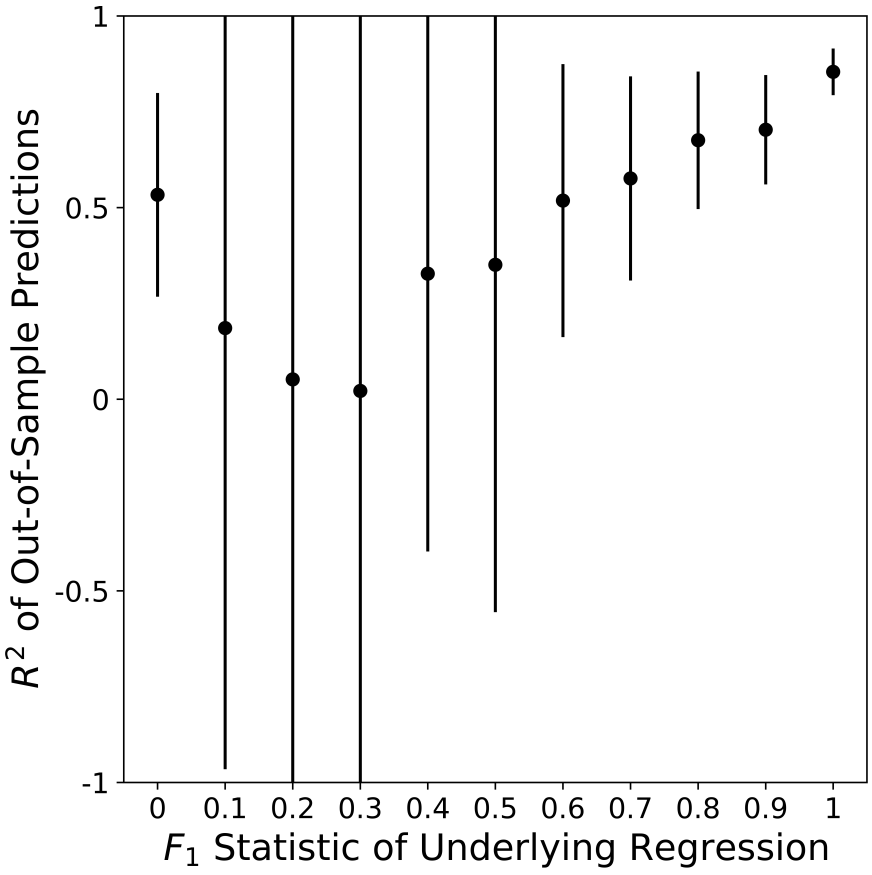


Figure 2.4: **Simulations Show Rugged Landscapes are Harder to Infer**. R^2 of leave-one-out cross-validation predictions as a function of the underlying ruggedness of the landscape. For a system of 60 communities created from 10 strain, 100 trials were generated for each level of ruggedness of the underlying landscape. Average R^2 and the standard deviation are shown for each underlying F_1 statistic.

Fig. 2.4 shows the R^2 of regression fits as a function of the true F_1 of the underlying

regression. On average, model fits become more unreliable as the ruggedness increases ($F_1 < 0.6$) but become reliable with a fully epistatic model. This result further confirms that when an underlying landscape is governed by epistasis, a low-order regression will have difficulty inferring the functional landscape

2.3 Discussion

The key result of this chapter is the demonstration that functional properties of microbial communities can be described by simple statistical models fed only information on strain presence and absence, without requiring measurement of microscopic dynamic properties. By developing the functional landscapes as analogous to the genetic fitness landscapes, we showed that regressions can quantitatively describe empirical landscapes for a wide range of community functions. The landscapes for all six functions were smooth, thereby allowing most of the variation in the datasets to be captured by additive and pairwise epistatic terms.

Our simulations of generalized Lotka-Volterra and consumer-resource models demonstrated that we can expect low-order approximations of landscapes to work well across a range of ecological contexts. We identified a clear exception, however, in functions that are strongly contingent on the simultaneous presence of multiple highly specialized community members [66], or are otherwise rampant with epistatic interactions [12, 21]. For example, the metabolism of a toxic compound could require the presence of a large number of strains to enable various steps of the degradation pathway while also enabling toxicity resistance. Both such cases require extensive interactions between strains [31, 67] and will likely lead to high ruggedness in the underlying landscape. In Chapter 3, we will explore such a landscape and identify modifications to the regression approach to enable the functional landscape to be inferred.

The most immediate downstream impact of the results of this chapter is the possibility of using a functional landscape regression approach for rationally designing communities

with desirable functional properties. The nature of learning community function through strain presence/absence alone overcomes the issues with bottom-up and top-down design approaches discussed in Chapter 1. In addition, because the functional landscapes can be approximated by sampling only a subset of all possible species combinations, it is conceivable that even large synthetic consortia can be constructed and optimized given a small number of measurements. In Chapters 3 and 4, we will observe two separate functional landscapes that are inferrable via regressions and can be utilized for the rational design of functional communities.

2.4 Methods

2.4.1 Datasets

Datasets were compiled from six experimental efforts to measure various community functions in defined synthetic microbial consortia. Clark et al. [7] measured the production of the short-chain fatty acid butyrate; Langenheder et al. [60] measured a combination of biomass and redox activity on the monosaccharide xylose; Sanchez-Gorostiaga et al. [21] measured the breakdown of the polysaccharide starch; Diaz-Colunga et al. [59] measured the total production of iron-scavenging siderophores (Fig. 2.1). In addition, we considered biomass-related community functions: work by the Sanchez lab measured total community biomass, and Kehe et al [61] measured the abundance of a single target species. Details about the size and species pools for each dataset are given in Table 2.2

The dataset by Langenheder et al. [60] was generated by measuring the activity of synthetic communities with xylose provided as the sole carbon source. Metabolic activity was measured via the absorbance of a redox dye (tetrazolium violet) at 600 nm over multiple time points. It should be noted that, because cell scattering at 600 nm likely also contributed to absorbance, the functional values collected in this study reflect a combination of both redox

activity and biomass growth. While this detail complicates the mechanistic interpretation of the data, it does not present a problem for our analysis, as the data still represent a complex functional property for the regression approach to predict. Here, the community-function landscape we sought to approximate was constructed from the functional values collected at the endpoint of the experiment (48 h).

The dataset by Sanchez-Gorostiaga et al. [21] included 53 community configurations out of the possible 63. We note that, as we confirmed with the authors of the original study, the potentially ambiguous phrasing in the original manuscript ("every combination of six amylolytic soil bacteria") referred to every pairwise combination being inlcuded in the list, not that the dataset was combinatorially complete.

Total community biomass data collected by Diaz-Colunga et al. were not included in the original manuscript [59]. Detailed experimental protocols for isolation, culturing, and community assembly are given in Ref. [59], with the biomass dataset differing from these methodological details in the following ways: (1) the eight isolates used in the biomass dataset were distinct from the siderophore production isolates (though isolated using the same methodology), (2) the medium used in biomass growth experiments lacks a trace mineral supplement, (3) biomass was quantified by measuring the OD at 600 nm using 100 μL of endpoint cultures in an AccuSkan FC plate reader (Fisher Scientific), and (4) monoculture measurements for four strains were omitted, resulting in a dataset of 160 unique configurations (instead of 164, as should otherwise result from the methodology in Ref. [59]). Though isolates used in the biomass dataset were not sequenced or taxonomically identified, colonies either possessed distinct morphologies, possessed distinct color profiles when grown on chromogenic agar plates (CHROMagar Mastitis GN), and/or were isolated from separate environmental samples, and were therefore likely to be genomically distinct.

In cases where datasets contained experimental replicates, the mean over replicates was taken to obtain a single community-function value for each unique species combination.

Function	# species	# communities	Ref.
Butyrate production	25	577	Clark *et al.* [68]
Growth on xylose	6	63	Langenheder *et al.* [60]
Starch hydrolysis	6	53	Sanchez-Gorostiaga *et al.* [69]
Siderophore production	8	225	Diaz-Colunga *et al.* [59]
Total biomass	8	160	Diaz-Colunga *et al.* [59]
Biomass of a single strain	14	4239	Kehe *et al.* [61]

Table 2.2: Information about the six datasets used in this chapter

2.4.2 Regressions

Community-function landscapes were approximated for empirical datasets by fitting equations of the form Eqn. 2.1 truncated at first, second, and third order via LASSO regularization [62]. 10-fold cross-validation was used to estimate the regularization hyperparameter. All models were fit using the package glmnet in R version 4.1.2. For all datasets considered here, the computational demands of fitting our models were modest, and all model fitting and analysis was performed using a personal computer.

To fit the data, two strategies were employed. First, all available data points were used to fit landscape models. The coefficients of determination (R^2) for these fits are shown in gray bars in Fig. 2.1 and predicted versus observed values are shown in Fig. S.2.3. Second, to obtain an estimate of out-of-sample model accuracy, a leave-one-out (LOO) procedure was employed. Individual data points corresponding to each distinct experimental community were systematically left out of sample, and models were then fit to all remaining data points using 10-fold cross-validation. The observed versus predicted values of left-out points estimated via this approach for a second-order model are shown in the scatter plots of Fig. 2.1. The prediction quality (R^2) for left-out points are shown in white bars in Fig. 2.1.

2.4.3 Calculating Epistasis

The convention of $x_i = \pm 1$ allows Eqn. 2.1 to correspond to the Fourier expansion of the fitness landscape [56, 58]. This mathematical feature enables the simplification the Fourier

amplitude equation [57, 58] of the coefficients of any order p as:

$$A_p = \sum_i \beta_i^2 \tag{2.3}$$

.

For example, for coefficients of the first order (i.e. additive coefficients), the Fourier amplitude is simply $A_1 = \sum_i \beta_i^2$.

It can be shown that sum of amplitudes across all orders is equal to the total variance of the landscape. This permits the calculation of the fraction of total variance explained by additive terms (V_1) as

$$V_1 = \frac{A_1}{\sum_p A_p} \tag{2.4}$$

which is equivalent to the equation for the F_1 statistic (Eqn. 2.2).

2.4.4 Simulations of the generalized Lotka-Volterra model

Community-function landscapes were synthetically generated using the generalized Lotka-Volterra (gLV) model, with total abundance serving as the community function of interest. A four-parameter "reference model" formulated by Barbier *et al.* [70] was used to explore important dimensions of the gLV parameter space. The model can be written as follows:

$$\dot{N}_i = \frac{r_i}{K_i} N_i \left(K_i - N_i - \sum_j \alpha_{ij} N_j \right), \tag{2.5}$$

where, for species i in a pool of N total species, N_i is the abundance, r_i is the intrinsic growth rate, and K_i is the carrying capacity. The parameters α_{ij} are interaction coefficients between species i and j.

Nonzero equilibria of Eqn. 2.5 are determined by the values of K_i and α_{ij}. For each

synthetic landscape, these parameters were randomly generated. The carrying capacities K_i were independently drawn from a normal distribution with mean 1 and variance ζ^2, while the interaction coefficients were drawn from a normal distribution with mean μ/N and variance σ^2/N. A broad, ecologically-relevant region of the interaction parameter space was explored in Fig. 2.3A ($\mu \in [-2, 4]$ and $\sigma \in [0, 1.5]$). The standard deviation of carrying capacities was fixed at $\zeta = 0.3$. In principle, interaction asymmetry can be encoded via a parameter $\gamma = \text{corr}(\alpha_{ij}, \alpha_{ji})$, but interactions here were set to be symmetric by taking $\gamma = 1$. All r_i were fixed to 1.

Landscapes were generated over a grid of points in μ-σ space. Each landscape was generated through the following steps:

1. A set of carrying capacity parameters K_i and interaction parameters α_{ij} were drawn for a pool of $N = 10$ species as described above.

2. Eqn. 2.5 was simulated to equilibrium for all species combinations.

3. Total endpoint abundances $\sum_i N_i(\infty)$ were computed for each simulation.

4. Exact, full-order landscapes (Eqn. 2.1) were computed using total endpoint abundance as the community function.

Initial species abundances were drawn from an exponential distribution with mean 0.1. Numerical integration was performed using ode15s (*MATLAB*). At each point μ-σ space, 10 trials of landscapes were generated. Note that for larger values of σ, species abundances in Eqn. 2.5 are more likely to diverge. Parameter combinations for which divergences were encountered in more than half of trials are indicated by grey values in Fig. 2.3A.

2.4.5 Simulations of the consumer-resource model with random cross-feeding networks

Community-function landscapes were synthetically generated using a consumer-resource model (CRM), taking the equilibrium concentration of a terminal waste product as the function of interest. The CRM is given as follows:

$$\begin{aligned} \dot{N}_i &= N_i(r_i - m_i), \\ r_i &= \sum_\alpha C_{i\alpha} R_\alpha, \\ \dot{R}_\beta &= \frac{K_\beta - R_\beta}{\tau_\beta} - \sum_i C_{i\beta} R_\beta N_i + \sum_i \gamma r_i N_i D_{i\beta}. \end{aligned} \tag{2.6}$$

Here, N_i, r_i, and m_i are the abundance, total resource uptake, and maintenance costs, respectively, of species i in a pool of N total species. R_α is the concentration of resource α. The matrices $C_{i\alpha}$ and $D_{i\beta}$ describe which resources a species consumes and secretes, respectively, with secretions assumed proportional to the metabolic uptake r_i. An efficiency factor $\gamma < 1$ ensures that energy cannot be gained, but only lost. Here γ was set to 0.5. The resource carrying capacities K_β, decay rates τ_β and species maintenance costs m_i are all set to 1 for simplicity. The total number of resources was fixed to $L + 1$, with resource $L + 1$ representing a terminal waste product that no species can consume.

Random cross-feeding networks were generated in order to explore how chains of trophic dependencies impact the ruggedness of community-function landscapes. To do this, it was assumed that each species i can consume and secrete exactly one resource, denoted in_i and out_i, respectively. These were selected through the following steps, performed independently for each species:

1. The identity of the consumed resource in_i was chosen randomly between 1 and L with equal probability.

2. The secreted resource out_i can range from 1 to $L+1$, where resource $L+1$ is the terminal waste product; however it must be distinct from in_i, leaving L possible choices. With probability p, $\text{out}_i = \text{in}_i + 1$; otherwise out_i was drawn from any of the remaining values at random.

The parameter p thus allows the exploration of a range of network topologies, from long linear pathways with a high degree of trophic dependency (at $p \approx 1$) to random graphs (at $p \approx 1/L$).

After selecting in_i and out_i for all species, it was verified that resource $L+1$ is "reachable" through the network from resource 1, i.e., whether there exists at least one path from resource 1 to resource $L+1$. If this was not the case, the functional landscape (the concentration of resource $L+1$ for a given set of species) would be identically zero; such networks were discarded as invalid and the steps above were repeated until a valid circuit was obtained.

For Fig. 2.3B, random cross-feeding networks were generated by carrying out the steps above across a range of values of p, fixing $N = 10$ total species and $L = 8$ total resources. 20 valid random trials over 15 values of $p \in [1/L, 1)$ were generated in order to create 300 total random cross-feeding networks. The community-function landscapes for these networks were then computed by simulating Eqn. 2.6 to equilibrium for all species combinations (ode15s, MATLAB), taking the equilibrium concentration of resource $L+1$ as the function of interest, and fitting Eqn. 2.1 exactly. Initial species abundances were set to 0.1, and initial resource concentrations were set to $R_\beta(0) = K_\beta$.

2.4.6 Inferrability of Rugged Landscapes

Community function landscapes were synthetically generated from a system of $N = 10$ strains formed into 60 communities. x_i^c defines the presence or absence of strain i in community c. For each community c, we defined the corresponding functional output y_c using a second-order model of the same form as Eqn. 2.1:

$$y_c = \sum_{i=1}^{N} \beta_i x_i^{(c)} + \sum_{i=1}^{N} \sum_{j>i}^{N} \beta i, j x_i^{(c)} x_j^{(c)} + \eta_c \tag{2.7}$$

where y_c is the functional output of community c, β_i and $\beta i, j$ are additive and pairwise coefficients respectively, and $\eta_c \sim \mathcal{N}(0, \sigma^2)$ is independent Gaussian noise with variance σ^2. We set $\sigma^2 = 0.25$, which is on the order of each of the coefficients.

We generated a family of regressions and varied the underlying F_1 statistic to be between 0 and 1, in increments of 0.1. This was done by generating a single set of coefficients, and then normalizing them such that:

$$\begin{aligned} A &= F_1 \\ B &= 1 - F_1 \\ \beta_i &= \beta_i \times \sqrt{\frac{A}{\sum_i \beta_i^2}} \forall i \\ \beta_{ij} &= \beta_{ij} \times \sqrt{\frac{B}{\sum_{i,j} \beta_{ij}^2}} \forall ij \end{aligned} \tag{2.8}$$

Using the set of coefficients with the target F_1 statistic set, 100 iterations of y_c were generated, each representing an experimental measurement of a different function with the same level of ruggedness. A regression was fit to all c, y_c in the same way described above, and the R^2 of leave-one-out cross-validation predictions were calculated. Average R^2 over the 100 different experiments for each target F_1 were compared.

2.4.7 Data and Code Availability

Data analyzed for this chapter is either available from the original studies or in the following repository: https://github.com/abbyskwara2/regression_on_landscapes.

All code created to run the analyses in this chapter is available in the above repository.

2.5 Supplementary Discussion

2.5.1 Low Ruggedness in Landscapes is not an Artifact of LASSO Regularization

A key result of this chapter is that epistasis is rare in these community function landscapes, and this observation is key in enabling low-order regressions to accurately approximate the landscapes. To avoid overfitting, we employed LASSO regularization during the fitting process. LASSO regularization performs 'feature selection', in which coefficients deemed unimportant are driven to zero, resulting in a 'sparse solution' [62, 71, 72]. It is reasonable to fear, then, that the result of low epistasis may be a statistical artifact of LASSO regularization's feature selection.

To address this, we repeated the regressions on all six datasets, but utilized a Ridge regularization. Often thought of as the opposite of LASSO, Ridge regularizations perform 'feature shrinkage', which decreases the values of all coefficients while preserving as many of them as possible [71, 73]. We fit Ridge regularized regressions as described in Section 2.4.2 truncated to third-order terms, and calculated the F_1 statistics for each (Tbl. 2.3). We recall that the F_1 statistic is also a measure the the percent of variation captured by additive terms within the model, which can be visualized by Fourier amplitude spectra (Fig. 2.5A).

Dataset	**Calculated F_1 Statistic**
Butyrate Production	0.035
Growth on Xylose	0.633
Starch Hydrolysis	0.883
Siderophore Production	0.945
Total Biomass	0.323
Abundance of Single Strain	0.900

Table 2.3: **Calculated F_1 Statistics for Six Inferred Landscapes with Ridge Regularized Regressions up to Third-Order Terms**

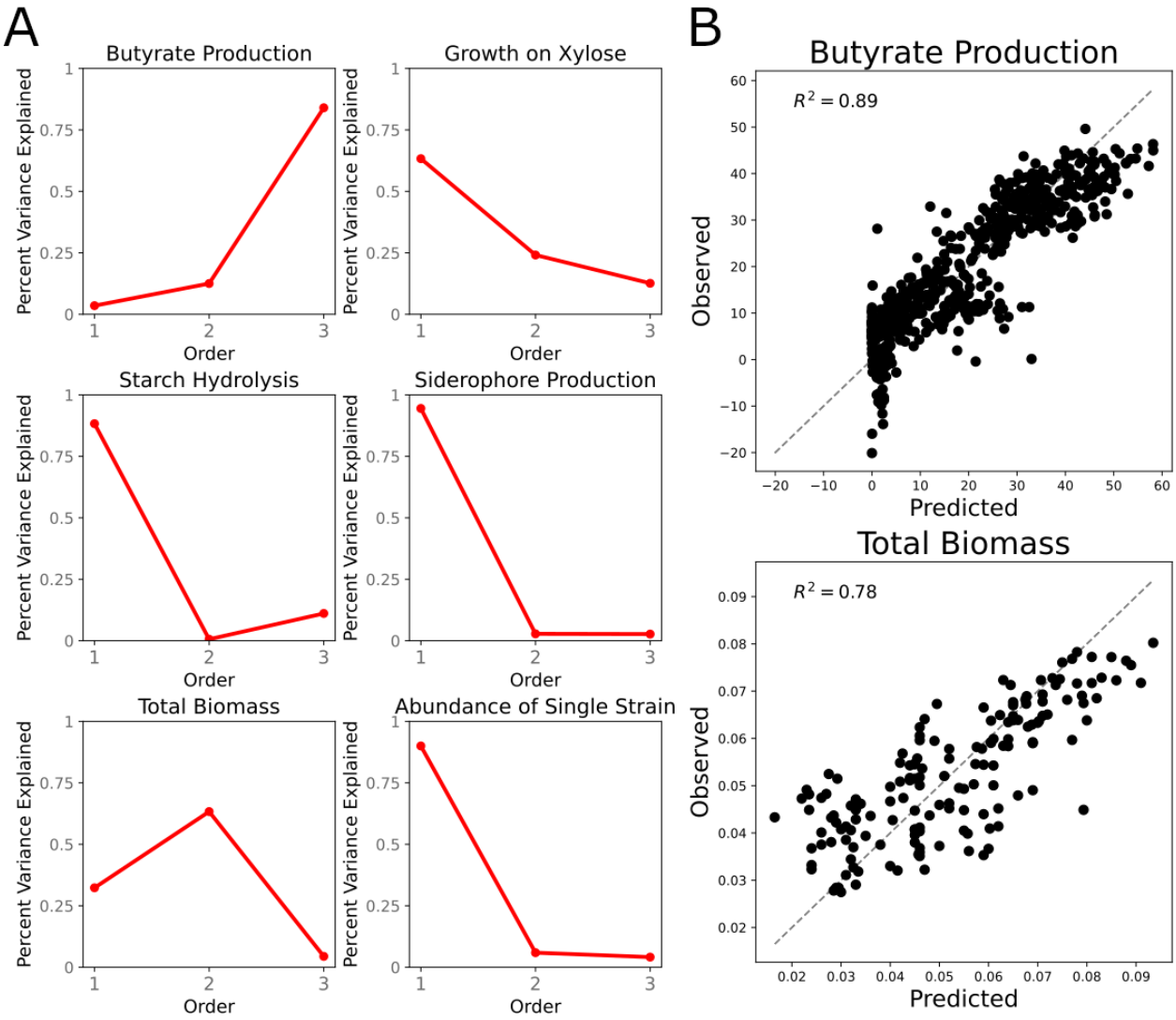


Figure 2.5: **Ridge Regularization Does Not Lead to Rampant Epistasis in Inferred Landscapes (A)** Fourier amplitude spectra for each of the six datasets when fit to a linear regression up to third-order terms with Ridge regularization. Spectra indicate low ruggedness for four functions. Butyrate production and total biomass show high epistasis. However, a regression only including additive terms for each of these two functions retains strong predictive power (**B**), indicating that epistasis is not the dominant driver of variation in those datasets.

Four of the functions - growth on xylose, starch hydrolysis, siderophore production, and single strain abundance - exhibit low epistasis, with majority of the variation captured by the additive terms ($F_1 > 0.6$). This indicates the true nature of the smoothness of the landscapes for these four functions.

Butyrate production and total biomass experienced increased epistasis through Ridge regularization ($F_1 < 0.33$), with butyrate production showing extreme ruggedness in its landscape ($F_3 = 0.84$). It appears that these two functions may be affected heavily by epistasis that was not captured in the regressions regularized by LASSO. However, if we ignore higher-order terms and refit the two respective datasets with models truncated at first-order terms, we find that a strong predictive power is conserved ($R^2 > 0.75$) (Fig. 2.5B). This result shows that despite the computed Fourier amplitude spectra, additive terms alone capture enough variation to enable accurate predictions, thus showing that the increased dominance of epistatic terms is an artifact of Ridge regularization's 'feature shrinkage', and that the smoothness of the landscapes inferred in Figures 2.1 and 2.2 is real.

2.5.2 Understanding the Impact of Undersampling on Regression Fits

One of the unique strengths of the framework presented in this chapter is the ability to infer a functional landscape from a small subsample of communities. For example, the landscape of butyrate production was inferred from 577 communities, which is much less than $2^{25} \approx 33$ million potential possible communities. However, we have no indication of the limit of undersampling that still enables regressions to infer the functional landscape.

To tackle this question, we considerd the same *in-silico* system considered in Section 2.4.6: a system of $N = 10$ strains, meaning a maximum of $2^{10} = 1024$ communities. We also considered the same family of regressions (Eqn. 2.7) with varying levels of ruggedness.

Instead of generating 60 random communities, we considered datasets consisting of 5, 10, 50, 100, 500, and 1000 randomly generated communities (to represent varying levels of

undersampling). For each dataset size, we sampled communities 100 times to represent 100 experimental procedures, and inferred the functional landscape for each regression in the family of regresions (as described in Section 2.4.6). The average predictive power (R^2) of the 100 randomizations is shown in Fig. 2.6. Note that in some cases, the standard deviation of predictive power was much greater than the average, indicating strong instability of the model and therefore unreliable landscape inference. These instances are depicted in gray in the heatmap.

Results show that, with as low as 100 communities sampled out of 1024 ($\sim 10\%$ of communities), the regressions are able to accurately infer the underlying landscape, regardless of ruggedness. With half the number of communities sampled, smooth landscapes are inferrable with some difficulty, while rugged landscapes cannot be reliably inferred at all. For many experiments that involve constructing synthetic microbial consortia, only a relatively small number of consortia need to be manually built in order to learn about the space of all possible consortia and their functional performances.

This result is surprising given the success with predicting butyrate production from only 577 communities (much less than 0.001% of all possible communities with 25 strains). Such a small sampling should be unable to lead our regressions to infer the functional landscape with any accuracy. However, we note that the methodology of Clark *et al* [7] included a design strategy for their consortia: beginning with co-cultures and using model predictions to inform the design of more complex communities; additionally, the community of all 25 strains and all 24-member communities were included to learn the effects of removing any given strain on butyrate production. This rational approach to design communities allowed for the targeted sampling of the landscape to enable capturing many successful communities, which can be learned by our regressions. This underpins the importance of implementing a rational methodology for the design of initial communities for the functional landscape to be properly inferred.

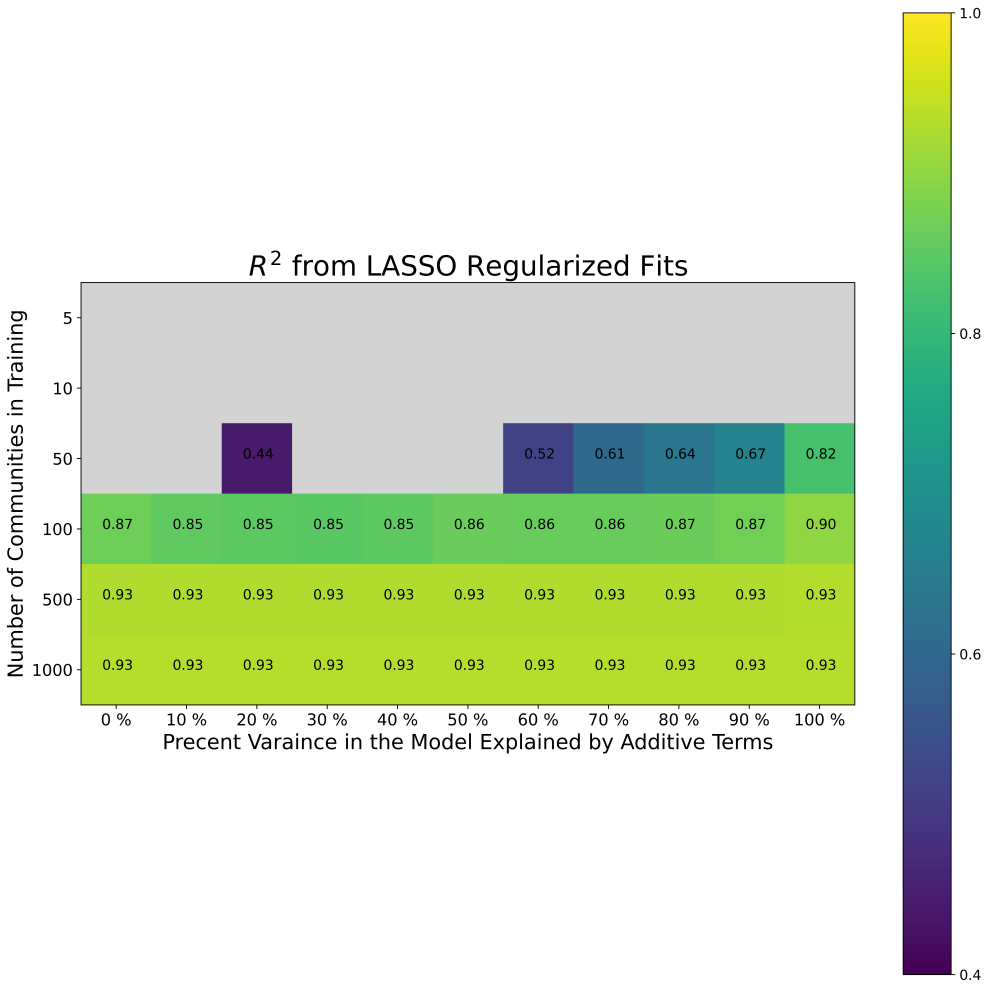


Figure 2.6: **Regressions Tolerate a Moderate Level of Undersampling**. Heatmap showing the average predictive power (R^2) of regressions as a function of the ruggedness of the underlying model (x-axis) and the number of communities used to fit the model (y-axis). R^2 shown is the average of 100 randomly generated datasets; grayed-out squares indicate highly unstable regression fits.

2.5.3 Supplementary Figures

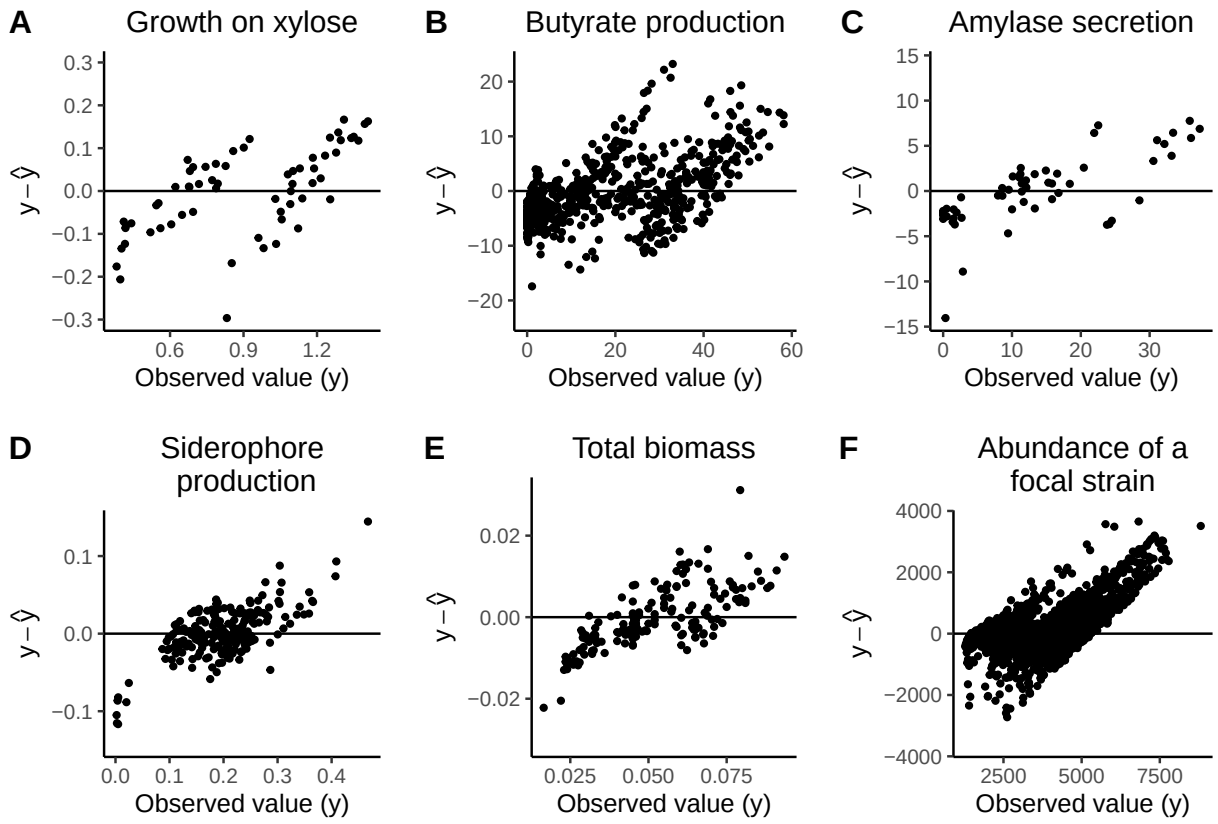


Figure S.2.1: **Residuals obtained from second-order regressions** for data from all six datasets (See Fig. 2.1)

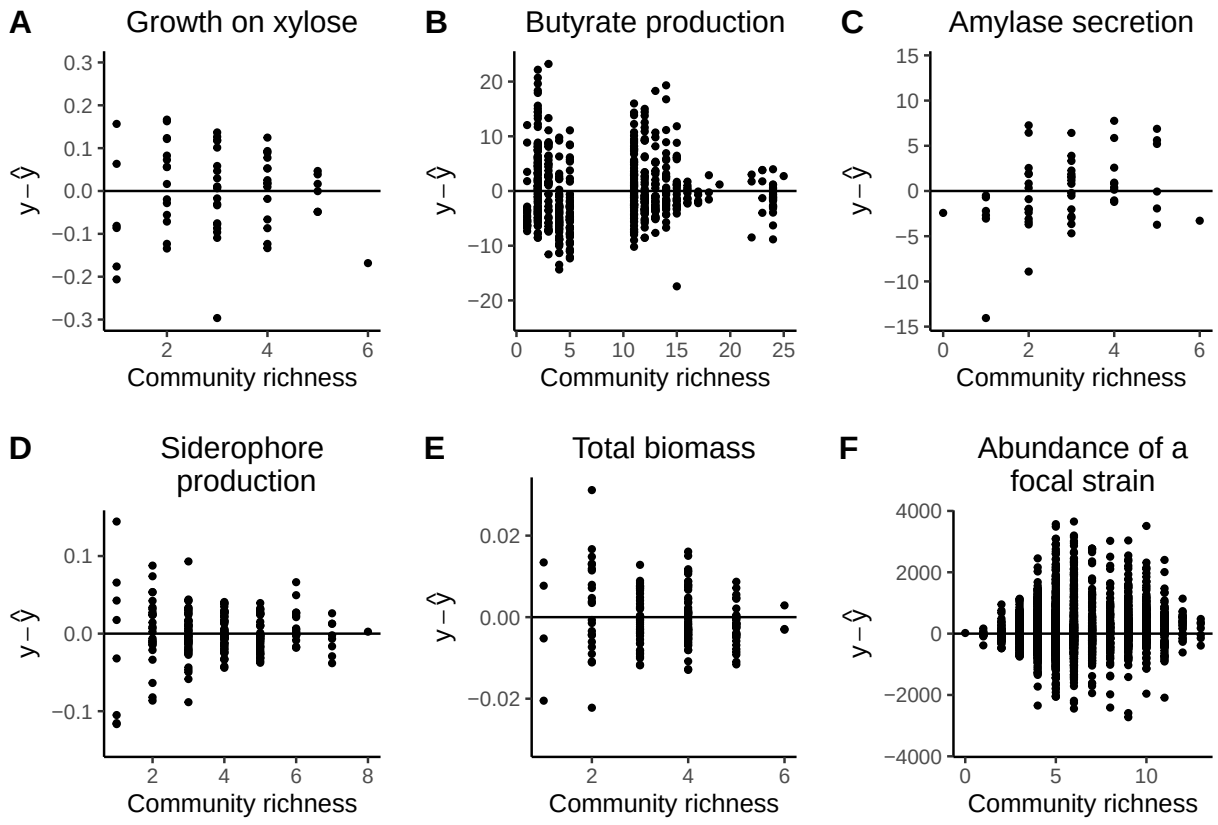


Figure S.2.2: **Model residuals obtained from second-order regressions as a function of community richness** for data from all six datasets

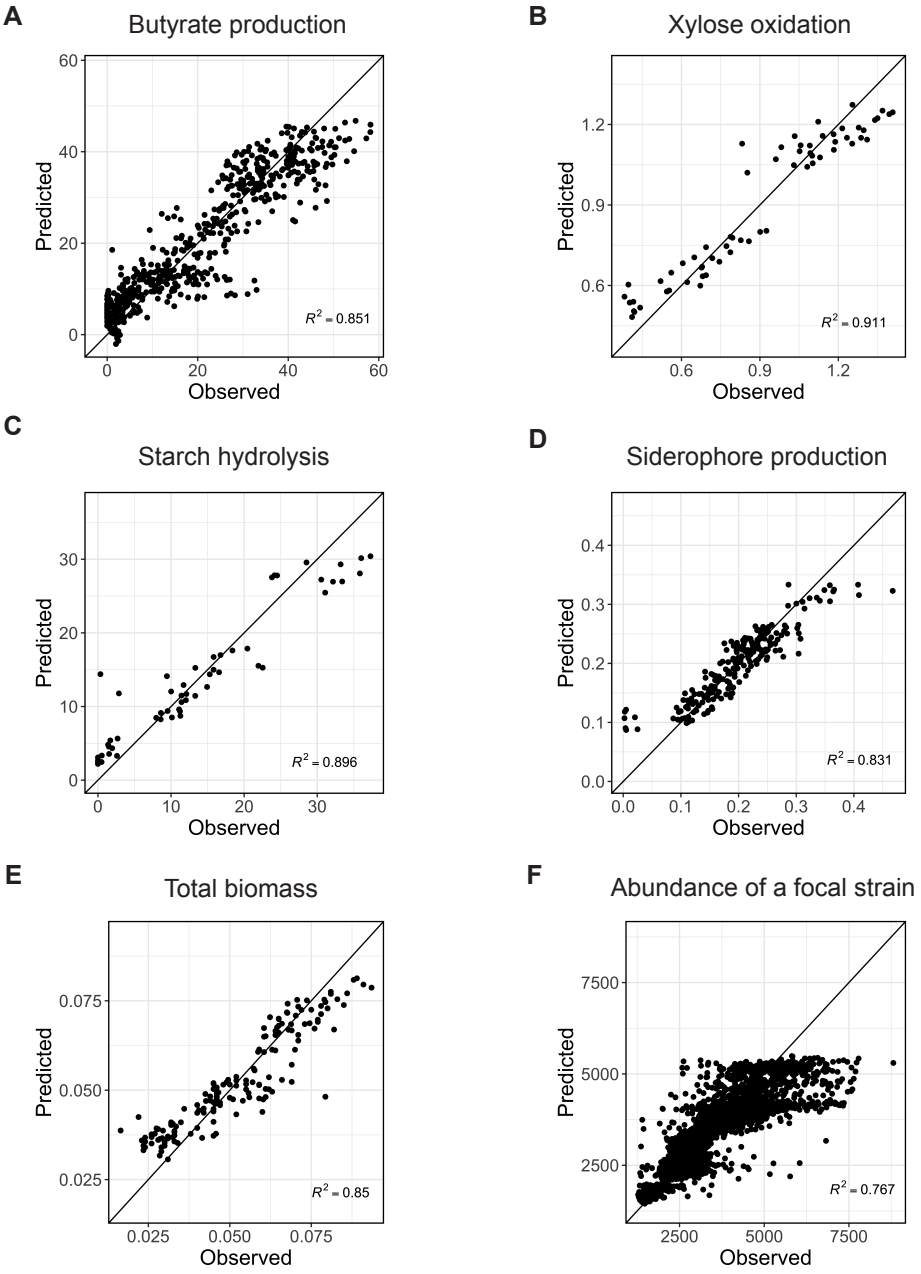


Figure S.2.3: **Model predictions from second-order regression fits with all data in-sample** for data from all six datasets

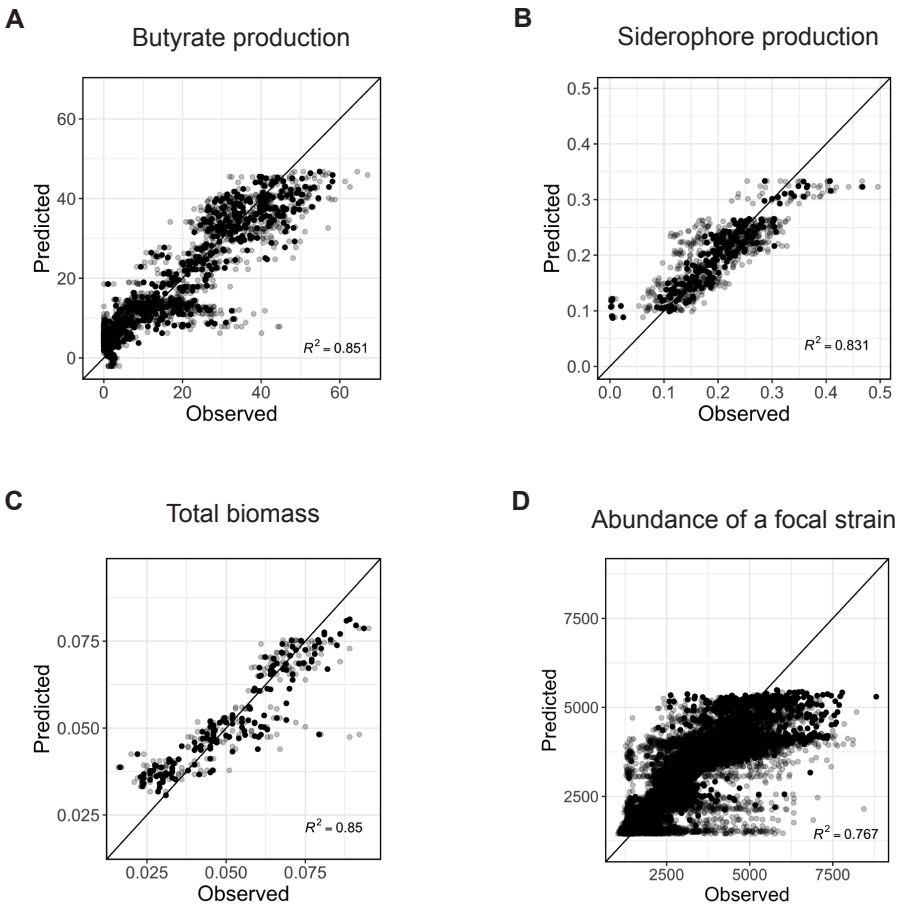


Figure S.2.4: **Model predictions from second-order regression fits with all data in-sample and experimental replicates included** Data from (A) Clark et al [7], (B) Diaz-Colunga et al [59], (C) data from the Sanchez lab [59], and (D) Kehe et al. [61]. Black points correspond to the mean observed value for each distinct experimental community, and are identical to the points shown in Fig. S.2.3. Gray points correspond to values for distinct experimental replicates. The datasets here include only datasets for which experimental replicates were measured

CHAPTER 3

EPISTASIS DRIVES TOXICITY RESISTANCE IN BIOREMEDIATION AND ENABLES RATIONAL DESIGN

3.1 Introduction

Bisphenol-A (BPA) is an organic molecule widely used industrially as a plastic hardening agent which recently emerged as a significant health hazard due to its carcinogenic and estrogenic proprerites. BPA levels detected in various soil, freshwater, and sewage samples often exceed safety thresholds set by regulatory bodies. As it naturally persists in these environments, scientists employed many strategies to attempt to remove BPA from these contaminated sites, including oxidative reactions, chemical reactions, and photocatalysis; however, many of these strategies suffer from limitations such as high costs, potential safety hazards, and inadequate efficiency.

Microbial bioremediation - the degradation of pollutants from natural environments - presents a promising alternative that is sustaianble, environmentally-friendly, and often affordable. Bioremediation is a function of many native natural microbiomes, although the wide functional diversity of natural microbiomes often prevent strong occupancy of the bioremediation niche [24, 32]. A range of toxins - including hydrocarbons and heavy metals - have been successfully remediated through both single-strain and community-level bioremediation *in-vitro* and *in situ*. However, many of the communal bioremediation attempts are limited due to a lack of rational design principles to engineer communities [17, 26, 45].

Many microbial strains can utilize BPA as a carbon source during aerobic respiration[31, 74, 75], providing the avenue for BPA bioremediation. Although BPA degradation by individual bacterial strains has been extensively studied [76, 77], many fail to achieve complete degradation [76] due to direct BPA toxicity and the accumulation of toxic intermediates [78–82]. Some studies suggest that communities can overcome the toxicity of BPA and its inter-

mediates [83], however, there are no design principles to building BPA degrading consortia. Part of the challenge of selecting strains is that we lack definitive knowledge of the genes responsible for every step of a BPA degradation pathway, as only a handful of steps have been characterized [31, 74, 75, 80, 84, 85]. Furthermore, there has been little study of the role of environmental factors in BPA bioremediation [76].

The challenges of BPA bioremediation present a useful application for evaluating predictive models for microbial community function. In particular, the lack of definitive mechanistic knowledge of BPA degradation along with the concentration-dependent toxicity of BPA introduces an environmental gradient that can potentially be climbed and exploited rationally. In this chapter, we utilize the functional landscape approach developed in Chapter 2 to design BPA-degrading communities across a range of increasing BPA concentrations, using soil-derived microbial communities as our model system. Inferred landscapes show that collective effects, which arise in landscapes through epistatic statistical interactions between taxa, increase with higher levels of the contaminant. However, despite the rising epistasis, the structure of the landscape changes smoothly as the BPA concentration rises, which enables landscape inference. We then use this landscape to design communities with desired rates of BPA degradation *in-vivo*. Finally, we show that our designed communities remediate BPA in contaminated soils. Together, our results show a successful attempt at describing a highly epistatic microbial function through a functional landscape approach which enables rational design of BPA-degrading communities both *in-vitro* and in soils. Furthermore, even in rugged functional landscapes, global patterns of community function remain relatively smooth, enabling rational community design in harsh regimes.

3.2 Results

3.2.1 A Library of Isolates that Enable Community Design for BPA Degradation

We set out to design synthetic communities of bacteria capable of degrading BPA. First, we assembled communities from a library of bacterial soil isolates available in our laboratory [42] (largely proteobacteria) and tested their ability to degrade BPA in minimal media. BPA degradation was assessed by sampling cultures in time and using a colorimetric assay to measure BPA concentration in samples (see Section 3.4.2). We found that none of these communities could degrade BPA (Fig. S.3.2), indicating that the ability to degrade BPA is not a common trait for readily cultured soil isolates.

Therefore, we sought to construct a strain bank of bacterial isolates with the capacity for BPA degradation. Our goal was to acquire 10-20 strains that could be used to assemble synthetic communities. To accomplish this, we identified five sites with potential BPA contamination—two sites in proximity to BPA-producing factories and three sites downstream of these factories (for details of the sampling sites and methods see Section 3.4.1 and Table 3.1, rows 1-5).

We enriched each of the five soil samples for BPA degradation by aerobically incubating 3 grams of soil in 60 mL M9 media with 30 ppm BPA as the sole carbon source (see Section 3.4.1). All of the samples exhibited BPA degradation and continued to do so upon serial dilutions. From these five samples, we isolated and identified 16 bacterial strains, ten of which degrade BPA from 30 ppm in monoculture (degraders) and 6 that do not degrade on their own (non-degraders) but were retained due to being co-isolated with BPA degraders (Fig. 3.1B, Tbl. 3.2). Importantly, the ten BPA degraders did not all fully degrade BPA, giving some variation to our strain bank (Fig. S.3.3). We confirmed BPA degradation occurs via aerobic respiration by measuring CO_2 production in monocultures of each of our ten

degrading isolates (Fig. 3.9).

Armed with a library of isolates capable of BPA degradation we next set out to assemble synthetic communities and test their BPA degradation capabilities. We assembled 70 communities from our 16 isolates and inoculated them into BPA minimal medium. These communities had a mean richness of 6 (with a standard deviation of 3) and included a community of all 16 strains, a community of all 10 degraders, and a community of all 6 non-degraders. The remaining 67 communities were assembled randomly, although care was taken to ensure relatively equal representation of each of the ten BPA degraders to provide some sampling diversity. To assemble communities, each bacterial isolate was grown axenically, and then the appropriate strains were combined with a fixed initial OD of 0.02 for each strain irrespective of community richness.

For each community, we evaluated its capacity to degrade BPA by incubating it for 72 hours in a BPA minimal medium and monitoring BPA concentrations over time (see Section 3.4.3). Previous work suggests that increasing BPA concentration is toxic to bacteria[78, 79], so we expect that increasing the initial concentration renders degradation more challenging. Therefore, we measured every community's BPA degradation dynamics at five initial BPA concentrations (60, 80, 100, 125, and 150 ppm). Note that while typical environmental concentrations are often below 1 ppm [86], contaminated BPA sites often range in BPA concentrations between approximates 10 and 200 ppm [87].

Figure 3.1 C shows the degradation dynamics for all 70 communities in five initial BPA concentrations. Five representative communities are highlighted in color across all concentrations. From these data, we made several initial observations. First, the community consisting of the six non-degrading isolates (Mix19) degrades BPA from an initial concentration of 60 ppm, suggesting the important role of collective effects in degradation. Second, as initial BPA concentration increases, it becomes increasingly harder for our communities to degrade. For example, most communities fail to degrade BPA when incubated at 150 ppm (Fig 3.1 C,

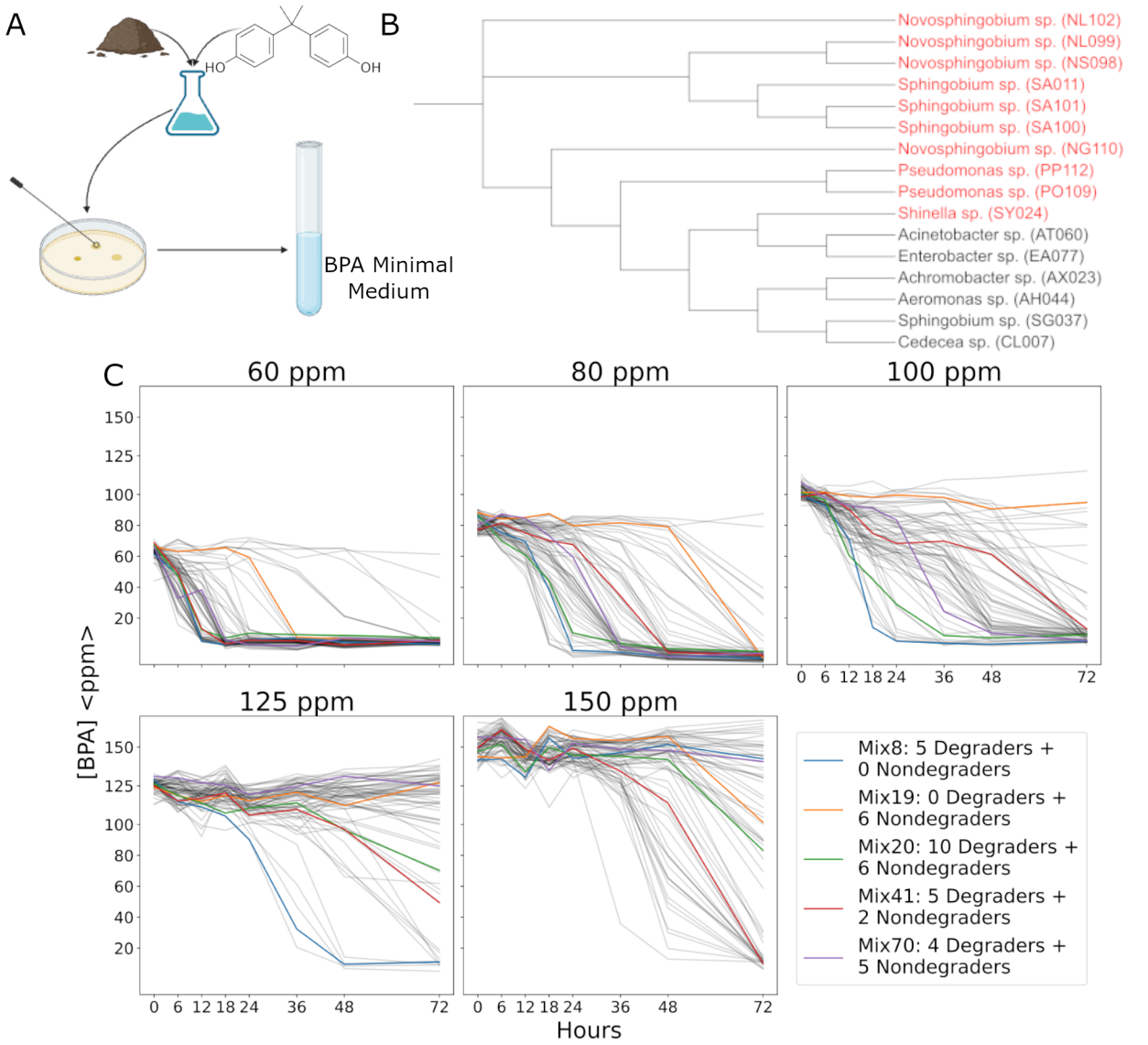


Figure 3.1: **Isolates enable construction of synthetic bacterial communities that degrade BPA across varying initial concentrations.** **(A)** Protocol for enrichment of soil samples. After enrichment, isolates are streaked to purity, and then screened for BPA degradation. **(B)** Phylogenetic tree of the 16 strains isolated. Red strains denote strains that degrade BPA on their own from an initial concentration of 30 ppm. **(C)** BPA degradation curves of 70 communities in five initial concentrations of BPA. Five communities are highlighted: Mix8 (blue), Mix19 (yellow), Mix20 (green), Mix41 (red), and Mix70 (purple). All measurements are corrected for evaporation (see Section 3.4.2).

bottom-center). Third, we observe that many communities that were able to degrade BPA at low concentrations cannot do so at higher concentrations (Mix19 and Mix70, orange and green traces, Figure 3.1 C).

3.2.2 Low-Dimensional Variation in Community Function

To quantify variation in community BPA degradation dynamics, we computed the area under the curve (AUC) for the BPA concentration in time. Communities that degrade BPA rapidly have a low AUC and the converse (Fig. 3.2 A). We chose this metric over explicit modeling of BPA degradation dynamics due to the complexity of the BPA dynamics we observed (such as differing timescales of BPA degradation and plateaus mid-way through degradation).

The AUC metric enables us to directly compare the performance for each community between a low initial BPA concentration (60 ppm) and higher, more stressful concentrations. (Fig. S.3.4). Remarkably, every community that did not degrade BPA from 60 ppm also did not degrade at any of the higher concentrations. We also do not observe any community that degrades BPA faster as concentration increases (points below the dashed lines, Fig. 3.2 B). This shows us that a community's ability to degrade BPA from a lower initial concentration is necessary but not sufficient for that community to degrade BPA from a higher initial concentration.

More strikingly, we observe a relatively simple structure of variation in community function as the BPA concentration increases, consisting of two axes (Fig. 3.2 B 60 vs 125ppm, black arrows). One axis comprises communities that are fast at the lower concentration but spread across AUC values at the higher concentration, and a second axis of communities that are poor degraders at the higher concentration but exhibit variable AUC at the lower concentration. These patterns are also observed when comparing higher concentrations of BPA (Fig. S.3.4). The relatively simple structure in the variation of AUC with BPA concentration suggested a deeper statistical interrogation of how community degradation changed

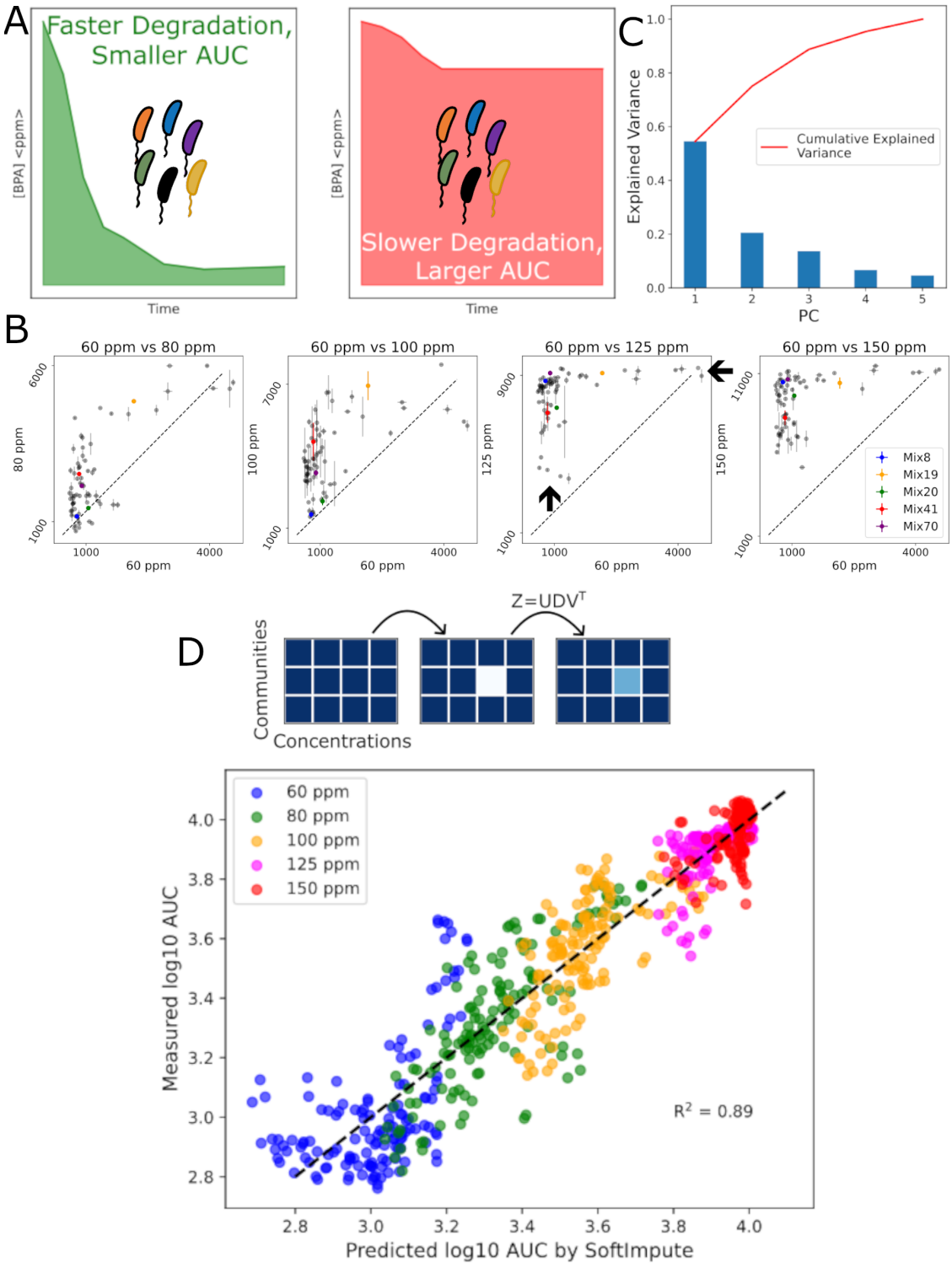


Figure 3.2: **BPA Degradation Varies in a Low Dimensional Manner Across Concentrations** See following page for caption.

Figure 3.2: (Continued from previous page) **(A)** Area Under the Curve (AUC) allows us to quantify BPA degradation of communities. A smaller AUC indicates faster and more complete degradation (left), while a higher AUC indicates the opposite (right). **(B)** Comparing the AUCs of our 70 communities between 60 ppm initial BPA and each of the other four initial concentrations (80, 100, 125, 150 ppm). Vertical error bars represent the standard deviation of AUC across technical replicates. Dotted lines indicate normalized equivalent AUC. Two modes of variation become clearer as the difference in concentrations increases (arrows, bottom left). **(C)** PCA analysis shows that the first two Principal Components explain 79% of the variation in communities' AUCs across concentrations. **(D)** (TOP) Schematic of `SoftImpute` predictions for AUC . A matrix where each row represents a community and each column an initial BPA concentration with entries being BPA degradation AUC. A single entry is deleted from the matrix, and SVD is used to predict the deleted entry (see Section 3.4.4). (BOTTOM) Predictions of all AUCs for all communities in all 5 initial concentrations. Predictions are performed with a log_{10} transformation of AUC. Points are colored by the initial concentration. $R^2 = 0.89$.

with initial concentration.

To further investigate the structure of variation in community function across initial BPA concentrations, we constructed a matrix with rows defined by community composition and columns by initial BPA concentration. The entries of this matrix were the AUC for each community in each BPA concentration. We performed Principal Component Analysis (PCA) on this matrix, which revealed low-dimensional structure in community performance across BPA concentrations with just two dimensions capturing nearly 80% of the variation in the data (Fig. 3.2C). This result suggests that the changes in community function (AUC) across BPA concentration are low dimensional, akin to soft modes observed widely in biological systems [88].

The observations led us to ask whether the low dimensionality of community AUCs can be exploited to make statistical inferences. In particular, the results in Fig. 3.2C suggest that low-dimensional variation in the data might be exploited via matrix imputation methods that rely on low-rank approximations[89]. To test this, we took the matrix of our BPA AUCs across the five concentrations for each community and removed a single entry corresponding to an AUC for a single community composition at a single initial BPA concentration. We then imputed (predicted) the missing value via matrix imputation that exploits a low-rank

approximation to matrices with missing data (`SoftImpute` [89]). We then repeated this iteratively for each community at each concentration (Fig. 3.2D, see Section 3.4.4). Fig. 3.2 D shows the measured versus predicted AUC using the `SoftImpute`. These imputed values are very close to the measured AUC values ($R^2 = 0.89$). The result shows that the BPA degradation performance of a given community at a given initial BPA concentration can be predicted from knowledge of how that community performs at other BPA concentrations. In particular, the decomposition matrix (see Section 3.4.4) was consistently rank 2, meaning that two-dimensional variation between AUCs was sufficient to accurately predict a community's BPA degradation. We also show that this method works when up to 10% of the communities are held out for prediction (Fig. S.3.5). These results strongly a simple structure of the underlying community-function landscape as a function of BPA toxicity.

3.2.3 Inferring the Functional Landscape of BPA Degradation Despite High Epistasis

We have shown that variation in BPA degradation is low-dimensional. The imputation method shown in Fig. 3.2 enables predictions of the performance of a community across BPA concentrations so long as any given community has been measured at least one BPA concentration. However, the imputation approach cannot reveal the roles of individual strains within a community in driving BPA degradation. Therefore, imputation cannot be used to rationally design communities with a composition that has not already been measured. To answer these questions, we turn to the community function landscape.

Our data fits the experimental conditions outlined in Section 2.2.1: a defined pool of 16 species constructed into 70 communities, which is much less than the number of all possible communities, $2^{16} - 1 \approx 65$ thousand. These communities have a single measurement indicative of its functionality - here, we use the $log10$ transformation of AUC as the response variable. However, a number of challenges complicate the process of inferring land-

scapes. The apparently vital role of collective effects in BPA degradation (Fig. 3.1C) indicate some level of epistasis to be inferred, which potentially makes the inferred landscape unreliable (Fig. 2.4). In addition, the rarity of BPA degrading communities as concentration increases means that we have fewer functional communities to learn from. However, the low variation in the structure of AUCs between concentrations (Fig. 3.2C) informs us that a low-dimensional global picture of the community functional landscape exists and is easy to infer.

In order to infer community function landscapes for BPA degradation, we make three critical modifications to the approach in Section 2.2.2. Firstly, we employ a Ridge regularization rather than a LASSO regularization when performing cross-validation. Unlike LASSO regularization, Ridge regularization performs feature shrinkage rather than feature selection [73], and it thus more likely to capture epistasis in higher-order coefficients (See Section 2.5.1, Fig. 2.5) In addition, LASSO regularization's 'feature selection' can be problematic when regressors are correlated [71] (see Section 3.5.2). Secondly, we perform leave-one-out cross-validation during the entire fitting stage rather than 10-fold validation (Fig. 3.11 A). This is to allow the regressions to learn from as much of the data as possible in each fold while still minimizing the risks of overfitting. Thirdly, we employ a weighting scheme during the fitting procedure to prioritize rare communities within each initial BPA concentration. This was done to ensure the regressions prioritize learning functional communities even if they are rare. See the discussion in Section 3.4.5 for more details on the fitting procedure including these modifications.

With these modifications, we inferred community-function landscapes of second-order models (Fig. 3.3A). We note that the regression formalism is identical to Eqn. 2.1 truncated at second-order terms, where the epistatic coefficients β_{ij} are simply renamed as $\gamma_{i,j}$ (this naming convention is adopted within this Chapter for ease of comparison between additive and epistatic coefficients, see Eqn 3.1). We find that adding third-order epistatic terms do

not improve model fits (See discussion in Section 3.5.4). Furthermore, we note that because lower AUC corresponds to faster degradation, more negative terms ($\beta_i, \gamma_{i,j} < 0$) indicate better BPA degradation. We infer functional landscapes for each of the five concentrations independently.

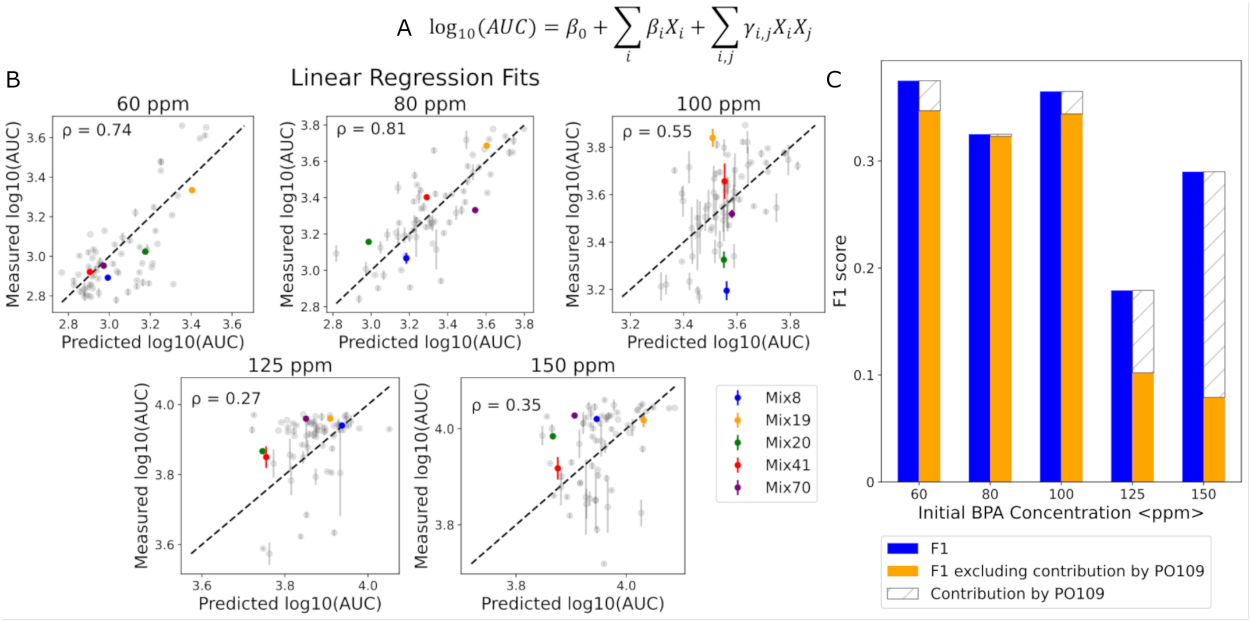


Figure 3.3: **Inferring community-function landscapes shows prominent epistasis** **(A)** The linear regression model for inferring community-function landscapes via regression. The response variable is $log_{10}(AUC)$. The β_i capture additive effects of each strain on function, and the $\gamma_{i,j}$ are epistatic contributions. To avoid overfitting, the model is regularized via a Ridge penalty term. Independent regressions are performed for each initial BPA concentration. See Section 3.4.5 for details. **(B)** For each concentration, a second-order model of the form shown in (A) is inferred. Hyperparameter tuning is performed via cross-validation. Each panel plots measured versus predicted $log_{10}(AUC)$ for each community at a single initial BPA concentration. For each community, at each concentration, the model is trained on the remaining 69 communities (leave-one-out), and this model is used to predict the $log_{10}(AUC)$ for the held-out community. Both duplicates for each community composition are held out at once. This predicted $log_{10}(AUC)$ is plotted on the axis and the measured $log_{10}(AUC)$ on the y-axis. The dotted lines denote perfect predictions. Vertical error bars represent the standard deviation of $log_{10}(AUC)$ across technical replicates. **(C)** Calculated F_1 statistic for each of the five models (blue). On average, F_1 decreases as concentration increases, showing rising epistasis as toxicity increases. Although the F_1 statistic increases at 150 ppm, this increase is due to the contribution of PO109 alone, and accounting for it maintains the pattern of decreased epistasis (orange and striped)

The leave-one-out predictions are shown in Fig 3.3B. Generally, we observe good out-

of-sample predictions (ρ between 0.27 and 0.81) for all five individual models, although the quality of model predictions decreases significantly as initial BPA concentration increases (Fig. 3.3B). Nevertheless, the variation explained by all five models are statistically significant (p-value < 0.05, see Fig. 3.5.3), and therefore are all reliable despite the difficulty with inferring the landscapes from the higher initial concentrations.

Computing F_1 statistics for each model shows us that epistasis is consistently dominant between 60 and 100 ppm BPA but rises considerably at 125 ppm, showing that epistatic terms dictate the landscape further as the initial concentration increases (Fig 3.3C). The increase in F_1 at 150 ppm is solely explained by a large β_i term corresponding to the presence/absence of strain PO109 (which, as we discuss below, actually represents a significant degree of epistasis in the model). Removing the contribution from this strain reveals epistasis increasing even further at 150 ppm. This result is consistent with the increased difficulty in fitting the models to the data for the higher initial concentration of BPA.

3.2.4 Epistasis is the Primary Driver of Toxicity Resistance

In Chapter 2, we observed that many landscapes are relatively smooth (Fig. 2.2), and that epistasis makes inferring functional landscapes difficult (Fig. 2.4). This explains the difficulty in inferring the landscapes from the higher initial BPA concentrations; however, it raises an important question regarding the success of the imputation methods shown in Fig. 3.2 D. In particular, the success of imputation (Fig. 3.2 D) necessarily means that the structure of variation in AUC across BPA concentrations is low-dimensional. It is unclear how this low dimensionality manifests in our inferred community-function landscapes because models for each concentration were inferred independently. Thus, we suspected that the variation in community-function landscapes, changes in coefficients with concentration, should also be low-dimensional. To test this idea, we performed PCA on a matrix of regression coefficients (see Section 3.4.6) and found that 75% variance in coefficients is explained by the first two

components (Fig. S.3.6). The result suggests that, despite the ruggedness in each individual landscape, the global functional landscape of BPA degradation is smooth with respect to initial BPA concentration. To elucidate this connection further, we compared the regression coefficients across initial BPA concentrations (Fig. S.3.6A)

The first property of these landscapes we note is that the additive terms (β_i) decline in magnitude as BPA concentration rises. Specifically, at 60 ppm, many of these terms are large and negative, while at concentrations >100 ppm additive coefficients become closer to zero or positive (Fig. 3.4A, with the exception discussed below). What this means is that the average effects of individual strains on BPA degradation are becoming less important in a statistical sense as BPA concentrations rise. This observation is the first hint that the ruggedness of the landscape changes with BPA concentration.

Interestingly, the additive coefficients at 150 ppm show a singular large negative coefficient, which corresponds to strain PO109 (Fig. 3.4 A, 150 ppm and Fig. S.3.6). This strongly negative coefficient implies that PO109 on average contributes strongly to BPA degradation at a highly toxic concentration whenever it is present in a community; however, we note that PO109 is unable to degrade 150 ppm BPA on its own (Fig. 3.4 D). Thus, we understand this phenomenon as PO109 strongly contributing to BPA degradation from 150 ppm with high effective interactions with all other strains on average; therefore, this additive term actually captures an overall epistatic effect regarding PO109. We also note that the additive coefficient for PO109 is also strongly negative at 125 ppm (but not as negative as 150 ppm) (Fig. S.3.6 A). This further supports PO109's supporting role in degrading toxic concentrations of BPA.

Next, we plotted epistatic coefficients representing degrader-degrader and non-degrader-non-degrader statistical interactions across the range of BPA concentrations (Fig. 3.4B). Epistatic effects ($\gamma_{i,j}$) between degrading strains are positive at 60 ppm initial BPA and zero on average for the higher concentrations. This means that at 60 ppm, degraders tend

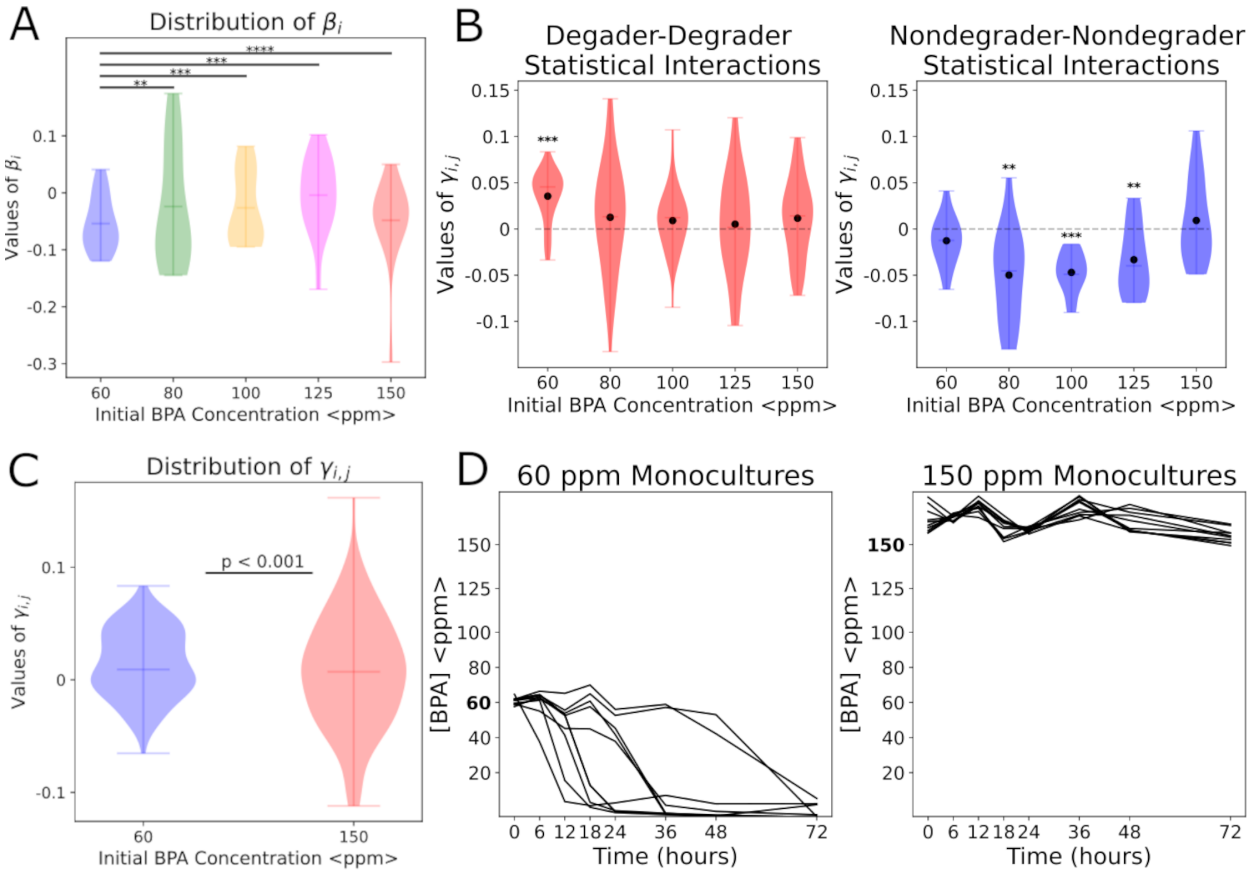


Figure 3.4: **Epistasis Increases Gradually as Toxicity Increases** **(A)** Distribution of additive coefficients (β_i) for linear models fit to each of five initial BPA concentrations (Fig. 3.3). Mean of coefficients rises towards zero as concentration increases. Horizontal lines within the distributions represent the mean of coefficients. $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.001$ when strain PO109 is excluded, all determined by a t-test comparing the distribution of additive coefficients at 60 ppm to the additive coefficients of the other four concentrations. **(B)** Breakdown of epistatic coefficients ($\gamma_{i,j}$) by degrader-degrader coefficients (left) and nondegrader-nondegrader coefficients (right). Degrader-degrader terms are competitive (positive, slowing down degradation) at 60 ppm, but become neutral on average as toxicity increases. Nondegrader-nondegrader effective interactions are cooperative (negative, promoting degradation) on average for the intermediate concentrations but neutral at 60 and 150 ppm. Solid dots represent the mean of coefficients. $^{**}p < 0.01$, $^{***}p < 0.001$ from 1000 bootstrapping iterations rejects the null hypothesis that the mean epistatic coefficient is 0. **(C)** Distribution of all epistatic coefficients for models inferred at 60 and 150 ppm. Epistatic coefficients are more dispersed at 150 ppm, indicating the greater presence of epistasis at high BPA toxicity. Horizontal lines within the distributions represent the mean of coefficients. $p < 0.001$ determined via an f-test. **(D)** Monocultures of the ten BPA-degrading strains in an initial BPA concentration of 60 ppm (left) and 150 ppm (right). None of the monocultures degraded BPA at 150 ppm, indicating the necessity of community effects.

to inhibit each other in terms of BPA degradation, but at higher concentrations of BPA, this is not the case. Moreover, the spread in epistatic terms increases between degraders as the concentration rises. The increasing variance in the distribution of epistatic terms means that there are more extremely negative and positive $\gamma_{i,j}$ as BPA concentration rises, indicating stronger epistasis. This is further shown by the increased spread of epistatic terms at 150 ppm compared to epistatic terms at 60 ppm (Fig 3.4C)

For nondegrader-nondegrader coefficients (Fig. 3.4B, right panel), we observe a decline in the mean epistatic interaction between non-degrading strains as the concentration rises. This decline means that there is more synergistic epistasis between strains. We note that the trend is true up to 125ppm, but at 150ppm, this epistasis rises (Fig 3.4 C).

The conclusion that epistasis rises with BPA concentration strongly suggests that BPA degradation becomes more of a collective, community-level process as toxicity increases. To test this hypothesis, we measured the ability of our ten degraders to degrade BPA in monoculture at 60 ppm and 150 ppm. In agreement with our landscape inferences, we found that while individual strains could degrade BPA at 60 ppm, no single-isolate could degrade BPA at 150 ppm, showing that communities perform better as concentrations rise (Fig. 3.4D).

The low-dimensional variation in landscape coefficients suggests that we might be able to infer landscapes across BPA concentrations in a single regression, assuming low-dimensional variation in community-function landscape coefficients. We developed a regularized optimization procedure to accomplish this and found that the variation in AUC across BPA concentrations enabled such an approach (see discussion in Section 3.5.5).

3.2.5 Functional Landscape Enables Generation of Unseen Communities

Community function landscapes enable the prediction of the function of additional communities by predicting from their composition alone, as seen during cross-validation and

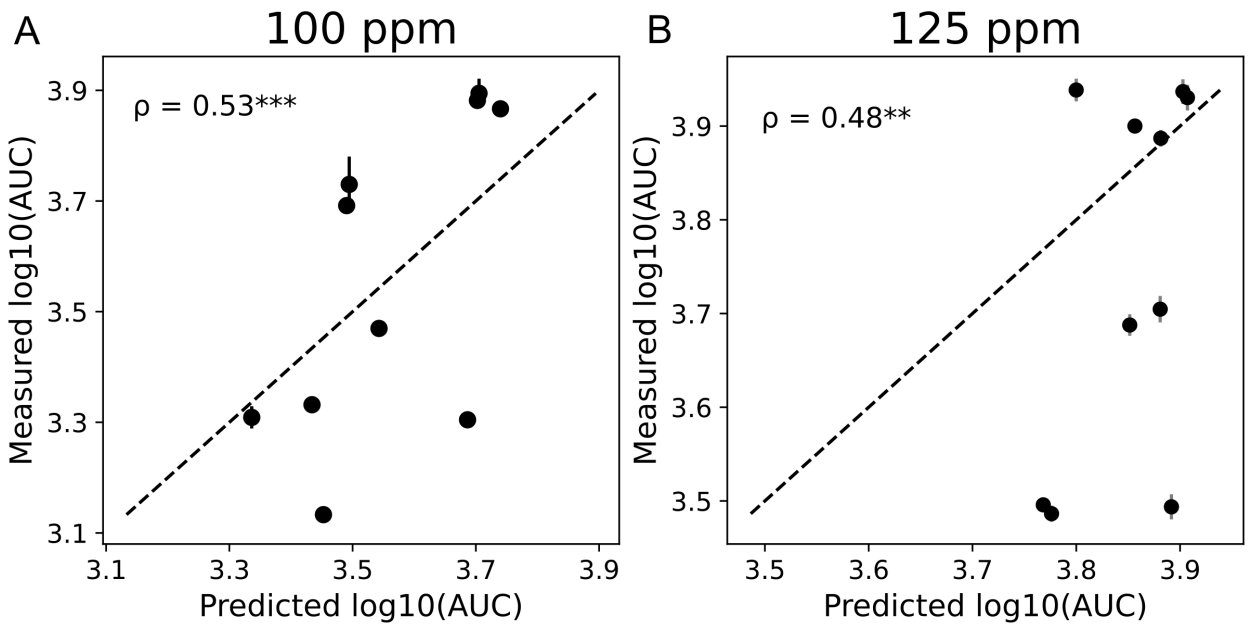


Figure 3.5: **Community Function Landscape Successfully Guides the Design of New, Functional Communities**. Using the model fit to 100 ppm AUC data, ten communities were computationally sampled with a range of predicted functions. (**A**) Scatter plot of predicted vs measured log_{10} AUC for these 10 communities at the initial concentration of 100 ppm. Vertical error bars represent the standard deviation of $log_{10}(AUC)$ across technical replicates. $\rho = 0.53$, $^{***}p < 0.001$ from 1000 bootstrapping iterations rejects the null hypothesis that the Pearson's correlation coefficient is ≤ 0. (**B**) Scatter plot of predicted vs measured log_{10} AUC for the same 10 communities, but at the initial concentration of 125 ppm. Error bars represent the standard deviation of technical replicates. Dotted line shows $y = x$, perfect predictions. $^{**}p < 0.01$ from 100 bootstrapping iterations rejects the null hypothesis that the Pearson's correlation coefficient is <0.

leave-one-out validation attempts (Figures 2.1 and 3.3). As the landscapes rely only on community composition for predictions, the generation of unseen communities with predicted optimal function is a natural next step to take.

Using the model fit to 100ppm, we computationally sampled a range of previously unmeasured communities with a range of predicted performance ($log10$ AUC). We then constructed these communities and measured their BPA degradation dynamics at an initial concentration of 100 ppm. Remarkably, we find the simple model provides good predictive power of the performance of entirely new communities (Fig 3.5A). This establishes the fidelity of the functional landscape as a generative tool for microbial engineering.

Using the same ten communities, we measured their BPA degradation dynamics from an initial concentration of 125 ppm, and correlated their function with the predicted functions from the model fit to 125 ppm data. Although the correlation is significant ($\rho = 0.48$, $p-$value < 0.01, Fig. 3.5B), the quality of predictions is less than ideal - the range of predicted values falls between 3.75 and 3.9, while the measured values fell between 3.5 and 3.95. However, we note that these communities were designed from the 100 ppm model, meaning that the spread in their activity was more in-line with their predicted functions at the lower initial concentration of BPA. Shockingly, the 100 ppm accurately predicted the activity of these communities at 125 ppm (Fig. S.3.7). This result highlights the practical impact of gradually rising epistasis - although the model fit at 125 ppm was relatively poorer due to increased epistasis (Fig. 3.3 B&C), the smooth, low-dimensional shift across concentrations enables generalization (Figures 3.2D and S.3.1).

3.2.6 Enabling Design for Bioremediation in Soils

All of the experiments described above were performed with synthetic communities in a minimal medium with BPA as the sole carbon source. In these conditions, organisms have replete nutrients (nitrogen, phosphorus, sulfur, etc.), and do not compete for resources with other strains. In a remediation context, organisms are often added to settings such as the gut [7, 90] or soils [32, 91], where they compete with or displace an extant microbiome. A major problem with biological remediation is getting the inoculated microbes to engraft in a complex natural environment.

Recently, Silverstein *et al.* suggested that engraftment might be more successful when the inoculated community can strongly occupy a niche that is only weakly occupied, or unoccupied, by the native microbiome [24]. Motivated by this hypothesis, and the fact that BPA degradation is a metabolically complex process not readily performed by all bacterial taxa [92], we wanted to test whether our *in-vitro*-designed BPA degrading consortia could

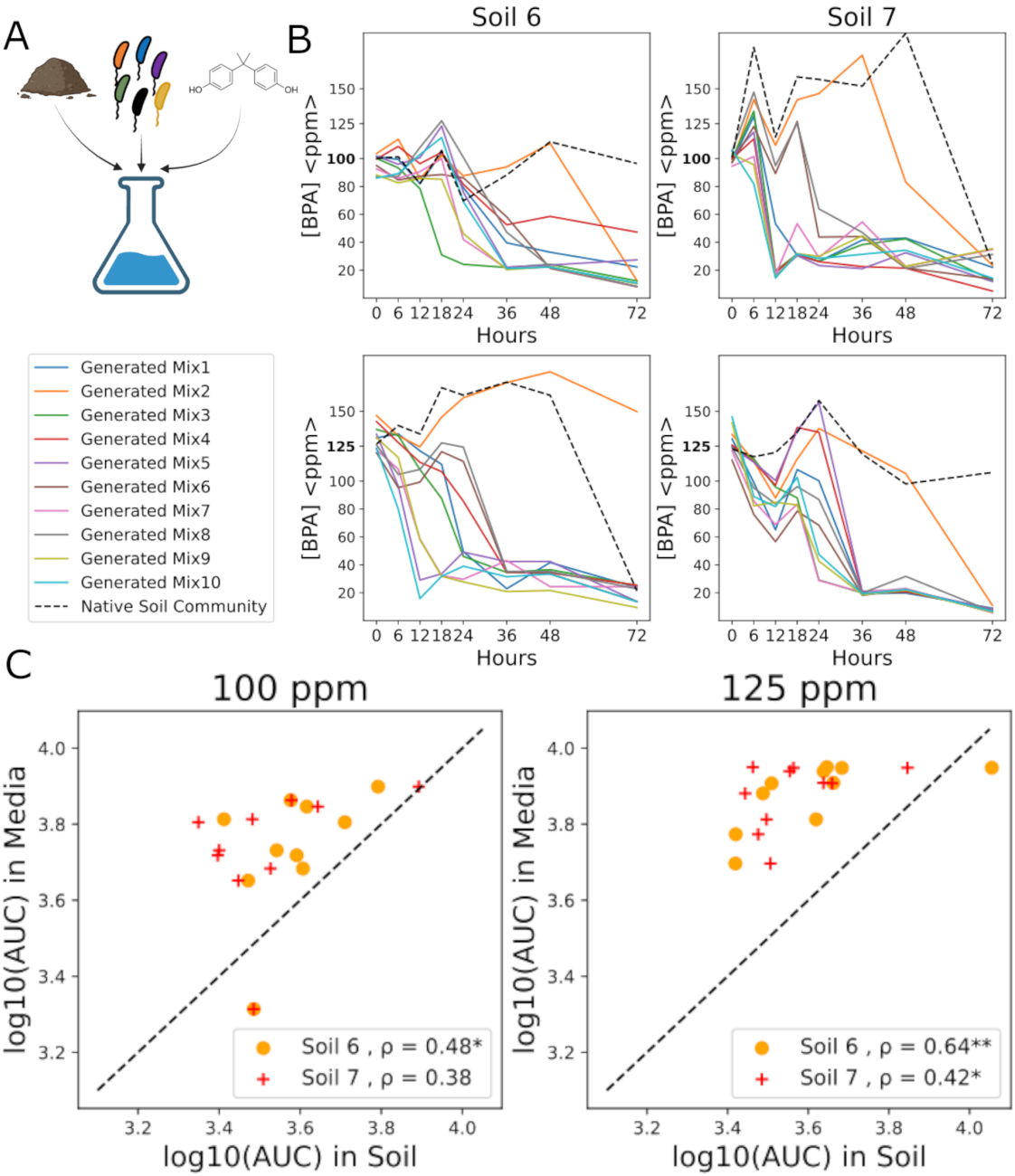


Figure 3.6: **Designing for BPA Remediation in Soils** **(A)** Schematic for BPA remediation protocol. Soil was combined with BPA-contaminated water to form a slurry. Contaminated slurries were inoculated with synthetic communities, and BPA degradation was assayed (see Section 3.4.7). **(B)** BPA degradation curves for the ten synthetic communities in two soils and two initial BPA concentrations: 100 ppm (top) and 125 ppm (bottom). The black line shows BPA degradation by the native soil microbiome without an added synthetic community. **(C)** Scatter plots showing the ten communities' AUC (Fig. 3.5) in soil (x-axis) and *in-vitro* (y-axis) for two soils and two initial BPA concentrations. Dotted lines denote y=x. ρ in the legend denotes Pearson's correlation between *in-vitro* and in soil degradation of BPA. $^*p < 0.05$, $**p < 0.01$ determined from 100 bootstrapping iterations rejects the null hypothesis that the Pearson's correlation coefficient is <0.

drive BPA degradation in soils.

To accomplish this, we created BPA-contaminated soil slurries by mixing soil with BPA-contaminated water. We modified our BPA measurement method to enable extraction from these slurries (see Section 3.4.7). We then inoculated each of the ten communities designed for a range of *in-vitro* degradation performances (Fig. 3.5) in slurries made with two different soil samples, each contaminated with two different BPA concentrations (100 and 125 ppm). As in the experiments above, we measured BPA concentrations in time via sampling (Fig. 3.6A). To assess the ability of the native soil microbiome to degrade BPA, we included a control where no community of isolates was added.

We found that adding communities to soils increased the rate of BPA degradation. In some cases, the native microbiome was not able to degrade BPA on its own (Soil 6, 100ppm, Soil 7, 125ppm). Furthermore, the BPA degradation dynamics in soils depended on which synthetic community was inoculated, suggesting that community composition is important for determining BPA degradation even in the presence of a complex indigenous microbiome. We conclude that designed synthetic communities can impact BPA remediation in soils.

Finally, given that our landscape models enable the rational design of degrading communities (Figures 3.3 and 3.5) we wanted to know whether the performance of a given community *in-vitro* provided any insight into the performance when the same community was used for remediation in soils. To accomplish this, we compared the AUC for each community *in-vitro* with the AUC for BPA degradation by that community in soils. We note that given the challenges of extracting BPA from soil, we do not expect a 1:1 relationship between these two AUC values. However, we find that the AUCs *in-vitro* correlate with the AUC in soils for both soils (Fig. 3.6). This result shows that the landscapes learned for synthetic communities in-vitro are informative of community performance in soils. Our results establish an example of successful bioremediation by designing communities to strongly occupy a niche (in this case, BPA degradation) as posited by Silverstein *et al.* [24].

3.3 Discussion

Using a library of BPA-degrading isolates, we showed that the community-function landscape of BPA degradation can be learned from ensembles of communities where BPA degradation dynamics are quantified. Inspection of BPA degradation dynamics revealed low-dimensional variation with initial BPA concentration. Inspection of the functional landscapes showed that epistasis rises with toxicity, allowing us to conclude that collective effects are key for community toxicity resistance. We then showed that what we learned from the landscapes is translatable into remediation of BPA from contaminated soils.

The functional landscapes described in Chapter 2 are limited to a stable environment and exploring changes in community composition. Our results here particularly emphasize the power of functional landscapes as a tool for understanding environmental changes. In particular, we showed that the functional landscape changes with initial BPA concentration smoothly, to the extent that `SoftImpute` accurately predicts BPA degradation of communities (Fig. 3.2D) and that regression coefficients vary in a low-dimensional manner (Fig. S.3.6). This fact enables prediction of the function of unseen communities despite the ruggedness of each individual landscape (Fig. 3.5). Our work serves as a proof of principle for analyzing community function landscapes across other environmental gradients or even multiple environmental gradients simultaneously (such as the effect of pH on BPA degradation landscape).

We showed that the smooth variation in the landscape is accompanied by epistasis that increases with BPA toxicity S.3.1. At low BPA concentrations, the presence of degraders is the primary driver of degradation, with statistical interactions between them generally competitive (Fig. 3.4). The importance of epistatic interactions for BPA degradation rises with toxicity, where at the highest BPA concentration measured, no individual strain can degrade BPA (Fig. 3.4 D). This is especially evident in the community of all strains - Mix20 - not being the fastest to degrade BPA in all five concentrations measured (Fig. 3.1C). This

result shows that community composition matters in a non-trivial manner. The observation is consistent with the presence of many instances of both statistical cooperation (negative γ coefficients, speeding up degradation) and competitive (positive γ coefficients, slowing down degradation) effective interactions between strains. Our work provides unique insight into the power of a functional landscape in determining epistatic interactions that are biologically relevant.

Previous efforts in bioremediation have encountered difficulties in engrafting bacteria into soils [23, 93], resulting in low bioremediation efficiency [94]. Some studies have attempted to model complex interactions between invaders and resident consortia to design communities for remediation [28]. Our work highlights a practical application of the theory posited by Silverstein *et al.* [24] in combination with community function landscapes to address this problem. This success paves the way for simplifying similar endeavors for other *in-vivo* bioremediation targets.

In this chapter, we have shown that inferring community function landscapes enable broad generalization of community functions, even in varying environmental gradients, that enable rational engineering of further communities. However, strong interactions between the microbiome and the environment were not explicitly considered in this chapter. In a microbial system that relies heavily on interacting with its many external environmental factors, what would the community function need to be able to accurately predict community function based on structure alone? And if a 'highly optimally functional' community was found, can it generalize to other environments beyond where it was engineered to function? In the next chapter, we will analyze these questions by looking at the function of microbes within the Human Gut Microbiome.

3.4 Methods

3.4.1 Isolation of Strains

Five sites were identified as being potentially BPA contaminated—two directly adjacent to BPA-producing factories, and three directly downstream from these factories (Tbl. 3.1, sites 1 —5). Approximately 2 kg of topsoil from each site was harvested and refrigerated at 4 °C for approximately one month before use. Sample #3 is from a riverbank, and is a 1 : 1 mixture of soil and freshwater.

Soil enrichments were performed in 30 ppm BPA M9 media. Enrichments were performed in 250 ml Erlenmeyer flasks, containing 3 g of soil in 60 mL of BPA media. 400 µg/mL of cycloheximide and 40 µg/mL of nystatin were added to inhibit eukaryotic growth. The solution was vortexed, and a 40 µL sample was taken for initial BPA concentration measurement. Flasks were incubated aerobically at 30°C, shaking at 220 RPM.

BPA concentration was measured daily. Whenever the measured BPA concentration fell below half of the starting concentration, the sample would be diluted into fresh BPA media adjusted for a final volume of 60 ml, up to three rounds of dilution. The dilution factor increased with each round —the first was 1:16 (3.75 mL sample with 56.25 mL BPA media), then 1:32, then 1:64. The purpose of dilutions was to select for microbes that either degrade BPA or co-exist with degraders in the presence of BPA.

From soils that degraded BPA after the third dilution, samples were diluted in PBS to a

Soil Site #	**Soil Name**	**Latitude/Longitude**	**Date of Sampling**
Soil 1	Bedford 1	(41.761921,-87.838646)	2022/05/10
Soil 2	Bridgeport 1	(41.82671,-87.657069)	2022/10/17
Soil 3	Park 571	(41.8434044,-87.6636085)	2022/10/17
Soil 4	Alsip 1	(41.65538739,-87.73754314)	2022/10/17
Soil 5	Bedford 6	(41.7620076,-87.8348817)	2022/10/17
Soil 6	LaBaghWoods-240826-1	(41.97780502,-87.74251079)	2024/08/26
Soil 7	Orland-240826-1	(41.57189832,-87.86993291)	2024/08/27

Table 3.1: Information about the seven soil samples obtained in this chapter

10^{-5} dilution and plated on R2A plates (R2 Broth, Research Prodcuts International, product number R24260; Agarose, Fischer Scientific, product number BP1356). Plates were incubated at 30 °C to encourage growth. Colonies were picked and then streaked to purity, after which isolates were mixed 1:1 with 50% glycerol and stored at -80°C. Isolates were sequenced via Sanger 16s sequencing using the universal 16S primers 27F (AGAGTTTGATCMTGGCTCAG) and 1492R (TACGGTTACCTTGTTACGACTT)

16 strains were isolated and identified from soil enrichments. Ten of these strains degraded 30 ppm BPA on their own, while six did not. Our BPA degrading strains came from two soil samples, while our nondegraders came from the other three samples. 16S Sanger sequencing was performed for each strain (as mentioned above), and DNAStar Lasergene 17 [95] was used to construct consensus sequences. Strain taxonomy was identified through NCBI BLAST [96]. The EMBL-EBI Job Dispatcher [97] was used to align sequences via Muscle and construct the phylogenetic tree (Fig. 3.1 B) and the Interactive Tree of Life [98] was used to visualize and edit the tree.

Strain ID #	**Genus**	**Soil Sample of Origin**	**Degraded in 30 ppm BPA?**
CL007	Cedecea	Bedford1	No
SA011	Sphingobium	Bedford1	Yes
AX023	Achromobacter	Bedford1	No
SY024	Shinella	Bedford1	Yes
SG037	Sphingobium	Bridgeport1	No
AH044	Aeromonas	Bridgeport1	No
AT060	Acinetobacter	Park 571	No
EA077	Enterobacter	Alsip1	No
NS098	Novosphingobium	Bedford6	Yes
NL099	Novosphingobium	Bedford6	Yes
SA100	Sphingobium	Bedford6	Yes
SA101	Sphingobium	Bedford6	Yes
NL102	Novosphingobium	Bedford6	Yes
PO109	Pseudomonas	Bedford6	Yes
NG110	Novosphingobium	Bedford6	Yes
PP112	Pseudomonas	Bedford6	Yes

Table 3.2: Information about the 16 strains isolated

3.4.2 BPA Concentration Assay

BPA concentrations were measured using a modified version of the phenol colorimetric assay [99] optimized for high-throughput use in a 96-well plate. Under basic pH conditions, 4-Aminoantipyrine (4-AAP) reacts with BPA in an electrophilic substitution reaction, which can then be oxidized by potassium ferricyanide to produce a red dye that absorbs at 506 nm in proportion to the initial BPA concentration in the sample. Reagents were added in the following order: 120 µL of 0.25 M sodium bicarbonate (Sigma-Aldrich, product number S6014) to serve as a basic buffer, 40 µL of a BPA containing sample sample, 40 µL of 20.8 mM 4-AAP (Sigma-Aldrich, product number A4382), and 40 µL of 83.4 mM potassium ferricyanide (Arcos Organics, product number 424125000). The plate is incubated at room temperature, shaking at 750 RPM for 45 seconds, and then incubated statically for ten minutes. Afterward, the absorbance of each sample was read at 506 nm. A standard curve was made from solutions of known BPA concentration in BPA minimal media as described above for every timepoint in every experiment, and each 96-well plate was prepared with its own standard curve. The standard curves had a slope of 133.36 ± 13.14 and an intercept of -0.88 ± 3.1(Fig. 3.7).

3.4.3 Synthetic Community Design

To prepare our strains for mixing into synthetic communities, each of the strains was sampled from its frozen stock and grown anaerobically in 5 ml Tryptic Soy Broth (TSB) media for 44 hours in tightly-closed falcon tubes, incubating at 30°C and 220 RPM. After incubation, samples were centrifuged at 4000 RPM for 15 minutes and the supernatant was pipetted out. Pellets from each sample were resuspended in 1ml M9 media as a wash step, and centrifuged again at 4000 RPM for 8 minutes, after which the supernatant was pipetted out again. Finally, pellets were suspended in 2 ml of BPA media, and samples were ready for community assembly.

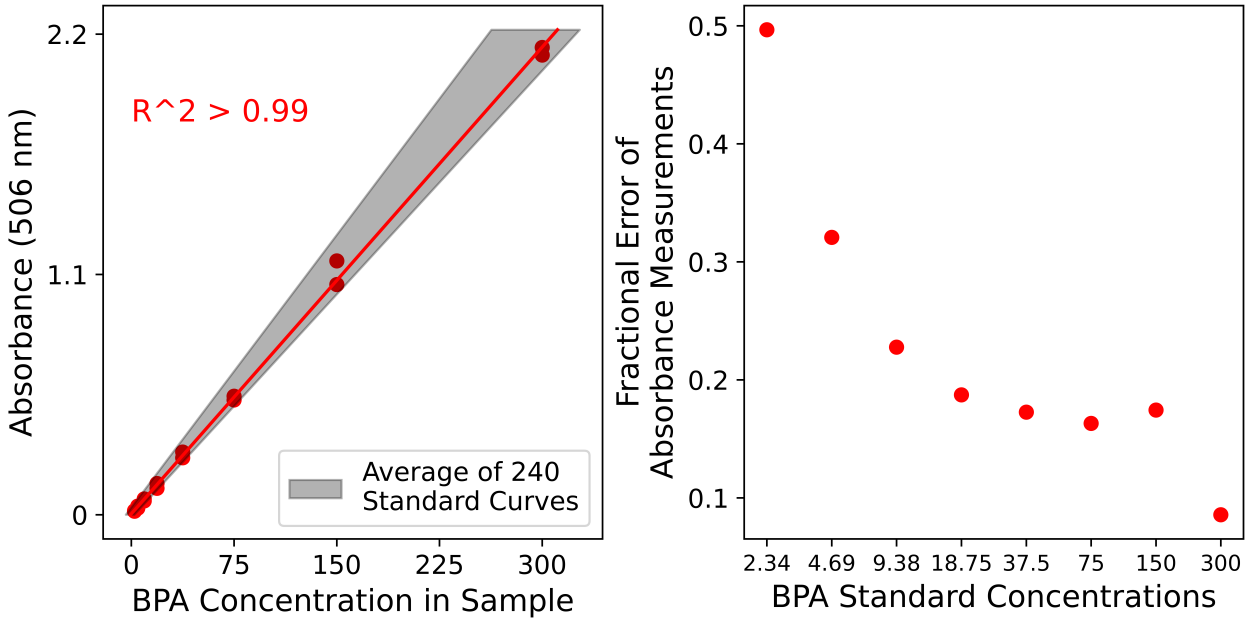


Figure 3.7: **Standard Curve of BPA Concentration Assay (LEFT)** A sample standard curve of BPA concentration, along with the line of best fit (red). $R^2 = 0.9998$ as calculated by the `scikit-learn` function `r2_score`. Gray shaded region indicates the error of the slope of the average standard curve. **(RIGHT)** The fractional error of absorbance measurement as a function of the prepared BPA standard concentration. Fractional error is large at the lowest concentrations and small at the highest concentrations.

All communities were mixed and incubated in Fischer 48 deep well plates (part number 1162B95), with one well housing one community. Each well was filled with 2 mL of M9-BPA medium, and each isolate was added at an initial nominal OD of 0.02 irrespective of the community richness. All wells were mixed thoroughly by pipetting up and down repeatedly and then incubated at 30°C in a Fisher Scientific benchtop microplate shaker (Catalog number 02-217-757) shaking at 900 RPM, covered by a breathe-easy membrane (Sigma-Aldrich, part number Z380059). Every plate also contained at least one well with no bacteria added to it (as an evaporation control) and one well with TSB medium (as a contamination/splashing control). Every community was run in technical duplicates.

Samples were taken at time points 0, 6, 12, 18, 24, 36, 48, and 72 hr, and the BPA concentration was measured for each sample at each time point as previously described. Positive controls ensured degradation dynamics across different batches of experiments were

comparable, while negative controls detected contamination and also served as evaporation controls.

3.4.4 Imputation to predict AUCs across concentrations of BPA

In Fig. 3.2D we predicted community AUCs via matrix imputation. Imputing entries from our AUC matrix Z (Fig. 3.2D) was done via the R package `SoftImpute`. Every entry of Z is denoted $z_{c,k}$, which is the $log_{10}(AUC)$ of community c in concentration k. We iteratively dropped out every entry of the matrix (along with all technical replicates), to achieve $Z_{-c,k}$ with community c in concentration k dropped out. We then process $Z_{-c,k}$ to be $\tilde{Z}_{-c,k}$, where every column k is z-scored. and fit `SoftImpute` using the SVD: $\tilde{Z}_{-c,k} = UDV^\top$, where D is soft-thresholded (D^*)to minimize its nuclear norm. The Z matrix is reconstructed by D^* as $Z^* = UD^*V^\top$, where $z^*_{c,k}$ is the imputed value of the dropped entry. Leave-one-out cross-validation was done to choose the soft-thresholding regularization parameter over a range of 40 values between 0 and 10 which maximizes the total R^2 of predictions (Fig. S.3.5 A). For all fits, the resulting D^* was rank 2.

We also performed a similar imputation by randomly dropping out 10% of the entire matrix (accounting for technical replicates), repeated over 100 iterations of 10% dropout (Fig. S.3.5 B).

3.4.5 Regressions

Regressions were performed using the Python `scikit-learn` package. . For each concentration k, models were fit of the form

$$y_c^{(k)} = \beta_0^{(k)} + \sum_i \beta_i^{(k)} X_i + \sum_{i,j} \gamma_{ij}^{(k)} X_i X_j \tag{3.1}$$

, where $y_c^{(k)}$ is the z-scored $log_{10}(AUC)$ of community c in concentration k (similar to our imputation method above, $y_c^{(k)}$ is z-scored within each model for concentration k), $\beta_0^{(k)}$ is the intercept that is fit from the data for concentration k, X_i represents the presence and absence of strain i in the community c, denoted in +1/-1 notation, $\beta_i^{(k)}$ is the additive coefficient for stain i in the model fit from concentration k, and $\gamma_{i,j}^{(k)}$ is the pairwise coefficient for strains i and j together in the community c in the model fit from concentration k. The objective function for each model was

$$\underset{\beta,\gamma}{\operatorname{argmin}} ||w_c^{(k)}(y_c^{(k)} - (\beta_0^{(k)} + \sum_i \beta_i^{(k)} X_i + \sum_{i,j} \gamma_{ij}^{(k)} X_i X_j))||^2 + \lambda(\sum_i (\beta_i^{(k)})^2 + \sum_{i,j} (\gamma_{ij}^{(k)})^2) \quad (3.2)$$

where $w_c^{(k)}$ is the weight applied to community c in the model for concentration k and λ is the L_2 penalty term. Weights were determined through a Gaussian Kernel Density Estimate of community AUCs using `scikit-learn`. For each $y_c^{(k)}$, the estimator computes

$$\rho_k(y_c^{(k)}) = \sum_{i=1}^{N} \frac{1}{h\sqrt{2\pi}} \exp(-\frac{(y_c^{(k)} - y_i^{(k)})^2}{2h^2}) \quad (3.3)$$

where $y_i^{(k)}$ represents all measured $y^{(k)}$, and h represents the bandwidth of the curves. For each concentration, a grid search of h between 10^{-3} and 1 was performed to find the bandwidth that maximized the log-likelihood of the estimated distribution of all $y_c^{(k)}$. For all other parameters, the default `scikit-learn` settings were used. The final weight $w_c^{(k)}$ was set to be $\frac{1}{\rho_k(y_c^{(k)})}$ so that a community whose AUC is more 'rare' within concentration k are given greater weight in the regression.

The optimal λ was chosen through leave-one-out cross-validation over a range of 10,000 values between 10^{-5} and 10^{10} and finding the value that minimized the objective function above (similar to our imputation method above, z-scoring of $y_c^{(k)}$ was only performed after excluding the validation community). Once the optimal λ was identified, the predicted

Initial Concentration	Value of Optimal L_2 Penalty
60 ppm	82.07
80 ppm	1.10×10^{-5}
100 ppm	133.57
125 ppm	32.41
150 ppm	16.64

Table 3.3: L_2 **Penalty Values for Five Regressors on BPA Degradation**

value for each community within each concentration was plotted against the measured value (Fig. 3.3 B). A separate model was fit to all $y^{(k)}$ within each concentration k, each with their own optimal λ (Tbl. 3.3). The cross-validation for the model fit to data from 80 ppm BPA yielded the smallest tested value for λ (near 0), in stark contrast to the other four models. The strong model fit (Fig. 3.3 B) indicates the general ease of fitting the model to this data compared to data from the other four initial concentrations.

3.4.6 PCA Analysis

PCA analysis was performed on our matrix of regression coefficients (Fig. S.3.6 C) to determine the variation in community function landscapes. For each concentration k, the column of our matrix becomes $[\beta_0^{(k)}, \beta_1^{(k)}, \beta_2^{(k)}, \ldots, \beta_{16}^{(k)}, \gamma_{1,2}^{(k)}, \gamma_{1,3}^{(k)}, \ldots, \gamma_{16,15}^{(k)}]$ The coefficients were scaled by subtracting the mean and normalizing by the standard deviation within each concentration before being subject to PCA transformation.

3.4.7 BPA Extraction and Community Inoculation in Soils

Two soils were sampled in amounts of roughly 2 kg from prairie fields (Tbl. 3.1, sites 6 & 7). Soils were sieved through to 2mm, and a portion of each was sterilized by autoclave at 121°C for 99 minutes three times, with 24 hours in between each time and the last occurrence being 8 hours before preparing the communities.

0.85 g of each soil was suspended in 1.7 mL of 100 ppm BPA media in 48 deep well plates

and vortexed to ensure sufficient mixing of BPA. Strains were added in an identical manner to community mixing experiments. Incubation was also performed similarly. For each soil, a sample was prepared with no community added, and another sample with non-BPA media for controls. Two technical replicates for each condition were prepared.

To measure BPA concentration at each sampling time point, BPA needed to be extracted from the soil slurry. 100 µL of the slurry was sampled and mixed with 150 µL of 100% ethanol. The samples were then vortexed and then spun down at 4000 RPM for 2 minutes. 40 µL of the supernatant was sampled to use for our BPA concentration assay, with an extraction efficiency of approximately 60% (Fig. 3.8)

3.4.8 Data and Code Availability

All datasets generated in our study are available within the following Open Science Framework (OSF) project: https://osf.io/gcfr8/. All code written for models and analysis was written in either python or R, and is available at the above OSF project as well.

3.5 Supplementary Discussion

3.5.1 Measuring CO_2 Production from BPA Degradation

CO_2 production in our ten degrading isolates by BPA Degradation was measured by a Microresp assay (Fig. 3.9). A monoculture of each strain was prepared as described in Section 3.4.3. Four technical replicates were prepared: two in 60 ppm BPA media, and two in M9 media with no carbon source. 700 µL of each replicate was transferred into a 96-deep well plate, which was then sealed with a pH-sensitive dye in an agarose gel. Respiration of CO_2 reduces the pH of the dye, thus changing its color (See de Jesús Astacio *et al* [100] for more information on the setup of the Microresp system). The plate was incubated at 30°C shaking at 900 RPM for three days, after which CO_2 respiration was measured by an

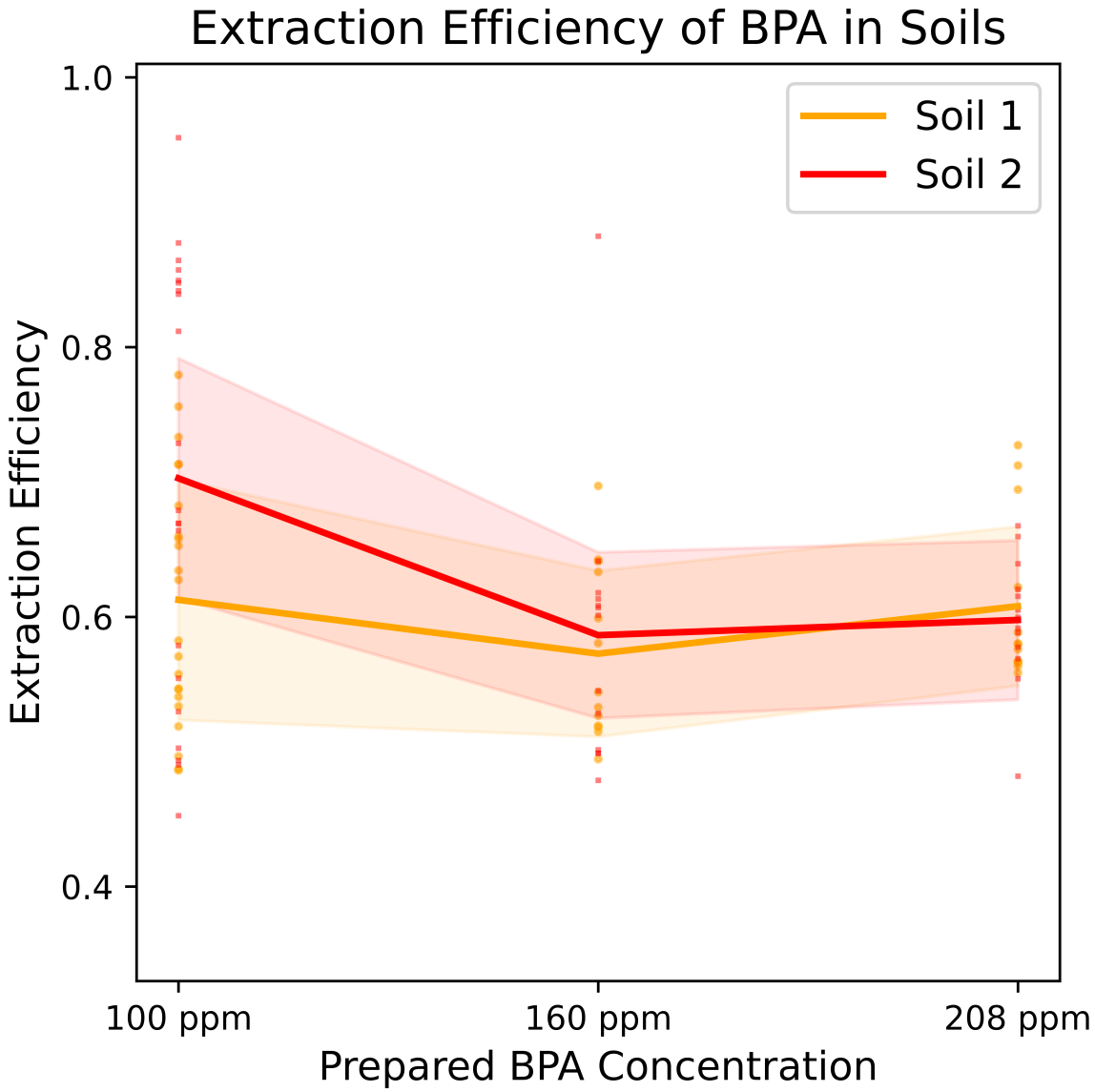


Figure 3.8: **BPA Extraction Efficiency from Soil Slurries** Extraction efficiency of BPA from both soil slurries from three initial BPA concentrations. Solid colored lines represent the mean extraction efficiency for each soil at each concentration, and shaded regions denote one standard deviation from the mean. Number of replicates varied between 20 and 40 for each condition.

absorbance reading of the dyes.

3.5.2 Stability Analysis Shows Covariance Between Features Impedes LASSO Regularization

In this chapter, Ridge regularizations were chosen over LASSO regularizations for two reasons: (1) Ridge regularization's 'feature shrinkage' encourages capture of epistatic terms (see Section 2.5.1), and (2) LASSO regularization's 'feature selection' is affected by correlated features. This second reason is of special concern when dealing with epistatic functions, as epistatic interactions co-vary strongly with their additive counterparts [101].

An effective way to determine the extent of LASSO regularization's susceptibility to feature co-variance is through a Stability Analysis [6]. In this method, a LASSO-regularized regression is fit to the data multiple times, and coefficients are analyzed for how frequently they were 'selected' by the model. Features that are not selected frequently enough are indicators of unstable features. More unstable features indicate a more unstable regression.

We performed a Stability Analysis of a LASSO-regularized regression on our data. For our matrix of taxa presence/absence X and the AUC of degradation for each concentration $y^{(k)}$, we performed the following:

1. A LASSO-regularized regression was fit to $X, y^{(k)}$ in the same method as described in Section 2.4.2. This was performed 100 times, and for each iteration i, the array of coefficients $S_k^{(i)}$ was stored. Coefficients are ordered as $[\beta_1^{(i)}, \beta_2^{(i)}, \ldots, \beta_{16}^{(i)}, \gamma_{1,2}^{(i)}, \gamma_{1,3}^{(i)}, \ldots, \gamma_{15,16}^{(i)}]$. For simplicity, we will refer to coefficients as s_b, such that $S_k^{(i)} = [s_1^{(i)}, s_2^{(i)}, \ldots, s_{136}^{(i)}]_k$

2. For each coefficient s_b, for all iterations i, the number of $s_b^{(i)}$ that is nonzero was counted. Note that, as implemented in the `scikit-learn` package, LASSO regularizations do not always set a coefficient to exactly 0 but rather a small number on the order of 10^{-7}. Therefore, we counted $s_b^{(i)} > 10^{-5}$. Let C_b be the selection frequency of

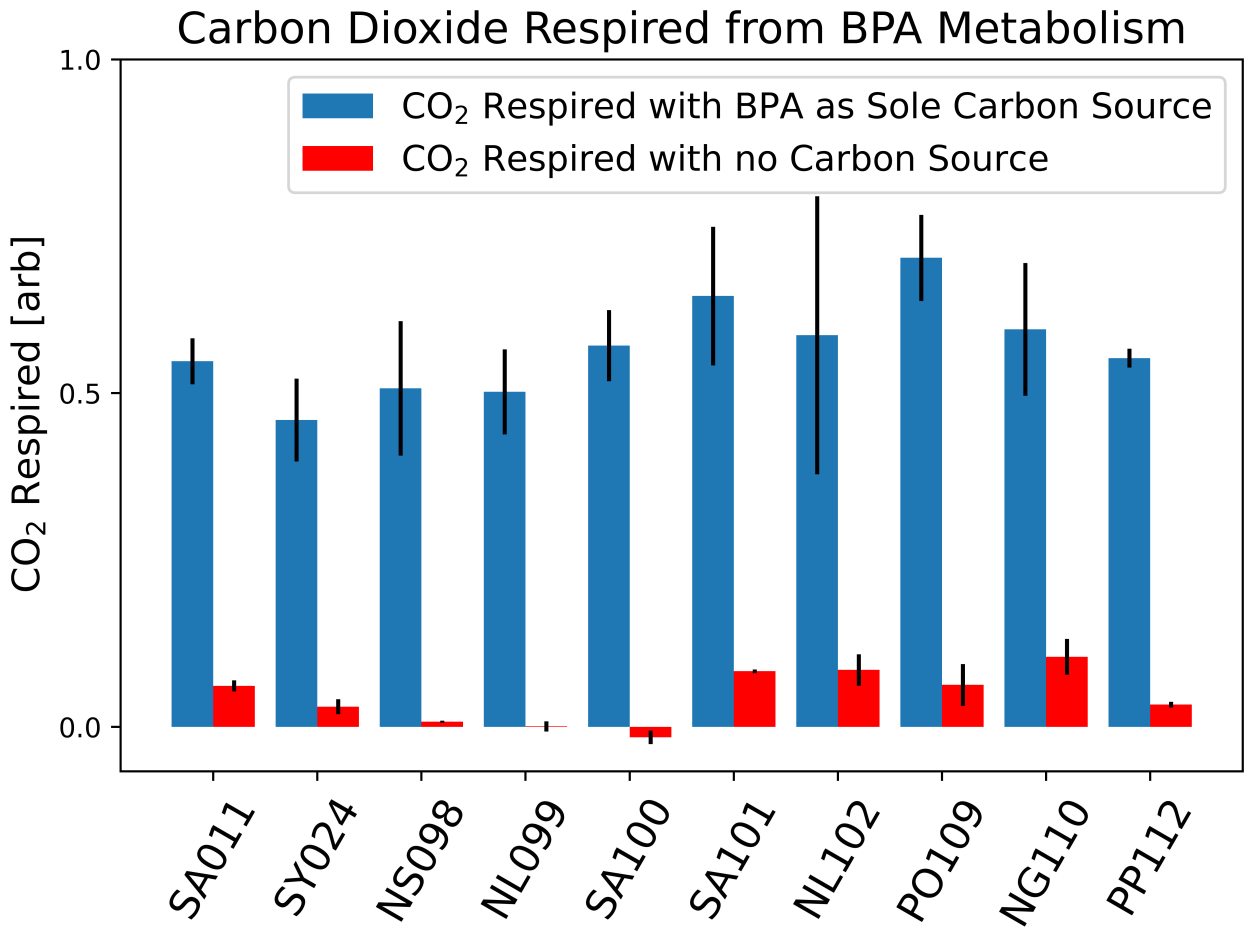


Figure 3.9: **CO_2 production by BPA Degraders during BPA degradation** Net CO_2 production by the 10 BPA-degrading strains when individually inoculated in 60 ppm BPA (blue) and when inoculated in minimal media with no carbon source (red) for three days. Error bars represent the standard deviation of technical duplicates. CO_2 measurements were performed with the microresp platform as described.

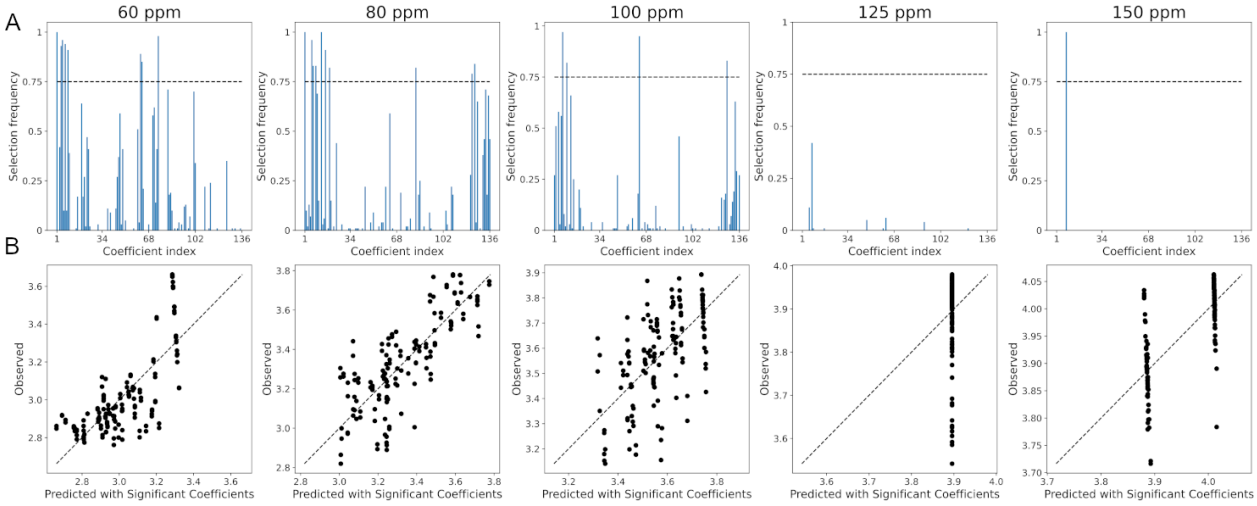


Figure 3.10: **LASSO Regularizations are Unstable with BPA Data** (**A**) Selection frequency (C_b) for all coefficients for each dataset (from left to right): 60 ppm, 80 ppm, 100 ppm, 125 ppm, and 150 ppm. Dotted line shows $C_b = 0.75$, the frequency cutoff for stability. (**B**) Regression predictions using only significant featurs identified from **A** (x-axis) versus observed AUCs (y-axis) for each of the five datasets. Dotted black lines denote y=x in all cases.

$s_b^{(i)} > 10^{-5}$ (divided by 100, such that $C_b = 1 \rightarrow s_b^{(i)} > 10^{-5} \forall i$ within concentration k)

3. For all coefficients s_b, if $C_b < 75$, then the coefficient was considered to be unstable.

4. Let $\hat{S}_k$ be the vector of all stable coefficients, i.e. $\hat{S}_k = [s_b]_k$ s.t. $C_b \geq 75$

We note that while a cutoff frequency of 0.9 is recommended, cutoff frequencies larger than 0.6 have shown little difference in overall analysis [6]. As such, we have chosen the halfway point as our cutoff value

Performing this analysis on our dataset revealed shockingly low stability in the features: 8 stable coefficients in the 60 ppm data, 10 stable coefficients in the 80 ppm data, 4 stable coefficients in the 100 ppm data, 0 stable coefficients in the 125 ppm data, and a single stable coefficient in the 150 ppm data (Fig. 3.10A).

These results indicate highly unstable LASSO regularizations on our datasets. However, it is possible that these features were selected because they are truly the only important

features for predictions [71]. To test this, we considered the dataset $X^{\hat{S}_k}$, which is the dataset X only including the columns that correspond to the significant features found in $\hat{S}_k$. We performed leave-one-out cross-validation on $X^{\hat{S}_k, y^{(k)}}$ on unregularized regressions. The leave-one-out predictions are shown in Fig. 3.10. While the data from 80 ppm is resilient to low coefficients and the data from 60 ppm experiences a mild drop in predictive power, the data from the higher three concentrations show a massive loss in predictive power. This proves the overall instability of LASSO regularizations for our data.

3.5.3 Regression Models are Statistically Significant

To evaluate the significance of our model fits (Fig. 3.3) and all resulting analyses, we sought to employ a null model of random AUCs for each community. To do this, within each initial BPA concentration k, we de-coupled y_c^k from community composition c by randomly shuffling the order of the response variable y_c^k (see Eqn. 3.1). This randomization enforces no relationship between a community composition and associated AUC. After random shuffling, models were fit in the same way as described in Section 3.4.5. The weighting scheme was not changed; therefore, every y_c^k retained its weighting w_c^k, ensuring that rare communities still prioritized.

For each model, 100 random shuffles and subsequent fits were performed, and ρ (Pearson's) between predicted and "true" AUC were compared to that of the original models. The data (Fig. 3.11B) shows that models fit to random shuffles of the response variable perform significantly worse than the original models, where p-value < 0.01 for the lower three concentrations and p-value < 0.05 for the higher two concentrations. The result suggests that our linear models are utilizing community composition to predict AUC in a statistically significant fashion.

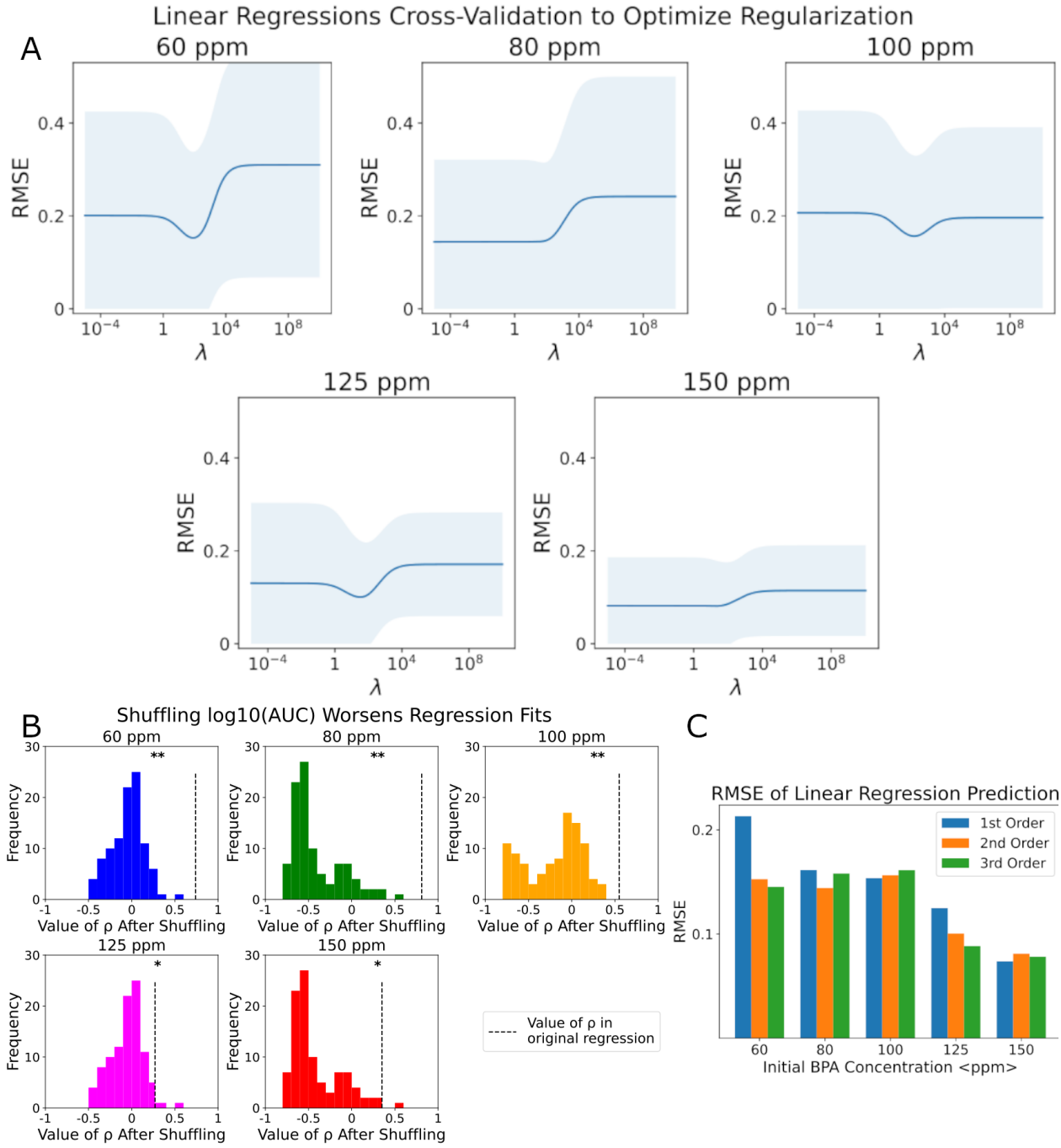


Figure 3.11: **Linear Regression Models.** See following page for caption.

Figure 3.11: (Continued from previous page) **(A)** Root Mean Squared Error (RMSE) (Eqn. 3.2) vs λ during leave-one-out cross-validation for all five regressions (see Section 3.4.5, Tbl. 3.3, and Fig. 3.3). The blue line is the average across all folds, and the shaded region is the standard deviation across all folds. The optimal penalty term was selected for each model to minimize the RMSE. **(B)** Shuffling the response variable reduces the predictive power of the model. Each panel corresponds to different initial BPA concentrations. For each concentration, we shuffled the response variable (log10(AUC)) 100 times, for each shuffle we performed a regularized linear regression as described in Section 3.4.5. The histogram shows the distribution of Pearson correlation (ρ) between measured and predicted log10(AUC) for the 100 randomizations. Dotted black lines indicate the ρ of the original regression fit (Fig. 3.3 B). *p-value $< .05$, **p-value < 0.01. Model fits to randomized data are significantly worse than the true data fits in Fig. 3.3. **(C)** Root Mean Squared Error (RMSE) for each of the five concentrations fit to linear regressions of models truncated at first-order terms (β_i), second-order terms ($\gamma_{i,j}$), and third-order terms ($\zeta_{i,j,k}$). For each model at each order, leave-one-out cross-validation was done to pick the optimal penalty term to minimize out-of-sample RMSE (see see Section 3.4.5).

3.5.4 Second-Order Models Outperform Models Truncated at First- and Third- Order Terms

In Chapter 2 (Fig. 2.1), we observed that for low-ruggedness landscapes, first-order models explained most of the variance in the functions measured, while second-order coefficients improved model fits and third-order coefficients mostly did not improve model performance. However, landscapes inferred for BPA degradation are significantly more rugged than those inferred in Chapter 2 (Fig. 2.2, Tbl. 2.1). Therefore, we wanted to test whether we would see similar results with BPA degradation landscapes.

We fit AUC data from all five initial BPA concentrations into two additional regressions: one only including first-order terms, and one truncated at third-order epistatic terms. To be explicit, the model including only first-order terms was:

$$y_c^{(k)} = \beta_0^{(k)} + \sum_i \beta_i^{(k)} X_i \tag{3.4}$$

while the model truncated at third-order terms was:

$$y_c^{(k)} = \beta_0^{(k)} + \sum_i \beta_i^{(k)} X_i + \sum_{i,j} \gamma_{ij}^{(k)} X_i X_j + \sum_{i,j,k} \zeta_{ijk} X_i X_j X_k \tag{3.5}$$

All models were fit in the same way and regularized with Ridge regularization (see Section 3.4.5), and model fits were compared by comparing the resulting RMSEs of leave-one-out cross-validation.

Results (Fig. 3.11 C) show us that, on average, adding third-order terms (green) does not substantially improve model fits compared to a model up to second-order coefficients (orange), with the only major difference seen at 125 ppm. However, we do see that fitting the data to only first-order coefficients (blue) does not lose significant predictive power compared to the second-order model, with the models fit to 100 ppm and 150 ppm data being nearly equivalent. The only major difference is seen at 60 ppm, the least epistatic model of the five, in which second-order coefficients greatly improve model performance.

At face value, it may appear that the results of Fig. 3.11C contradict the importance of epistasis as concentration increases. However, we note that the quality of the fit (RMSE, y-axis Fig. 3.11C) does not indicate the *relative importance or magnitude* of additive vs epistatic coefficients in the regression [57]. This is quantified via the F_1-statistic.

3.5.5 Fitting a Low Rank Regressor

Inspired by the success of `SoftImpute` as well as the low-dimensional structure of community AUCs (Fig. 3.2 D), we sought to pursue a regression model that would exploit the low-dimensionality of our data to inform the regressions.

The low-rank regressor is a model that attempts to learn coefficients from all initial BPA concentrations simultaneously. It assumes the relationship

$$Z \approx X\Theta Y, \quad \text{for an approximately low-rank matrix } \Theta \in^{p\times r}, \tag{3.6}$$

where $Z \in^{n_{\text{comm}} \times n_{\text{BPA}}}$ is the matrix of community AUCs under all available n_{BPA} concentrations, $X \in \pm 1^{n_{\text{comm}} \times p}$ is the community composition matrix and $Y \in^{r \times n_{\text{BPA}}}$ is a polynomial basis matrix, with each row comprising monomials evaluated at the available BPA concentrations:

$$Y = \begin{bmatrix} 1 & 1 & \dots & 1 \\ \phi_1(\mathsf{PPM}_1) & \phi_1(\mathsf{PPM}_2) & \dots & \phi_1(\mathsf{PPM}_{n_{\text{BPA}}}) \\ \phi_2(\mathsf{PPM}_1) & \phi_2(\mathsf{PPM}_2) & \dots & \phi_2(\mathsf{PPM}_{n_{\text{BPA}}}) \\ \vdots & & & \vdots \\ \phi_r(\mathsf{PPM}_1) & \phi_r(\mathsf{PPM}_2) & \dots & \phi_r(\mathsf{PPM}_{n_{\text{BPA}}}) \end{bmatrix}. \tag{3.7}$$

Here, r is the degree of the polynomial, $\phi_1, \dots, \phi_r$ are the elements in the polynomial basis, and PPM_j denotes the jth BPA concentration. In our experiments, we choose ϕ_j to be the monomials, with $\phi_k(x) := x^k$; however, alternative choices may be able to provide better fits. Under this formalism, the product $\Theta Y \in^{p \times n_{\text{BPA}}}$ is the matrix of coefficients for all observations and satisfies $(\Theta Y) \leq r$ by construction.

Fitting

We calculate an estimate $\widehat{\Theta}$ from data by solving the following optimization problem:

$$\widehat{\Theta} := \underset{\Theta}{\operatorname{argmin}}\, W \odot (Z - X\Theta Y)^2 + \lambda_1 \Theta_* + \lambda_2 \Theta^2, \tag{3.8}$$

where $W \in_+^{n_{\text{comm}} \times n_{\text{BPA}}}$ is a nonnegative weighting matrix (see Section 3.4.5 for our weighting scheme) and:

- $\odot$ denotes the elementwise (Hadamard) product;
- $\Theta_* := \sum_{i=1}^{r} \sigma_i(\Theta)$ is the *nuclear* norm, equal to the sum of all singular values of Θ;

- $\Theta = \sqrt{\sum_{i,j} \Theta_{i,j}^2}$ is the so-called *Frobenius* norm;
- λ_1 and λ_2 are nonnegative, tunable parameters controlling the regularization strength.

The sum of the last two terms in (3.8) is a regularizer akin to the well-known Elastic Net [102]. The nuclear norm acts as an ℓ_1 penalty on the singular values, promoting low rank in the solution $\widehat{\Theta}$, while the squared norm induces shrinkage in the learned coefficients like an ℓ_2 penalty.

The optimization problem (3.8) is convex (in fact, *strongly* convex owing to the presence of the squared Frobenius norm). As such, it is amenable to several convex optimization solvers. For our purposes, we use the `SCS` solver [103] via its Python interface available in `cvxpy` [104]. The `SCS` solver is a general-purpose conic optimization solver: it first transforms (3.8) into an equivalent conic representation (which typically yields a constrained optimization problem) and applies an operator splitting method with algorithmic embellishments, including a variant of Anderson Acceleration [105], to solve the resulting optimization problem. The `SCS` interface provides access to several options that control its behavior during a solve; our code uses the default settings except for the relative and absolute feasibility tolerances, ϵ_{rel} and ϵ_{abs}: we set these to $\epsilon_{\mathsf{rel}} = \epsilon_{\mathsf{abs}} = 0.005$.

Model selection

We use cross-validation across BPA concentrations to select the regularization coefficients λ_1 and λ_2. Each regularization coefficient was scanned across 20 values logarithmically spaced between 0.001 and 1000. In particular, for each candidate (λ_1, λ_2) pair, we iteratively select one row of Z to act as the "held-out" community, use the remaining rows to fit $\widehat{\Theta}$ by solving (3.8) with the given (λ_1, λ_2) pair, and compute the RMSE score of the fit on the held-out community; finally, we compute the average RMSE scores over held-out communities (Fig. 3.12 bottom right panel). We select the (λ_1, λ_2) pair that achieved the lowest average

score and retrain the model using all available communities. The optimal regularization coefficients were $(\lambda_1 = 4.833, \lambda_2 = 784.760)$, and the resulting $\widehat{\Theta}$ was rank-2.

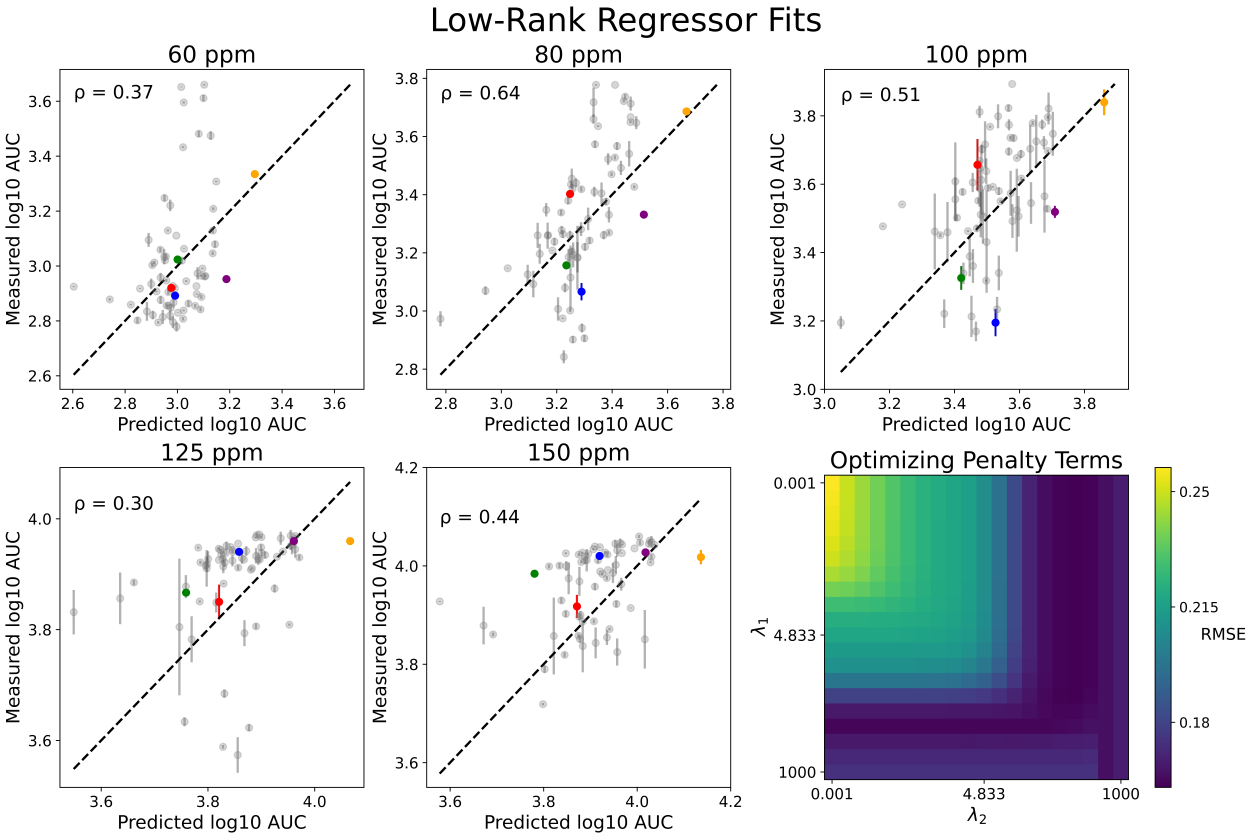


Figure 3.12: **Low-Rank Regressor Fits to BPA AUC Data** Low Rank Regressor fits to our AUC matrix (see Supplementary Information). *Top Row and Bottom Left and Center Panels* Fits for each of the five initial concentrations. Dotted black lines denote $y = x$ (perfect predictions) within each concentration. Vertical error bars represent the standard deviation of AUCs in technical replicates. *Bottom Right Panel* Total RMSE of all fits as a function of the regularization parameters λ_1 and λ_2 (Eqn. 3.8). Darker shades represent smaller RMSEs.

Fig. 3.12 shows the fit of the LRR to the five BPA concentrations. In comparison to our standard linear regressions (Fig. 3.3), within each individual concentration, the LRR slightly underperforms the linear regression. However, the LRR has two constraints not shared with the linear regression: 1) the LRR only fits complete matrices Z and X and therefore leave-one-out cross-validation can only be performed by dropping a community out of all five concentrations simultaneously versus one-by-one, 2) the low-rank nature of $\widehat{\Theta}$ restricts the

potential coefficients for all five concentrations, whereas the linear regression is free to have different parameters for each of the five concentrations. Given these restrictions, the fit of the LRR to the data is surprisingly well. However, further work is needed to enable the LRR to capture the variation between concentrations - for instance, non-polynomial functions may need to be considered to improve LRR predictions.

3.5.6 Supplementary Figures

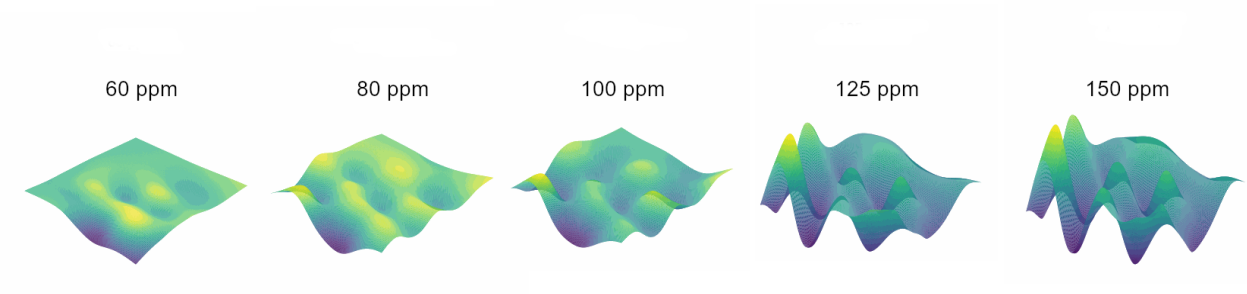


Figure S.3.1: **Depiction of the BPA Degradation Landscape Across the Five Initial Concentrations** from left to right. Surfaces are interpolated and smoothed.

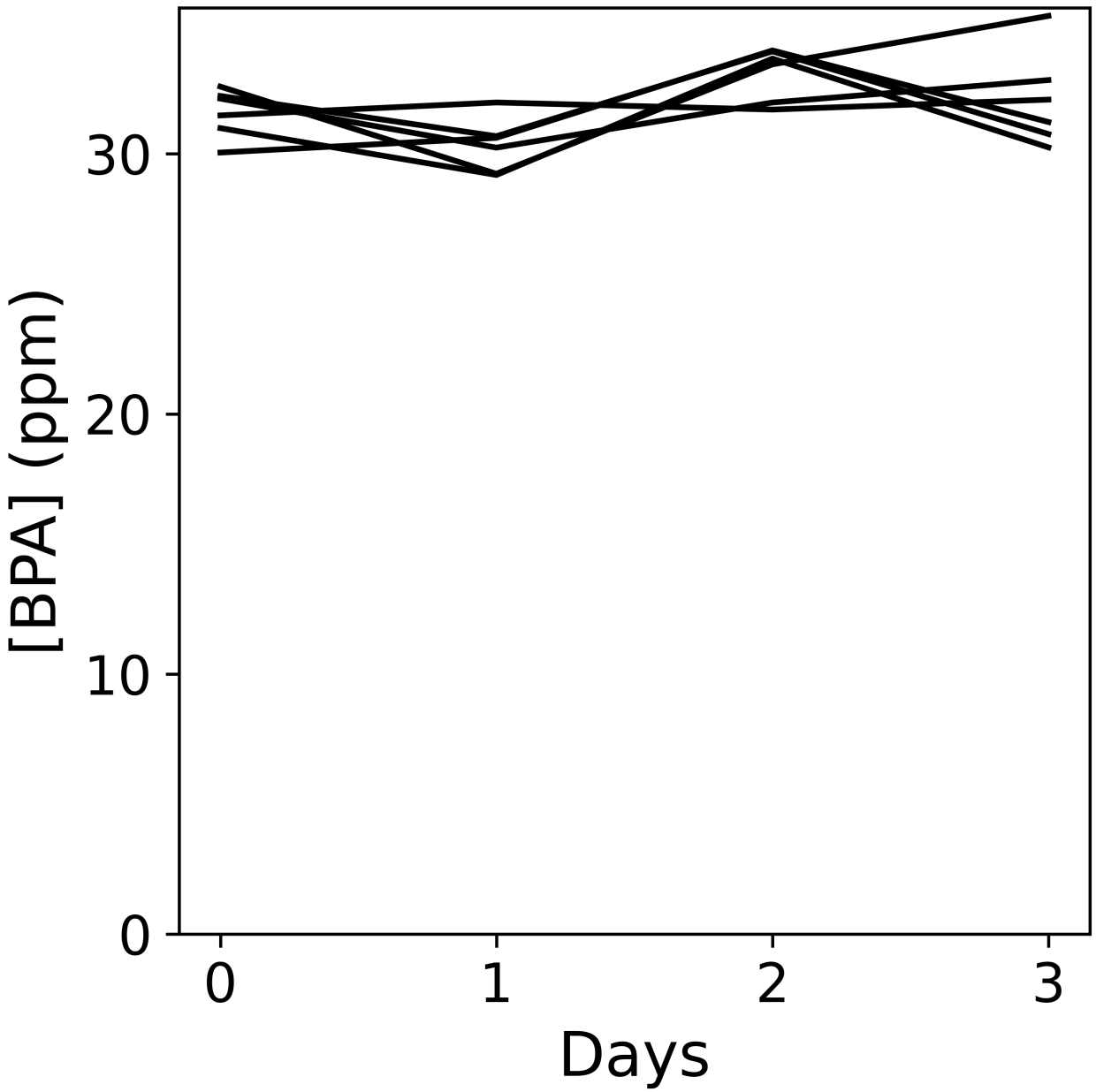


Figure S.3.2: **Random Communities from Soil Isolates Do Not Degrade BPA** 5 random communities were assembled from six Proteobacterial strains isolated from various forests and plains within Illinois. These strains have been previously characterized for denitrification [42]. Communities were inoculated in 30 ppm BPA media, and the BPA concentration was measured every 24 hours for each community as described in the Section 3.4.3.

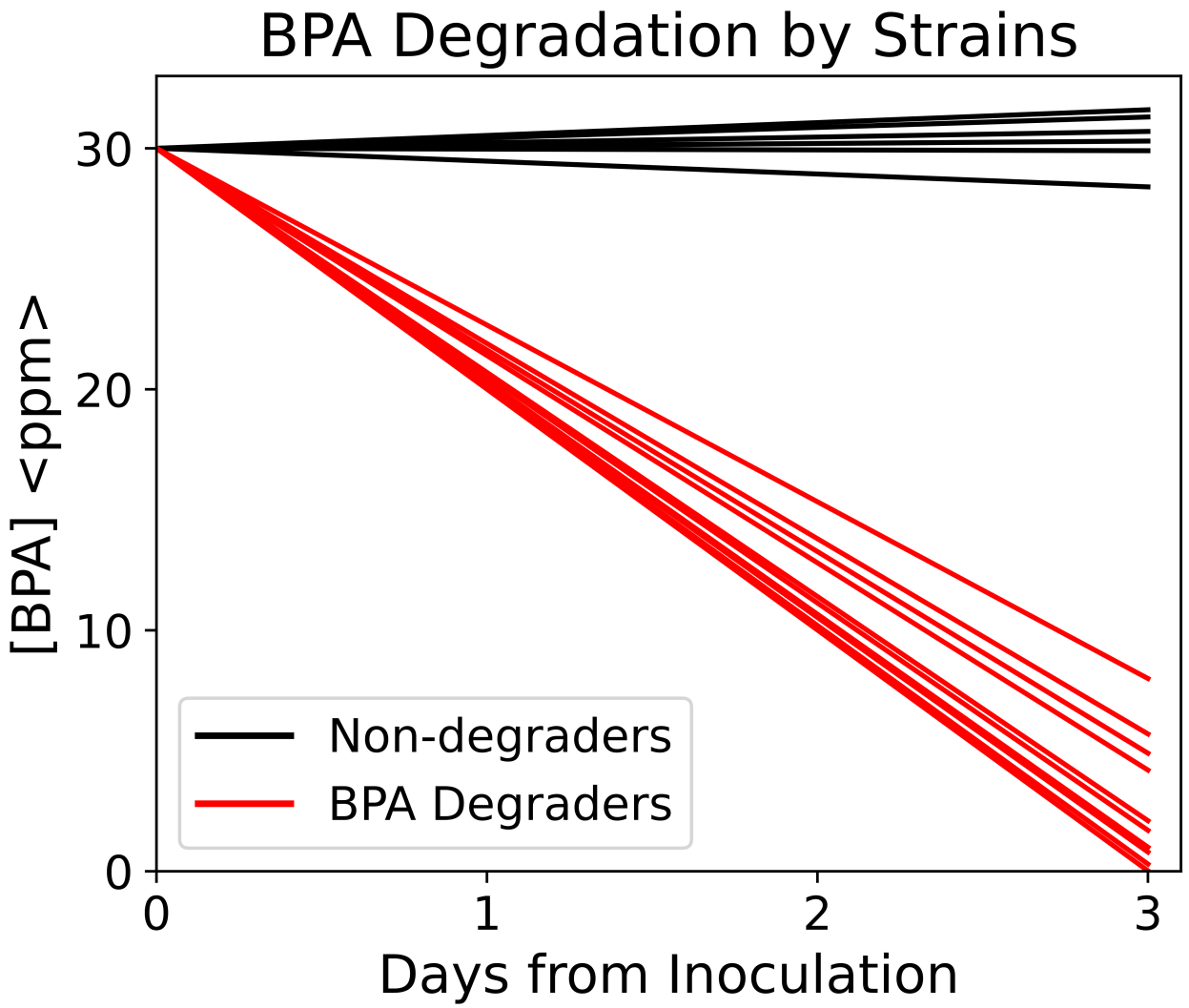


Figure S.3.3: **BPA Degradation of the isolated 16 strains from 30 ppm BPA** in three days. Red lines show strains that degraded BPA, black lines show strains that did not.

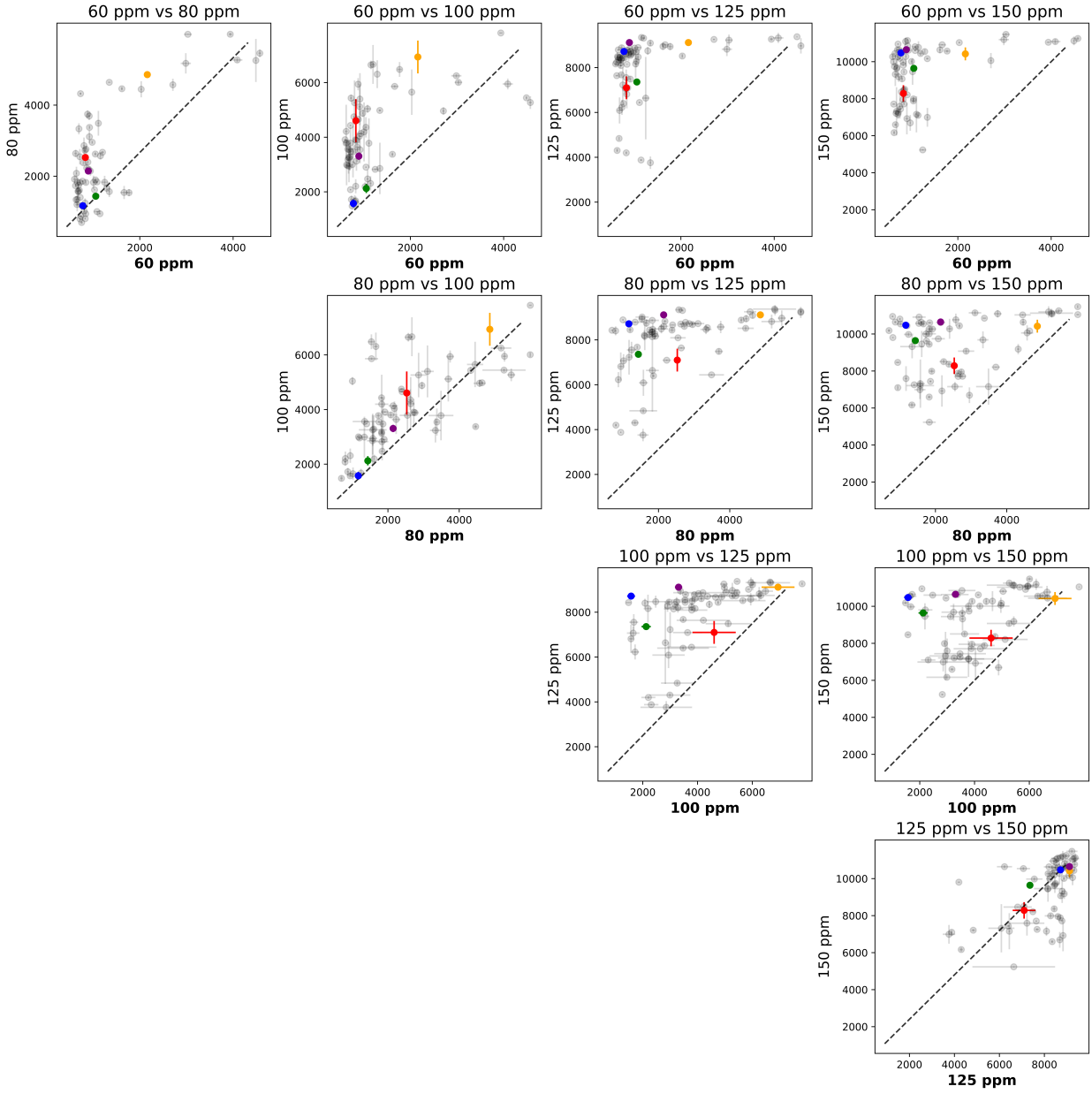


Figure S.3.4: **Comparing Degradation AUC Across the 5 Initial Concentrations** Comparing the AUCs of our 70 communities across the five different initial concentrations. Error bars represent the standard deviation of technical replicates. Dotted lines indicate normalized equivalent AUC.

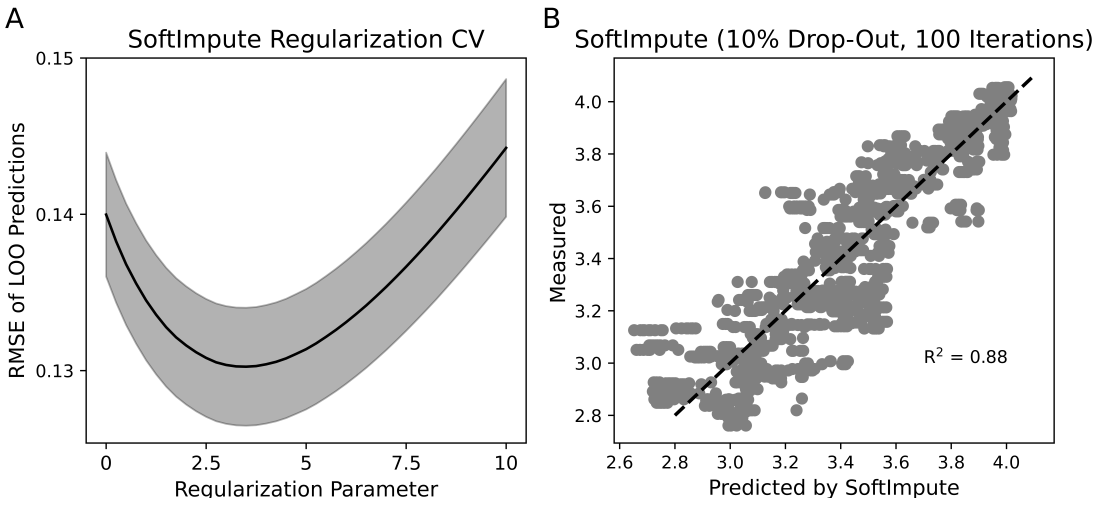


Figure S.3.5: **SoftImpute Optimization and Further Validation** **(A)** Root Mean Squared Error of Leave-One-Out predictions as a function of the `SoftImpute` regularization parameter (see Section 3.4.4). Regularization parameters were tested between 0 and 10 in increments of 0.25, and R^2 was calculated from the complete leave-one-out predictions of the AUC matrix. The optimal regularization of 3.5 was chosen for all imputations. Shaded region indicates the standard error of all folds. **(B)** `SoftImpute` predictions from randomly dropping 10% of community AUCs. Random drops were performed 100 times, and all resulting predictions are plotted. Dotted line represents y=x. $R^2 = 0.88$, p-value < 0.01 as computed by the `scikit-learn` function `r2_score`.

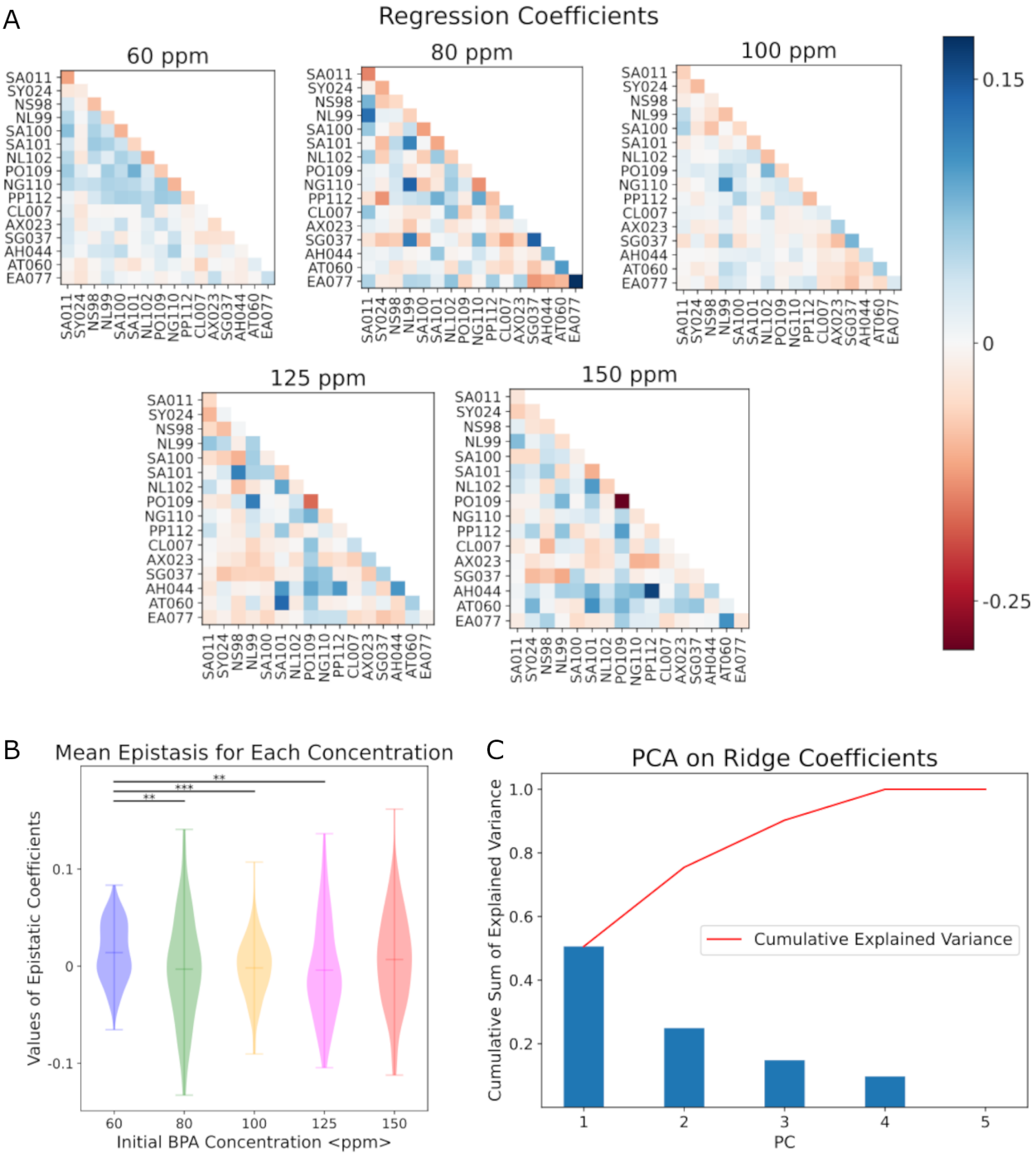


Figure S.3.6: **Analyzing Regression Coefficients to Understand the Functional Landscape**

Figure S.3.6: (Continued from previous page) **(A)** Heatmaps of regression coefficients from each of the five regression models (Fig. 3.3)B. Diagonal entries represent additive (β_i) coefficients, while off-diagonal entries represent epistatic ($\gamma_{i,j}$) coefficients. **(B)** Violin plots of epistatic coefficients ($\gamma_{i,j}$ terms) for each initial BPA concentration. More positive values indicate more antagonistic epistatic interactions (one strain inhibiting another), while more negative values indicate more cooperative epistasis (improving BPA degradation). Horizontal lines within distributions represent the mean coefficient. Stars indicate p-values for statistical tests of differences in average epistasis via two-sided t-test: $^{**}p < 0.01$, $^{***}p < 0.001$. At 150 ppm, the mean epistasis is not significantly different from 60 ppm, but the variance of the distribution of epistatic terms is larger (Fig. 3.4D). **(C)** PCA on Regression Coefficients show that the first two principal components explain 75% of the variation between coefficients of the five models, while the first four PCs explain all of the variation.

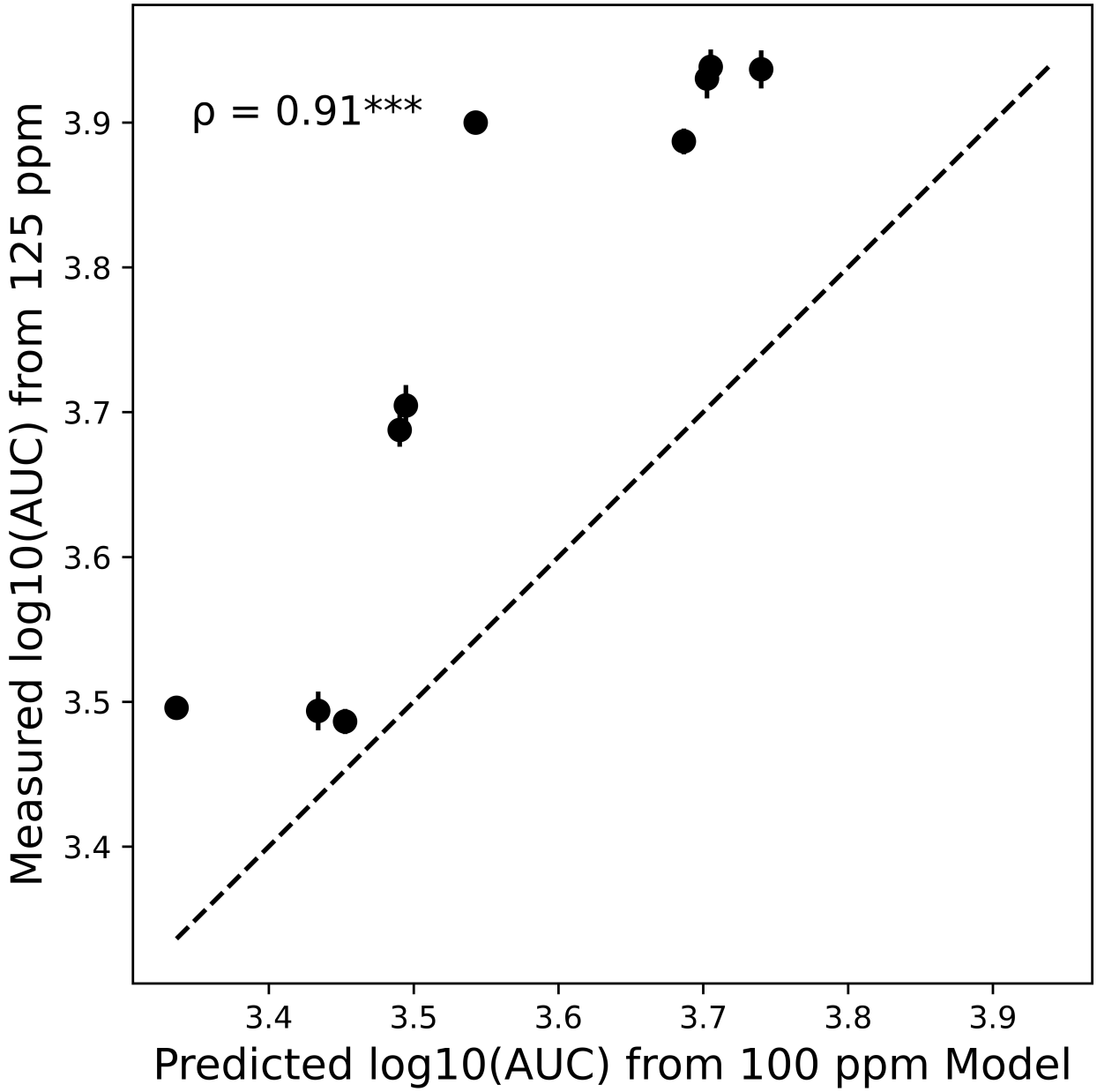


Figure S.3.7: **Smooth Variation Between Landscapes Enable Cross-Predictions** Scatter plot of predictions from the model trained on data from 100 ppm (x-axis) versus measured BPA degradation of the additional ten communities from 125 ppm (y-axis). Error bars represent the standard deviation of technical replicates. Dotted line shows $y = x$, perfect predictions. $^{***}p < 0.001$ from 1000 bootstrapping iterations rejects the null hypothesis that the Pearson's correlation coefficient is <0

CHAPTER 4

RATIONAL DESIGN OF A SYNTHETIC COMMUNITY THAT SUPPRESSES A MULTI-DRUG RESISTANT HUMAN GUT PATHOGEN

4.1 Introduction

The Human Gut Microbiome is a large network of microbial organisms that heavily affect the health of the human host, through processes such as vitamin production [106], nutrient metabolism [107], and immune regulation [46]. These functions primarily arise through large-scale interactions between strains within the microbiome as well as with the human host [41, 108, 109]. In particular, the human gut microbiome is essential for the sequestration of multi-drug resistant (MDR) invasive pathogens [110]. These pathogens often exist at low abundances in healthy gut microbiomes [111]; however, upon exposure to an antibiotic treatment which kills the healthy microbiome, these pathogens are left to proliferate freely [34], inducing diseases that often do not have known reliable treatments [33, 111]. This strongly implicates the role of a healthy microbiome in mitigating pathogen suppression [112, 113]. However, while a number of efforts have successfully created functional communities for pathogen suppression [45], developing a generative model (one that can reliably predict the functional activity of unseen communities) has remained elusive.

One such MDR pathogen is the *Klebisella pneumoniae*, a gram-negative proteobacteria that causes pneumonia and can lead to systemic infections if left unchecked [111, 114]. Fecal microbiota transplantation (FMT) has been shown to alleviate the symptoms of a *K. pneumoniae* infection, although inconsistently [33]. Furthermore, the undefined nature of microbiomes within successful FMTs make it impossible to learn what enables their effectiveness over unsuccessful attempts [34].

To address this, we leverage the community function landscape to implement a DBTL

approach for optimizing *K. pneumoniae* by microbial consortia (Fig. 1.1C). By initially learning from 96 synthetic consortia built from 46 human gut microbiome isolates, we created a generative model for *K. pneumoniae* suppression that captures high epistatic interactions within the communities. Statistical inference performed on our model identified a set of 15 key strains that when combined into a community ('SynCom15') sustainably suppressed *K. pneumoniae* across various diverse in-vitro environments, and could not be obviously deconstructed. We show that SynCom15 is capable of clearing *K. pneumoniae* infections in an *in-vivo* mouse model and that interactions between members enabled engraftment into the mouse model, showing the broad potential of the super community. Our work describes a potentially therapeutic, sparse synthetic microbiome made of human gut commensal bacteria for treatment of MDR *K. pneumoniae* infections and, more generally, displays the effectiveness of the community function landscape in designing a core community with optimal performance both *in-vitro* and *in-vivo*

4.2 Results

4.2.1 Designing Communities for the Clearance of K. pneumoniae

To begin, we first isolated and whole-genome sequenced 848 gut commensal strains from fecal samples of healthy donors (see 4.4.1). Since the space of all possible communities is 2^{848}, we chose 46 strains to constitute our reduced strain bank (Tbl. 4.2). This number was chosen due to the practical limitations of experimental design. The 46 strains were chosen in such a way as to represent the genetic diversity of the original 848 strains (Fig. 4.5A&B).

Unlike the degradation of BPA, gut pathogen clearance requires extensive microbiome diversity [13, 45, 110]. This function, also known as colonization resistance, is known to arise from nutrient competition, niche exclusion, and host immune system modulation, among many other factors. Therefore, randomly designed communities with no attention to func-

tional diversity are likely to miss the ecological functions necessary for *K. pneumoniae* suppression. Thus, for our first set of communities, we sought to design communities comprised of a diverse set of strains, rather than a set of strains that were closely related to each other, to maximize the potential for functional diversity of a given community. Therefore, we used the UMAP space of the 46 strains to design diverse communities (See 4.4.2). In brief, for a chosen community size N, $10,000$ communities were randomly generated, and all pairwise distances (dispersal) between constituent strains were computed based on their UMAP distances. After considering the 30% of pairs with the shortest dispersal for all communities, the community with the maximum mean dispersal was chosen to be built. By choosing the community with the maximal mean dispersal of the most closely related strains (i.e. the lowest 30%), this algorithm enforces the constraint of diversity across the whole community. We labeled these consortia as 'Designed Microbial Communities' (DMCs), as they were designed to maximize functional diversity.

Using this approach led us to a total of 96 DMCs for our first round of design: 92 diverse communities, three replicates, and one community containing all 46 strains (Fig. 4.1A). These communities had a mean richness of 15 (with a standard deviation of 5), but ranged from two-member communities all the way to the DMCs of all 46 strains. All consortia were tested for their ability to suppress *K. pneumoniae* MH258 — an MDR strain isolated from a patient sample obtained from Memorial Sloan Kettering Hospital (MH) representative of the epidemic multilocus sequence type (ST) 258 clone harboring the bla_{KPC}-encoded carbapenemase. We chose this strain to use as our target for suppression because it was amongst the most multi-drug resistant strains that have been previously characterized, exhibiting resistance against a diversity of antibiotics [115]. DMCs were co-cultured with a GFP-tagged MH258 in Brain-Heart-Infused media with cysteine (BHIS) for 120 hours in an anaerobic chamber. The abundance of *K. pneumoniae* during co-culture with consortia was quantified through time by plating (See Section 4.4.3).

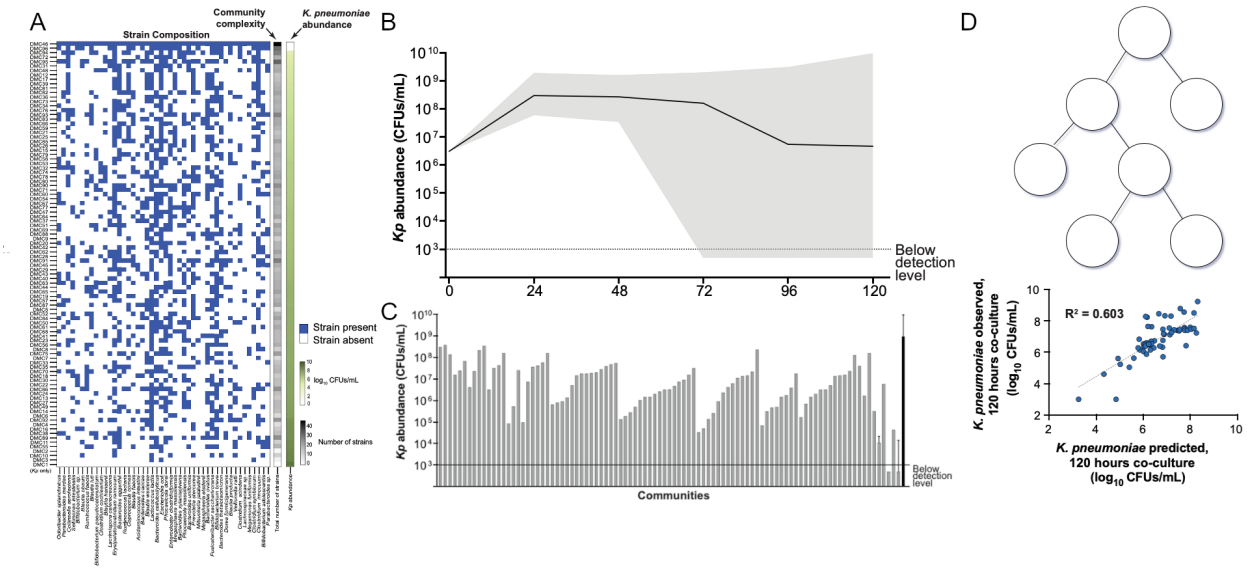


Figure 4.1: **Creating a Generative Model for the Suppression of *K. pneumoniae*** (**A**) Engineered consortia (rows, labeled as 'DMCs - Designed Microbial Communities') described by strain composition (columns). Blue pixels mean that strain is included in designed community; white pixels mean that strain is not included in the designed community. Each row is labeled by the number of strains within the DMC ('Community complexity') and the *K. pneumoniae* abundance after 120 hours of co-culture in BHIS media ('*K. pneumoniae* abundance'). Rows are ordered by their ability to suppress *K. pneumoniae* after 120 hours of co-culture. 'DMC1', the last row, is *K. pneumoniae* in monoculture ('Kp only'). (**B**) Timecourse of *K. pneumoniae* abundance (y-axis) for all 96 communities shown in **A**. Solid line represents median, shade represents range. (**C**) *K. pneumoniae* abundance (y-axis) after co-culture with each DMC (x-axis) for 120 hours where the x-axis is ordered by least complex (left) to most complex (right) DMCs. The black bar is *K. pneumoniae* grown in monoculture in BHIS. (**D**) (*Top*) A RandomForest model model was trained and validated to predict *K. pneumoniae* abundance after 120 hours of co-culture using only the designed strain presence-absence matrix in **A** as input. 60 new communities were generated from the model. (*Bottom*)*K. pneumoniae* abundance predicted by RF model for the new 60 new consortia (x-axis) versus *K. pneumoniae* abundance observed after 120 hours of co-culture with the 60 new communities (y-axis).

We found that across all DMCs, *K. pneumoniae* grew for the first 24 hours from an abundance between 10^6 and 10^7 to an abundance of 10^8 on average and remained constant through the next 24 hours (Fig. 4.1B). After the first 48 hours of co-culture and up to 120 hours, the 96 consortia reproducibly exhibited a range of capacity to suppress *K. pneumoniae* spanning no suppression to suppression greater than seven orders of magnitude equivalent to clearing *K. pneumoniae* given the lower limit of detection for our assay (Fig. 4.1A&B). The DMC containing all 46 strains ('DMC46') suppressed *K. pneumoniae* the most, while *K. pneumoniae* cultured alone maintained the highest abundance. Moreover, we found that the suppressive capacity of DMCs was unrelated to the size of community composition (Fig. 4.1C). This result suggested it was likely not mere resource competition that mediated the suppression of *K. pneumoniae*, but rather a complex set of microbial interactions.

4.2.2 A RandomForest Regressor Infers the Landscape of K. pneumoniae Suppression and Enables Community Design

The complex interactions seemingly necessary for *K. pneumoniae* clearance is highly indicative of an epistasis-dominated function, potentially more-so that of BPA degradation (See Chapter 3). Indeed, a regression model of the form Eqn. 2.1 when fit in the same method described in 2.2.1 truncated at second-order terms struggles to accurately describe the data (See Section 4.5.1). To circumvent this problem, we turned to a nonlinear regression that captures complex interactions by nature. We relied on the RandomForest (RF) Regressor - a regressor that utilizes decision trees to capture nonlinear interactions between members of a consortium [116–118]. The inputs for the RF model remained the presence and absence of all 46 species for each consortium (Fig. 4.1D). Note that with this approach, the convention of $x_i = \pm 1$ complicates the interpretability of the RF regressor, and thus we defaulted to the typical $x_i = [0, 1]$ convention. For the functional response variable, we used the end-point log-transformed *K. pneumoniae* abundance. The RF model was trained on 90% of the

data and validated on the remaining 10% 100 times for bootstrap support, resulting in a validation R^2 value of 0.98 (4.4.4).

Using the RF model, we created 60 new consortia spanning different membership sizes that were predicted to span a large dynamic range of *K. pneumoniae* clearance in our assay. We compared the abundance of *K. pneumoniae* for the 60 new consortia as predicted by our RF model versus the *K. pneumoniae* abundance we experimentally observed after co-culture of each of the 60 new consortia with *K. pneumoniae* for 120 hours. We found that our RF model was predictive of the resulting *K. pneumoniae* abundance to an R^2 value of 0.6 ($p < 10^{-3}$) (Fig. 4.1D bottom).

Collectively, our results showed that our RF model could accurately predict the capacity of a complex microbial community defined by our 46 strains to suppress *K. pneumoniae*, thereby enabling engineering of new communities with desired suppressive capacity. Thus, in a single round of DBTL where the first round of design was constrained by genomic diversity of strain combinations, we derived a generative model of community design for suppressing *K. pneumoniae* in BHIS media relying only on strain presence/absence, despite the epistatic nature of the community function.

4.2.3 Community Structure Reveals 15-member Group Essential ('SynCom15') for K. pneumoniae Suppression

The structure of the RF model is designed to highlight individual features important for predictions. The feature importance scores derived from the RF model are continuous, and therefore unable to delineate groups of important strains (Fig. S.4.2A). Furhtermore, unlike a linear regression approach, we cannot explicitly isolate the impact of any one strain (or even groups of strains) on the target function. Thus, to understand the community structure underlying the functional landscape and enabling *K. pneumoniae* clearance, different approaches were needed.

To determine the extent of ruggedness of the RF-inferred landscape, we relied on a simple measure for epistasis for RF models: comparing the cross-validation prediction performance of the RF to a linear regression of additive terms alone. In theory, an increased predictive performance by the RF compared to an additive linear model indicates that the RF is capturing higher-order behavior that is 'hidden' from the additive model [119]. We fit an additive linear regression in the same method described in Section 2.4.2 and compared the R^2 of validation and of the out-of-sample DMCs to the RF (Tbl. 4.1). The large improvement by the RF over the linear additive model shows that the landscape of *K. pneumoniae* is highly epistatic (see further discussion in Section 4.5.1).

	RandomForest Regression	Additive Linear Regression
Validation R^2	0.98	0.44
Out-of-Sample R^2	0.6	0.46

Table 4.1: **RandomForest Improves over Linear Additive Model**

To define the critical strains responsible for clearing *K. pneumoniae*, we used statistical co-variation as a proxy for identifying key interactions between strains. This approach has been used to identify collectives across different scales of complexity: groups of stocks defining economic 'sectors', groups of amino acids defining functional units of proteins ('protein sectors'), and groups of microbes within microbiomes defining covarying units of therapeutic importance ('ecogroups') [44, 120–122].

We generated 100,000 consortia *in-silico* and predicted their capacity to suppress *K. pneumoniae* after co-culture using our RF model. We then selected the set of consortia predicted to suppress *K. pneumoniae* at least five orders of magnitude (log_{10} CFUs/ml < 5 at end point, Fig. 4.2A&B), resulting in 5,752 consortia. We took the matrix of strain presence/absence for all synthetic consortia and performed Principal Components Analysis (PCA), yielding 46 principal components (PCs) of data-variance. We regressed the contribution of each of the 5,752 consortia onto each PC against the predicted *K. pneumoniae* abundance to identify PCs that most associate with *K. pneumoniae* suppression in a data-

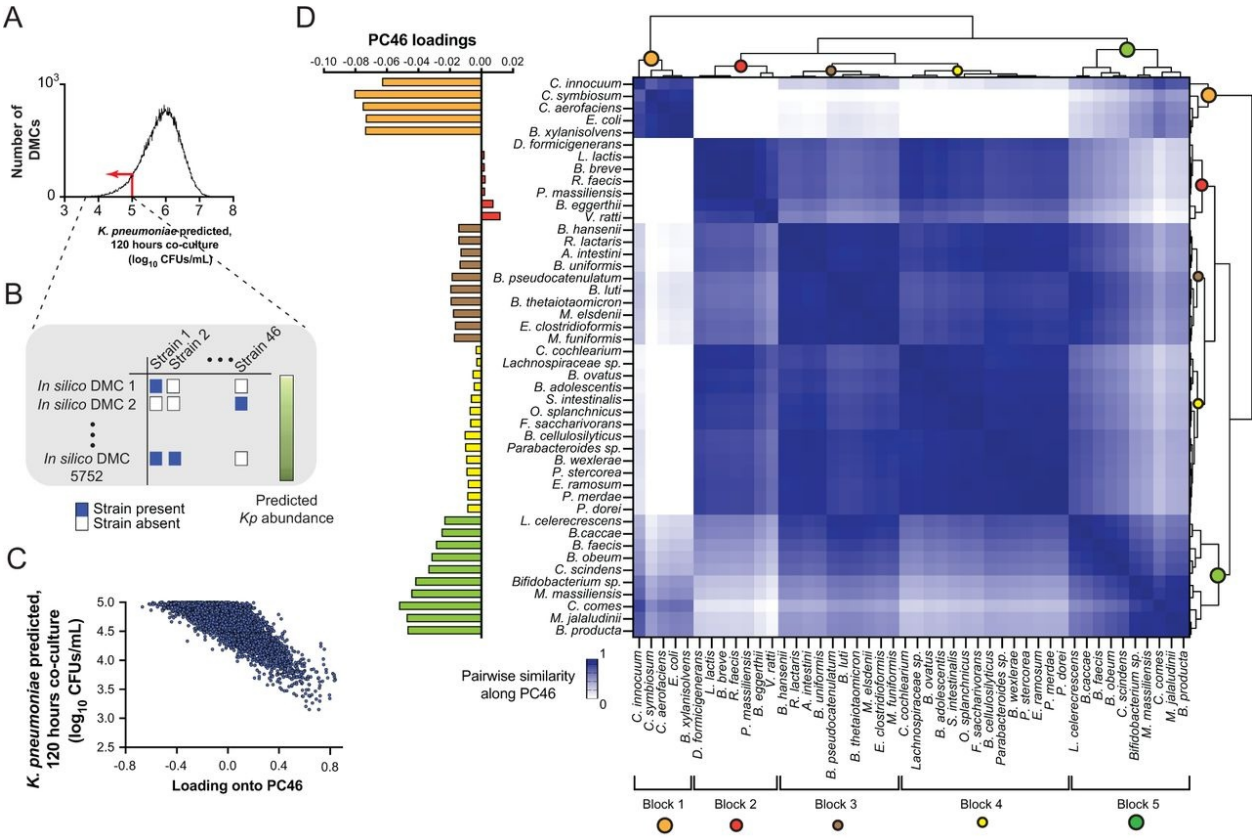


Figure 4.2: **Statistical Analysis Reveals Blocks of Strains Correlated with *K. pneumoniae* Clearance** (**A**) Histogram of predicted *K. pneumoniae* abundance for 100,000 *in-silico* generated consortia. Red arrow is the chosen threshold of predicted *K. pneumoniae* suppression of 10^5; communities to the left of the arrow for further analysis. (**B**) The matrix of strain presence/absence for all 5,752 *in-silico* consortia, labeled with the predicted *K. pneumoniae* abundance after 120 hours of co-culture (green bar). (**C**) Contribution of each of the 5,752 consortia onto the 46th principal component (PC46) of the matrix in panel B (x-axis) versus predicted *K. pneumoniae* abundance associated with each community (y- axis). (**D**) *Left* Contribution of each strain onto PC46. *Right* Hierarchically clustered strain-strain matrix where each entry is the similarity in contribution onto PC46 between two strains. Blocks 1 through 5 are defined according to the clustering pattern (colored dots in dendrogram). Bars in left panel are colored according to which block each strain belongs

driven, unbiased manner. We found that PC46, containing <0.1% data-variance, was the most associated with *K. pneumoniae* abundance (Fig. 4.2C) (See Section 4.4.5, Tbl. 4.3).

Similar to the distribution of RF importance scores, the contribution of strains onto PC46 was continuous, precluding the ability to define groups of strains to construct communities (Fig. 4.2 left panel). Interestingly, the contribution of strains onto PC46 did not resemble the

distribution of feature importance scores from the RF model (Fig. S.4.2B), suggesting that PC46 contained information that was different from the feature importance scores (Table 4.3). To use the information in PC46 to define groups of statistically interacting strains, we computed the statistical similarity between all pairs of strains on PC46 (See Section 4.4.6). The concept behind this measure is that two strains that significantly contribute to PC46 and are close together along PC46 are, on average, co-present in consortia predicted to suppress *K. pneumoniae*. Hierarchical clustering of the pairwise similarity between strains illustrated a distinct block structure amongst five separate groups (Fig. 4.2). Five strains that contributed the most to defining PC46—Clostridium innocuum, Clostridium symbosium, Collinsella aerofaciens, Escherichia coli, and Bacteroides xylanisolvens—formed a group that we term 'Block 1' (Fig. 4.2D). The following ten strains that contributed to PC46—Lacrimispora celerecrescens, Bacteroides caccae, Blautia faecis, Blautia obeum, Clostridium scindens, a Bifidobacterium species, Megasphera massiliensis, Coprococcus comes, Mitsuokella jalaludinii, and Blautia producta—formed a group that we term 'Block 5' (Fig. 4.2D). Blocks 1 and 5 exhibited collective similarity amongst each other; we term this group of strains 'SynCom15'—a 15-member group comprised of statistically interacting strains that are co-present in communities predicted to clear *K. pneumoniae*. In contrast to SynCom15, three other groups of strains were statistically inferred to be co-absent in communities predicted to clear *K. pneumoniae*. These groups were comprised of 7, 10, and 14 strains; we term these groups 'Block 2', 'Block 3', and 'Block 4' respectively (Fig. 4.2D).

Interestingly, of our initial 96 designed DMCs, only two communities contained all 15 strains identified within SynCom15 - DMC46 (the community of all strains) and DMC96 (a 28-member community, our third largest). These are also the only communities that suppressed *K. pneumoniae* to levels below detection limit (Fig. 4.1A-C). This validation further signifies the importance of these strains for the suppression of *K. pneumoniae*.

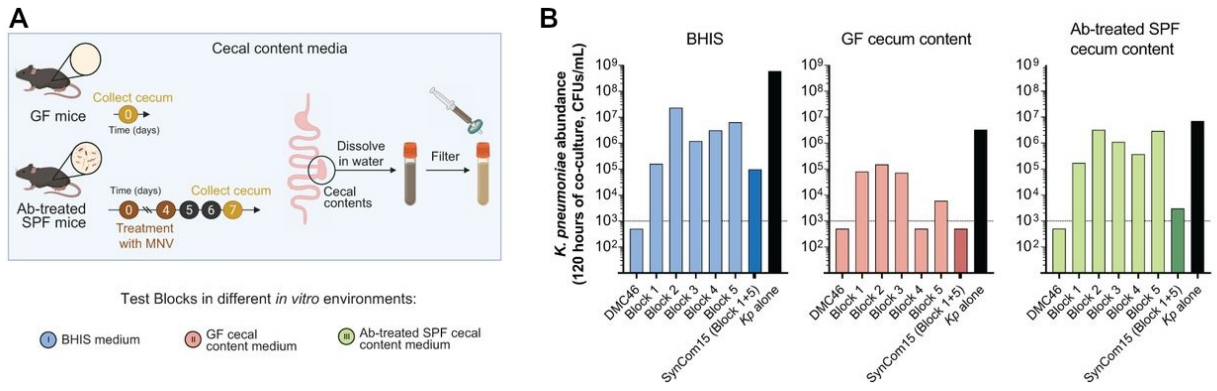


Figure 4.3: **SynCom15 Clears *K. pneumoniae* in Multiple *in-vitro* Environments** (**A**) Workflow for creating cecal extract media from germ-free ('GF') and antibiotic treated specific pathogen free ('Ab-treated SPF) mice. (**B**) *K. pneumoniae* abundance (y-axis) for DMC46, all Blocks, and SynCom15 (darker shade) after 120 hours of co-culture in BHIS (blue), GF cecal extract media (salmon), and Ab-treated SPF cecal extract media (green). *K. pneumoniae* abundance after 120 hours of monoculture ('Kp alone') in each media is shown in black. Dashed line is detection limit of assay.

4.2.4 SynCom15 Clears K. pneumoniae in Different In-Vitro *Environments*

SynCom15 was discovered by identifying key blocks of strains that co-varied in their presence in consortia that successfully clear *K. pneumoniae*. We wanted to test whether these 15 strains alone are sufficient to suppress *K. pneumoniae* growth. In addition, because these strains represent a core ecological group necessary for function, we hypothesized that SynCom15 would be able to suppress *K. pneumoniae* in environments beyond the one our RF model was trained on (BHIS media). Furthermore, since SynCom15 was composed of Blocks 1 and 5 (Fig. 4.2D), we wanted to see whether both blocks were needed to maintain function, or if only one of the block was sufficient.

We built and tested SynCom15 as well as all other Blocks individually for their capacity to clear *K. pneumoniae* across three unrelated media conditions: BHIS, media created from the cecal extracts of germ-free (GF) mice, and media created from the cecal extracts of specific-pathogen-free (SPF) mice treated with broad spectrum antibiotics (Ab-treated SPF) (Fig. 4.3A) (See Section 4.4.8). As a comparator to SynCom15 and the other Blocks,

we also tested DMC46 — the community of all strains that suppressed *K. pneumoniae* the most in BHIS media. Notably, our results clearly illustrated the environmental dependence of community efficacy. Blocks 2 and 3 were consistently ineffective at suppressing *K. pneumoniae* across environments while Blocks 1, 4 and 5 were able to suppress *K. pneumoniae* depending on the environment in which they were tested: Block 1 was effective in BHIS and Blocks 4 and 5 were effective in GF cecal extract. Thus, Blocks 1, 4 and 5 were all conditionally effective. In contrast, we found that DMC46 and SynCom15 suppressed *K. pneumoniae* across all three conditions tested. DMC46 cleared *K. pneumoniae* across all environments. SynCom15 suppressed *K. pneumoniae* five orders of magnitude in BHIS, cleared *K. pneumoniae* in GF cecal extracts, and suppressed *K. pneumoniae* greater than four orders of magnitude in Ab-SPF cecal extracts. (Fig. 4.3B).

At face value, it seems that the success of SynCom15 is trivial given the success of DMC46. However, we highlight that the significance of SynCom15 isn't purely in its ability to suppress *K. pneumoniae*. Rather, a statistical analysis of the functional landscape from a single round of DBTL led directly to the design of SynCom15, which performs just as well as DMC46. Thus, our strategy of statistical inference performed on the RF model of community design defined SynCom15—a phylogenetically diverse 15-member community—that suppressed *K. pneumoniae* across diverse environmental contexts in a manner similar to DMC46 — the community containing all 46 strains.

4.2.5 SynCom15 Clears K. pneumoniae in Clinically Relevant Mouse Model of Infection

Because DMC46 and SynCom15 were unconditionally effective at clearing *K. pneumoniae* in-vitro, we sought to test the ability of both communities to clear *K. pneumoniae* in a more complex, clinically relevant environment. We evaluated the efficacy of DMC46 and SynCom15 in a mouse model of infection. To mimic a clinically relevant scenario, we did

not use germ-free mice (mice without a microbiome). Rather, we treated SPF mice with broad spectrum antibiotics to deplete their gut microbiota then infected them with *K. pneumoniae*—a sequence of events commonly encountered in patients who acquire MDR *K. pneumoniae* infection (this mechanism - clearance of the healthy gut microbiome followed by *K. pneumoniae* colonization - is the primary mechanism by which *K. pneumoniae* occur in humans [111, 114]). Additionally, we singly-housed mice to ensure that no sharing of microbes by coprophagia amongst animals would affect microbiome composition during and post-antibiotic treatment [123]. Singly-housed antibiotic-treated SPF mice infected with *K. pneumoniae* MH258 were given either (i) saline (PBS) as a negative control, (ii) a heterologous whole stool transplant derived from mice ('Fecal Microbial Transplant', FMT) as a positive control [111], (iii) Block 1, (iv) Block 2, (v) DMC46, or (vi) SynCom15 as interventions for three sequential days after infection (see Section 4.4.9). Blocks 1 and 2 were given as bacterial communities that were either conditionally efficacious across in-vitro conditions or unable to clear *K. pneumoniae* across any in-vitro condition respectively. Fecal samples were collected and *K. pneumoniae* abundances were tracked through the course of the experiment by plating (Fig. 4.4A).

We found that Block 1 and Block 2 did not suppress *K. pneumoniae* relative to saline. DMC46 suppressed *K. pneumoniae* by up to six orders of magnitude, matching the suppression caused by the FMT control but with slower kinetics of response. In contrast, SynCom15 rapidly suppressed *K. pneumoniae*, outperforming the FMT control in rate of suppression. Additionally, SynCom15 cleared *K. pneumoniae* four days after the last gavage and maintained clearance through nine days after the last gavage (Fig. 4.4B). These results highlighted the rapid and sustained efficacy of SynCom15 in clearing *K. pneumoniae* in-vivo as well as the utility of reducing the community size from the 46 strains defining DMC46 to the inferred key 15 strains defining SynCom15.

Taxonomic profiling of fecal samples procured through the experiment revealed that 10 of

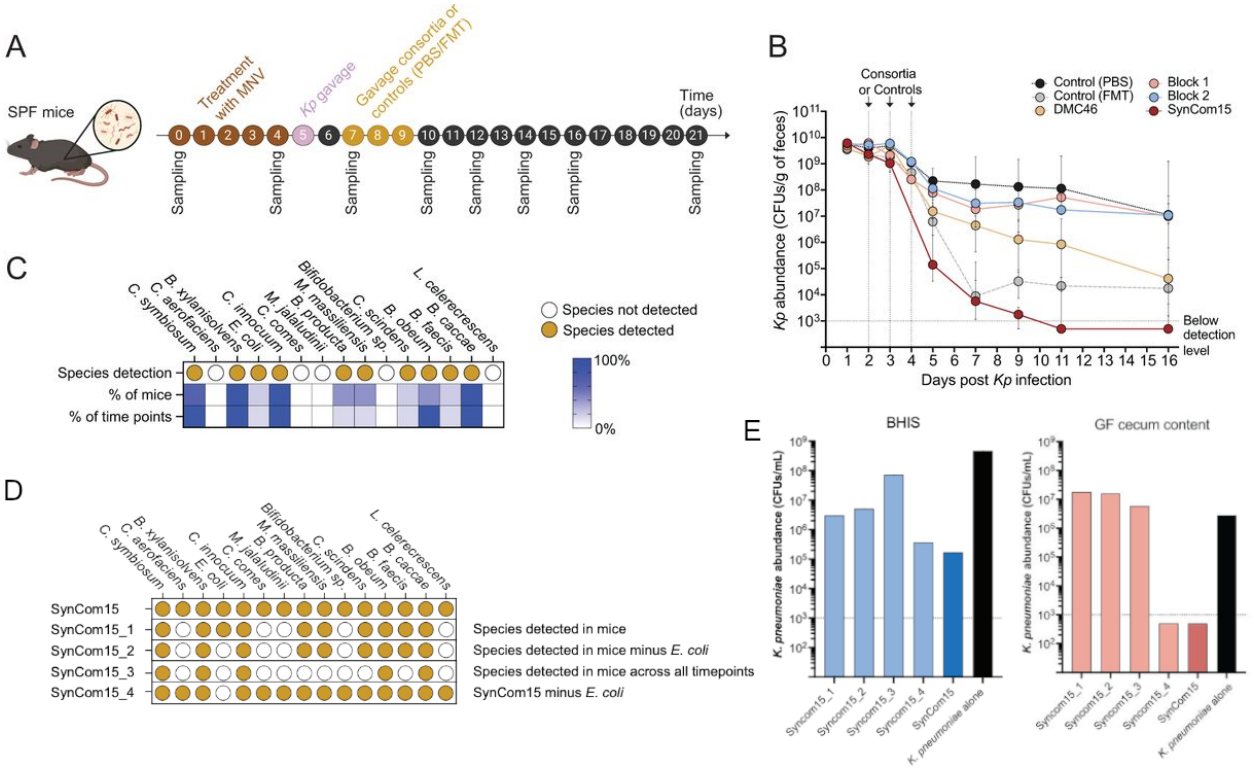


Figure 4.4: **SynCom15 Sustainably Clears *K. pneumoniae* in a Clinically Relevant Mouse Model of Infection** (**A**) Specific pathogen free (SPF) mice are treated with metronidazole, neomycin, and vancomycin (MNV) (brown), then infected with *K. pneumoniae* ('Kp gavage', pink), then given either a mouse fecal microbial transplant (FMT), saline (PBS), DMC46, Block 1, Block 2, or SynCom15 (beige). Fecal samples are collected at select days delineated in the schematic as 'Sampling'; mice are sacrificed after day 21. (**B**) Median fecal abundance of *K. pneumoniae* (y-axis) versus time (x-axis). Vertical dashed lines on days 2, 3, and 4 reflect gavage of bacterial communities or controls ('Consortia or Controls'). Error bars indicate interquartile range. (**C**) Engraftment statistics and relative presence of SynCom15 strains through the experiment. (**D**) Composition of (i) SynCom15, (ii) strains of SynCom15 that engrafted mice ('SynCom15_1'), (iii) strains of SynCom15 that engrafted mice without E. coli ('SynCom15_2'), (iv) strains of SynCom15 found in mice across all timepoints of the experiment shown in panel **C** ('SynCom15_3'), (v) SynCom15 without E. coli ('SynCom15_4'). (**E**) All communities defined in panel **D** were assayed for *K. pneumoniae* clearance in BHIS (left) and GF cecum content (right) media. Black bars indicate *K. pneumoniae* abundance in BHIS and GF cecum content when cultured alone.

the 15 strains in SynCom15 engrafted in at least one of the mice within the cohort (Fig. 4.4C) (see Section 4.4.11). Dynamics of SynCom15 strains showed that 5 of the 10 strains were present at detectable fractional abundances throughout the course of the experiment—C. symbiosum, B. xylanisolvens, C. innocuum, B. obeum, and B. caccae. Together, these results illustrated that the engraftment and strain dynamics of SynCom15 in mice did not follow obvious phylogenetic trends.

Given the surprising success and safety of SynCom15, we sought to further characterize its compositional content. First, we tested each strain of SynCom15 individually for its ability to suppress *K. pneumoniae* in BHIS. We found that no individual strain suppresses *K. pneumoniae* greater than two orders of magnitude and eleven of the strains suppressed *K. pneumoniae* only up to one order of magnitude (Fig. S.4.1). Moreover, the four strains that suppressed *K. pneumoniae* two orders of magnitude were found in Block 1, a Block that suppressed *K. pneumoniae* comparable to SynCom15 in BHIS but was less efficacious by several orders of magnitude in other environments without the addition of the other ten strains comprising SynCom15. Thus, including the eleven strains that have only a modest individual effect on suppressing *K. pneumoniae* in BHIS media was important for achieving the suppressive capacity of SynCom15 in other environments. These findings highlight the complex nature of the ability of SynCom15 to suppress *K. pneumoniae* across environments.

Next, we interrogated whether data from our mouse experiment could inform which strains of SynCom15 are important for functionality. We built two communities—(i) a community constituting strains that consistently engrafted the mice (10 species) and (ii) a community constituting strains that were consistently detected in mice across all timepoints (5 species) (Fig. 4.4D). The first community suppressed *K. pneumoniae* two orders of magnitude in BHIS and did not suppress *K. pneumoniae* in GF cecal extract media; the second community suppressed *K. pneumoniae* one order of magnitude in BHIS and did not suppress *K. pneumoniae* in GF cecal extract media (Fig. 4.4E). Thus, the inclusion of strains

constituting SynCom15 that were not statistically detectable in the mouse fecal pellets was important for achieving the clearance of *K. pneumoniae* we observed across environments.

Recent results have claimed the critical importance of E. coli in clearing *K. pneumoniae* [110]. This motivated us to test the importance of our E. coli strain for SynCom15. We therefore built two more communities—SynCom15 without E. coli and the community comprised of strains that engrafted the mouse without E. coli (Fig. 4.4D). Removing E. coli from either community resulted in a decrease in *K. pneumoniae* suppression by just half an order of magnitude in BHIS and no difference in suppression in GF cecal extract media (Fig. 4.4E). Additionally, we note that the Block 1 community—a five-member community containing E. coli—was unable to suppress *K. pneumoniae* in mice more than a saline intervention at day 11 and day 16 post infection (Fig. 4.4B). Recent studies have also suggested augmenting E. coli with large, diverse communities to clear *K. pneumoniae* [110]. Our data provide a contrasted result: DMC46, a diverse community comprised of 46 strains, contained a strain of E. coli but was not as effective as SynCom15, comprised of 15 strains including the same E. coli strain, at suppressing *K. pneumoniae* in mice (Fig. 4.4B).

Collectively, these observations illustrated that the efficacy of SynCom15 as a community that suppresses *K. pneumoniae* across different environments cannot be solely ascribed to the presence of any single strain, including E. coli, or an obvious subset of strains gleaned from analysis of our mouse experiments. Moreover, coarse community descriptions, like community diversity for instance, do not provide an explanation for our results. In contrast, our findings highlight the utility of evaluating community function through our statistical approach that considers emergent, and potentially non-obvious properties of the structure-function relationship for communities.

4.3 Discussion

Using the clearance of *K. pneumoniae* as a target function, we developed a generative model using a functional landscape approach to engineer a community - SynCom 15 - to optimally suppress *K. pneumoniae* in a range of *in-vitro* media conditions, and *in-vivo* experimental conditions. SynCom15 was discovered by analyzing strains that co-varied with successful *K. pneumoniae* suppression. All 15 strains were essential for SynCom15's function, indicating the necessity of the complex ecological interactions between the 15 strains.

In Chapter 2, we showed that low epistasis enabled a functional landscape approach to designing predictive and generative regression models. In Chapter 3, we successfully inferred an epistatic functional landscape by discovering smooth variation with an environmental gradient (toxicity). Pathogen suppression within the human gut microbiome relies heavily on large-scale interactions between strains and with the human host [43, 45, 113], and there is no environmental gradient to exploit. However, we have shown that a nonlinear regression is able to infer the functional landscape and enable the generation of performative communities, outperforming the linear regression in out-of-sample predictions. This result paves the way for approaching various other microbial community functions that rely heavily on complex interactions.

Analysis of the RF model led to the design of SynCom15, a community of strains built entirely by identifying key strains that co-varied in successful consortia. SynCom 15 successfully sequestered *K. pneumoniae* in multiple media conditions, proving its more general efficacy. It is important to reiterate that SynCom15 was generated by only considering strain presence and absence across the initial 96 communities built, and not through any underlying metabolic or dynamic processes. Furthermore, the greater effectiveness of SynCom15 over DMC46 - a 46-member community that contained the 15 strains defining SynCom15 - ighlights the functional power of defined small bacterial communities in contrast to recent studies advocating engineering large communities spanning 50 to greater than 100

strains [90]. The discovery of a core ecological, functional group of strains is a pivotal result for the future of rational design of microbial communities. The ability to engineer sparse, functional bacterial communities that retain their function beyond the environment designed in is a tremendous advantage from a manufacturing and regulatory standpoint for creating therapeutic consortia for clinical use [41].

The 15 strains comprising SynCom15 formed a core group of interacting strains that were necessary and sufficient for the clearance of *K. pneumoniae*. Although not all members of SynCom15 engrafted into mice guts, removing the unengrafted strains inhibited SynCom15's activity. This result is striking, indicating that the community function landscape applied to DBTL discovered a community of strains where interactions between them enabled survivability and engraftment alongside the actual intended function. Engraftment into natural systems eluded microbiome engineering efforts [24], and our results open the door to designing specifically for functional activity in natural systems.

4.4 Methods

4.4.1 Creation and whole genome sequencing of strain bank

Fecal samples were obtained from 28 human donors that fell within the age range of 18 to 63 with a median age of 35. Donors were selected as those with no antibiotic use in the past year, no known history of diabetes, colitis, autoimmune disease, cancer, pneumonia, dysentery, or cellulitis at time of consent. Institutions that approved protocols of fecal sample collection were Memorial Sloan Kettering (MSK) and the University of Chicago. Fresh fecal samples were immediately reduced in an anaerobic chamber upon collection and diluted and cultured on various growth media. Agar media types vary, but include any of the following: Columbia Blood Agar, Brain Heart Infusion + Yeast, Brain Heart Infusion + Mucin, Brain Heart Infusion + Yeast + Acetate or N-acetylglucosamine, reinforced Clostridial Agar, Peptone

Yeast Glucose, Yeast Casitone Fatty Acids, Defined media M5. Colonies were selected and grown to be sufficiently turbid, 20% glycerol/PBS stocks were created and stored in a -80°C freezer.

Colonies were selected for whole-genome based on pyro-sequencing of the 16S region which provides a rough estimate of genus level designation. For each donor, only colonies that had a sequence identity threshold of less than 99% from CD-Hit (v. 4.8.1) were selected for whole-genome sequencing [124, 125]. Bacterial genomic DNA was extracted using QIAamp DNA Mini Kit (QIAGEN) according to manufacturer's manual. The purified DNA was quantified using a Qubit 2.0 fluorometer. 1000ng of each sample was prepared for sequencing using the QIAseq FX DNA Library Kit (QIAGEN). The protocol was carried out for a targeted fragment size of 550bp. Sequencing was performed on the MiSeq or NextSeq platform (Illumina) with a paired-end (PE) kit in pools designed to provide 1-3 million PE reads per sample with read length of 250 or 150 bp.

Adapters were trimmed off with Trimmomatic (v0.39) with following parameters: the leading and trailing 3 bp of the sequences were trimmed off, quality was controlled by a sliding window of 4, with an average quality score of 15 (default parameters of Trimmomatic) [126]. Moreover, any read that was less than 50 bp long after trimming and quality control were discarded. The remaining high-quality reads were assembled into contigs using SPAdes (v3.15.4) [127].

Taxonomic classification of the assembled contigs was performed with the following methods: (a) Kraken2 (v2.1.2; (b) full/partial length 16S rRNA gene from each isolated colony's assembled contigs is extracted and input into BLASTn (v2.10.1+) to query against NCBI's RNA RefSeq database [96, 128, 129]. Top five hits for each query are manually curated to determine an isolate's identity, with identity and coverage cutoff both at 95%; (c) GTDB-Tk (v1.5.1) [130]. The final taxonomy is determined by the consensus of the three methods. Any colony that did not match initial pyro-sequencing taxonomy or lacked consensus was

excluded from the commensal strain bank.

Table 4.2: **Strain identification of the 46 strains comprising the restricted strain bank** used to create DMCs

Strain Identity	Genus	Species
MSK.1.10	Collinsella	Collinsella aerofaciens
MSK.13.1	Bifidobacterium	Bifidobacterium sp.
MSK.13.37	Blautia	Blautia obeum
MSK.15.50	Lacrimispora	Lacrimispora celerecrescens
MSK.17.59	Coprococcus	Coprococcus comes
MSK.17.74	Blautia	Blautia faecis
MSK.18.53	Bacteroides	Bacteroides caccae
MSK.19.28	Escherichia	Escherichia coli
MSK.20.21	Megasphaera	Megasphaera massiliensis
MSK.20.34	Bacteroides	Bacteroides xylanisolvens
MSK.21.47	Mitsuokella	Mitsuokella jalaludinii
MSK.4.17	Blautia	Blautia producta
MSK.5.24	Lachnoclostridium	[Clostridium] scindens
MSK.7.21	Lachnoclostridium	[Clostridium] symbosium
MSK.7.7	Erysipelatoclostridium	[Clostridium] innocuum
DFI.1.157	Odoribacter	Odoribacter splanchnicus
DFI.4.154	Parabacteroides	Parabacteroides merdae
MSK.1.25	Sellimonas	Sellimonas intestinalis
MSK.14.27	Clostridium	Clostridium cochlearium
MSK.16.69	Bacteroides	Bacteroides eggerthii
MSK.17.22	Mediterraneibacter	Ruminococcus lactaris
MSK.17.8	Acidaminococcus	Acidaminococcus intestini

Strain Identity	Genus	Species
MSK.18.8	Bacteroides	Bacteroides cellulosilyticus
MSK.2.98	Lachnoclostridium	Enterocloster clostridioforme
MSK.20.40	Bacteroides	Phocaeicola massiliensis
MSK.20.79	Bacteroides	Bacteroides uniformis
MSK.21.24	Prevotella	Prevotella stercorea
MSK.21.52	Megasphaera	Megasphaera elsdenii
MSK.22.62	Fusicatenibacter	Fusicatenibacter saccharivorans
MSK.22.64	Bifidobacterium	Bifidobacterium breve
MSK.23.29	Dorea	Dorea formicigenerans
MSK.5.16	Veillonella	Veillonella ratti
MSK.13.4	Mediterraneibacter	Ruminococcus faecis
MSK.13.40	Blautia	Blautia luti
MSK.13.8	Bifidobacterium	Bifidobacterium pseudocatenulatum
MSK.15.233	unclassified Lachnospiraceae	Lachnospiraceae sp.
MSK.15.26	Blautia	Blautia hansenii
MSK.15.55	Erysipelatoclostridium	Erysipelatoclostridium ramosum
MSK.18.71	Blautia	Blautia wexlerae
MSK.18.79	Lactococcus	Lactococcus lactis
MSK.19.36	Bacteroides	Phocaeicola dorei
MSK.22.29	Bacteroides	Bacteroides ovatus
MSK.23.25	Bacteroides	Bacteroides thetaiotaomicron
MSK.7.16	Megamonas	Megamonas funiformis
MSK.8.32	Bifidobacterium	Bifidobacterium adolescentis
MSK.9.14	Parabacteroides	Parabacteroides sp.

4.4.2 Initial design strategy for bacterial communities

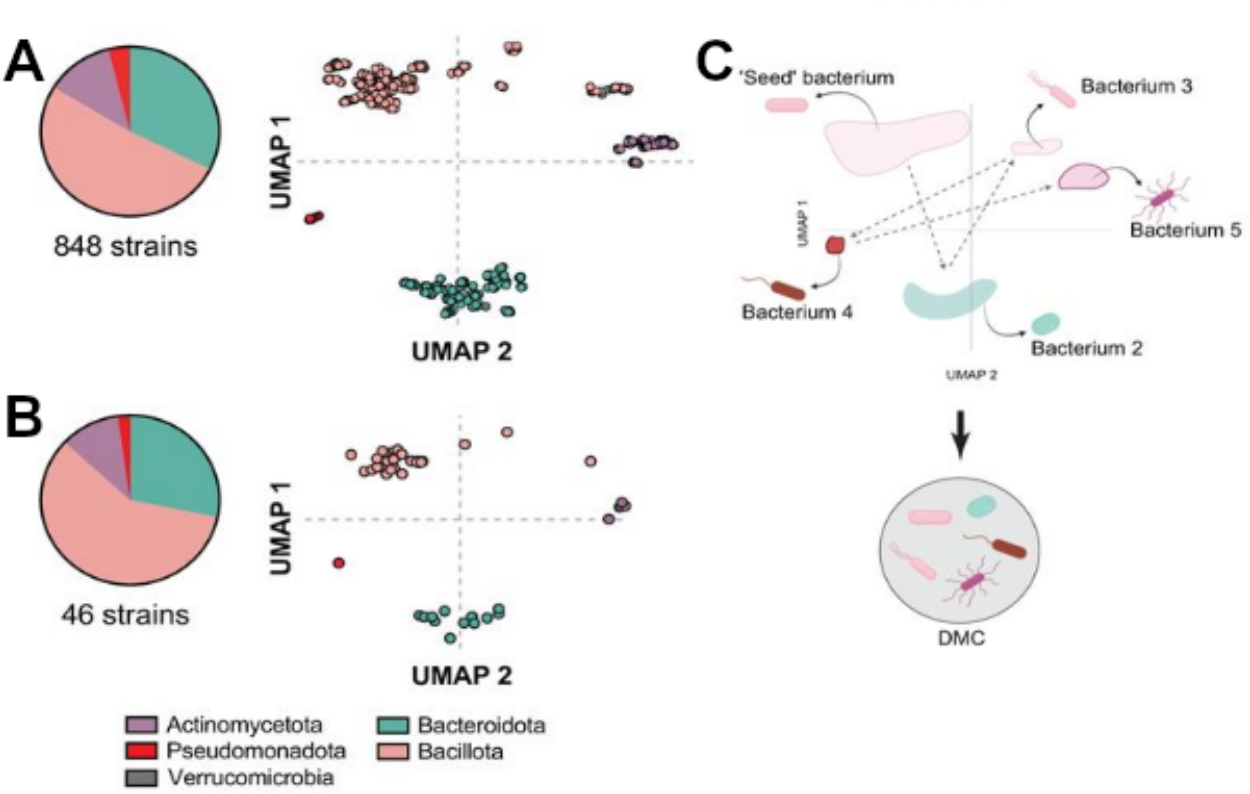


Figure 4.5: **Algorithm for Creative Diverse Communities Via UMAP** (**A**,**B**) Diversity of the full strain bank (panel **A**) and the restricted strain bank of 46 strains used to design communities (panel **B**). (**C**) Schematic for design of a five-member consortium. 'Seed' bacterium is a randomly chosen member of our strain bank.

To construct the UMAP plot, all gut commensal strains were annotated by their Prokka annotations and an alignment was created (848 rows comprising commensal strains, 150181 columns comprising Prokka annotated features). Each entry in the alignment is a '1' or a '0' indicating the presence or absence of a specific feature in a particular bacterial proteome.

To create a diverse community of size N, one option could be to choose the set of N strains that maximize dispersion across the UMAP space. This problem has been encountered in the field of facilities optimization and is known as 'the discrete p-dispersion problem' [131, 132]. However, this problem is considered 'NP complete'—a class of problem in computer science that is formally hard to solve and whose solutions can be verified only in non-polynomial

time. Therefore, we created an algorithm to design diverse communities comprised of N strains.

We perform the following steps using the UMAP plot based on bacterial genomes of 46 strains shown in Fig. 4.5B as the basis of our approach:

Step 1: Create 10,000 communities randomly of size N. The ensemble of all 10,000 communities of size N is represented as $C_{size\ N} = [c_1, ..., c_{10,000}]$

Step 2: Each community c_i is defined as $c_i = [x_1, ..., x_N]$ where x_i is the presence of strain i in c_i. Compute all pairwise distances in the UMAP space for all strains in c_i. For instance, the pairwise distance between strain 1 and 2 would be $pd_{1,2} = dist(x_1, x_2)$ where $dist$ is the function that computes the space between x_1 and x_2 in the UMAP space. We can then define the distribution of all pairwise spaces for community c_i as $PD_i = [pd_{1,2}, pd_{1,3}, ..., pd_{N-1,N}]$

Step 3: Order PD_i for a given c_i from largest to smallest values, then compute the mean pairwise distance across the lower 30% of values comprising PD_i. We term this value the 'mean adjusted dispersal'.

Step 4: Compute the mean adjusted dispersal for all communities in $C_{size\ N}$

Step 5: Identify the community within the 10,000 communities comprising $C_{size\ N}$ with the maximum mean adjusted dispersal. This community is the designed community comprising N strains

As an example, implementing this algorithm to engineer a five-member community would result in a bacterial community spanning different regions of the UMAP space (Fig. 4.5C).

4.4.3 Experimental workflow for K. pneumoniae clearance assay

The 46 bacterial strains that comprised our refined strain bank (Tbl. 4.2) were individually inoculated from a frozen stock into 900μL of BHI supplemented with cysteine 0.1% (BHIS) previously reduced. Strains were incubated at 37°C in static conditions for 48h in anaerobiosis to ensure that the most fastidious strains reach stationary phase. *K. pneumoniae* was

also inoculated in the same conditions, but only 24h after commensal isolates inoculation due to the fast growth capacity of this species and was incubated for 24h. All strain densities were assessed by taking 100 μL of each culture and measuring OD_{600} in a Biotek Cytation 5. To build all DMCs, isolates were inoculated in 900 μL of BHIS previously reduced in different combinations with an initial OD_{600} of 0.001, so that the densest community reaches a maximum total initial OD_{600} of approximately 0.05.*K. pneumoniae* was added at the same initial OD_{600} of 0.001 to all consortia. Cultures were incubated at 37°C in static conditions and anaerobiosis for 5 days. To assess *K. pneumoniae* abundance, 10 μL of each culture were collected daily and homogenized in 90 μL of PBS and serially diluted. Diluted samples were plated in BHIS with kanamycin ($50\mu g$/mL). Plates were incubated at 37°C overnight in aerobiosis. GFP expressing *K. pneumoniae* colony forming units (CFUs) were enumerated. In parallel, $100\mu L$ of each culture was also collected to recover the cell phase and the supernatants at 72h, 96h, and 120h. These samples were stored at -80°C to be later processed for shotgun metagenomics and metabolomics.

4.4.4 Training and validation of RandomForest (RF) model

We used a RandomForestRegressor, available with *scikit-learn* python package. Tree Depth was set to 12 levels per tree, the number of trees was set to 100, and the maximum number of features was set to "sqrt" (square-root of the number of strains total). Out-of-bag error was measured by a combination of R^2 (where numbers less than 1 indicate more error) and Mean Squared Error (where larger numbers indicate more error). To train and validate our model, we randomly split our dataset into 90% training and 10% true-out-of-sample 100 times. The input data was a vector of 46 1's and 0's corresponding to the pattern of presence-absence for each consortium.

In each iteration, the RandomForestRegressor was fit to the training set via 6-fold cross validation. Cross-validation accuracy was measured through Pearson Correlation. The true

out of sample set was then predicted, and prediction accuracy was measured by computing Mean Squared Error and Pearson Correlation of the predicted versus measured *K. pneumoniae* abundances after 120 hours of co-culture with the consortium. Feature Importance Scores for all features were observed and stored. This process was repeated 100 times, and prediction accuracies and feature significance scores were averaged. An additional RandomForestRegressor model was then trained on the entirety of the dataset with 6-fold cross-validation. Cross-validation accuracy was measured by calculating Mean Squared Error, Pearson Correlation, and R^2. Averaged prediction accuracies and feature significance scores were used to estimate prediction error.

4.4.5 PC analysis of in-silico consortia

The matrix of the 5,572 in-silico consortia was subject to PCA resulting in 46 principal components of data-variance (eigenvectors). To isolate the effect of *K. pneumoniae* clearance from confounding parameters, we first performed a series of steps to 'regress out' the effect of community size. First, let s_i be the community size of consortium i. Let y_i be the predicted *K. pneumoniae* clearance from the RF model for consortium i. A linear model is then created regressing community size against *K. pneumoniae* clearance taking the form:

$$y_i = \beta_1 s_i + \beta_0 + \epsilon \tag{4.1}$$

and

$$\hat{y_i} = \beta_1 s_i + \beta_0 \tag{4.2}$$

where $\hat{y_i}$ is the *K. pneumoniae* clearance of consortium i as a function of its size. The residuals of this linear model are given by $\epsilon_i = y_i - \hat{y_i}$ where ϵ_i is the degree of *K. pneumoniae* clearance of consortium i after removing linearly modeled information related to the size of

the community. All principal components were regressed against ϵ_i and principal component 46 (PC46) was found to be the most significantly associated with predicted, residualized *K. pneumoniae* clearance (Tbl. 4.3).

Table 4.3: Regression of principal components against predicted *K. pneumoniae* abundance

Principal Component	β_s	$q-$value
PC46	-1.19741742	$<$ e-30
PC42	-0.094923307	7.7290954916787e-21
PC39	-0.051681078	7.14711045634282e-07
PC15	-0.038966449	2.04860065834578e-05
PC18	-0.027653248	0.007412869
PC38	-0.02890851	0.013070101
PC22	-0.024975835	0.019219324
PC7	0.022949866	0.019732623
PC5	-0.019715046	0.044701679
PC30	0.022065149	0.044701679
PC32	-0.022728029	0.044701679
PC36	-0.023099879	0.044701679
PC16	-0.020290875	0.046933242
PC27	-0.020620294	0.053224398
PC3	-0.017141521	0.077531739
PC45	-0.024198535	0.088816739
PC1	0.012078782	0.26883872
PC17	0.013516972	0.26883872
PC6	-0.010035721	0.395213454
PC13	0.010984913	0.395213454
PC31	-0.011560292	0.395213454

Principal Component	β_s	$q-$value
PC35	0.01204232	0.395213454
PC44	-0.013660414	0.395213454
PC33	0.010802939	0.42403292
PC37	0.010317101	0.475060492
PC19	-0.009007958	0.489774719
PC4	-0.006895169	0.631834059
PC10	-0.005966014	0.749259327
PC43	-0.007174808	0.763243104
PC2	0.004869009	0.766364069
PC25	0.005382092	0.766364069
PC40	-0.005967165	0.766364069
PC8	-0.004142526	0.834050935
PC26	0.004182922	0.844740281
PC12	-0.003375313	0.862419699
PC14	0.003529667	0.862419699
PC9	0.001365272	0.882185345
PC11	0.002920493	0.882185345
PC20	0.002246982	0.882185345
PC21	0.002618705	0.882185345
PC23	-0.001620745	0.882185345
PC24	0.001730291	0.882185345
PC28	-0.002639369	0.882185345
PC29	0.001678534	0.882185345
PC41	0.001723672	0.882185345
PC34	0.000868522	0.922323735

4.4.6 Computing Similarity of Strain Projections onto PC46

Let $u \in \mathbb{R}^{46}$ be the vector of column projections of each strain on PC46 of the matrix of in-silico community strain presence/absence (Fig. 4.2D left panel). Let s be the scalar denoting the maximum value of u. Then,

$$\begin{aligned} A^{46\times 46} &= a_{ij} \\ a_{ij} &= d_{ij}^2 = ||u_i - u_j||^2 \end{aligned} \tag{4.3}$$

where $||\cdot||$ is the Euclidian norm

$$\begin{aligned} A_{ij} &= \begin{bmatrix} 0 & \dots & d_{46,1}^2 \\ \vdots & \ddots & \vdots \\ d_{1,46}^2 & \dots & 0 \end{bmatrix} \\ S_{ij} = s - A_{ij} &= \begin{bmatrix} s & \dots & s - d_{46,1}^2 \\ \vdots & \ddots & \vdots \\ s - d_{1,46}^2 & \dots & s \end{bmatrix} \end{aligned} \tag{4.4}$$

The resulting matrix S_{ij} is a symmetric similarity matrix with rows and columns indicating each strain and each element representing the similarity between strains i and j describes how strains are related to one another based on their projections along PC46. Hierarchical clustering on the resulting similarity matrix was then performed to identify groups of strains. Strains that are more similar are often found in communities that suppress *K. pneumoniae* and those that are more distant are rarely found in communities that suppress *K. pneumoniae*

4.4.7 Characterization of Mice Used for All Experiments Spanning Figures 4.3 and 4.4

All mouse experiments were performed in accordance with and approved by the Institutional Animal Care and Use Committee of the University of Chicago under protocol 72599. Male specific-pathogen-free C57BL/6J mice, aged 8 weeks to 10 weeks, from Jackson Laboratories were used for all experiments. Mice were kept within a facility that maintained a 12 hour light and 12 hour dark cycle and controlled humidity (30–70%) and temperature (68– 79 °F). Mice were housed in sterile, autoclaved cages with irradiated feed (LabDiets 5K67) and acidified, autoclaved water upon arriving at the on-site mouse facility. Mouse handling and cage changes were performed by investigators wearing sterile gowns, masks and gloves in a sterile biosafety hood. Mice were cohoused with their original shipment group until starting the experiment.

For germ-free (GF) studies, 8–10-week-old wild-type male C57BL/6J mice were used for all studies. Mice were initially obtained from The Jackson Laboratory and subsequently bred and raised in a GF isolator. After removal from the GF isolator, mice were handled in a sterile manner and individually housed in sealed negative pressure bio-containment unit isolators. Throughout breeding, mice were housed within the University of Chicago Gnotobiotic Research Animal Facility (GRAF) and maintained at a 12 hour light and 12 hour dark cycle and controlled humidity (30–70%) and temperature (68–79 °F). Gnotobiotic mice were fed an ad libitum diet of autoclaved Teklad Global 18% Protein Rodent Diet (Sterilizable) (2018S/2018SC).

4.4.8 Creating GF and antibiotic (Ab)-SPF cecal extract media

To create GF cecal extract media, 8–10-week-old wild-type male C57BL/6J GF mice were euthanized and cecal contents were collected, weighted, and homogenized in 10mL of sterile distilled water on a of per gram of content. Cecal suspension was centrifuged, and super-

natants were filtered through a 0.22 mm filter. GF cecal extract media was stored at -80°C.

To create ab-SPF cecal extract media C57BL/6J SPF male mice at 8-10 weeks of age were singly housed and placed under an antibiotic regime (0.25g MNV – metronidazole, neomycin, vancomycin) in the drinking water (day 0). Four days later, antibiotic treatment was halted and mice were placed on normal acidified water (day 4). Cages and food were also changed. On day 7 were euthanized and cecal contents were collected, weighted, and homogenized in 10mL of sterile distilled water on a of per gram of content. Cecal suspension was centrifuged, and supernatants were filtered through a 0.22 mm filter. Ab-SPF cecal extract media was stored at -80°C.

4.4.9 Preparation of Mice Stool Samples for Fecal Microbiota Transplant (FMT)

Fecal samples from 15-20 mice SPF mice from different cages (to increase sample diversity) were collected to a 50 mL tube. Samples were transferred immediately to the anaerobic chamber (anaerobic exposure was kept under 30 min). Samples were dissolved in 1 mL of PBS 20% glycerol 0.1% cysteine (previously filtered and reduced) per fecal pellet (1mL per 20 mg of fecal sample) using a mechanical pestle and vortexing. Samples were aliquoted in cryovials and stored -80 °C until use.

4.4.10 SPF Mouse Model of K. pneumoniae Infection

C57BL/6J male at 8-10 weeks of age were singly housed and placed under an antibiotic regime (0.25g MNV – metronidazole, neomycin, vancomycin) in the drinking water (day 0). Four days later, antibiotic treatment was halted and mice were placed on normal acidified water (day 4). Cages and food were also changed. On day 5 all mice were gavaged with 100μL of PBS containing 500 CFUs of *K. pneumoniae*, prepared as previously explained. On days 7, 8, and 9 mice were gavaged with 100μL of either selected defined bacterial consortia,

a fecal microbiota transplant from naive healthy mice, or PBS. Fecal samples were collected on days 0, 4, 7, 10, 12, 14, 16, and 21 (final day of the experiment) for 16s rRNA sequencing and on day 10 and 12 for metabolomics. These were immediately place on dry ice after collection and later stored at -80 °C. To assess for *K. pneumoniae* levels, fecal samples were collected on days 7, 10, 12, 14, 16, and 21. Fecal samples were homogenized in 1mL of PBS and serially diluted. Undiluted and diluted samples were plated in BHIS and kanamycin (50μg/mL).

4.4.11 Determining Engraftment of SynCom15 Strains in SPF Mice

To determine SynCom15 strain engraftment, 16s rRNA sequences from all 15 strains were blasted against 16S rRNA sequences derived from fecal samples of antibiotic-treated SPF mice gavaged with SynCom15 consortium. Fecal-derived sequences were assigned to a SynCom15 strain if their 16s rRNA percentage sequence identity was 100% with a minimum of a 95% coverage.

4.4.12 Data and Code Availability

The datasets generated in this are available within the Supplementary Tables of the original study [133]. Metagenomic data generated from profiling of human fecal microbiomes used in this study are publicly available on NCBI under BioProject ID PRJNA838648. 16S data generated from mouse experiments used in this study will be publicly available on NCBI under BioProject ID PRJNA1074807. Raw data files associated with metabolomic data used in this study will be found on MassIVEW repository MSV000094183

All code was written in either Python or R, and can be found at the following Github repository: https://github.com/aramanlab/Oliveira_et_al_2024

4.5 Supplementary Discussion

4.5.1 RandomForest Regressor Outperforms Linear Regressions

The RandomForest Regressor is a highly non-linear regression that captures nonlinear interactions between different strains within consortia. This model structure lends itself well to capturing highly epistatic interactions [116, 117, 119].

We compared the RF model with two versions of the linear regression - one fit exactly as described in Section 2.4.2 with LASSO regularization (LR #1), and one fit exactly as described in Section 3.4.5 with Ridge regularization (LR #2). We compared each models' performance in-sample (Validation on the 96 original DMCs) and out-of-sample (the generated 60 DMCs) to the performance of the RF (Tbl. 4.4).

Table 4.4: **Comparing Linear Regressions Performance to RandomForest**

	RandomForest	LR #1	LR #2
Validation R^2	0.98	0.45	0.28
Out of Sample R^2	0.60	0.44	0.29

Both versions of the linear regression perform worse than the RandomForest, but two surprising results arise. Firstly, both linear regressions perform poorly during in-sample validation, after the models have been fit to the entire dataset. Secondly, the performance out-of-sample is nearly the same as during validation for both linear regressions.

To understand why this occurred, we inspected the residuals of predictions (Fig. 4.6 for each of the three models. All three models behave similarly with nonfunctional communities, overpredicting *K. pneumoniae* suppression. However, as the observed *K. pneumoniae* abundance decreases, the RF model more accurately predicts the bulk of communities that lie in the observed range of $6-7$. We recall that LR #2 is weighted during the fitting procedure to prioritize communities with 'rarer' functions (see Section 3.4.5). Interestingly though, LR #1 similarly struggles within this range. We attribute this difficulty to the highly epistatic nature of *K. pneumoniae* clearance that makes it difficult for the linear regressions to learn

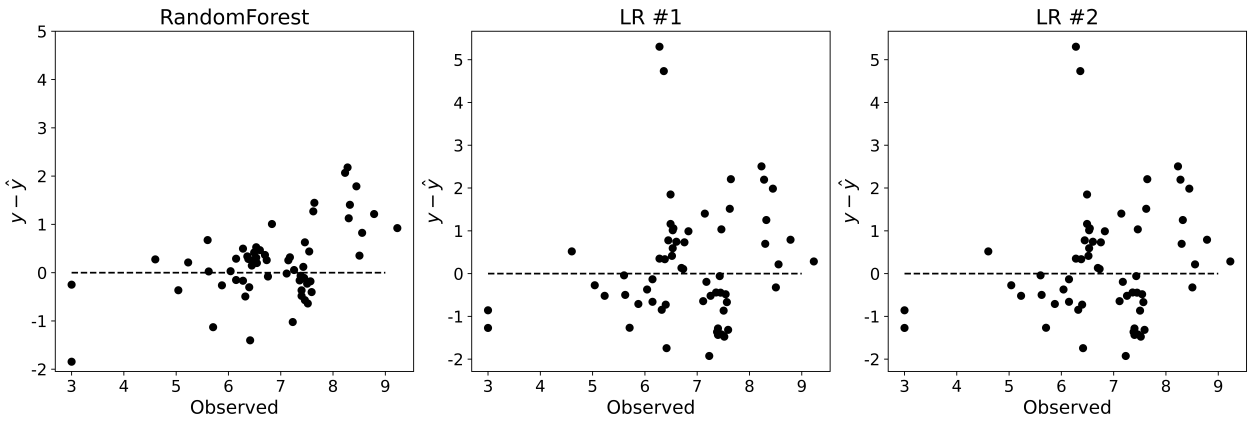


Figure 4.6: **RandomForest Regressor Outperforms Linear Regressions** Scatter plots of observed *K. pneumoniae* log_10 CFUs (x-axis) versus residuals of prediction ($y - \hat{y}$, y-axis) for the RF model (left), linear regression with LASSO regularization (middle), and linear regression with Ridge regularization (right). Dotted lines show $y = \hat{y}$, perfect predictions.

from a heavily undersampled landscape.

4.5.2 Supplementary Figures

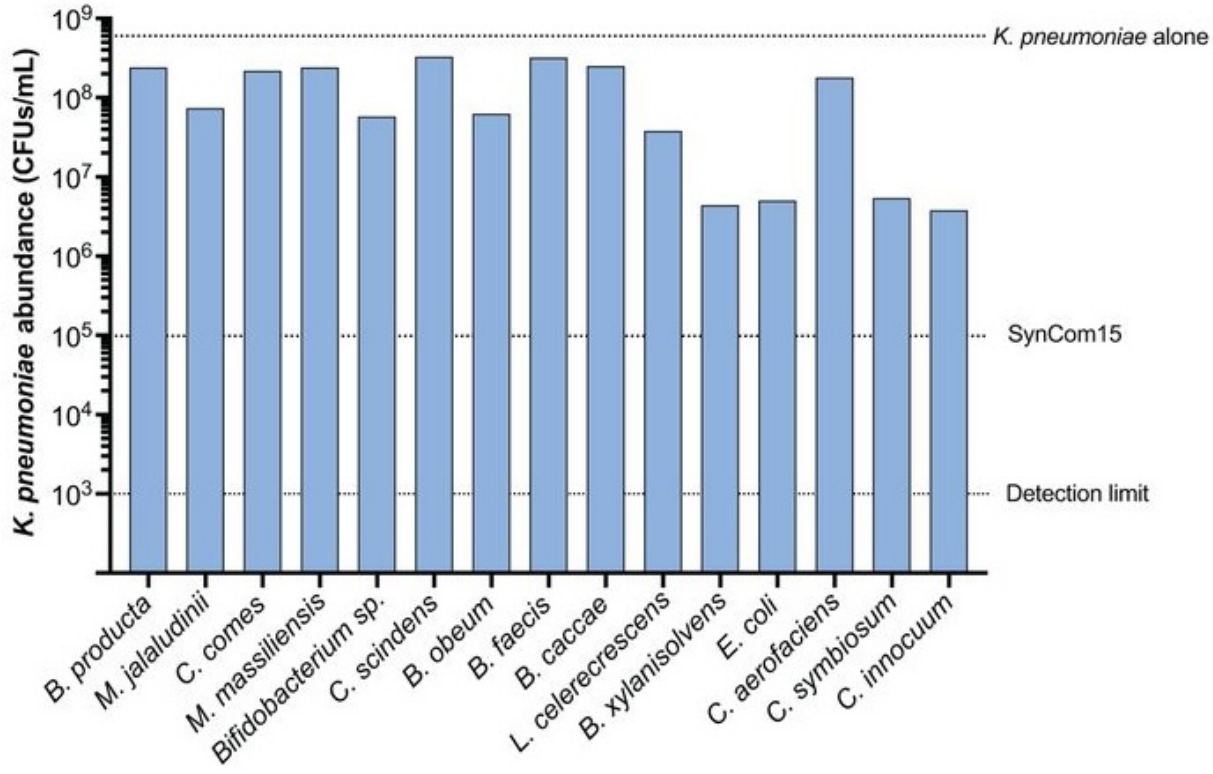


Figure S.4.1: **No Individual Strain Comprising SynCom15 Suppresses *K. pneumoniae* On Its Own**. *K. pneumoniae* abundance (y-axis) when co-cultured with each of the SynCom15 strains (x-axis) for 120 hours in BHIS media. 'K. pneumoniae alone' labels the *K. pneumoniae* abundance in monoculture for 120 hours in BHIS shown in Fig. 4.3B. 'SynCom15' labels the *K. pneumoniae* abundance of *K. pneumoniae* in co-culture with SynCom15 for 120 hours in BHIS shown in Fig. 4.3B. 'Detection limit' labels the lower limit of *K. pneumoniae* abundance for our assay.

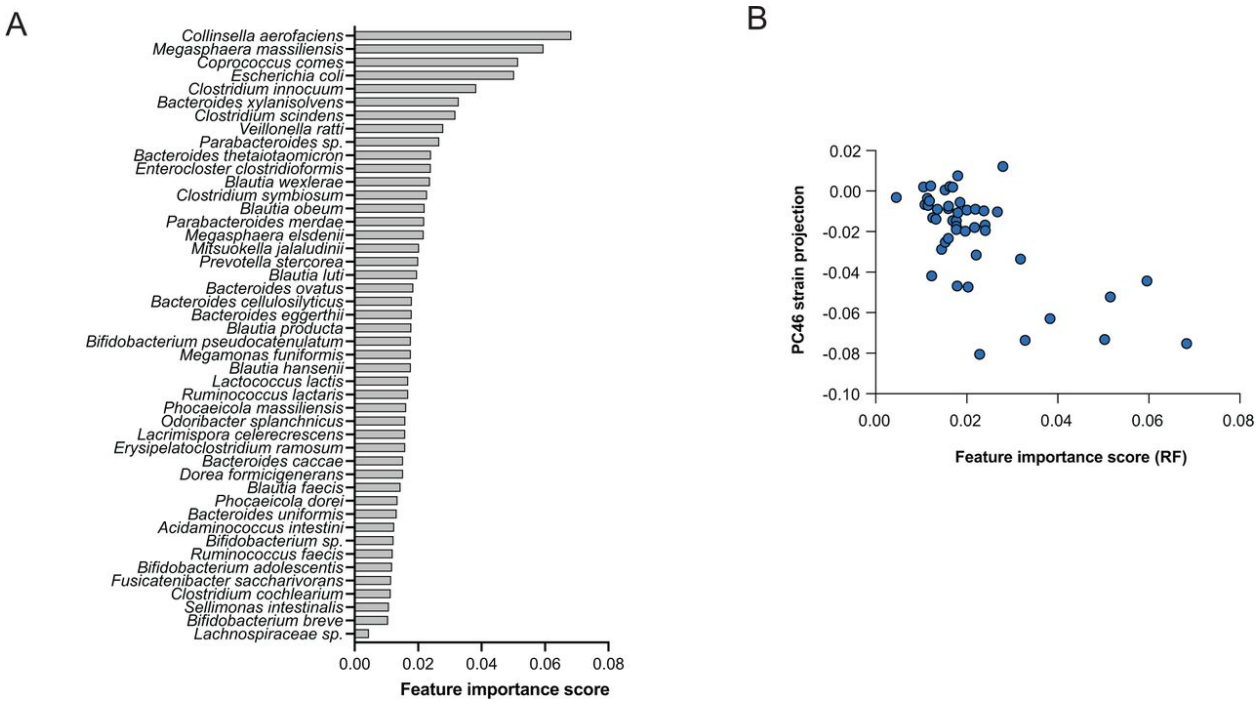


Figure S.4.2: **RandomForest Regressor Feature Importance Scores** (**A**) Feature importance score of each strain resulting from the RF model built on strain presence/absence. (**B**) Feature importance scores (x-axis) versus the projection onto PC46 (y-axis) for each individual strain. $|\rho| < 0.01$, $p-$value < 0.05 as calculated from a two-sided t-test

CHAPTER 5

CONCLUSION

Microbial communities perform many functions that directly affect human and environmental health. Exploiting these functions through rational community design has been one of the central endeavors of microbial ecology.

In this thesis, we presented the Community Function Landscape as a direct mapping between the structure of a microbial community and its function. Throughout the different computational and translational applications, we have shown that the functional landscape is applicable to a wide variety of microbial ecological systems and enable the engineering of defined microbiomes with targeted functional properties, bypassing the need to measure abundance dynamics or model specific interactions between strains. In particular, while the most basic form of the functional landscape occurs when the underlying landscape is smooth, many modifications can be made to infer rugged landscapes. This flexibility opens the door for a much wider range of microbial functional landscapes to be inferred through regression approaches.

In both translational applications explored in this thesis, the community function landscape enabled rational design for functional communities in natural systems. The translation of microbial functions from *in-vitro* settings to *in-vivo* systems has historically been a significant challenge. However, the ability of the functional landscapes to generalize community function enabled this translation. Understanding the change in BPA degradation across toxicity enabled the design of communities that carried epistatic interactions to resist toxicity while also degrade BPA at the less toxic concentrations. The statistical analysis of thousands of communites from the RF model enabled learning key ecological interactions that enabled *K. pneumoniae* clearance, thus overcomming the specificity of the model's training on data from BHIS media. This finding is consistent with emerging evidence suggesting that a way that the evolutionary process can generate adaptable systems is not selecting for individual

systems that function per se, but by selecting for underlying structural regularities amongst ensembles of systems that function [134]. This key principle of generalization opens the door for translational microbiome research by enabling a route of rational design that is environmentally-contextual and tunable.

The functional landscape can be thought of as a 'mechanism-second' approach to microbiome engineering: by inferring the landscape, one can identify functional communities and then analyze the mechanism behind their activity [35, 39]. The landscape uncovered the necessary nature of epistatic interactions for BPA degradation at toxic concentrations and the vital role of strain PO109, despite it not having any activity on its own. Furthermore, the core group of strains comprising SynCom15 was not established through a bottom-up approach, but rather finding strains that were co-varied with *K. pneumoniae* suppression. This important feature enables the rational design targeting many microbial functions in which no mechanism is known definitively, and even led to the discovery of these mechanisms.

5.1 Future Directions

The Community Function Landscape, despite its many successes, makes a few critical assumptions which must be analyzed deeply in context of further generalization of its scope.

Because the functional landscape approach predicts community function from species presence/absence, it explicitly assumes that replicate communities with the same composition will have the same function. This assumption could fail if communities exhibit alternative stable states [135, 136] with distinct final abundances and functions despite identical presence/absence compositions. While alternative stable states have been documented in synthetic consortia. [137], they do not appear to be widespread or have large impacts on function in the communities studied here. Data from four of the datasets analyzed in Chapter 2 with replicate communities are well predicted by the landscape (Fig. S.2.4). Duplicates of BPA-degrading communities were almost all nearly performing to each other and also

predicted equally well by the regressions (Fig. 3.3B). Furthermore, of the replicates prepared in measuring *K. pneumoniae* clearance, all replicates performed similarly well (Fig. 4.1). Thus, we conclude that alternative stable states do not drive substantial functional variation across any of the landscapes inferred here. However, none of the data throughout this thesis systematically varied the initial relative abundances for communities of fixed composition, and this may result in communities not reaching alternative stable states despite their existence. In ecological contexts where alternative stable states are pervasive and drive large functional variation, our predictions would begin to degrade in quality.

Another assumption made by the functional landscape is that no functional redundancy exists between taxa, as every strain as assumed to have a unique and independent role in contributing to community function. In cases of extensive functional redundancy between strains, the regressions would likely fail to infer a landscape. Thus, the excellent performance of our method on all data shown throughout this thesis is probably aided by the fact that the labor-intensive nature of the combinatorial experiments favors synthetic communities with relatively low redundancy. The experimental cost associated with increasing the number of species discourages the inclusion of taxa highly similar to those already in the pool. This is best seen with the initial DMCs created to measure *K. pneumonia* clearance - in order to extract the best value out of our time, we sought to maximize functional diversity within every DMC (Fig. 4.5), enabling the RandomForest regression to much more easily infer the landscape. This is the opposite regime of the natural communities, which are often phylogenetically under-dispersed [138, 139] and which often rely on functional redundancy to maintain stability [140, 141].

Whereas the regressions as presented (Eqn. 2.1) would probably struggle in communities with extensive redundancy, the argument above indicates how this limitation could be remedied, namely by grouping redundant taxa before performing regressions, or by coarse-graining functional units. In fact, our results suggest that the high predictive power of a

simple regression on the variables describing the presence or absence of any member in a group could be taken as the criterion indicating that the grouping was chosen appropriately. This offers a path towards quantitative prediction of complex community functions in the high-diversity regime and echoes recently proposed ideas from multiple groups [142–144].

The result of low-dimensional variation between the landscapes of BPA degradation as a function of toxicity begs the question of whether low-dimensional variation exists across different environmental gradients. Indeed, many studies have suggested that the activity of microbes is gradually affected by shifts in pH and temperature [145, 146]. However, the current formalism of the functional landscape does not explicitly enable direct comparisons across a landscape, but rather requires manual inspection of variation between different environmental conditions. It is theoretically possible to consolidate the landscapes inferred from different environmental conditions into a single model that captures the gradient within its coefficients (see Section 3.5.5); however, much more work is needed to determine how to best capture the effects of the gradient within a linear model.

Because functional landscapes can be approximated by sampling only a subset of all possible species combinations, large synthetic consortia with predefined functions can be designed and optimized computationally given only a small number of measurements, as seen with SynCom15 (Fig. 4.2). However, a smart initial sampling methodology is needed, since randomly sampling the space of all communities is not sufficient for undersampled landscapes (see the discussion in Section 2.5.2). Determining the optimal sampling strategy to accurately infer landscapes remains an important avenue for future work. It is unclear if the variation in single strain BPA degradation (Fig. S.3.3) enabled us to infer the landscape from only 70 communities per condition, or if the functional diversity within all initial 96 DMCs were the main factor in experiencing the range of endpoint *K. pneumoniae* abundances. However, even without extensive optimization, our models throughout this thesis were able to identify the communities with the highest functional output, even when these communities were left

out of sample. With functions that are more difficult for microbial communities to perform, understanding sampling strategies remains a key point of improvement.

BIBLIOGRAPHY

1. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nature Biotechnology* **39,** 499–509. ISSN: 1087-0156, 1546-1696. https://www.nature.com/articles/s41587-020-0718-6 (2025) (Apr. 2021).
2. Dubilier, N., McFall-Ngai, M. & Zhao, L. Microbiology: Create a global microbiome effort. *Nature* **526,** 631–634. ISSN: 0028-0836, 1476-4687. https://www.nature.com/articles/526631a (2025) (Oct. 29, 2015).
3. Godfrey, L. & Glass, J. in *Methods in Enzymology* 1st ed. C, 483–506 (Elsevier Inc., 2010). http://dx.doi.org/10.1016/B978-0-12-381294-0.00027-4.
4. Stein, L. Y. & Klotz, M. G. The nitrogen cycle. *Current Biology* **26,** R94–R98. ISSN: 09609822. http://dx.doi.org/10.1016/j.cub.2015.12.021 (2016).
5. Bardgett, R. D., Freeman, C. & Ostle, N. J. Microbial contributions to climate change through carbon cycle feedbacks. *ISME Journal* **2,** 805–814. ISSN: 17517362 (2008).
6. Meinshausen, N. & Bühlmann, P. Stability Selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **72,** 417–473. ISSN: 1369-7412, 1467-9868. https://academic.oup.com/jrsssb/article/72/4/417/7076513 (2025) (Sept. 1, 2010).
7. Clark, R. L. *et al.* Design of synthetic human gut microbiome assembly and butyrate production. *Nature Communications* **12,** 3254. ISSN: 2041-1723. http://www.nature.com/articles/s41467-021-22938-y (2022) (Dec. 2021).
8. McKenney, P. T. & Pamer, E. G. From Hype to Hope: The Gut Microbiota in Enteric Infectious Disease. *Cell* **163,** 1326–1332. ISSN: 00928674. https://linkinghub.elsevier.com/retrieve/pii/S0092867415015512 (2022) (Dec. 2015).
9. Grosen, A. K. *et al.* Faecal Microbiota Transplantation Eradicated Extended-Spectrum Beta-Lactamase-Producing ***Klebsiella pneumoniae*** from a Renal Transplant Recipient with Recurrent Urinary Tract Infections. *Case Reports in Nephrology and Dialysis* **9,** 102–107. ISSN: 2296-9705. https://www.karger.com/Article/FullText/502336 (2022) (Aug. 20, 2019).
10. Claus, S. P., Guillou, H. & Ellero-Simatos, S. The gut microbiota: a major player in the toxicity of environmental pollutants? *npj Biofilms and Microbiomes* **2,** 16003. ISSN: 2055-5008. http://www.nature.com/articles/npjbiofilms20163 (2022) (Nov. 23, 2016).
11. Genovese, M. *et al.* Bioremediation of benzene, toluene, ethylbenzene, xylenes-contaminated soil: a biopile pilot experiment. *Journal of Applied Microbiology* **105,** 1694–1702. ISSN: 13645072, 13652672. https://academic.oup.com/jambio/article/105/5/1694/6720271 (2025) (Nov. 2008).
12. Mickalide, H. & Kuehn, S. Higher-Order Interaction between Species Inhibits Bacterial Invasion of a Phototroph-Predator Microbial Community. *Cell Systems* **9,** 521–533.e10. ISSN: 24054712. https://linkinghub.elsevier.com/retrieve/pii/S2405471219303904 (2022) (Dec. 2019).

13. Morin, M. A., Morrison, A. J., Harms, M. J. & Dutton, R. J. Higher-order interactions shape microbial interactions as microbial community complexity increases. *Scientific Reports* **12,** 22640. ISSN: 2045-2322. https://www.nature.com/articles/s41598-022-25303-1 (2025) (Dec. 31, 2022).

14. Gould, A. L. *et al.* Microbiome interactions shape host fitness. *Proceedings of the National Academy of Sciences* **115.** ISSN: 0027-8424, 1091-6490. https://pnas.org/doi/full/10.1073/pnas.1809349115 (2022) (Dec. 18, 2018).

15. Ramírez Rojas, A. A., Swidah, R. & Schindler, D. Microbes of traditional fermentation processes as synthetic biology chassis to tackle future food challenges. *Frontiers in Bioengineering and Biotechnology* **10,** 982975. ISSN: 2296-4185 (2022).

16. Wagner, M. *et al.* Microbial community composition and function in wastewater treatment plants. *Antonie van Leeuwenhoek* **81,** 665–680. ISSN: 00036072. http://link.springer.com/10.1023/A:1020586312170 (2025) (2002).

17. Eng, A. & Borenstein, E. Microbial community design: methods, applications, and opportunities. *Current Opinion in Biotechnology* **58,** 117–128. ISSN: 09581669. https://linkinghub.elsevier.com/retrieve/pii/S0958166918301927 (2025) (Aug. 2019).

18. San León, D. & Nogales, J. Toward merging bottom–up and top–down model-based designing of synthetic microbial communities. *Current Opinion in Microbiology* **69,** 102169. ISSN: 13695274. https://linkinghub.elsevier.com/retrieve/pii/S1369527422000534 (2022) (Oct. 2022).

19. Hogle, S. L., Ruusulehto, L., Cairns, J., Hultman, J. & Hiltunen, T. Localized coevolution between microbial predator and prey alters community-wide gene expression and ecosystem function. *The ISME Journal* **17,** 514–524. ISSN: 1751-7362, 1751-7370. https://academic.oup.com/ismej/article/17/4/514-524/7474033 (2025) (Apr. 1, 2023).

20. Gopalakrishnappa, C., Gowda, K., Prabhakara, K. H. & Kuehn, S. An ensemble approach to the structure-function problem in microbial communities. *iScience* **25,** 103761. ISSN: 25890042. https://linkinghub.elsevier.com/retrieve/pii/S2589004222000311 (2022) (Feb. 2022).

21. Sanchez-Gorostiaga, A., Bajić, D., Osborne, M. L., Poyatos, J. F. & Sanchez, A. High-order interactions distort the functional landscape of microbial consortia. *PLOS Biology* **17** (ed Shou, W.) e3000550. ISSN: 1545-7885. https://dx.plos.org/10.1371/journal.pbio.3000550 (2022) (Dec. 12, 2019).

22. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic Gene Expression in a Single Cell. *Science* **297,** 1183–1186. ISSN: 0036-8075, 1095-9203. https://www.science.org/doi/10.1126/science.1070919 (2022) (Aug. 16, 2002).

23. Goldstein, R. M., Mallory, L. M. & Alexander, M. Reasons for possible failure of inoculation to enhance biodegradation. *Applied and Environmental Microbiology* **50,** 977–983. ISSN: 0099-2240 (Oct. 1985).

24. Silverstein, M. R., Segrè, D. & Bhatnagar, J. M. Environmental microbiome engineering for the mitigation of climate change. *Global Change Biology* **29,** 2050–2066. ISSN: 1354-1013, 1365-2486. https://onlinelibrary.wiley.com/doi/10.1111/gcb.16609 (2023) (Apr. 2023).

25. Albright, M. B. N. *et al.* Solutions in microbiome engineering: prioritizing barriers to organism establishment. *The ISME Journal* **16,** 331–338. ISSN: 1751-7362, 1751-7370. https://academic.oup.com/ismej/article/16/2/331-338/7474227 (2025) (Feb. 1, 2022).

26. Lawson, C. E. *et al.* Common principles and best practices for engineering microbiomes. *Nature Reviews Microbiology* **17,** 725–741. ISSN: 1740-1526, 1740-1534. http://www.nature.com/articles/s41579-019-0255-9 (2022) (Dec. 2019).

27. Gallardo-Navarro, O., Aguilar-Salinas, B., Rocha, J. & Olmedo-Álvarez, G. Higher-order interactions and emergent properties of microbial communities: The power of synthetic ecology. *Heliyon* **10,** e33896. ISSN: 24058440. https://linkinghub.elsevier.com/retrieve/pii/S2405844024099274 (2025) (July 2024).

28. Ruan, Z. *et al.* Engineering natural microbiomes toward enhanced bioremediation by microbiome modeling. *Nature Communications* **15,** 4694. ISSN: 2041-1723. https://www.nature.com/articles/s41467-024-49098-z (2025) (June 1, 2024).

29. Chandran, H., Meena, M. & Sharma, K. Microbial Biodiversity and Bioremediation Assessment Through Omics Approaches. *Frontiers in Environmental Chemistry* **1,** 570326. ISSN: 2673-4486. https://www.frontiersin.org/article/10.3389/fenvc.2020.570326/full (2023) (Sept. 25, 2020).

30. Lawson, C. E. Retooling Microbiome Engineering for a Sustainable Future. *mSystems* **6,** 10.1128/msystems.00925–21. ISSN: 2379-5077. https://journals.asm.org/doi/10.1128/msystems.00925-21 (2025) (Aug. 31, 2021).

31. Yu, K. *et al.* An integrated meta-omics approach reveals substrates involved in synergistic interactions in a bisphenol A (BPA)-degrading microbial community. *Microbiome* **7,** 16. ISSN: 2049-2618. https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-019-0634-5 (2022) (Dec. 2019).

32. Wołejko, E., Wydro, U. & Łoboda, T. The ways to increase efficiency of soil bioremediation. *Ecological Chemistry and Engineering S* **23,** 155–174. ISSN: 1898-6196. https://www.sciendo.com/article/10.1515/eces-2016-0011 (2023) (Mar. 1, 2016).

33. Wilson, B. C., Vatanen, T., Cutfield, W. S. & O'Sullivan, J. M. The Super-Donor Phenomenon in Fecal Microbiota Transplantation. *Frontiers in Cellular and Infection Microbiology* **9,** 2. ISSN: 2235-2988. https://www.frontiersin.org/article/10.3389/fcimb.2019.00002/full (2022) (Jan. 21, 2019).

34. Gargiullo, L., Del Chierico, F., D'Argenio, P. & Putignani, L. Gut Microbiota Modulation for Multidrug-Resistant Organism Decolonization: Present and Future Perspectives. *Frontiers in Microbiology* **10,** 1704. ISSN: 1664-302X. https://www.frontiersin.org/article/10.3389/fmicb.2019.01704/full (2025) (July 25, 2019).

35. Diaz-Colunga, J., Skwara, A., Vila, J. C. C., Bajic, D. & Sánchez, Á. *Emergent ecosystem functions follow simple quantitative rules* preprint (Ecology, June 26, 2022). http://biorxiv.org/lookup/doi/10.1101/2022.06.21.496987 (2022).

36. Izabel-Shen, D. *et al.* Repeated introduction of micropollutants enhances microbial succession despite stable degradation patterns. *ISME Communications* **2,** 48. ISSN: 2730-6151. https://www.nature.com/articles/s43705-022-00129-0 (2023) (June 7, 2022).

37. Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nature Reviews Microbiology* **10,** 538–550. ISSN: 1740-1526, 1740-1534. https://www.nature.com/articles/nrmicro2832 (2025) (Aug. 2012).

38. Van den Berg, N. I. *et al.* Ecological modelling approaches for predicting emergent properties in microbial communities. *Nature Ecology and Evolution* **6,** 855–865. ISSN: 2397334X (2022).

39. Sanchez, A. *et al.* The community-function landscape of microbial consortia. *Cell Systems* **14,** 122–134. ISSN: 24054712. https://linkinghub.elsevier.com/retrieve/pii/S2405471222004999 (2025) (Feb. 2023).

40. Zomorrodi, A. R. & Segrè, D. Synthetic Ecology of Microbes: Mathematical Models and Applications. *Journal of Molecular Biology* **428,** 837–861. ISSN: 00222836. https://linkinghub.elsevier.com/retrieve/pii/S0022283615006129 (2025) (Feb. 2016).

41. Vrancken, G., Gregory, A. C., Huys, G. R. B., Faust, K. & Raes, J. Synthetic ecology of the human gut microbiota. *Nature Reviews Microbiology* **17,** 754–763. ISSN: 1740-1526, 1740-1534. https://www.nature.com/articles/s41579-019-0264-8 (2025) (Dec. 2019).

42. Gowda, K., Ping, D., Mani, M. & Kuehn, S. Genomic structure predicts metabolite dynamics in microbial communities. *Cell* **185,** 530–546.e25. ISSN: 00928674. https://linkinghub.elsevier.com/retrieve/pii/S0092867421015427 (2025) (Feb. 2022).

43. Baranwal, M. *et al.* Recurrent neural networks enable design of multifunctional synthetic human gut microbiome dynamics. *eLife* **11,** e73870. ISSN: 2050-084X. https://elifesciences.org/articles/73870 (2025) (June 23, 2022).

44. Raman, A. S. *et al.* A sparse covarying unit that describes healthy and impaired human gut microbiota development. *Science* **365,** eaau4735. ISSN: 0036-8075, 1095-9203. https://www.sciencemag.org/lookup/doi/10.1126/science.aau4735 (2022) (July 12, 2019).

45. Cao, Z., Yan, W., Ding, M. & Yuan, Y. Construction of microbial consortia for microbial degradation of complex compounds. *Frontiers in Bioengineering and Biotechnology* **10,** 1051233. ISSN: 2296-4185. https://www.frontiersin.org/articles/10.3389/fbioe.2022.1051233/full (2023) (Dec. 6, 2022).

46. Faith, J. J., Ahern, P. P., Ridaura, V. K., Cheng, J. & Gordon, J. I. Identifying Gut Microbe–Host Phenotype Relationships Using Combinatorial Communities in Gnotobiotic Mice. *Science Translational Medicine* **6.** ISSN: 1946-6234, 1946-6242. https://www.science.org/doi/10.1126/scitranslmed.3008051 (2025) (Jan. 22, 2014).

47. Wright, S. *The roles of mutation, inbreeding, crossbreeding and selection in evolution.* 1932.

48. Ferguson, A. L. *et al.* Translating HIV Sequences into Quantitative Fitness Landscapes Predicts Viral Vulnerabilities for Rational Immunogen Design. *Immunity* **38,** 606–617. ISSN: 10747613. http://dx.doi.org/10.1016/j.immuni.2012.11.022 (2013).

49. Chou, H. H., Chiu, H. C., Delaney, N. F., Segrè, D. & Marx, C. J. Diminishing returns epistasis among beneficial mutations decelerates adaptation. *Science* **332,** 1190–1192. ISSN: 00368075 (2011).

50. Khan, A. I., Dinh, D. M., Schneider, D., Lenski, R. E. & Cooper, T. F. Negative epistasis between beneficial mutations in an evolving bacterial population. *Science* **332,** 1193–1196. ISSN: 00368075 (2011).

51. Kauffman, S. & Levin, S. Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology* **128,** 11–45. ISSN: 10958541 (1987).

52. Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M. & Tans, S. J. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* **445,** 383–386. ISSN: 14764687 (2007).

53. Kryazhimskiy, S., Tkăcik, G. & Plotkin, J. B. The dynamics of adaptation on correlated fitness landscapes. *Proceedings of the National Academy of Sciences of the United States of America* **106,** 18638–18643. ISSN: 00278424 (2009).

54. Diaz-Colunga, J. *et al.* Global epistasis on fitness landscapes. arXiv:2210.03677 [q-bio]. http://arxiv.org/abs/2210.03677 (2023) (Oct. 2022).

55. Phillips, P. C. Epistasis - The essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics* **9,** 855–867. ISSN: 14710056 (2008).

56. Poelwijk, F. J., Krishna, V. & Ranganathan, R. The Context-Dependence of Mutations: A Linkage of Formalisms. *PLOS Computational Biology* **12** (ed Nussinov, R.) e1004771. ISSN: 1553-7358. https://dx.plos.org/10.1371/journal.pcbi.1004771 (2025) (June 23, 2016).

57. Szendro, I. G., Schenk, M. F., Franke, J., Krug, J. & De Visser, J. A. G. M. Quantitative analyses of empirical fitness landscapes. *Journal of Statistical Mechanics: Theory and Experiment* **2013,** P01005. ISSN: 1742-5468. https://iopscience.iop.org/article/10.1088/1742-5468/2013/01/P01005 (2025) (Jan. 16, 2013).

58. Hordijk, W. & Stadler, P. F. Amplitude Spectra of Fitness Landscapes. *Advances in Complex Systems* **01,** 39–66. ISSN: 0219-5259 (1998).

59. Diaz-Colunga, J., Skwara, A., Vila, J. C. C. & Bajic, D. Global epistasis and the emergence of ecological function. *bioRxiv,* 2022.06.21.496987 (2023).

60. Langenheder, S., Bulling, M. T., Solan, M. & Prosser, J. I. Bacterial Biodiversity-Ecosystem Functioning Relations are Modified by Environmental Complexity. *PLoS ONE* **5.** ISSN: 19326203 (2010).

61. Kehe, J. *et al.* Massively parallel screening of synthetic microbial communities. *Proceedings of the National Academy of Sciences of the United States of America* **116,** 12804–12809. ISSN: 10916490 (2019).

62. Hastie, T., Tibshirani, R. J. & Friedman, J. *The elements of statistical learning* 2nd (Springer, 2008).

63. Otwinowski, J. & Plotkin, J. B. Inferring fitness landscapes by regression produces biased estimates of epistasis. *Proceedings of the National Academy of Sciences* **111.** ISSN: 0027-8424, 1091-6490. https://pnas.org/doi/full/10.1073/pnas.1400849111 (2022) (June 3, 2014).

64. Barbier, M., Arnoldi, J. F., Bunin, G. & Loreau, M. Generic assembly patterns in complex ecological communities. *Proceedings of the National Academy of Sciences of the United States of America* **115,** 2156–2161. ISSN: 10916490 (2018).

65. Bunin, G. Ecological communities with Lotka-Volterra dynamics. *Phys. Rev. E* **95,** 042414. https://link.aps.org/doi/10.1103/PhysRevE.95.042414 (4 Apr. 2017).

66. Vanwonterghem, I. *et al.* Deterministic processes guide long-term synchronised population dynamics in replicate anaerobic digesters. *ISME Journal* **8,** 2015–2028. ISSN: 17517370 (2014).

67. Tong, T., Li, R., Chen, J., Ke, Y. & Xie, S. Bisphenol A biodegradation differs between mudflat and mangrove forest sediments. *Chemosphere* **270,** 128664. ISSN: 00456535. https://linkinghub.elsevier.com/retrieve/pii/S0045653520328599 (2023) (May 2021).

68. Clark, R. L. *et al.* Design of synthetic human gut microbiome assembly and butyrate production. *Nature Communications* **12,** 1–16. ISSN: 20411723. http://dx.doi.org/10.1038/s41467-021-22938-y (2021).

69. Sanchez-Gorostiaga, A., Bajić, D., Osborne, M. L., Poyatos, J. F. & Sanchez, A. High-order interactions distort the functional landscape of microbial consortia. *PLoS Biology* **17,** 1–34 (2019).

70. Barbier, M., Arnoldi, J.-F., Bunin, G. & Loreau, M. Generic assembly patterns in complex ecological communities. *Proceedings of the National Academy of Sciences* **115.** Publisher: Proceedings of the National Academy of Sciences, 2156–2161. https://www.pnas.org/doi/abs/10.1073/pnas.1710352115 (2023) (Feb. 2018).

71. Freijeiro-González, L., Febrero-Bande, M. & González-Manteiga, W. A Critical Review of LASSO and Its Derivatives for Variable Selection Under Dependence Among Covariates. *International Statistical Review* **90,** 118–145. ISSN: 0306-7734, 1751-5823. https://onlinelibrary.wiley.com/doi/10.1111/insr.12469 (2025) (Apr. 2022).

72. Hastie, T., Tibshirani, R. & Wainwright, M. Statistical learning with sparsity: The lasso and generalizations. *Statistical Learning with Sparsity: The Lasso and Generalizations,* 1–337 (2015).

73. Ahmed, E., Belaghi, R. A., Asar, Y. & Hussein, A. *Ridge-Type Shrinkage Estimators in Low and High Dimensional Beta Regression Model with Application in Econometrics and Medicine* Version Number: 1. 2021. https://arxiv.org/abs/2111.13887 (2025).

74. López-Moreno, A., Torres-Sánchez, A., Acuña, I., Suárez, A. & Aguilera, M. Representative Bacillus sp. AM1 from Gut Microbiota Harbor Versatile Molecular Pathways for Bisphenol A Biodegradation. *International Journal of Molecular Sciences* **22,** 4952. ISSN: 1422-0067. https://www.mdpi.com/1422-0067/22/9/4952 (2023) (May 7, 2021).

75. Zhou, N. A., Kjeldal, H., Gough, H. L. & Nielsen, J. L. Identification of Putative Genes Involved in Bisphenol A Degradation Using Differential Protein Abundance Analysis of *Sphingobium* sp. BiD32. *Environmental Science & Technology* **49,** 12232–12241. ISSN: 0013-936X, 1520-5851. https://pubs.acs.org/doi/10.1021/acs.est.5b02987 (2025) (Oct. 20, 2015).

76. Hemavarshini, S. *et al.* Bacterial bioremediation as a sustainable strategy for the mitigation of Bisphenol-A. *Environmental Geochemistry and Health* **46,** 386. ISSN: 0269-4042, 1573-2983. https://link.springer.com/10.1007/s10653-024-02154-5 (2024) (Oct. 2024).

77. Razia, S., Hadibarata, T. & Lau, S. Y. A review on biodegradation of Bisphenol A (BPA) with bacteria and fungi under laboratory conditions. *International Biodeterioration & Biodegradation* **195,** 105893. ISSN: 09648305. https://linkinghub.elsevier.com/retrieve/pii/S0964830524001641 (2025) (Nov. 2024).

78. Li, G. *et al.* Biodegradation and detoxification of bisphenol A with one newly-isolated strain Bacillus sp. GZB: Kinetics, mechanism and estrogenic transition. *Bioresource Technology* **114,** 224–230. ISSN: 09608524. https://linkinghub.elsevier.com/retrieve/pii/S0960852412005317 (2023) (June 2012).

79. Noszczyńska, M., Chodór, M., Jałowiecki, Ł. & Piotrowska-Seget, Z. A comprehensive study on bisphenol A degradation by newly isolated strains Acinetobacter sp. K1MN and Pseudomonas sp. BG12. *Biodegradation* **32,** 1–15. ISSN: 0923-9820, 1572-9729. http://link.springer.com/10.1007/s10532-020-09919-6 (2022) (Feb. 2021).

80. Vijayalakshmi, V. *et al.* Bio-degradation of Bisphenol A by Pseudomonas aeruginosa PAb1 isolated from effluent of thermal paper industry: Kinetic modeling and process optimization. *Journal of Radiation Research and Applied Sciences* **11,** 56–65. ISSN: 16878507. https://linkinghub.elsevier.com/retrieve/pii/S1687850717301115 (2022) (Jan. 2018).

81. Tian, K. *et al.* Mechanisms of BPA Degradation and Toxicity Resistance in Rhodococcus equi. *Microorganisms* **11,** 67. ISSN: 2076-2607. https://www.mdpi.com/2076-2607/11/1/67 (2023) (Dec. 26, 2022).

82. Drzewiecka, M., Beszterda, M., Frańska, M. & Frański, R. 2,2-Bis(4-Hydroxyphenyl)-1-Propanol—A Persistent Product of Bisphenol A Bio-Oxidation in Fortified Environmental Water, as Identified by HPLC/UV/ESI-MS. *Toxics* **9,** 49. ISSN: 2305-6304. https://www.mdpi.com/2305-6304/9/3/49 (2023) (Mar. 5, 2021).

83. Eio, E. J., Kawai, M., Tsuchiya, K., Yamamoto, S. & Toda, T. Biodegradation of bisphenol A by bacterial consortia. *International Biodeterioration & Biodegradation* **96,** 166–173. ISSN: 09648305. https://linkinghub.elsevier.com/retrieve/pii/S0964830514002819 (2025) (Dec. 2014).

84. Gu, C. *et al.* Aerobic degradation of bisphenol A by Pseudomonas sp. LM-1: characteristic and pathway. *Biodegradation* **34,** 73–81. ISSN: 0923-9820, 1572-9729. https://link.springer.com/10.1007/s10532-022-10003-4 (2023) (Feb. 2023).

85. Eltoukhy, A. *et al.* Transcriptome analysis and cytochrome P450 monooxygenase reveal the molecular mechanism of Bisphenol A degradation by Pseudomonas putida strain YC-AE1. *BMC Microbiology* **22,** 294. ISSN: 1471-2180. https://bmcmicrobiol.biomedcentral.com/articles/10.1186/s12866-022-02689-6 (2023) (Dec. 9, 2022).

86. Corrales, J. *et al.* Global Assessment of Bisphenol A in the Environment: Review and Analysis of Its Occurrence and Bioaccumulation. *Dose-Response* **13,** 155932581559830. ISSN: 1559-3258, 1559-3258. http://journals.sagepub.com/doi/10.1177/1559325815598308 (2022) (July 1, 2015).

87. Tappert, L. *et al.* Bisphenol A in surface waters in Germany: Part I. Reassessment of sources and emissions pathways for FlowEQ modeling. *Integrated Environmental Assessment and Management* **20,** 211–225. ISSN: 1551-3777, 1551-3793. https://academic.oup.com/ieam/article/20/1/211/7725040 (2025) (Jan. 1, 2024).

88. Husain, K. & Murugan, A. Physical Constraints on Epistasis. *Molecular Biology and Evolution* **37** (ed Wilke, C.) 2865–2874. ISSN: 0737-4038, 1537-1719. https://academic.oup.com/mbe/article/37/10/2865/5839750 (2025) (Oct. 1, 2020).

89. Mazumder, R., Hastie, T. & Tibshirani, R. Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *Journal of machine learning research: JMLR* **11,** 2287–2322. ISSN: 1532-4435 (Mar. 1, 2010).

90. Cheng, A. G. *et al.* Design, construction, and in vivo augmentation of a complex gut microbiome. *Cell* **185,** 3617–3636.e19. ISSN: 00928674. https://linkinghub.elsevier.com/retrieve/pii/S0092867422009904 (2022) (Sept. 2022).

91. Sales Da Silva, I. G. *et al.* Soil Bioremediation: Overview of Technologies and Trends. *Energies* **13,** 4664. ISSN: 1996-1073. https://www.mdpi.com/1996-1073/13/18/4664 (2023) (Sept. 8, 2020).

92. Mahesh, N. *et al.* Trends and thresholds on bacterial degradation of bisphenol-A endocrine disruptor — a concise review. *Environmental Monitoring and Assessment* **194,** 886. ISSN: 0167-6369, 1573-2959. https://link.springer.com/10.1007/s10661-022-10558-y (2023) (Dec. 2022).

93. Van Veen, J. A., Van Overbeek, L. S. & Van Elsas, J. D. Fate and activity of microorganisms introduced into soil. *Microbiology and Molecular Biology Reviews* **61,** 121–135. ISSN: 1092-2172, 1098-5557. https://journals.asm.org/doi/10.1128/mmbr.61.2.121-135.1997 (2025) (June 1997).

94. Jiang, W. & Fan, W. Bioremediation of Heavy Metal–Contaminated Soils by Sulfate-Reducing Bacteria. *Annals of the New York Academy of Sciences* **1140,** 446–454. ISSN: 0077-8923, 1749-6632. https://nyaspubs.onlinelibrary.wiley.com/doi/10.1196/annals.1454.050 (2025) (Oct. 2008).

95. DNASTAR. *SeqBuilder Pro®* version 17.0. Madison, WI.

96. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421. ISSN: 1471-2105. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-421 (2025) (Dec. 2009).

97. Madeira, F. *et al.* The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024. *Nucleic Acids Research* **52,** W521–W525. ISSN: 0305-1048, 1362-4962. https://academic.oup.com/nar/article/52/W1/W521/7643289 (2025) (W1 July 5, 2024).

98. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research* **49,** W293–W296. ISSN: 0305-1048, 1362-4962. https://academic.oup.com/nar/article/49/W1/W293/6246398 (2025) (W1 July 2, 2021).

99. Modaressi, K., Taylor, K., Bewtra, J. & Biswas, N. Laccase-catalyzed removal of bisphenol-A from water: Protective effect of PEG on enzyme activity. *Water Research* **39,** 4309–4316. ISSN: 00431354. https://linkinghub.elsevier.com/retrieve/pii/S0043135405004562 (2022) (Nov. 2005).

100. De Jesús Astacio, L. M., Prabhakara, K. H., Li, Z., Mickalide, H. & Kuehn, S. Closed microbial communities self-organize to persistently cycle carbon. *Proceedings of the National Academy of Sciences* **118,** e2013564118. ISSN: 0027-8424, 1091-6490. https://pnas.org/doi/full/10.1073/pnas.2013564118 (2025) (Nov. 9, 2021).

101. López-Fanjul, C., Fernández, A. & Toro, M. A. EPISTASIS AND THE TEMPORAL CHANGE IN THE ADDITIVE VARIANCE-COVARIANCE MATRIX INDUCED BY DRIFT. *Evolution* **58,** 1655–1663. ISSN: 0014-3820, 1558-5646. https://academic.oup.com/evolut/article/58/8/1655/6756864 (2025) (Aug. 2004).

102. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **67,** 301–320 (2005).

103. O'Donoghue, B., Chu, E., Parikh, N. & Boyd, S. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications* **169,** 1042–1068 (2016).

104. Diamond, S. & Boyd, S. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research* **17,** 1–5 (2016).

105. Zhang, J., O'Donoghue, B. & Boyd, S. Globally convergent type-I Anderson acceleration for nonsmooth fixed-point iterations. *SIAM Journal on Optimization* **30,** 3170–3197 (2020).

106. Tarracchini, C. *et al.* Exploring the vitamin biosynthesis landscape of the human gut microbiota. *mSystems* **9** (ed Gilbert, J. A.) e00929–24. ISSN: 2379-5077. https://journals.asm.org/doi/10.1128/msystems.00929-24 (2025) (Oct. 22, 2024).

107. Devaraj, S., Hemarajata, P. & Versalovic, J. The Human Gut Microbiome and Body Metabolism: Implications for Obesity and Diabetes. *Clinical Chemistry* **59,** 617–628. ISSN: 0009-9147, 1530-8561. https://academic.oup.com/clinchem/article/59/4/617/5621987 (2025) (Apr. 1, 2013).

108. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214. ISSN: 0028-0836, 1476-4687. https://www.nature.com/articles/nature11234 (2025) (June 2012).

109. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568,** 499–504. ISSN: 0028-0836, 1476-4687. https://www.nature.com/articles/s41586-019-0965-1 (2025) (Apr. 25, 2019).

110. Spragge, F. *et al.* Microbiome diversity protects against pathogens by nutrient blocking. *Science* **382,** eadj3502. ISSN: 0036-8075, 1095-9203. https://www.science.org/doi/10.1126/science.adj3502 (2025) (Dec. 15, 2023).

111. Bassetti, M. & Peghin, M. How to manage KPC infections. *Therapeutic Advances in Infectious Disease* **7,** 204993612091204. ISSN: 2049-9361, 2049-937X. http://journals.sagepub.com/doi/10.1177/2049936120912049 (2022) (Jan. 2020).

112. Dong, Q. *et al.* Protection against Clostridioides difficile disease by a naturally avirulent strain. *Cell Host & Microbe* **33,** 59–70.e4. ISSN: 19313128. https://linkinghub.elsevier.com/retrieve/pii/S1931312824004098 (2025) (Jan. 2025).

113. Sorbara, M. T. *et al.* Inhibiting antibiotic-resistant Enterobacteriaceae by microbiota-mediated intracellular acidification. *Journal of Experimental Medicine* **216,** 84–98. ISSN: 0022-1007, 1540-9538. https://rupress.org/jem/article/216/1/84/42433/Inhibiting-antibioticresistant-Enterobacteriaceae (2022) (Jan. 7, 2019).

114. Organization, W. H. *Prioritization of pathogens to guide discovery, research and development of new antibiotics for drug-resistant bacterial infections, including tuberculosis* ().

115. Xiong, H. *et al.* Distinct Contributions of Neutrophils and CCR2^{+} Monocytes to Pulmonary Clearance of Different Klebsiella pneumoniae Strains. *Infection and Immunity* **83** (ed Young, V. B.) 3418–3427. ISSN: 0019-9567, 1098-5522. https://journals.asm.org/doi/10.1128/IAI.00678-15 (2025) (Sept. 2015).

116. Jiang, R., Tang, W., Wu, X. & Fu, W. A random forest approach to the detection of epistatic interactions in case-control studies. *BMC bioinformatics* **10 Suppl 1,** S65. ISSN: 1471-2105 (Suppl 1 Jan. 30, 2009).

117. Smith, P. F., Ganesh, S. & Liu, P. A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *Journal of Neuroscience Methods* **220,** 85–91. ISSN: 01650270. https://linkinghub.elsevier.com/retrieve/pii/S0165027013003026 (2025) (Oct. 2013).

118. Thompson, P. L., MacLennan, M. M. & Vinebrooke, R. D. An improved null model for assessing the net effects of multiple stressors on communities. *Global Change Biology* **24,** 517–525. ISSN: 1354-1013, 1365-2486. https://onlinelibrary.wiley.com/doi/10.1111/gcb.13852 (2025) (Jan. 2018).

119. Nurin Dureh, Attachai Ueranantasan & Mayuening Eso. A comparison of multiple linear regression and random forestfor community concern of youth and young adults survey. *Songklanakarin Journal of Science an Technology (SJST)* **44.** Publisher: Research and Development Office, Prince of Songkla University, 481487. http://doi.nrct.go.th/?page=resolve_doi&resolve_doi=10.14456/sjst-psu.2022.66 (2025) (2022).

120. Plerou, V. *et al.* Random matrix approach to cross correlations in financial data. *Physical Review E* **65,** 066126. ISSN: 1063-651X, 1095-3787. https://link.aps.org/doi/10.1103/PhysRevE.65.066126 (2025) (June 27, 2002).

121. Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell* **138,** 774–786. ISSN: 00928674. https://linkinghub.elsevier.com/retrieve/pii/S0092867409009635 (2025) (Aug. 2009).

122. Lutsiv, T. *et al.* Compositional Changes of the High-Fat Diet-Induced Gut Microbiota upon Consumption of Common Pulses. *Nutrients* **13,** 3992. ISSN: 2072-6643. https://www.mdpi.com/2072-6643/13/11/3992 (2025) (Nov. 9, 2021).

123. Oliveira, R. A. *et al.* Klebsiella michiganensis transmission enhances resistance to Enterobacteriaceae gut invasion by nutrition competition. *Nature Microbiology* **5,** 630–641. ISSN: 2058-5276. https://www.nature.com/articles/s41564-019-0658-4 (2025) (Jan. 20, 2020).

124. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659. ISSN: 1367-4811, 1367-4803. https://academic.oup.com/bioinformatics/article/22/13/1658/194225 (2025) (July 1, 2006).

125. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26,** 680–682. ISSN: 1367-4811, 1367-4803. https://academic.oup.com/bioinformatics/article/26/5/680/212234 (2025) (Mar. 1, 2010).

126. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120. ISSN: 1367-4811, 1367-4803. https://academic.oup.com/bioinformatics/article/30/15/2114/2390096 (2025) (Aug. 1, 2014).

127. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics* **70,** e102. ISSN: 1934-3396, 1934-340X. https://currentprotocols.onlinelibrary.wiley.com/doi/10.1002/cpbi.102 (2025) (June 2020).

128. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biology* **20,** 257. ISSN: 1474-760X. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1891-0 (2025) (Nov. 28, 2019).

129. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44,** D733–D745. ISSN: 0305-1048, 1362-4962. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1189 (2025) (D1 Jan. 4, 2016).

130. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36** (ed Hancock, J.) 1925–1927. ISSN: 1367-4803, 1367-4811. https://academic.oup.com/bioinformatics/article/36/6/1925/5626182 (2025) (Mar. 1, 2020).

131. Kuby, M. J. Programming Models for Facility Dispersion: The p -Dispersion and Maxisum Dispersion Problems. *Geographical Analysis* **19,** 315–329. ISSN: 0016-7363, 1538-4632. https://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1987.tb00133.x (2025) (Oct. 1987).

132. Erkut, E. The discrete p-dispersion problem. *European Journal of Operational Research* **46,** 48–60. ISSN: 03772217. https://linkinghub.elsevier.com/retrieve/pii/0377221790902970 (2025) (May 1990).

133. Oliveira, R. A. *et al.* Statistical design of a synthetic microbiome that clears a multidrug resistant gut pathogen. *bioRxiv,* 2024.02.28.582635. http://biorxiv.org/content/early/2024/02/29/2024.02.28.582635.abstract (Jan. 1, 2024).

134. Kouvaris, K., Clune, J., Kounios, L., Brede, M. & Watson, R. A. How evolution learns to generalise: Using the principles of learning theory to understand the evolution of developmental organisation. *PLOS Computational Biology* **13** (ed Rzhetsky, A.) e1005358. ISSN: 1553-7358. https://dx.plos.org/10.1371/journal.pcbi.1005358 (2025) (Apr. 6, 2017).

135. Goyal, A., Dubinkina, V. & Maslov, S. Multiple stable states in microbial communities explained by the stable marriage problem. *The ISME Journal* **12,** 2823–2834. ISSN: 1751-7362, 1751-7370. https://academic.oup.com/ismej/article/12/12/2823-2834/7475323 (2025) (Dec. 1, 2018).

136. Dubinkina, V., Fridman, Y., Pandey, P. P. & Maslov, S. Multistability and regime shifts in microbial communities explained by competition for essential nutrients. *eLife* **8,** e49720. ISSN: 2050-084X. https://elifesciences.org/articles/49720 (2025) (Nov. 22, 2019).

137. Amor, D. R., Ratzke, C. & Gore, J. Transient invaders can induce shifts between alternative stable states of microbial communities. *Science Advances* **6,** eaay8676. https://www.science.org/doi/full/10.1126/sciadv.aay8676 (2023) (Feb. 2020).

138. Darcy, J. L. *et al.* A phylogenetic model for the recruitment of species into microbial communities and application to studies of the human microbiome. *The ISME Journal* **14,** 1359–1368. ISSN: 1751-7362, 1751-7370. https://academic.oup.com/ismej/article/14/6/1359-1368/7474899 (2025) (June 1, 2020).

139. O'Dwyer, J. P., Kembel, S. W. & Green, J. L. Phylogenetic Diversity Theory Sheds Light on the Structure of Microbial Communities. *PLoS Computational Biology* **8** (ed Guttman, D. S.) e1002832. ISSN: 1553-7358. https://dx.plos.org/10.1371/journal.pcbi.1002832 (2025) (Dec. 20, 2012).

140. Cross, K., Beckman, N., Jahnes, B. & Sabree, Z. L. Microbiome metabolic capacity is buffered against phylotype losses by functional redundancy. *Applied and Environmental Microbiology* **91** (ed Rudi, K.) e02368–24. ISSN: 0099-2240, 1098-5336. https://journals.asm.org/doi/10.1128/aem.02368-24 (2025) (Feb. 19, 2025).

141. Allison, S. D. & Martiny, J. B. H. Resistance, resilience, and redundancy in microbial communities. *Proceedings of the National Academy of Sciences* **105,** 11512–11519. ISSN: 0027-8424, 1091-6490. https://pnas.org/doi/full/10.1073/pnas.0801925105 (2025) (supplement_1 Aug. 12, 2008).

142. Louca, S. *et al.* Function and functional redundancy in microbial systems. *Nature Ecology and Evolution* **2,** 936–943. ISSN: 2397334X (2018).

143. Moran, M. A. *et al.* Microbial metabolites in the marine carbon cycle. *Nature Microbiology* **7,** 508–523. ISSN: 2058-5276. http://www.ncbi.nlm.nih.gov/pubmed/35365785%20https://www.nature.com/articles/s41564-022-01090-3 (Apr. 2022).

144. Shan, X., Goyal, A., Gregor, R. & Cordero, O. X. Annotation-free discovery of functional groups in microbial communities. *Nature Ecology & Evolution* **7,** 716–724. ISSN: 2397-334X. https://www.nature.com/articles/s41559-023-02021-z (2025) (Mar. 30, 2023).

145. Crocker, K. *et al.* Environmentally dependent interactions shape patterns in gene content across natural microbiomes. *Nature Microbiology* **9,** 2022–2037. ISSN: 2058-5276. https://www.nature.com/articles/s41564-024-01752-4 (2025) (July 8, 2024).

146. Lee, K. K. *et al. Functional regimes define the response of the soil microbiome to environmental change* Mar. 17, 2024. http://biorxiv.org/lookup/doi/10.1101/2024.03.15.584851 (2025).

147. Yousef, M. *et al. Collective Microbial Effects Drive Toxin Bioremediation and Enable Rational Design* Mar. 28, 2025. http://biorxiv.org/lookup/doi/10.1101/2025.03.28.645802 (2025).