Model-based computational methods to aid the design of synthetic microbial communities

Alexander Eng

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Elhanan Borenstein, Chair

Samuel Miller

Wenying Shou

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Model-based computational methods to aid the design of synthetic microbial communities

Alexander Eng

Chair of the Supervisory Committee:
Elhanan Borenstein, Associate Professor
Department of Genome Sciences

Microbial communities permeate the world around us, inhabiting a range of environments where they play crucial roles in environmental resources cycles, agriculture, and human health, highlighted by recent interest in the human gut microbiome. These influences have marked microbial communities as potential tools for biotechnology and therapeutics to both enhance existing benefits and develop novel microbiome-based applications. This has led to the creation of several design techniques that manipulate these communities to optimize for specific beneficial functions, termed microbiome engineering. Previous engineering methods broadly fall into three categories: modulating the environment, modifying individual species' genomes, and constructing synthetic compositions of existing species. This last category has relied primarily on experimental approaches to select for or identify desirable compositions. However, the explosion of mechanistic knowledge regarding microbial function now offers an alternative approach to designing synthetic compositions, i.e., computational model-based tools that can evaluate and optimize compositions *in silico*. Here, I introduce computational methods that provide novel design capabilities by utilizing models of the link between microbial species and their functional capacities. In chapter

1, I review the current state of microbiome engineering techniques, their shortcomings, and the availability of mechanistic knowledge to support model-based design approaches. In chapter 2, I introduce CoMiDA, an algorithm for identifying minimal communities that provide sufficient metabolic capacities to perform a specific function. I then use CoMiDA to analyze the communities designed from common gut species for a wide array of metabolic tasks, finding that, when feasible, the designed communities tend to contain very few species. In chapter 3, I describe a perturbation-based framework for estimating the robustness of a community's functional capacities to perturbations in its species composition. I apply this framework to estimate measures of taxa-function robustness for communities from several human body and environmental sites, demonstrating that robustness varies depending on the source of the community. I also characterize the distribution of genes associated with functional orthology groups across community member genomes, showing that gene distribution features are related to robustness and revealing environment-specific signatures of gene distribution features. Finally, in chapter 4, I discuss other recently introduced computational design methods and describe the current challenges that face computational methods. I then propose potential avenues to advance the current state of computational methods, including rule-based frameworks that incorporate biological observations and integrated approaches that pair computational methods with complementary experimental techniques, which may provide comprehensive and powerful design tools for future microbiome engineering efforts.

ACKNOWLEDGEMENTS

I am extremely grateful to my family for all of their encouragement, and advice throughout my studies. Thank you for always showing an interest in my work and supporting my decisions. I also want to thank my friends, and particularly my classmates and fellow lab members, for making this experience significantly more enjoyable both in lab and beyond. Finally, I want to thank my advisor Elhanan for guidance and mentorship throughout this process. His high standards and dedication to good science have been an inspiration and were instrumental in my growth as a scientist. Furthermore, his help in better defining my research interests and his aid in identifying the more promising threads from among my ideas have made this work possible, and his interest in my success will always be greatly appreciated.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. Introduction

Sections of this chapter are based on the following manuscript submitted to *Current Opinion in Biotechnology*:

Eng A, Borenstein E. Microbial Community Design: Methods, Applications, and Opportunities. *Current Opinion in Biotechnology* Submitted.

Single-celled organisms, i.e., microbial species, permeate the world around us. These bacteria, archaea, and unicellular eukaryotes have found homes ranging from on and within the human body [1], to some of the most inhospitable or inaccessible habitats on the planet, such as deserts [2] or deep below the earth's surface [3]. Survival in these distinct and varied environments has necessitated extensive diversification, leading to specializations in individual microbial function that correspond to the unique circumstances of their surroundings. These specializations can encompass various tasks important to microbial life including metabolism [4], physical interaction with the environment [5], and response to extreme conditions [6]. Similarly, microbial communities, or mixtures of distinct species, also display a diversity in collective function that, perhaps most notably, encompasses the synthesis and degradation of a range of compounds [7–10]. These community-level functions result from the combination of functions performed by each individual species within the community as well as the complex interactions between those species. Via these diverse community functions, microbial communities play crucial roles in many settings including human health, industry, and the environment, suggesting that they could serve as promising and versatile tools for novel biotechnologies. Here I review some of the key benefits that microbial communities provide and existing approaches for engineering microbial community function.

1.1 The beneficial impacts of microbial communities

Microbial communities have myriad important and beneficial impacts on our lives, and perhaps the most immediate and personal influences result from the communities that inhabit our bodies. The human body is host to a multitude of microbial species, the human microbiome, with

estimates suggesting that the number of these microbial cells is comparable to the number of cells in the human body [11]. A person's microbiome begins to develop from birth when newborns are exposed to maternal and environmental species [12]. Evidence suggests that, even at this early stage of life, the microbiome plays an important role by supporting proper immune system development. Observational studies have noted that allergies developed during childhood are associated with early differences in fecal microbiota composition when compared to non-allergic children [13]. Moreover, it appears that specific early gut colonizers influence particular aspects of immune system development. For example, germ-free mice exhibit a deficiency in $CD4^+$ T cells, which are important for modulating various immune responses, but this deficiency is rescued when germ-free mice are colonized with *Bacteroides fragilis* [14]. In another case, *Enterococcus faecalis*, present in the newborn gut, was shown to help regulate PPARγ1, a transcription factor involved in inflammation, in a human colonic cell line [15]. These interactions with the immune system represent some of the significant benefits that the human microbiome can provide to its host.

After initial development, the gut microbiome continues to influence diverse aspects of our well-being throughout our lives. In terms of metabolism, the gut microbiome provides access to a variety of biosynthetic and digestive capabilities that we lack, including vitamin production [16] and the fermentation of dietary fiber to short chain fatty acids (SCFAs) [17]. SCFAs, such as acetate, propionate, and butyrate, serve a multitude of purposes in the gut ranging from appetite control to glucose regulation, and butyrate in particular is both an anti-inflammatory compound [18] and an important energy source for colonic cells [19]. Some of the metabolic benefits provided by the gut microbiome also become clear when normal function is disrupted, as in the case of obesity. The gut microbiomes of lean and obese individuals encode different functional capacities [20], and when lean or obese gut communities are transplanted into germ-free mice, the obese community recipients display relative increases in body mass and fat [21]. This negative effect of a disrupted gut microbiome demonstrates the importance of normal gut microbiome function in supporting host health. Besides metabolism, the healthy gut microbiome can also protect against pathogens through various mechanisms. For example, commensal gut bacteria can help the development of host immune responses that enable activation in response to pathogens, exclude pathogens via competition for nutrients, and produce antimicrobials that target pathogens [22].

Microbial communities also provide many benefits in industrial settings by supplying important biosynthetic and degradation capabilities. Communities present in municipal and industrial wastewater harbor species that can break down a range of pollutants under the right conditions, making them ideal for use in wastewater treatment plants [23]. In this setting, wastewater is passaged through tanks designed to promote the growth of these communities [24,25]. There, they remove assorted nitrogen-, phosphorus-, and carbon-containing compounds, and sometimes even more toxic byproducts of industrial processes [26]. Marine sediment communities, on the other hand, have displayed the ability to generate electrical current by reducing various elements [27], leading to the development of microbial fuel cells (MFCs), a biotechnology that uses microbial communities for electricity production [28,29]. MFC communities can consume a wide range of substrates and this has enabled the coupling of wastewater treatment and MFCs to generate electricity while also degrading pollutants [30,31]. Additionally, the biosynthetic capabilities of microbial communities have enabled numerous biotechnologies primarily focused on synthesizing biofuels such as methane [32], ethanol [33], and hydrogen [34], which offer promising renewable alternative fuel sources.

Environmental resource cycles and agriculture also benefit from the activity of microbial communities, which transform and recycle several elements, making them available to other organisms. For instance, the nitrogen cycle is an important process in which nitrogen is extracted from environmental reservoirs, utilized by organisms for essential functions, and released back into the environment. Microbial communities are a critical component to this process because they alone can “fix” atmospheric nitrogen gas, converting it into ammonia and thereby producing a more accessible nitrogen source [35]. This is especially important in agriculture because crop-accessible nitrogen comes from two main sources, nitrogen-fixing species and nitrogen-containing fertilizers, and nitrogen fixers offer the more environmentally sustainable alternative [36]. Similar to nitrogen, microbial species also contribute to the natural phosphorus cycle by decomposing organic matter to release phosphorus into the environment [37]. Some species also support crop growth via their ability to convert inorganic phosphates into plant-accessible soluble forms [38,39].

1.2 Microbiome engineering for improving community function

These varied and valuable functions suggest that microbial communities are a prime target for medical, industrial, and agricultural advancement and have sparked interest in the development of microbiome-based biotechnologies and therapeutics [40–42]. Specifically, microbiome engineering – the *in situ* manipulation of existing microbial communities or the construction of synthetic communities to achieve a specific community function – is a promising tool for improving and innovating upon various applications. Microbiome engineering relies on the ability to modulate community function, which can be done via three main approaches, altering a community's environmental conditions, editing the genomes of community members, or modifying a community's species composition.

1.2.1 Environment adjustment

One approach to influence community function is by controlling environmental conditions, and this has been applied to optimize community performance in industrial biotechnologies. Factors such as substrate availability, pH, and temperature can affect various aspects of microbial metabolism and protein expression [43–45] and thus can serve as adjustable operating parameters for engineers to tune overall community function. This has primarily been applied in industrial settings where operators have direct control over these factors. By systematically testing different environmental configurations, operators can identify conditions that optimize specific functions such as microbial hydrogen production [46,47], nitrification [48,49], and pigment production [50]. While this general approach is amenable to industrial applications, it is less suitable for therapeutic and agricultural technologies because precise control of operating conditions and comprehensive evaluation of potential environmental configurations are more difficult to achieve. In human gut microbiome therapeutics, for example, determining the optimal gut environment and how to appropriately achieve those conditions will depend in part on host genetics and how the individual's immune system interacts with the microbiome.

1.2.2 Genome editing

Another tool for altering microbial function is synthetic biology via genome editing, which relies on the modification, insertion, and/or deletion of DNA sequences in a species' genome and

thus may avoid the need for adjusting and maintaining specific environmental conditions. This technique has been applied to develop synthetic individual microbial strains that specialize in the production of various substances including biofuels [51,52] and plant compounds [53]. Synthetic strains are generally created by first identifying a metabolic pathway that converts some available feed material into a desired compound. If a strain that encodes the desired pathway does not already exist, then a strain of a well-studied species such as *Escherichia coli* or *Saccharomyces cerevisiae* is used instead [54]. In the latter case, missing components of the desired pathway are introduced by inserting the associated genes via either plasmids or direct genome editing. Once all genes associated with the desired pathway are present, the strain's genome is modified to overexpress this pathway to increase the desired activity. This engineered single strain technique has proven successful for creating organisms that can generate biofuels, pharmaceuticals, and food-related compounds at commercial scale [55], but solitary strains entail certain disadvantages. For instance, some multi-step metabolic processes may be challenging to perform efficiently in a single cell [56]. Additionally, operating conditions and substrate quality must often be kept to high, sometimes cost-prohibitive, standards because single strain systems may not be resilient to contamination of other species or changes in available substrate concentrations [57–59]. Thus, as with environmental manipulation, the synthetic single strain approach is likely unsuitable for broader applications beyond industrial bioreactors.

Synthetic ecology, or the engineering of multiple synthetic strains that form a community, is a related approach that can address some of the shortcomings of single strain engineering, including its reliance on tightly controlled operating conditions. In a community context, metabolic pathways can be distributed across multiple engineered strains, which reduces the genetic requirements for any one strain and can enable cooperative extra-cellular processes [60]. This metabolic compartmentalization has been used for applications such as contaminant removal [61] and certain fermentations [62], displaying increased efficiency compared to engineered single strain systems. Distributed metabolism can also improve utilization of multiple substrates, e.g., mixtures of different sugars, and can allow the community to respond to changes in substrate composition through shifts in community composition [63]. This resilience to substrate variability, as well as observed resistance to invasion by undesirable species [59], suggests that precise regulation of environmental conditions may be less crucial for synthetic communities, displaying yet another advantage over environmental manipulation and single strain techniques. Besides

distributed metabolism, synthetic strains can also be engineered to communicate and coordinate their individual functions by the intercellular transfer of signaling molecules and metabolites [64]. Due to these advantages, synthetic ecology is a promising tool for further developments in microbial community engineering. However, extensive effort is still required to design individual synthetic strains and community interactions. This may not scale well when creating larger, more complex communities that can be desirable due to the benefits of increased diversity observed in settings such as the gut microbiome [65] and soil communities [66].

1.2.3 **Synthetic composition design**

Microbial communities can also be engineered by designing synthetic compositions of existing microbial strains that perform or optimize a desired function. This alternative to synthetic strain engineering aims to exploit the available diversity in individual- and community-level microbial function by either manipulating natural community compositions or creating novel combinations of isolated strains from scratch. For example, by building upon natural communities, a resulting synthetic community can preserve important community functions that are not specifically the target of optimization. This will be important in therapeutics since the gut microbiome interacts with host health via a variety of mechanisms and synthetic therapeutic communities should maintain these positive interactions. These same tasks would be more challenging via genome engineering. More specifically, effectively altering existing communities via the introduction of synthetic strains will require consideration, and likely further design, of how the complex interactions these strains will have with other community members. Furthermore, designing entire communities from synthetic strains would, in addition to optimizing a targeted function, necessitate the identification and proper implementation of other necessary community functions in the final community. As an alternative, many synthetic composition design methods are able to maintain or approximate the high diversity of naturally occurring communities, unlike current genome editing approaches.

1.2.3.1 **Composition perturbation**

Perturbing a community's composition, whether through the addition of beneficial species or removal of undesirable ones, provides a simple tool for developing synthetic compositions based on existing communities. Antibiotics are perhaps the most prevalent example of this

approach in therapeutics, acting as a tool for removing pathogenic species [67]. Probiotics, on the other hand, are an example of an additive tool, aiming to improve gut microbiome function via supplementation with beneficial species [68,69]. Composition perturbation through species addition has also been used in biotechnology, where it is referred to as bioaugmentation, and has shown some success in improving wastewater treatment [70], biogas production [71], and treatment of contaminated soils [72]. However, these tools are not always ideal, especially in therapeutics, as antibiotics can be involved in subsequent increased pathogen susceptibility [73,74] and the persistence of probiotic strains can be inconsistent [75,76]. These shortcomings arise from a focus on specific species within a community, and more sophisticated methods that consider the entire community composition, may be more powerful.

1.2.3.2 **Community enrichment**

Existing community compositions can also be manipulated to achieve desired synthetic compositions by subjecting them to environmental conditions that favor the growth of species with particular functional capabilities. By selecting for these species, this technique, named community enrichment, aims to reach a community composition that has optimized the target function. To date, enrichment has primarily been used for biotechnological applications, including MFCs, biopolymer production, and hydrogen production. Marine sediment communities inoculated into MFCs often undergo compositional changes and exhibit gradual improvement in efficiency over time as the community adapts to operating in the MFC environment. Such changes were shown to include enrichment for species potentially related to current generation [77] and degradation of the supplied substrate [78], and the observed changes in community composition during extended operation of an MFC were demonstrated to be linked to concurrent increases in MFC efficiency [79].

While MFC communities experience inherent and appropriate selective pressures, enrichment for other biotechnologies may require the application of artificial selection procedures to optimize community function. For example, to increase the yield and efficiency of microbial communities grown and harvested for biopolymers used in biodegradable plastics, researchers have applied artificial feast-famine cycles [57,80]. A feast-famine cycle selects for communities that store energy (in the form of biopolymers) more efficiently during the feast phase so that energy is available during the famine phase. Artificial selection procedures have also improved microbial

community hydrogen production, though in this context, artificial section is often applied as a pretreatment rather than as part of post-enrichment operating conditions. Such pretreatments include heat shock, acidic or basic incubation, freeze drying, and chloroform treatment [81], which aim to enrich for hydrogen-producing species in the original community while excluding hydrogen consumers.

1.2.3.3 **Community reduction**

Some microbial community applications may impose specific restrictions on the species that can be present in the synthetic community. For example, microbiome therapeutics must meet various regulatory guidelines [82], and be devoid of pathogenic species so as to avoid inadvertently infecting the recipient [41]. However, it may not be possible to fully satisfy such restrictions using enrichment approaches due to the relatively broad and unspecified nature of environmental selection. For example, applying environmental conditions that inhibit the growth of pathogens may simultaneously negatively impact the growth of desirable species. This challenge can be addressed by a complementary design approach, referred to here as community reduction, wherein individual members of some initial community are isolated and characterized to rationally determine whether they should be used in the synthetic community. This approach provides better control over community composition and enables a more principled selection of desirable species and the explicit exclusion of undesirable ones. However, some community members may be lost during the isolation step due to our current inability to culture certain species [83], and thus there is some risk of missing community members with a crucial role in the target function.

This design paradigm has been used, for example, to reconstruct synthetic communities for treating *Clostridium difficile* infection (CDI). CDI, a gastrointestinal infection where *C. difficile*, a spore-forming, antibiotic-resistant enteric pathogen, dominates the gut microbiome and causes inflammation and diarrhea [84], has previously been effectively treated with fecal microbiota transplantation (FMT) from healthy donors [85]. In an attempt to recapitulate these treatments with engineered compositions, two studies have used community reduction to formulate synthetic CDI treatment communities [86,87]. In both studies, treatment displayed similar success, and in one case more success, compared to standard FMTs from donors. The community reduction approach has also shown success when used for non-CDI therapeutic applications. For example, rather than targeting *C. difficile*, a reduced gut community of ampicillin-resistant isolates demonstrated

resistance against invasion from vancomycin-resistant *Enterococcus* [88]. Reduced gut communities have also shown promising results for inducing T cells in mice [89,90], further illustrating the potential for community reduction to recapitulate various important gut microbiome functions and its promising future as a microbiome therapeutics design tool. One important caveat to note, however, is that community reduction inherently cannot design synthetic compositions with novel functions. This deficiency may restrict its wider applicability in augmenting the functional capacities of human-associated communities or developing new biotechnologies.

1.2.3.4 **Combinatorial evaluation**

One of the benefits of synthetic community compositions is that they can include combinations of species that have never been observed co-occurring in naturally occurring communities, potentially facilitating the novel coupling of diverse metabolic capacities. Such synthetic communities may therefore be able to outperform existing communities (or communities obtained via enrichment and reduction), and may even be capable of entirely new functions. As with engineered synthetic strains, this is of particular interest for industrial applications such as the production of biofuel and other biological compounds [91]. To go beyond a simple trial and error approach for identifying beneficial combinations, community engineers can employ a more comprehensive process referred to here as combinatorial evaluation. This entails the systematic enumeration, construction, and evaluation of possible combinations of a set of species to identify the best-performing composition. When the number of species to consider is small, combinatorial evaluation can be performed in its ideal form, constructing and assessing all possible combinations of the species of interest. For example, a synthetic community was optimized for dye degradation in textile wastewater by considering all possible combinations of the three candidate species, including variations in relative abundance [92].

Importantly, however, as the number of candidate species grows, the number of potential compositions grows exponentially, quickly rendering the evaluation of all possible combinations impossible. This setting calls for techniques that can drastically reduce the number of evaluated compositions. One such technique is fractional factorial design (FFD) [93], which can reduce the number of evaluated compositions by carefully selecting a subset of potential community compositions to isolate the contributions of specific species and inter-species interactions to the target community-level function. Microbial community function optimization via FFD has

historically focused on factors external to the community for environmental adjustment [94–96]. However, more recent efforts have used FFD to investigate the potential estimation of individual species effects. For example, FFD has been applied to develop wastewater treatment communities by estimating the species and interaction contributions to total organic carbon degradation and substrate utilization rate [97,98]. These contributions were then employed to identify optimal synthetic compositions for each of these two biodegradation tasks, determining that certain three- and four-strain communities performed better than a baseline mixture of all six evaluated strains.

Another technique for efficiently evaluating potential compositions is the definition of microbial consortia that will be treated as single units when enumerating possible species combinations. More specifically, each combination will either include or exclude all species in a given consortium. This technique is particularly useful when a microbial consortium has previously demonstrated a desirable emergent function. For instance, a consortium of marine species, named the NPMC, can efficiently fix CO_2 [99], and has been used as a single candidate community member in combinatorial evaluation to develop a synthetic community for CO2 fixation [100]. In addition to reducing the pool of available species to six candidate community members (one of which was the NPMC), this approach also enabled the inclusion of species that could not be isolated from the NPMC in the final community. In another example, a synthetic community was designed for lignocelluloytic enzyme activity by considering a synthetic consortium previously designed for cellulolytic activity [101] alongside several fungal strains [102]. These studies highlight the benefits of this approach, which allows engineers to evaluate higher complexity communities without drastically increasing the number of evaluated compositions.

1.2.3.5 **Growing support for computational model-based design methods**

The synthetic composition design paradigms described above focus on experimental methods for identifying optimal compositions, but prior ecological, genomic, and metabolic information regarding microbial function can enable complementary computational approaches. Databases such as NCBI [103] and IMG [104] provide access to an ever increasing number of sequenced microbial genomes, which can be used to identify the genes each species encodes. This information can be employed to infer the functional capacities of each species when coupled with various curated gene annotation databases such as KEGG [105] and MetaCyc [106], which link

genes and protein domains to functions using literature-based evidence. These links between species and function can support model-based methods to predict and evaluate individual- and community-level function *in silico*, allowing engineers to rationally design synthetic compositions prior to experimental validation. For example, knowledge of individual species functional capacities could be used to identify species combinations that collectively encode sufficient functional capacities to support a desired community function. The stoichiometry of metabolic reactions encoded by a species can also aid in predicting function, as has previously been used in constraint-based modeling and flux balance analysis (FBA) [107,108]. Such models can predict the steady state metabolic activity of a given species by identifying a set of metabolic fluxes that maximize microbial growth while adhering to a set of thermodynamic constraints [109,110]. Predicting community function in this way could serve to focus experimental efforts on synthetic compositions that are expected to optimize the desired function, thereby reducing time and labor required to engineer synthetic communities.

1.3 Development of computational methods for synthetic community composition design

Despite the availability of resources for linking taxonomic composition to community function, computational methods to aid synthetic composition design are lacking. Model-based methods can provide two major benefits to microbiome engineering: rapid identification of optimized synthetic compositions based on predicted function, and evaluation of community characteristics that are challenging to comprehensively assess experimentally. Methods that address this first task can help reduce the experimental effort required to engineer synthetic communities. On the other hand, methods that support this second goal instead offer insight into new facets of microbial community function, which engineers can then consider during the design process. In this dissertation, I introduce two model-based computational methods that provide support in these areas by utilizing models linking microbial species to their functional capacities. The objective of the first method is to identify minimal sets of species that collectively encode the functional capacities for desired community-level metabolic functions, which can provide a starting point for further community engineering. This method also offers the opportunity to explore the community composition constraints imposed by different metabolic functions, and I apply this to examine minimal communities of human gut species. The objective of the second

method is to evaluate the robustness of a community's functional capacities with respect to changes in its taxonomic composition, which can provide novel insights into the consequences of community structure. One question is whether this taxa-function robustness varies by environment, which I investigate by using this method to characterize robustness in communities from human-associated, soil, and marine sources.

2. An algorithm for designing minimal microbial communities with desired metabolic capacities

This chapter is based on the following manuscript published in *Bioinformatics*:

Eng A, Borenstein E. An algorithm for designing minimal microbial communities with desired metabolic capacities. *Bioinformatics* 2016; 32:2008-2016

2.1 Summary

2.1.1 Motivation

Recent efforts to manipulate various microbial communities, such as fecal microbiota transplant and bioreactor systems' optimization, suggest a promising route for microbial community engineering with numerous medical, environmental, and industrial applications. However, such applications are currently restricted in scale and often rely on mimicking or enhancing natural communities, calling for the development of tools for designing synthetic communities with specific, tailored, desired metabolic capacities.

2.1.2 Results

Here, I present a first step toward this goal, introducing a novel algorithm for identifying minimal sets of microbial species that collectively provide the enzymatic capacity required to synthesize a set of desired target product metabolites from a predefined set of available substrates. This method integrates a graph theoretic representation of network flow with the set cover problem in an integer linear programming framework to simultaneously identify possible metabolic paths from substrates to products while minimizing the number of species required to catalyze these metabolic reactions. I apply this algorithm to successfully identify minimal communities both in a set of simple toy problems and in more complex, realistic settings, and to investigate metabolic capacities in the gut microbiome. This framework adds to the growing toolset for supporting informed microbial community engineering and for ultimately realizing the full potential of such engineering efforts.

2.1.3 Availability

The algorithm source code, compilation, usage instructions, and examples are available under a non-commercial research use only license at https://github.com/borenstein-lab/CoMiDA.

2.2 Background

Complex microbial communities can be found everywhere on our planet, spanning marine communities inhabiting the deep ocean to symbiotic communities living on and within host organisms. These communities impact a broad set of processes ranging from environmental resource cycles to host organism health. For example, deep sea rock and vent communities play a fundamental role in oxidizing environmental methane [111], whereas the human gut microbiome crucially aids in drug metabolism, energy harvest, and immune system response [112]. Microbial communities affect these processes through a variety of metabolic reactions catalyzed by enzymes encoded in the member species' genomes, and ultimately through the diverse compounds each community can degrade or produce.

These critical roles microbial communities play in shaping their environment, combined with the potential to manipulate these communities, suggest a promising route for numerous medical and environmental applications [40]. Specifically, several such efforts to shift target communities toward preferred states have used samples from naturally occurring communities as an inoculation source. For instance, transplanting healthy donor microbiome samples into patient guts has recently been used to treat a variety of gut disorders [113]. Such fecal microbiota transplants (FMTs) have been shown to perturb a patient's dysbiotic gut community, shifting it to a healthier state and ameliorating their condition [114,115]. These FMT-based therapies have had a >90% success rate at curing recurrent *Clostridium difficile* infections and have promising results for addressing other gut disorders including inflammatory bowel disease and metabolic syndrome [85]. Similarly, wastewater treatment bioreactors are often seeded by microbial communities cultivated from naturally occurring wastewater microbes or from previously established bioreactors [24]. These seed communities colonize the new bioreactor and thereby provide the metabolic processes necessary to degrade biological matter in wastewater.

Following the success of such transplants, recent efforts have further aimed to use engineered, rather than naturally occurring, communities in an attempt to increase control over

transplant outcomes. For example, a synthetic stool substitute was recently developed using a mixture of cultured bacterial isolates to mimic a healthy gut community [87]. Such a synthetic community removes the need for sample donors, allows greater regulation over the bacteria present in the transplant community, and reduces the risk for inadvertent transfer of pathogens. This synthetic and markedly simpler community was shown to still be effective in treating *C. difficile* infections. Another effort applied a simple selection-based approach to optimize the species composition of a bioreactor seed community for increased biopolymer production from glycerol [80]. The final community's biopolymer production rate was demonstrated to be noticeably increased compared to the original community.

Such community engineering approaches are clearly an important first step towards customizing microbial community composition, yet they still largely rely on imitating natural community structures or enhancing existing community capabilities and cannot, for example, produce communities with potentially desired abilities absent from the initial community. Indeed, even optimizing an existing community function involves developing a carefully controlled selection procedure tailored to the preferred function and may require a long time for the community to reach an optimal state. The applications of such engineering efforts are therefore inherently constrained and are often very system-specific and hard to generalize.

One approach to address these challenges is to rationally design and construct synthetic communities with desired and predefined metabolic capabilities. Such a design process would involve the careful selection of member species and their abundances, hopefully defining a community composition that would achieve the desired metabolic task in the target environment. The ability to design such communities would significantly broaden the applicability of community engineering, could alleviate the reliance on naturally occurring community functions, and would ultimately support the construction of communities tailored to perform specific tasks within the context of various environmental settings.

Designing microbial communities, however, is a daunting task. Microbial species are endowed with tremendously diverse and complex capacities, which may not be trivial or easy to discern. Moreover, the various species comprising each community do not function independently, and each community impacts its environment through the orchestrated activity of its members. Interaction between species can lead to emergent behaviors that cannot be attributed to the function of just a single species or to additive species functions [116,117]. One species can, for example,

provide the necessary precursors that allow a second species to produce metabolites that it could not produce when growing in isolation [118]. Similarly, costly metabolic tasks could be distributed among community members such that each member performs a specific part of a complex metabolic pathway [119]. A successful design framework should therefore account for such interactions and their impact on the metabolism of the community as a whole.

As a first step to address this challenge, here I present CoMiDA (Community Metabolism Design Algorithm), an algorithmic framework for designing simple communities with some predefined metabolic capacities. Specifically, this algorithm aims to identify a set of species that, as a community, has the metabolic potential to convert a set of metabolic substrates to a set of desired target metabolites. It further aims to discern the smallest set of species required to provide this desired metabolic potential, reducing downstream complexities and providing more streamlined communities. In other words, the goal is to identify a minimal set of species whose genomes collectively encode a set of enzymatic genes that can catalyze a collection of metabolic reactions forming metabolic paths to each desired product metabolite from the available substrates.

Communities designed with this framework will therefore have the required metabolic potential to achieve the specified metabolism. Obviously, there are additional factors and processes that should be ultimately considered in designing a stable and functional community that carries out a specific task. First and foremost, possessing the set of reactions leading from substrates to products does not necessarily imply that the community would actively and efficiently perform the desired metabolic function. Toxin production, signaling between microbes, the capacity to transport metabolites between cells, and the ability of the selected species to survive in the target environment could further affect the community behavior, stability, and dynamics. Yet, having the metabolic potential to carry out the desired function is an important and essential *prerequisite* for any community that could achieve the specified task, and is therefore a natural first step in rational community design and a critical component of any design task (see also 2.5 Discussion).

2.3 Methods

2.3.1 Problem statement and approach

The goal of this design task is, given a set of substrate metabolites and a set of target product metabolites, to find a minimal subset of the available species that can collectively

synthesize this set of target products using the available substrates. Specifically, each microbial species is viewed as a simple assemblage of metabolic reactions, corresponding to the set of enzymatic genes encoded in its genome. Each reaction is represented as a hyperedge, linking the reaction's substrates to the reaction's products. Furthermore, we will initially assume that metabolites can transfer freely between species, a common simplifying assumption in various community models [120–123], though we will relax this assumption later. The metabolic potential of each community can accordingly be viewed as the aggregate set of metabolic reactions of the member species. A solution to this design task is therefore a minimal set of species that collectively include some set of metabolic reactions sufficient to form valid paths to all target products from available substrates. This design task is depicted in Figure 2.1.

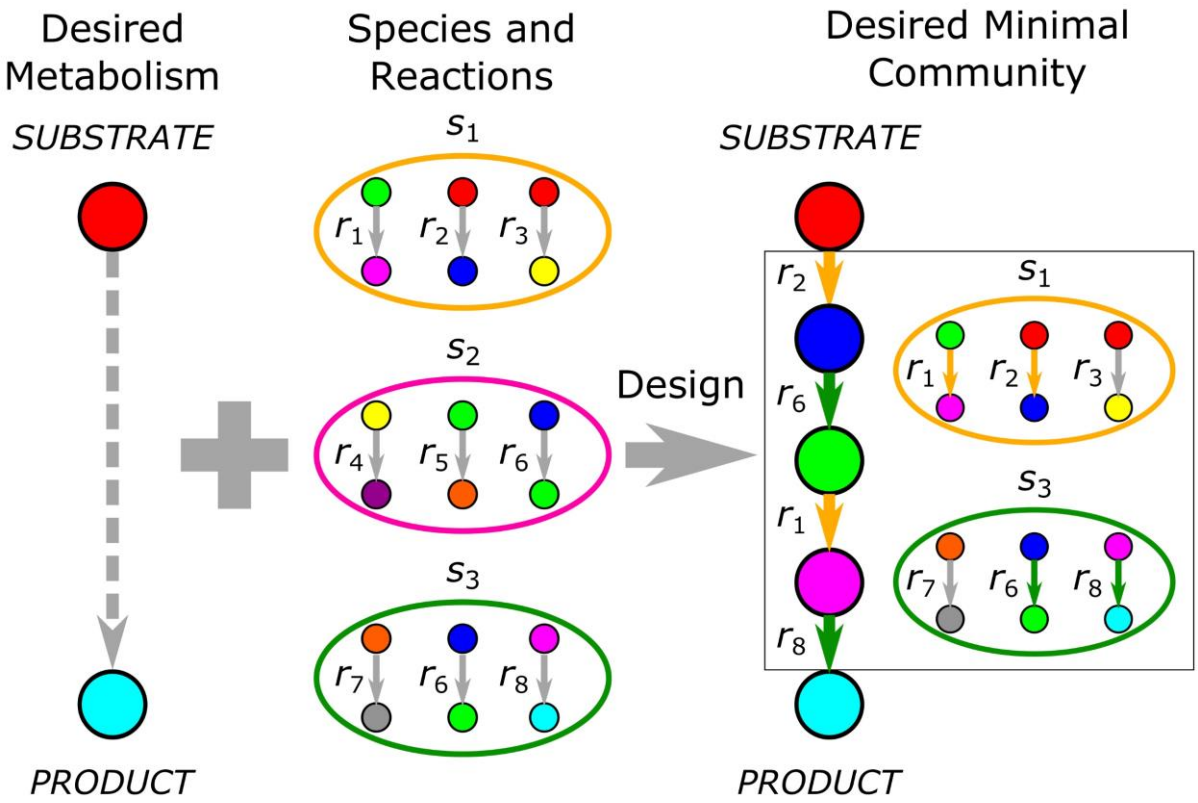

Figure 2.1. Schematic representation of the design task. Circles represent metabolites, with arrows between metabolites representing metabolic reactions and ovals representing species. The presence of a reaction within a species indicates that this species can catalyze that reaction. Given desired products and available substrates (left) and a set of available species (middle), CoMiDA aims to identify a minimal subset of species that can collectively synthesize the desired products from the available substrates (right).

To solve the above design problem, I used an integer linear programming (ILP) formulation. ILP is a framework for defining a linear expression of variables to maximize/minimize, along with a set of linear equations and inequalities that constrain those variables. ILP is a well-established framework, with several efficient solvers and numerous applications [124].

Below, I introduce an ILP formulation of this design task, inspired by ILP-based solutions to both the set cover problem and the network flow problem. To outline the different conceptual parts of the algorithm, I construct this ILP formulation in multiple steps. I first assume that all reactions are simple (connecting a single substrate to a single product) and that a set of reactions necessary to form paths from available substrates to all target products is specified. I show that, with these assumptions, identifying a minimal set of species that collectively encode this required set of reactions can be represented as a *set cover* problem and solved using an ILP formulation. Next, I relax the assumption of specified paths (or a specified set of reactions), introducing an array of *network flow*-inspired ILP constraints that defines possible paths from available substrates to target products using terms that can be linked to the set cover formulation. Finally, I consider the presence of multiple-substrate multiple-product reactions and adjust the network flow constraints to account for such edges with multiple inputs and multiple outputs, or hyperedges, in the metabolic network.

2.3.2 **Species, reactions, and metabolites in a simple metabolic network representation**

There are three main components to this community design problem: the set of available species, the metabolic reactions catalyzed by each species, and the metabolites these reactions consume and produce. Let $M = \{m_1, m_2, \dots, m_n\}$ denote the set of possible metabolites where *n* is the number of metabolites. We additionally define $R = \{r_1, r_2, \dots, r_p\}$ to be the set of reactions, where p is the number of reactions. Each reaction can then be defined as an ordered pair of metabolites, representing the reaction's substrate and product, respectively:

$$r_j = (m_{j\_substrate}, m_{j\_product}).$$

For now, assume that each reaction has one substrate metabolite and one product metabolite. This assumption will be relaxed later. Similarly, let $S = \{s_1, s_2, \dots, s_q\}$ denote the set

of species, where q is the number of species. Each species, s, in this formulation can be defined as the set of reactions it can catalyze:

$$s_i = \{r_{i\_1}, r_{i\_2}, \dots, r_{i\_a}\}.$$

We additionally define the set of available substrate metabolites and set of target product metabolites as:

$$SUBSTRATE = \{m_{substrate\_1}, m_{substrate\_2}, \dots, m_{substrate\_b}\},$$

$$PRODUCT = \{m_{product\_1}, m_{product\_2}, \dots, m_{product\_c}\},$$

where b is the number of substrates and c is the number of products. Notably, with these definitions, metabolites and reactions can also be viewed as a graph or a network, where nodes represent metabolites and edges represent reactions connecting substrates to products. Notice also that each species can be associated with some subgraph of this metabolic graph based on the set of reactions that species can catalyze.

2.3.3 Finding a minimal set of species with a pre-specified collection of metabolic capacities

To focus on the minimization aspect of the algorithm, first assume that there is a specified set of necessary metabolic reactions, $N \subseteq R\text{P} \subseteq \text{R}$, that provides valid paths from $SUBSTRATE$ to $PRODUCT$, such as the set of metabolic reactions in Figure 2.1 (right). Given this assumption, the aim is to identify a solution set of species that both can collectively catalyze this set of necessary reactions, and is minimal (in terms of the number of species). Since each species is viewed as some subset of the possible reactions, this task corresponds to identifying the minimal set of such subsets whose union contains the specified set of necessary reactions N. This representation of the task forms an instance of the well-defined set cover (SC) problem, which can be solved using an ILP formulation [125]. Specifically, first we define a set of binary ILP species variables $I\_S = \{I\_s_1, I\_s_2, \dots, I\_s_q\}$ such that each ILP species variable, I_s, corresponds to a species s:

$$I\_s_i \in \{0,1\} : i \in \{1,2, \dots, q\},$$

with $I\_s_i = 1$ indicating that the ith species is included in the solution species set, and $I\_s_i = 0$ indicating that the ith species is not included. Given these ILP species variables, the objective function of minimizing the number of species included in the solution species set can be defined as:

$$\min \sum_{i=1}^{q} I_{s_i}. \tag{2.1}$$

To link the set of species to be included in the solution set to the sets of reactions each species can catalyze and the set of specified necessary reactions N, we define an additional set of binary ILP reaction variables $I\_R = \{I\_r_1, I\_r_2, \dots, I\_r_p\}$ such that each ILP reaction variable, I_r, corresponds to a reaction r:

$$I\_r_j \in \{0,1\}: j \in \{1,2, \dots, p\},$$

with $I\_r_j = 1$ indicating that the jth reaction is included in P, and $I\_r_j = 0$ indicating that the jth reaction is not included. Given these ILP reaction variables, the constraints ensuring that each necessary reaction can be catalyzed by at least one species can be defined as:

$$\sum_{\substack{\forall i \ s.t. \\ r_j \in s_i}} [I_{s_i}] \geq I_{r_j} : j \in \{1,2, \dots, p\}. \tag{2.2}$$

In other words, these constraints require that if a reaction is necessary ($I\_r_j = 1$), then there must be at least one species in the solution species set that catalyzes that reaction. The objective function (2.1) and the set of constraints (2.2) thus fully define an ILP formulation of the SC component of the algorithm, minimizing the number of species required to catalyze a known set of necessary metabolic reactions.

As a brief example of such a formulation, consider the set of available species in Figure 2.1 (middle) and the set of reactions in the displayed solution (right). By appropriately assigning ILP species variables, the objective function for this instance would be:

$$\min[I\_s_1 + I\_s_2 + I\_s_3].$$

Similarly, when we assign the ILP reaction variables and their values, we can formulate the set cover constraints associated with this instance (following the general form of constraint (2.2)):

$$\begin{array}{llllll} I\_r_2 = 1 \rightarrow & I\_s_1 & & & & \geq 1 \\ I\_r_6 = 1 \rightarrow & & I\_s_2 & + & I\_s_3 & \geq 1 \\ I\_r_1 = 1 \rightarrow & I\_s_1 & & & & \geq 1 \\ I\_r_8 = 1 \rightarrow & & & & I\_s_3 & \geq 1 \end{array} .$$

Together, this specific objective function and these specific constraints define ILP problem associated with the task depicted in Figure 2.1 assuming the reactions in the presented path are necessary.

2.3.4 **Considering all possible paths from available substrates to target products**

When defining the SC component of the algorithm above, a set of necessary metabolic reactions connecting $SUBSTRATE$ to $PRODUCT$ was assumed to be predefined. Clearly, however, when considering the complex network of metabolic reactions that can be catalyzed by microbial species, there are likely numerous alternative paths connecting the available substrates to the desired target products. Since one cannot know *a priori* which paths require the fewest species to catalyze, a complete solution to this design problem must consider all possible paths when minimizing the number of species. To address this challenge and to remove the assumption of a specified set of necessary reactions, we use network flow (NF)-inspired constraints. Specifically, instead of predefining the values of the I_r variables to denote which reactions are necessary, we allow I_r values to vary freely and introduce a set of constraints that guarantee that the collection of reactions for which $I\_r = 1$ form valid paths from $SUBSTRATE$ to $PRODUCT$. Intuitively, an NF problem considers a graph as a network of pipes where the task is to push the maximal flow through these pipes from a source node to a sink node. Here, we use this NF-based approach to define a valid path in the metabolic network as a set of reactions that allow flow to pass from $SUBSTRATE$ to $PRODUCT$.

To define such a valid path, first we define a set of ILP flow variables, $F\_R = \{F\_r_1, F\_r_2, \dots, F\_r_p\}$, where F_r_j denotes the amount of flow passing through the jth reaction. Since flow has to be non-negative and since real-valued flow variables are unnecessary and slow computation, we further limit the values for flow variables to non-negative integers:

$$F\_r_j \in \mathbb{N}: j \in \{1,2, \dots, p\}.$$

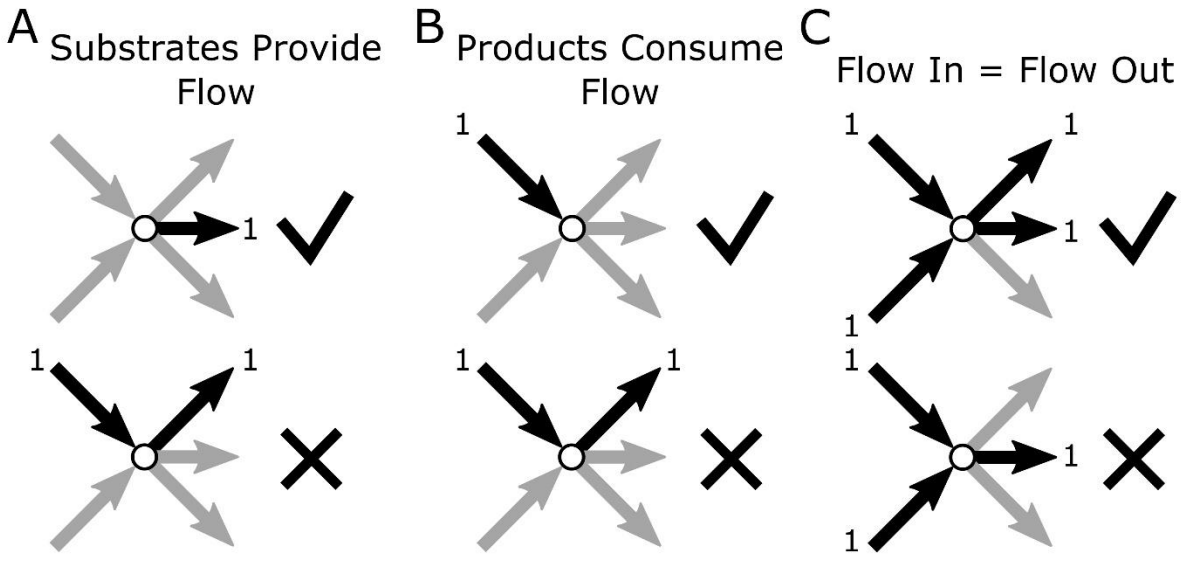

Figure 2.2. The network flow constraints. (A) The net flow out of all available substrates must be equal to the number of target products. **(B)** The net flow into any target product must be equal to 1. **(C)** The net flow for any intermediate metabolite must be 0. Together these constraints define any viable set of metabolic reactions that form paths from available substrates to target products.

The first NF constraint requires that only metabolites in $SUBSTRATE$ can be sources of flow, hence forcing any viable path to start from an available substrate metabolite (Figure 2.2A):

$$\sum_{\substack{\forall j\ s.t. \\ m_j \in SUBSTRATE}} \left[- \sum_{\substack{\forall in\ s.t. \\ r_{in}=(m_i,m_j)}} F_{r_{in}} + \sum_{\substack{\forall out\ s.t. \\ r_{out}=(m_j,m_k)}} F_{r_{out}} \right] = |PRODUCT|, \qquad (2.3)$$

where $i, j, k \in \{1,2,\dots,n\}, i \neq j, j \neq k$, and $in, out \in \{1,2,\dots,p\}$. In other words, the sum of flow leaving all available substrate metabolite nodes must be greater than the sum of flow entering these metabolites. Specifically, we require that the difference in flow be equal to the number of target product metabolites ($|PRODUCT|$), ensuring that each target product can receive one unit of flow if a viable path exists. Note here that even though flow would not usually need to enter a substrate metabolite node, it may be necessary for problems involving forced substrate usage (see 6.1 Supplementary Text).

The second NF constraint requires that metabolites in $PRODUCT$ be flow sinks, forcing every viable path to end at a target product (Figure 2.2B):

$$\sum_{\substack{\forall in\ s.t. \\ r_{in}=(m_i,m_j)}} F_{r_{in}} - \sum_{\substack{\forall out\ s.t. \\ r_{out}=(m_j,m_k)}} F_{r_{out}} = 1 : \forall j\ s.t.\ m_j \in PRODUCT. \quad (2.4)$$

This forces the flow into any target metabolite node to be greater than the flow leaving that node by one unit of flow. Thus, each product metabolite must be reached by some viable path. It should be noted that the network flow solution does not necessarily reflect *all* metabolic activity and, for example, intermediate reactions' by-products could still be generated even when no flow is associated with these by-products.

The third NF constraint asserts that all metabolites not in $SUBSTRATE$ or $PRODUCT$ have zero net flow (i.e., neither sources nor sinks of flow), allowing such metabolites to serve as intermediate nodes in any viable path (Figure 2.2C):

$$\sum_{\substack{\forall in\ s.t. \\ r_{in}=(m_i,m_j)}} F_{r_{in}} = \sum_{\substack{\forall out\ s.t. \\ r_{out}=(m_j,m_k)}} F_{r_{out}} : \forall j\ s.t.\ m_j \in R, m_j \notin SUBSTRATE \cup PRODUCT. \quad (2.5)$$

Given these NF constraints, any viable set of metabolic reactions for converting $SUBSTRATE$ to $PRODUCT$ will have non-zero associated ILP flow variables.

Finally, to appropriately set the I_r reaction variables to 1 if the reaction is used in the NF task and 0 otherwise, an additional set of conversion constraints is added:

$$F_{r_j} \leq |PRODUCT| \times I_{r_j} : \forall j \in \{1,2,\dots,p\}, \quad (2.6)$$

ensuring that if a reaction's flow variable is greater than 0, then the ILP reaction variable for that reaction must be 1.

Combining the objective function (2.1) and the constraints (2.2)-(2.6) for the SC and NF tasks therefore provides a complete ILP formulation for minimizing the number of species in the solution species set while ensuring the existence of viable paths from available substrates to all target products.

2.3.5 **Allowing for metabolic reactions with multiple substrates and products**

The ILP formulation above relies on the assumption that each metabolic reaction has a single substrate and a single product. This is a common simplification in metabolic network analysis and various protocols exist to reconstruct metabolic networks in which this assumption

holds [126,127]. Yet, a more complete and more accurate metabolic network formulation allows metabolic reactions to have multiple substrates (accounting, for example, for co-factors) and/or multiple products. To account for such metabolic reactions, we modify the metabolic network representation and instead of connecting substrate nodes to product nodes directly, we introduce a new type of node, representing reactions, and connect the (potentially multiple) substrates of each reaction to the (potentially multiple) products through the appropriate reaction node (Figure 2.3A). In this representation, a reaction, r, is therefore no longer representing an edge in the network, but rather a node that connects

$$\{m_{j\_substrate\_1}, m_{j\_substrate\_2}, \dots, m_{j\_substrate\_d}\}$$

to

$$\{m_{j\_product\_1}, m_{j\_product\_2}, \dots, m_{j\_product\_e}\},$$

where d is the number of substrates and e is the number of products for the jth reaction. $\{\mathrm{m_{j_product_1}}, \mathrm{m_{j_product_2}}, \dots, \mathrm{m_{j_product_e}}\}$Specifically, we define two new classes of edges: a set of reaction input edges, $I = \{i_1, i_2, \dots, i_t\}$ where t is the number of substrate metabolites across all reactions, connecting a substrate metabolite m to a reaction r:

$$i_j = (m_{j\_input}, r_{j\_reaction}),$$

and a set of reaction output edges, $O = \{o_1, o_2, \dots, o_u\}$ where u is the number of product metabolites across all reactions, similarly connecting a reaction r to its product metabolite m:

$$o_j = (r_{j\_reaction}, m_{j\_output}).$$

Together, these new nodes and edges thus define a bipartite graph where edges only exist between one metabolite node and one reaction node, but never between two metabolites or two reactions (Figure 2.3A).

Now we redefine the set flow variables in this network as two sets, $F\_I = \{F\_i_1, F\_i_2, \dots, F\_i_p\}$ and $F\_O = \{F\_o_1, F\_o_2, \dots, F\_o_p\}$ where F_i and F_o represent flow along input and output edges respectively. Notably, most of the flow constraints defined above are still valid when applied to both metabolite and reaction nodes; however, constraint (2.6), which aimed to link the flow variables to the ILP reaction variables I_r, needs to be updated to represent the link between reactions' multiple substrates and products. Specifically, one set of constraints is

introduced to ensure that a reaction can be active only if all its substrates are present (in other words, if all reaction input edges have flow):

$$F_{i_j} \geq I_{r_k} : \forall j, k \; s.t. \; i_j = (m_l, r_k). \tag{2.7}$$

Then, another set of constraints is introduced to allow active reactions to generate products (by providing flow on the reaction output edges):

$$\sum_{\substack{\forall i \; s.t. \\ o_i = (r_j, m_k)}} \left(F_{o_i}\right) \leq (|I| + |O|) \times I_{r_j} : \forall j \in \{1,2, \dots, p\}. \tag{2.8}$$

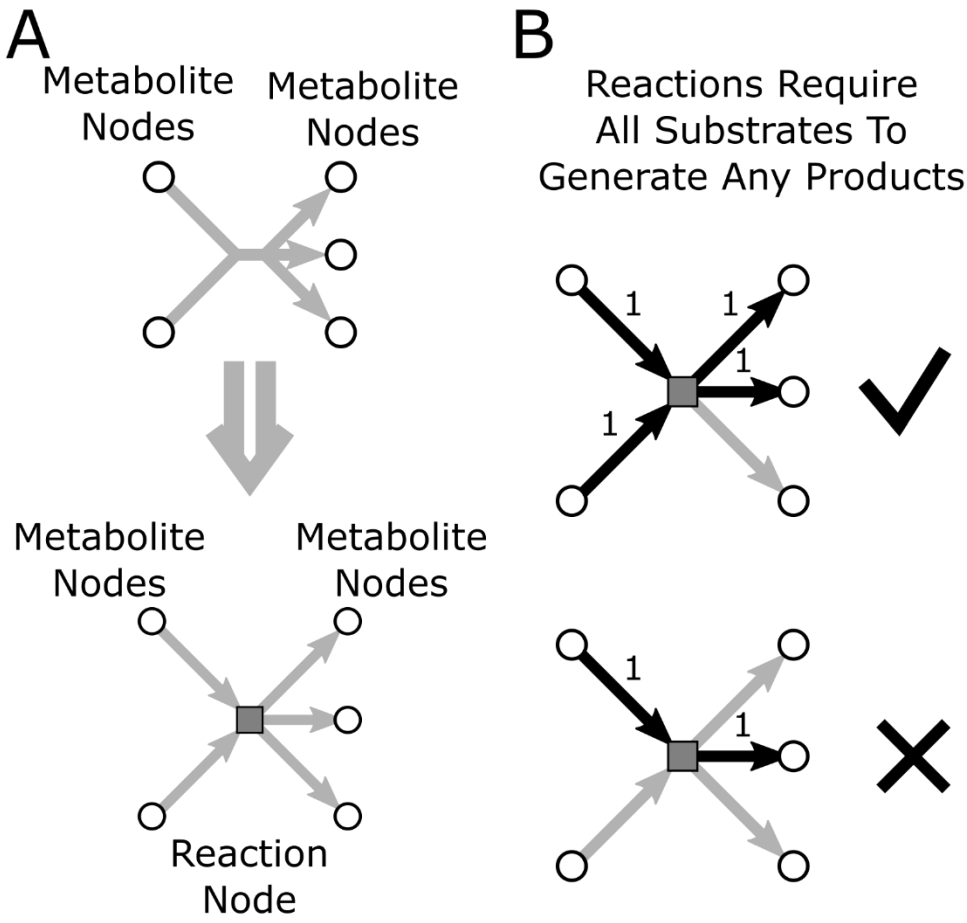

Figure 2.3. The bipartite graph representation of multi-substrate, multi-product metabolic reactions and associated flow constraints. **(A)** Each metabolic reaction is replaced by a new node type representing the enzyme that catalyzes that reaction. New edges are added to indicate the input and output metabolites for each reaction. **(B)** The new flow constraints require that all reaction substrates can provide flow to use a reaction and that reaction output edges can have flow only if all input edges have flow.

The only difference between constraints (2.6) and (2.8) is that the maximum flow across a single edge is no longer bounded above by just the number of target products but instead the number of edges in the network. This difference is due to the need for all substrates of a reaction to provide flow instead of just a single substrate, which may require multiple units of flow to reach a single target product. Combined, constraints (2.7) and (2.8) guarantee that a reaction's products can be available only if all of the reaction's substrates are available, and that if the reaction is required for generating flow the ILP reaction variable for that reaction must be 1.

2.3.6 **Compartmentalizing species and defining transport reactions**

The algorithm so far has treated the community as a single super-organism, whereby metabolites can transfer freely between species. A more realistic scenario, however, assumes that metabolites are compartmentalized within each species and requires species to have the necessary transport reactions to allow environmental metabolite uptake and secretion. Such a compartmentalized problem can in fact be solved by the algorithm as currently defined simply by modifying the sets of metabolites and reactions. Specifically, rather than having a single variable to denote each metabolite (regardless of its compartment), a *set* of variables should be defined to denote the metabolite in each compartment in which it exists (be it a specific species or the shared environment). Given this extended set of metabolite variables, metabolic reactions are now viewed as operating within a species (and accordingly connect species-specific substrates to species-specific products). An additional set of transport reactions (which correspond to each species' uptake and secretion capacities) can then convert species-specific metabolites to environmental metabolites and vice-versa. In this compartmentalized setting, one species can only use metabolites produced by another species if both species have the appropriate transport reactions. For example, for a given metabolite m to transfer from species A to species B, species A must include a transport reaction to convert the A-specific version of m to an environmental version, and similarly species B must include a transport reaction to convert the environmental version of m to a B-specific version.

More formally, we now define $M = \{m_{0,1}, m_{0,2}, \dots, m_{0,n}, m_{1,1}, \dots, m_{q,n}\}$ as the set of metabolites such that $m_{i,k}$, where $i \in \{0,1, \dots, q\}$ and $k \in \{1,2, \dots, n\}$, denotes metabolite k in species i. We interpret the 0^{th} species as the shared environment. We then replace each reaction, r_j, present in species I in the previous formulation with a new reaction:

$$r_{i,j} = \left(\left\{m_{i,j\_substrate\_1}, m_{i,j\_substrate\_2}, \ldots, m_{i,j\_substrate\_d}\right\},\right.$$
$$\left.\left\{m_{i,j\_product\_1}, m_{i,j\_product\_2}, \ldots, m_{i,j\_product\_e}\right\}\right).$$

Each species may also include a set of transport reactions that convert environmental metabolites to species-specific metabolites (reflecting uptake reactions):

$$r_{i,transport\_l} = \left(m_{0,k}, m_{i,k}\right),$$

or species-specific metabolites to environmental metabolites (representing secretion):

$$r_{i,transport\_l} = \left(m_{i,k}, m_{0,k}\right).$$

Together, these new metabolite and reaction definitions relax the assumption of freely transferred metabolites and allow the algorithm to solve problems in a compartmentalized species setting.

2.3.7 **Forcing substrate usage and incorporating species costs**

The above ILP-based formulation can be further extended to force the obtained solution to meet additional requirements or to consider additional factors. Specifically, I have developed and implemented algorithm extensions to handle two biologically-relevant considerations, the first forcing the solution to utilize (or degrade) specified substrates, and the second to weigh species' predefined desirability when constructing a community. For a detailed description of the associated constraints and modifications, see 6.1 Supplementary Text.

2.4 **Results**

2.4.1 **Algorithm Implementation and Availability**

I implemented the algorithm outlined above as a C++ program which takes as input a file describing the various parameters of the design task, including available substrates, target products, and the set of available species with their associated metabolic reactions. The program then generates the associated ILP instance in the Mathematical Programming System (MPS) format (default) or the CPLEX format (depending on the requirements of the ILP solver used). The source code for the algorithm is available under a non-commercial research use only license

at https://github.com/borenstein-lab/CoMiDA. To obtain solutions for the test cases and dataset analysis, I used the COIN-OR Branch and Cut (CBC) solver [128].

2.4.2 Unit Test Validation

I first aimed to verify the algorithm using a set of simple design problems. Specifically, I generated a suite of toy problems as unit test cases for the algorithm. These toy problems test whether the algorithm identifies an optimal solution under different scenarios that cover a variety of edge cases. These test cases focus on simple design tasks, with up to five species and up to seven associated metabolic reactions. For example, some cases examined scenarios in which the minimal species solution requires a longer metabolic path from substrate to product than a non-minimal species solution. Other cases examined scenarios in which a solution does not exist (e.g., because no path exists from substrate to product, regardless of which species are used). I have applied the algorithm to each of these test cases and confirmed that the algorithm correctly produces the ILP formulation and ultimately identifies an optimal solution for each design task (or the absence of one). These toy problems, along with their expected ILP formulations, can be found (with the source code) at https://github.com/borenstein-lab/CoMiDA, providing users with simple examples of the expected input/output format and allowing users to confirm that the algorithm is working properly.

2.4.3 Glycolysis Pathway Validation

The toy problems described above are limited in size and may not be comparable in scale to many real-world scenarios. To examine the algorithm's performance on datasets of a more practical size, I next focused on a well-characterized metabolic pathway, the Embden-Meyerhof glycolysis pathway (KEGG entry *M00001*) [105], defining *glucose* and *pyruvate* as the available substrate and target product respectively (Figure 6.1). For the set of available species, I selected all 284 species identified from the 2013 Human Microbiome Project (HMP) [129] stool sample datasets that contained the entire set of metabolic reactions in the glycolysis pathway as predicted by PICRUSt [130]. Combined, this set of species corresponds to an aggregate metabolic network containing 1803 metabolites and 3120 metabolic reactions. Since each species in this set can catalyze the entire pathway from glucose to pyruvate, the algorithm identifies, as expected, a single species solution (one of the 284 possible choices). To test the algorithm's performance when

minimal solutions required multiple species, I next therefore modified the metabolic network of each species, deleting various reactions and forcing a multi-species solution. Specifically, I first removed all alternate reaction paths between glucose and pyruvate by removing the first reaction in each alternate path that was not also part of the glycolysis pathway (Figure 6.1), filtering out 39 reactions and leaving a total of 3081 metabolic reactions in the aggregate network. I then removed selected reactions in the glycolysis pathway from subsets of the available species such that no single species contained all reactions in the path (e.g., by removing one reaction in the pathway from half of the available species and a different reaction from the other half). Through numerous such modifications, I forced minimal solutions for providing the glycolysis pathway to require multiple species, fully controlling the size of these minimal solutions. I confirmed that the algorithm was able to handle such cases and to provide a correct minimal solution in each such scenario.

2.4.4 **Analysis of Minimal Communities of Gut Microbiome Species**

Naturally occurring microbial communities often comprise an extremely complex and diverse collection of species [1,131]. This diversity can be the product of numerous factors, including a variety of niches species can occupy [132], metabolic specialization of individual species within the community [133,134], intricate multi-species interactions [135], or functional redundancy [136,137]. Yet, when designing synthetic communities, markedly fewer species may be required [87]. To explore this possibility and to characterize potential redundancy in naturally occurring communities, I used CoMiDA to identify minimal communities required to perform various simple metabolic syntheses within the context of a diverse natural community. Specifically, given the promise of gut microbiome-based therapies, I focused on minimal communities that consist of gut dwelling species. To this end, I selected a set of 2051 species (represented as Operational Taxonomic Units; OTUs) detected via 16S sequencing of HMP stool samples. The set of metabolic reactions each OTU could catalyze was determined using PICRUSt [130]. Combined, the aggregate metabolic network of this set of species included 2225 unique metabolites. I then selected 10,000 random pairs of metabolites from this set, one as the available substrate and one as the target product, and used CoMiDA to identify minimal communities that could provide a pathway from substrate to product. I specifically used CoMiDA in three different settings: one with the metabolic network simplified such that each reaction has a

single substrate and a single product (see above), one with the full bipartite graph representation of the metabolic network (allowing each reaction to have multiple substrates and/or multiple products), and one with the full bipartite graph representation but also including a set of common currency metabolites [20] as available substrates.

As shown in Figure 2.4, for most random metabolite pairs, no set of species had the capacity to perform the desired synthesis (potentially owing to various gaps in the aggregate metabolic network and incomplete annotation of the various species). Yet, when a solution existed, it generally required only very few species (≤5 for all metabolite pairs tested). Notably, since the simplified graph representation requires only one substrate of a reaction to be available to generate any of that reaction's products (ignoring, for example, the need for additional co-factors), many more solutions existed when this simple graph representation was used compared with the complete bipartite representation. Making currency metabolites available evidently allowed

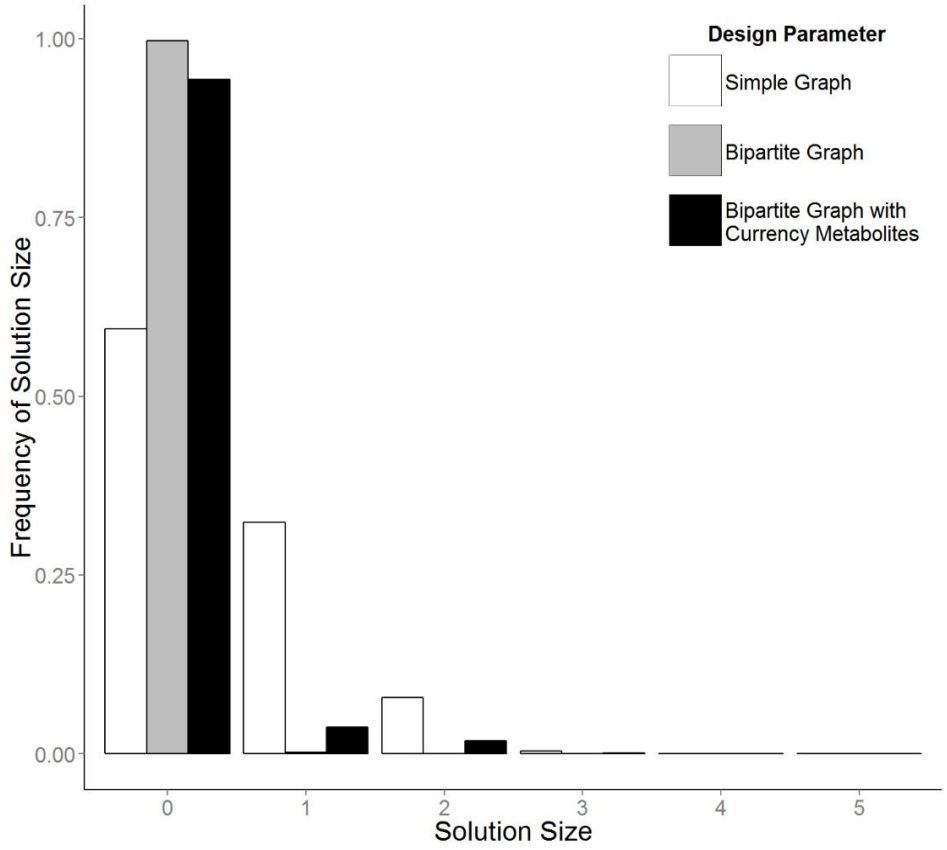

Figure 2.4. Solution sizes identified for 10,000 random substrate/product metabolite pairs, using species from the Human Microbiome Project.

additional reactions to be active and therefore recovered some of the metabolic capacity that could not be realized in the bipartite graph representation. Given the small minimal communities identified and the small number of unique species used in these communities across all pairs (379 OTUs), one might suspect that a small number of metabolic generalist species are responsible for providing the required metabolic capacity in many of these minimal communities.

2.5 Discussion

Recent efforts to manipulate various naturally occurring communities and to impact their activities have shown tremendous promise. For example, efforts to modify the human gut microbiome have demonstrated that properly perturbing this community can treat or ameliorate certain conditions [113]. Expanding this approach to effectively treat a wider variety of diseases, as well as alter the functions of environmentally and industrially-relevant microbial communities, requires methods for rationally designing communities with specific metabolic capacities. Above, I take a first step towards this goal, introducing and validating a novel algorithm which identifies minimal microbial communities that provide specified and desired metabolic capacities.

Clearly, various biological factors are currently not considered by CoMiDA, including, for example, species-level interactions [138], the expected flux through each metabolic reaction, and whether member species, or even the community as a whole, can survive in the target environment [136]. Ignoring such factors may render communities designed by CoMiDA markedly different than naturally occurring communities and synthetic communities constructed based on such designs may consequently fail to survive or to perform specific desired tasks. For example, the discrepancy between the small communities identified in the analysis above and the extreme diversity observed in many naturally occurring communities [1,131] may be likely accounted for, at least partly, by such factors. Yet, CoMiDA provides a starting point for such design efforts and for future method development in this area. Specifically, selecting a community based initially on the presence of desired metabolic capacities provides a simple way to address an important prerequisite for community metabolism; any community designed to consume or produce given metabolites or to have some metabolic activity must obviously also have the metabolic capacity to carry out those functions. This attempt to identify minimal communities may again not necessarily be aligned with biological assembly rules, but offers simple candidate communities for further design refinement. Moreover, by formulating CoMiDA as an ILP

algorithm, I provide an easy way to introduce additional design considerations. As our understanding of the various constraints affecting community assembly improves, such considerations can be added to this framework by devising equations and inequalities that encode these constraints.

Of the various considerations that could be implemented to further refine any design approach, two stand out as logical next steps. First, the expected stability of designed communities could be improved by examining the likelihood that a combination of species will coexist in a community. Specifically, information on species co-occurrence in natural communities can be used to estimate the tendency of various species pairs to co-exist or the exclude one another from a shared environment [126,139]. Such information would allow an algorithm to prioritize communities that minimize the risk of losing member species due to antagonistic species interactions, ultimately stabilizing community structure. Second, considering the *predicted* activity of candidate communities, rather than just the presence of specific metabolic capacities, could increase the likelihood that designed communities would perform the desired task. Several frameworks for predicting the metabolic activity of microbial communities have recently been introduced [118,140–142], potentially allowing future design algorithms to consider predicted rates of metabolite consumption and production and predicted changes in species abundances over time. CoMiDA could be used, for example, as an initial filtering step, providing a set of candidate minimal communities that have the capacity for some desired metabolism, followed by a metabolic model-based prediction of the metabolic activity of each candidate community to further refine the design process. Moreover, such metabolic modeling could allow the design process to account for important factors that CoMiDA may not be able consider. For instance, CoMiDA does not explicitly prevent community members from degrading one or more of the specified target products. Such inadvertent target metabolite degradation may depend on the set of microbes present, other available substrates, and various environmental conditions, and could therefore be predicted and potentially avoided using metabolic modeling-based design.

The ability to computationally design microbial communities will be a useful tool for many purposes. For example, designed synthetic communities could be ultimately used in place of FMTs, removing the need for screening donor samples while also optimizing treatments to target specific conditions. Communities could also be created for industrial resource and pharmaceutical production, potentially obviating the need for extensive microbial genetic engineering and

providing novel mechanisms for production control in the form of inter-species signaling [40]. Clearly, such applications are not yet feasible and the development of a comprehensive, general-purpose design framework may still be out of our reach for years to come. I hope, however, that CoMiDA will encourage future developments of such design methodologies and will lay the foundation for future efforts in microbial community design.

3. Taxa-function robustness in microbial communities

This chapter is based on the following manuscript published in *Microbiome*:

Eng A, Borenstein E. Taxa-function robustness in microbial communities. *Microbiome* 2018; 6:45.

3.1 Summary

3.1.1 Background

The species composition of a microbial community is rarely fixed and often experiences fluctuations of varying degrees and at varying frequencies. These perturbations to a community's taxonomic profile naturally also alter the community's functional profile – the aggregate set of genes encoded by community members – ultimately altering the community's overall functional capacities. The magnitude of such functional changes and the specific shift that will occur in each function, however, are strongly dependent on how genes are distributed across community members' genomes. This gene distribution, in turn, is determined by the taxonomic composition of the community, and would markedly differ, for example, between communities composed of species with similar genomic content vs. communities composed of species whose genomes encode relatively distinct gene sets. Combined, these observations suggest that community functional robustness to taxonomic perturbations could vary widely across communities with different compositions, yet, to date, a systematic study of the inherent link between community composition and robustness is lacking.

3.1.2 Results

In this study, I examined how a community's taxonomic composition influences the robustness of that community's functional profile to taxonomic perturbation (here termed *taxa-function robustness*), across a wide array of environments. Using a novel simulation-based computational model to quantify this taxa-function robustness in host-associated and non-host-associated communities, I find notable differences in robustness between communities inhabiting different body sites, including significantly higher robustness in gut communities compared to

vaginal communities that cannot be attributed solely to differences in species richness. I additionally find between-site differences in the robustness of specific functions, some of which are potentially related to site-specific environmental conditions. These taxa-function robustness differences are most strongly associated with differences in overall functional redundancy, though other aspects of gene distribution also influence taxa-function robustness in certain body environments, and are sufficient to cluster communities by environment. Further analysis revealed a correspondence between my robustness estimates and taxonomic and functional shifts observed across human-associated communities.

3.1.3 **Conclusions**

This analysis approach revealed intriguing taxa-function robustness variation across environments and identified features of community and gene distribution that impact robustness. This approach could be further applied for estimating taxa-function robustness in novel communities, and for informing the design of synthetic communities with specific robustness requirements.

3.2 **Background**

The examination and characterization of microbial communities have become increasingly important as their impacts on human health, industrial processes, and the environment have been recognized. These communities have been studied both in terms of their taxonomic and functional compositions, elucidating important community features and revealing intriguing disease- and environment-associated variation. A community's taxonomic composition is often determined via targeted 16S rRNA sequencing [143], a technique that uses hyper-variable regions of the 16S rRNA gene to identify the microbes present in a given community and estimate their relative abundances. Such taxonomic information can provide insight into inter-microbial or host-microbe interactions and facilitate the detection of shifts in community ecology that may be associated with host disease [144–148]. The functional composition of a community, in turn, can be estimated via whole metagenome shotgun sequencing followed by gene annotation. Using such data, gene-level analyses have provided insight into the functional capacities of various microbial communities [149,150] and how those capacities change over time or vary with altered environmental conditions [151,152].

Indeed, these two facets of microbiome composition, namely its taxonomic structure and its functional capacities, offer different but complementary views into microbial communities. These two aspects of microbiome organization, however, are clearly not independent as the composition of genes in the metagenome is a direct derivative of the genes encoded by the community members' genomes and the relative abundance of each member in the community. Moreover, this link can be represented as a simple set of linear equations wherein the abundance of each gene in the metagenome is the sum of that gene's copy number in each community member's genome weighted by the relative abundance of each community member [153] (Figure 3.1A). This inherent link between a community's taxonomic composition and its functional profile, here referred to as the *taxa-function relationship* [154], has many practical applications in the analysis of microbial communities. For example, this relationship is explicitly utilized by tools such as PICRUSt [130] and Tax4Fun [155] for predicting overall community gene content based on the community taxonomic profile and available reference genomes. Other studies have similarly used the taxa-function relationship to identify taxonomic drivers of disease-associated functional shifts [156] or to estimate differences in community metabolic capacities [157]. Such methods for integrated analysis of multi-omic microbiome data offer unique, mechanistically-driven insights into how taxonomic composition affects community function through features such as gene abundances and metabolism.

Conceptually, the taxa-function relationship can be viewed as a structure-to-function landscape, whose topology is determined by the distribution of genes across community genomes. As such, it is similar to the fitness landscape concept used in evolutionary biology [158–161], but instead of describing how changes in genotype map to changes in phenotype, it describes how changes in a microbial community's composition map to changes in the community's functional capacities. Characterizing the topology of this taxa-function landscape is similarly crucial for understanding how constraints on community ecology restrict community function, and should be considered when designing the targeted manipulation of community composition. One important manifestation of the taxa-function landscape is the degree to which a shift in a community's taxonomic composition will impact its functional capacities (a property that I refer to here as *taxa-function robustness*). Specifically, depending on the local topology of the taxa-function landscape around a given community, changes in the abundance of its members could result in a minor or major alteration to the community's functional profile [136,162–164] (Figure 3.1B; analogous to

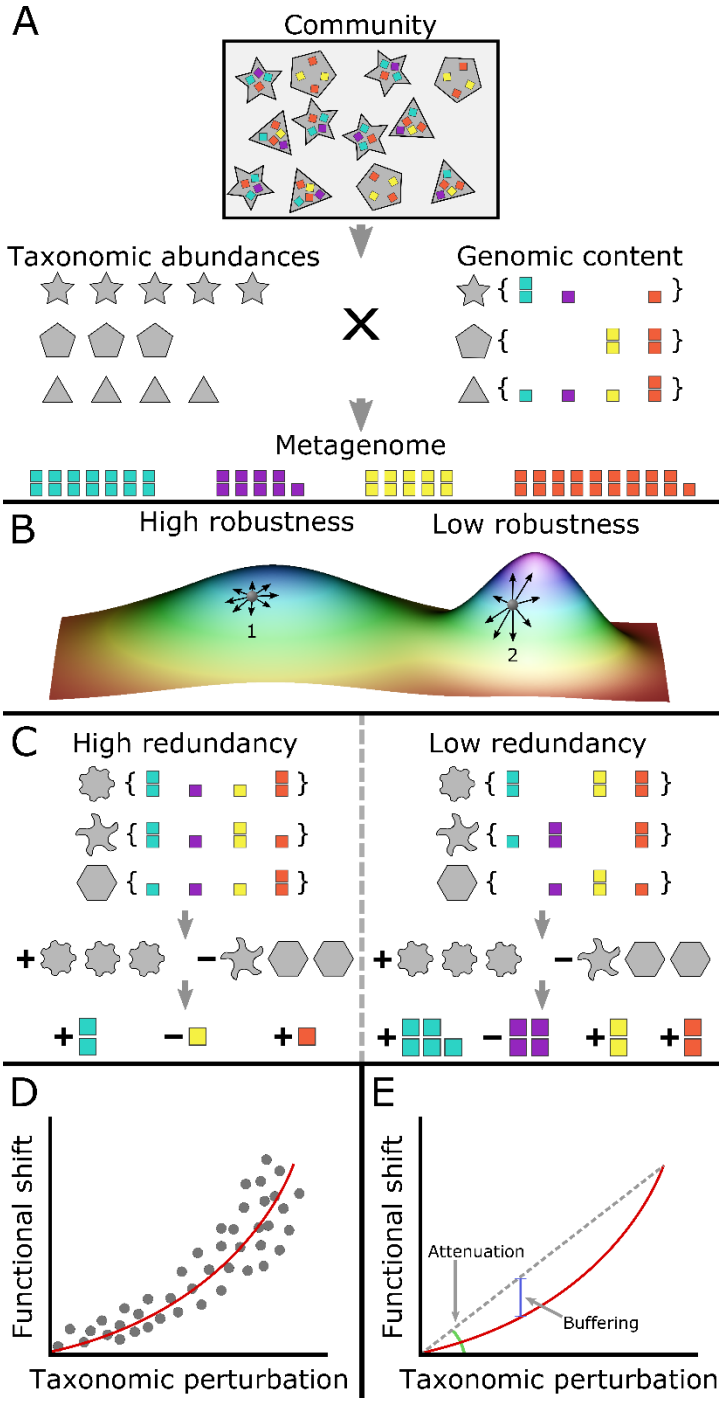

Figure 3.1. An illustration of the taxa-function relationship and response curves. (A) A community's functional profile is a linear combination of each taxon's functional profile (the copy number of each gene in each taxon's genome) weighted by the taxon's abundance. Here a taxon can represent any subpopulation of the community with shared genomic content (such as a strain). **(B)** The taxa-function relationship can be modeled as a high-dimensional landscape, linking each community composition to the corresponding functional profile. Here, an extremely simplified two-dimensional abstraction of this model illustrates the landscape's impact on taxa-function robustness. Each coordinate corresponds to a specific taxonomic composition, with points close to one another corresponding to communities with similar taxonomic compositions. The height represents the functional profile of the community (for example, the abundance of some function). The local topology of this landscape around a specific taxonomic composition (e.g., point 1) determines whether minor changes in that composition (black arrows) will induce small (point 1) or large (point 2) functional shifts. **(C)** Depending on the distribution of genes across species' genomes, changes to a community's taxonomic composition can produce functional shifts of varying magnitudes. For example, if the distribution of genes differs markedly between two communities (e.g., high vs. low redundancy), similar taxonomic composition perturbations may produce drastically different functional shifts. **(D)** To model the relationship between taxonomic perturbations and functional shifts in a given community, a taxa-function response curve is obtained by fitting a power function to an array of measured functional shifts associated with different taxonomic perturbations of varying magnitudes. **(E)** The taxa-function response curve can be decomposed into two factors: attenuation, which describes how quickly functional shifts increase in magnitude as taxonomic perturbations increase, and buffering, which indicates how well functional shifts are suppressed at smaller taxonomic perturbations.

the impact of the fitness landscape on genetic robustness [165,166]). As noted above, the topology of the landscape around a given community, and consequently its taxa-function robustness, depends solely on the manner in which genes are distributed among the genomes of that community's members (Figure 3.1C). For example, if a specific gene family (or pathway) is encoded by a single species in the community, any perturbation to that species' abundance will

directly translate to changes in the abundance of this gene family in the metagenome and ultimately in the community's functional profile. If, however, this gene family is encoded by multiple species in the community, its abundance is less likely to substantially change in the face of taxonomic perturbations as a decrease in the abundance of one species that encodes this gene family could be compensated for by an increase in the abundance of another [136,162–164]. Moreover, if multiple genes tend to co-occur across the various genomes in a community [167], then those genes (and the functions associated with them) will shift in a similar manner as the taxonomic composition is perturbed. Combined, these observations suggest that the taxonomic composition of a community (which in turn also determines the distribution of genes across community members) directly impacts the taxa-function *robustness* of that community, and may therefore vary substantially from community to community.

Importantly, this definition of taxa-function robustness aims to capture an important aspect of the taxa-function relationship and is independent of the specific variation the community actually experiences or its natural dynamics (just as a fitness landscape is determined by the genotype-to-phenotype relationship regardless of observed evolutionary trajectories). As such, it may not necessarily correspond to conventional notions of ecological robustness, such as how well a community can tolerate environmental disturbances before undergoing significant changes in composition or function. Neither does it exactly match definitions of functional resilience (how quickly community function returns to normal after environmental perturbation) or functional stability (how well normal community function is maintained over time). Yet, characterizing such underlying taxa-function robustness and its determinants is essential for gaining a profound understanding of community dynamics and function. For example, this feature of a community's local taxa-function landscape could indicate how susceptible a community's functional capacities are to stochastic fluctuations in the community's composition, and can be used as a null model when studying community dynamics in response to environmental change. More generally, while microbiome perturbations, be they minor stochastic fluctuations or major shifts in response to environmental modulation, are often characterized as ecological changes, in many cases the most important consequences of these perturbations are shifts in overall community function. In the context of disease-associated dynamics, taxa-function robustness will also determine whether the functional capacities of a community could be maintained in the face of ecological dysbiosis. For example, the function of the gut microbiome is robust enough to maintain normal function despite

day-to-day fluctuations in taxonomic composition [168], but may become disrupted following a more drastic perturbation as in the case of *Clostridium difficile* infection [74]. Determining a community's functional robustness can further help estimate the functional impact of a planned targeted perturbation (e.g., via a particular probiotic), or evaluate candidate synthetic communities during the design process to gauge how susceptible they are to disruption of function.

Here, I set out to systematically characterize and study the taxa-function robustness of communities from diverse environments. This requires a comprehensive, systematic, and unbiased mapping of the local topology of the underlying taxa-function landscape around each community composition. Unfortunately, currently available experimental data does not adequately or comprehensively survey functional shifts associated with small changes to a particular taxonomic composition. I address this challenge by using a simulation-based approach, uniformly and systematically simulating a range of possible perturbations (including small perturbations) to each community's taxonomic composition. This approach, combined with the prediction of community functional profiles associated with each such perturbation, allows me to generate a large set of perturbed compositions relative to each original community and to sample the taxa-function landscape around these communities. Given this simulation-based approach, I first define two factors that characterize a community's taxa-function robustness. I then present an analysis of how taxa-function robustness varies within and between body sites in human-associated communities as well as across several non-host-associated communities. I extend this analysis to the robustness of individual functions and pathways, identifying universally robust functions and noting that environment-specific pressures may influence robustness variation in specific functions. Next, I investigate how the manner in which genes are distributed across microbial genomes is associated with a community's taxa-function robustness and use this information to predict robustness directly from taxonomic composition. Finally, I confirm that these robustness estimates are in agreement with observed taxonomic and functional shifts measured from experimental data.

3.3 Methods

3.3.1 Samples and data processing

Samples were obtained from the Human Microbiome Project (HMP) [1] and the Earth Microbiome Project (EMP) [169]. This included four sites from the human microbiome collected

as part of the HMP: 128 gut communities, 1141 oral communities (pooled from 9 subsites), 285 skin communities (pooled from 3 subsites), and 209 vaginal communities (pooled from 3 subsites). Additionally, these samples were supplemented with 132 aquatic communities (pooled from 3 subsites) and 199 soil communities (from 4 subsites) collected in association with the Earth Microbiome Project [169]. 16S rRNA-based Operational Taxonomic Unit (OTU) tables and metadata files were downloaded from the QIITA website (http://qiita.microbio.me), which provides OTU tables generated with the QIIME workflow [170] using Greengenes OTUs [171]. OTU tables were filtered to remove read counts mapped to plant chloroplasts. To improve quality and comparability between taxonomic profiles, taxonomic profiles with fewer than 10 OTUs or fewer than 5000 reads were removed and the remaining taxonomic profiles were rarefied to 5000 reads. When analyzing body subsites, left and right bilaterally symmetric body subsites (antecubital fossa or retroauricular crease) were pooled together. For each subsite, communities were selected from different hosts.

3.3.2 **Functional profile prediction**

The Kyoto Encyclopedia of Genes and Genomes (KEGG) [105] was used to define orthologous gene functions in terms of KEGG Orthology (KO) groups. KO abundances were predicted using PICRUSt [130] tables to first normalize each OTU's relative abundance by its 16S rRNA copy number and then infer KO abundance from the genomic content of each OTU. KO abundances were then normalized using inter-sample MUSiCC [172]. Pathway-level functional summaries, as defined by the KEGG BRITE hierarchy [173], were obtained by evenly distributing each KO's average copy-number across all pathways that contain that KO. Both KO and pathway tables were filtered to remove non-bacterial orthologs.

3.3.3 **Taxonomic composition perturbation**

Community taxonomic composition perturbations were designed to simulate stochastic OTU relative abundance fluctuations (assuming no migration or introduction of OTUs absent in the original community). Perturbation size for each OTU was proportional to the abundance of that OTU. Formally, for a given community with N OTUs with non-zero relative abundances $a_i \; \forall \; i \in \{1, 2, \ldots, N\}$, perturbation multipliers $m_i \; \forall \; i \in \{1, 2, \ldots, N\}$ were sampled from a uniform distribution over the interval $(0, M]$, M being the maximum perturbation magnitude, and

perturbation directions d_i were chosen such that $d_i \in \{-1, 1\}\ \forall\ i \in \{1, 2, \ldots, N\}$ with an equal chance of either direction. Given these values, the perturbed taxonomic composition OTU abundances $p_i\ \forall\ i \in \{1, 2, \ldots, N\}$ were calculated as

$$p_i = a_i \times m_i^{d_i}$$

and then renormalized such that $\sum_i^N p_i = 1$. Using this method, each community was perturbed at 45 maximum perturbation magnitudes evenly spaced between 1.2 and 10 inclusive with 100 perturbations at each maximum perturbation magnitude.

3.3.4 **Perturbation magnitude calculations**

Changes in taxonomic composition between an original community composition and a perturbed composition were measured using the weighted UniFrac metric [174], a common, phylogeny-aware metric for estimating dissimilarity between microbial community compositions. The shift in a community's functional profile was defined as the cosine dissimilarity between the original and perturbed functional profiles, as done in Das et al. [175]. Specifically, given an original community and perturbed community with N unique pathways with average copy number $a_j\ \forall\ j \in \{1, 2, \ldots, N\}$ and $b_j\ \forall\ j \in \{1, 2, \ldots, N\}$, respectively, the cosine dissimilarity between the two functional profiles is

$$1 - \frac{\sum_j^N a_j b_j}{\sqrt{\sum_j^N a_j^2}\sqrt{\sum_j^N b_j^2}}.$$

3.3.5 **Robustness metric definition and fitting**

For this study, the taxa-function robustness of a microbial community is defined as the average shift in the functional profile given a perturbation to the community's taxonomic composition. To allow quantitative robustness comparisons between communities, I first evaluated several models for their ability to fit the relationship between taxonomic and functional differences (7.1 Supplementary Text; Figure 7.1; Table 7.1). Among the models evaluated, I found that the following model best captured the relationship between taxonomic perturbation magnitude and functional profile shift:

$$f = \frac{1}{e^{a}} t^{b}$$

where t denotes the magnitude of the taxonomic perturbation, f denotes the expected shift in functional profile, and a and b are community-specific coefficients. I term this function the *taxa-function response curve*. I further term the coefficient a '*attenuation'* since it describes the expected rate at which increases in the taxonomic perturbation magnitude are expected to increase functional profile shifts. I similarly term the exponent b '*buffering'* since it indicates how large a perturbation must be before a functional profile shift becomes noticeable and approaches the expected shift magnitude defined by attenuation. Function-specific robustness was defined in a similar manner to taxa-function robustness except that instead of the cosine dissimilarity between the original and perturbed functional profiles, the functional shift of a single function was measured as the relative change in the abundance of that function.

Attenuation and buffering were fit by first transforming weighted UniFrac and cosine dissimilarities to the natural log scale to reduce heteroscedasticity observed in the simulated perturbation data (variance of cosine dissimilarity increased as weighted UniFrac distance increased). A uniform sampling of simulated perturbations across weighted UniFrac dissimilarities was obtained by subsampling perturbations across 50 non-overlapping windows evenly spaced on the natural log scale between minimum and maximum distances. For each community, in each window, perturbations were subsampled to 50 when ≥50 perturbations were present and all perturbations were kept when <50 were present. The transformation to the natural log scale also transformed the proposed taxa-function robustness curve function to the following form:

$$\ln(f) = -a + b\ln(t)$$

which was then fit using the linear least-squares best fit to calculate attenuation and buffering.

Due to the asymmetric distributions of attenuation and buffering, the pseudomedian was used rather than the median, and pseudomedian attenuation and buffering estimates were calculated using the Hodges-Lehmann statistic [176]. The 95% confidence interval for a pseudomedian estimate was calculated as the range of values for which the Wilcoxon statistic (given the observed attenuation or buffering values) was between the 0.025 and 0.975 quantiles of the standard normal distribution.

3.3.6 **Function-specific most robust environment determination**

A function was determined to be most robust in a particular environment (or environments) by comparing the median attenuation of that function across all environments using Mood's Median test. Specifically, a function was defined as most robust in a set of environments if, for each environment in that set, the median attenuation of the function in that environment was both significantly higher than its median attenuation in all of the environments not in the set and not significantly different than its median attenuation in any of the other environments in the set.

3.3.7 **OTU subsampling procedure**

To determine the relationship between diversity and robustness when comparing vaginal and gut communities, communities were subsampled from each environment to obtain similar species richness and the analysis of between-environment robustness differences was repeated. Specifically, each community was randomly subsampled to 10 OTUs such that the probability of an OTU remaining in the subsampled community was proportional to its relative abundance in the original community. This subsampling procedure aimed to achieve a similar distribution of relative abundance among community members between the original and subsampled communities. Once subsampled, OTU abundances were renormalized.

3.3.8 **Gene distribution feature (GDF) definitions**

Five main gene distribution features (GDFs) were used in correlation and PCA analysis, consisting of average functional redundancy, average functional similarity, average genome size, genome size variability, and unique function abundance. Each GDF captures a different, though potentially related, aspect of the distribution of genes across the genomes of species in a community.

The functional redundancy of a given function was defined here as the evenness (Shannon's diversity index) of the abundances of each species that encodes that function weighted by the copy number of the function in each species' genome respectively:

$$-\sum_{i=1}^{N}[(s_i c_i)\ln(s_i c_i)]$$

where N species encode a function, s_i is the abundance of species i that encodes that function, and c_i is the copy number of that function in species i's genome. This definition aims to capture how evenly species contribute to a function's abundance, such that a function should be considered less redundant if one species contributes the majority of that function's abundance while it will be more redundant if many species all contribute similarly to its abundance. The average functional redundancy of a community was then defined as the average redundancy of all functions present in a community, weighted by the relative abundance of each function in the community's functional profile.

The functional similarity between two microbes aimed to capture how well two different species could compensate functionally for one another, and thus was defined as the cosine similarity between the functional profiles of two species:

$$\frac{\sum_{i=1}^{M} a_i b_i}{\sqrt{\sum_{i=1}^{M} a_i^2}\sqrt{\sum_{i=1}^{M} b_i^2}}$$

where M is the number of functions encoded by encoded by at least one of the species, a_i is the copy number of function i in species a and b_i is the copy number of function i in species b. The average functional similarity within a community was then defined as the unweighted average of all pairwise functional similarities between species present in the community.

Genome size for a given species was defined as the total abundance of functions encoded by the species (i.e., the sum of the copy number of each function in that species genome). Average genome size was calculated as the unweighted average of each species' genome size and genome size variability was calculated as the coefficient of variation of species genome size.

Unique function abundance was defined as the total abundance of functions in a community's functional profile that are each encoded by a single species (though each unique function need not be encoded by the same species).

3.3.9 **Metagenome-based data and functional shifts**

Shotgun metagenome-based KO profiles for 94 communities were obtained from HMP. KO profiles were corrected using inter-sample MUSiCC [172] and summarized to the pathway-level using the same protocol used for predicted profiles. To obtain shotgun metagenome-based functional profile differences, 47 random community pairs were assigned from the 94 communities

with both 16S rRNA and shotgun metagenome profiles. Pairs were restricted to contain two communities from the same subsite. Between-community taxonomic and functional dissimilarities were calculated (as described above) between the communities in each pair.

3.3.10 **Mixed community perturbations**

Community mixing was performed using the same community pairs as for metagenome-based functional shift measurements. For each pair, one community was designated the original community and the other the mixing community. For a taxon i with abundance a_i in the original community and abundance b_i in the mixing community, taxon i's abundance in the mixed community perturbation (relative to the original community) with mixing fraction m was $[a_i(1-m)] + [b_i m]$. Similarly, for a function j with average copy number c_j in the original community and average copy number d_j in the mixing community, function j's abundance in the mixed community perturbation (relative to the original community) with mixing fraction m was $\left[c_j(1-m)\right] + \left[d_j m\right]$. Mixed community perturbations were generated at specific weighted UniFrac dissimilarities by using a binary search to identify the mixing fraction that achieved a mixed community perturbation with the desired weighted UniFrac dissimilarity from the original community within a tolerance of 10^{-9}. To fit robustness curves for the original communities, mixed community perturbations were generated at weighted UniFrac distances of 0.01 to 0.1 in intervals of 0.01 for each community pair (except for one community pair, which could not achieve a weighted UniFrac distance of 0.1 through community mixing). Robustness curves were then fit to the resulting perturbation taxonomic and functional dissimilarities as described above.

3.3.11 **Longitudinal functional shift calculations**

Longitudinal data with 16S rRNA and shotgun metagenomic profiles collected at two time points were obtained for 8 HMP communities. Taxonomic and functional dissimilarities between the community compositions at each time point were calculated as above. Expected functional shifts based on robustness estimates were calculated using the robustness curve formula defined above, the estimated attenuation and buffering values for the community, and the weighted UniFrac distance between the community compositions at the two time points.

3.4 Results

3.4.1 Characterizing and defining taxa-function robustness

Consider the taxa-function mapping discussed above, linking a community's taxonomic and functional compositions. To rigorously characterize how this mapping impacts taxa-function robustness, I define the *taxa-function response curve*, describing the average shift in the functional profile of a community as a function of the taxonomic perturbation magnitude and the stability of a community's functional profile when faced with taxonomic perturbations (Figure 3.1C). Response curves are commonly used in biology to describe how the changes in an organism vary as the magnitude of a particular stimulus is modulated (e.g., drug dosage-dependent physiological effects) and allow quantitative comparisons between the response curves of different individuals [177–180]. Interpreting a community's taxa-function robustness via these response curves could offer insights such as the potential impact of antibiotics on a community's functional capacities or the expected stability of candidate synthetic communities. For example, a community's taxa-function response curve could be used to identify an antibiotics dosage threshold above which there would be significant disruption to community function. The specific form of the taxa-function response curve was chosen by comparing the fit of various functions to the relationship between taxonomic perturbation and functional shift magnitudes across all communities examined (7.1 Supplementary Text; Figure 7.1).

To provide a direct, quantitative, and interpretable comparison of taxa-function robustness differences between communities, I will further specifically focus on two robustness factors that can be derived from the response curve (Figure 3.1D): The first factor is *attenuation*, or how rapidly functional shift increases as perturbation magnitude increases. Attenuation describes the slope of the response curve and thus models the intuitive expectation that larger perturbations should generate larger functional shifts. Technically, attenuation is defined as inversely proportional to the response curve slope, implying that increased attenuation leads to smaller functional shifts and thus higher robustness. The second factor is *buffering*, or how well functional shifts are suppressed at smaller perturbation magnitudes. Buffering determines how large a taxonomic perturbation must be before noticeable functional shifts occur. Higher buffering thus indicates that relatively large perturbations are required before a substantial functional shift could be observed. This factor is especially important when considering community robustness in the

absence of major external changes, as buffering will determine whether small fluctuations in composition due to stochastic variation, neutral community dynamics, or minor environmental variation will significantly impact the community's function. Indeed, many biological systems can completely buffer small perturbations, but may be more susceptible when exposed to larger perturbations [181,182].

3.4.2 **Taxa-function robustness varies within and between environments**

To comprehensively characterize taxa-function robustness across the human microbiome and certain non-host-associated ecosystems, I obtained taxonomic compositions from previously assayed communities representing several distinct environments, including human-associated gut, oral, skin, and vaginal communities from the HMP as well as soil and water communities from the EMPT. For each community, I generated 4,500 perturbed taxonomic compositions at varying magnitudes (using the weighted UniFrac distance between the original and perturbed community to measure the magnitude of perturbation). Compositional perturbations were simulated by randomly modifying the abundances of individual taxa in the original community such that the expected magnitude of change in each taxon's abundance was proportional to its original abundance. I further filtered the obtained perturbations to uniformly sample perturbation at a range of taxonomic distances, resulting in an average of 933 ± 186 perturbations per community and a total of 1,954,447 perturbations. To determine the functional profiles of both the original and perturbed compositions, I used PICRUSt [130], a computational framework for inferring the functional profile of a given community based on its taxonomic composition as described above. Using these inferred functional profiles, I measured the functional shift associated with each taxonomic perturbation and obtained a taxa-function response curve for each community and calculated the associated attenuation and buffering values based on these response curves. With this approach, I was able to compare response curves and robustness factors between communities and examine how taxa-function robustness varied within and between environments.

To gain an intuition of how perturbation magnitudes affect the degree of functional shift, I first examined the taxa-function response curve of a single human vaginal community. As expected, the degree of the functional shift generally increased with the magnitude of the taxonomic perturbation (Figure 3.2A). I observed some variation in the extent of functional shifts associated with taxonomic perturbations of a similar magnitude despite using a phylogeny-aware

metric for perturbation magnitude. Interestingly, comparing this response curve to the response curve for a community from a different environment – the gut – revealed marked differences, with the gut community displaying noticeably smaller functional shifts at similar taxonomic perturbation magnitudes (Figure 3.2B). These differences are also apparent in the corresponding robustness factors, with the vaginal community having lower attenuation (1.57 compared to 3.487 in the gut) and comparable buffering (2.06 compared to 2.03). Moreover, the vaginal community also displayed more drastic deviations from its taxa-robustness response curve.

To examine whether the differences in robustness between vaginal and gut communities extended beyond these two specific examples, I compared the taxa-function response curves for all communities from these two body sites. This confirmed that vaginal communities indeed exhibited larger functional shifts on average compared to gut communities for similar taxonomic perturbation magnitudes (Figure 3.2C-D). The gut communities also appeared to have more similar response curves across communities, whereas vaginal community response curves were more diverse. Examining the robustness factors of these communities further revealed a clear difference between these two body sites, with gut communities having significantly higher attenuation compared to vaginal communities (Figure 3.2E; $p < 10^{-23}$; Wilcoxon rank-sum test). Interestingly, however, I found only slightly higher buffering values in gut communities (Figure 3.2F; $p < 0.01$; Wilcoxon rank-sum test). I further examined whether this marked difference in robustness between the vaginal and gut microbiomes can be attributed solely to the substantially lower diversity of vaginal microbiomes. To this end, I subsampled all communities from the vagina and gut to obtain communities with comparable diversity. I found that, in these subsampled communities, attenuation was still significantly higher in the gut compare to the vagina, which suggested that the difference in robustness could be attributed, at least partly, to some environment-specific features that go beyond community diversity (Figure 7.2; $p < 10^{-6}$; Wilcoxon rank-sum test).

To extend this analysis beyond vaginal and gut communities, I next examined the robustness factors of every community from all environments I analyzed. For this analysis, I also separated HMP communities by subsite to determine how between-subsite robustness differences compared to differences between more distantly related environments. My analysis revealed substantial variation in attenuation between environments, potentially suggesting different

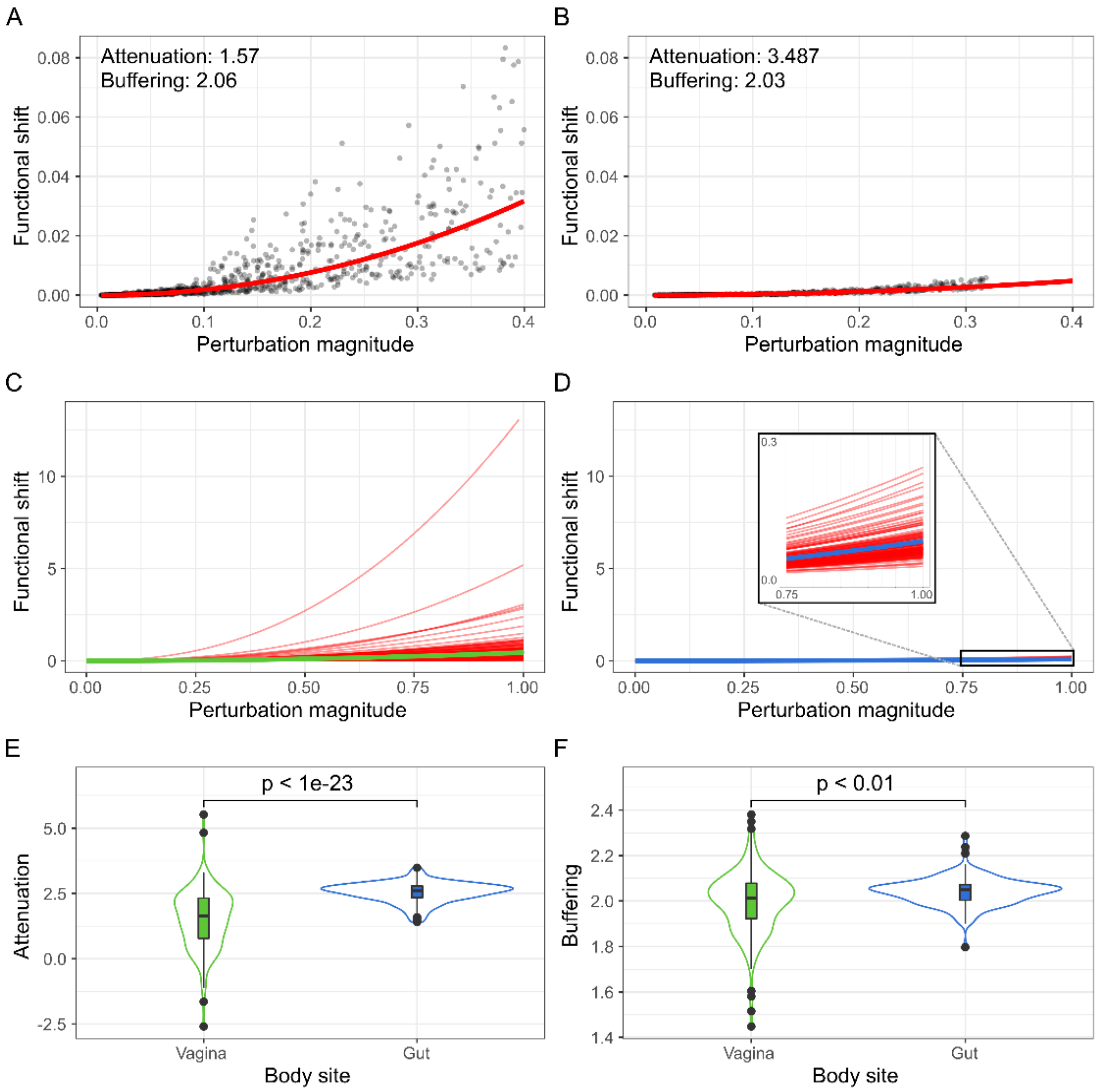

Figure 3.2. Examples of taxonomic perturbations, their corresponding functional shifts, and the associated response curves. The taxonomic perturbation and functional shift magnitudes generated for a single vaginal **(A)** and a single gut **(B)** community. Each point represents a single perturbation. The red lines indicate the taxa-function response curve fit to these points. The response curves for all vaginal **(C)** and gut **(D)** communities are overlaid to compare general body site trends. Green and blue lines represent the mean response curve for all vaginal and gut communities respectively. The robustness factor distributions associated with these response curves are shown as violin plots with inlaid boxplots for attenuation **(E)** and buffering **(F)**. The width of the violin plot indicates the density of robustness factor values, the middle of the box displays the median robustness factor value, the upper and lower edges of the box represent the 75th and 25th percentiles respectively, and the whiskers extend to 1.5 times the inner quartile range (range between the 75th and 25th percentiles) past either end of the box. Outliers are shown as individual black circles.

pressures to maintain robust functional profiles under different environmental conditions (Figure 3.3). Notably, communities from all 3 vaginal subsites appear to the have the lowest attenuation,

with communities from other body sites exhibiting higher (and more comparable) attenuation. In contrast, communities from two of the three skin subsites appear to be among the most robust body site communities. Soil communities tend to have the highest attenuation, whereas marine communities have intermediate attenuation values. Interestingly, subsites from the same environment tend to cluster by attenuation and buffering, suggesting that spatially-distinct subsites (such as different locations in the mouth) still exhibit similar taxa-function robustness factors, potentially reflecting shared environment-specific conditions. Buffering similarly exhibited some variation between environments but not as extreme as the variation seen in attenuation values; in the analyses below, I therefore focus mainly on attenuation to examine differences in robustness.

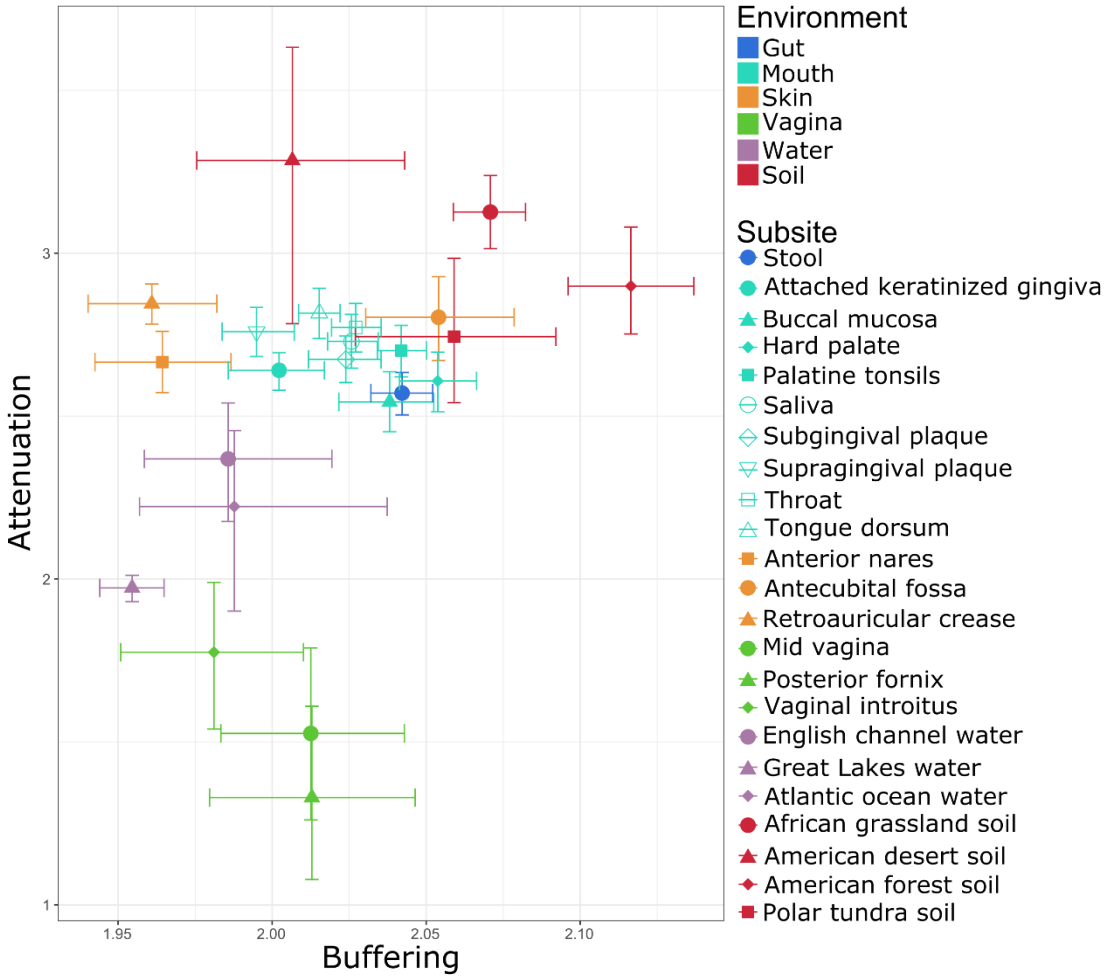

Figure 3.3. Comparison of attenuation and buffering between environments and subsites. Points represent the median attenuation and buffering values of samples from the indicated subsite, with error bars indicating the 95% confidence interval of the median in both dimensions. Colors are shared by subsites from the same environment, while shape indicates the specific subsite within the environment.

3.4.3 **Function-specific robustness reflects environmental conditions**

Given the variation in overall taxa-function robustness observed above, I next set out to examine whether robustness also varied across different functions and whether such *function-specific robustness* is consistent across environments. To this end, I calculated *function-specific response curves* that described the average magnitude of the relative shift in a particular function's abundance due to a taxonomic perturbation. The resulting function-specific response curves can be analyzed as described above to obtain function-specific attenuation measurements.

Examining function-specific attenuation at the superpathway level, I found marked variation in the robustness of different functions. Perhaps not surprisingly, superpathways associated with universal housekeeping activities, such as translation, nucleotide metabolism, and cell growth, were among the most robust functions, likely reflecting high redundancy in these functions across genomes in all communities and environments (Figure 3.4A). In contrast, functions associated with a more specialized lifestyle, such as cell motility, transport, secondary metabolite biosynthesis, and glycan metabolism were generally less robust.

To further characterize differences in function-specific robustness across environments, I next compared the attenuation of each function between environments, this time analyzing functions at the pathway level. This analysis revealed similarly intriguing between-environment differences. Notably, the difference observed in a function's robustness between environments often does not seem to be associated with the difference in function abundance between environments (Figure 3.4B, Figure 7.3). This finding suggests that the abundance and robustness of a given function may be driven by different pressures. Put differently, some functions may be beneficial at high capacity but can tolerate variation in their abundance, hence exhibiting high abundance but low robustness. In contrast, other functions may be advantageous at a more stable capacity, even though they are required at a relatively low capacity, and will hence exhibit high robustness but low abundance. For example, while cysteine and methionine metabolism occurs at low abundance in gut communities (1.5% median relative abundance), it is one of the most robust functions specifically in the gut (2.43 median attenuation; 7^{th} most robust gut function; Figure 3.4B). Indeed, cysteine and methionine deficiencies are associated with malnutrition and can be influenced in part by the gut microbiome [183], and accordingly, maintaining a stable capacity of

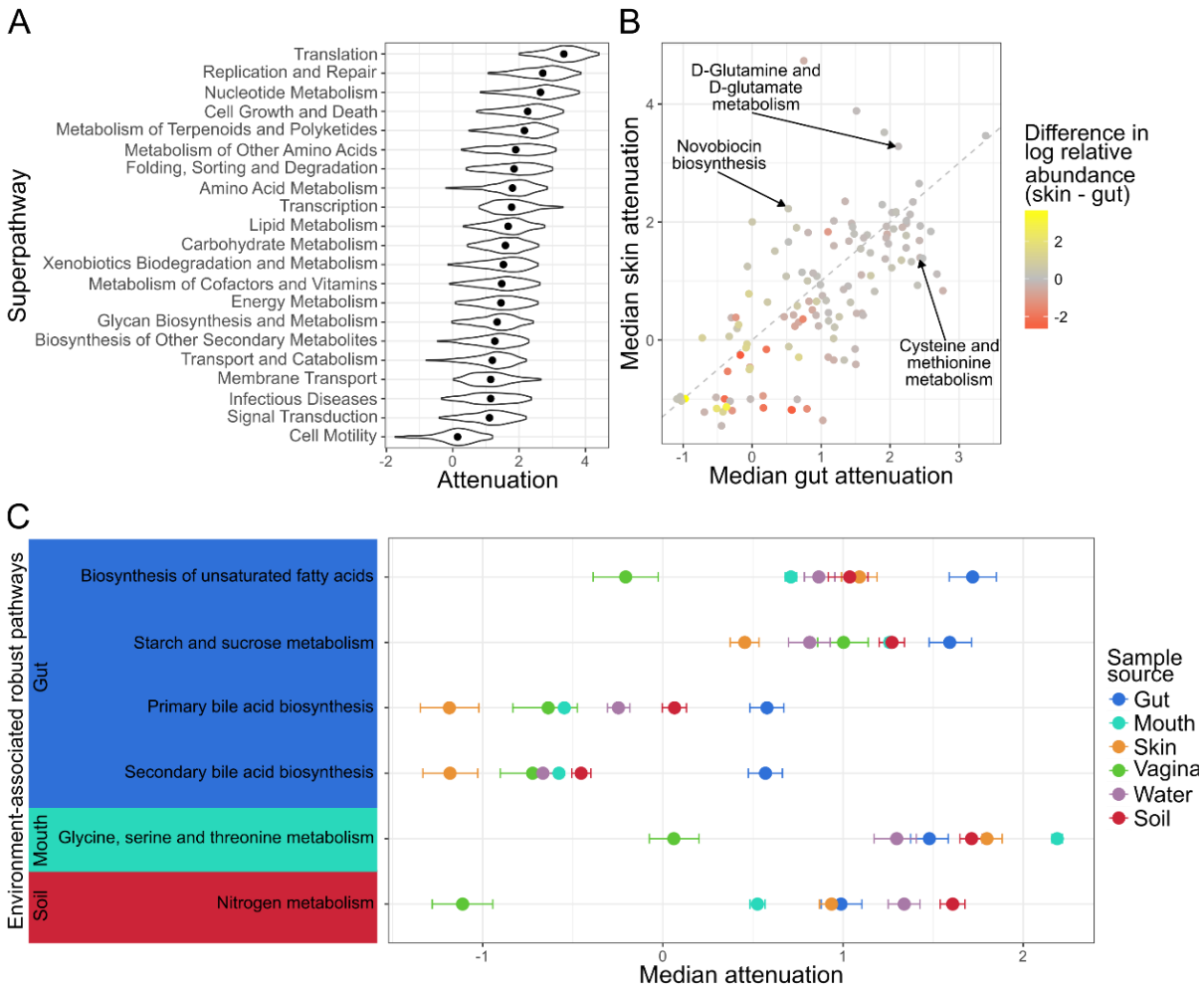

Figure 3.4. Attenuation of individual functions. (A) The density of attenuation values for each KEGG superpathway across all communities, with dots indicating the median attenuation. **(B)** A scatterplot of the median attenuation of each function in skin communities vs. gut communities. The color of each point indicates the difference in log median relative abundance of that function between skin and gut communities. The gray line indicates the 1:1 relationship in median attenuation. Only differentially abundant functions are show (FDR < 0.01; Wilcoxon rank-sum test). **(C)** Each point shows a function's median attenuation in a particular environment with error bars displaying the 95% confidence interval and color indicating the environment.

genes from this pathway in the gut microbiome may be beneficial. Similarly, novobiocin biosynthesis had high robustness specifically in skin communities (2.23 median attenuation; 10th most robust skin function) and its high robustness could be related to novobiocin's antibiotic activity against *Staphylococcus epidermidis* [184]. D-Glutamine and D-glutamate metabolism also had high robustness in skin communities (3.28 median attenuation; 5th most robust skin function)

while being very low in abundance (0.3% median relative abundance). Enrichment for D-glutamine and D-glutamate metabolism in skin communities has been linked to individuals prone to atopic dermatitis [185]. This may suggest that rather than having high robustness to maintain baseline capacity, D-glutamine and D-glutamate metabolism could have high robustness to prevent dramatic increases in abundance.

I finally investigated which functions displayed noticeably higher robustness in one or more environments compared to the others. This analysis revealed many pathways that had significantly higher robustness in specific environments (Figure 7.4), including biosynthesis of unsaturated fatty acids, primary and secondary bile acid biosynthesis, and starch and sucrose metabolism, which were all more robust in the gut than any other environment (Figure 3.4C) and potentially reflected the increased occurrence of metabolites related to these functions in the gut compared to the other environments. In contrast, glycine, serine, and threonine metabolism was most robust in oral communities and could be related to the role of these amino acids in pH recovery in the oral environment after microbial fermentation of carbohydrates [186]. Interestingly, nitrogen metabolism had higher robustness in soil communities compared to the various body site communities and may reflect the role that these communities play in the nitrogen cycle.

3.4.4 **Gene distribution impacts taxa-function robustness across body sites**

As discussed above, the observed variation in taxa-function robustness between and within environments likely reflects differences in the way various genes/functions are distributed across community members in each environment. To characterize how the distribution of functions contributes to robustness variation, I formulated a set of gene distribution features (GDFs), including *average functional redundancy*, *average functional similarity*, *average genome size*, *genome size variability*, and *unique function abundance*, to describe particular aspects of this distribution. *Functional redundancy*, defined here as the redundancy of each function weighted by the relative abundance of that function, has been proposed as an important contributor to the robustness of a community's functional capacities [136,162–164]. *Functional similarity*, defined here as the pairwise similarity in genomic content between species, captures the interchangeability of microbes and the potential for a change in the abundance of one species to be compensated for by an opposite change in the abundance on another. *Genome size*, measured here as the number of

genes in a genome, accounts for the possibility that abundance changes in species with larger genomes will produce larger functional shifts. *Genome size variability* measured the presence of microbes with significantly different genome sizes, which could potentially decrease robustness due to the inability for such microbes to compensate for abundance changes in one another. Finally, *unique function abundance* measures the total abundance of functions that are encoded by a single species, aiming to capture the prevalence of functions with no redundancy (and hence with no potential for compensatory changes). Notably, while some of these GDFs are correlated with each other, these correlation magnitudes are <0.5, suggesting that they indeed capture different aspect of functional distribution (Figure 7.5). Importantly, though previous work has found some association between robustness and species diversity [187,188], here I wish to focus on the impact of functional distribution and hence exclude species diversity as a feature in this analysis.

I calculated these GDFs for each community and examined how they correlate with robustness both within and across environments, aiming to identify universal or environment-specific relationships with robustness. As expected, I found that functional redundancy positively correlates with attenuation both when communities from all environments are pooled together ($r = 0.38$; $p < 10^{-72}$) and within each environment individually (Figure 3.5A). Additionally, functional similarity among community members appears to be positively associated with attenuation when environments are pooled together ($r = 0.24$; $p < 10^{-29}$), as well as in several individual environments (e.g., the gut). Genome size variability, which I hypothesized may negatively impact robustness, does exhibit a negative association with robustness when pooling communities from all environments together, and both genome size and genome size variability also have negative associations in specific environments (i.e., the gut or vagina). Interestingly, however, these associations are not consistent across all environments and may indicate that the relationship between these GDFs and robustness are in part influenced by other features of the community.

Obviously, the impact of each of the GDFs described above on robustness is not independent of the impact of other GDFs, and thus their individual associations with robustness may not reflect how the combination of all five GDFs contributes to robustness. To examine how variation in attenuation is associated with the major axes of variation separating communities based on all GDFs simultaneously, I performed a principal component analysis (PCA) of the five GDFs described above (Figure 3.5B). Remarkably, I found that the first two principal components of these GDFs clearly separate communities by environment, even though they are based on only

five simple summary metrics of gene distribution without directly accounting for the presence or absence of specific microbes or functions. This finding indicates a strong environment-specific

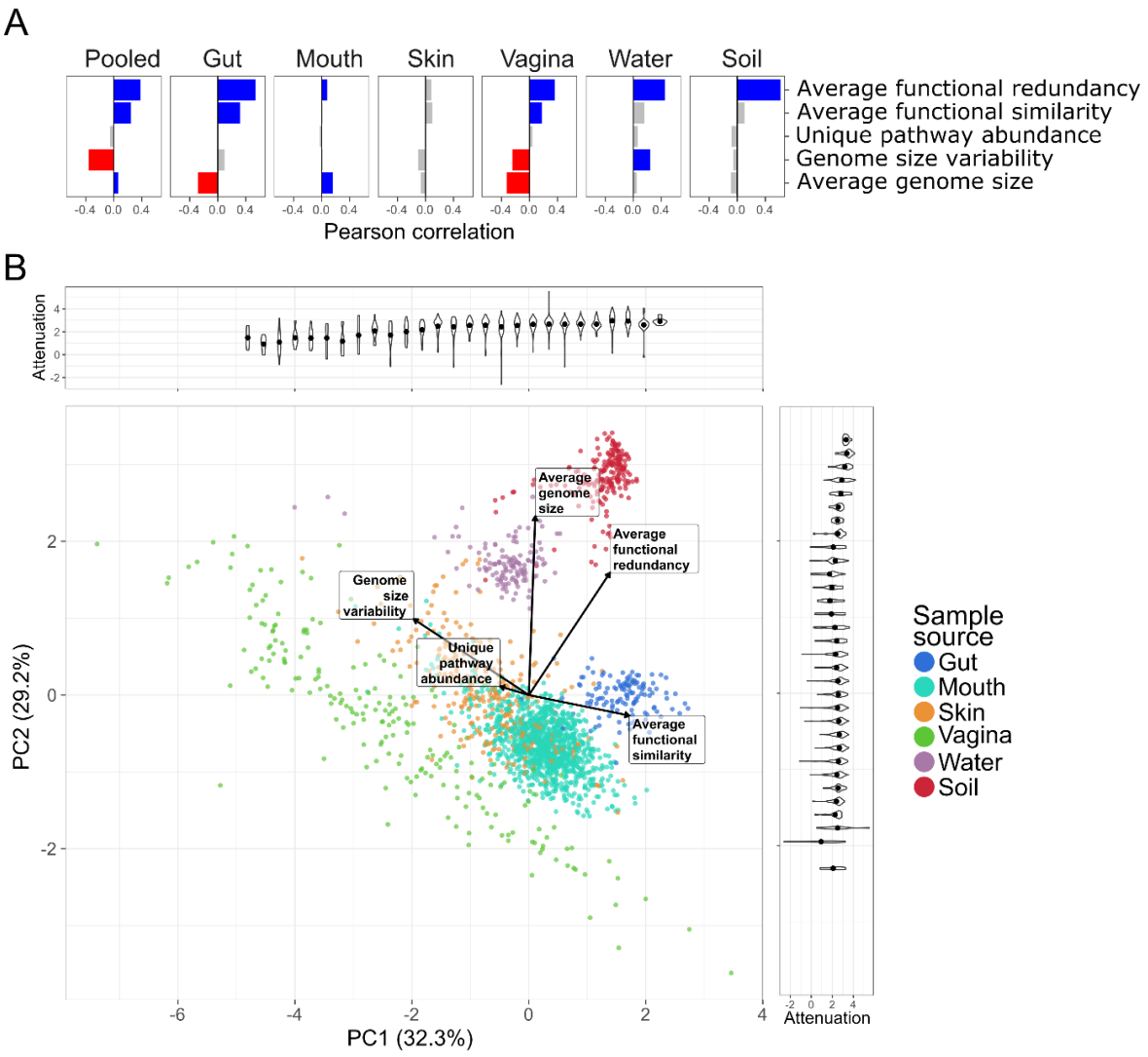

Figure 3.5. Associations between gene distribution features (GDFs) and taxa-function robustness. (A) Pearson correlations between GDFs and attenuation; blue and red bars indicate significant ($p < 0.01$) positive and negative correlation coefficients respectively whereas grey bars show non-significant correlation coefficients. Each panel corresponds to correlations when considering communities from all environments (Pooled) or the subset of communities from a particular environment. **(B)** Communities plotted by the first two principal components (PCs) determined by Principle Component Analysis (PCA) of the five GDFs. The percent variance explained by each PC is indicated on the axis labels. Loadings for GDFs are indicated by the direction and magnitude of the labeled vectors. Communities were binned along both axes and the density of attenuation values is displayed by the width of the violin plots in the plots along the top and right margins of the PCA plot. Dots indicate the median attenuation of communities in each bin.

signature of gene distribution. I further found a significant positive correlation between attenuation and both the first and second principal components (Figure 3.5B; $r = 0.44$; $p < 10^{-99}$and $r = 0.09$; $p < 10^{-4}$ respectively), confirming that the variation in the combination of these GDFs is inherently associated with variation in community robustness. Given these findings, I additionally examined whether GDFs can be used to predict the robustness of each community, using both the set of 5 GDFs above, as well as an expanded set of 45 GDFs (Table 7.2). This analysis again demonstrated the strong association between gene distribution profiles and taxa-function robustness (though predictive power varied markedly between environments) and highlighted the importance of functional redundancy in determining robustness (7.1 Supplementary Text; Figure 7.6; Figure 7.7).

3.4.5 Taxa-function robustness estimations are in agreement with observed functional shifts

The above results rely on simulated taxonomic perturbations and predicted functional profiles (via PICRUSt), rather than on observed taxonomic shifts and shotgun metagenomics-based functional profiling. Clearly, there are caveats involved in both simulated perturbations (which may not reliably capture natural community fluctuations) and prediction accuracy. Yet, as described above, this simulation-based approach is crucial for comprehensively mapping the local taxa-function landscape. Specifically, this approach allows me to survey a large set of perturbed community compositions uniformly distributed around each community and sampled across a range of weighted UniFrac dissimilarities (ranging from 0.001 to 0.4) – a challenging task using available metagenomic data. Nonetheless, to confirm that these simulation-based estimates of robustness are biologically meaningful, here I further examine whether they agree with observed shifts in communities' taxonomic and functional profiles. To this end, I examined functional profiles that are based directly on shotgun metagenomic sequencing for a subset of the HMP communities used in the above analyses. Specifically, considering the HMP communities for which both a shotgun metagenome was available and the 16S rRNA data passed the quality filtering process, I was able to obtain 94 HMP communities from 5 different subsites with both taxonomic and metagenome-based functional profiles. Below, I use these communities to assess the agreement between predicted and metagenome-based functional profiles and between simulation-based robustness estimates and measured functional shifts. For the functional shift comparisons described below, communities from the same subsite were paired (resulting in 47

pairs), providing independent observations of functional shifts within each environment. Additionally, I was able to consider longitudinal shifts for a small number of communities (including 8 communities with 16S rRNA and metagenomic profiles sampled at two time points).

Using these data, I first verified that the compositions of the predicted functional profiles used in this analysis recapitulate metagenome-based functional profiles. To this end, I compared the dissimilarity between each predicted functional profile and its corresponding metagenome-based functional profile to the dissimilarity between each predicted functional profile and a different, randomly chosen metagenome-based functional profile from the same subsite. I found that indeed the median cosine dissimilarity between the corresponding predicted and metagenome-based functional profiles (at the pathway-level) was significantly lower compared to the dissimilarity between predicted functional profiles and other metagenome-based profiles from the same subsite ($p=0.001$, Wilcoxon rank-sum test). A Procrustes analysis [189] of the first two principal components of the functional profiles further demonstrated a significant fit between predicted and metagenome-based profiles (Procrustes measure of fit 0.89; $p=0.00001$).

Next, I set out to examine the degree to which these simulation-based robustness estimates agree with observed functional shifts. In lieu of perturbed community compositions (and comparing the original and perturbed community compositions), I used pairs of communities (from the same subsite), considering one community as representing the original community and the other as representing a perturbed version of that original community. I then measured (for each pair of communities) the ratio between the dissimilarity in their functional profiles and the dissimilarity in their taxonomic profiles. This ratio is expected to be lower for more robust communities since taxonomic perturbations of similar magnitude should produce relatively smaller functional shifts in communities with higher taxa-function robustness. The findings of this analysis confirm the expectation above, demonstrating a noticeably lower robustness in vaginal communities (larger functional shifts relative to taxonomic perturbation magnitudes) compared to gut and oral communities (Figure 3.6A-B). This suggests that the differences in attenuation values estimated from simulated perturbations reflect an inherent difference in how taxonomic changes induce functional shifts in communities from different environments.

Notably, as also discussed above, the taxonomic dissimilarities between different communities are substantially higher than the range of local taxonomic perturbations that are the focus of this study. Given this, the analysis above, using pairs of communities to represent original

and perturbed communities may fail to capture subtler properties of the local landscape of functional shifts around a community. To address this potential shortcoming and examine the impact of small taxonomic differences that are still rooted in metagenome-based functional profiles, I used a community-mixing approach. Specifically, given a pair of communities, a perturbed community composition at a specific (and small) taxonomic distance from the first community can be generated by identifying a mixing fraction and 'replacing' this fraction of the first community's taxonomic and functional profiles with a corresponding fraction of the second community's taxonomic and functional profiles respectively. This community-mixing approach allowed me to generate a set of perturbed community compositions at varying weighted UniFrac dissimilarities with metagenome-based functional profiles. I used this set of multiple perturbed compositions for each community to estimate the robustness of the first community using the same approach described in Figure 3.1D. I found that these community-mixing metagenome-based attenuation values recapitulated the robustness trends identified using simulation-based attenuation estimates, with vaginal communities exhibiting significantly lower robustness than gut or oral communities, further supporting the agreement between simulation-based and metagenome-based robustness estimates (Figure 3.6C).

Finally, I examined whether these robustness estimates agree with temporal changes in community composition. To this end, I used 8 HMP communities with taxonomic and metagenomic data available at two time points to measure longitudinal taxonomic and functional shifts (notably, relatively few HMP communities have both 16S rRNA and metagenome data available at two time points). As above, given a pair of communities from the same subsite from an individual obtained at two time points, I use the dissimilarity ratio between their functional and taxonomic profiles as a measure of observed taxa-function robustness, with the expectation that this ratio will decrease as robustness increases. Indeed, I found that attenuation was negatively correlated with this ratio (r = -0.75). Furthermore, for each individual, I used the simulation-based robustness curve obtained for the community composition at the first time point, the metagenome-based functional profile of that community, and the observed taxonomic dissimilarly between the two time points to predict the expected metagenome-based functional shift between the two time points. Reassuringly, the expected functional shifts were positively correlated with the observed functional shifts (r = 0.86, p=0.006), suggesting that these robustness curves are predictive of the

relationship between the taxonomic perturbations and functional shifts that might be expected to occur over time in microbial communities.

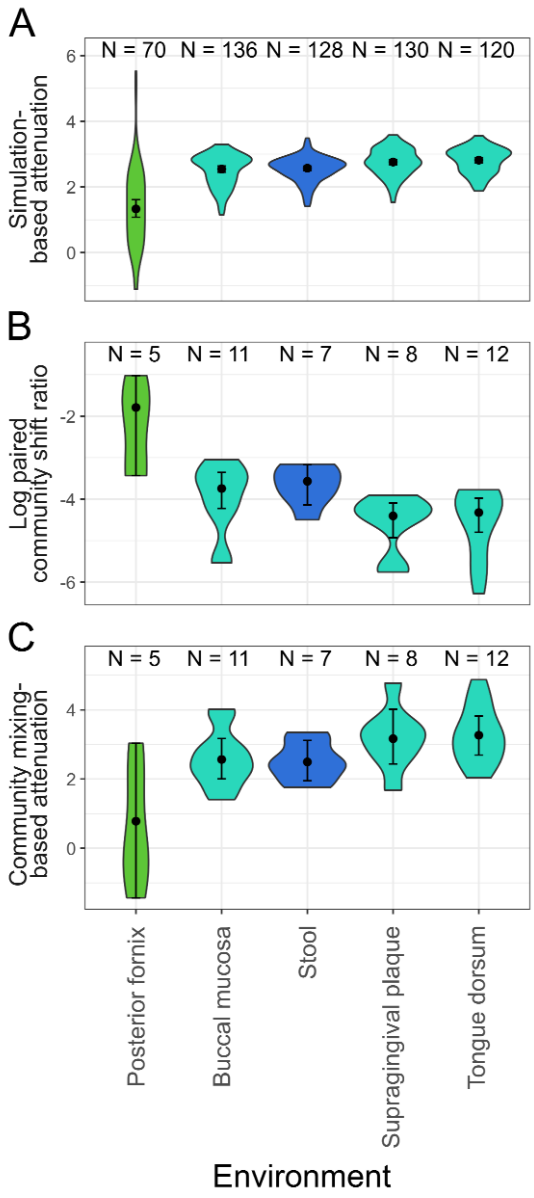

Figure 3.6. Robustness and metagenome-based functional shift trends across subsites. (A) Distribution of simulation-based attenuation estimates across 5 subsites, ordered by median attenuation in each subsite. **(B)** Distribution of log scale ratio of metagenome-based functional profile dissimilarity to taxonomic profile dissimilarity observed between pairs of communities from the same subsite across 5 subsites. **(C)** Distribution of attenuation estimates from mixing community compositions between pairs of communities from the same subsite across 5 subsites.

3.5 Discussion

Microbial communities are recognized as important components of various systems including human health, environmental resource cycling, and industrial processes. Given these varied and significant roles, improving our understanding of these systems requires a better comprehension of how these communities are structured taxonomically, how they function, how they react to change, and the relationships between these features. Previous work using the inherent link between taxonomic structure and function in microbial communities has already led to intriguing results and powerful tools for analyzing microbiome data [130,155–157]. However, a crucial property that has not yet been comprehensively studied is how a community's underlying taxonomic structure modulates functional robustness in response to taxonomic perturbation. This taxa-function robustness is a direct derivative of community composition and of the distribution of genes across genomes, and while it may not correspond directly to conventional concepts of microbial community resilience and stability, it allows me to quantify an inherent component of the structure-function relationship between taxonomic and functional profiles.

My analysis of robustness across various environments revealed intriguing differences between communities from different body sites. One of the more marked differences was between gut and vaginal communities, and while robustness to taxonomic perturbation has not been directly compared between gut and vaginal communities experimentally, there may be some evidence that supports vaginal communities being more susceptible to disrupted function than gut communities [157,190]. I also observed that skin communities from two subsites were among the most robust host-associated communities. This functional robustness to taxonomic perturbation is especially intriguing given recent observation concerning the taxonomic stability of the skin microbiome despite virtually continuous perturbation [191]. Indeed, the link between ecological stability and taxa-function robustness is likely complex and multi-faceted, since on the one hand high taxonomic stability may render taxa-function robustness unnecessary for maintaining community function, but on the other high taxa-function robustness may promote taxonomic stability via maintenance of the community niche.

I further found that functional redundancy was strongly associated with taxa-function robustness, in agreement with previously suggested hypotheses regarding the role of functional redundancy in microbial communities [136,162–164]. Other GDFs also showed some associations

with robustness, but these associations were inconsistent across environments and may point to between-GDF interactions in determining robustness that masks a more consistent association with robustness. Notably, however, the five GDFs I examined appear to separate environments along different principal components, suggesting that they capture key information about environment-specific differences in community structure. This is even more striking when considering that these GDFs do not directly mirror the presence or absence of certain taxa or functions. This GDF-based separation may therefore suggest that communities inhabiting different environments substantially differ in the way community members contribute to overall community functional capacities. For example, the higher functional redundancy and functional similarity in gut communities might indicate that many community members in the gut are performing similar functions with relatively few specialized roles, whereas the increased presence of pathways unique to specific microbes and lower function redundancy in vaginal communities could point to a more well-defined and distinct functional niche in this environment.

To some extent, the extreme variation observed above in taxa-function robustness across environments may not be surprising. Some of this variation, for example, can be attributed to stochastic aspects of community assembly, such as priority effects or the pool of species from which community founders could originate. Such stochastic factors could lead to marked variation in community composition, and consequently to variation in taxa-function robustness. However, the similarity in robustness within a similar environment or subsite compared to between-environment robustness differences suggests that some of this observed variation may be selected for and that selection for robustness varies across environments. Selective pressure for robustness could vary based on the relative needs for consistency versus plasticity in community function when communities respond to changing environmental conditions or based on how consistent the environment might be. Indeed, community-level functional plasticity could be driven by changes in the metabolism or behavior of individual microbes in response to environmental changes, but could also be achieved by shifts in community composition that modulate the community functional profile in a desired direction. Indeed, noticeable alterations to taxonomic composition often accompany large environmental changes, such as in the gut microbiota following diet switches [192], in skin communities during and after atopic dermatitis flares [193], or in the vaginal microbiome during pregnancy [194], and in certain cases these composition alterations have also been associated with changes in community functional capacities [195,196]. In such

cases, lower taxa-function robustness may be selected for to enable plasticity in community function, thereby allowing community function to adapt to a changing environment.

Notably, there are a few caveats to this taxa-function robustness estimation approach. First, the simulation-based perturbation method I used was fairly simple and restricted perturbations to only modify the abundances of present microbes without considering the possible addition of novel microbes, e.g., via migration. In reality, the underlying taxa-function landscape depends not only on the taxa that are present in a community but also on the taxa that could be introduced to the community from an environmental reservoir or transferred from some other source. Yet, the correspondence between these robustness estimates and the observed taxonomic and functional shifts in real communities suggests that the local topology of the taxa-function landscape around a community may be relatively similar when considering the possible influx of new taxa. Beyond the simulation-based perturbation method, these robustness calculations also depend on community functional profile prediction via a database of annotated microbial genomes. These predicted functional profiles may not accurately reflect the true functional profile of the community or variation between communities. For example, the analyses above have considered community members at the species level, ignoring potential strain-level variation [197]. While the taxa-function mapping representation and this approach for robustness estimation can be applied, in principle, at higher taxonomic resolution (e.g., profiling communities at the strain level and associating each strain with a corresponding distinct genome) when such data is available, failing to do so may introduce inaccuracies to the predicted functional profiles and subsequently to these robustness estimates. More generally, robustness estimates can be inflated if rarer functions (that are likely less redundant) are left unannotated or deflated if certain functions are actually highly redundant but missing from some genome annotations. It is worth noting, however, that despite these caveats, these robustness estimates were shown to agree with differences in taxonomic and functional profiles between naturally occurring community compositions, both across communities and over time.

My analysis of simulated community perturbations indicates that taxa-function robustness of microbial communities varies by environment, though subsites from within a given environment tend to share similar taxa-function robustness signatures. Furthermore, function-specific robustness at the pathway level is associated with the universality of the pathway, with microbial housekeeping pathways displaying higher robustness than pathways associated with more

specialized lifestyles. Interestingly, the variation in the robustness of a pathway across communities was not associated with differences in pathway abundance. I also found that environment-specific characteristics of gene distribution across community member genomes account for between-environment differences in taxa-function robustness and suggest potential drivers for functional robustness. Finally, a comparison between simulation-based robustness estimates and metagenome-based taxonomic and functional shifts suggested that these robustness estimates agree with observed community dynamics.

Importantly, the applications for computational robustness estimation could extend beyond the analyses and results presented here. Function-specific robustness estimation, for example, could inform the analysis of novel communities from a particular environment, highlighting functions whose capacities are more or less robust than expected. Taxa-function robustness estimation could also be incorporated into the construction of synthetic microbial communities to inform the design of more resilient community compositions. As we further explore and analyze the temporal dynamics of microbial community structure and function, being able to determine how and why robustness varies will continue to be of interest.

4. Concluding Remarks

Sections of this chapter are based on the following manuscript submitted to *Current Opinion in Biotechnology*:

Eng A, Borenstein E. Microbial Community Design: Methods, Applications, and Opportunities. *Current Opinion in Biotechnology* Submitted.

In this dissertation I have developed novel model-based computational methods that utilize the link between taxonomy and microbial functional capacities to aid the design and evaluation of synthetic microbial communities. Computational methods such as these offer a promising prospect for synthetic community design because they can reduce the labor and time required by previous approaches by exploiting our growing mechanistic knowledge of microbial function. This can also provide a more nuanced understanding of how designed communities function, which could allow microbiome engineers to consider the conditions under which the desired function might be disrupted. The work I have described here represents an initial foray into this area, laying the groundwork for future progress, providing tools that can support current design efforts, and offering insights into the structure of naturally occurring communities.

4.1 Summary of presented work

In the second chapter, I detailed CoMiDA, an algorithm for identifying minimal sets of species with the capacity to perform a desired metabolic conversion. Utilizing a database of functionally-annotated genomic content, CoMiDA translates this design problem into an integer linear programing representation of an integrated network flow and set cover framework. After verifying its ability to correctly identify non-intuitive minimal communities, I used CoMiDA to analyze minimal communities for metabolic conversions using gut-associated species and found that, when possible, conversions required only a small number of species.

In the third chapter, I presented a perturbation-based method for estimating the robustness of a community's functional capacities with respect to changes in taxonomic composition. This method is a novel computational technique that offers insight into a previously unconsidered facet

of community structure that can be informative when evaluating community compositions during the design process. Using this method, I demonstrated differences in taxa-function robustness factors for naturally occurring communities associated with different body sites and environmental sources. I also showed that the community-level capacities for specific functions varied in taxa-function robustness in a manner that may be related to environment-specific pressures. Finally, I developed metrics to describe the distribution of genes across species in a community and found that, in addition to being associated with taxa-function robustness, gene distribution features had environment-specific signatures.

4.1.1 Potential role of functional redundancy in engineering taxa-function robustness

Based on the results presented in the third chapter, it appears that taxa-function robustness is notably associated with functional redundancy, defined here as the evenness of contributions to a function's abundance across different taxa. This finding agrees with previous discussions regarding the role of functional redundancy in promoting the robustness and stability of microbial community function [136,162–164]. Furthermore, this lends support to the concept that functional redundancy will be an important factor to consider when engineering communities for persistence of function. In this work I mainly considered functional redundancy across all functional capacities in a community, but targeted applications will likely benefit from focusing on the redundancy of specific, desirable functions. For example, therapeutics that aim to promote the production of SCFAs may prefer the inclusion of multiple different species or strains with SCFA-production capabilities, rather than a single SCFA-producing species. This could buffer against a perturbation, such as a change in diet or an infection, that might negatively affect a specific SCFA producer because other present SCFA producers may be less affected by that particular perturbation.

4.2 Recent advances in computational design methods

Following and concurrent with publication of the above works, additional novel computational model-based design methods have been developed and introduced, further advancing the field and providing complementary approaches to the methods presented here. For example, MultiPus uses a framework similar to CoMiDA, but instead of minimizing community size, it minimizes the number of metabolic reactions and inter-species metabolite transfers required for the desired metabolic conversion [198]. Thus, while CoMiDA focuses on providing the

smallest community to provide the simplest starting point for further development with other design tools, MultiPus instead aims to provide a community with the capacity to more efficiently perform the desired function. Both of these are valid concerns when starting to engineer a synthetic community, and which method is appropriate will depend on the needs of the specific application.

Network-based models, such as those used in CoMiDA and MultiPus, are easy to construct and analyze, but they generally only account for the potential metabolic capacity of each species, rather than for the way each species will behave in a given environment. Accordingly, communities designed by CoMiDA or MultiPus are indeed guaranteed to have the metabolic *potential* to carry out the desired function, but may not actually perform this function in reality. To address this shortcoming, a recent design method, termed FLYCOP [199], utilizes a previously introduced FBA-based community modeling framework to evaluate synthetic composition function *in silico*. The underlying modeling framework accounts for community dynamics and spatial community organization, as well as metabolism-mediated species interactions via changes to metabolite concentrations in the shared environment [140]. Using this framework, FLYCOP explores potential synthetic compositions using a stochastic search procedure and identifies an optimized composition. Due to the community FBA framework used, FLYCOP is not restricted to optimizing metabolic activity and can also consider a community's growth over time. As a demonstration of this, the FLYCOP developers identified an initial synthetic composition of four cross-feeding strains that optimized community stability. These capabilities make FLYCOP an exceptionally comprehensive and valuable design tool, and will serve as a strong basis for future developments in FBA-based computational design methods.

Another recently applied modeling approach aims to engineer communities based on models of ecological dynamics, rather than community metabolism, which can be useful stability is an important consideration. These models capture how the abundance of each community member impacts the abundances of others over time [200,201]. Such models can be especially useful when interactions between species are not mediated via metabolism or when detailed metabolic models are unavailable. For example, a synthetic community was optimized for both a non-metabolic function (T_{reg} induction) and community stability through a combination of ecological modeling and experimental characterization of individual microbial activity [202]. To achieve this, the authors first created a model of community induction effectiveness for a set of *Clostridia* strains using data on T_{reg} induction contributions. They then simulated the community's

ecological dynamics using a previously published ecological model for those strains [200], and used their induction model to estimate T_{reg} induction over time. This enabled them to predict each potential composition's inductive effect and stability simultaneously. Such integration of different modeling frameworks may be a promising avenue for future expansion and improvement of computational design capabilities.

4.3 Future directions

The power and flexibility of computational model-based design, combined with the recent expansion of available mechanistic data, make computational design efforts an especially promising route toward rapid advancement in synthetic community design. For example, computational methods that can concurrently optimize multiple community functions could enable synthetic therapeutic communities to simultaneously treat diverse health concerns, such as metabolic deficiencies and pathogenic infections, while also ensuring community stability. There are, however, substantial obstacles that must be overcome for model-based design to reach its full potential. Here I describe current challenges that computational models face and discuss the potential future of computational design methods, and synthetic community design techniques in general.

4.3.1 Challenges for model-based techniques

One key challenge is the inability of most current modeling frameworks to directly incorporate effects of non-metabolic microbial interactions into models of community function. Recent evidence suggests that non-metabolic interactions, such as the microbial secretion of toxins to inhibit the growth of competitors, can be important factors in shaping the human gut microbiome [203]. Ecological models can attempt to capture the impact of such interactions on community dynamics by using time-series data to infer interaction parameters between species. However, this approach may only partially capture the outcomes of such interactions, especially when multiple interaction mediators are involved, and it is likely that the nuanced effects of key mediators are ignored [204]. Instead, adequately integrating these effects into current methods may require augmentation with mechanistic models of these non-metabolic influences on community function and dynamics.

The potentially important functional impact of higher-order interactions (i.e., interactions involving multiple species in the community) [205–207] poses another challenge for computational methods. Evaluating a subset of possible compositions, as with FLYCOP's stochastic sampling, may miss the effects of certain higher-order interactions because the specific combinations of species required were never present in the evaluated compositions. Eventually, it may become convenient to evaluate potential species combinations, in terms of presence or absence, given sufficient computing resources and parallelized evaluation. However, functions and interactions may be determined by species abundances, as in quorum sensing [208], and it is impossible to enumerate, let alone evaluate, all conceivable community compositions when considering relative abundance differences. Unfortunately, there are no available methods that can extrapolate from species-specific mechanistic models to predictions of emergent functions due to higher-order, abundance-dependent interactions. If such techniques are feasible, then their development could allow computational methods to avoid extensive community evaluation while simultaneously enabling a more comprehensive consideration of possible interactions.

4.3.2 **Potential avenues for advancement**

The ability to integrate prior biological knowledge of microbial behavior is currently difficult when corresponding mechanistic or ecological models are unavailable, and yet there is a wealth of observational data on microbial activity that could inform community design. For instance, various studies detail the environmental and ecological effects of species under specific conditions, including influences on host systems [209,210] and impacts on the functions of other species [211,212]. Rule-based frameworks offer an approach that can take advantage of this information to design communities by intuitively translating observed biological phenomena into expectations of, and constraints on, community function. Rules can, for example, take the form of logical expressions that indicate the observed outcomes of microbial function, e.g., translating the results from Mathur et al. [210] to "presence of *Bacillus cereus* in the gut implies activation of the NLRP3 inflammasome, secretion of interleukin-1β, and secretion of interleukin-18". This example portrays a fairly simplistic rule, but more nuance is possible via higher resolution definitions of microbiome states, such as species and compound abundance thresholds. Another benefit of these rule articulations is their intuitive representation, which can enable experts on specific microbial systems to easily express literature-based knowledge.

Designing microbial communities based on rules of the above form can be achieved by converting them into an instance of a SATISFIABILITY (SAT) problem [213]. Briefly, a SAT problem asks whether a set of variables can be assigned true or false values such that they satisfy a collection of clauses that relate the variables to one another and constrain their values. In terms of microbiome engineering, these variables would represent community and environmental components, including species, metabolites, and signaling molecules, as well as effects resulting from community activity such as protection against pathogens or gut inflammation. The true/false value of each variable would indicate presence/absence for microbiome components or occurrence/absence for community effects. The clauses relating these variables would encode identified rules (e.g., if the *B. cereus* variable is true, then the NLRP3 activation, interleukin-1β secretion, and interleukin-18 secretion variables must also be true). Given these clauses and variables, a satisfying set of variable assignments will define a community composition (the true species variables). Additionally, these values will also describe the community's influence on the environment, including produced compounds and active community effects (again indicated by the corresponding true variables). Communities can be designed using this framework by pre-specifying the values of specific variables, i.e., interleukin-18 secretion must be true, and then finding a satisfying assignment for the remaining variables. Notably, this approach could offer an extremely flexible framework for incorporating knowledge of multiple distinct facets of microbial community function. However, it is currently challenging to assemble an adequate body of rules because the necessary data often being presented in terms unamenable to automatic and large-scale rule formulation, and future work will be necessary to overcome this obstacle.

Improved design capabilities may also be achieved by combining rule-based frameworks with mechanistic models, utilizing each in tandem to simultaneously consider the facets of community function that they are individually most suitable for. Similar to the integrated ecological and T_{reg} induction model mentioned above [202], community-level dynamic FBA models could employ rule-based frameworks to evaluate and incorporate non-metabolic effects. More specifically, at each time point, available rules could be evaluated based on current community composition and environmental state to determine the consequences of non-metabolic functions and interactions. This data would then inform the standard dynamic FBA-based adjustments to species and metabolite abundances while also providing an estimate of the

community's contributions to non-metabolic environmental conditions such as gut inflammation, MFC electricity generation, or antimicrobial production.

Complementary design method integration may also offer promising opportunities more generally in the field of synthetic design by considering both experimental and computational methods. Indeed, there is already evidence that communities designed using one design framework can be further improved via orthogonal techniques. The aforementioned synthetic T_{reg} induction community designed using computational model-based methods [202] was developed from a T_{reg} induction community originally engineered using community reduction [89]. In this case, community reduction identified a set of culturable species that could form a synthetic T_{reg} induction community and computational design optimized the composition to improve its function. Pairing community enrichment with combinatorial evaluation or computational design could also offer potentially fruitful composite approaches. Specifically, the enrichment procedure could begin with compositions identified and pre-optimized by other design techniques, rather than environmentally sampled or randomly constructed initial communities. Such method integration could potentially reduce the time required to select for an optimal composition since the initial community is hopefully closer in composition to the final community that enrichment would achieve. Additionally, this could help augment combinatorial evaluation or computational design, which may result in sub-optimal communities due to insufficient coverage of evaluated compositions or insufficient mechanistic and ecological knowledge respectively. Such innovative combinations of design approaches could enhance or enable the development of synthetic compositions that may have previously been challenging due to the various limitations of each individual approach.

4.4 Closing thoughts

Past microbiome engineering efforts have resulted in effective synthetic microbial communities for a variety of applications, and the future of microbial community-based biotechnologies and therapeutics is promising. Here, I have presented computational methods that begin to explore the engineering possibilities available when considering the genomic link between microbial species and their function. While these methods may be limited in some respects regarding the modeling of community behavior, they provide an important stepping stone towards future method development. Further development of such computational design methods should

provide powerful tools to support advances in microbiome technology, especially as quality and availability of mechanistic data related to microbial function continues to grow.

5. References

1. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* 2012; 486:207–214.
2. Makhalanyane TP, Valverde A, Gunnigle E, Frossard A, Ramond J-B, *et al.* Microbial ecology of hot desert edaphic systems. *FEMS Microbiol Rev* 2015; 39:203–221.
3. Pedersen K. The deep subterranean biosphere. *Earth-Science Rev* 1993; 34:243–260.
4. Corzett CH, Elsherbini J, Chien DM, Hehemann J-H, Henschel A, *et al.* Evolution of a vegetarian Vibrio: metabolic specialization of Vibrio breoganii to macroalgal substrates. *J Bacteriol* 2018; 200:e00020-18.
5. Hall-Stoodley L, Costerton JW, Stoodley P. Bacterial biofilms: from the natural environment to infectious diseases. *Nat Rev Microbiol* 2004; 2:95–108.
6. Lebre PH, De Maayer P, Cowan DA. Xerotolerant bacteria: surviving through a dry spell. *Nat Rev Microbiol* 2017; 15:285–296.
7. Rakoff-Nahoum S, Coyne MJ, Comstock LE. An ecological network of polysaccharide utilization among human intestinal symbionts. *Curr Biol* 2014; 24:40–49.
8. Gonzalez-Gil G, Lens PNL, Saikaly PE. Selenite reduction by anaerobic microbial aggregates: microbial community structure, and proteins associated to the produced selenium spheres. *Front Microbiol* 2016; 7:571.
9. Boetius A, Ravenschlag K, Schubert CJ, Rickert D, Widdel F, *et al.* A marine microbial consortium apparently mediating anaerobic oxidation of methane. *Nature* 2000; 407:623–626.
10. Chirwa EMN, Bezza FA. Petroleum hydrocarbon spills in the environment and abundance of microbial community capable of biosurfactant production. *J Pet Env Biotechnol* 2015; 6:237.
11. Sender R, Fuchs S, Milo R. Revised estimates for the number of human and bacteria cells in the body. *PLOS Biol* 2016; 14:e1002533.
12. Orrhage K, Nord CE. Factors controlling the bacterial colonization of the intestine in breastfed infants. *Acta Paediatr* 1999; 88:47–57.
13. Björkstén B, Sepp E, Julge K, Voor T, Mikelsaar M. Allergy development and the intestinal microflora during the first year of life. *J Allergy Clin Immunol* 2001; 108:516–

520.

14. Mazmanian SK, Liu CH, Tzianabos AO, Kasper DL. An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell* 2005; 122:107–118.
15. Are A, Aronsson L, Wang S, Greicius G, Lee YK, *et al. Enterococcus faecalis* from newborn babies regulate endogenous PPARγ activity and IL-10 levels in colonic epithelial cells. *Proc Natl Acad Sci U S A* 2008; 105:1943–8.
16. LeBlanc JG, Milani C, de Giori GS, Sesma F, van Sinderen D, *et al.* Bacteria as vitamin suppliers to their host: a gut microbiota perspective. *Curr Opin Biotechnol* 2013; 24:160–168.
17. Morrison DJ, Preston T. Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. *Gut Microbes* 2016; 7:189–200.
18. Canani RB, Costanzo M Di, Leone L, Pedata M, Meli R, *et al.* Potential beneficial effects of butyrate in intestinal and extraintestinal diseases. *World J Gastroenterol* 2011; 17:1519–28.
19. Donohoe DR, Garge N, Zhang X, Sun W, O'Connell TM, *et al.* The microbiome and butyrate regulate energy metabolism and autophagy in the mammalian colon. *Cell Metab* 2011; 13:517–526.
20. Greenblum S, Turnbaugh PJ, Borenstein E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc Natl Acad Sci U S A* 2012; 109:594–9.
21. Ridaura VK, Faith JJ, Rey FE, Cheng J, Duncan AE, *et al.* Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science* 2013; 341:1241214.
22. Buffie CG, Pamer EG. Microbiota-mediated colonization resistance against intestinal pathogens. *Nat Rev Immunol* 2013; 13:790–801.
23. Buswell A, Long H. Microbiology and theory of activated sludge. *J Am Water Work Assoc* 1923; 10:309–321.
24. Alleman JE, Prakasam TBS. Reflections on seven decades of activated sludge history. J. Water Pollut. Control Fed. 1983; 55:436–443.
25. Narihiro T, Sekiguchi Y. Microbial communities in anaerobic digestion processes for waste and wastewater treatment: a microbiological update. *Curr Opin Biotechnol* 2007;

18:273–278.

26. Grady, C P L J. Biodegradation of hazardous wastes by conventional biological treatment. *Hazard Waste Hazard Mater* 1986; 3:333–365.
27. Reimers CE, Tender LM, Fertig S, Wang W. Harvesting energy from the marine sediment−water interface. *Environ Sci Technol* 2001; 35:192–195.
28. Lovley DR. Microbial fuel cells: novel microbial physiologies and engineering approaches. *Curr Opin Biotechnol* 2006; 17:327–332.
29. Logan BE, Hamelers B, Rozendal R, Shröder U, Keller J, *et al.* Microbial fuel cells: methodology and technology. *Environ Sci Technol* 2006; 40:5181–5192.
30. Kargi F, Eker S. Electricity generation with simultaneous wastewater treatment by a microbial fuel cell (MFC) with Cu and Cu–Au electrodes. *J Chem Technol Biotechnol* 2007; 82:658–662.
31. Moon H, Chang IS, Kim BH. Continuous electricity production from artificial wastewater using a mediator-less microbial fuel cell. *Bioresour Technol* 2006; 97:621–627.
32. Mao C, Feng Y, Wang X, Ren G. Review on research achievements of biogas from anaerobic digestion. *Renew Sustain Energy Rev* 2015; 45:540–555.
33. Mamma D, Koullas D, Fountoukidis G, Kekos D, Macris BJ, *et al.* Bioethanol from sweet sorghum: simultaneous saccharification and fermentation of carbohydrates by a mixed microbial culture. *Process Biochem* 1996; 31:377–381.
34. Yun Y-M, Lee M-K, Im S-W, Marone A, Trably E, *et al.* Biohydrogen production from food waste: current status, limitations, and future perspectives. *Bioresour Technol* 2018; 248:79–87.
35. Kuypers MMM, Marchant HK, Kartal B. The microbial nitrogen-cycling network. *Nat Rev Microbiol* 2018; 16:263–276.
36. Peoples MB, Herridge DE, Ladha JK. Biological nitrogen fixation: an efficient source of nitrogen for sustainable agricultural production? *Plant Soil* 1995; 174:3–28.
37. Föllmi KB. The phosphorus cycle, phosphogenesis and marine phosphate-rich deposits. *Earth-Science Rev* 1996; 40:55–124.
38. Pradhan N, Sukla LB. Solubilization of inorganic phosphates by fungi isolated from agriculture soil. *African J Biotechnol* 2006; 5.
39. Lebeis SL. The potential for give and take in plant-microbiome relationships. *Front Plant*

Sci 2014; 5:287.

40. Brenner K, You L, Arnold FH. Engineering microbial consortia: a new frontier in synthetic biology. *Trends Biotechnol* 2008; 26:483–9.
41. Mimee M, Citorik RJ, Lu TK. Microbiome therapeutics — advances and challenges. *Adv Drug Deliv Rev* 2016; 105:44–54.
42. Marchesi JR, Adams DH, Fava F, Hermes GDA, Hirschfield GM, *et al.* The gut microbiota and host health: a new clinical frontier. *Gut* 2016; 65:330–9.
43. Wong P, Gladney S, Keasling JD. Mathematical model of the lac operon: inducer exclusion, catabolite repression, and diauxic growth on glucose and lactose. *Biotechnol Prog* 1997; 13:132–143.
44. Blum J-M, Su Q, Ma Y, Valverde-Pérez B, Domingo-Félez C, *et al.* The pH dependency of N-converting enzymatic processes, pathways and microbes: effect on net N_2O production. *Environ Microbiol* 2018; 20:1623–1640.
45. Mosier AC, Li Z, Thomas BC, Hettich RL, Pan C, *et al.* Elevated temperature alters proteomic responses of individual organisms within a biofilm community. *ISME J* 2015; 9:180–194.
46. Cui M, Yuan Z, Zhi X, Shen J. Optimization of biohydrogen production from beer lees using anaerobic mixed bacteria. *Int J Hydrogen Energy* 2009; 34:7971–7978.
47. Fang HHP, Liu H. Effect of pH on hydrogen production from glucose by a mixed culture. *Bioresour Technol* 2002; 82:87–93.
48. Bassin JP, Kleerebezem R, Rosado AS, van Loosdrecht MCM, Dezotti M. Effect of different operational conditions on biofilm development, nitrification, and nitrifying microbial population in moving-bed biofilm reactors. *Environ Sci Technol* 2012; 46:1546–1555.
49. Pholchan MK, Baptista J de C, Davenport RJ, Curtis TP. Systematic study of the effect of operating variables on reactor performance and microbial diversity in laboratory-scale activated sludge reactors. *Water Res* 2010; 44:1341–1352.
50. Simova ED, Frengova GI, Beshkova DM. Effect of aeration on the production of carotenoid pigments by *Rhodotorula rubra-Lactobacillus casei* subsp. *casei* co-cultures in whey ultrafiltrate. *Z Naturforsch* 2003; 58:225–229.
51. Lee SK, Chou H, Ham TS, Lee TS, Keasling JD. Metabolic engineering of

microorganisms for biofuels production: from bugs to synthetic biology to fuels. *Curr Opin Biotechnol* 2008; 19:556–563.

52. Peralta-Yahya PP, Zhang F, del Cardayre SB, Keasling JD. Microbial engineering for the production of advanced biofuels. *Nature* 2012; 488:320–328.
53. Kirby J, Keasling JD. Biosynthesis of plant isoprenoids: perspectives for microbial engineering. *Annu Rev Plant Biol* 2009; 60:335–355.
54. Patnaik R. Engineering complex phenotypes in industrial strains. *Biotechnol Prog* 2008; 24:38–47.
55. Jullesson D, David F, Pfleger B, Nielsen J. Impact of synthetic biology and metabolic engineering on industrial production of fine chemicals. *Biotechnol Adv* 2015; 33:1395–1402.
56. Wu G, Yan Q, Jones JA, Tang YJ, Fong SS, *et al.* Metabolic burden: cornerstones in synthetic biology and metabolic engineering applications. *Trends Biotechnol* 2016; 34:652–664.
57. Reis MAM, Serafim LS, Lemos PC, Ramos AM, Aguiar FR, *et al.* Production of polyhydroxyalkanoates by mixed microbial cultures. *Bioprocess Biosyst Eng* 2003; 25:377–385.
58. Choi J, Lee SY. Process analysis and economic evaluation for Poly(3-hydroxybutyrate) production by fermentation. *Bioprocess Eng* 1997; 17:335.
59. Kazamia E, Aldridge DC, Smith AG. Synthetic ecology – A way forward for sustainable algal biofuel production? *J Biotechnol* 2012; 162:163–169.
60. Arai T, Matsuoka S, Cho H-Y, Yukawa H, Inui M, *et al.* Synthesis of *Clostridium cellulovorans* minicellulosomes by intercellular complementation. *Proc Natl Acad Sci U S A* 2007; 104:1456–60.
61. Martínez I, Mohamed ME-S, Rozas D, García JL, Díaz E. Engineering synthetic bacterial consortia for enhanced desulfurization and revalorization of oil sulfur compounds. *Metab Eng* 2016; 35:46–54.
62. Shin H-D, McClendon S, Vo T, Chen RR. *Escherichia coli* binary culture engineered for direct fermentation of hemicellulose to a biofuel. *Appl Environ Microbiol* 2010; 76:8150–9.
63. Eiteman MA, Lee SA, Altman E. A co-fermentation strategy to consume sugar mixtures

effectively. *J Biol Eng* 2008; 2:3.

64. Brenner K, Karig DK, Weiss R, Arnold FH. Engineered bidirectional communication mediates a consensus in a microbial biofilm consortium. *Proc Natl Acad Sci U S A* 2007; 104:17300–4.
65. Sommer F, Anderson JM, Bharti R, Raes J, Rosenstiel P. The resilience of the intestinal microbiota influences health and disease. *Nat Rev Microbiol* 2017; 15:630–638.
66. Philippot L, Spor A, Hénault C, Bru D, Bizouard F, *et al.* Loss in microbial diversity affects nitrogen cycling in soil. *ISME J* 2013; 7:1609–1619.
67. Leekha S, Terrell CL, Edson RS. General principles of antimicrobial therapy. *Mayo Clin Proc* 2011; 86:156–67.
68. Walsh CJ, Guinane CM, O'Toole PW, Cotter PD. Beneficial modulation of the gut microbiota. *FEBS Lett* 2014; 588:4120–4130.
69. Derrien M, van Hylckama Vlieg JET. Fate, activity, and impact of ingested bacteria within the human gut microbiota. *Trends Microbiol* 2015; 23:354–66.
70. Limbergen H V., Top EM, Verstraete W. Bioaugmentation in activated sludge: current features and future perspectives. *Appl Microbiol Biotechnol* 1998; 50:16–23.
71. Čater M, Fanedl L, Malovrh Š, Marinšek-Logar R. Biogas production from brewery spent grain enhanced by bioaugmentation with hydrolytic anaerobic bacteria. *Bioresour Technol* 2015; 186:261–269.
72. Innemanová P, Filipová A, Michalíková K, Wimmerová L, Cajthaml T. Bioaugmentation of PAH-contaminated soils: a novel procedure for introduction of bacterial degraders into contaminated soil. *Ecol Eng* 2018; 118:93–96.
73. Bartlett JG. Narrative review: the new epidemic of *Clostridium difficile* –associated enteric disease. *Ann Intern Med* 2006; 145:758.
74. Theriot CM, Koenigsknecht MJ, Carlson PE, Hatton GE, Nelson AM, *et al.* Antibiotic-induced shifts in the mouse gut microbiome and metabolome increase susceptibility to Clostridium difficile infection. *Nat Commun* 2014; 5:3114.
75. Zmora N, Zilberman-Schapira G, Suez J, Mor U, Dori-Bachash M, *et al.* Personalized gut mucosal colonization resistance to empiric probiotics is associated with unique host and microbiome features. *Cell* 2018; 174:1388–1405.e21.
76. Maldonado-Gómez MX, Martínez I, Bottacini F, O'Callaghan A, Ventura M, *et al.* Stable

engraftment of *Bifidobacterium longum* AH1206 in the human gut depends on individualized features of the resident microbiome. *Cell Host Microbe* 2016; 20:515–526.

77. Bond DR, Holmes DE, Tender LM, Lovley DR. Electrode-reducing microorganisms that harvest energy from marine sediments. *Science* 2002; 295:483–5.
78. Lu L, Huggins T, Jin S, Zuo Y, Ren ZJ. Microbial metabolism and community structure in response to bioelectrochemically enhanced remediation of petroleum hydrocarbon-contaminated soil. *Environ Sci Technol* 2014; 48:4021–4029.
79. Rabaey K, Boon N, Siciliano SD, Verhaege M, Verstraete W. Biofuel cells select for microbial consortia that self-mediate electron transfer. *Appl Environ Microbiol* 2004; 70:5373–82.
80. Moralejo-Gárate H, Mar'atusalihat E, Kleerebezem R, van Loosdrecht MCM. Microbial community engineering for biopolymer production from glycerol. *Appl Microbiol Biotechnol* 2011; 92:631–9.
81. Wang J, Wan W. Comparison of different pretreatment methods for enriching hydrogen-producing bacteria from digested sludge. *Int J Hydrogen Energy* 2008; 33:2934–2941.
82. Petrof EO, Khoruts A. From stool transplants to next-generation microbiota therapeutics. *Gastroenterology* 2014; 146:1573–1582.
83. Pham VHT, Kim J. Cultivation of unculturable soil bacteria. *Trends Biotechnol* 2012; 30:475–484.
84. Leffler DA, Lamont JT. *Clostridium difficile* infection. *N Engl J Med* 2015; 372:1539–1548.
85. Rossen NG, MacDonald JK, de Vries EM, D'Haens GR, de Vos WM, *et al.* Fecal microbiota transplantation as novel therapy in gastroenterology: a systematic review. *World J Gastroenterol* 2015; 21:5359–71.
86. Tvede M, Rask-Madsen J. Bacteriotherapy for chronic relapsing Clostridium difficile diarrhoea in six patients. *Lancet* 1989; 333:1156–1160.
87. Petrof EO, Gloor GB, Vanner SJ, Weese SJ, Carter D, *et al.* Stool substitute transplant therapy for the eradication of *Clostridium difficile* infection: "RePOOPulating" the gut. *Microbiome* 2013; 1:3.
88. Caballero S, Kim S, Carter RA, Leiner IM, Sušac B, *et al.* Cooperating commensals restore colonization resistance to vancomycin-resistant *Enterococcus faecium*. *Cell Host*

Microbe 2017; 21:592–602.e4.

89. Atarashi K, Tanoue T, Oshima K, Suda W, Nagano Y, *et al.* T_{reg} induction by a rationally selected mixture of Clostridia strains from the human microbiota. *Nature* 2013; 500. doi:10.1038/nature12331
90. Atarashi K, Tanoue T, Ando M, Kamada N, Nagano Y, *et al.* Th17 cell induction by adhesion of microbes to intestinal epithelial cells. *Cell* 2015; 163:367–380.
91. Bader J, Mast-Gerlach E, Popović MK, Bajpai R, Stahl U. Relevance of microbial coculture fermentations in biotechnology. *J Appl Microbiol* 2010; 109:371–87.
92. Ayed L, Achour S, Khelifi E, Cheref A, Bakhrouf A. Use of active consortia of constructed ternary bacterial cultures via mixture design for Congo Red decolorization enhancement. *Chem Eng J* 2010; 162:495–502.
93. Gunst RF, Mason RL. Fractional factorial design. *Wiley Interdiscip Rev Comput Stat* 2009; 1:234–244.
94. Skonieczny MT, Yargeau V. Biohydrogen production from wastewater by *Clostridium beijerinckii*: effect of pH and substrate concentration. *Int J Hydrogen Energy* 2009; 34:3288–3294.
95. Jiménez J, Guardia-Puebla Y, Romero-Romero O, Cisneros-Ortiz ME, Guerra G, *et al.* Methanogenic activity optimization using the response surface methodology, during the anaerobic co-digestion of agriculture and industrial wastes. Microbial community diversity. *Biomass and Bioenergy* 2014; 71:84–97.
96. Kikot P, Viera M, Mignone C, Donati E. Study of the effect of pH and dissolved heavy metals on the growth of sulfate-reducing bacteria by a fractional factorial design. *Hydrometallurgy* 2010; 104:494–500.
97. Chen Y, Lin C-J, Jones G, Fu S, Zhan H. Enhancing biodegradation of wastewater by microbial consortia with fractional factorial design. *J Hazard Mater* 2009; 171:948–953.
98. Chen Y, Lin C-J, Jones G, Fu S, Zhan H. Application of statistical design for the optimization of microbial community of synthetic domestic wastewater. *Biodegradation* 2011; 22:205–213.
99. Hu J, Wang L, Zhang S, Le Y, Fu X. Feasibility of a two-step culture method to improve the CO_2-fixing efficiency of nonphotosynthetic microbial community and simultaneously decrease the spontaneous oxidative precipitates from mixed electron donors. *Appl*

Biochem Biotechnol 2014; 173:2307–2320.

100. Hu J, Xue Y, Li J, Wang L, Zhang S, *et al.* Characterization of a designed synthetic autotrophic–heterotrophic consortia for fixing CO_2 without light. *RSC Adv* 2016; 6:78161–78169.
101. Poszytek K, Ciezkowska M, Sklodowska A, Drewniak L. Microbial consortium with high cellulolytic activity (MCHCA) for enhanced biogas production. *Front Microbiol* 2016; 7:324.
102. Hu J, Xue Y, Guo H, Gao M, Li J, *et al.* Design and composition of synthetic fungal-bacterial microbial consortia that improve lignocellulolytic enzyme activity. *Bioresour Technol* 2017; 227:247–255.
103. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2018; 46:D8–D13.
104. Chen I-MA, Markowitz VM, Chu K, Palaniappan K, Szeto E, *et al.* IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res* 2017; 45:D507–D516.
105. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 2015; 44:D457–D462.
106. Caspi R, Billington R, Fulcher CA, Keseler IM, Kothari A, *et al.* The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res* 2018; 46:D633–D639.
107. Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, *et al.* Creation and analysis of biochemical constraint-based models: the COBRA Toolbox v3.0. *arXiv Prepr* 2017; :arXiv:1710.04038.
108. Magnúsdóttir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, *et al.* Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat Biotechnol* 2016; 35:81–89.
109. Kauffman KJ, Prakash P, Edwards JS. Advances in flux balance analysis. *Curr Opin Biotechnol* 2003; 14:491–496.
110. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? *Nat Biotechnol* 2010; 28:245–8.
111. Marlow JJ, Steele JA, Ziebis W, Thurber AR, Levin LA, *et al.* Carbonate-hosted methanotrophy represents an unrecognized methane sink in the deep sea. *Nat Commun*

2014; 5:5094.

112. Sekirov I, Russell SL, Antunes LCM, Finlay BB. Gut microbiota in health and disease. *Physiol Rev* 2010; 90:859–904.
113. Aroniadis OC, Brandt LJ. Fecal microbiota transplantation: past, present and future. *Curr Opin Gastroenterol* 2013; 29:79–84.
114. Hamilton MJ, Weingarden AR, Unno T, Khoruts A, Sadowsky MJ. High-throughput DNA sequence analysis reveals stable engraftment of gut microbiota following transplantation of previously frozen fecal bacteria. *Gut Microbes* 2013; 4:125–35.
115. Song Y, Garg S, Girotra M, Maddox C, von Rosenvinge EC, *et al.* Microbiota dynamics in patients treated with fecal microbiota transplantation for recurrent *Clostridium difficile* infection. *PLoS One* 2013; 8:e81330.
116. Pelz O, Tesar M, Wittich R-M, Moore ERB, Timmis KN, *et al.* Towards elucidation of microbial community metabolic pathways: unravelling the network of carbon sharing in a pollutant-degrading bacterial consortium by immunocapture and isotopic ratio mass spectrometry. *Environ Microbiol* 1999; 1:167–174.
117. Pettit RK. Mixed fermentation for natural product drug discovery. *Appl Microbiol Biotechnol* 2009; 83:19–25.
118. Chiu H-C, Levy R, Borenstein E. Emergent biosynthetic capacity in simple microbial communities. *PLoS Comput Biol* 2014; 10:e1003695.
119. Moran NA. Symbiosis as an adaptive process and source of phenotypic complexity. *Proc Natl Acad Sci U S A* 2007; 104 Suppl:8627–33.
120. Gordon JI, Klaenhammer TR. A rendezvous with our microbes. *Proc Natl Acad Sci U S A* 2011; 108 Suppl:4513–5.
121. Raymond J, Segrè D. The effect of oxygen on biochemical networks and the evolution of complex life. *Science* 2006; 311:1764–7.
122. Song H-S, Cannon W, Beliaev A, Konopka A. Mathematical modeling of microbial community dynamics: a methodological review. *Processes* 2014; 2:711–752.
123. Taffs R, Aston JE, Brileya K, Jay Z, Klatt CG, *et al. In silico* approaches to study mass and energy flows in microbial consortia: a syntrophic case study. *BMC Syst Biol* 2009; 3:114.
124. Nemhauser G, Wolsey L. Integer programming and combinatorial optimization. In: *IPCO:*

International Conference on Integer Programming and Combinatorial Optimization.New York, New York, USA: Springer Berlin Heidelberg; 1988. pp. 8–20.

125. Ye Y, Doak TG. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput Biol* 2009; 5:e1000465.
126. Levy R, Borenstein E. Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. *Proc Natl Acad Sci U S A* 2013; 110:12804–9.
127. Parter M, Kashtan N, Alon U. Environmental variability and modularity of bacterial metabolic networks. *BMC Evol Biol* 2007; 7:169.
128. Lougee-Heimer R. The Common Optimization INterface for Operations Research: promoting open-source software in the operations research community. *IBM J Res Dev* 2003; 47:57–66.
129. Methé BA, Nelson KE, Pop M, Creasy HH, Giglio MG, *et al.* A framework for human microbiome research. *Nature* 2012; 486:215–21.
130. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* 2013; 31:814–21.
131. Lozupone CA, Knight R. Global patterns in bacterial diversity. *Proc Natl Acad Sci U S A* 2007; 104:11436–40.
132. Escalante AE, Rebolleda-Gómez M, Benítez M, Travisano M. Ecological perspectives on synthetic biology: insights from microbial population biology. *Front Microbiol* 2015; 6:143.
133. Johnson DR, Goldschmidt F, Lilja EE, Ackermann M. Metabolic specialization and the assembly of microbial communities. *ISME J* 2012; 6:1985–91.
134. Zhou J, Xia B, Treves DS, Wu L-Y, Marsh TL, *et al.* Spatial and resource factors influencing high microbial diversity in soil. *Appl Environ Microbiol* 2002; 68:326–334.
135. Doebeli M, Ispolatov I. Complexity and diversity. *Science* 2010; 328:494–7.
136. Ley RE, Peterson DA, Gordon JI. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* 2006; 124:837–48.
137. Nemergut DR, Schmidt SK, Fukami T, O'Neill SP, Bilinski TM, *et al.* Patterns and processes of microbial community assembly. *Microbiol Mol Biol Rev* 2013; 77:342–56.

138. Hansen SK, Rainey PB, Haagensen JAJ, Molin S. Evolution of species interactions in a biofilm community. *Nature* 2007; 445:533–6.

139. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, *et al.* Microbial co-occurrence relationships in the human microbiome. *PLoS Comput Biol* 2012; 8:e1002606.

140. Harcombe WR, Riehl WJ, Dukovski I, Granger BR, Betts A, *et al.* Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Rep* 2014; 7:1104–15.

141. Zhuang K, Izallalen M, Mouser P, Richter H, Risso C, *et al.* Genome-scale dynamic modeling of the competition between *Rhodoferax* and *Geobacter* in anoxic subsurface environments. *ISME J* 2011; 5:305–16.

142. Zomorrodi AR, Islam MM, Maranas CD. d-OptCom: dynamic multi-level and multi-objective metabolic modeling of microbial communities. *ACS Synth Biol* 2014; 3:247–57.

143. Hamady M, Knight R. Microbial community profiling for human microbiome projects: tools, techniques, and challenges. *Genome Res* 2009; 19:1141–52.

144. Giloteaux L, Goodrich JK, Walters WA, Levine SM, Ley RE, *et al.* Reduced diversity and altered composition of the gut microbiome in individuals with myalgic encephalomyelitis/chronic fatigue syndrome. *Microbiome* 2016; 4:30.

145. Rehman A, Rausch P, Wang J, Skieceviciene J, Kiudelis G, *et al.* Geographical patterns of the standing and active human gut microbiome in health and IBD. *Gut* 2016; 65:238–48.

146. Stewart CJ, Embleton ND, Marrs ECL, Smith DP, Nelson A, *et al.* Temporal bacterial and metabolic development of the preterm gut reveals specific signatures in health and disease. *Microbiome* 2016; 4:67.

147. Ley RE, Bäckhed F, Turnbaugh P, Lozupone CA, Knight RD, *et al.* Obesity alters gut microbial ecology. *Proc Natl Acad Sci U S A* 2005; 102:11070–5.

148. Murphy EF, Cotter PD, Hogan A, O'Sullivan O, Joyce A, *et al.* Divergent metabolic outcomes arising from targeted manipulation of the gut microbiota in diet-induced obesity. *Gut* 2013; 62:220–6.

149. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 2010; 464:59–65.

150. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, *et al.* Structure and function of the global ocean microbiome. *Science* 2015; 348:1261359.

151. Bengtsson-Palme J, Hammarén R, Pal C, Östman M, Björlenius B, *et al.* Elucidating selection processes for antibiotic resistance in sewage treatment plants using metagenomics. *Sci Total Environ* 2016; 572:697–712.
152. Thompson LR, Williams GJ, Haroon MF, Shibl A, Larsen P, *et al.* Metagenomic covariation along densely sampled environmental gradients in the Red Sea. *ISME J* 2017; 11:138–151.
153. Carr R, Shen-Orr SS, Borenstein E. Reconstructing the genomic content of microbiome taxa through shotgun metagenomic deconvolution. *PLoS Comput Biol* 2013; 9:e1003292.
154. Vieira-Silva S, Falony G, Darzi Y, Lima-Mendez G, Yunta RG, *et al.* Species–function relationships shape ecological properties of the human gut microbiome. *Nat Microbiol* 2016; 1:16088.
155. Aßhauer KP, Wemheuer B, Daniel R, Meinicke P. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics* 2015; 31:2882–4.
156. Manor O, Borenstein E. Systematic characterization and analysis of the taxonomic drivers of functional shifts in the human microbiome. *Cell Host Microbe* 2017; 21:254–267.
157. Noecker C, Eng A, Srinivasan S, Theriot CM, Young VB, *et al.* Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. *mSystems* 2016; 1:e00013-15.
158. Orr HA. The genetic theory of adaptation: a brief history. *Nat Rev Genet* 2005; 6:119–127.
159. Poelwijk FJ, Kiviet DJ, Weinreich DM, Tans SJ. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* 2007; 445:383–386.
160. Romero PA, Arnold FH. Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol* 2009; 10:866–876.
161. Hartl DL. What can we learn from fitness landscapes. *Curr Opin Microbiol* 2014; 21:51–57.
162. Moya A, Ferrer M. Functional redundancy-induced stability of gut microbiota subjected to disturbance. *Trends Microbiol* 2016; 24:402–413.
163. Naeem S, Kawabata Z, Loreau M. Transcending boundaries in biodiversity research. *Trends Ecol Evol* 1998; 13:134–135.
164. Little AEF, Robinson CJ, Peterson SB, Raffa KF, Handelsman J. Rules of engagement:

interspecies interactions that regulate microbial communities. *Annu Rev Microbiol* 2008; 62:375–401.

165. Borenstein E, Ruppin E. Direct evolution of genetic robustness in microRNA. *Proc Natl Acad Sci U S A* 2006; 103:6593–8.
166. Wilke CO, Wang JL, Ofria C, Lenski RE, Adami C. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature* 2001; 412:331–333.
167. Kim P-J, Price ND. Genetic co-occurrence network across sequenced microbes. *PLoS Comput Biol* 2011; 7:e1002340.
168. David LA, Materna AC, Friedman J, Campos-Baptista MI, Blackburn MC, *et al.* Host lifestyle affects human microbiota on daily timescales. *Genome Biol* 2014; 15:R89.
169. Gilbert JA, Jansson JK, Knight R. The Earth Microbiome project: successes and aspirations. *BMC Biol* 2014; 12:69.
170. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010; 7:335–336.
171. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 2006; 72:5069–72.
172. Manor O, Borenstein E. MUSiCC: a marker genes based framework for metagenomic normalization and accurate profiling of gene abundances in the microbiome. *Genome Biol* 2015; 16:53.
173. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res* 2007; 36:D480–D484.
174. Lozupone CA, Hamady M, Kelley ST, Knight R. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Appl Environ Microbiol* 2007; 73:1576–85.
175. Das A, Srinivasan M, Ghosh TS, Mande SS. Xenobiotic metabolism and gut microbiomes. *PLoS One* 2016; 11:e0163099.
176. Hodges JL, Jr., Lehmann EL. Estimates of location based on rank tests. *Ann Math Stat* 1963; 34:598–611.
177. Sverrisdóttir E, Lund TM, Olesen AE, Drewes AM, Christrup LL, *et al.* A review of morphine and morphine-6-glucuronide's pharmacokinetic–pharmacodynamic

relationships in experimental and clinical pain. *Eur J Pharm Sci* 2015; 74:45–62.

178. Gumbo T, Angulo-Barturen I, Ferrer-Bazaga S. Pharmacokinetic-pharmacodynamic and dose-response relationships of antituberculosis drugs: recommendations and standards for industry and academia. *J Infect Dis* 2015; 211:S96–S106.
179. Chang S, Zhuang D, Guo W, Li L, Zhang W, *et al.* The antiviral activity of approved and novel drugs against HIV-1 mutations evaluated under the consideration of dose-response curve slope. *PLoS One* 2016; 11:e0149467.
180. Novák B, Tyson JJ. Design principles of biochemical oscillators. *Nat Rev Mol Cell Biol* 2008; 9:981–991.
181. Queitsch C, Sangster TA, Lindquist S. Hsp90 as a capacitor of phenotypic variation. *Nature* 2002; 417:618–624.
182. Shilo B-Z, Barkai N. Buffering global variability of morphogen gradients. *Dev Cell* 2017; 40:429–438.
183. Smith MI, Yatsunenko T, Manary MJ, Trehan I, Mkakosya R, *et al.* Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. *Science* 2013; 339:548–54.
184. Raad I, Darouiche R, Hachem R, Sacilowski M, Bodey GP. Antibiotics and prevention of microbial colonization of catheters. *Antimicrob Agents Chemother* 1995; 39:2397–400.
185. Chng KR, Tay ASL, Li C, Ng AHQ, Wang J, *et al.* Whole metagenome profiling reveals skin microbiome-dependent susceptibility to atopic dermatitis flare. *Nat Microbiol* 2016; 1:16106.
186. Edlund A, Yang Y, Yooseph S, Hall AP, Nguyen DD, *et al.* Meta-omics uncover temporal regulation of pathways across oral microbiome genera during in vitro sugar metabolism. *ISME J* 2015; 9:2605–2619.
187. Gamfeldt L, Hillebrand H, Jonsson PR. Multiple functions increase the importance of biodiversity for overall ecosystem functioning. *Ecology* 2008; 89:1223–1231.
188. Peter H, Ylla I, Gudasz C, Romaní AM, Sabater S, *et al.* Multifunctionality and diversity in bacterial biofilms. *PLoS One* 2011; 6:e23225.
189. Gower JC. Generalized procrustes analysis. *Psychometrika* 1975; 40:33–51.
190. Srinivasan S, Morgan MT, Fiedler TL, Djukovic D, Hoffman NG, *et al.* Metabolic signatures of bacterial vaginosis. *MBio* 2015; 6:e00204-15.
191. Oh J, Byrd AL, Park M, NISC Comparative Sequencing Program, Kong HH, *et al.*

Temporal stability of the human skin microbiome. *Cell* 2016; 165:854–866.

192. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 2014; 505:559–563.
193. Kong HH, Oh J, Deming C, Conlan S, Grice EA, *et al.* Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. *Genome Res* 2012; 22:850–9.
194. Aagaard K, Riehle K, Ma J, Segata N, Mistretta T-A, *et al.* A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. *PLoS One* 2012; 7:e36466.
195. Wilhelm SW, LeCleir GR, Bullerjahn GS, McKay RM, Saxton MA, *et al.* Seasonal changes in microbial community structure and activity imply winter production is linked to summer hypoxia in a large lake. *FEMS Microbiol Ecol* 2014; 87:475–485.
196. Smits SA, Leach J, Sonnenburg ED, Gonzalez CG, Lichtman JS, *et al.* Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. *Science* 2017; 357:802–806.
197. Greenblum S, Carr R, Borenstein E. Extensive strain-level copy-number variation across human gut microbiome species. *Cell* 2015; 160:583–94.
198. Julien-Laferrière A, Bulteau L, Parrot D, Marchetti-Spaccamela A, Stougie L, *et al.* A combinatorial algorithm for microbial consortia synthetic design. *Sci Rep* 2016; 6:29182.
199. García-Jiménez B, García JL, Nogales J. FLYCOP: metabolic modeling-based analysis and engineering microbial communities. *Bioinformatics* 2018; 34:i954–i963.
200. Bucci V, Tzen B, Li N, Simmons M, Tanoue T, *et al.* MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. *Genome Biol* 2016; 17:121.
201. Bucci V, Xavier JB. Towards predictive models of the human gut microbiome. *J Mol Biol* 2014; 426:3907–3916.
202. Stein RR, Tanoue T, Szabady RL, Bhattarai SK, Olle B, *et al.* Computer-guided design of optimal microbial consortia for immune system modulation. *Elife* 2018; 7:e30916.
203. Verster AJ, Ross BD, Radey MC, Bao Y, Goodman AL, *et al.* The landscape of type VI Secretion across human gut microbiomes reveals its role in community composition. *Cell Host Microbe* 2017; 22:411–419.e4.

204. Momeni B, Xie L, Shou W. Lotka-Volterra pairwise modeling fails to capture diverse pairwise microbial interactions. *Elife* 2017; 6:e25051.
205. Werner EE, Peacor SD. A review of trait-mediated indirect interactions in ecological communities. *Ecology* 2003; 84:1083–1100.
206. Gould AL, Zhang V, Lamberti L, Jones EW, Obadia B, *et al.* High-dimensional microbiome interactions shape host fitness. *bioRxiv* 2018; :232959.
207. Sanchez-Gorostiaga A, Bajić D, Osborne ML, Poyatos JF, Sanchez A. High-order interactions dominate the functional landscape of microbial consortia. *bioRxiv* 2018; :333534.
208. Miller MB, Bassler BL. Quorum sensing in bacteria. *Annu Rev Microbiol* 2001; 55:165–199.
209. Zhao K, Li W, Li J, Ma T, Wang K, *et al.* TesG is a type I secretion effector of *Pseudomonas aeruginosa* that suppresses the host immune response during chronic infection. *Nat Microbiol* 2019; :1.
210. Mathur A, Feng S, Hayward JA, Ngo C, Fox D, *et al.* A multicomponent toxin from *Bacillus cereus* incites inflammation and shapes host outcome via the NLRP3 inflammasome. *Nat Microbiol* 2019; 4:362–374.
211. Cameron EA, Curtis MM, Kumar A, Dunny GM, Sperandio V. Microbiota and pathogen proteases modulate type III secretion activity in enterohemorrhagic *Escherichia coli*. *MBio* 2018; 9:e02204-18.
212. Russell AB, Peterson SB, Mougous JD. Type VI secretion system effectors: poisons with a purpose. *Nat Rev Microbiol* 2014; 12:137–148.
213. Schaefer TJ, J. T. The complexity of satisfiability problems. In: *Proceedings of the tenth annual ACM symposium on Theory of computing*.New York, New York, USA: ACM; 1978. pp. 216–226.
214. Liaw A, Wiener M. Classification and regression by randomForest. *R News* 2002; 2:18–22.

6. Appendix A: Supplementary material for Chapter 2

6.1 Supplementary Text

6.1.1 Forcing substrate usage

Notably, the ILP formulation described in section 2.3 Methods provides a solution that can generate all target metabolites from available substrates, but does not require that all available substrates will be used. Forcing specific available substrate to be utilized in target product synthesis could be desired, for example, if one were to design a community to consume specific environmental metabolites, such as for pollutant or toxin conversion. Below, I therefore extend CoMiDA, allowing users to define a specific subset of the available substrates that the designed community *must* utilize in some metabolic path that generates the target products.

To this end, I define a new set of metabolites,

$$FORCED \subseteq SUBSTRATE,$$

that contains all the available substrates that the community must be able to use in the production of the target product metabolites. Given this set, I define the following constraint:

$$\sum_{\substack{\forall j \; s.t. \\ i_j=(m_k,r_l)}} \left(F\_i_j\right) \geq 1 : \forall m_k \in FORCED,$$

which requires that there must be some amount of flow out of all metabolite nodes in $FORCED$ and so the designed community must contain metabolic reactions capable of utilizing those metabolites. As mentioned in 2.3 Methods, forced substrate usage introduces a scenario that may require flow to enter substrate metabolite nodes. In particular, normally if a viable path exists from a substrate to a product, there would never be a need for flow to enter that substrate node since the substrate can provide any amount of flow. Additionally, any path leading from one substrate to a product that passes through another substrate could be shortened by starting at the intermediate substrate. Forced substrates, however, must provide flow and thus may require paths to pass through other substrates.

6.1.2 **Prioritizing or avoiding specific species**

Another consideration I have incorporated into CoMiDA is the desirability of particular available species. As an example usage of species desirability, one might want to factor in the availability or ease of culturing various species when designing a community such that readily available species are preferred to those that might be more difficult to obtain. Such desirability can be encoded as a cost for each species in the objective function using an ILP cost variable, I_c, for each species that represents the cost for including a species in the solution species set. These cost variables can then be included in the optimization function:

$$\min \sum_{\substack{\forall i \ s.t. \\ s_i \in S}} (I\_c_i \times s_i).$$

This encoding would use higher costs for less desirable species and low costs for more desirable species. Thus, when minimizing the sum of species variables, more desirable species would have a smaller impact the sum of species variables than less desirable species.

6.2 Supplementary Figures

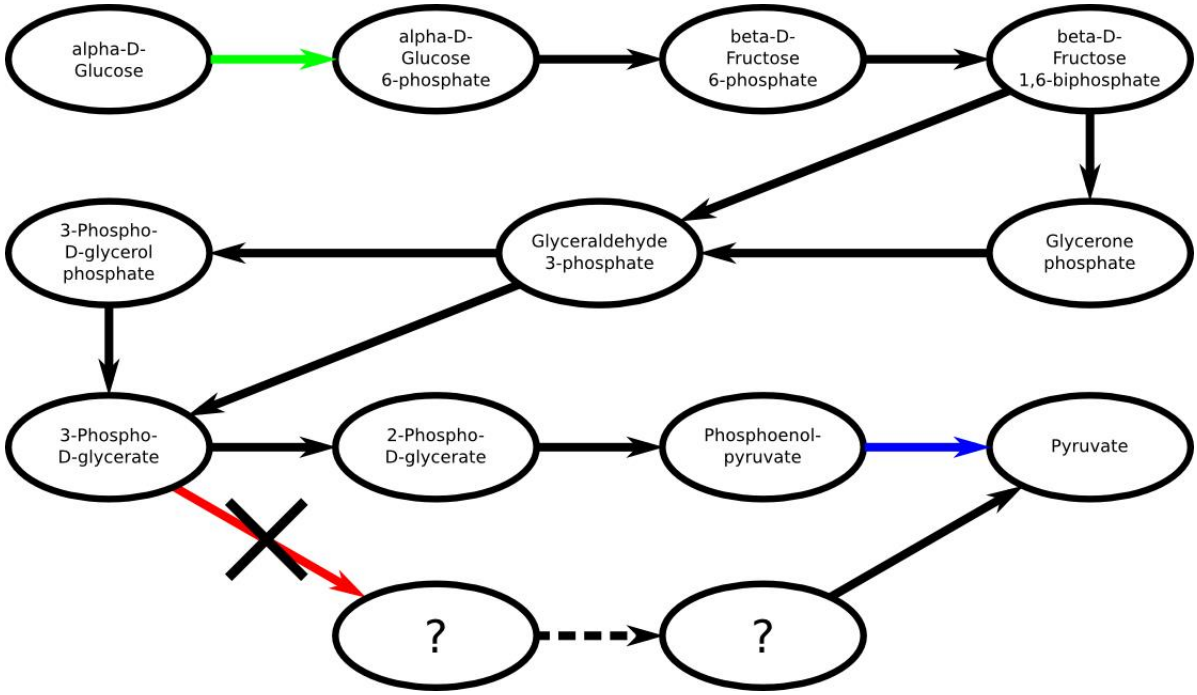

Figure 6.1 Schematic of the Embden-Meyerhof glycolysis pathway used for validation. I modified the set of available metabolic reactions to control the minimal solution sizes during validation. First, I removed alternate paths to the target product (pyruvate) by removing the first reaction in each alternate path (red). Next, to adjust the minimal solution sizes, I deleted metabolic reactions from subsets of the available species. For example, to force a two species solution, I removed one metabolic reaction (green) from half of the available species and a different metabolic reaction (blue) from the other half of the available species.

7. Appendix B: Supplementary material for Chapter 3

7.1 Supplementary Text

7.1.1 Taxa-function response curve model evaluation suggests a power function relationship

I considered eight different models describing the relationship between a taxonomic perturbation magnitude, t, and the associated degree of functional shift, f. These models were constructed by first considering the expected increase in functional shift as taxonomic perturbation magnitudes increase, which suggests a potential linear component and is represented by the a parameter in each model. The second consideration for each model was whether there would be a polynomial or generally non-linear component to the relationship. For this second component, I tested: the absence of a second component (model 1), a quadratic relationship (model 2), a more general power function relationship (model 3), a cosine transformation to match the measure of functional shift (model 4), a combination of linear and quadratic relationships (model 5), an exponential relationship (model 6), a combination of linear and exponential relationships (model 7), and an exponential relationship with a proportionality constant (model 8), using the variable b to parameterize this second component when necessary (Table 7.1).

To evaluate these models, for every community I fit a response curve using each model and calculated the error for every perturbation of that community. I then aggregated communities by environment, by subsite, and pooled together across environments and determined the root mean squared error (RMSE) of each aggregate. Across all aggregates, model 3 had the lowest RMSE and model 2 consistently had the second lowest RMSE (Figure 7.4). Given that model 3 was a generalization of model 2, I decided to use model 3 to examine whether the variation in b was associated with any biological factors.

7.1.2 Predictive models support the association between robustness and functional redundancy

In addition to correlation analysis and PCA, I further examined associations between GDFs and robustness using predictive models. Specifically, I attempted to predict a community's

attenuation using that community's GDFs and determined GDF associations with robustness based on the importance of each GDF in the model (i.e., a GDF being more informative suggests a stronger association). In this analysis, I tested both the 5 GDFs used in the main results as well as an expanded set of 45 GDFs (Table 7.2). The additional 40 GDFs aimed to capture other aspects of gene distribution such as how many species encode unique functions, how many species encode each function, and the similarity in function co-occurrence across genomes. I evaluated models by using 70% of the available communities for model fitting and then measuring the correlation between observed and predicted attenuation, as well as the root mean squared error in attenuation prediction, with the remaining 30%. I also accounted for potential variation in performance due to the specific communities in model fitting and evaluation sets by repeating this process for 500 different paired sets of model fitting and evaluation communities, taking the average of these evaluation metrics across all 500 models. To ensure comparability between GDFs, GDFs were centered and scaled before model construction.

I found that predictions were significantly improved when I used the expanded GDF set to fit unregularized linear models to the pooled set of communities (Figure 7.6; Figure 7.7; $p < 10^{-101}$; F-test) and all following analyses are based on models fit using the expanded GDF set. Unregularized linear model predictions on evaluation communities were somewhat correlated with observed community attenuation (Figure 7.6A; $r = 0.63$; RMSE = 0.54), indicating that at least a subset of these 45 GDFs were associated with robustness. When I attempted to regularize the linear models, I found no noticeable improvement in performance (Figure 7.6B; $r = 0.62$; RMSE = 0.54). I suspected that the relationship between GDFs and robustness might not necessarily be linear, so I also attempted to fit another model, a Regularized Random Forest [214], using 5000 trees for model construction. The Regularized Random Forest did have better predictive power than the linear models (Figure 7.6C; $r = 0.7$; RMSE = 0.49), suggesting that some GDFs may contribute to robustness in a nonlinear fashion. I also considered the possibility that GDF associations with robustness might vary by environment, so I fit Regularized Random Forests for each environment individually. This revealed noticeable variation in model performance across environments, with the best correlation between observed and predicted attenuation in water-associated communities (Figure 7.6D; $r = 0.78$) and the lowest prediction error for gut communities (RMSE = 0.33).

To examine robustness associations with specific GDFs, I next examined the importance of each GDF in each model. For linear models, feature importance was measured as the absolute

value of the feature's coefficient in the model, while Regularized Random Forests feature importance was measured as the mean decrease in accuracy when the feature's values are permuted. Functional redundancy appeared as the most important feature in the Regularized Random Forest, further supporting its strong association with robustness, while the most important features in the linear model were related to function co-occurrence across genomes.

7.2 Supplementary Figures

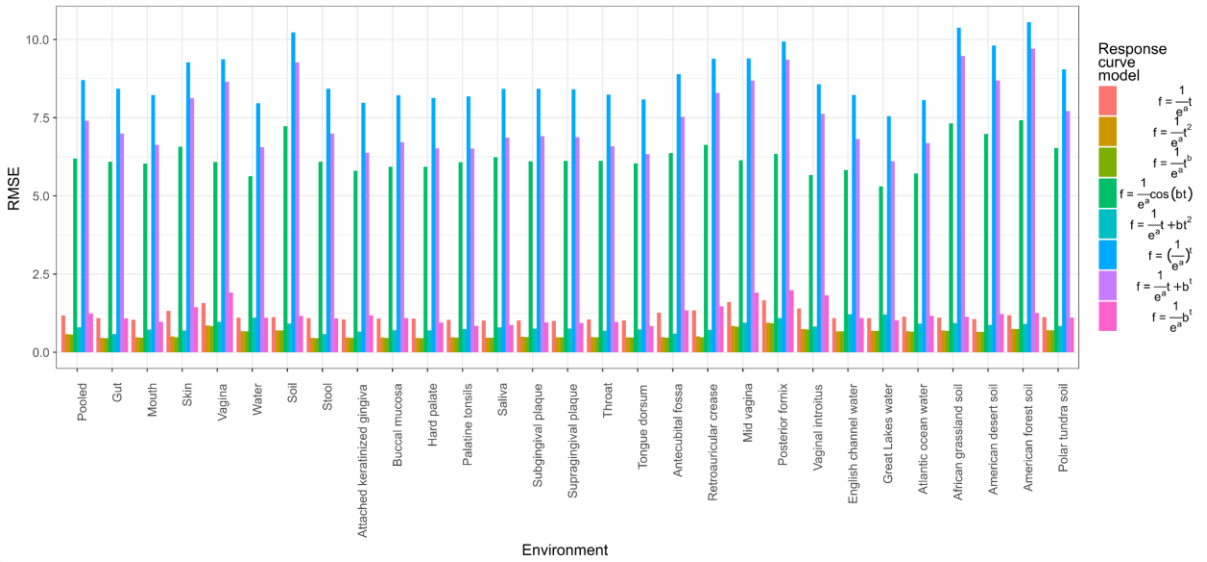

Figure 7.1. Candidate taxa-function response curve model fits. Bar height indicates the root mean squared error (RMSE) between the observed functional shifts and the taxa-function response curves across all communities in the group labeled on the x-axis.

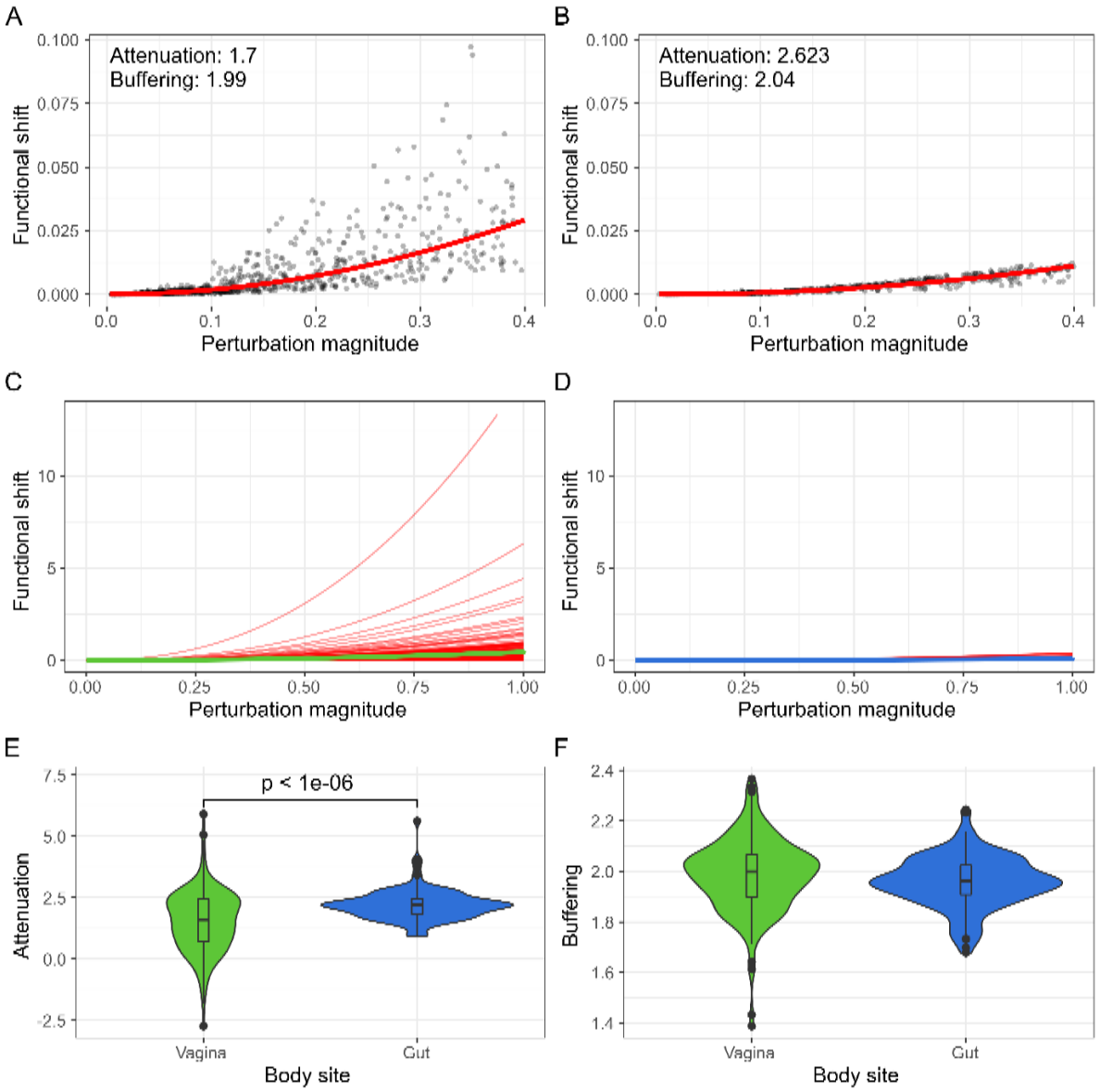

Figure 7.2. Taxonomic perturbations, their corresponding functional shifts, and the response curves fit in gut and vaginal sub-sampled communities. All panels are the same as in Figure 3.2 except that communities were subsampled to 10 OTUs as described in 3.3 Methods.

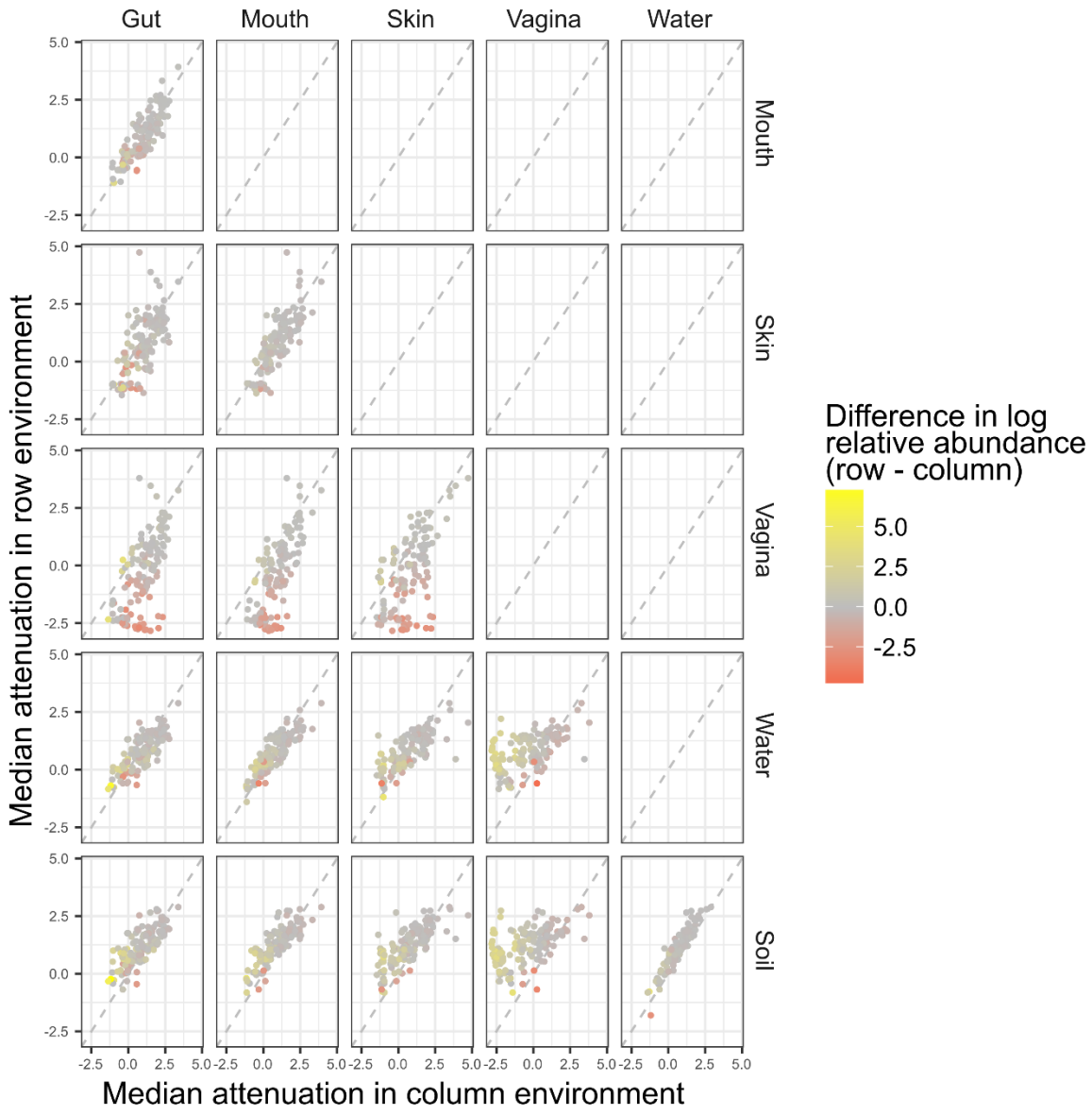

Figure 7.3. Comparison of pathway-specific attenuation trends between environments. Each column and row represents a different environment where each panel shows a comparison of the median attenuation of all pathways present in both environments. Position on the y-axis indicates median attenuation in the row environment while position on the x-axis indicates median attenuation in the column environment. The dotted line indicates the 1:1 median attenuation relationship. Color of each point is determined by the difference in log median abundance of the pathway between the two environments.

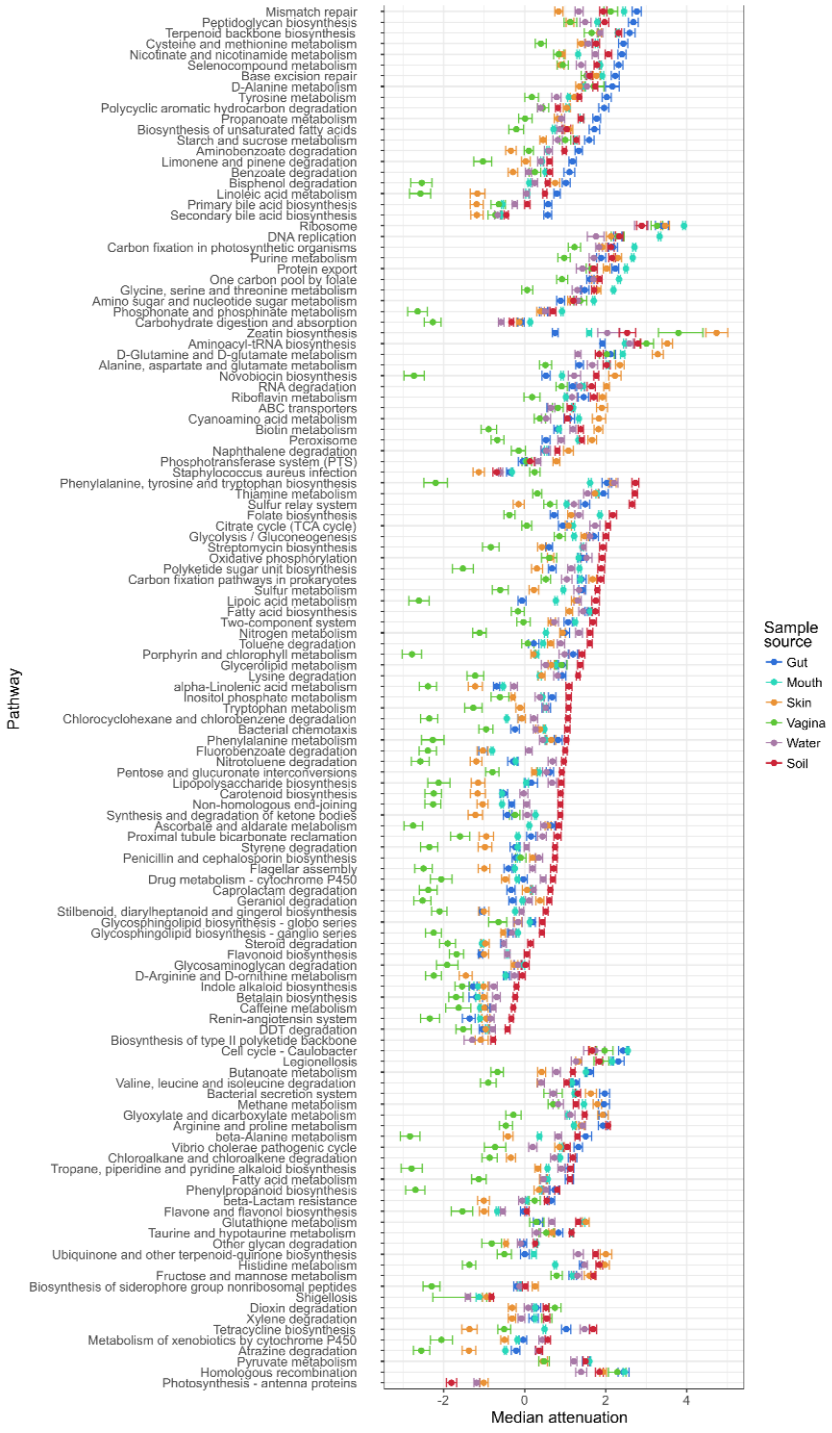

Figure 7.4. Pathway-specific attenuation of all pathways by environment. Each point represents the median attenuation of a pathway in the environment indicated by the color. Error bars show the 95% confidence interval of the median. Pathways are grouped vertically by the environment(s) in which the pathway is (are) most robust as determined by Mood's median test (3.3 Methods) and then by highest median attenuation within those groupings.

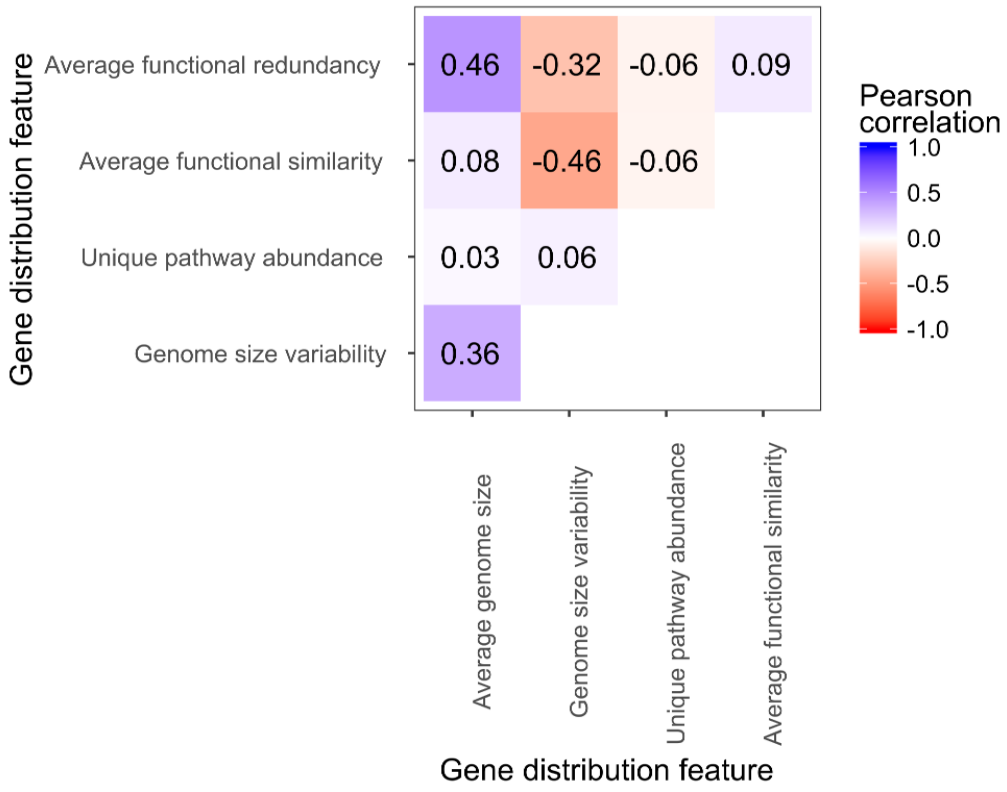

Figure 7.5. Correlations between gene distribution features. The heatmap shows the strength of correlation between the five GDFs used in the correlation analysis and PCA.

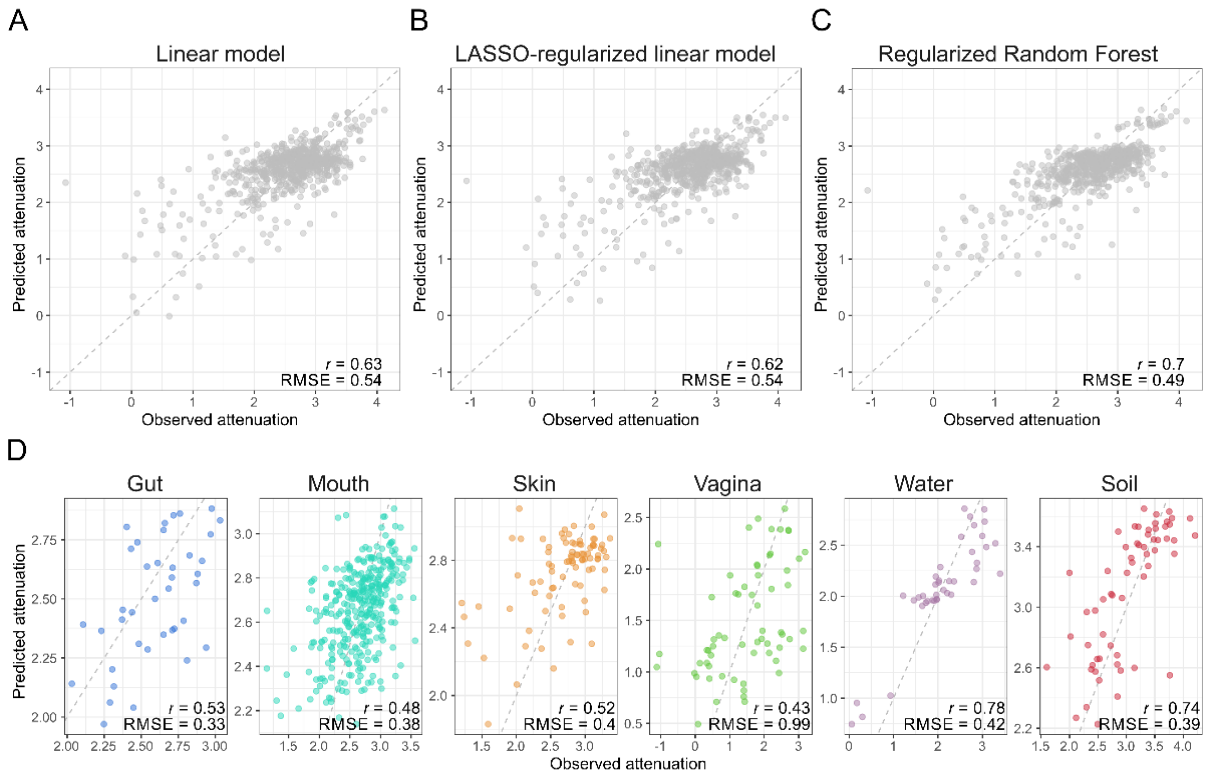

Figure 7.6. Performance of GDF-based predictive models of attenuation, using an extended set of 45 GDFs. **(A)** Points indicate the predicted vs. observed attenuation of evaluation communities for the linear model (left) and Regularized Random Forest (right) from one constructed model for each method. The grey lines show the 1:1 relationship. The average correlation between observed and predicted attenuation and the average root mean squared error (RMSE) across 500 models constructed on different training and evaluation partitions are displayed in the lower right corners of both panels. **(B)** Similar to **(A)** except each panel shows the predicted vs. observed attenuation for Regularized Random Forests constructed using communities from a single environment.

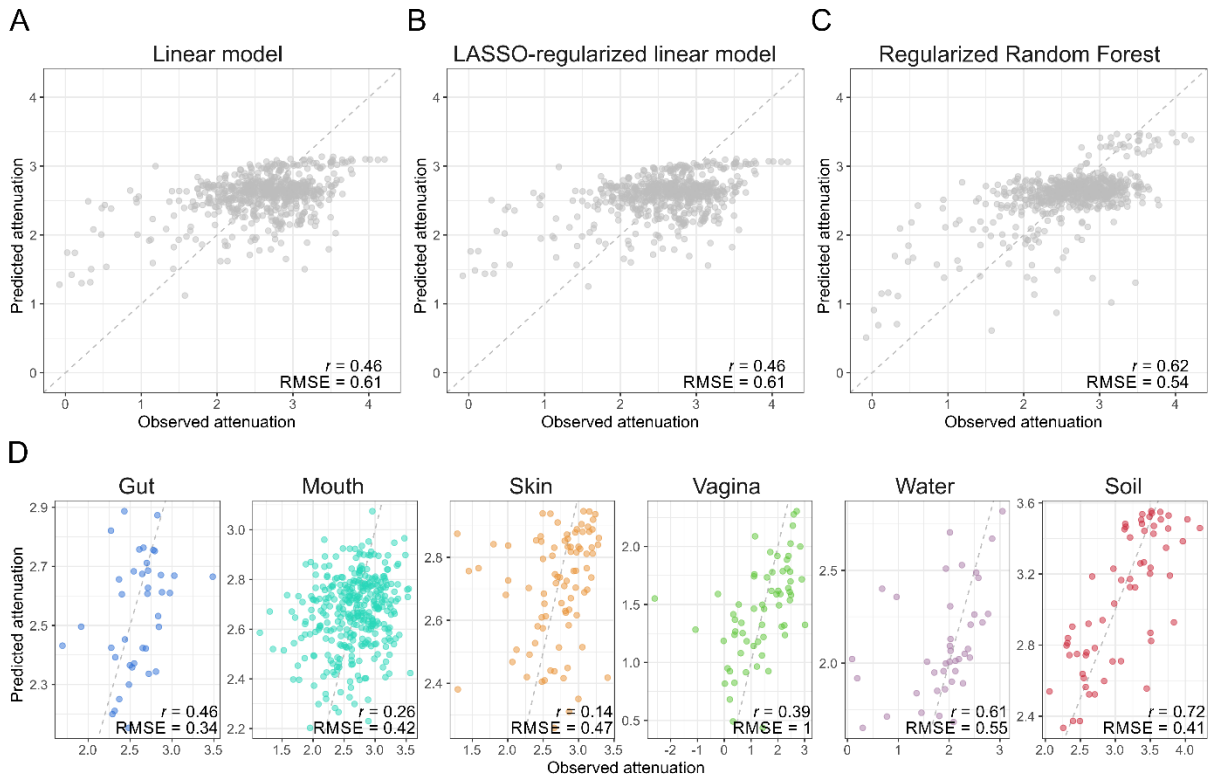

Figure 7.7. Performance of GDF-based predictive models of attenuation with only 5 GDFs. Similar to Figure 7.6 except that models were fit using only the 5 GDFs described in the correlation analysis and PCA.

7.3 Supplementary Tables

Table 7.1. Candidate taxa-function response curve function definitions.

Model	Definition
Model 1	$f = \frac{1}{e^a} t$
Model 2	$f = \frac{1}{e^a} t^2$
Model 3	$f = \frac{1}{e^a} t^b$
Model 4	$f = \frac{1}{e^a} \cos(bt)$
Model 5	$f = \frac{1}{e^a} t + bt^2$
Model 6	$f = \left(\frac{1}{e^a}\right)^t$
Model 7	$f = \frac{1}{e^a} t + b^t$
Model 8	$f = \frac{1}{e^a} b^t$

Table 7.2. Extended definitions of variables used in gene distribution features.

Variable description	Variable definition
Number of taxa present in the sample	$N \in \mathbb{N}$
Number of functions present in the sample	$M \in \mathbb{N}$
Taxon index	$i \in \{1,2,\dots,N\}$
Function index	$j \in \{1,2,\dots,M\}$
Relative abundance of taxon i	$t_i \in [0,1]; \sum_{i=1}^{N} t_i = 1$
Relative abundance of function j	$f_j \in [0,1]; \sum_{j=1}^{M} f_j = 1$
Copy number of function j in taxon i	$c_{ij} \in [0,\infty)$
Redundancy of function j	$r_j = -\sum_{i=1}^{N} [t_i c_{ij} \ln[t_i c_{ij}]]$
Average functional redundancy	$\bar{r} = \frac{\sum_{j=1}^{M} [f_j \times r_j]}{\sum_{j=1}^{M} f_j}$
Variation in functional redundancy	$\frac{\sqrt{\sum_{j=1}^{M} \left[f_j^2 \times [r_j - \bar{r}]^2\right]}}{\bar{r}}$
Genome size of taxon i	$\sum_{j=1}^{M} c_{ij}$
Uniqueness of function j	$UF_j = \begin{cases} 1 \text{ if function } j \text{ is encoded} \\ \text{by a single taxon} \\ 0 \text{ otherwise} \end{cases}$
Unique pathway abundance	$\sum_{j=1}^{M} [f_j \times UF_j]$
Euclidean distance between genomic content of taxon i_1 and taxon i_2	$\sqrt{\sum_{j=1}^{M} [c_{i_1 j} - c_{i_2 j}]^2}$
Jaccard dissimilarity between genomic content of taxon i_1 and taxon i_2	$1 - \frac{\sum_{j=1}^{M} \min(c_{i_1 j}, c_{i_2 j})}{\sum_{j=1}^{M} \max(c_{i_1 j}, c_{i_2 j})}$
Bray-Curtis dissimilarity between genomic content of taxon i_1 and taxon i_2	$\frac{\sum_{j=1}^{M} \lvert c_{i_1 j} - c_{i_2 j} \rvert}{\sum_{j=1}^{M} [c_{i_1 j} + c_{i_2 j}]}$
Manhattan distance between genomic content of taxon i_1 and taxon i_2	$\sum_{j=1}^{M} \lvert c_{i_1 j} - c_{i_2 j} \rvert$

Number of different functions encoded by taxon i	$\sum_{j=1}^{M} \mathbf{1}_{\{c_{ij}>0\}}$
Presence of unique functions encoded by taxon i	$UT_i = \begin{cases} 1 \text{ if at least one function encoded by taxon } i \text{ is only encoded by taxon } i \\ 0 \text{ otherwise} \end{cases}$
Number of species with unique functions	$\sum_{i=1}^{N} UT_i$
Abundance of species with unique functions	$\sum_{i=1}^{N} [t_i \times UT_i]$
Number of different taxa encoding function j	$\sum_{i=1}^{N} \mathbf{1}_{\{c_{ij}>0\}}$
Total copy number of function j	$\sum_{i=1}^{N} c_{ij}$
Total abundance of taxa encoding function j	$\sum_{i=1}^{N} \left[t_i \times \mathbf{1}_{\{c_{ij}>0\}}\right]$
Total copy number of function j weighted by species abundance	$\sum_{i=1}^{N} [c_{ij} \times t_i]$

Averages are unweighted and variation is measured as the coefficient of variation unless otherwise stated.