2020

Machine learning for microbial ecology: predicting interactions and identifying their putative mechanisms

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

AND

COLLEGE OF ENGINEERING

Dissertation

MACHINE LEARNING APPLICATIONS FOR MICROBIAL ECOLOGY: PREDICTING INTERACTIONS AND IDENTIFYING THEIR PUTATIVE MECHANISMS

by

DEMETRIUS M. DIMUCCI

B.S., University of California San Diego, 2011
MS., Boston University, 2014

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2020

Approved by

First Reader _____

Daniel Segrè, Ph.D.
Professor of Biology

Second Reader _____

Mark Kon, Ph.D.
Professor of Mathematics

DEDICATION

Dedicado a mi abuela.

ACKNOWLEDGMENTS

My experiences the past five years at Boston University as a graduate student have been overwhelmingly positive. I found the community that has been cultivated in the bioinformatics program to be both extraordinarily accepting and supportive of its current and prospective students. For that, I owe thanks to our administrative team members Dave King, Mary-Ellen, Johana, and Caroline.

I'd like to thank all of my committee members Drs. Kirill Korolev, Jenny Bhatnagar, Mark Kon, Mo Khalil. I extend a special thanks to my advisor Daniel invaluable feedback and conversation regarding my research projects and most critically, giving me the freedom to explore so many disparate ideas. I also thank all of the members of the Segrè lab who were there during my stay with a special thanks to all of the PhD students. It is hard to say how much of an influence our regular conversations had on my projects, but it was significant to be sure.

I would like to extend my heartfelt thanks to those people who provided me with opportunities that directly led to today: my undergraduate research advisor Dr. Paul Insel and Andrea Wilderman for teaching me how to do molecular biology in practice. Thanks to Mark and Candace Byrnes for taking me under their wing at the beginning of my academic journey and giving me such impactful guidance. Thanks to Stacy Jeambert whose decision to transfer me to 11th Marines Regiment saved me probably 2 years.

Finally, and most importantly I would like to thank family without whom none of this would have been possible. Primera, muchas gracias a mí mama, Marina Argueta, quien sacrificado mucho por mi educacion y mi futoro. Gracias a mis primos Victor y

Christopher por ser tu. Finalmente, muchisimo thanks to my wife, Nisha Rajagopal, for her unwavering support for me throughout graduate school in all its forms.

MACHINE LEARNING FOR MICROBIAL ECOLOGY: PREDICTING INTERACTIONS AND IDENTIFYING THEIR PUTATIVE MECHANISMS

DEMETRIUS DIMUCCI

Boston University Graduate School of Arts and Sciences

And

College of Engineering, 2020

Major Professor: Daniel Segrè, Professor of Biology, Bioinformatics, Biomedical engineering, and Physics

ABSTRACT

Microbial communities are key components of Earth's ecosystems and they play important roles in human health and industrial processes. These communities and their functions can strongly depend on the diverse interactions between constituent species, posing the question of how such interactions can be predicted, measured and controlled. This challenge is particularly relevant for the many practical applications enabled by the rising field of synthetic microbial ecology, which includes the design of microbiome therapies for human diseases. Advances in sequencing technologies and genomic databases provide valuable datasets and tools for studying inter-microbial interactions, but the capacity to characterize the strength and mechanisms of interactions between species in large consortia is still an unsolved challenge. In this thesis, I show how machine learning methods can be used to help address these questions.

The first portion of my thesis work was focused on predicting the outcome of pairwise interactions between microbial species. By integrating genomic information and observed

experimental data, I used machine learning algorithms to explore the predictive relationship between single-species traits and inter-species interaction phenotypes. I found that organismal traits (e.g. annotated functions of genomic elements) are sufficient to predict the qualitative outcome of interactions between microbes. I also found that the relative fraction of possible experiments needed to build acceptable models drastically shrinks as the combinatorial space grows. In the second part of my thesis work, I developed an algorithmic method for identifying putative interaction mechanisms by scoring combinations of variables that random forest uses in order to predict interaction outcomes. I applied this method to a study of the human microbiome and identified a previously unreported combination of microbes that are strongly associated with Crohn's disease. In the last part of my thesis, I utilized a regression approach to first identify and then quantify interactions between microbial species relevant to community function. The work I present in this dissertation provides a general framework for understanding the myriad interactions that occur in natural and synthetic microbial consortia.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC ..... Akaike Information Criterion

AUC ..... Area Under the Curve

CD ..... Crohn's Disease

IBD ..... Inflammatory Bowel Disease

KEGG ..... Kyoto Encyclopedia of Genes and Genomes

KO ..... KEGG Orthology

OLS ..... Ordinary Least Squares

OTU ..... Operational Taxonomic Unit

PICRUSt ... Phylogenetic Investigation of Communities by Reconstruction of Unobserved States

PR ..... Precision Recall

Qiime ..... Quantitative Insights into Microbial Ecology

RAST ..... Rapid Annotation using Subsystem Technology

RF ..... Random Forest

RNA ..... Ribonucleic Acid

rRNA ..... Ribosomal Ribonucleic Acid

ROC ..... Receiver Operator Characteristic

SMRT ..... Single-Molecule Real-Time Sequencing

UC ..... Ulcerative Colitis

CHAPTER ONE

Introduction and Background

Microbial communities have been found in nearly every environment on Earth [1]. The survival of an individual species hinges on the nature of its interactions with other species in the community [2–5]. From the perspective of a single species these interactions can exert positive, neutral, or negative influences on its ability to survive [6]. In turn, multicellular life is critically dependent on the biochemical processes that mediate the many inter-species interactions found within microbial communities [7–9]. Microbial communities can also have profoundly personal effects on individuals by causing or preventing disease [10–12], providing camouflage [13], and adding flavor to our food [14, 15]. Humans are no strangers to the use of microbial communities for industrial purposes. We have evidence that our ancestors used fermentation for cheese making nearly 10,000 years ago [16]. Despite our millennia-long relationship with microbial communities [17], the development of methods for intentionally manipulating their composition to obtain desired effects is still nascent.

In recent decades, strategies for improving the productivity of microbial communities for human purposes have focused on manipulating the genomes of member strains [18–20], altering media components in order to optimize communal productivity [21–24], or identifying the best starting compositions of several bacteria through exhaustive experimentation [25][26]. Studies of this last type have discovered that consortia of

multiple microbial species are often able to accomplish tasks that no individual species can on its own [27].

The success of fecal microbiota transplants in treating infection by *Clostridium difficile* [28] has contributed to the realization that communities can do things monocultures cannot. This realization has helped inspire the rise of an entirely new biomedical industry – microbiome engineering [29]. Ever since the scientific community turned its attention toward understanding the myriad inter-species interactions taking place in our guts, many in our society have hoped that understanding would bring with it the ability to engineer novel therapies [30]. Dozens of companies have appeared on the market, each promising to alter the trajectory of human health through the rational design of the bacterial communities living in our digestive tracts and on our food. To date, the grand promise of microbiome therapy for human health has yet to materialize [31]. Nevertheless, the scientific community is undaunted in its quest to understand the world of the microbes and their communities.

High throughput sequencing technologies, such as PacBio's SMRT sequencing [32], have provided researchers with an immensely powerful tool with which to spy on the microbial composition of the world around us [33]. Additionally, recent descriptions of novel culturing methods have provided tools for the rapid characterization of thousands of combinations microbial assemblages [34, 35]. For example, some of the emerging methodologies leverage microfluidics to produce many compositions of microbes [36]

while others enable the culturing of fastidious species [37]. The synthesis of these tools and existing methods has set the stage for improved effectiveness in the engineering of microbial consortia. The outputs of studies that use these tools are massive data sets relating community composition to function that can yield a plethora of insights when analyzed with specialized bioinformatics tools. Yet, significant challenges remain. In the following sections, I briefly discuss three challenges whose solutions will benefit the design of synthetic microbial consortia. Then, in the subsequent chapters of this dissertation, I present my work that focused on addressing these issues.

Challenge 1: Predicting interactions

The first challenge I address is that of predicting what the outcome of interactions between microbes will be in pairwise co-culture experiments. The implications of the answer to this are significant; if interactions between microbes are largely predictable then there is potential for considerable savings in terms of resources and time when designing consortia from the bottom up. In the design of synthetic consortia, candidate species are selected based on their known properties and expected cooperative or antagonistic interactions with the other species that will be in the community [38–41].

For a set of microbes, the number of possible pairwise combinations grows as the square of the number of species. Exhaustively characterizing every interaction is both resource intensive and likely to produce many redundant qualitative outcomes. In this context, there is much to be gained from the application of supervised machine learning. The

utility of machine learning to address this challenge is demonstrated in chapter two. By leveraging a relatively small fraction of the possible experimental space it is possible to construct a model that has high predictive performance on the qualitative, if not quantitative, outcomes of unobserved pairwise interactions. An accurate model will allow researchers to efficiently allocate resources toward those experiments that are most likely to result in interesting outcomes.

The actual implementation of machine learning algorithms is not challenging; however the features used for prediction should be carefully considered. The relationship of the predictive features to the predicted quantity has direct implications on the interpretation of model parameters. Various annotation tools are freely available that can generate feature sets from genomic content. One such tool is PICRUSt [42] which produces a set of predicted metabolic functions based on the 16s rRNA sequence of a bacterium. Another tool that can be effectively utilized is the Rapid Annotation using Subsystem Technology server (RAST) [43]. RAST takes as input the sequenced genome of an organism and returns to the user a draft of a stoichiometric metabolic model. The outputs produced by both of these tools, and others, are easily converted into feature sets for machine learning purposes. The specific representation scheme (e.g.presence/absence of metabolic functions, copy number of functions) is a choice that must be made with the study goals in mind.

Challenge 2: Identifying interactions that driving community-level phenotypes

Sometimes prediction of how microbes will interact or how they will function as a community is insufficient. We may also need to know the likely causal mechanisms so that we can better understand of the system we are studying. In this scenario the complex architecture of machine learning algorithms becomes a weakness because many methods are notoriously difficult to interpret, hence the phrase “black box” [44]. Some methods such as ridge regression [45] produce parsimonious models through regularization that retain most of their predictive power, but the catch is that important information about variable interactions may be lost in the process.

We also know that the causes of many biological processes are non-linear interactions of multiple constituent parts [46]. Due to the vast degree of biochemical diversity in microbial communities there may be multiple mechanisms by which different consortia achieve a given qualitative function. Provided we have selected a feature set that faithfully represents the underlying mechanisms (e.g. composition of bacteria in the gut, genes of individual species), multiple machine learning approaches will be capable of predicting which compositions will perform a given function – the challenge now becomes understanding why the algorithms make the predictions that they do. Developing an approach to answer this question, even for a single algorithm, would serve to help our experimental colleagues to generate novel mechanistic hypotheses and deepen our collective understanding of the phenomena we are witnessing. To this end, I have developed a novel set of algorithms for identifying these potentially complex interactions between species in large consortia. These algorithms are described in Chapter three.

Challenge 3: Identification of interactions between species in microbial consortia

One of the most commonly encountered hurdles in the analysis of microbial communities is to identify and quantify interactions between species as they relate to overall community characteristics. Examples of these include net biomass or typical disease status of a host organism. When the measured community characteristic is net biomass then an ideal data set for the relationship of biomass to community composition would be a time series of the absolute abundances of each species across multiple experimental conditions. With a data set like this, it is possible to identify interactions between species and to establish causality [40]. Because labor and resources are finite, these types of data sets will not always be available. It is much more likely that we will be in possession of a cross-sectional data set. For example, one may have abundance profiles of bacterial taxa found in the guts of patient samples of an observational study with corresponding disease status [47]. A popular approach in this case is to implement pairwise correlation analyses of the healthy and disease sample subsets in order to infer interactions between taxa. While conceptually easy, pairwise correlation methods are prone to producing many false positives [48] and the resulting network can become indecipherably complex when species have interactions with more than one other microbe [49]. Recent work has taken advantage of conditional independencies in the data to discard uninformative pairwise associations [50–52].

An additional complication that arises when dealing with cross-sectional data is a lack of information regarding the abundances of constituent species at the time when the target function is measured. This is a common occurrence in studies of synthetic consortia where the response of interest is a community level function and not the population densities of individual species per se [53–55]. In contrast to observational studies (e.g. microbiome studies from recruited donors), the controllable nature of experiments allows us to know the starting compositions of each microcosm, which can be used to optimize consortia for community function [56] and build explanatory models of community function [57]. When interactions between species assumed to not be significantly influential on the outcome, then simple additive models are expected to explain the data satisfactorily well. With this assumption in mind, feature selection methods such as LASSO [58], elastic net [59], or stepwise regression [60] are typically employed to produce parsimonious explanatory models. These methods are incapable of detecting interactions [61] and thus are unable, by themselves, to inform users when interactions between species should be considered. In order to evaluate the presence of interaction terms in parametric regression models, we must add each term of interest individually and estimate its influence.

Most current statistical methods are limited to evaluating pairwise interactions between microbes [62][63] but attention is turning to the identification of higher-order interactions [64, 65]. Efficiently identifying and quantifying high order interactions will allow us to characterize core community modules relative to a particular function that we could, in

principle, use as a foundation for further mechanistic studies. In Chapter four I present an approach that allows us to identify high-order interactions in microbial consortia and demonstrate on publicly available data set that their influence should be considered as a matter of course.

The future is bright for the marriage of machine learning and microbial ecology. There have already been several studies demonstrating the kinds of medical and agricultural riches we can expect to reap from engineered communities in the near future [66–69]. As the scientific community continues to build and release tools for annotating genomes and statistical methods for detecting interactions continue to disseminate through the biological sciences our collective prowess in engineering microbial communities will surely become formidable.

CHAPTER TWO

Machine Learning of Microbial Ecosystem Networks

Summary

This thesis chapter was published as the following research article:

Dimucci, D., Kon, M., Segrè, D. *Machine Learning Reveals Missing Edges and Putative Interaction Mechanisms in Microbial Ecosystem Networks.* mSystems. *Sep/Oct 2018 volume 3 issue 5*

Abstract

Microbes affect each other's growth in multiple, often elusive ways. The ensuing interdependencies form complex networks, believed to reflect taxonomic composition, as well as community-level functional properties and dynamics. Elucidation of these networks is often pursued by measuring pairwise interaction in co-culture experiments. However, combinatorial complexity precludes the exhaustive experimental analysis of pairwise interactions even for moderately sized microbial communities. Here, we use a machine-learning random forest approach to address this challenge. In particular, we show how partial knowledge of a microbial interaction network, combined with trait-level representations of individual microbial species, can provide accurate inference of missing edges in the network and putative mechanisms underlying interactions. We applied our algorithm to three case studies: an experimentally mapped network of interactions between auxotrophic *E. coli* strains, a community of soil microbes, and a large *in silico* network of metabolic interdependencies between 100 human gut-associated bacteria. For this last case, 5% of the network is enough to predict the remaining 95% with 80%

accuracy, and mechanistic hypotheses produced by the algorithm accurately reflect known metabolic exchanges. Our approach, broadly applicable to any microbial or other ecological network, can drive the discovery of new interactions and new molecular mechanisms, both for therapeutic interventions involving natural communities and for the rational design of synthetic consortia.

Importance

Different organisms in a microbial community may drastically affect each other's growth phenotype, significantly affecting the community dynamics, with important implications for human and environmental health. Novel culturing methods and decreasing costs of sequencing will gradually enable high-throughput measurements of pairwise interactions in systematic co-culturing studies. However, a thorough characterization of all interactions that occur within a microbial community is greatly limited both by the combinatorial complexity of possible assortments, and by the limited biological insight that interaction measurements typically provide without laborious specific follow-ups. Here we show how a simple and flexible formal representation of microbial pairs can be used for classification of interactions with machine learning. The approach we propose predicts with high accuracy the outcome of yet to be performed experiments, and generates testable hypotheses about the mechanisms of specific interactions.

Introduction

The collective behavior of microbial ecosystems across biomes is an outcome of the many interactions between members of the community [70–76]. These interactions include exchange of metabolites, signaling and quorum sensing processes, as well as growth

inhibition and killing. Understanding the interspecific interactions within microbial communities is essential for understanding the function of natural ecosystems [70–72, 75, 77] and for the design of synthetic consortia [74, 78–81]

A powerful and increasingly employed method for assessing microbial interactions is the direct measurement of phenotypes of microbial species grown in co-culture [81, 82][81, 82]. A fundamental challenge in this endeavor is the huge diversity of many natural communities, which could count up to several hundred strains or species of microbes. Performing experiments for all possible pairwise interactions constitutes a herculean, and likely insurmountable task for even a moderately sized community. It is however, conceivable that new computational approaches could systematically complement existing tools such as high-throughput sequencing and genome annotation [83–87]. to help extract as much information as possible from interaction datasets, providing both insight on yet-to-be-measured interactions, and on possible biological mechanisms mediating specific partnerships.

Here we present a conceptual framework for the mathematical representation of microbial interactions and subsequent use of supervised learning to build a classifier with high predictive accuracy. While any algorithm may be used, we obtained our best results with random forest [88–90]. Random forests are ensembles of many decision trees that individually are poor classifiers but can be democratically pooled to create a very good classifier. Random forests have two attributes that we found particularly attractive for our purposes. First, they are non-parametric and thus require no *a priori* definitions or

assumptions about underlying relationships between predictive variables. Second, recent methodological developments in the interpretation of random forests have been made that allow users to query why specific examples are classified as they are, through the calculation of feature contributions [91]. Feature contributions can be exploited to develop new hypotheses about the mechanisms that mediate specific interactions. In order to demonstrate a proof of principle for the classification of microbial interactions using organism traits and the utility of feature contributions for developing insight into the underlying mechanisms, we applied this approach to three communities where all pairwise experiments had been performed. The first is an *in silico* community of 100 metabolic models of human gut associated bacteria. The second community involves 14 amino acid auxotrophic strains of *Escherichia coli.* The third community is a collection of 20 microbial strains that were isolated from the same soil sample. Our results show that combining random forests with trait level representations results in high-performance classifiers. Furthermore, feature contributions have the potential to facilitate the discovery of new interaction mechanisms.

Results

Representing Pairwise Interactions

Our objective in this study was twofold; first, we sought to predict the qualitative outcomes of unobserved pairwise interactions in microbial communities; second, we wanted to identify predictive variables that suggest potential mechanisms of interaction. In order to achieve both of these goals it was important to establish a representation that can

be used by an algorithm to make good predictions and that can also be easily parsed for interpretation. Our approach relies on the availability of trait-level descriptions for each organism in the community under consideration. These trait descriptions are used to construct feature vectors for each organism (see Methods). Specific interactions are represented as the concatenation of the relevant trait vectors (Figure 2.1). Trait vectors may be constructed from any set of biologically relevant features such as presence/absence of a certain gene or metabolic function, phylogenetic classifications, or even characteristics of the environment where the organism was found. In our analyses, different case studies are based on different trait vector representations: in particular, we used (i) presence/absence of metabolic reactions for the *in silico* community case study, (ii) binary vectors of biosynthetic capabilities for each *E. coli* strain in the auxotroph community case study, and (iii) metabolic functions predicted from 16s sequences for the soil community case study.

These trait vectors, together with the known outcome of a subset of interactions, can be fed into machine learning algorithms that can separate outcome classes and subsequently predict the outcome of unobserved interactions. Here we use the random forest algorithm, based on an ensemble of many decision trees that individually ask a series of yes or no questions about randomly selected subsets of predictive features in order to classify samples. In order to find potential mechanisms of interaction we take advantage of the structure of individual trees in order to identify which variables are the most influential for the classification of specific samples.

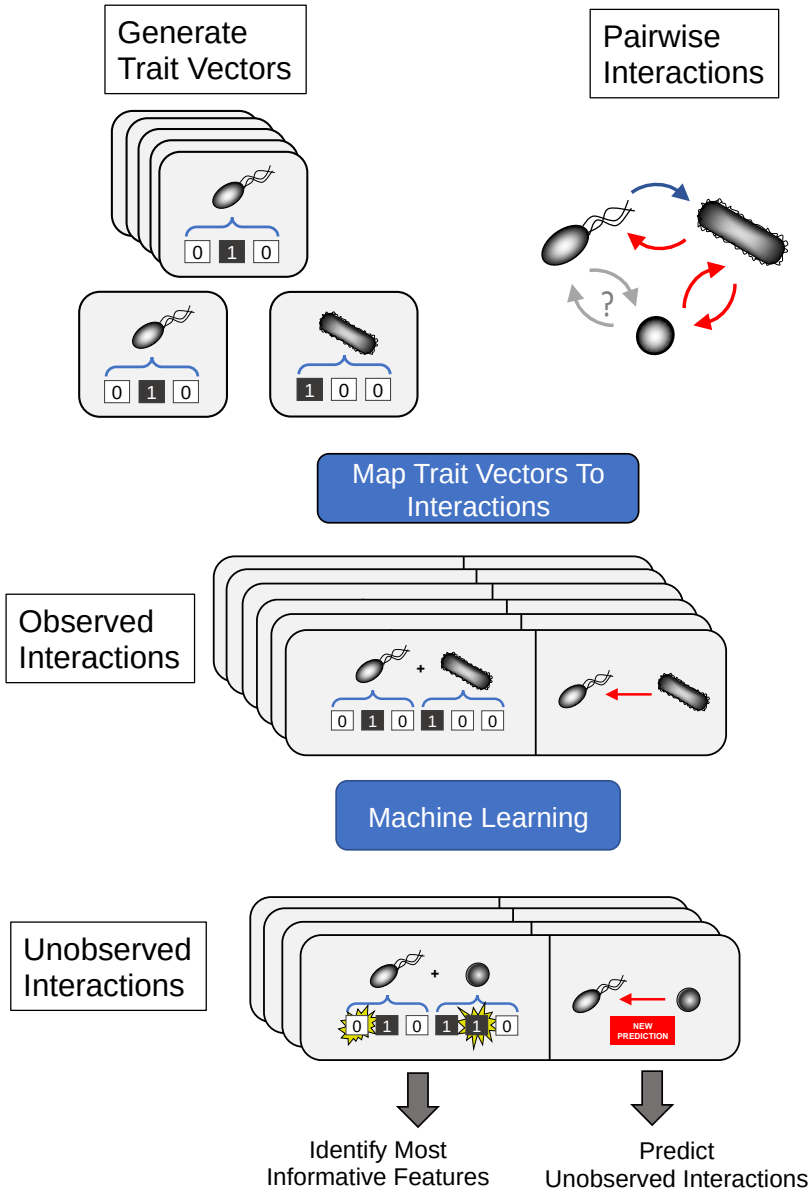


Figure 2.1. A schematic representation of our machine learning approach for inferring interactions among microbes. A trait vector captures the characteristics of each organism in the community of interest. The presence or absence of a trait in a given organism is encoded (as a binary number) in the corresponding element of the trait vector. For every possible pairwise interaction among community members we construct a composite vector that is the concatenation of the corresponding trait vectors. The vector of the organism whose response is being predicted is concatenated to the front of the trait vector of its interaction partner. For the set of observed interactions each composite vector is then mapped to the measured response of the interacting species. All observed interactions are then used to train a model that predicts the outcome of unobserved interactions. If random forest is used then feature contributions can be calculated on a

case-by-case basis in order to identify which elements of the composite genome contribute most strongly to the prediction.

Application to computationally predicted interactions between human gut microbes

We first applied our approach to a large *in silico* dataset that we generated by simulating time course microbial co-culture experiments with dynamic flux balance analysis [92, 93] using Computation of Microbial Ecosystems in Time and Space (COMETS) [74] (see Methods). Dynamic flux balance analysis enables the computation of approximate growth curves based on the complete metabolic networks of microbes (derived from their sequenced genomes), and the abundance of each nutrient present in the medium at the beginning of the experiment. At the end of a simulated experiment one obtains an estimate of the final biomass for each organism and the exchange fluxes during exponential growth. Possible interactions between different species in co-culture can emerge due to the exchange of secreted byproducts or the competition for common nutrients. In order to generate a large set of observations for machine learning we selected metabolic models of 100 human gut-associated bacteria [94] and used COMETS to simulate all pairwise co-culture interactions between them under the same rich medium, in a well-mixed batch culture scenario.

The trait vectors we used to represent each organism were simply binary vectors indicating the presence or absence of various nutrient exchange reactions in the metabolic network models (see Methods and Figure 2.2A). Interactions in the network were computed by determining the influence of every organism on every other organism in

COMETS co-culture simulations. In particular, the simulations provide the final biomass of each organism in co-culture and monoculture. A normalized difference between these two yields (which we refer to as relative yield, see Methods), is used as the phenotypic metric for classifying the interaction (negative or non-negative, see Figure 2.2B).

We first applied the random forest algorithm to the full dataset, and found that its out-of-bag (OOB) accuracy (which is roughly equivalent to five-fold cross-validation, see [95] and Methods) is approximately 90.5 %. The receiver operator characteristic (ROC) curve for the random forest algorithm (Figure. 2.2C and Methods) compares favorably to a naïve prediction based on the Jaccard distance [96] between the different trait vectors (see Methods, and [97, 98] for similar use of Jaccard distance in microbial communities studies).

High predictive accuracies are encouraging but are of little use if they can only be achieved when the vast majority of the experiment outcomes are already known. We thus constructed a series of learning curves to visualize how the balanced accuracy of the random forest classifier is affected by the size of the community and by the amount of training data available (Figure 2.2D). For small communities (for example, $N_{organisms} = 10$) there is little gain in predictive performance until the experimental space is nearly totally known. However, when $N_{organisms}$ is increased to 20 (which amounts to 190 pairwise experiments, corresponding to 380 individual responses to co-culture), as little as 5% of the total data (~9-10 experiments, i.e. 18-20 responses) is enough to obtain useful

predictions. ROC curves and comparison with a Jaccard distance classifier for selected points along the learning curve show a similar trend to what seen for the full dataset (Fig. A.3). The general trend indicates that the larger a community is, the smaller the relative fraction of experiments needed to get a high accuracy. In general, learning curves can be used as guidelines to determine how many experiments should be implemented in order to reach a target performance.

In addition to confirming that the algorithm can accurately classify unobserved interactions, we wanted to investigate whether the top feature vector components used as predictors are biologically interpretable. The variable importance plot (Methods and Figure. 2.2E) visualizes the globally most informative trait vector components. In this case, the most important predictors for the classification of a given organism is a feature of the interaction partner (Figure 2.2E). In other words, the predicted growth phenotype of organism *i* in presence of organism *j* is best described by features that are in the vector for organism *j*. In addition to analyzing the global contributions of variables on classification across all data, one can exploit the tree-based approach of random forests to determine why specific samples were classified as they were by examining the feature contributions for specific interactions. A feature contribution (see details in Methods) quantifies how much a given variable typically influences the classification probability of a single sample. Feature contributions were originally developed for analysis of regression models [99] but have since been adapted for binary classification models [91]. We wondered whether we could use the simulated data to illustrate the possible value of feature contributions in

identifying putative biological mechanisms underlying a given interaction. In particular, we envisaged that the random forest algorithm, trained only based on the trait profiles and the relative yields in co-cultures, could be used to suggest which metabolites may be more likely to mediate a given competitive (Figure. 2.3) or facilitative (Fig. A.1) interaction. As opposed to an *in vitro* system, where such prediction would need to be validated with new experiments, in our *in silico* system we can check the value of the random forest prediction by comparing it with simulated exchange fluxes across the two species (which, importantly, were not used in training the random forest).

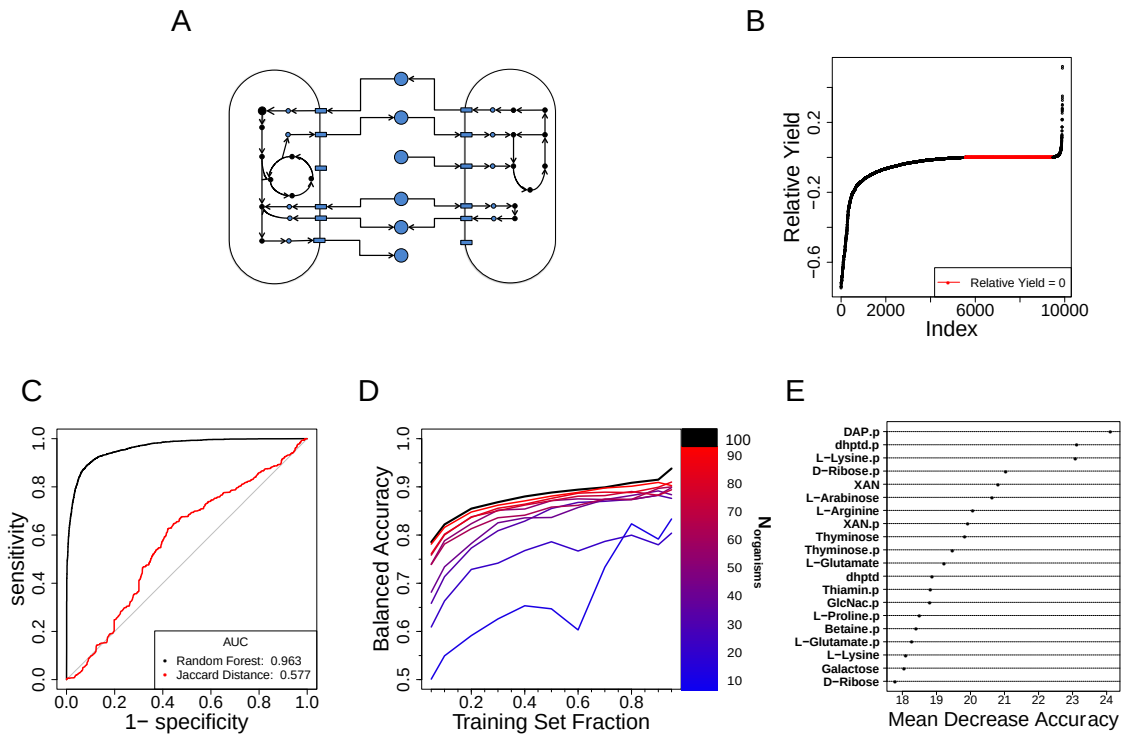


Figure 2.2. Classification of pairwise interactions for an *in silico* model of a community of human gut microbes. **A.** Organisms are represented *in silico* as large networks of metabolic reactions that take up metabolites (blue circles) from the environment (arrows leading to model) and release by-products (arrows leading to metabolite). Organisms may

interact with one another during simulation when both organisms compete for the uptake of a metabolite or through cross feeding where one model consumes a by-product of the other. **B.** Relative yields from all experiments are plotted in ascending order. There were 5563 samples with a negative relative yield. Neutral interactions, a relative yield of zero, occurred 3917 times, and positive relative yield happened 420 times. Samples were classified as negative or non-negative. **C.** For all 9900 *in silico* observations we determined the ROC curve of a random forest classifier using 388 exchange reactions as predictors and compared to the ROC curve obtained from using Jaccard Distance as a simple threshold to predict negative versus non-negative relative yields. Values for the ROC curve were obtained by evaluating the class voting ratios on out-of-bag samples (methods). ROC curves for classifiers trained on subsets of the data can be seen in supplemental figure A.3. **D.** Learning curves for sub-communities of the full *in silico* community. These learning curves are the median learning curve evaluated with 10 fold cross-validation on held out test sets at each point (methods) for 5 sub-communities selected at random for each value of $N_{organisms}$. **E.** Globally the 20 most influential predictors as determined by mean decrease in accuracy. The '.p' suffix indicates that the predictor belongs to the interaction partner. Some metabolite names are shortened: DAP = meso-2,6-Diaminopimelate, dhptd = 4-5-dihydroxy-2-3-pentanedione, XAN = xanthine, GlcNac = N-Acetylglucosamine. The results of an alternative representation scheme using phylogenies is presented in supplemental figure S5.

Towards this goal, for each pair of organisms, we ranked - by their net feature contributions (see Methods, Figure 2.3A and Supplemental Table A.1) - the 194 metabolites involved in exchange reactions. We found that metabolites ranking highly based on this criterion were much more likely than random to be among the metabolites truly exchanged in the COMETS simulations (Figure. 2.3B). This is particularly valuable if the interaction is due to a single exchanged metabolite (Figure. 2.3C). In practice, if this criterion were to be used on *in vitro* data, it would imply a significant reduction in the number of tests that would have to be performed to identify at least one mechanism of interaction.

It is also instructive to look in more detail at a specific case of feature contribution analysis. In particular, we observed that fructose exchange was most frequently the strongest predictor of competitive interactions (it was the top ranking true feature in ~18.7% of all competitive interactions, Supplemental Table A.1) and it corresponded to the 15th most common true mechanism based on the COMETS-simulated fluxes (Supplemental Table A2). Interestingly, fructose has been implicated in altering the gut microbiome in connection to a number of diseases, including antibiotic treatable [100] metabolic syndrome [101, 102], liver disease [103], and obesity [104]. Our approach is also readily applicable for the discovery of metabolites that mediate positive interactions, which comprise a small minority of all interactions (420/9900). Due to the scarcity of their occurrence and the dearth of metabolites that mediate positive interactions, discovery of these mechanisms is more challenging. Nevertheless, using ranked feature contributions to find facilitative metabolites was a powerful improvement over a naive approach (Supplemental Figure A.1).

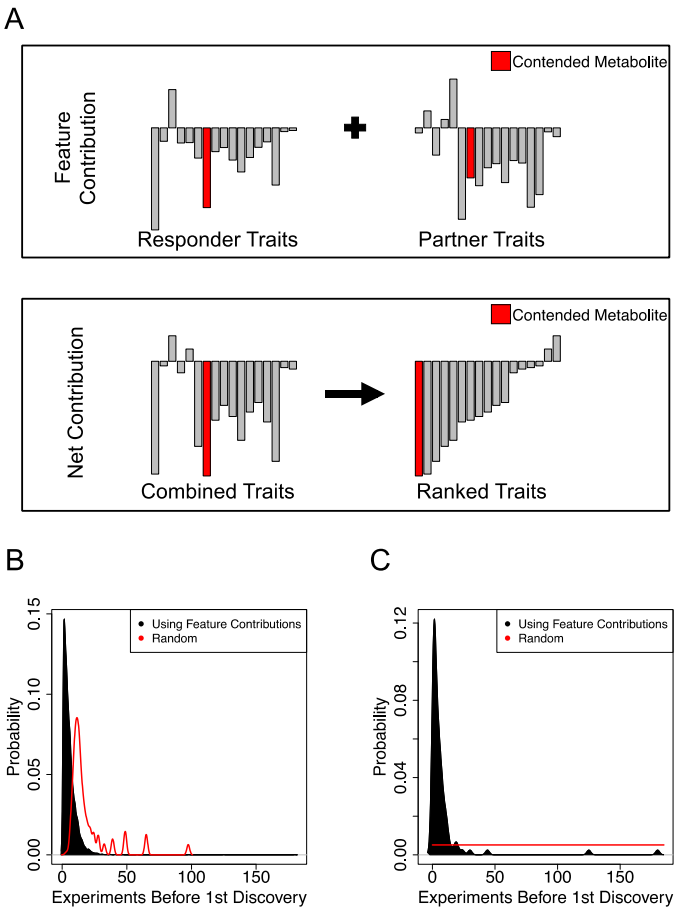


Figure 2.3. Using feature contributions to find a metabolite for which two organisms compete. **A.** The first half of the composite trait vector corresponds to metabolite transporters belonging to the organism of interest while the second half corresponds to metabolite transporters belonging to its interaction partner. We were interested in identifying a metabolite that is associated with the negative relative yield for the organism of interest. To establish a ranking of metabolites we took the summation of feature contributions from both halves of the composite trait vector and then sorted the new vector according to the net contribution. Proceeding from the negative end, the rank and identity of the first contended metabolite encountered relative to the negative end of the new vector was recorded. **B.** The probability distribution of the average rank at which the first mechanistic metabolite would be encountered by sampling metabolites randomly one at a time was calculated for each sample and compared to the observed distribution obtained by using feature contributions. By chance the first metabolite would typically be encountered after 13 random queries. Feature contributions reduce the median number of queries to 4. **C.** 99 samples produced a negative relative yield through the competition for exactly one metabolite. Randomly investigating the 194 candidate metabolites one at a time results in an average of 97.5 experiments before discovering the metabolite. Using feature contributions to prioritize the order in which to investigate metabolites instead

would typically reveal the contended metabolite on or before the fourth experiment (median = 4).

Application to a community of auxotrophic *Escherichia coli* strains

We next applied random forest to experimental data on auxotrophic *E. coli* co-cultures. In particular, we used previously published data from all possible co-cultures of 14 *E. coli* strains, each auxotrophic for a given amino acid [105]. Interactions between any given pair of *E. coli* strains are presumably dependent on the direct exchange of the missing amino acids, or related precursors (Figure 2.4A). The total growth of each strain in the 91 experiments was measured after 84 hours and reported as the net fold change relative to the initial inoculum, resulting in 182 total observations (see Methods for additional comments on the experimental setup). We built trait vectors based on the 14 amino acids, and labeled growth phenotypes based on the fold change response a given *E. coli* auxotroph strain had in co-culture with another auxotrophic strain (using 2 as the fold change cutoff for distinguishing between "strong" and "weak" interactions phenotypes, Figure 2.4B).

Random forest yielded a balanced accuracy of ~79.2% in predicting this interaction phenotype. Examination of the corresponding ROC curves shows that random forest is a much better predictor than simpler metrics based on biosynthetic costs [105] of the different amino acids (Figure. 2.4C). The learning curve for this test case (Figure. 2.4D) resembles the trajectory of the learning curve for *in silico* communities of 20 members (Figure 2.2D). Variable importance rankings show that, in general, the identity of the

amino acid needed by the receiver is more impactful on classification accuracy than the amino acid that its partner needs, suggesting that specificity of interaction is dominated by auxotrophies, whereas most mutants can in principle provide the missing amino acid (Figure 2.4F).

As for the *in silico* simulations, also in this case we analyzed the feature contributions, and asked whether they reflect the underlying mechanisms. In particular, we asked how often one of the two amino acids missing in a pair of organisms has the strongest contribution in random forest. As expected, the random forest is more strongly influenced by the absence of an amino acid feature than by its presence. Of all 182 observations, the amino acid missing from the receiver had the largest feature contribution 140 times and the amino acid from the giver had the largest contributor 40 times (Supplemental Table S3). Thus the pair of most influential predictors tended to correspond to the underlying mechanism of the interaction, even in instances where the predicted class was incorrect. Scenarios where the presumed mechanisms are the strongest contributors sometimes result in misclassification, presenting opportunities for direct research of interesting outliers. The response of the methionine auxotroph (**Δ**Met) in co-culture with the cysteine auxotroph (**Δ**Cys) was one such case, which we describe in detail in Fig. A.2.

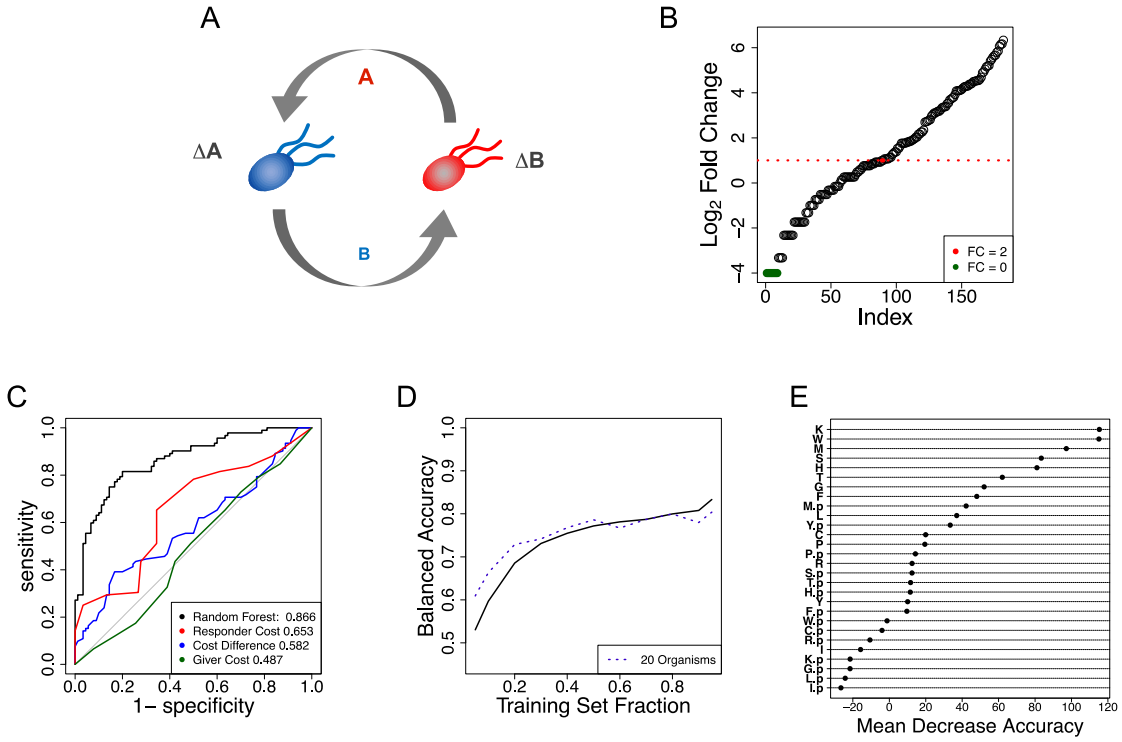


Figure 2.4. Data representation and results for the case study of a network of auxotrophic *E. coli* strains. **A.** In the original experiment, single gene knockout *E. coli* auxotrophs were co-cultured in a minimal medium. In order for ΔA to grow it must receive amino acid A from ΔB, which in turn must receive another amino acid, B, in order to grow itself. Auxotroph strains were constructed for the following amino acids: cysteine, phenylalanine, glycine, histidine, isoleucine, leucine, methionine, proline, arginine, serine, threonine, tryptophan, and tyrosine. **B.** Auxotroph strain fold changes in ascending order. *E. coli* strains had a weak response (fold change ≤ 2) 90 times and failed to grow at all 9 times (green circles). In 92 instances the *E. coli* auxotroph population more than doubled over the course of 84 hours. **C.** For all 182 observations we determined the ROC curve for a random forest classifier using 28 amino acids as predictors. Single value thresholds based on the biosynthetic costs of knocked out amino acids resulted in poorer performance than random forest. **D.** The trajectory of a learning curve built for the *E. coli* interactions closely resembles that of the learning curve for *in silico* communities with 20 organisms, dashed line. **E.** The 28 amino acids ranked according to their affects on prediction accuracy when randomly permuted. Amino acids corresponding to the receiver strain are enriched near the top of the list. Amino acids are represented by their single letter codes. The suffix '.p' indicates that the predictive feature belongs to the giver strain. In supplemental figure A.2 we examine the single case of ΔMethionine co-cultured with ΔCysteine.

Application to a community of soil bacteria

For our final test case we analyzed the results of a study featuring all pairwise co-culture experiments of 20 bacterial strains isolated from the same soil sample [72] (see Methods). For each experiment the authors reported whether each species was present at detectable levels at the final time point. We built trait vectors based on the presence or absence of KEGG modules [106] as predicted by PICRUSt [42]. Random forest trained on the full data set resulted in an out-of-bag balanced accuracy of 79.4%. The ROC curve shows that random forest performs much better than a simple decision rule based on the differences in the reported initial growth rates of each species (Figure 2.5A). The learning curve for this community closely resembles that of the 20-member communities from our *in silico* case study (Figure 2.5B). The variable importance plot shows that predictions are most strongly influenced by transport of teichoic acids (which are found in the walls of several gram positive bacteria [107]), both in the strain being predicted and in its interaction partner (Figure 2.5C, see supplemental table S4 for KEGG Module names).

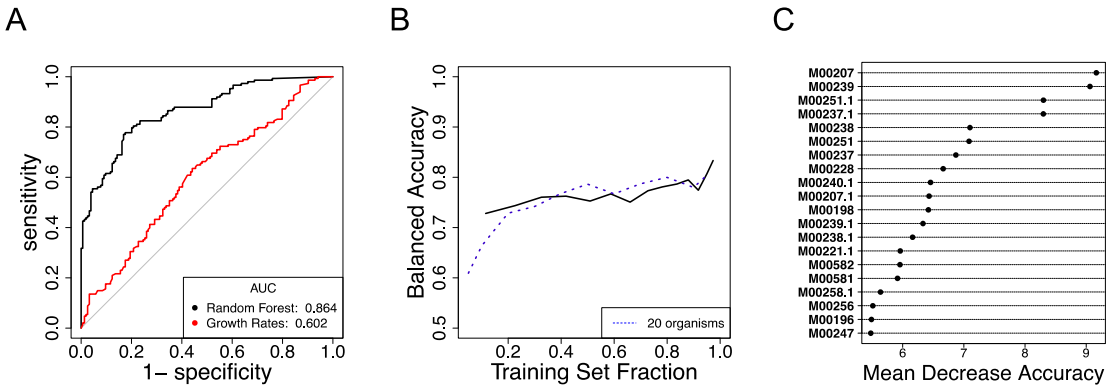


Figure 2.5. A. We determined the ROC curve from the random forest trained on all 302 observations using 79 predicted KEGG modules as features. Difference in the initial

growth rates of both strains was used as a baseline simple predictor. **B.** The learning curve built on this data set starts at ~72% balanced accuracy and tops out at ~78% balanced accuracy. The learning curve for the *in silico* communities with 20 organisms is displayed for comparison (dashed line). **C.** The IDs of the most important modules for predictive accuracy of the forest. See supplemental table S3 for the full module names.

Discussion

Exhaustive pairwise co-culture studies of microbial strains are an increasingly common avenue for estimating an ecosystem interaction network. While such pairwise interactions do not necessarily capture all possible interdependencies in a community [73, 108], they have been shown to be a dominant factor [81], making the reliable prediction and interpretation of predictive models matters of great importance. In this study, we have described a conceptual framework for the representation of microbes and their pairwise interactions in order to address both of these challenges.

Ideal datasets for testing our approach would include a large number of pairs of microbes, and genotypes or multi-dimensional phenotypes for each species. While we envisage that a multitude of such datasets will be available in the future, existing datasets are either limited in size or in trait vector accessibility. We thus tested our approach on three datasets, each with a different set of advantages and limitations. The first, and largest dataset was obtained by simulating 4950 microbial co-cultures with dynamic flux balance metabolic modeling (COMETS). An important caveat about this specific test case is that metabolic models may not capture the full biochemical details of the real system they approximate, and they do not incorporate any of the non-metabolic processes that one may expect to observe in real communities [109]. However, these models have been used to

successfully help understand the physiology of specific organisms [110] and communities [74, 111, 112]. The two experimental studies we used next are not affected by these issues, but they are limited in the number of organisms and pairs analyzed. The first experimental dataset is the outcome of a study involving 14 *E. coli* amino acid auxotrophs. In this case, the trait vectors are a straightforward representation of the auxotrophies, but the random forest has a chance to highlight the complexity of the underlying interdependencies. The second experimental dataset is from a community of soil microbes, whose trait vectors were derived from the available 16s rRNA sequences, suggesting a broad applicability of our approach to future similar studies.

Qualitatively predicting the outcome of unobserved interactions is most valuable if those predictions can lead to a reduction in the usage of precious resources and time. To this end the construction of learning curves is an important step in identifying how much data is required in order to achieve desired prediction accuracy from machine learning. This may be particularly useful for planning large-scale studies of naturally co-occurring species or synthetic consortia, e.g. for searching communities with specific properties relevant for biomedical or engineering applications [113].

Despite the common perception that random forest algorithms are merely un-interpretable "black boxes", we showed here that feature contributions provide a clear window into the decision-making process of a random forest. If the features are defined based on clearly identifiable biological entities (e.g. genes, reactions, or phenotypic traits) then feature

contributions can be effectively used for guiding experiments that can help reveal the underlying mechanisms.

In the current implementation of our algorithm we concatenated the binary trait vectors of two organisms in order to form a new composite trait representation. However, alternative representations of microbes and their interactions are possible, and should be explored. These could also include more quantitative information, such as gene copy number or mean transcriptional levels. While in our current work the environment for each case study was fixed, it is also possible to apply our method to data coming from heterogeneous environments, provided that the environmental parameters are encoded in the trait vector.

The current study focused entirely on demonstrating the possible benefits of applying machine learning to the study of inter-species interactions in microbial communities. In this context, our use of mechanistic models (based on dynamic flux balance analysis) was limited so far to the generation of *in silico* datasets meant to enable testing of our approach. However, we envisage that in the future it will be possible to integrate machine learning and mechanistic approaches towards a better characterization and design of microbial consortia. More broadly, we foresee that the interplay of quantitative approaches with high-throughput genotypic and phenotypic measurements will constitute a very valuable instrument for future microbiome research and synthetic ecology.

Data Availability. Pointers to datasets obtained from previous work, and used in our analysis are reported in the Materials and Methods Section.

The code and data tables necessary to reproduce all of our figures and analyses is hosted at: https://github.com/ddimucci/MicrobialCommunities

Materials and Methods

Representation of interactions with trait-derived features

For a given community *C*, the observed co-culture response of each species *i* in the presence of species *j* is encoded into the element X_{ij} of a community matrix $\boldsymbol{X}$. X_{ij} could represent the appropriately normalized abundance of species *i* at the end of a co-culture experiment with species *j*, or a binary variable describing whether or not species *i* will survive after inoculation with species *j*. To define a set of trait vectors for each organism in *C*, we start by obtaining a list of *n* features that can be assigned systematically across all organisms. These features could be the presence/absence of specific genes, metabolic functions, or any other relevant trait, so long as these features are not dependent on or derived from the quantities being measured. Thus, each organism *i* is assigned a *n*-long vector $\boldsymbol{F^{(i)}}$, whose k-th element $F_k^{(i)}$ is 0 or 1 depending on whether or not the corresponding trait is absent or present in the organism. Each pair of organisms (*i*,*j*) is then associated with a co-culture feature vector, defined as the concatenation of vectors $\boldsymbol{F^{(i)}}$ and $\boldsymbol{F^{(j)}}$, indicated as $\boldsymbol{F^{(i,j)}} = [\boldsymbol{F^{(i)}}, \boldsymbol{F^{(j)}}]$ (see Fig. 1). The behavior of a specific organism in a pair in co-culture, is thus formally described by the concatenated feature vector $\boldsymbol{F^{(i,j)}}$ and the corresponding phenotype X_{ij}. Note that in general $\boldsymbol{F^{(i,j)}} \neq \boldsymbol{F^{(j,i)}}$ and $X_{ij} \neq X_{ji}$.

Data generation for case study of *in silico* gut microbe interactions

Metabolic reconstructions of human gut-associated microbes were obtained from Bauer et al [94]. At the time of this writing these models can be downloaded directly from the following URL:

https://wwwen.uni.lu/content/download/86230/1056013/file/Bauer_et_al_301_microbe_models.rar

Each metabolic reconstruction encompasses the stoichiometry of virtually all metabolic reactions present in an organism, including uptake/secretion. Flux balance analysis (FBA) is a constraint-based steady state approach that uses this stoichiometry to predict fluxes and growth capacity under a given boundary condition of nutrient availability, and has been described in detail before [93, 109, 111, 114]. Briefly: The set of reactions contained in a model is derived from the organism's genome annotation. Reactions are then used to construct the stoichiometric matrix S for the metabolic model, whose element S_{ij} indicates the number of molecules of type i used or produced by reaction j. Identification of feasible metabolic fluxes ($\mathbf{v}$) for the system is achieved by imposing a steady state ($S\mathbf{v} = 0$), as well as upper/lower bound constraints that define the environmental nutrient availability. Standard flux balance analysis calculations then use linear optimization to identify feasible flux states that maximize a given objective function, usually the growth flux of the cell, i.e. the production of a balance biomass composition of the organism.

Dynamic flux balance analysis (dFBA) [92] extends classical FBA to perform dynamic simulations in which intracellular metabolites are still assumed to be at steady states, but

total biomass and environmental metabolites are treated as time-dependent variables in a discretized approximation. Crucially, in a dFBA simulation of multiple species, competition or facilitation (e.g. cross feeding) are emergent properties of the flux dynamics of individual organisms. Thus, no a priori assumptions need to be made about the existence or nature of ecological interactions. We performed dFBA simulations using our platform for Computation of Microbial Ecosystems in Time and Space (COMETS), which has been previously used to model microbial communities [74] . We selected 100 metabolic models [94] and identified a common medium that would permit the growth of nearly all models in a monoculture scenario. We then performed all pairwise co-culture simulations of the 100 models using the common media set in a well-mixed batch culture scenario (approximated by using COMETS without spatial structure). For each scenario we saved the record of biomass accumulation and fluxes in order to calculate relative yield and identify mechanisms of interaction, respectively. In this case, X_{ij} corresponds to the relative yield of strain i in co-culture with strain j at the final timepoint. X_{ij} can be directly computed from the amounts of biomass for different species at the end of the COMETS simulations. If B_{ij} is the final amount of biomass for organism i in co-culture with organism j, and the diagonal element B_{ii} is the biomass of i in monoculture, then the relative yield is defined as:

$$X_{ij} = (B_{ij} - B_{ii})/B_{ii}$$

$X_{ij} < 0$ indicates that strain *i* (which we also call the responder) is detrimentally affected by its partner. Correspondingly, an $X_{ij} = 0$ indicates no effect and an $X_{ij} > 0$ indicates a positive effect of *j* on *i*.

For this case study, the feature profile $\boldsymbol{F}^{(i)}$ for species *i* encodes the presence (1) or absence (0) of each of 194 possible exchange reactions (corresponding to the columns in the **S** matrix). It is important to note that these feature vectors are equivalent to functional annotations based on genomes, e.g. profiles of presence/absence of specific genes. They do not depend on the fluxes that can be eventually computed for each of the corresponding reactions.

In addition to implementing random forest, as described below, we also built a simple classifier based on the Jaccard distance (JD) between two feature vectors $\boldsymbol{F}^{(i)}$ and $\boldsymbol{F}^{(j)}$, defined as:

$$\mathrm{JD}(\boldsymbol{F}^{(i)}, \boldsymbol{F}^{(j)}) = 1 - (\boldsymbol{F}^{(i)} \cap \boldsymbol{F}^{(j)}) / (\boldsymbol{F}^{(i)} \cup \boldsymbol{F}^{(j)}).$$

Data for case study of auxotrophic *E. coli*

We obtained the measured growth response of individual *E. coli* strains and biosynthetic costs of amino acids from the supplemental files provided by [105]. In this study 14 amino acid auxotrophic strains of *E. coli* were generated knocking out single genes. Co-cultures were reported as being inoculated in 200 μL of M9 glucose media in 96 well microtiter plates at an initial cell density of 10^7/mL and incubated at 30 °C for 84 hours at which point the fold change in growth relative to initial inoculum for each strain was determined by plating, counting colonies, and qPCR to identify strain proportions. In this

case the feature vector $\boldsymbol{F}^{(i)}$ (of length n=14) encodes the presence/absence of biosynthetic capabilities for each of the 14 amino acids, and the co-culture phenotype X_{ij} corresponds to the fold change of strain i in co-culture with strain j at the final time point, which may represent the final growth yield. Based on the original dataset, batch effects (e.g. evaporation) or mutations seem not to have affected the quantitative estimate of the reported yield, and thus the outcome of our analysis. However, further scrutiny of the level of precision in yield measurements, and corresponding estimates of how experimental errors could affect machine learning outcomes would be an important subject for future follow up studies.

Data for case study of soil community

We downloaded the results of an experimental study of 20 soil microbial strains in which all pairwise co-culture experiments were performed in a yeast extract nutrient broth media [72]. Survival of different strains after 5 dilution cycles were estimated by plating co-culture media and counting colonies and verified with next-generation sequencing. For our analysis, we encoded in X_{ij} the reported persistence ($X_{ij} = 1$) or exclusion ($X_{ij} = 0$) of strain i when co-cultured with strain j. To generate feature vectors $\boldsymbol{F}^{(i)}$ for this community we downloaded the 16s rRNA sequences of each strain from GenBank [115] and used PICRUSt [42] to predict the presence of KEGG modules. We obtained KEGG modules for 18 strains and represented each strain with a binary trait vector of 79 modules (Supplementary Table S3).

Using PICRUSt to generate features

Full 16S sequences were obtained from the supplemental files of [72]. For each strain, we ran PICRUSt [42] to identify the number of genes within orthologous gene families in KEGG (KO numbers) [106]. We then assigned each strain a taxonomic identity and reference sequence with the Qiime [116] function pick_closed_reference_otus.py using the greengenes database (version 13.5) [117] at 97% similarity.

We next computed the fraction of genes within each reference genome that belonged to transport modules found in KEGG. We identified genes belonging to 154 transport modules from the KEGG database using the Restful Web API. More explicitly, let g_{ij} represent the copy number of ortholog *j* in strain *i.* We then computed the fractional abundance of each ortholog using the following equation:

$$g'ij = g_{ij} / \sum g_{ij}$$

We next computed the fractional abundance each transport module *k* using the following equation:

$$m_{ik} = \sum_{j \in Mk} g'_{ij}$$

Where M_k represents the set of genes in module *k.* If the fractional abundance was greater than 0 we represented the module as a 1 in feature space and as a 0 otherwise.

Implementation of random forest

We used the randomForest R library [95]. Random forests are ensemble classifiers that aggregate the results of many individual decision trees. This specific algorithm makes use

of two hyper-parameters: the number of training trees (nTree) and the number of predictors to consider at each split point (mTry). We determined that the default settings of nTree and mTry were near optimal for our *in silico* data set (Supplementary Fig. S4) and therefore used only the default setting for the remainder of the study. Each tree in the random forest is assigned a synthetic data set that is of the same size as the training set but generated through sampling with replacement. The result is that the average tree is trained on approximately 2/3 of the observations; these observations are referred to as in-bag samples. The remaining 1/3 of observations not in the synthetic data sets are referred to as out-of-bag samples. The new synthetic data set is placed at the root node of a new tree; next a randomly selected subset of predictive features is queried for the best split of the data into two child nodes. This process is repeated at each node until a stopping criterion is met. The classification accuracy of individual trees is assessed by using them to predict their out-of-bag samples and recording the results. The random forest then makes a classification call for individual samples based on what class the majority of trees predicted them to be. Accuracy was evaluated on the full training set with out-of-bag performance metrics and has been shown to be equivalent to 5-fold cross validation [95]. The ratio of the votes of the out-of-bag trees can be used to construct ROC curves (see below). See [88] for a full description of the algorithm.

Balanced accuracy

We report the balanced accuracy to evaluate the performance of classifiers on independent held out test sets and on the OOB samples when the model was trained using the full data

set. This metric is based on the values from the confusion matrix: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Balanced accuracy is calculated as follows:

$$\text{Balanced Accuracy} = [TP/(TP+FN) + TN/(TN+FP)]/2.$$

ROC curve

To evaluate the random forest classifiers for each case study we determined the receiver operator curve (ROC) from the model trained on the full set of available data. Using the out-of-bag voting proportions we plotted the true positive rate (sensitivity) against the false positive rate (1 – specificity) as the classification threshold was increased from the minimum value to the maximal value. In the context of random forest, the classification threshold is the fraction of out-of-bag votes for the positive class. After generating the ROC curve, we calculated the area under the curve (AUC) with the 'AUC' package in R [118].

Learning curves

To construct learning curves we defined a set of fractions, $\boldsymbol{r}$ = [.05, .1, .2, .3, .4, .5, .6, .7, .8, .9, .95] where we would evaluate balanced accuracy of the model using cross validation. For all cross-validation experiments we ensured that observations X_{ij} and X_{ji} were either both in the training set or in the test set. For each fraction in $\boldsymbol{r}$ we randomly selected a subset of the community matrix of the corresponding size to use as a training set and reserved the remaining data as an independent test set. This process was repeated until

at least 10 subsets of training data were selected for each value in ***r***. The median balanced accuracy of classifiers was then calculated for each fraction. In order to investigate the effect of the community size on the learning curve we defined a set of community sizes ***c*** = [10, 20, 30, 40, 50, 60, 70, 80, 90]. For each community size in ***c*** we randomly selected five community sub-matrices from the full *in silico* community matrix. Then for each sub-community we determined the learning curve. For each size in ***c*** the median learning curve for balanced accuracy of each community size was calculated and reported in Fig 2D.

Variable importance plots

Variable importance plots are commonly used with random forests to evaluate which variables are the most important for the model by comparing their mean decrease in accuracy scores. Mean decrease in accuracy is a measurement of the change in the accuracy of the forest's predictions when the variable in question is randomly permuted [89]. Here, we use it as a relative ranking of the global importance of each feature. The randomForest package automatically generates the variable importance plots which we visualize in Figs. 2E, 4E and 5C.

Feature contributions for binary classifications

The calculation of feature contributions has been described in [91]. The goal of this calculation is to quantify the effect of a given variable on the classification of a specific sample *j*. After training of a random forest with *T* trees, one can count, for each tree *t* and node *k* in the path followed by sample *j* in tree *t*, how many training samples at node *k*

belong to each of the two classes (C_1 and C_2). The fraction of samples belonging to C_1 is indicated by $Y^j_{t,k}$. Next, in order to evaluate the contribution of an individual feature *f* in classifying a specific sample, we perform the following steps: (i) At each node where feature *f* is the splitting variable, we calculate the local increment ($L^j_{t,k,f}$) in the fraction of samples belonging to class C_1, defined as $L^j_{t,k,f} = Y^j_{t,k+1} - Y^j_{t,k}$. (ii) We obtain a mean sample-specific contribution of a given feature *f* across all trees, by averaging over all the local increments, i.e.

$$\phi_f^j = \frac{\sum_{t=1}^{T} L^j_{t,k,f}}{T}$$

Feature contributions for all case studies were computed on out-of-bag trees using the forestFloor package available in R [119].

CHAPTER THREE

BowSaw: Discovering Explanatory High-Order Interactions for Biological Phenotypes

Summary

This thesis chapter will be submitted as the following manuscript:

Dimucci, D., Kon, M., Segrè, D. *BowSaw: Discovering Explanatory High-Order Interactions for Biological Phenotypes*

Abstract

Machine learning is revolutionizing biology by enabling the learning (prediction) of phenotypes from large datasets produced by studies utilizing high throughput methods including, metagenomic, genomic, and transcriptomic ones. Algorithms search data sets to discover complex and often nonlinear patterns in measured variables such as gene expressions or microbial taxa. There is a need to understand the underlying decision processes of algorithms, since doing so can enable generation of new mechanistic hypotheses. We have developed a set of algorithmic methods, collectively called BowSaw, that take advantage of the structure of a trained random forest (RF) algorithm to identify patterns frequently used by RF. We first demonstrate the utility of our approach by showing that it recovers causal patterns in simulated data sets, and explore the perfomance of BowSaw under increasingly difficult scenarios. We next apply our method to data from the Human Microbiome Project and find previously unreported high-order combinations of microbial taxa putatively associated with Crohn's disease. By

leveraging the structure of trees within a RF, BowSaw provides a new way of using decision trees to generate plausible hypotheses.

Introduction

The production of large life sciences data sets with high throughput techniques has increased the utilization of supervised machine learning algorithms to produce accurate predictions of phenotypes (e.g healthy vs disease). These algorithms use measurements of relevant traits such as gene variants, the presence/absence of microbial taxa, or metabolic consumption variables as predictors. Categorical prediction of phenotypes is typically the end goal of these applications. However, an additional benefit of these algorithms is the potential to extract explanatory classification rules. In this context, rules are defined by the values that a specific set of traits have when they are associated with a given phenotype. Identifying the relationships between the traits involved in classification rules may yield key insights into the biological processes associated with important phenotypes [120, 121] . This realization is creating demand for methods that assist in the interpretation of supervised machine learning methods [122–124], especially when the measured traits are expected to be causal agents such as genes or microbial taxa [125]. Identifying classification rules associated with a phenotype of interest is valuable because they are likely to carry information about the causal mechanisms that generate the phenotype.

Algorithms that are particularly valuable in this respect, are those involving decision trees, such as random forests, because the decision trees are easily interpretable

[126]. Decision trees are rule-based classifiers, where rules arise from a series of "yes-no" questions that can efficiently divide the data into categorical groups. In a biological context, such rules may arise from sets of genes whose simultaneous modulation could affect a phenotype, or sets of microbial species whose co-occurrence may be associated with a disease state. While in several cases it seems like disease phenotypes are uniquely associated with a single specific pattern (e.g. retinoblastoma [127]), there is increasing evidence for cases in which multiple distinct patterns can be associated with (and potentially causing) the same high-level phenotype [128, 129]. A particular example we will explore in this work is the multiplicity of distinct microbial presence/absence patterns which may be associated with Crohn's disease [130]. Crohn's disease has five clinically defined sub-types [131] but studies of the associated microbiome do not usually indicate which form of Crohn's disease a donor has been diagnosed with. Each sub-type of the disease may have different mechanisms, each requiring different treatment regimes. Thus, identifying rules associated with sub-populations within a particular phenotype label are of great interest due to potential therapeutic implications.

The fact that there may be multiple rules generating the same or similar phenotypes complicates the straightforward interpretation of parameter coefficients or variable importance scores [132, 133]. Uncovering the multiple interactions between predictive variables as they relate to phenotypic labels remains a challenging statistical endeavor, but one that is of paramount importance. Identifying multiple interactions associated with a disease enables the development of mechanistic hypotheses for follow up-studies. This challenge, and an overview of the key strategy we propose, are

illustrated in Figure 3. In figure 3.1A we depict a toy-model where measured variables (traits) have only two possible values (e.g.: present/absent), the high-level phenotype (category) is binary (e.g.: no disease/disease), and two distinct rules can both generate the phenotype. The goal in this case is to identify each of the rules that are associated with the phenotype. In this work we will show how this can be achieved by in-depth analyses of a random forest (RF) (Fig. 3.1B).

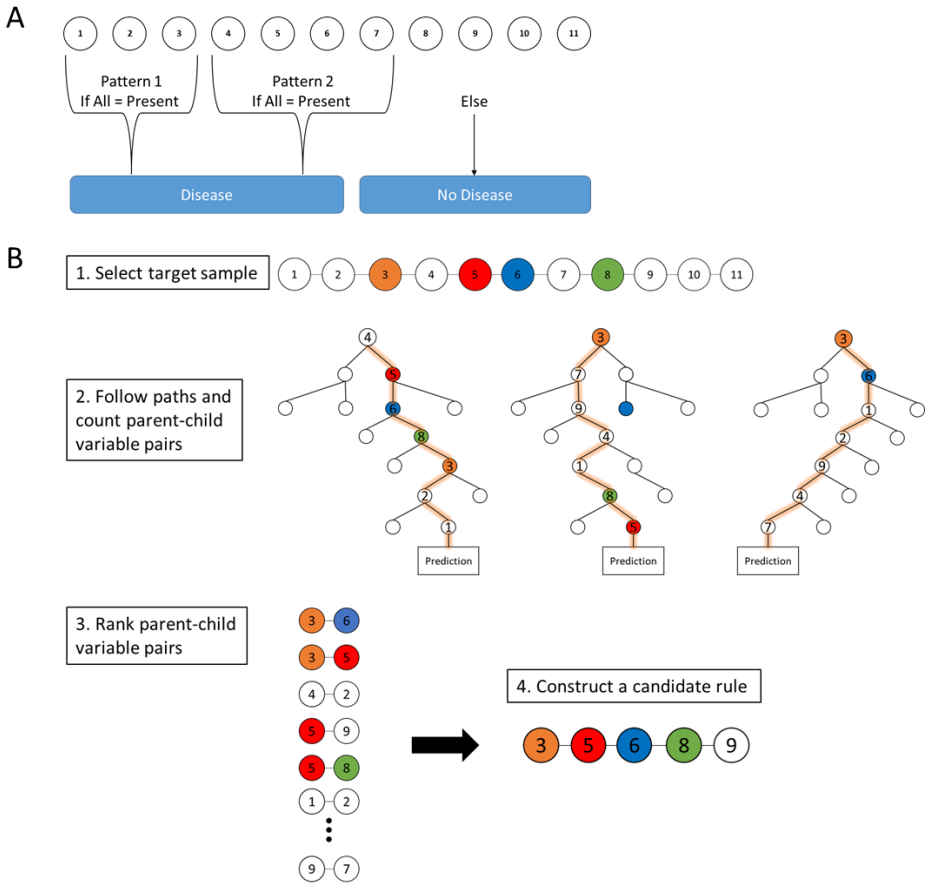


Figure 3.1 A) In a hypothetical data set there may be two phenotype labels – "Disease" and "No Disease" that we wish to discriminate based on input predictor variables. In this example, there are two distinct high-order patterns that both confer the "Disease" phenotype. B) Conceptual pipeline of BowSaw. In (1) we begin by identifying a target vector, in this case the colored nodes indicate the true pattern. In (2) we follow the path of the sample through each of its out-of-bag trees and record how often the sample

encounters sequential pairs of variables. (3) Each ordered pair sequence is sorted in descending order by frequency. (4) Pair sequences are used to generate a candidate rule that is maximally associated with the observed phenotype of the target vector.

The random forest algorithm intrinsically takes advantage of non-linear relationships between variables and is widely used in the life sciences [134–136]. RFs, when used to distinguish between disease states known to have multiple causes, often result in excellent classifiers [137, 138]. It has also been reported that RFs capture subtle statistical interactions between variables [132]. Unfortunately, a RF is not straightforwardly interpretable despite its hierarchical structure, and recovering those interactions is notoriously difficult [133] due in large part to the method's reliance on ensembles of trees [89]. The difficulties in interpretation created by these properties has led many to refer to RF as a 'black-box' model [44].

Identifying the rules that a RF utilizes in classification tasks is an active area of research, and many strategies have been developed to address this problem. Effective strategies have focused on evaluating how individual variables influence the classification probabilities of specific samples [91, 119], pruning existing decision rules found in the tree ensemble to produce a compact model [139], computing conditional importance scores [140], or iteratively enriching the most prevalent variable co-occurrences through regularization [141]. Each of these approaches offer valuable methods for the identification of statistical interactions between variables. However, we and others have observed that while these methods are capable of recovering true causal rule in simulated data when exactly one such rule is present, the existence of multiple rules associated with one phenotype can confound interpretation efforts [141].

Here we describe BowSaw, a new set of algorithms that utilizes variable interactions in a trained RF model in order to extract multiple candidate explanatory rules. With BowSaw, we set out to develop a *post hoc* method intended to aid in the discovery of these rules when the input variables are categorical in nature. The main idea of BowSaw is to systematically quantify the co-occurrence of specific pairs of traits across multiple trees, and assemble these co-occurring pairs into larger sets of traits that best account for a phenotype. We first demonstrate that BowSaw is capable of recovering true rules by applying the algorithms to simulated data sets of varying complexity. We then apply BowSaw to a study on the role of the gut microbiome on Crohn's disease [130], and show that it can find a previously unreported combination of microbial taxa that is fully associated with Crohn's disease. BowSaw can be broadly applied to any dataset with categorical or discrete predictors.

Methods

Overview of the pipeline

Provided with a trained random forest and a training set, BowSaw goes through three steps in order to generate a candidate rule (variable-value combination) for each observation associated with the phenotype of interest (Figure 3.1. B). First, for each observation, the *Count* algorithm counts the frequency of unique ordered pairs of variables encountered along each relevant tree in the forest. Second, for each observation, the *Construct* algorithm generates a list of ordered pairs of variables, ranked by their frequencies, and uses this list as a guide to construct a decision rule (which could consist

of two or more variables) that is maximally associated with the observed phenotype. Finally, the *Curate* algorithm pools all of the rules together in order to select a subset of rules that collectively account for all of the samples with the desired phenotype. Optionally, the *Sub-rule* algorithm can be used to generate pruned versions of candidate rules prior to applying the Curate algorithm in order to obtain a more concise, albeit less specific, set of candidate rules. The Count and Curate algorithms generate the candidate rules for individual observations while the Curate and Sub-rule algorithms produce a combined set of rules that account for all observations with the chosen phenotype.

In the following section, we provide a description of the inputs BowSaw takes and the algorithms that implement these steps along with pseudocode.

Inputs

BowSaw takes as inputs a dataset, $\boldsymbol{D}$, composed of $\boldsymbol{N}$ observed vectors each of $\boldsymbol{p}$ categorical variables. There are assumed to be $\boldsymbol{K}$ possible class labels for each vector in $\boldsymbol{D}$ which for the purposes of this discussion denote different phenotypes. A random forest is assumed to be trained on $\boldsymbol{D}$ to distinguish the classes $\boldsymbol{k = 1, \dots, K}$. Additionally, BowSaw takes a target vector $\boldsymbol{n_i}$ with observed phenotype $\boldsymbol{k_i}$ for which the goal is to identify a set of simplified association rules.

Counting stubs

Given an RF machine $\boldsymbol{M}$ trained on dataset $\boldsymbol{D}$ and a feature vector $\mathbf{x} = (x_1, \dots, x_p) \in \boldsymbol{D}$ the first sub-routine of our method (the *count algorithm*) proceeds as follows. It starts by identifying among the set of trees in $\boldsymbol{M}$ sub-paths (sequences of successive variable

indices) encountered by **x** as it travels through $\boldsymbol{T}$, its set of out-of-bag trees. An out-of-bag tree is a tree in which **x** was not including in the training set. For the specific path $\boldsymbol{T_i}$ with $\boldsymbol{v + 1}$ variable indices, each stub (ordered pairwise variable sequence) along $\boldsymbol{T_i}$ (e.g. $\boldsymbol{T_iT_{i+1}}$) from $\boldsymbol{i = 1, \ldots v}$ is accounted for in an $\boldsymbol{n}$ x $\boldsymbol{n}$ matrix, $\boldsymbol{C}$, where the element $\boldsymbol{C_{ij}}$ records the number of instances of the stub in all paths of $\boldsymbol{T}$.

Algorithm 1: Count Algorithm Pseudocode

For $\boldsymbol{t}$ in $\boldsymbol{T}$**:**

$\boldsymbol{v}$ = {1,2, …, $|\boldsymbol{t}| - 1$}

For $\boldsymbol{i}$ in $\boldsymbol{v}$

$\boldsymbol{C_{ti,ti+1} = C_{ti,ti+1} + 1}$

End loop

End loop

Return $\boldsymbol{C}$

Constructing a candidate rule

The second sub-routine (the *construct algorithm*) builds a candidate rule for **x** by first placing all of the stubs with non-zero counts in a list $\boldsymbol{L}$ sorted in descending order by their values in $\boldsymbol{C}$. We define the candidate rule $\boldsymbol{R}$ as a set of paired lists $\{\boldsymbol{R_{indices}, R_{values}}\}$ and initialize it by adding the stub from $\boldsymbol{L}$ to $\boldsymbol{R_{indices}}$ and the corresponding values from **x** to $\boldsymbol{R_{values}}$. Next, we identify all of the observations in $\boldsymbol{D}$ whose values at $\boldsymbol{R_{indices}}$ are exact matches for the corresponding values in $\boldsymbol{R_{values}}$ and store their indices in $\boldsymbol{I}$. The fraction of observations in $\boldsymbol{I}$ that have phenotype $\boldsymbol{k}$ is $\boldsymbol{F}$. Stubs and values are added to $\boldsymbol{R}$ until $\boldsymbol{F}$ is

either equal to 1 or all of the stubs in $\boldsymbol{L}$ have been exhausted. We retain in $\boldsymbol{R}$ only those variable pairs whose additions increase $\boldsymbol{F}$.

Algorithm 2: Construct Algorithm Pseudocode

Make ranked list $\boldsymbol{L}$ of stubs from $\boldsymbol{C}$

Initialize $\boldsymbol{R}$ as the paired lists $\{\boldsymbol{R_{indices}}, \boldsymbol{R_{values}}\}$

$\boldsymbol{I} = 0$, $\boldsymbol{F} = 0$

For stub i in $\boldsymbol{L}$:

    If $\boldsymbol{F} = 1$:

        Exit loop

    Else:

        $\boldsymbol{R'_{indices}} = (\boldsymbol{R_{indices}}, \boldsymbol{L_i})$

        $\boldsymbol{R'_{values}} = (\boldsymbol{R_{values}}, \mathbf{x}_{\boldsymbol{Li}})$

        $\boldsymbol{I'} = \boldsymbol{I} \cup (\boldsymbol{D_{R'indices}} = \boldsymbol{R'_{values}})$

        $\boldsymbol{F'} = |\boldsymbol{k_{M'}} = \boldsymbol{k}| \,/\, |\boldsymbol{I'}|$

    If $\boldsymbol{F'} > \boldsymbol{F}$:

        $\boldsymbol{R_{indices}} = \boldsymbol{R'_{indices}}$

        $\boldsymbol{F} = \boldsymbol{F'}$

        $\boldsymbol{I} = \boldsymbol{I'}$

End loop

Return $\boldsymbol{R, I, F}$

Curating candidate rules:

The *count* and *construct* algorithms are the heart of BowSaw. In our workflow, we apply these algorithms to each observation in $\boldsymbol{D}$ that has the desired observed phenotype $\boldsymbol{K_i}$. We call the set of these vectors $\mathbf{x_{all}}$. By default, we produce a single candidate rule for each vector in $\mathbf{x_{all}}$. We store each candidate rule in list $\boldsymbol{Q}$ and rank them by their respective values of $|\boldsymbol{I}|$. Since $\boldsymbol{Q}$ may include many redundant rules, we developed another sub-routine (the *curate algorithm*) to generate a concise set of candidate rules that collectively account for $\mathbf{x}_{\boldsymbol{all}}$. Briefly, we create a new empty list, $\boldsymbol{E}$, to which we add the top ranked rule from $\boldsymbol{Q}$ (by default this is the rule with the greatest value for $\boldsymbol{I}$), and record the index of samples in $\boldsymbol{D}$ that match any rule in $\boldsymbol{E}$ and also have the desired observed phenotype class, $\boldsymbol{K_d}$, in the set $\boldsymbol{A}$. Next, we determine how many samples remain unaccounted for $\boldsymbol{U} = \boldsymbol{x_{all}} - \boldsymbol{A}$, then we determine which of the remaining rules in $\boldsymbol{Q}$ minimizes $|\boldsymbol{U}|$, add it to $\boldsymbol{E}$, and repeat these steps until $\boldsymbol{U}$ is an empty set.

Algorithm 3: Curate algorithm pseudocode

$\boldsymbol{Q}$ = ranked list of all candidate rules for Φ_t

$\boldsymbol{E} = \boldsymbol{Q_{best}}$ (user defined, default is maximum $\boldsymbol{M}$)

$\boldsymbol{I^*} =$ which $\boldsymbol{D}$ match any rule in $\boldsymbol{E}$ and $\boldsymbol{k} = \mathbf{K_d}$

$\boldsymbol{A} = \boldsymbol{x_{all}} \cap \boldsymbol{M^*}$

$\boldsymbol{U} = \boldsymbol{x_{all}} - \boldsymbol{A}$

While $\boldsymbol{U}$ is not empty**:**

$\boldsymbol{B}$ = { }

For rule $\boldsymbol{i}$ in $\boldsymbol{Q}$:

$\boldsymbol{E^* = E + Q_i}$

$\boldsymbol{I^*}$ = which $\boldsymbol{D}$ match any rule in $\boldsymbol{E^*}$ and $\boldsymbol{k = K_d}$

$\boldsymbol{A^* = x_{all} \cap I^*}$

$\boldsymbol{B_i = |U - A^*|}$

End loop

$\boldsymbol{best}$ = which min $\boldsymbol{B_i}$

$\boldsymbol{E = E + Q_{best}}$

$\boldsymbol{M^*}$ = which $\boldsymbol{D}$ match any rule in $\boldsymbol{E}$ and $\boldsymbol{k = K_d}$

$\boldsymbol{A = x_{all} \cap M^*}$

$\boldsymbol{U = U - A}$

End while loop

Return $\boldsymbol{E}$

Constructing sub-rules

Since rules are rarely 100% associated with any given phenotype, we devised a strategy for selecting a set of candidate sub-rules that account for all samples with desired observed phenotype class $\boldsymbol{K_d}$. Candidate sub-rules are shorter candidate rules derived from larger candidate rules by omitting one or more variables. For each candidate rule in $\boldsymbol{E}$, we identify sub-rules that meet a user-defined complexity criteria, e.g. only produce sub-rules that are composed of three or four variables and their corresponding values. We place each of the unique sub-rules into a new list $\boldsymbol{E_{sub}}$. Then the corresponding number of

identical matches, ***I***, and proportion of ***I*** that have the phenotype $\boldsymbol{K_d}$, ***F***, are determined. At this stage, we can apply our third sub-routine (the *Curate* algorithm) to $\boldsymbol{E_{sub}}$ to obtain a parsimonious list of sub-rules that accounts for $\mathbf{x_{all}}$. In our pipeline, we also choose thresholds based on desired levels of ***I*** and/or ***F*** in order to eliminate poor candidate sub-rules from consideration. In this study, we decided on the thresholds after visually inspecting a plot of ***F*** against ***I*.**

Algorithm 4: Sub-rule algorithm pseudocode

$\boldsymbol{E_{sub}}$ = { }

Complexity = {user defined numeric values}

For ***rule*** in ***E***

    For ***i*** in ***Complexity***

        $\boldsymbol{Esub} = \boldsymbol{E_{sub}} \cup (\frac{\boldsymbol{rule}}{\boldsymbol{i}})$

    End loop

End loop

The algorithms described above are generalizable to multi-classification tasks but are currently limited to discretized or categorical representations of the feature space. Pseudocode for implementing each of the algorithms described above along with an implementation of the algorithms in R [142] can be found in the supplemental files and on github: https://github.com/ddimucci/BowSaw.

Results

Application to simulated Data

To test the capacity of BowSaw to recover multiple decision rules, we applied it to increasingly challenging simulated data sets. These data set consists of binary vectors representing different observations. The phenotype associated with each observation is a function of the corresponding vector. The function consists of a set of multiple mutually distinct rules, such that if a rule is satisfied, it will cause the observation to have the phenotype with a certain probability (which we call here "penetrance" because of its resemblance to the genetics concept). The first dataset (IDEALIZED) we use is relatively simple, and includes multiple equally prevalent rules. It is also generated under the assumption that there are no unmeasured confounders, i.e. that if an observation does have a phenotype, then it must be satisfying at least one of the above rules. We then apply BowSaw to a more challenging scenario (INTERMEDIATE) in which the phenotype-generating rules differ in their relative prevalence and the assumption of unmeasured confounders is violated. Finally, is a set of data sets with complex co-varying parameters (COMPLEX), we systematically varied the underlying parameters of the simulation and examined the relationship between summary statistics of the RF performance and the ability of BowSaw to generate candidate rules containing the true phenotype-generating rules.

For the IDEALIZED scenario, we simulated data set of 100 independent random binary variables and 2,000 observations. We randomly defined five rules that each required four randomly selected variables each to have specific values (e.g. all variables

equal to 1) in order to assign a hypothetical phenotype with likelihood between .8 and .9. Here we present the results of this scenario with a specified random seed, but other seeds and parameters can be explored using the scripts provided in the supplemental files. Using these parameters 479 samples were assigned the phenotype and BowSaw produced a set of 135 unique candidate rules ranging in complexity from six to fourteen variables. From these rules, we produced all sub-rules ranging involving anywhere from two to five variables, which resulted in unique 50,034 sub-rules. We calculated the number of matches $|\boldsymbol{I}|$, the proportion of samples with the phenotype, $\boldsymbol{F}$, for each sub-rule, and visualized these values in order to select an association threshold (Figure 2A). To reduce the number of sub-rules that the curate algorithm would need to examine, we eliminated from consideration any rules that had an $\boldsymbol{F}$ below 80%. We selected an 80% threshold because in the cluster centered around 125 matching samples there is a small cloud of rules that are clearly segregating the phenotype more efficiently than the others are. We selected the sub-rule with largest $|\boldsymbol{I}|$ among these as the top candidate rule. This produced a final list consisting of five candidate rules that accounted for all of the samples with the phenotype and were each one of the true phenotype generating rules (Figure 3.2A red points). These results demonstrate that in an ideal scenario with no phenotype diagnosis errors BowSaw is indeed capable of recovering multiple true rules.

For the more challenging scenario (INTERMEDIATE), we generated the data set the same as before except this time we allowed the five underlying rules to vary in complexity from three to five variables. Varying the complexities of rules resulted in different prevalence among them, as rules that are more complicated are less likely to

appear in the data. In this case, we had one rule of complexity five, two that required four variables, and two that used three variables. We also added background noise by randomly assigning the phenotype to 2% of samples that did not possess any of the rules. BowSaw produced 176 unique candidate rules involving between six to thirteen variables. From this list we generated 68,938 sub-rules and chose an association threshold of 75% because there are two clusters at ~$|\boldsymbol{I}|$ = 125 that begin to clearly separate in that range and the two outlier points at ~$|\boldsymbol{I}|$ = 250 do not combine to account for all of the phenotype (Figure 3.2B). Applying the curate algorithm to the rules meeting this threshold produced 20 candidate sub-rules the top four (when ranked by $|\boldsymbol{I}|$) of which were true rules. The rule of five variables was not recovered. These results show that BowSaw is able to recover strongly associated patters (and in this case, causal patterns) even in the presence of noise, but low prevalence rules can be masked by high prevalence rules.

We used the same data generation method to investigate BowSaw's ability to produce candidate rules containing true rules when the underlying parameters change. We applied BowSaw to 20,000 simulated data sets where we randomly altered the number of features, sample size (200 or 2,000 samples), complexity of the rules, number of rules, the likelihood of each rule assigning the phenotype, and the background noise. We identified scenarios where rule recovery with BowSaw performs very well and situations in which it fails to recover any rules at all. Additionally, we found a strong linear relationship between BowSaw's performance measured as the average fraction of rules recovered and the of number of samples, number of features, and two evaluation

metrics for RF model – the area under the curve for both the receiver operator characteristic and precision recall curves (Figure 3.2C).

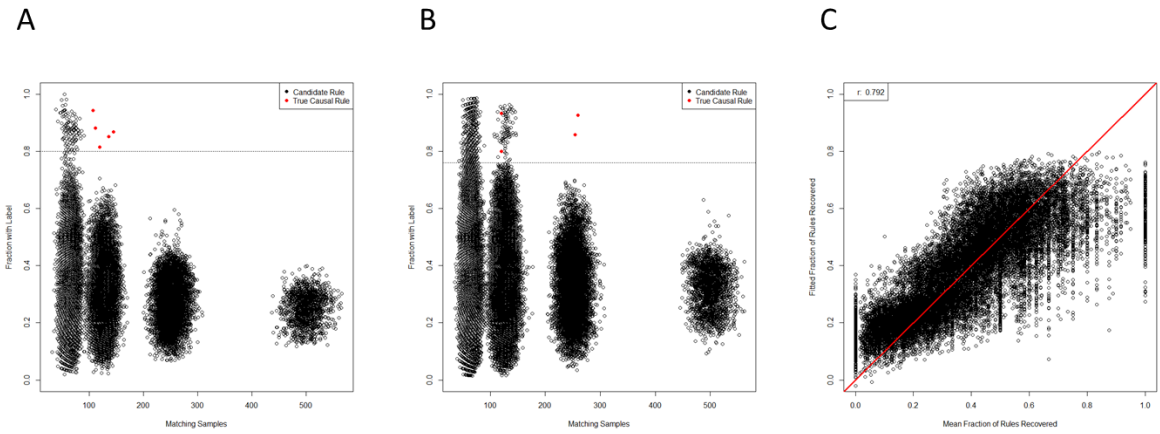


Figure 3.2 A) Candidate sub-rules generated for the ideal scenario. Each point represents a unique sub-rule. X-axis is the number of samples that exactly match the pattern defined by the rule. Y-axis is the fraction of matching samples with the observed phenotype. Each cluster corresponds to decreasing rule complexity from 5 variables per rule to 2 on the right most cluster. These clusters appear because the data is produced by a binary distribution. Dashed line is the association threshold we set. Red points are the causative sub-rules we defined. BowSaw identified by all five red points in this scenario. B) Candidate sub-rules generated for the more challenging scenario. We defined 5 causative rules of varying lengths in this scenario but BowSaw was only able to recover 4 of them completely. The longest rule which was 5 variables long was not recovered. In this scenario BowSaw recovered the 4 red points. C) There is a strong linear relationship between the performance of BowSaw and observable metrics measured here as the average fraction of a complete rule recovered in any candidate rule (y). The linear model we specified was: y = sampleSize + #features + ROCAUC + PRAUC.

Application to Human Microbiome Data

Irregular distributions of microbial taxa within the gut are often associated with serious illnesses such as Crohn's disease or ulcerative colitis [143, 144]. Human microbiome studies regularly use 16s sequencing methods and extensive reference databases to report on microbial taxa found in samples as operational taxon units (OTUs).

RF classifiers are frequently built using counts of OTUs to accurately discriminate between disease and healthy patient samples [145, 146]. Despite their demonstrated effectiveness as good classifiers of Crohn's disease, studies that look to discover associations with disease status typically focus on individual OTUs while specific microbial association rules found by RF are not discussed, as a result it is uncertain how heterogeneous study cohorts are. To investigate potential rule heterogeneity in a human microbiome cohort we downloaded processed files from the Human Microbiome Project for inflammatory bowel disease (IBD) [130] which contain information on the taxonomic profiles of 982 OTUs in 178 patients – 86 of which have been diagnosed with Crohn's disease, 46 diagnosed with ulcerative colitis, and 46 diagnosed as non-IBD. We were specifically interested in finding rules that separate the Crohn's disease samples from ulcerative colitis and non-IBD, so we framed the problem as a binary classification task with Crohn's disease as the target phenotype.

Since the current implementation of BowSaw is limited to finding rules when the variables have categorical values, we first converted the OTU counts of each taxon to a simple presence/absence scheme. This resulted in nearly equivalent RF performance relative to training RF with the original continuous OTU inputs: ROC AUC of .862 (binary) vs .882 (continuous) and PR AUC of .846 (binary) vs .886 (continuous) (Figure 3.3A-B). This is an important result because it allows us to think about associations just in terms of presence or absence of an OTU without sacrificing much in model performance. We applied BowSaw to the Crohn's disease samples and visualized 56,902 resultant sub-rules ranging in complexity from 2 to 7 variables (Figure 3.3C). There were

1,941 sub-rules with $\boldsymbol{F} = 1$. We selected the most general of these rules ($\max|\boldsymbol{I}|$) to be the top candidate for the curate algorithm and found that it considers the status of 5 OTUs and accounts for 38 of 86 Crohn's disease samples (Figure 3.3C). We set an association threshold of 90% and ended up with 10 sub-rules that together account for all 86 Crohn's disease samples and an additional 11 non-Crohn's disease samples (4 non-IBD, 7 ulcerative colitis). The top five rules combine to account for 78 of 86 Crohn's disease samples and include 10 non-Crohn's disease samples (Table 3.1).

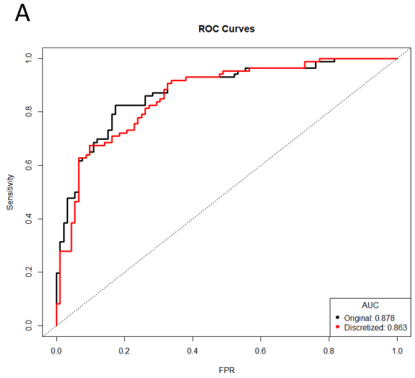
Rule	CD Samples	Non CD Sample	New Samples Covered	Taxonomy	Presence
1	38	0	38	*Bacteroides (genus)*	y
				Lachnolostridium (genus)	y
				Tyzzerella (genus)	n
				Lachnospira (genus)	n
				Lachnospiricae UCG-001(genus)	n
2	41	4	20	*Dialister (genus)*	y
				Christensenellacea R7 group (genus)	n
				Christensenellacea R7 group (genus)	n
				Collinsella (genus)	n
				Ruminococcaceae (family)	n
				Finegoldia (genus)	n
				Ruminococcus 1 (genus)	n
3	9	1	9	*Ruminococcus 1 (genus)*	y
				Ruminococcaceae UCG-002 (genus)	n
				Lachnospiraceae (family)	n
4	24	2	6	*Streptococcus (genus)*	y
				Tyzzerella (genus)	n
				Lachnospiraceae (family)	n
				Hafnia Obesumbacterium	n
5	27	3	5	*Lachnospiraceae UCG-008 (family)*	y
				Ruminococcus 1 (genus)	n
				Eubacterium eligens group	n
6	5	0	2	*Ruminococcus 1 (genus)*	y
				Dorea (genus)	n
7	7	0	2	*Bacteroides (genus)*	y
				Dialister (genus)	n
				Eubacterium rectale group	n
8	15	0	2	*Lachnospiraceae NK4A136 group*	y
				Eubacterium eligens group	y
				Tyzzerella (genus)	n
				Christensenellacea R7 group (genus)	n
				Lachnospira (genus)	n
9	3	0	1	*Ruminococcus gnavus group*	y
				Veillonella (genus)	n
				Bacteroides (genus)	n
				Finegoldia (genus)	n
10	10	1	1	*Parabacteroides (genus)*	y
				Eubacterium eligens group	y
				Ruminococcaceae UCG-003 (genus)	n
				Lachnospiraceae ND3007 group	n

Table 3.1 Association rules identified by BowSaw that account for all Crohn's disease samples.

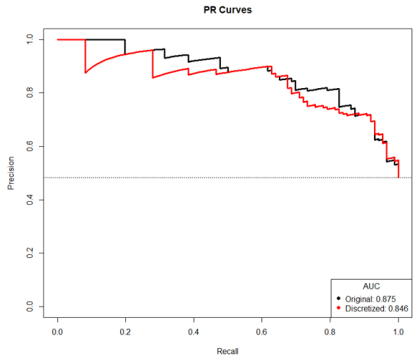
The top candidate rule is comprised of the presence of *bacteroides* and *lachnoclostridium* and the absence of three genera from the family *lachnospiraceae: lachnospira*, *tyzerrella,* and *lachnospiracea UCG 001* (Figure 3.3D). Detection of *bacteroides* was nearly

ubiquitous within the cohort, it was found in 170 of 178 total samples, but only 3 of the samples in which it was missing are diagnosed as Crohn's disease. *Lachnoclostridium* was frequently found in Crohn's disease (67/86) but not in ulcerative colitis (27/46, p = .02) and was detected at roughly the same rate in non-IBD samples (34/46, p = .616). Detection of *lachnospira* was depleted in Crohn's disease samples (20/86) relative to ulcerative colitis (20/46, p = .022) and to non-IBD samples (31/46, p = 9.9^{-7}). *Tyzzerella* was also detected at a lower rate in Crohn's disease (63/86) relative to ulcerative colitis (24/46, p = .019) and non-IBD (24/46, p = .019). *Lachnospiracea UCG 001* was rarely detected in Crohn's disease (4/86) which is a lower rate than it was detected in ulcerative colitis (9/46, p = .022) and in non-IBD samples (19/46, p = 1.45^{-5}).

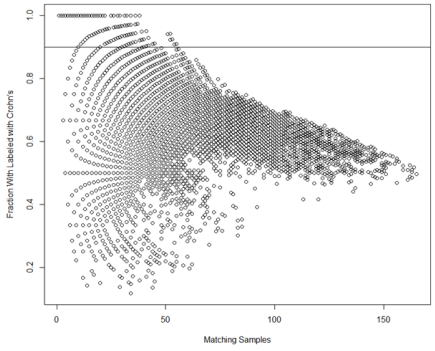
A



B



C



D

Taxonomy	Present?
Bacteroides (genus)	y
Lachnolostridium (genus)	y
Tyzzerella (genus)	n
Lachnospira (genus)	n
Lachnospiricae UCG-001 (genus)	n

Figure 3.3 A) Performance of the random forest classifier as measured by area under the receiver operator curve (ROC-AUC) is not strongly perturbed by simplifying OTU representation to a presence/absence scheme versus the original continuous count. Dashed line indicates the performance of a perfectly random classifier. B) The area under the curve of the precision recall curve is similarly not strongly affected by the new representation scheme. Dashed horizontal line is the random performance line. C) Each point represents a unique candidate sub-rule. On the x-axis is the number of samples in the data matrix that are subject to that rule. The y-axis represents what fraction of matching samples were diagnoses as Crohn's disease. D) The taxon identities of the OTUs that make up the most generally applicable of the sub-rules where all matching samples have the Crohn's disease label.

Discussion

Interpretation of random forest models for classification may be confounded when there are multiple rules (combinations of variables and their specific values) associated with a phenotype of interest. We have developed BowSaw, which is an algorithmic approach for identifying the rules that a trained random forest model uses to make classifications when the values are categorical in nature. By taking advantage of the structure of trees found within a random forest BowSaw produces a set of multiple decision rules that combine to account for each sample with a given observed phenotype. When the variables are the presumed causal agents, these rules represent plausible mechanistic relationships.

Results on simulated data demonstrate that when there are multiple rules associated with a single phenotype label that BowSaw is capable of faithfully identifying them. Application to data from the human microbiome project offers further evidence that BowSaw provides an efficient way of generating plausible hypotheses for high through put metagenomics studies. In particular we identified a rule that utilizes a presence/absence pattern of five microbial taxa (present: *bacteroides, lachnoclostridium*,

absent: *lachnospira,lachnospiracea, tyzerrella*) that accounts for nearly half of all Crohn's disease samples in the cohort (38/86). This specific pattern of microbial colonization in the guts of Crohn's disease patients is unreported, but each taxon's respective enrichment or depletion status and association with disease status has been reported. If the cohort of patients in the human microbiome study are representative of all people afflicted by Crohn's disease then this rule represents a significantly large sub-set of those suffering. Inquiries into the relationship of the taxa included in this rule with disease status may yield important insights into the mechanisms of the disease and potential therapeutic strategies for this sub-population. Of the five associated taxa, we suspect that the absence of *lachnospira, lachnospiracea UCG 001,* and *tyzzerella* are biologically meaningful. We have reason to believe so because it has been reported that the *lachnospiraceae* family is generally suppressed in Crohn's disease [147–149]. *Lachnospira* has been reported as depleted with respect to Crohn's disease several times [150, 151]. The depletion of *tyzzerella* has been associated with chronic intestinal inflammation and supplementation suggested as a probiotic for Crohn's disease [152, 153]. While the relationship of *lachnospiracea UCG 001* with Crohn's disease is still unclear, its depletion has been reported in mice displaying symptoms of anhedonia and it was significantly enriched in anhedonia resilient mice [154]. Partly because IBD is frequently accompanied by depression, anhedonia has been suggested as an important symptom in the diagnosis of IBD [155]. The associations of the individual OTUs defined by this rule are consistent with previously reported findings in the existing literature and describe a taxonomic profile that exclusively identifies a large sub-population of Crohn's

disease samples within this cohort. The presence of *bacteroides* does not appear to be particularly useful and in this context is probably preserved because it causes a perfect association, although high levels of some species are implicated in the pathology of Crohn's disease [156]. *Lachnoclostridium*, is differentially distributed across the three classes. Notably it is less frequently detected in ulcerative colitis relative to Crohn's and non-IBD samples, which roughly resemble one another. Increased levels of this genus was detected in rats that showed relief of colitis symptoms after treatment with a proposed therapeutic agent [157].

The current implementation of the algorithms are restricted to classification tasks with categorical predictor values, this is a challenge that we will need to address in order to make the approach more generally applicable. Future work will also focus on extending these for the interpretation of regression models. Such additions will greatly increase the number of systems to which we can apply BowSaw.

CHAPTER FOUR

A Strategy for Identifying the Presence of Microbial Interactions in Mixed Consortia and Quantifying Them

Summary

This thesis chapter will be submitted as the following article:

Dimucci, D., Bhatnagar, J., Segrè, D. *Identification and Quantification of Microbial Interactions in Synthetic Consortia.* Manuscript in preparation

Abstract

Consortia of microbes can be constructed combinatorically in order to achieve the dual goals of predicting emergent community level functions and identifying interactions between species. Identifying the functional relationships between species as they relate to the measured quantity provides a foothold for further mechanistic studies and the eventual rational engineering of communities to perform desired functions. Although combinatorial studies of microbial communities is common, the identification of interactions between species in those communities is often neglected. Here, we applied simple additive regression models to two such published studies. We first use additive assumptions to build predictive models of net community productivity in a system of soil bacteria and a system of wood rot fungi. We analyze patterns in the residual errors generated by the additive assumption in order to identify modules of interacting microbes and subsequently quantify their statistical significance. We were able to identify high

order interactions between microbes that systematically bias community function and generate new hypotheses regarding the role of interspecies interactions in these systems.

Introduction

Productivity of a microbial community is a function of its species composition. Studies measuring the productivity of microbial communities as a function of their composition often ignore the effects of interspecies interactions or consider them to be inconsequential [57, 158]. The decision to omit interaction effects from models of community productivity can be justified by the fact that main effects typically explain the bulk of variability in experimental observations. Interspecies interactions can have profound effects on community level attributes [5, 73, 108, 159, 160] and for this reason they should be thoroughly examined in any study.

The assembly of microbial consortia betrays an implicit appreciation for the importance of interspecies interactions. For example, the emergence of “bugs as drugs” therapies rely on interactions between bacteria to treat diseases via the gut [161, 162], it is understood that ecosystem functioning is critically dependent on interactions [163], and interactions can drastically improve community productivity [35, 164, 165]. In spite of these observations, analytical strategies tend to either only assume additivity or they fail to identify statistical interactions found by their chosen modeling algorithm.

Incorporating the appropriate interaction terms in our models will improve both our ability to predict community traits and advance our mechanistic understanding of the relationships causing those traits. Once one or more interactions between species have been identified, incorporating the correct interaction terms into models is trivial. However, even the detection of pairwise interaction terms can be a significant challenge [133, 141, 166], particularly when dealing with synthetic microbial consortia because there is rarely an *a priori* rationale for including interaction terms. Knowledge of which pairwise interaction terms to include in studies of microbial communities is commonly derived from an exhaustive set of pairwise experiments [81, 82] although these relationships may not be relevant in larger consortia [167].

An effective strategy for modeling community productivity as a function of community composition is linear regression. In its simplest form, only assumptions of additive effects are made. In practice researchers often keep only the terms for the main effects that are statistically significant in order to generate a parsimonious model [57, 168]. Analysts employ this approach because prediction is often the primary or only objective of these studies. However, the absent discussion of interspecies interactions in these studies prompted us to ask the question, how can interactions between microbial species be efficiently identified? The ideal strategies for identifying significant interaction terms is to exhaustively evaluating all possible interaction terms, but this approach rapidly becomes infeasible in high dimensional settings. For the high-dimensional case it has been shown that tree based models such as random forest can identify important

interactions between variables [169–171] but tree-based ensembles are notoriously difficult to interpret [172].

Here we systematically analyze the residual errors from linear regression models fit to two publicly available data sets of synthetic microbial consortia. We identify sets of interacting species within the trees and evaluate their statistical significance. In both data sets, we find evidence of significant interactions between two or more species. Inclusion of the corresponding interaction terms significantly improves predictive performance. Our results indicate that evaluation of interspecies interactions should be a standard analytic step in any study of microbial communities.

Results

Evidence of Interactions

If interspecies interactions are present, we expect that their associated residuals would show a non-random bias. For example, a simulated system with a positive pairwise interaction between species *A* and species *B* produces biased results even though the distribution of all residuals appears unbiased in visual diagnostic plots. Provided the effect is strong enough to overcome noise, this interaction would result in an enrichment for positive valued residuals whenever both species are present in a community ($A = 1$, $B = 1$). To test this hypothesis we first convert all residuals into classes based on their sign. Then, if the feature space is a manageable size, we can exhaustively evaluate the strength of the association of every pairwise motif with any class. In this context, a motif is a

specific pattern of presence or absence for a subset of microbial species. We add the interaction terms for the species involved in motifs that are significantly associated with any type of residual to the linear model and quantify their effects. We apply this evaluation to motifs of increasing size so long as it is computationally feasible. The number of possible motifs to evaluate grows roughly at the rate of $3^{\#species}$, thus even for moderately sized systems more sophisticated machine learning approaches will need to be employed for motif identification.

Analysis of communities comprised of six soil bacteria

In the first data set we analyzed, the authors of the original study collected 6 soil microbial strains and inoculated all 63 possible combinations into a liquid growth medium with xylose as the sole carbon source [164]. Each species was giving a strain designation (SL-68, SL-104, SL106, SL187, SL-197, SL-WC2) for simplicity we refer to each species by the index of its name in this list. For each community, they estimated metabolic activity with a colorimetric assay. Since they only reported one reading for each community combination, we proceed with our analysis as though the reading is the true population mean.

We fit a regression model to the measured xylose oxidation using just the presence or absence of each species in communities as predictors with ordinary least squares regression (OLS). The linear model produced a good fit to the data (adjusted $r^2 \sim .8$, figure 4.1A). Homoscedasticity and lack of obvious bias of the residuals in relation to the

fitted values indicates that the model fit was good as does their normal distribution (Figure 4.1 B,C). Next, we produced a learning curve to evaluate how the average predictive performance of the linear model would be affected by data availability in terms of r^2. For this system, there is a significant improvement in predictive performance when the number of randomly selected samples is increased to 13 from 9, at which point the expected r^2 is .66 (95% CI .46 - .87). These results indicate that using about 1/5 of the possible information will result in a model that has an r^2 that is nearly 84% as good as the full fitted model (Figure 4.1D). Plotting the residuals as we did in figures 4.1B and 4.1C provides us with a visual diagnostic that the assumptions made by our model are satisfied. The learning curve in 4.1D gives us a sense of how much data is required to fairly represent the full experimental space.

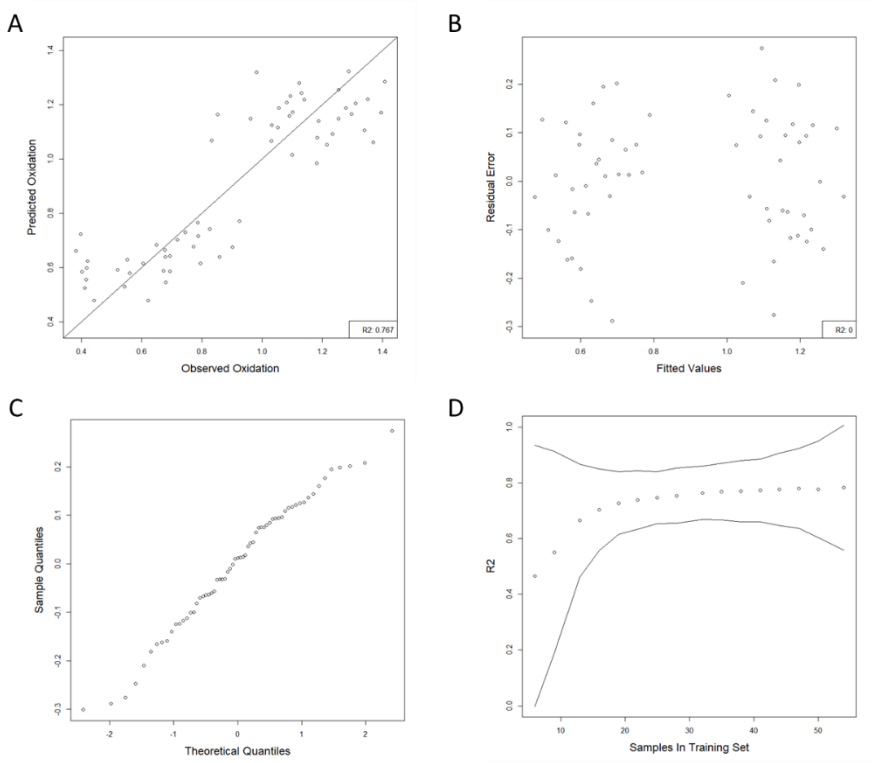


Figure 4.1 Performance of a model with additive assumptions only. (A) An ordinary least squares linear regression model using only the presence or absence of the 6 inputsconstituent species produces a strong fit. (B) Homoscedasticity (equal variance of residuals across predicted values) and the lack of an obvious bias is evidence of a good model fit as is (C) the normal distribution of errors. (D) The predictive performance of the model improves as the number of samples made available for fitting increases. There is a sharp jump in the mean r^2 when the number of samples in the training set increases from 9 to 13. Black lines are the 95% confidence interval.

To search for lurking interactions we first classified the residuals based on their sign (30 negative, 33 positive). Next for each motif in the design matrix, we determined how many of its associated residuals belonged to each class and calculated the probability of that association occurring by chance for a motif of the same size. Due to the nature of this data, even the most extreme associations fail to reach significance after multiple hypothesis correction at $\alpha = .05$. Therefore, we decided to examine in detail those motifs that were fully associated with a single class and had a $p < .01$. We identified five motifs representing three unique combinations of species each involving three species using this criterion (Table 4.1).

$Sp1$	$Sp2$	$Sp3$	$Sp4$	$Sp5$	$Sp6$	$-Residuals$	$+Residuals$
-1	-1	0	-1	0	0	7	0
1	-1	0	1	0	0	0	8
1	1	0	1	0	0	8	0
1	-1	0	0	0	1	0	8
0	-1	0	0	1	-1	0	8

Table 4.1. Motifs fully associated with one residual class. For columns 1 through 6, a 1 indicates the species is present, -1 that it is absent, and 0 that it is ignored.

These microbial combinations suggested that there are potentially several significant three-way interactions. Alternatively, each three-way motif may represent the co-occurrence of two or more significant pairwise interactions. To distinguish between these

possibilities we first quantified the strength of all pairwise interaction effects by calculating how the effect of species *A* on net biomass of the community is affected by the presence or absence of species *B* and then calculating the probability of the interaction effect being non-zero (Methods – Quantification of Interactions). Our analysis showed that there were four significant pairwise interactions – species 2 interacts antagonistically with species 1, 4, and 5 and species 6 interacts negatively with species 4 (Figure 4.2). Then for each unique three species combination, we tested the likelihood that a third species, *C*, modifies the effect of the corresponding pairwise interaction *AB* on net biomass. Although we are able to quantify the strength of the net interaction, it is worth noting that with this approach we are unable to determine the directionality of interactions (i.e. *A* affects *B* while *B* does not affect *A*).

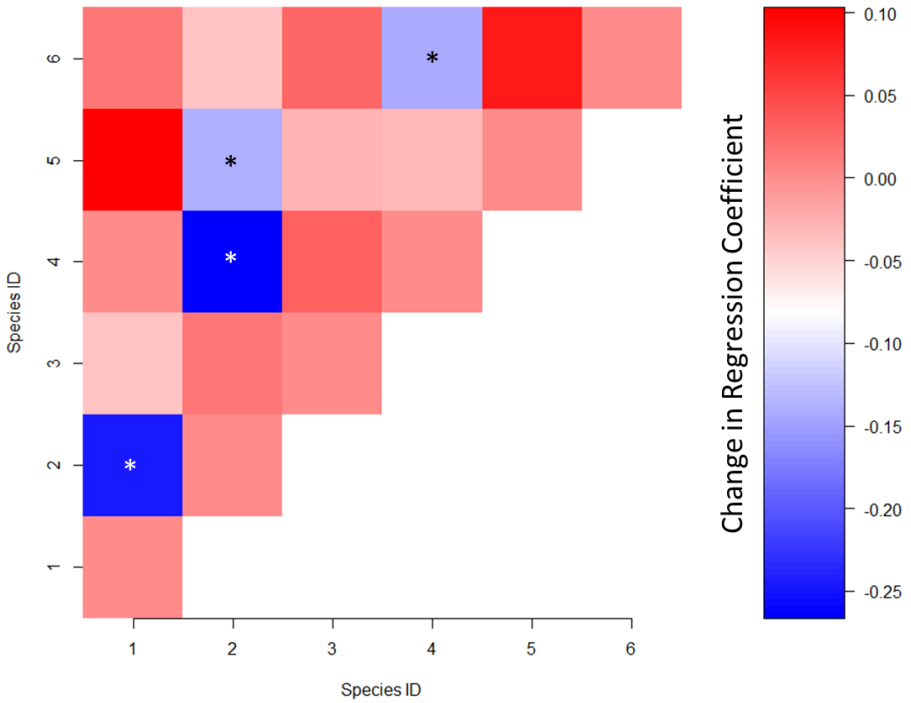


Figure 4.2 Adding an interaction term between two species influences the value of the constituent regression coefficients. From this data, we can determine the net effect of an interaction between two species but not the directionality or relative contributions. Most interaction terms appear to be cooperative (although not significantly different from zero). Species 2 has evidence for negative interactions with three other species {1, 4, 5} as does the interaction between species 4 and 6. * = $p < .05$.

Next, we tested the hypothesis that there are significant interactions among the unique three species combinations involved in the motifs we identified. In order for a three-way interaction to be significant, the addition of a three-way interaction term should affect the coefficients of the two-way interaction terms. For each of these combinations highlighted by the motifs we discovered, we quantified the effect adding a three-way interaction term had on the coefficient for the interaction of the pairwise interactions. None of the three-way interactions effects we investigated were statistically different from zero, leading us to reject the prospect of a large influence coming from three way interactions in this system. We specified a new form of the linear model which included terms for the four pairwise interaction of species. The inclusion of these four interaction terms improved the fit (adjusted $r^2 \sim .92$) and predictive power of the model (predicted $r^2 \sim .9$) (Figure 4.3A). The new model with interaction terms was also more parsimonious than the form without interaction terms ($AIC_{original}$: -60.52, $AIC_{plusInteractions}$: -118.38). The residuals displayed both constant variance and no bias (Figure 4.3 B) but the errors appeared to almost normally distributed (Figure 4.3 C). The learning curve displayed the same shape as it did for the no interaction model (Figure 4.3D).

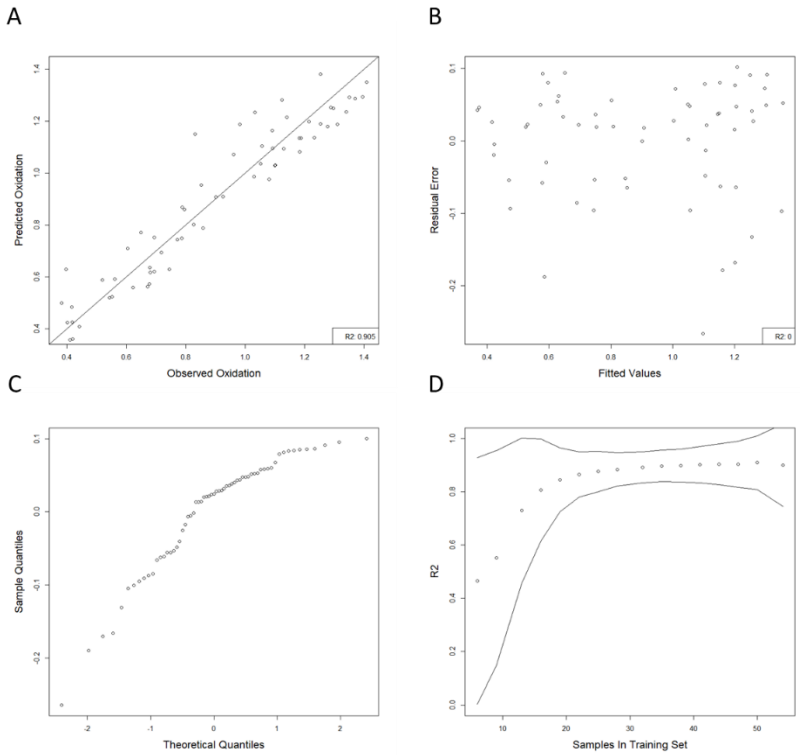


Figure 4.3. **Performance of a model with significant interaction terms.** (A) Leave one out prediction of net xylose oxidation when four pairwise interaction terms were included in the model (2x1, 2x4, 2x5, and 4x6). (B) Residuals of the more complicated model are unbiased and homoscedastic. (C) Residuals of the new model are not quite normally distributed. (D) The predictive performance again shows sharp jump in the mean r^2 when the number of samples in the training set increases from 9 to 13. Predictive accuracy begins to saturate around 20 training samples. Black lines are the 95% confidence interval.

Analysis of communities of 18 fungi

The second data set is a system of 18 wood decay basidiomycete fungi grown in petri dishes by Maynard et al [158]. The authors set up 147 combinations out of a possible 262,144 combinations and assayed them for net biomass production. We fit an OLS model to the measured net biomass using the presence or absence of each species as binary predictors. The performance of the model was weak in contrast to what we observed with the bacterial communities, but still useful (fitted r^2 ~ .36 Figure 4.4A). The

residuals display a random pattern (Figure 4.4B) and are normally distributed but with a long positive tail (Figure 4.4C). The learning curve shows a steadily increasing r^2 as the number of training samples increases, but there is not an obvious elbow along it (Figure 4.4D).

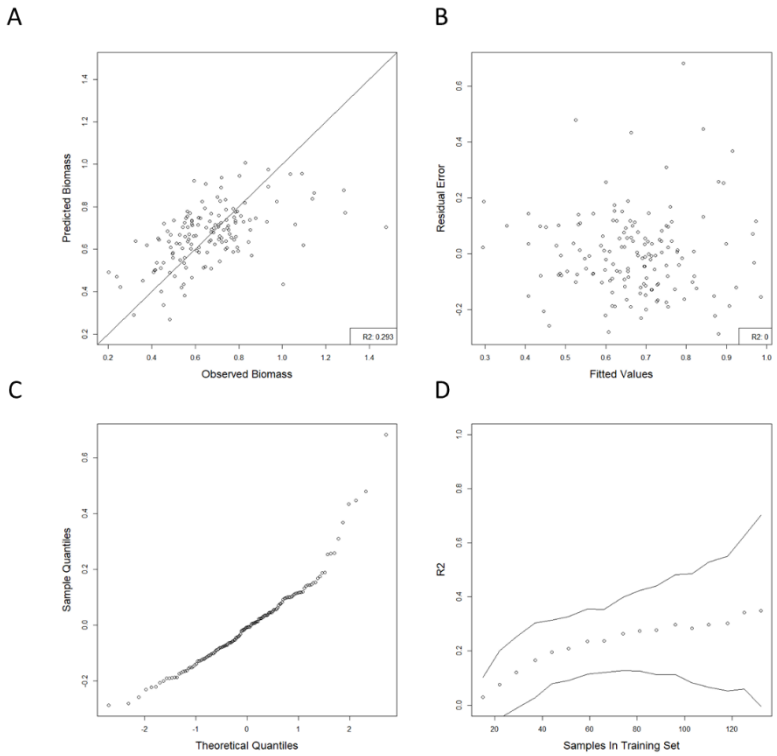


Figure 4.4. (A) The fitted values against observed net biomass for communities of 18 fungal species. (B) Residuals appear to be unbiased with even variance across fitted values. (C) Residuals also appear to be normally distributed, indicating that the assumptions of the linear model are satisfactory. (D) The trajectory of the learning curve is smooth.

We evaluated the strength and likelihood of all 153 pairwise interactions. We found five significant pairwise interactions (non-overlapping 95% confidence intervals), two net positive interactions and three net negative interactions (Figure 4.5). Species 2 was involved in both positive interactions -one with species 1 and one with species 6. Species

14 was involved in all three negative interactions, its interaction partners were species 11, 15, and 16. Because of the overlap between two-way interactions, we added the triplet combination of species 1, 2, and 6 to the set of prospective high order combinations we wanted to evaluate along with the four-way combination of species 11, 14, 15, and 16. We next searched the higher order combinatorial space to identify motifs of species that stand out. For 18 species, there are greater than $387x10^6$ motif combinations that could occur in the design matrix. Since we do not normally expect higher order interactions to be significant and the evidence that supports them becomes scarcer as they become more complicated we exhaustively evaluated all motifs involving up to four species. We fully investigated only the most widely applicable motifs that were fully associated with one residual class (restricted to motifs with greater than 10 instances). We identified three candidate motifs this way, each involving the presence or absence of 4 species. Along with the two combinations of species found when we quantified pairwise interactions, we now had five high order combinations to evaluate for the presence of interactions (Table 4.2).

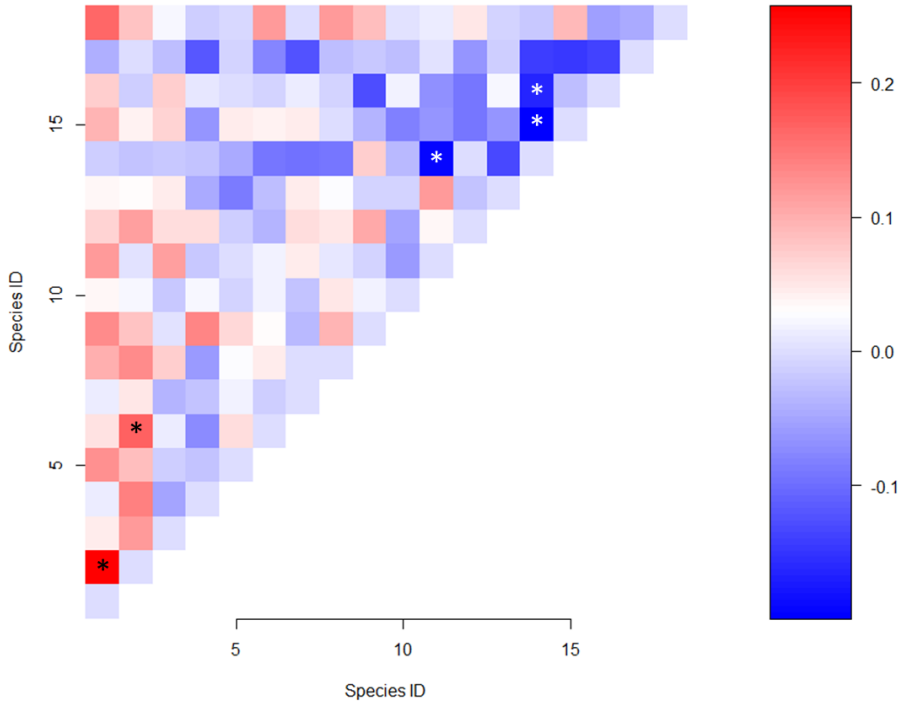


Figure 4.5 Adding an interaction term between two species influences the value of the constituent regression coefficients. Interaction terms appear evenly split between antagonistic and cooperative (81 positive, 72 negative). Species 2 has evidence for positive interactions with species 1 and 6. Species 14 has negative interactions with species 11, 15, and 16. * = p < .05 and 95% confidence intervals are non-overlapping.

$Spc1\ ID$	$Spc2\ ID$	$Spc3\ ID$	$Spc4\ ID$
1	2	5	9
1	2	9	15
4	7	14	17
11	14	15	16
1	2	6	na

Table 4.2. Combinations of species identified for in depth analysis.

For each triplet combination within the larger sets, we calculated the probability of a third species affecting the interaction between two species. We did the same for the full four species interactions in the applicable combinations by measuring the effect of the fourth

species on the coefficient for the three species interaction. There was no evidence for significant influences occurring in any of the three way interactions nor most of the 4 species interactions. However, we found evidence supporting the existence of a net positive 4-way interaction involving species 1, 2, 5, and 9 ($p = .035$). This combination is one of the three combinations we identified via motif analysis. A new linear model that incorporated the 4-way interaction and the five significant pairwise interactions we identified earlier resulted in an improved performance (predicted $r^2 \sim .488$) (Figure 4.6A). The new model was significantly parsimonious compared to the original additive model ($AIC_{original}$: -102.74, $AIC_{plusInteractions}$: -144.8). The residuals follow the same patterns as they did for the no interaction model (Figure 4.6B,C). The learning curve displays an upward bending elbow between 29 and 37 training samples then smoothly increases until the full data set is used (Figure 4.6D). It appears that given our choice of representation, more observations are needed in order to reach a saturation point, signaled by a sharp decrease in the rate of improved prediction. While none of the 3-way interactions that could be made from the set of species 1, 2, 5 and 9 was statistically significant, they all produced net negative interactions, which are the likely causes of the motif's strong association with negative residuals.

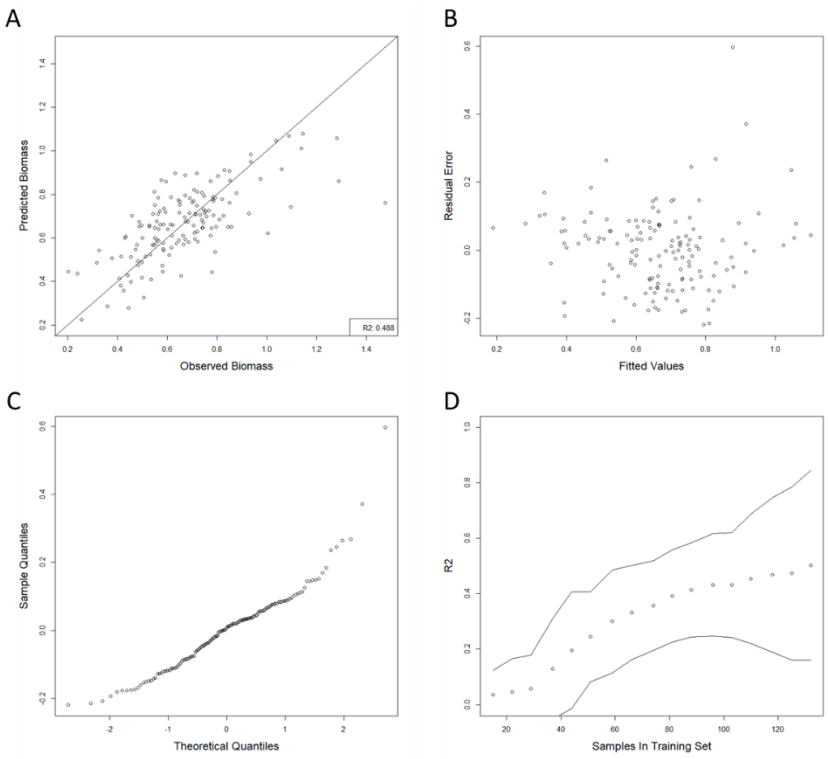


Figure 4.6 (A) Leave one out prediction of net biomass when five pairwise interaction terms and the order-four interaction term are included in the model. (B) Residuals of the more complicated model are unbiased and largely homoscedastic. (C) Residuals of the new model are not quite normally distributed. (D) The predictive performance shows an inflection point between 29 and 37 training samples when a sharp upward trend in r^2 begins. After this point, the smooth monotonic growth in performance indicates that there more samples are needed in order to reach predictive saturation.

Discussion

High throughput experiments of hundreds and sometimes thousands of synthetic microbial consortia are becoming commonplace [35]. Building reliable predictive models for the functions of these communities is the first step to being able to engineer them. While prediction of community functioning is a powerful tool, we will also need to understand the nature of the interactions within communities that cause community phenotypes. By identifying and quantifying interspecies interactions in synthetic

consortia, we will be able to build models with improved predictive performance. Not only will our predictive powers increase, but also by better understanding the statistical interactions that are present in our experimental systems, we will be able better guide our search for the mechanisms of interactions.

Here we have demonstrated how to use a motif analysis of residual errors to identify putative high order interactions. We further demonstrated how one could use regression models methods to quantify the strength of interactions contained within high order combinations of microbes and test their significance. Our results on the xylose oxidizing soil bacteria reveals that even when additive assumptions are sufficient to build a model with good predictive properties that a detailed analysis of patterns associated with error types can reveal significant interactions that would otherwise be missed. Adding these interactions to the linear regression model resulted in a significant increase in terms of predictive power at a relatively cost in terms of extra parameters.

The results of our analysis on the 18 fungal species community data set illustrates the potential impact a motif analysis of residuals can have. When the combinatorial space is too large to exhaustively search, limiting the scope of the inquiry can still reveal interesting species combinations worthy of in depth investigation. In this data set, we found three motifs involving four fungal species each that were fully associated with negative residual errors – that is to say, whenever these particular combinations of variables were present the model overestimated the net productivity of the community. Of these we found none of the 3-way interaction terms to be significant but interestingly

found that one of the 4-way interaction terms was and that the sign of its coefficient was opposite to the sign of its constituent 3-way interactions.

To date most of the experimental studies we have found in the literature that try to predict community functions as linear functions of the constituent species neglect to search for significant interactions. Feature selection with methods such as LASSO [173] and stepwise regression [57] are often used to produce a parsimonious model with high predictive accuracy that is also amenable to interpretation. While predictive power is not adversely affected by these methods, they introduce significant biases into the parameter estimates [174]. If the goal of the study is interpretation of the coefficients then the coefficients for all species in the experiments should be kept, even when their coefficients are not significant, and when an interaction term is being evaluated all of its constituent terms should be as well [174].

This study focused on identifying and quantifying interspecific interactions across many synthetic consortia. However, in the future we should combine these methods with genomic analyses to develop plausible hypotheses regarding the biological mechanisms driving the interactions.

Methods

Data Sets: The xylose oxidation dataset was produced by Langeheder et al [164]. We obtained the design matrix and measurements data table from the supplemental files of

Jaillard et al [175]. The design matrix and measurements for the experiments with 18 fungal strains were obtained from the supplemental files from Maynard et al [158].

Linear Regression

To build a predictive model of community productivity as a function of community composition of *n* species we specified a linear regression model according to the equation:

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \varepsilon$$

Where y is the observed measurement (i.e. net oxidation of xylose, or net biomass), β_0 is the average community productivity, each β_i coefficient is the per-unit effect X_i has on y, each X_i is a binary variable indicating the presence or absence of species, and ε is the residual error term. We used ordinary least squares to estimate the values of the coefficients. When including interactions between species we add a new β_i coefficient and a new X_i term that is the product of the species in the interaction and increment n by one.

Cross-Validation

To obtain an estimate of the predictive utility of a given regression model we performed leave-one-out cross validation. In this process, we fit a model to all but one sample in the data set and then predict the value of that observation. We repeat this step until every sample in the data set has a predicted value. We then calculate the Pearson correlation of the predicted and observed values, which we report as the predicted r^2.

Learning Curves

To build learning curve we defined a set of fractions, *f = {.1, .15, .2, .25, .35, .4, .45, .5, .55, .6, .65, .7, .75, .8, .85, .9, .95}*, which we used to define the number of samples we allocate to the training set in order to evaluate test error as a function of data availability. For each fraction in *f*, we fit a regression model to 100 randomly selected sub-sets of the corresponding size and held the omitted data separate to use as an independent test set. We recorded the r^2 for each iteration of predictions.

Quantification of interactions

We adopt the methods and practices described by [174] for interpretation of interaction effects in linear models. When there are no interaction terms included in a linear regression model then we may interpret the coefficients associated with each X_i as the unconditional marginal effect of that species' presence on the quantity of interest. When an interaction term is included, such as in the simple two variable system below this interpretation can no longer be used because β_1 is now conditionally dependent on the value of *Z.*

$$y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon$$

When $Z = 0$ the above equation becomes:

$$y = \beta_0 + \beta_1 X + \varepsilon$$

In this form it is clear that β_1 captures the per unit effect of *X* on *y.* We can re-write equation (1) to capture the effect of *X* on *y* when *Z* is present ($Z = 1$):

$$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)X + \varepsilon$$

The contribution of *X* on *y* is represented by its marginal effect:

$$\frac{\partial y}{\partial X} = \beta_1 + \beta_3 Z$$

The standard error of the parameter estimate is

$$\hat{\sigma}_{\frac{\partial y}{\partial X}} = \sqrt{var(\hat{\beta}_1) + Z^2 var(\hat{\beta}_3) + 2Zcov(\hat{\beta}_1\hat{\beta}_3)}$$

With these equations, we derive a confidence interval for the effect that *X* has on *y* conditional on *Z*. When investigating an interaction effect all constitutive terms of the interaction must be included in the model, e.g. when the interaction being investigated is *WXZ* then the terms *W, X, Z, WX, WZ, XZ* should also be included in the model in order to obtain an unbiased estimate of the interaction effects. We implemented these calculations with the interplot package for R [].

To determine the significance of an interaction with overlapping confidence intervals we calculate the t-statistic:

$$t = \frac{\hat{\beta}_{1z=1} - \hat{\beta}_{1z=0}}{\sqrt{\hat{\sigma}^2_{\beta_{1z=1}} + \hat{\sigma}^2_{\beta_{1z=0}}}}$$

Definition of motifs

For our purposes, we define a motif to be a specific combination of variables and values found within the design matrix. For example, for a design matrix of three binary variables eight unique vectors can be constructed:

$V1$	$V2$	$V3$
0	0	0
1	0	0
0	1	0
0	0	1
1	1	0
0	1	1
1	0	1
1	1	1

A motif need not contain all of the variables in the matrix, e.g. $\{V1 = 0, V3 = 1\}$. The number of motifs that can be found in a complete binary matrix of n variables is roughly 3^n.

Identification of potential high-order interactions

In order to identify potential combinations of variables that are interacting we first convert the residual errors from the fitted model into a set of classes based on their sign. Next, we construct a matrix that stores every motif present in the design matrix. When the number of variables is small, this matrix contains the exhaustive list of all motifs in the design matrix; this was what we did for the six-member soil bacteria data set. When the combinatorial space is unwieldly, we define a cutoff in terms of motif complexity – in the case of the 18 fungal strain data set, we limited our exhaustive search to motifs involving at most four variables.

For each of our curated motifs we find all of the samples within the design matrix that possess that motif – e.g. identify all vectors where $V1 = 0$ and $V3 = 1$. We then enumerate the number of identified vectors that produced negative residual errors and the number that produced positive residual errors. The association of the motif with a particular error types is determined by calculating the cumulative hypergeometric probability of it being associated with that many samples of the given error type if its observed residual errors had been randomly assigned.

CHAPTER FIVE

Discussion

In this thesis, I reviewed the current state of synthetic microbial ecology and some of the applications that motivate research in the area. Throughout much of the field's history, the bulk of scientific research has been on the optimization of single species functions. This was achieved either through modification of the organisms' genomes or by optimization of the environmental growth conditions. These approaches have been extraordinarily successful as evidenced by the abundance of companies dedicated to using them.

The recognition of the importance of the microbiome in human health has resulted in the development of synthetic gut consortia as therapies for various ailments, a "bugs as drugs" approach. These methods necessarily rely on the power of communities to achieve their ends. Identifying and characterizing the interactions in these consortia is very challenging but is seen as a necessity in order to fully understand these communities and ultimately engineer them. A key step along this path is gaining the ability to predict the nature of interactions between microbes.

In chapter two, I described a conceptual framework that we can use for representing interactions between two microbes so that we may apply machine learning methods. To the best of my knowledge, this is the first application of a machine learning approach to

the prediction of interspecies interactions of microbes. As is the case with most machine learning applications, the choice of feature representation is more important than the specific algorithm chosen. The reason for this is that our ability to derive meaningful hypotheses from the trained models is directly affected by how we perceive the features.

In chapter three I built on the concepts of feature contributions in random forests that I introduced in chapter two. The main influenceof feature contributions to the concept that eventually became BowSaw is the idea of following the path individual samples take through the forest and track the frequency of pairwise variable co-occurrences as evidence for pairwise interactions.

The intuition behind BowSaw reflects the action we observe on a Galton board. A Galton board is a device, such as a board, that has many rows of pegs arranged triangularly, beginning with one peg at the top and ending with the greatest number of pegs on the bottom row. When we drop a ball down the board it hits the first peg and then gets sent to either the left or the right where it encounters another peg and this decision is made again. As an analogy to random forest we can imagine that the first peg is the splitting variable that is encountered at the root node of a tree and the subsequent peg is the next variable along the branch. At the first node there is some probability distribution for the likelihood that each variable will be selected as the splitting variable and this is true at the second node and so on. The probabilities are not constant though at each node, in fact the probability distribution at nodes other than the root node are conditionally dependent on

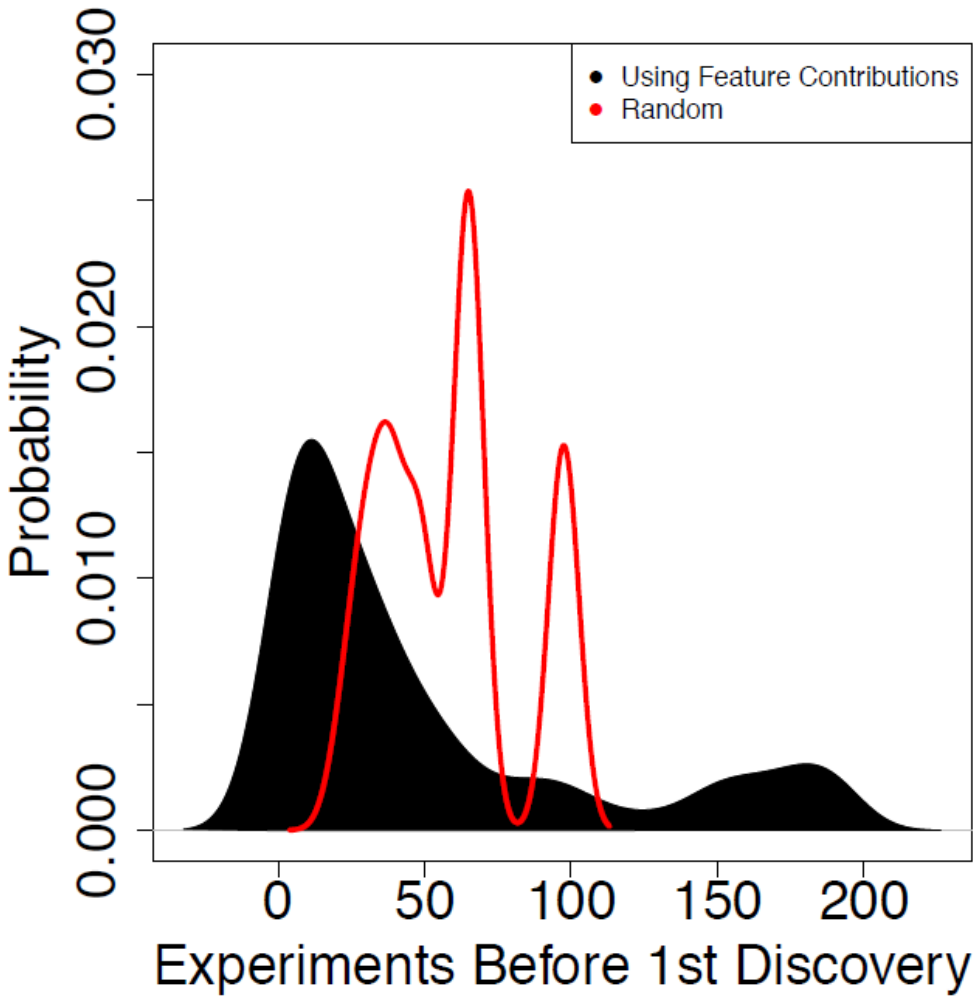
the history of splitting decisions that preceded them. If there is a true positive statistical relationship between two variables {A,B} then a split on variable A will increase the probability of variable B being selected as the child of the A node. We assume that giving many opportunities for this interaction to occur randomly by building many trees will cause the interaction pair A -> B to occur more frequently than a false interaction e.g. A -> C. In this work, I focus on the practical implications of finding high order interactions only. In the future attention should be paid to the statistical properties of using the BowSaw approach in order to formalize the assumptions it makes and better understand the mechanics of random forests.

The current form of BowSaw is subject to some considerable limitations. Firstly, due to its reliance on exactly matching values in order to generate candidate rules it is only meaningfully useful on data sets with discrete or categorical predictors. Second, while the relevant variables are identified, it does not provide any extra insight into the functional relationship between them. Finally, in order to apply it to regression forests we must first convert the forest into a classification forest.

Chapter four is an exploratory analysis of the residuals produced by linear regression and their associated patterns. In this study, I applied an additive linear model to microbial communities in order to predict community function (e.g. xylose oxidation, net biomass). These models appear to be well fit to the data and the assumptions of linear regression upheld. By looking at high order patterns and determining their associations with certain

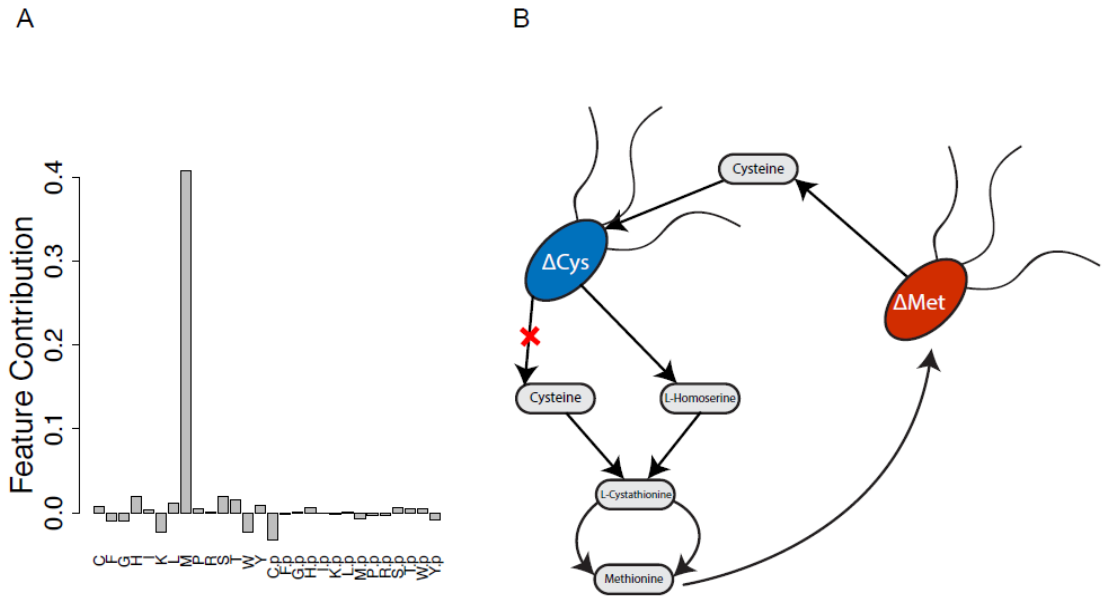
types of errors I was able to identify putative interactions. Statistical evaluation of these interactions revealed evidence that some of these interactions may be real and this manifests in the significant improvements to model predictions when their interaction terms were accounted for. The methods I described in this chapter, while quite simple, lay the foundation for a new line of inquiry in looking for and quantifying interspecies interactions in microbial consortia. This is an important avenue for future development since the current approaches tend to rely heavily on pairwise correlational analysis, which cannot be easily extended to higher order interactions, and are prone to producing many false positives [48].

APPENDIX



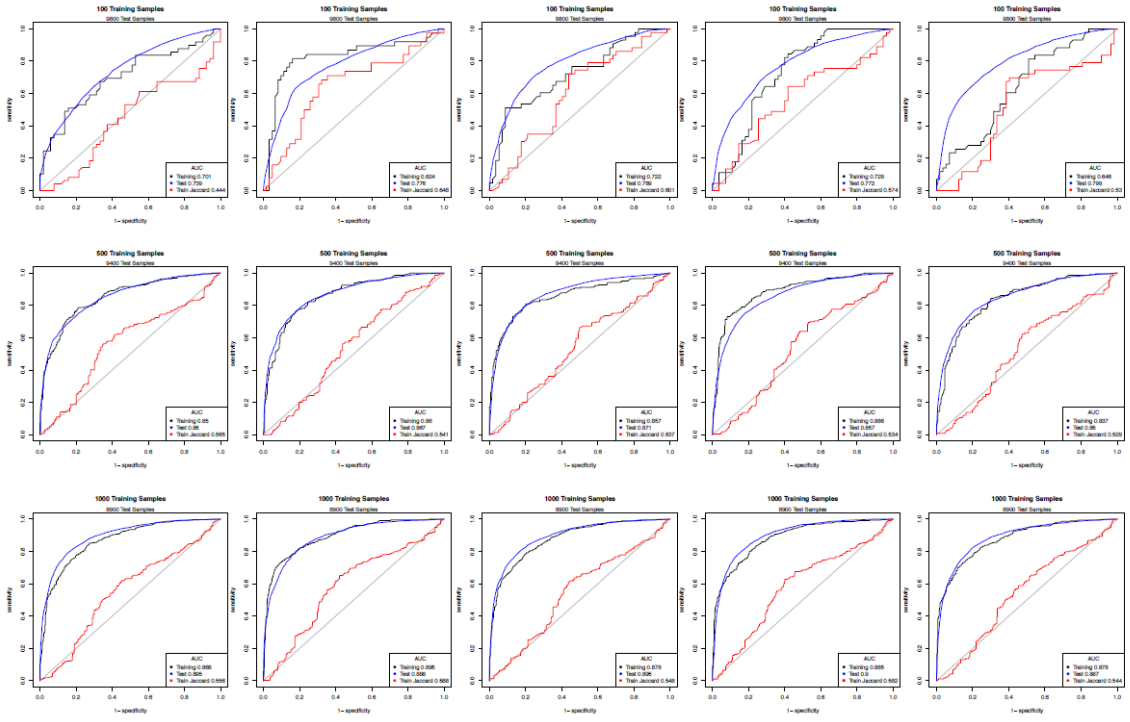
Supplemental Figure A.1. Feature contributions were used to find a facilitative metabolite in samples that had a positive relative yield (RY > 0). The empirical probability distribution of the average rank at which the first facilitative metabolite would be encountered by sampling metabolites randomly one at a time was calculated for each sample and compared to the observed probability distribution obtained from using ranked feature contributions. By chance, the median first metabolite is encountered after 65 queries (mean ≈ 58.7). With feature contributions, the median number of queries was 27 (mean ≈ 50). The number of experiments required by chance to find the first metabolite in a sample is a function of the number of real mechanisms in that sample and is the cause of the observed multimodality. Positive samples were scarce in the *in silico* data set

(420/9,900). A reliable classifier was developed via a balanced training set created by randomly sampling 420 nonpositive samples. This process was repeated 100 times, with an observed median balanced accuracy of ~85%. A single random forest model was then used to calculate the feature contributions for the identification of putative facilitative metabolites.

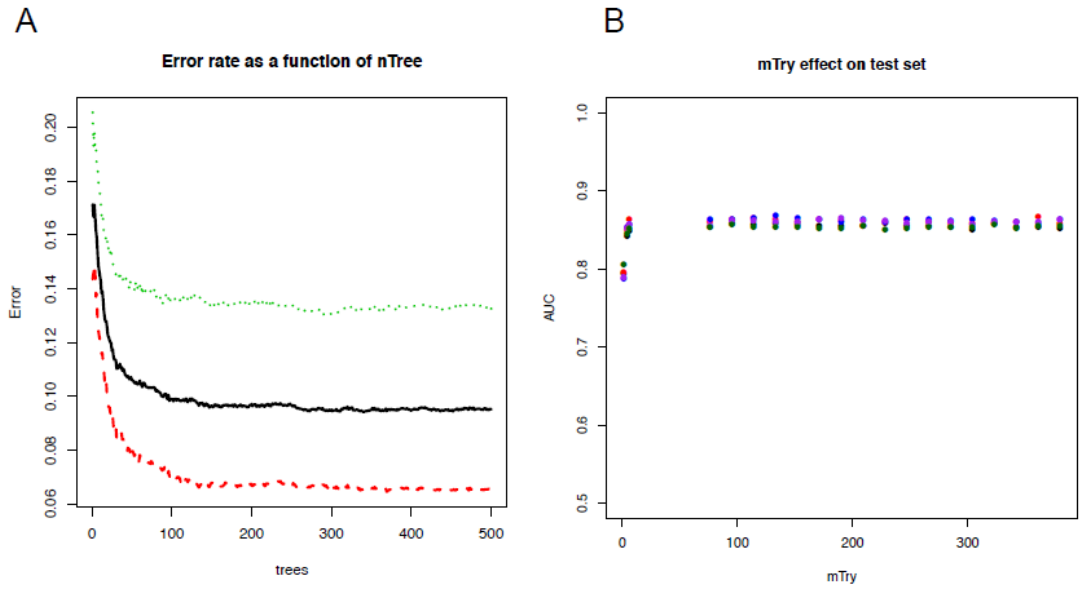


Supplemental Figure A.2. A. Bar plot of the calculated feature contributions for the growth response of a ΔMet mutant cocultured with a ΔCys mutant in the *E. coli* auxotrophs case study. The feature contribution from the receiver methionine is ≈0.41. The feature contribution from the giver cysteine is ≈−0.03. The net contribution from the remaining predictors is ≈0.02. (B) The ΔMet mutant typically had a strong response in coculture no matter the identity of its interaction partner; 12 interactions resulted in strong response type for the ΔMet mutant. When it was grown with a ΔCys mutant, however, it had a weak growth response. The use of feature contributions correctly identified the receiver's methionine and the giver's cysteine as the first and second most important predictors, respectively, in this interaction. The contribution from the receiver methionine is overwhelmingly positive, reflecting the fact that the ΔMet mutant typically benefits strongly from coculture and results in the strong response prediction in panel A. To develop a hypothesis for why this interaction defied the expectations of the random forest, we consulted the literature regarding the biosynthetic pathways for methionine and cysteine and learned that under the specified growth conditions, cysteine is necessary for the biosynthesis of cystathionine. Cystathionine is

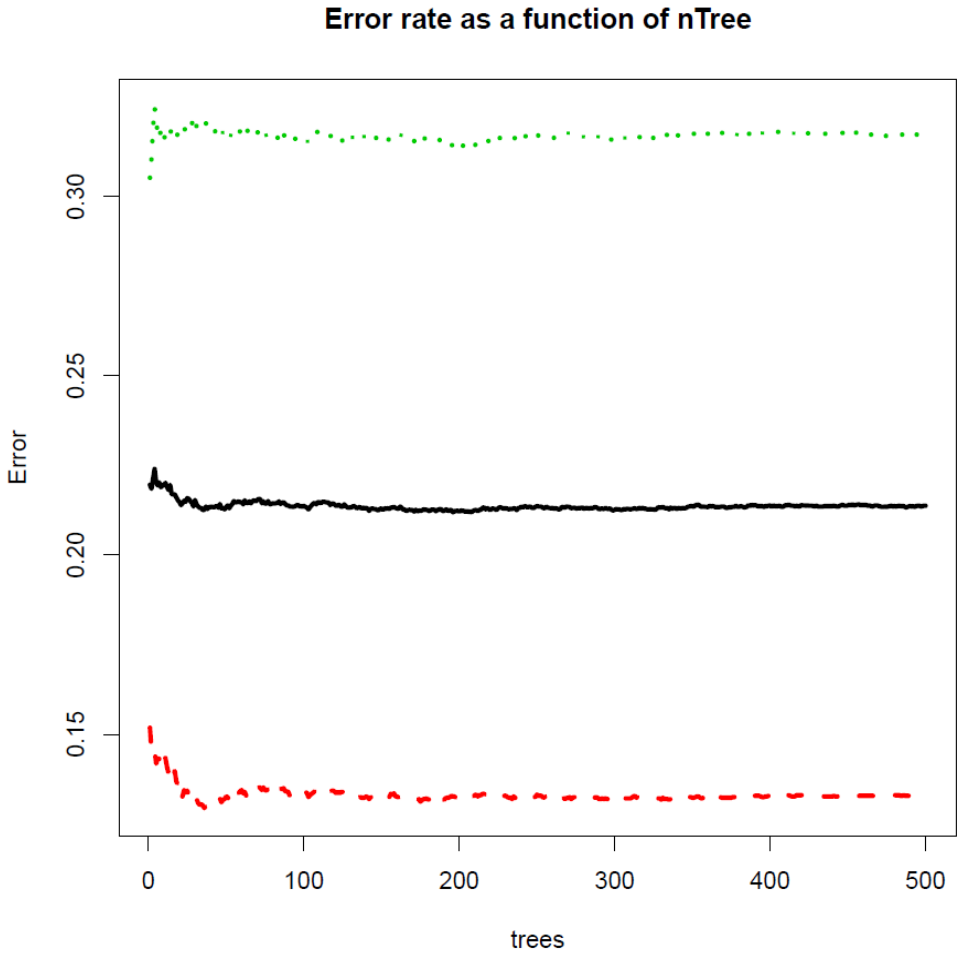
subsequently required for the biosynthesis of homocysteine, which is in turn required for the production of methionine. Given this knowledge, we suspect the ΔCys mutant is unable produce methionine until enough cysteine accumulates in the environment, and the ΔMet mutant must wait for the ΔCys mutant to use its excess cysteine for the production of excess methionine (B). The ΔMet mutant is likely not able to provide an abundance of extracellular cysteine for the ΔCys mutant early in the interaction, because cysteine biosynthesis is tightly regulated due to its toxicity, but (i) efflux of cysteine has been proposed as a potential regulatory mechanism and (ii) likely enables the ΔMet mutant to produce low levels of extracellular cysteine. The waiting time associated with ΔMet mutant-derived cysteine to accumulate in the environment would result in delayed growth for both strains relative to other interactions involving the ΔMet mutant. The reported fold changes in this experiment were 0.8 for the ΔMet mutant and 1.2 for the ΔCys mutant. The weak growth response of the ΔMet mutant is consistent with delayed growth



Supplemental Figure A.3 For 3 subset sizes of the *in silico* dataset (100, 500, and 1,000), the respective number of samples was randomly selected to use for training a random forest. The AUCs of the ROC curves were calculated using the Jaccard distance as a single threshold on the training set, the vote proportions on the out-of-bag samples from the random forest on the training set, and the vote proportions of the random forest on the samples in the test set. This process was repeated 5 times for each subset.



Supplemental Figure A.4 (A) The out-of-bag error rate as a function of nTree for the random forest model trained on the full *in silico* data. Error converges at ~150 trees; the green line is the error rate for nonnegative responses, the red line is the error rate for negative responses. (B) A random subset of 500 samples of the *in silico* data and for the series of possible values of mTry (1, 2, 4, 6, 76, 95, 114, 133, 152, 171, 190, 209, 228, 247, 266, 285, 304, 323, 342, 361, and 380) was selected to see if tuning the hyperparameter mTry resulted in benefits to performance. The AUCs were calculated for the ROC curves obtained from the subsequent class votes. This process was done for 5 random subsets of 500 samples. Points that are the same color in the figure correspond to AUC results from the same series.



Supplemental Figure A.5 Phylogenetic classifications for each metabolic model were used as an alternative set of features. The out-of-bag error for a model trained on the full 9,900 sample data set is ~ 21.18% (black line); the green line is the error rate for nonnegative responses and red line is the error rate for negative responses.

SeedID	Name	Times Found First	Median Rank
cpd00082	D-Fructose	1041	4
cpd11581	gly-asn-L	347	9
cpd11593	ala-L-asp-L	308	8
cpd11589	gly-asp-L	261	8
cpd00053	L-Glutamine	252	3
cpd00039	L-Lysine	245	2
cpd00277	Deoxyguanosine	197	2
cpd00246	Inosine	192	3
cpd00080	G3P	176	2
cpd00654	Deoxycytodine	161	3
cpd00179	Maltose	151	6
cpd00162	Aminoethanol	149	2
cpd00105	D-Ribose	145	2
cpd00047	Formate	142	2
cpd00129	L-Proline	116	6.5
cpd00027	D-Glucose	113	5
cpd00264	Spermidine	99	5
cpd03279	Deoxyinosine	98	3
cpd11585	L-alanylglycine	86	6
cpd00220	Riboflavin	82	5
cpd00023	L-Glutamate	80	4
cpd00516	meso-2,6-Diaminopimelate	75	1
cpd00182	Adenosine	66	3
cpd00367	Cytidine	66	4
cpd00156	L-Valine	59	4
cpd00106	Fumarate	58	1
cpd00122	N-Acetyl-D-glucosamine	58	3
cpd01017	Cys-Gly	51	3
cpd15605	Gly-Phe	47	12
cpd15606	Gly-Tyr	47	11
cpd00322	L-Isoleucine	44	5
cpd01080	ocdca	33	2
cpd00051	L-Arginine	32	2
cpd11586	ala-L-glu-L	29	9
cpd11588	gly-pro-L	28	11.5
cpd00438	Deoxyadenosine	26	6
cpd00159	L-Lactate	24	1

cpd15604	Gly-Leu	23	10
cpd00092	Uracil	20	2.5
cpd00644	PAN	20	6
cpd00108	Galactose	19	3
cpd00311	Guanosine	18	1
cpd00035	L-Alanine	17	2
cpd00118	Putrescine	17	3
cpd00249	Uridine	17	2
cpd11591	Gly-Met	16	2
cpd00054	L-Serine	15	4
cpd00794	TRHL	14	2
cpd00184	Thymidine	13	4
cpd00309	XAN	13	6
cpd11583	Ala-Leu	13	8
cpd00013	NH3	11	7
cpd11582	ala-L-Thr-L	11	9
cpd00036	Succinate	10	1
cpd00137	Citrate	10	7.5
cpd00064	Ornithine	9	12
cpd00117	D-Alanine	8	4
cpd11584	Ala-His	8	9.5
cpd11606	Menaquinone 7	8	4.5
cpd00060	L-Methionine	6	6.5
cpd00393	Folate	6	3
cpd00028	Heme	5	1
cpd00133	Nicotinamide	5	3
cpd00208	LACT	5	3
cpd00226	HYXN	5	10
cpd00305	Thiamin	5	3
cpd00355	Nicotinamide ribonucleotide	5	8
cpd00065	L-Tryptophan	4	7
cpd00076	Sucrose	3	4
cpd00107	L-Leucine	3	7
cpd00066	L-Phenylalanine	2	2.5
cpd00158	CELB	2	14.5
cpd00268	H2S2O3	2	2.5
cpd00276	GLUM	2	5.5
cpd00655	Dephospho-CoA	2	15.5
cpd00024	2-Oxoglutarate	1	1
cpd00130	L-Malate	1	12

cpd00136	4-Hydroxybenzoate	1	3
cpd00161	L-Threonine	1	1
cpd00185	D-Arabinose	1	2
cpd11590	met-L-ala-L	1	98
cpd16336	Isoprene	1	2

Supplemental Table A.1 Top metabolites for which pairs of organisms are predicted to compete based on ranked feature contributions.

Model Seed ID	Times Competed For
cpd15605	3205
cpd15606	3153
cpd01017	3057
cpd11584	2842
cpd11581	2836
cpd00322	2745
cpd11589	2723
cpd00027	2651
cpd00156	2521
cpd11593	2482
cpd00220	2399
cpd11582	2211
cpd00039	1881
cpd00226	1876
cpd00082	1875
cpd00129	1849
cpd00179	1740
cpd11586	1642
cpd11590	1595
cpd00644	1471
cpd00065	1145
cpd11588	1141
cpd00023	1028
cpd00080	931
cpd11583	918
cpd00092	886
cpd00013	871
cpd00107	856
cpd00105	855
cpd00051	835
cpd00367	827
cpd00118	824
cpd00654	800
cpd00053	790
cpd00122	757
cpd15604	729
cpd00182	715
cpd00054	691
cpd00246	685

cpd11585	634
cpd00249	568
cpd00355	560
cpd00794	528
cpd00516	518
cpd03279	489
cpd11591	485
cpd00264	461
cpd00438	443
cpd00277	426
cpd00305	392
cpd00159	389
cpd00106	363
cpd00393	363
cpd00036	339
cpd00117	339
cpd00064	337
cpd00137	325
cpd00276	320
cpd00047	319
cpd00311	296
cpd00108	280
cpd00309	280
cpd00033	279
cpd00162	268
cpd00028	253
cpd11606	201
cpd00069	188
cpd00655	163
cpd00060	162
cpd00184	145
cpd00130	125
cpd00041	121
cpd00035	104
cpd00066	101
cpd01080	82
cpd00307	58
cpd03847	57
cpd00208	55
cpd03198	50

cpd00550	46
cpd00136	30
cpd00133	27
cpd00161	22
cpd00218	22
cpd00158	20
cpd00224	12
cpd00492	12
cpd00268	11
cpd00100	10
cpd00359	10
cpd01217	6
cpd01914	6
cpd16336	6
cpd00132	5
cpd00215	5
cpd00075	4
cpd00076	4
cpd00084	4
cpd00142	4
cpd00176	4
cpd03422	4
cpd08636	4
cpd00239	3
cpd00024	2
cpd00185	2
cpd00793	2
cpd01741	2
cpd00119	1
cpd00314	1
cpd00006	0
cpd00012	0
cpd00079	0
cpd00098	0
cpd00104	0
cpd00121	0
cpd00138	0
cpd00139	0
cpd00154	0
cpd00164	0

cpd00209	0
cpd00210	0
cpd00211	0
cpd00214	0
cpd00216	0
cpd00221	0
cpd00222	0
cpd00232	0
cpd00235	0
cpd00244	0
cpd00266	0
cpd00280	0
cpd00281	0
cpd00298	0
cpd00308	0
cpd00338	0
cpd00357	0
cpd00395	0
cpd00396	0
cpd00412	0
cpd00423	0
cpd00441	0
cpd00531	0
cpd00540	0
cpd00573	0
cpd00588	0
cpd00609	0
cpd00635	0
cpd00637	0
cpd00652	0
cpd00653	0
cpd00681	0
cpd00797	0
cpd00811	0
cpd00870	0
cpd00971	0
cpd01012	0
cpd01015	0
cpd01030	0
cpd01048	0

cpd01092	0
cpd01155	0
cpd01171	0
cpd01242	0
cpd01262	0
cpd01329	0
cpd01912	0
cpd02227	0
cpd03048	0
cpd03343	0
cpd03424	0
cpd03696	0
cpd03724	0
cpd03725	0
cpd04097	0
cpd04098	0
cpd08023	0
cpd08305	0
cpd08306	0
cpd09878	0
cpd11574	0
cpd11575	0
cpd11576	0
cpd11578	0
cpd11579	0
cpd11580	0
cpd11587	0
cpd11592	0
cpd11595	0
cpd11596	0
cpd11597	0
cpd15269	0
cpd15302	0
cpd15603	0
cpd16062	0

Supplemental Table A.2 Numbers of times each of the 194 metabolites were consumed by both organisms in negative interactions. Metabolite cpd00082 corresponds to fructose. Legends for all metabolites can be found at http://modelseed.org/biochem/compounds.

Predictive Rank of Knocked Out (KO) Amino Acids	Total Occurrences (Fraction of samples)
Receiver KO 1st	140 (.769)
Giver KO 1st	40 (.22)
Receiver KO 2nd	35 (.192)
Giver KO 2nd	97 (.533)
Receiver KO 1st/Giver KO 2nd	97 (.533)
Giver KO 1st/Receiver KO 2nd	35 (.192)

Supplemental Table A.3 Counts of how often one or both auxotrophic amino acids were the strongest predictors for *E. coli* interactions.

Module.Name	Module.ID
Sulfate transport system	M00185
Tungstate transport system	M00186
NitT/TauT family transport system	M00188
Molybdate transport system	M00189
Iron(III) transport system	M00190
Putative thiamine transport system	M00192
Putative spermidine/putrescine transport system	M00193
Maltose/maltodextrin transport system	M00194
Raffinose/stachyose/melibiose transport system	M00196
Putative fructooligosaccharide transport system	M00197
Putative sn-glycerol-phosphate transport system	M00198
L-Arabinose/lactose transport system	M00199
Putative sorbitol/mannitol transport system	M00200
alpha-Glucoside transport system	M00201
N-Acetylglucosamine transport system	M00205
Cellobiose transport system	M00206
Putative multiple sugar transport system	M00207
Glycine betaine/proline transport system	M00208
Osmoprotectant transport system	M00209
Phospholipid transport system	M00210
Putative ABC transport system	M00211
Ribose transport system	M00212
D-Xylose transport system	M00215
Multiple sugar transport system	M00216
AI-2 transport system	M00219
Putative simple sugar transport system	M00221
Phosphate transport system	M00222
Phosphonate transport system	M00223
Putative glutamine transport system	M00228
Glutamate/aspartate transport system	M00230
Glutamate transport system	M00233
Putative polar amino acid transport system	M00236
Branched-chain amino acid transport system	M00237
D-Methionine transport system	M00238
Peptides/nickel transport system	M00239
Iron complex transport system	M00240
Zinc transport system	M00242
Manganese/iron transport system	M00243
Putative zinc/manganese transport system	M00244

Cobalt/nickel transport system	M00245
Nickel transport system	M00246
Putative ABC transport system	M00247
Lipopolysaccharide transport system	M00250
Teichoic acid transport system	M00251
Lipooligosaccharide transport system	M00252
Sodium transport system	M00253
ABC-2 type transport system	M00254
Cell division transport system	M00256
Hemin transport system	M00257
Putative ABC transport system	M00258
Heme transport system	M00259
Spermidine/putrescine transport system	M00299
Bacitracin transport system	M00314
Uncharacterized ABC transport system	M00315
Manganese/zinc/iron transport system	M00319
alpha-Hemolysin/cyclolysin transport system	M00325
RTX toxin transport system	M00326
Glutathione transport system	M00348
Microcin C transport system	M00349
Competence-related DNA transformation transporter	M00429
KdpD-KdpE (potassium transport) two-component regula	M00454
TctE-TctD (tricarboxylic acid transport) two-component r	M00457
BceS-BceR (bacitracin transport) two-component regulat	M00469
CitS-CitT (magnesium-citrate transport) two-component	M00487
MalK-MalR (malate transport) two-component regulatory	M00490
arabinogalactan oligomer/maltooligosaccharide transport	M00491
DctB-DctD (C4-dicarboxylate transport) two-component	M00504
AlgE-type Mannuronan C-5-Epimerase transport system	M00571
Biotin transport system	M00581
Energy-coupling factor transport system	M00582
Arabinosaccharide transport system	M00602
Glucose/mannose transport system	M00605
N,N'-Diacetylchitobiose transport system	M00606
Tetracycline resistance, TetA transporter	M00668
gamma-Hexachlorocyclohexane transport system	M00669
Mce transport system	M00670
Macrolide resistance, MacAB-TolC transporter	M00709
Bacitracin resistance, VraDE transporter	M00737
Bacitracin resistance, BceAB transporter	M00738

Supplemental Table A.4 KEGG module names and IDs identified with PICRUSt for the soil bacteria case study dataset.

BIBLIOGRAPHY

1. Alivisatos, A. P., Blaser, M. J., Brodie, E. L., Chun, M., Dangl, J. L., Donohue, T. J., … Consortium, U. M. I. (2015). A unified initiative to harness Earth's microbiomes. *Science*. doi:10.1126/science.aac8480
2. Gause, G. F. (1932). Experimental studies on the struggle for existence. *Journal of Experimental Biology*.
3. Volterra, V. (1926). Fluctuations in the abundance of a species considered mathematically. *Nature*. doi:10.1038/118558a0
4. Peacor, S. D., & Werner, E. E. (2001). The contribution of trait-mediated indirect effects to the net effects of a predator. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.071061998
5. Utsumi, S., Kishida, O., & Ohgushi, T. (2010). Trait-mediated indirect interactions in ecological communities. *Population Ecology*. doi:10.1007/s10144-010-0236-3
6. Foster, K. R., & Bell, T. (2012). Competition, not cooperation, dominates interactions among culturable microbial species. *Current Biology*. doi:10.1016/j.cub.2012.08.005
7. Tetu, S. G., Sarker, I., Schrameyer, V., Pickford, R., Elbourne, L. D. H., Moore, L. R., & Paulsen, I. T. (2019). Plastic leachates impair growth and oxygen production in Prochlorococcus, the ocean's most abundant photosynthetic bacteria. *Communications Biology*. doi:10.1038/s42003-019-0410-x
8. Dinnage, R., Simonsen, A. K., Barrett, L. G., Cardillo, M., Raisbeck-Brown, N., Thrall, P. H., & Prober, S. M. (2019). Larger plants promote a greater diversity of

symbiotic nitrogen-fixing soil bacteria associated with an Australian endemic legume. *Journal of Ecology*. doi:10.1111/1365-2745.13083

9. Harris, J. (2009). Soil microbial communities and restoration ecology: Facilitators or followers? *Science*. doi:10.1126/science.1172975

10. Furin, J., Cox, H., & Pai, M. (2019). Tuberculosis. *The Lancet*. doi:10.1016/S0140-6736(19)30308-3

11. Reid, G., & Burton, J. (2002). Use of Lactobacillus to prevent infection by pathogenic bacteria. *Microbes and Infection*. doi:10.1016/S1286-4579(02)01544-7

12. Brown, C. T., Davis-Richardson, A. G., Giongo, A., Gano, K. A., Crabb, D. B., Mukherjee, N., … Triplett, E. W. (2011). Gut microbiome metagenomics analysis suggests a functional model for the development of autoimmunity for type 1 diabetes. *PLoS ONE*. doi:10.1371/journal.pone.0025792

13. Jones, B. W., & Nishiguchi, M. K. (2004). Counterillumination in the Hawaiian bobtail squid, Euprymna scolopes Berry (Mollusca: Cephalopoda). *Marine Biology*. doi:10.1007/s00227-003-1285-3

14. Weimer, B., Seefeldt, K., & Dias, B. (1999). Sulfur metabolism in bacteria associated with cheese. In *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*. doi:10.1023/A:1002050625344

15. Wolfe, B. E., & Dutton, R. J. (2015). Fermented foods as experimentally tractable microbial ecosystems. *Cell*. doi:10.1016/j.cell.2015.02.034

16. McClure, S. B., Magill, C., Podrug, E., Moore, A. M. T., Harper, T. K., Culleton, B. J., … Freeman, K. H. (2018). Fatty acid specific δ13C values reveal earliest

Mediterranean cheese production 7,200 years ago. *PLoS ONE*. doi:10.1371/journal.pone.0202807

17. Patnaik, R. (2008). Engineering complex phenotypes in industrial strains. In *Biotechnology Progress*. doi:10.1021/bp0701214

18. Paine, J. A., Shipton, C. A., Chaggar, S., Howells, R. M., Kennedy, M. J., Vernon, G., … Drake, R. (2005). Improving the nutritional value of Golden Rice through increased pro-vitamin A content. *Nature Biotechnology*. doi:10.1038/nbt1082

19. Butler, P. R., Brown, M., & Oliver, S. G. (1996). Improvement of antibiotic titers from Streptomyces bacteria by interactive continuous selection. *Biotechnology and Bioengineering*. doi:10.1002/(SICI)1097-0290(19960120)49:2<185::AID-BIT7>3.0.CO;2-M

20. Leonard, E., Yan, Y., Fowler, Z. L., Li, Z., Lim, C. G., Lim, K. H., & Koffas, M. A. G. (2008). Strain improvement of recombinant Escherichia coli for efficient production of plant flavonoids. *Molecular Pharmaceutics*. doi:10.1021/mp7001472

21. Munroe, S., Sandoval, K., Martens, D. E., Sipkema, D., & Pomponi, S. A. (2019). Genetic algorithm as an optimization tool for the development of sponge cell culture media. *In Vitro Cellular and Developmental Biology - Animal*. doi:10.1007/s11626-018-00317-0

22. Etschmann, M. M. W., Sell, D., & Schrader, J. (2004). Medium optimization for the production of the aroma compound 2-phenylethanol using a genetic algorithm. In *Journal of Molecular Catalysis B: Enzymatic*.

doi:10.1016/j.molcatb.2003.10.014

23. Yuan, Y., Du, J., & Zhao, H. (2013). Customized optimization of metabolic pathways by combinatorial transcriptional engineering. *Methods in Molecular Biology*. doi:10.1007/978-1-62703-299-5_10

24. Chen, Q. H., He, G. Q., & Ali, M. A. M. (2002). Optimization of medium composition for the production of elastase by Bacillus sp. EL31410 with response surface methodology. *Enzyme and Microbial Technology*. doi:10.1016/S0141-0229(02)00028-5

25. Rawlings, D. E., & Johnson, D. B. (2007). The microbiology of biomining: Development and optimization of mineral-oxidizing microbial consortia. *Microbiology*. doi:10.1099/mic.0.2006/001206-0

26. Mikesková, H., Novotný, C., & Svobodová, K. (2012). Interspecific interactions in mixed microbial cultures in a biodegradation perspective. *Applied Microbiology and Biotechnology*. doi:10.1007/s00253-012-4234-6

27. Brenner, K., You, L., & Arnold, F. H. (2008). Engineering microbial consortia: a new frontier in synthetic biology. *Trends in Biotechnology*. doi:10.1016/j.tibtech.2008.05.004

28. Bakken, J. S., Borody, T., Brandt, L. J., Brill, J. V., Demarco, D. C., Franzos, M. A., … Surawicz, C. (2011). Treating clostridium difficile infection with fecal microbiota transplantation. *Clinical Gastroenterology and Hepatology*. doi:10.1016/j.cgh.2011.08.014

29. Foo, J. L., Ling, H., Lee, Y. S., & Chang, M. W. (2017). Microbiome engineering:

Current applications and its future. *Biotechnology Journal*. doi:10.1002/biot.201600099

30. Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The Human Microbiome Project. *Nature*. doi:10.1038/nature06244

31. Ilan, Y. (2019). Why targeting the microbiome is not so successful: Can randomness overcome the adaptation that occurs following gut manipulation? *Clinical and Experimental Gastroenterology*. doi:10.2147/CEG.S203823

32. Schloss, P. D., Jenior, M. L., Koumpouras, C. C., Westcott, S. L., & Highlander, S. K. (2016). Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. *PeerJ*. doi:10.7717/peerj.1869

33. Fanning, S., Proos, S., Jordan, K., & Srikumar, S. (2017). A review on the applications of next generation sequencing technologies as applied to food-related microbiome studies. *Frontiers in Microbiology*. doi:10.3389/fmicb.2017.01829

34. Lewis, K., Epstein, S., D'Onofrio, A., & Ling, L. L. (2010). Uncultured microorganisms as a source of secondary metabolites. *Journal of Antibiotics*. doi:10.1038/ja.2010.87

35. Kehe, J., Kulesa, A., Ortiz, A., Ackerman, C. M., Thakku, S. G., Sellers, D., … Blainey, P. C. (2019). Massively parallel screening of synthetic microbial communities. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1900102116

36. Hsu, R. H., Clark, R. L., Tan, J. W., Romero, P. A., & Venturelli, O. S. (2019).

Rapid microbial interaction network inference in microfluidic droplets. *bioRxiv*. doi:10.1101/521823

37. Traore, S. I., Khelaifia, S., Armstrong, N., Lagier, J. C., & Raoult, D. (2019). Isolation and culture of Methanobrevibacter smithii by co-culture with hydrogen-producing bacteria on agar plates. *Clinical Microbiology and Infection*. doi:10.1016/j.cmi.2019.04.008

38. Said, S. Ben, & Or, D. (2017). Synthetic microbial ecology: Engineering habitats for modular consortia. *Frontiers in Microbiology*. doi:10.3389/fmicb.2017.01125

39. Kong, W., Meldgin, D. R., Collins, J. J., & Lu, T. (2018). Designing microbial consortia with defined social interactions. *Nature Chemical Biology*. doi:10.1038/s41589-018-0091-7

40. Bucci, V., Tzen, B., Li, N., Simmons, M., Tanoue, T., Bogart, E., … Gerber, G. K. (2016). MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. *Genome Biology*. doi:10.1186/s13059-016-0980-6

41. Angulo, M. T. (2017). Controlling microbial communities: a theoretical framework. *arXiv*. doi:10.1073/pnas.xxxxxxxxxx

42. Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., … Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, *31*(9), 814–821. doi:10.1038/nbt.2676

43. Aziz, R. K., Bartels, D., Best, A., DeJongh, M., Disz, T., Edwards, R. A., …

Zagnitko, O. (2008). The RAST Server: Rapid annotations using subsystems technology. *BMC Genomics*. doi:10.1186/1471-2164-9-75

44. Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*. doi:10.1038/538020a

45. Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*. doi:10.1080/00401706.1970.10488634

46. Janson, N. B. (2012). Non-linear dynamics of biological systems. *Contemporary Physics*. doi:10.1080/00107514.2011.644441

47. Tuddenham, S. A., Koay, W. L. A., Zhao, N., White, J. R., Ghanem, K. G., Sears, C. L., … Goedert, J. J. (2019). The Impact of Human Immunodeficiency Virus Infection on Gut Microbiota α-Diversity: An Individual-level Meta-analysis. *Clinical Infectious Diseases*. doi:10.1093/cid/ciz258

48. Ness, R. O., Sachs, K., & Vitek, O. (2016). From Correlation to Causality: Statistical Approaches to Learning Regulatory Relationships in Large-Scale Biomolecular Investigations. *Journal of Proteome Research*. doi:10.1021/acs.jproteome.5b00911

49. Carr, A., Diener, C., Baliga, N. S., & Gibbons, S. M. (2019). Use and abuse of correlation analyses in microbial ecology. *The ISME Journal*. doi:10.1038/s41396-019-0459-z

50. Tackmann, J., Rodrigues, J. F. M., & Mering, C. von. (2018). Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data. *bioRxiv*. doi:10.1101/390195

51. Menon, R., Ramanan, V., & Korolev, K. S. (2018). Interactions between species introduce spurious associations in microbiome studies. *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1005939

52. Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., & Bonneau, R. A. (2015). Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1004226

53. Tilman, D., Reich, P. B., Knops, J., Wedin, D., Mielke, T., & Lehman, C. (2001). Diversity and productivity in a long-term grassland experiment. *Science*. doi:10.1126/science.1060391

54. Brophy, C., Dooley, Á., Kirwan, L., Finn, J. A., McDonnell, J., Bell, T., … Connolly, J. (2017). Biodiversity and ecosystem function: making sense of numerous species interactions in multi-species communities. *Ecology*. doi:10.1002/ecy.1872

55. Dooley, Á., Isbell, F., Kirwan, L., Connolly, J., Finn, J. A., & Brophy, C. (2015). Testing the effects of diversity on ecosystem multifunctionality using a multivariate model. *Ecology Letters*. doi:10.1111/ele.12504

56. Kucharzyk, K. H., Crawford, R. L., Paszczynski, A. J., Soule, T., & Hess, T. F. (2012). Maximizing microbial degradation of perchlorate using a genetic algorithm: Media optimization. *Journal of Biotechnology*. doi:10.1016/j.jbiotec.2011.10.011

57. Faith, J. J., Ahern, P. P., Ridaura, V. K., Cheng, J., & Gordon, J. I. (2014).

Identifying gut microbe-host phenotype relationships using combinatorial communities in gnotobiotic mice. *Science Translational Medicine*. doi:10.1126/scitranslmed.3008051

58. Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*. doi:10.1111/j.2517-6161.1996.tb02080.x

59. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*. doi:10.1111/j.1467-9868.2005.00503.x

60. Walker, E. (2003). Regression Modeling Strategies. *Technometrics*. doi:10.1198/tech.2003.s158

61. Barrera-Gómez, J., Agier, L., Portengen, L., Chadeau-Hyam, M., Giorgis-Allemand, L., Siroux, V., … Basagaña, X. (2017). A systematic comparison of statistical methods to detect interactions in exposome-health associations. *Environmental Health: A Global Access Science Source*. doi:10.1186/s12940-017-0277-6

62. Xia, L. C., Steele, J. A., Cram, J. A., Cardon, Z. G., Simmons, S. L., Vallino, J. J., … Sun, F. (2011). Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. *BMC Systems Biology*. doi:10.1186/1752-0509-5-S2-S15

63. Faust, K., & Raes, J. (2016). CoNet app: inference of biological association networks using Cytoscape. *F1000Research*. doi:10.12688/f1000research.9050.2

64. Senay, Y., John, G., Knutie, S. A., & Ogbunugafor, C. B. (2019). Deconstructing higher-order interactions in the microbiota: A theoretical examination. *bioRxiv*. doi:10.1101/647156

65. Lambiotte, R., Rosvall, M., & Scholtes, I. (2019). From networks to optimal higher-order models of complex systems. *Nature Physics*. doi:10.1038/s41567-019-0459-y

66. Netzker, T., Fischer, J., Weber, J., Mattern, D. J., König, C. C., Valiante, V., … Brakhage, A. A. (2015). Microbial communication leading to the activation of silent fungal secondary metabolite gene clusters. *Frontiers in Microbiology*. doi:10.3389/fmicb.2015.00299

67. Zarins-Tutt, J. S., Barberi, T. T., Gao, H., Mearns-Spragg, A., Zhang, L., Newman, D. J., & Goss, R. J. M. (2016). Prospecting for new bacterial metabolites: A glossary of approaches for inducing, activating and upregulating the biosynthesis of bacterial cryptic or silent natural products. *Natural Product Reports*. doi:10.1039/c5np00111k

68. Milshteyn, A., Schneider, J. S., & Brady, S. F. (2014). Mining the metabiome: Identifying novel natural products from microbial communities. *Chemistry and Biology*. doi:10.1016/j.chembiol.2014.08.006

69. Tshikantwa, T. S., Ullah, M. W., He, F., & Yang, G. (2018). Current trends and potential applications of microbial interactions for human welfare. *Frontiers in Microbiology*. doi:10.3389/fmicb.2018.01156

70. Røder, H. L., Sørensen, S. J., & Burmølle, M. (2016). Studying Bacterial

Multispecies Biofilms: Where to Start? *Trends in Microbiology*. doi:10.1016/j.tim.2016.02.019

71. Aziz, F. A. A., Suzuki, K., Ohtaki, A., Sagegami, K., Hirai, H., Seno, J., … Futamata, H. (2015). Interspecies interactions are an integral determinant of microbial community dynamics. *Frontiers in Microbiology*, *6*(OCT). doi:10.3389/fmicb.2015.01148

72. Higgins, L. M., Friedman, J., Shen, H., & Gore, J. (2017). Co-occurring soil bacteria exhibit a robust competitive hierarchy and lack of non-transitive interactions. *bioRxiv*, 175737. doi:10.1101/175737

73. Bairey, E., Kelsic, E. D., & Kishony, R. (2016). High-order species interactions shape ecosystem diversity. *Nature Communications*, *7*. doi:10.1038/ncomms12285

74. Harcombe, W. R., Riehl, W. J., Dukovski, I., Granger, B. R., Betts, A., Lang, A. H., … Segrè, D. (2014). Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Reports*, *7*(4), 1104–1115. doi:10.1016/j.celrep.2014.03.070

75. Coyte, K. Z., Schluter, J., & Foster, K. R. (2015). The ecology of the microbiome: Networks, competition, and stability. *Science*. doi:10.1126/science.aad2602

76. Taga, M. E., & Bassler, B. L. (2003). Chemical communication among bacteria. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1934514100

77. Datta, M. S., Sliwerska, E., Gore, J., Polz, M. F., & Cordero, O. X. (2016). Microbial interactions lead to rapid micro-scale successions on model marine particles. *Nature Communications*. doi:10.1038/ncomms11965

78. Johns, N. I., Blazejewski, T., Gomes, A. L. C., & Wang, H. H. (2016). Principles for designing synthetic microbial communities. *Current Opinion in Microbiology*. doi:10.1016/j.mib.2016.03.010

79. Shou, W., Ram, S., & Vilar, J. M. G. (2007). Synthetic cooperation in engineered yeast populations. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.0610575104

80. Zomorrodi, A. R., & Segrè, D. (2016). Synthetic Ecology of Microbes: Mathematical Models and Applications. *Journal of Molecular Biology*. doi:10.1016/j.jmb.2015.10.019

81. Venturelli, O. S., Carr, A. C., Fisher, G., Hsu, R. H., Lau, R., Bowen, B. P., … Arkin, A. P. (2018). Deciphering microbial interactions in synthetic human gut microbiome communities. *Molecular Systems Biology*, *14*(6), e8157. doi:10.15252/msb.20178157

82. Friedman, J., Higgins, L. M., & Gore, J. (2017). Community structure follows simple assembly rules in microbial microcosms. *Nature Ecology and Evolution*, *1*(5). doi:10.1038/s41559-017-0109

83. Goers, L., Freemont, P., & Polizzi, K. M. (2014). Co-culture systems and technologies: taking synthetic biology to the next level. *Journal of The Royal Society Interface*, *11*(96), 20140065–20140065. doi:10.1098/rsif.2014.0065

84. Lasken, R. S., & McLean, J. S. (2014). Recent advances in genomic DNA sequencing of microbial species from single cells. *Nature Reviews Genetics*, *15*(9), 577–584. doi:10.1038/nrg3785

85. Gawad, C., Koh, W., & Quake, S. R. (2016). Single-cell genome sequencing: Current state of the science. *Nature Reviews Genetics*. doi:10.1038/nrg.2015.16

86. Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. doi:10.1093/bioinformatics/btu153

87. Dam, P., Olman, V., Harris, K., Su, Z., & Xu, Y. (2007). Operon prediction using both genome-specific and general genomic information. *Nucleic Acids Research*, *35*(1), 288–298. doi:10.1093/nar/gkl1018

88. Ho, T. K. (1995). Random Decision Forests Tin Kam Ho Perceptron training. In *Proceedings of 3rd International Conference on Document Analysis and Recognition* (pp. 278–282). doi:10.1109/ICDAR.1995.598994

89. Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi:10.1023/A:1010933404324

90. Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

91. Palczewska, A., Palczewski, J., Robinson, R. M., & Neagu, D. (2013). Interpreting random forest classification models using a feature contribution method (extended). *2013 IEEE 14th International Conference on Information Reuse & Integration (IRI)*, 1–30. doi:10.1109/IRI.2013.6642461

92. Henson, M. A., & Hanly, T. J. (2014). Dynamic flux balance analysis for synthetic microbial communities. *IET Systems Biology*, *8*(5), 214–229. doi:10.1049/iet-syb.2013.0021

93. Maarleveld, T. R., Khandelwal, R. A., Olivier, B. G., Teusink, B., & Bruggeman,

F. J. (2013). Basic concepts and principles of stoichiometric modeling of metabolic networks. *Biotechnology Journal*. doi:10.1002/biot.201200291

94. Bauer, E., Laczny, C. C., Magnusdottir, S., Wilmes, P., & Thiele, I. (2015). Phenotypic differentiation of gastrointestinal microbes is reflected in their encoded metabolic repertoires. *Microbiome*, *3*, 55. doi:10.1186/s40168-015-0121-6

95. Svetnik, V., Liaw, A., Tong, C., Christopher Culberson, J., Sheridan, R. P., & Feuston, B. P. (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences*, *43*(6), 1947–1958. doi:10.1021/ci034160g

96. Jaccard, P. (1912). THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE. *New Phytologist*, *11*(2), 37–50. doi:10.1111/j.1469-8137.1912.tb05611.x

97. Bach, E. M., Williams, R. J., Hargreaves, S. K., Yang, F., & Hofmockel, K. S. (2018). Greatest soil microbial diversity found in micro-habitats. *Soil Biology and Biochemistry*, *118*, 217–226. doi:10.1016/j.soilbio.2017.12.018

98. Mainali, K. P., Bewick, S., Thielen, P., Mehoke, T., Breitwieser, F. P., Paudel, S., … Fagan, W. F. (2017). Statistical analysis of co-occurrence patterns in microbial presence-absence datasets. *PLoS ONE*, *12*(11). doi:10.1371/journal.pone.0187132

99. Kuz'min, V. E., Polishchuk, P. G., Artemenko, A. G., & Andronati, S. A. (2011). Interpretation of QSAR models based on random forest methods. *Molecular Informatics*, *30*(6–7), 593–603. doi:10.1002/minf.201000173

100. Di Luccia, B., Crescenzo, R., Mazzoli, A., Cigliano, L., Venditti, P., Walser, J. C., … Iossa, S. (2015). Rescue of fructose-induced metabolic syndrome by antibiotics

or faecal transplantation in a rat model of obesity. *PLoS ONE*, *10*(8). doi:10.1371/journal.pone.0134893

101. Khitan, Z., & Kim, D. H. (2013). Fructose: A key factor in the development of metabolic syndrome and hypertension. *Journal of Nutrition and Metabolism*. doi:10.1155/2013/682673

102. Bantle, J. P. (2009). Dietary Fructose and Metabolic Syndrome and Diabetes. *Journal of Nutrition*, *139*(6), 1263S-1268S. doi:10.3945/jn.108.098020

103. Lambertz, J., Weiskirchen, S., Landert, S., & Weiskirchen, R. (2017). Fructose: A dietary sugar in crosstalk with microbiota contributing to the development and progression of non-alcoholic liver disease. *Frontiers in Immunology*. doi:10.3389/fimmu.2017.01159

104. Payne, A. N., Chassard, C., & Lacroix, C. (2012). Gut microbial adaptation to dietary consumption of fructose, artificial sweeteners and sugar alcohols: Implications for host-microbe interactions contributing to obesity. *Obesity Reviews*, *13*(9), 799–809. doi:10.1111/j.1467-789X.2012.01009.x

105. Mee, M. T., Collins, J. J., Church, G. M., & Wang, H. H. (2014). Syntrophic exchange in synthetic microbial communities. *Proceedings of the National Academy of Sciences*, *111*(20), E2149–E2156. doi:10.1073/pnas.1405641111

106. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*. doi:10.1093/nar/27.1.29

107. Swoboda, J. G., Campbell, J., Meredith, T. C., & Walker, S. (2010). Wall teichoic

acid function, biosynthesis, and inhibition. *ChemBioChem*. doi:10.1002/cbic.200900557

108. Levine, J. M., Bascompte, J., Adler, P. B., & Allesina, S. (2017). Beyond pairwise mechanisms of species coexistence in complex communities. *Nature*. doi:10.1038/nature22898

109. Raman, K., & Chandra, N. (2009). Flux balance analysis of biological systems: Applications and challenges. *Briefings in Bioinformatics*. doi:10.1093/bib/bbp011

110. Bordbar, A., Monk, J. M., King, Z. A., & Palsson, B. O. (2014). Constraint-based models predict metabolic and associated cellular functions. *Nature Reviews Genetics*. doi:10.1038/nrg3643

111. O'Brien, E. J., Monk, J. M., & Palsson, B. O. (2015). Using genome-scale models to predict biological capabilities. *Cell*, *161*(5), 971–987. doi:10.1016/j.cell.2015.05.019

112. van der Ark, K. C. H., van Heck, R. G. A., Martins Dos Santos, V. A. P., Belzer, C., & de Vos, W. M. (2017). More than just a gut feeling: constraint-based genome-scale metabolic models for predicting functions of human intestinal microbes. *Microbiome*, *5*(1), 78. doi:10.1186/s40168-017-0299-x

113. Zomorrodi, A. R., & Segrè, D. (2017). Genome-driven evolutionary game theory helps understand the rise of metabolic interdependencies in microbial communities. *Nature Communications*, *8*(1), 1563. doi:10.1038/s41467-017-01407-5

114. Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis?

Nature Biotechnology, *28*(3), 245–248. doi:10.1038/nbt.1614

115. Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, *41*(D1). doi:10.1093/nar/gks1195

116. Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*. doi:10.1038/nmeth.f.303

117. DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., … Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*. doi:10.1128/AEM.03006-05

118. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*. doi:10.1186/1471-2105-12-77

119. Welling, S. H., Refsgaard, H. H. F., Brockhoff, P. B., & Clemmensen, L. H. (2016). Forest Floor Visualizations of Random Forests. *arXiv*. Retrieved from http://arxiv.org/abs/1605.09196

120. Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*. doi:10.1016/j.ajhg.2017.06.005

121. Furqan, M. S., & Siyal, M. Y. (2016). Inference of biological networks using Bi-

directional Random Forest Granger causality. *SpringerPlus*. doi:10.1186/s40064-016-2156-y

122. Le, V., Quinn, T. P., Tran, T., & Venkatesh, S. (2019). Deep in the Bowel: Highly Interpretable Neural Encoder-Decoder Networks Predict Gut Metabolites from Gut Microbiome. *bioRxiv*. doi:10.1101/686394

123. Azmi, M., Runger, G. C., & Berrado, A. (2019). Interpretable regularized class association rules algorithm for classification in a categorical data space. *Information Sciences*. doi:10.1016/j.ins.2019.01.047

124. Nguyen, M., Wesley Long, S., McDermott, P. F., Olsen, R. J., Olson, R., Stevens, R. L., … Davisa, J. J. (2019). Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal Salmonella. *Journal of Clinical Microbiology*. doi:10.1128/JCM.01260-18

125. LaPierre, N., Ju, C. J. T., Zhou, G., & Wang, W. (2019). MetaPheno: A critical evaluation of deep learning and machine learning in metagenome-based disease prediction. *Methods*. doi:10.1016/j.ymeth.2019.03.003

126. Brodley, C. E., & Friedl, M. A. (1997). Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*. doi:10.1016/S0034-4257(97)00049-7

127. Knudson, A. G. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America*. doi:10.1073/pnas.68.4.820

128. Emily, M., Mailund, T., Hein, J., Schauser, L., & Schierup, M. H. (2009). Using

biological networks to search for interacting loci in genome-wide association studies. *European Journal of Human Genetics*. doi:10.1038/ejhg.2009.15

129. Leem, S., Jeong, H. H., Lee, J., Wee, K., & Sohn, K. A. (2014). Fast detection of high-order epistatic interactions in genome-wide association studies using information theoretic measure. *Computational Biology and Chemistry*. doi:10.1016/j.compbiolchem.2014.01.005

130. Proctor, L. M., Creasy, H. H., Fettweis, J. M., Lloyd-Price, J., Mahurkar, A., Zhou, W., … Huttenhower, C. (2019). The Integrative Human Microbiome Project. *Nature*. doi:10.1038/s41586-019-1238-8

131. Reading, D. (2014). Crohn Disease: Pathophysiology, Diagnosis, and Treatment, *85*(3), 297–320.

132. Louppe, G. (2014). *Understanding Random Forests*. *Cornell University Library*.

133. Wright, M. N., Ziegler, A., & König, I. R. (2016). Do little interactions get lost in dark random forests? *BMC bioinformatics*. doi:10.1186/s12859-016-0995-8

134. Boulesteix, A. L., Janitza, S., Kruppa, J., & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. doi:10.1002/widm.1072

135. Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., & Sacha van Hijum, A. F. T. (2013). Data mining in the life science swith random forest: A walk in the park or lost in the jungle? *Briefings in Bioinformatics*. doi:10.1093/bib/bbs034

136. Nguyen, C., Wang, Y., & Nguyen, H. N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *Journal of Biomedical Science and Engineering*. doi:10.4236/jbise.2013.65070

137. Franzosa, E. A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H. J., Reinker, S., … Xavier, R. J. (2019). Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature Microbiology*. doi:10.1038/s41564-018-0306-4

138. Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., & Alm, E. J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature Communications*. doi:10.1038/s41467-017-01973-8

139. Deng, H. (2019). Interpreting tree ensembles with inTrees. *International Journal of Data Science and Analytics*. doi:10.1007/s41060-018-0144-8

140. Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*. doi:10.1186/1471-2105-9-307

141. Basu, S., Kumbier, K., Brown, J. B., & Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1711236115

142. Dessau, R. B., & Pipper, C. B. (2008). [''R"--project for statistical computing]. *Ugeskrift for laeger*.

143. Carding, S., Verbeke, K., Vipond, D. T., Corfe, B. M., & Owen, L. J. (2015). Dysbiosis of the gut microbiota in disease. *Microbial Ecology in Health &*

Disease. doi:10.3402/mehd.v26.26191

144. Levy, M., Kolodziejczyk, A. A., Thaiss, C. A., & Elinav, E. (2017). Dysbiosis and the immune system. *Nature Reviews Immunology*. doi:10.1038/nri.2017.7

145. Ai, D., Pan, H., Han, R., Li, X., Liu, G., & Xia, L. C. (2019). Using Decision Tree Aggregation with Random Forest Model to Identify Gut Microbes Associated with Colorectal Cancer. *Genes*, *10*(2), 112. doi:10.3390/genes10020112

146. Vangay, P., Hillmann, B. M., & Knights, D. (2019). Microbiome learning Repo (ML Repo): A public repository of microbiome regression and classification tasks. *GigaScience*. doi:10.1093/gigascience/giz042

147. Loh, G., & Blaut, M. (2012). Role of commensal gut bacteria in inflammatory bowel diseases. *Gut Microbes*. doi:10.4161/gmic.22156

148. Nagao-Kitamoto, H., & Kamada, N. (2017). Host-microbial Cross-talk in Inflammatory Bowel Disease. *Immune Network*. doi:10.4110/in.2017.17.1.1

149. Geirnaert, A., Calatayud, M., Grootaert, C., Laukens, D., Devriese, S., Smagghe, G., … Van De Wiele, T. (2017). Butyrate-producing bacteria supplemented in vitro to Crohn's disease patient microbiota increased butyrate production and enhanced intestinal epithelial barrier integrity. *Scientific Reports*. doi:10.1038/s41598-017-11734-8

150. Wang, Y., Gao, X., Ghozlane, A., Hu, H., Li, X., Xiao, Y., … Zhang, T. (2018). Characteristics of faecal microbiota in paediatric Crohn's disease and their dynamic changes during infliximab therapy. *Journal of Crohn's and Colitis*. doi:10.1093/ecco-jcc/jjx153

151. Wright, E. K., Kamm, M. A., Wagner, J., Teo, S. M., Cruz, P. De, Hamilton, A. L., … Kirkwood, C. D. (2017). Microbial Factors Associated with Postoperative Crohn's Disease Recurrence. *Journal of Crohn's & colitis*. doi:10.1093/ecco-jcc/jjw136

152. Y.-J., C., H., W., S.-D., W., N., L., Y.-T., W., H.-N., L., … Shen, X.-Z. (2018). Parasutterella, in association with irritable bowel syndrome and intestinal chronic inflammation. *Journal of Gastroenterology and Hepatology (Australia)*. doi:10.1111/jgh.14281

153. Berry, D., Rahman, S., Kaplan, J., & Gordon, N. (2018). Probiotic and prebiotic compositions, and methods of use thereof for treatment and prevention of graft versus host disease. *USPTO*.

154. Yang, C., Fang, X., Zhan, G., Huang, N., Li, S., Bi, J., … Hashimoto, K. (2019). Key role of gut microbiota in anhedonia-like phenotype in rodents with neuropathic pain. *Translational Psychiatry*. doi:10.1038/s41398-019-0379-8

155. Carpinelli, L., Bucci, C., Santonicola, A., Zingone, F., Ciacci, C., & Iovino, P. (2019). Anhedonia in irritable bowel syndrome and in inflammatory bowel diseases and its relationship with abdominal pain. *Neurogastroenterology and Motility*. doi:10.1111/nmo.13531

156. Rabizadeh, S., Rhee, K. J., Wu, S., Huso, D., Gan, C. M., Golub, J. E., … Sears, C. L. (2007). Enterotoxigenic Bacteroides fragilis: A potential instigator of colitis. *Inflammatory Bowel Diseases*. doi:10.1002/ibd.20265

157. Wang, K., Yang, Q., Ma, Q., Wang, B., Wan, Z., Chen, M., & Wu, L. (2018).

Protective effects of salvianolic acid a against dextran sodium sulfate-induced acute colitis in rats. *Nutrients*. doi:10.3390/nu10060791

158. Maynard, D. S., Crowther, T. W., & Bradford, M. A. (2017). Competitive network determines the direction of the diversity–function relationship. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1712211114

159. Mayfield, M. M., & Stouffer, D. B. (2017). Higher-order interactions capture unexplained complexity in diverse communities. *Nature Ecology and Evolution*. doi:10.1038/s41559-016-0062

160. Werner, E. E., & Peacor, S. D. (2003). A review of trait-mediated indirect interactions in ecological communities. *Ecology*. doi:10.1890/0012-9658(2003)084[1083:AROTII]2.0.CO;2

161. Cheadle, E. J., & Jackson, A. M. (2002). Bugs as drugs for cancer. *Immunology*. doi:10.1046/j.1365-2567.2002.01498.x

162. Allen-Vercoe, E., & Petrof, E. (2018). Using bugs as drugs: Microbial ecosystem therapeutics. *Canada Communicable Disease Report*. doi:10.14745/ccdr.v41is5a01

163. Thompson, J., Johansen, R., Dunbar, J., Munsky, B., Engineering, B., & Alamos, L. (2019). Machine learning to predict microbial community functions : An analysis of dissolved organic carbon from litter decomposition . *bioRxiv*. doi:10.1101/599704

164. Langenheder, S., Bulling, M. T., Solan, M., & Prosser, J. I. (2010). Bacterial Biodiversity-Ecosystem Functioning Relations are Modified by Environmental

Complexity. *PLoS ONE*. doi:10.1371/journal.pone.0010834

165. Herrera Paredes, S., Gao, T., Law, T. F., Finkel, O. M., Mucyn, T., Teixeira, P. J. P. L., … Castrillo, G. (2018). Design of synthetic bacterial communities for predictable plant phenotypes. *PLoS Biology*. doi:10.1371/journal.pbio.2003962

166. Wang, T., & Zhao, H. (2017). Structured subcomposition selection in regression and its application to microbiome data analysis. *Annals of Applied Statistics*. doi:10.1214/16-AOAS1017

167. Momeni, B., Xie, L., & Shou, W. (2017). Lotka-Volterra pairwise modeling fails to capture diverse pairwise microbial interactions. *eLife*, *6*. doi:10.7554/eLife.25051.001

168. Zheng, P., Zeng, B., Liu, M., Chen, J., Pan, J., Han, Y., … Xie, P. (2019). The gut microbiome from patients with schizophrenia modulates the glutamate-glutamine-GABA cycle and schizophrenia-relevant behaviors in mice. *Science Advances*. doi:10.1126/sciadv.aau8317

169. García-Magariños, M., López-De-Ullibarri, I., Cao, R., & Salas, A. (2009). Evaluating the ability of tree-based methods and logistic regression for the detection of SNP-SNP interaction. *Annals of Human Genetics*. doi:10.1111/j.1469-1809.2009.00511.x

170. Qi, Y. (2012). Random forest for bioinformatics. In *Ensemble Machine Learning: Methods and ApplicatiOns*. doi:10.1007/9781441993267_10

171. Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *Annals of Applied Statistics*. doi:10.1214/07-AOAS148

172. Hara, S., & Hayashi, K. (2018). Making tree ensembles interpretable: A Bayesian model selection approach. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018*.

173. Yang, J. H., Wright, S. N., Hamblin, M., McCloskey, D., Alcantar, M. A., Schrübbers, L., … Collins, J. J. (2019). A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action. *Cell*. doi:10.1016/j.cell.2019.04.016

174. Brambor, T., Clark, W. R., & Golder, M. (2006). Understanding interaction models: Improving empirical analyses. *Political Analysis*. doi:10.1093/pan/mpi014

175. Jaillard, B., Richon, C., Deleporte, P., Loreau, M., & Violle, C. (2018). An a posteriori species clustering for quantifying the effects of species interactions on ecosystem functioning. *Methods in Ecology and Evolution*. doi:10.1111/2041-210X.12920

CURRICULUM VITAE