Machine learning reveals missing edges and putative interaction mechanisms in microbial ecosystem networks

Demetrius DiMucci[a,b], Mark Kon[a,c], Daniel Segrè[a,b,d, #]

[a] Bioinformatics Graduate Program, Boston University, Boston, Massachusetts, USA

[b] Biological Design Center, Boston University, Boston, Massachusetts, USA

[c] Department of Mathematics and Statistics, Boston University, Boston, Massachusetts, USA

[d] Department of Biology, Boston University, Boston. Massachusetts, USA

Running Head: Machine Learning of Microbial Ecosystem Networks

[#] Correspondence to Daniel Segrè, dsegre@bu.edu, 617-358-2301

Abstract Word Count: 207 (Abstract)/ 134 (Importance)

Main Text Word Count: 4140

Abstract

Microbes affect each other's growth in multiple, often elusive ways. The ensuing interdependencies form complex networks, believed to influence 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 two case studies: an experimentally mapped network of interactions between auxotrophic *E. coli* strains, 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 (1, 2). 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 communities (3) and for the design of synthetic communities (4).

A powerful, and increasingly employed method for assessing intermicrobial interactions is the direct measurement of phenotypes of microbial species grown in co-culture (5–7). A fundamental

70 challenge in this endeavor is the huge diversity of many natural communities, which could count
71 up to several hundred strains or species of microbes. Performing experiments for all possible
72 pairwise interactions constitutes a herculean, and likely insurmountable task for even a
73 moderately sized community. It is however, conceivable that new computational approaches
74 could systematically complement existing tools such as high-throughput sequencing and genome
75 annotation (8–13) to help extract as much information as possible from interaction datasets,
76 providing both insight on yet-to-be-measured interactions, and on possible biological mechanisms
77 mediating specific partnerships.
78
79 Here we present a conceptual framework for the mathematical representation of microbial
80 interactions and subsequent use of supervised learning to build a classifier with high predictive
81 accuracy. While any algorithm may be used, we obtained our best results with random forest (14).
82 Random forests are ensembles of many decision trees that individually are poor classifiers but can
83 be democratically pooled to create a very good classifier. Random forests have two attributes that
84 we found particularly attractive for our purposes here. First, they are non-parametric and thus
85 require no *a priori* definition about underlying relationships between predictive variables. Second,
86 recent methodological developments in the interpretation of random forests have been made that
87 allow users to query why a specific example was classified the way that it was through the
88 calculation of feature contributions (15). Feature contributions can be exploited to develop new
89 hypotheses about the mechanisms that mediate specific interactions. In order to demonstrate a
90 proof of principle for the classification of microbial interactions using organism traits and the
91 utility of feature contributions for developing insight into the underlying mechanisms, we applied
92 this approach to two communities where all pairwise experiments had been performed and the

mechanisms of interaction identified. The first is an *in silico* community of 100 metabolic models of human gut associated bacteria and the second are the experimental results of a study involving 14 amino acid auxotrophic strains of *Escherichia coli* (*E. coli*). Our results show that the application of random forests and feature contributions to the study of microbial communities can significantly increase the capacity to estimate a large number of interactions based on a limited number of experiments, and has the potential to enable the discovery of new interaction mechanisms.

Results

Representing Pairwise Interactions

Our objective in this study was twofold; first, we sought to predict the qualitative outcome of unobserved pairwise interactions in a microbial community; 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 enables an algorithm to make good predictions and also be easily parsed for interpretation. Our approach relies of the availability of a trait-level description available for each organism in the community under consideration. Trait descriptions are used to construct trait vectors for each organism. Specific interactions are represented as the concatenation of the relevant trait vectors (Figure 1). Trait vectors may be constructed from any set of biologically relevant features such as body size or metabolic requirements; here we used genome-scale metabolic data for every organism in the *in silico* community case study, and vectors of biosynthetic capabilities for each *E. coli* strain in the auxotroph community case study.

115 These trait vectors, together with the known outcome of specific interactions, can be fed into a
116 machine learning algorithm capable of absorbing existing patterns and extending predictions to
117 unknown interactions. The approach we use here, a random forests classifier, is an ensemble of
118 many decision trees that individually ask a series of yes or no questions about randomly selected
119 subsets of predictive features in order to classify samples. Single trees tend to be poor classifiers;
120 however, when their decisions are pooled, the collective predictive power is often impressive. In
121 order to find potential mechanisms of interaction we took advantage of the binary nature of
122 individual trees in order to identify which variables are the most influential for the classification
123 of specific samples.
124
125 **Application to computationally predicted interactions between human gut microbes**
126 We first applied our approach to a large *in silico* data set that we generated by simulating with
127 dynamic flux balance analysis (16) (using COMETS (17), see Methods) all pairwise co-culture
128 interactions between 100 metabolic models of human-gut associated bacteria (18) under rich
129 medium, in a well-mixed batch culture. Each metabolic model is a network of over a thousand
130 interconnected metabolic reactions that fall into two categories: intracellular and exchange
131 reactions (19). Intracellular reactions correspond to the various pathways found within the cell,
132 while exchange reactions represent the ability of an organism to transport a metabolite across the
133 cell membrane. Among the 100 organisms, there were a total of 2083 unique metabolic reactions
134 that could serve as potential predictors. Out of these, 194 were exchange reactions. Since
135 interactions between organisms emerge through the exchange of or competition for one or more
136 extracellular metabolites in the environment, we used only the exchange reactions as predictive
137 features (Figure 2A).

138 Interactions in the network were computed by determining the influence of each organism on
139 each other organism in co-culture simulations. We compared the accumulated biomass of each
140 model in co-culture experiment to its accumulated biomass in monoculture and expressed the
141 phenotypic response as relative yield (Methods). A negative relative yield indicates that a model
142 produces less biomass in co-culture than it does when growing alone, a relative yield of 0
143 indicates that its growth is unaffected by interaction with another model, and a positive relative
144 yield means that the model benefits from the presence of another. For the purpose of binary
145 classification, we divided observations into two classes based on their relative yields. Samples
146 with a negative relative yield were classified as 'negative' responders (5563/9900). Samples with
147 a positive relative yield (420/9900) or a relative yield of zero (3917/9900) were classified as 'non-
148 negative' responders (Figure 2B). We refer to the model whose response is being predicted as the
149 'responder' and its interaction partner is the 'partner'. In order to develop an intuition for how
150 difficult this classification task might be we calculated what fraction of the 99 interactions each
151 responder took part resulted in a negative relative yield. Classification could become a trivial task
152 if the distribution of these fractions were bimodal with peaks far from each other, the most
153 extreme scenario being one where responders experience either only negative or non-negative
154 relative yields in co-culture. On the other hand, a normal or uniform distribution would indicate a
155 more complicated reality where classification would benefit from machine learning. This last
156 option was indeed confirmed to hold for our data, as the histogram of the negative fractions
157 follows a truncated normal distribution (Figure 2C).
158
159 In order to establish a baseline for comparing the effectiveness of the random forest as a classifier,
160 we calculated the Jaccard distance of each interaction pair and then used the Jaccard distance as a

161 decision threshold for predicting whether a model would have a negative or non-negative relative
162 yield in co-culture. Jaccard distance was chosen as a baseline predictor because it is easy to
163 calculate and is commonly used in statistical analyses of microbial communities to infer
164 interactions (20, 21). The resulting receiver operator characteristic (ROC) curve was then
165 compared to the ROC curve obtained from a random forest classifier trained using the 194
166 exchange reactions as features to predict the outcome the full set of 9900 observations. As a
167 classifier, the Jaccard distance does not perform much better than randomly guessing, while the
168 random forest was surprisingly good (Figure 2D). To evaluate the accuracy of the random forest
169 classifier on the full data set we examined the out of bag (OOB) error estimates. OOB estimates
170 in random forests have been shown to be roughly equivalent to five-fold cross-validation (22). We
171 therefore took the OOB error rate as an estimate for the overall test error rate. The random forest
172 resulted in a balanced accuracy of 90.4%.
173
174 High predictive accuracies are encouraging but will be of little use if they can only be achieved
175 when the vast majority of the experiment outcomes are already known. Learning curves visualize
176 how the performance of a classifier behaves as the data available for trainings increases. We
177 constructed a series of learning curves to measure how the balanced accuracy of the random forest
178 classifier is affected by the size of the community being studied and as the fraction of interaction
179 outcomes known increases (Figure 2E). As is typically seen with learning curves, classifier
180 performance improves when more of the data is made available. Interestingly we observe that for
181 the smallest communities ($N_{organisms}$ = 10) balanced accuracy was better than random when as
182 little as 10% of the data was available (2 experiments), but the trajectory with additional data
183 improves at a much slower rate than it does for larger communities. When $N_{organisms}$ is increased to

184 20, significant improvements to balanced accuracy manifest immediately and have a stronger
185 benefit with the outcome of additional experiments. Encouragingly, these benefits appear
186 moderate when compared to the trajectory of much larger communities ($N_{organisms} \geq 30$). The
187 general trend indicates that the larger a community is, the smaller the relative fraction of
188 experiments needed to get a high accuracy. In fact, similar learning curves could be used as
189 guidelines to determine how many experiments should be performed to reach a desired balanced
190 accuracy.
191
192 Variable importance plots are commonly used with random forests to evaluate which variables are
193 the most important to the model by comparing their mean decrease in accuracy scores. Mean
194 decrease in accuracy for each variable is the change in the mean accuracy of the forest predictions
195 when the variable in question is randomly permuted (23). Variable importance is then determined
196 by computing the ratio between their mean change in accuracy and the corresponding standard
197 deviation of accuracy. The randomForest package for R (14) performs this calculation, which we
198 then used to rank exchange reactions as predictors. Variable importance should be interpreted as a
199 ranked list of which variables are generally the most informative rather than a quantification of
200 their effects. These variable rankings are indicators of global predictive importance and can be
201 used to develop insights into why certain features are influential in a general sense. Visualization
202 of the top 20 predictors revealed that the most important predictors tend to belong to the
203 interaction partner (Figure 2F). Interestingly, features from both halves of the interaction vector
204 for the amino acids L-Lysine and L-Glutamate are highly ranked, as are the features for the
205 monosaccharides D-Ribose and Thyminose.
206

207 The tree-based approach of random forest to classification can be exploited to determine why
208 specific samples were classified the way they were by examining the feature contributions of each
209 predictor. A feature contribution (Methods) is the quantification of how much a given variable
210 influences the decision of the random forest when a single sample is evaluated. Feature
211 contributions were originally developed for analysis of regression models (24) but have since
212 been adapted for binary classification models (15). In the context of binary classification, feature
213 contributions can be interpreted as how much a feature changes the probability that a given
214 sample is classified as class 1 by the random forest. In the case of our *in silico* data class 1 refers
215 to the non-negative responder type. For our purposes this convention means that a negative
216 feature contribution increases the probability of a sample being classified as a negative responder
217 and a positive feature contribution increases the probability of classified as a non-negative
218 responder. We used the forestFloor package in R (25) to efficiently calculate feature contributions
219 for all 9900 samples.
220
221 If the choice of representation faithfully reflects the underlying nature of the system of interest,
222 then feature contributions may be used to gain insight into the underlying mechanisms of an
223 interaction. Were we investigating our *in silico* data as a novel microbial community via pairwise
224 interactions, we would wish to perform additional experiments to identify a metabolite that
225 mediates competition in each negative interaction and then design further detailed experiments to
226 describe molecular details. Absent any particular insight into the nature of the community,
227 metabolites mediating any particular interaction would have to be identified by querying them
228 randomly. In situations where there are few metabolites mediating competition such a process
229 would be a costly endeavor. We wondered if the experimental load could be reduced if for each

230 sample we ranked the 194 metabolites by their net feature contributions and then used the
231 rankings as a guide for the order in which metabolites should be examined (Figure 3A).
232
233 Because the role of every extracellular metabolite in an interaction is known with COMETS, we
234 were able to evaluate the effectiveness of feature contributions as an experimental guide. For each
235 negative interaction, we identified which metabolites were subject to competition and calculated
236 how many metabolites we would expect to examine at random before encountering any one of the
237 contended metabolites. We then determined the guided rate of discovery by proceeding along the
238 ranked list as determined by net feature contributions and recorded the position at which we first
239 find one of the contended metabolites. Using feature contributions to guide discovery compares
240 favorably to a random selection process. For this community, one would expect on average to
241 perform 19.55 experiments before discovering a relevant metabolite through random selection.
242 Using feature contributions as a guide instead shifted the probability distribution to the left,
243 reducing the expected number of inquiries to 5.54 (Figure 3B). Early discovery of metabolites is
244 particularly valuable when there are very few metabolites mediating an interaction. In particular
245 we wanted to determine how effective feature contributions were in the most challenging cases,
246 those in which a single contended metabolite caused the negative response. We found that for
247 those 99 samples the contended metabolite was found at median rank 4 and in 95 cases was
248 encountered within the top 20 positions (Figure 3C).
249
250 Since random forest classifies based on patterns it finds in the data, we can expect that
251 metabolites that tend to be high-ranking predictors are also more likely to be broadly significant
252 to the community than metabolites encountered through random sampling. Across all negative

253 samples there were 109 metabolites that were subjected to competition in at least one interaction.
254 82 of those metabolites were the first ones encountered using feature contributions, but the
255 majority of encounters were concentrated among eight of them (Supplemental Table 1). D-
256 Fructose was the most commonly found metabolite, accounting for ~18.7% of all discovered
257 metabolites. Interestingly, fructose has been shown to be implicated in altering the gut
258 microbiome in connection to a number of diseases, including antibiotic treatable (26) metabolic
259 syndrome (27, 28), liver disease (29), and obesity (30).
260
261
262 Notably, our approach is also readily applicable for the discovery of metabolites that mediate
263 positive interactions, which comprise a small minority of all interactions in this specific dataset
264 (420/9900). Due to the scarcity of their occurrence and the dearth of metabolites that mediate
265 positive interactions, discovery of these mechanisms is more challenging. Nevertheless, using
266 ranked feature contributions to find facilitative metabolites was a powerful improvement over a
267 naive approach; requiring a median of 27 queries before discovery versus 65 when selecting
268 metabolites randomly (Supplemental Figure 1).
269
270 **Application to a Community of Auxotrophic *Escherichia coli* strains**
271 We analyzed the results of a study in which all-pairwise co-culture experiments for 14 *E. coli*
272 MG1655 auxotrophs were performed (31). In this study, auxotrophic *E. coli* strains were
273 generated by knocking out a single gene that was essential for the production of each of 14 amino
274 acids. Interactions between any given pair of *E. coli* strains are presumably dependent on the
275 direct exchange of the knocked out amino acids, or related precursors (Figure 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. We classified growth phenotypes based on the fold change response a given *E. coli* auxotroph strain had in co-culture with another auxotrophic strain. We used a fold change of 2 as a threshold in order to separate response classes. Fold change 2 is an intuitive threshold since doubling is often used when discussing growth of bacterial populations (32) and classifying all samples with a fold change ≤ 2 as a separate class from those with a fold change greater than 2 results in a balanced data set (Figure 4B). We refer to the 90 growth responses with fold change ≤ 2 as 'weak' responses and the 92 growth responses with fold change > 2 as 'strong' responses. When discussing a particular *E. coli* interaction, the strain whose response is being predicted is referred to as the 'receiver' and its interaction partner is the 'giver'. We calculated the fraction of interactions that resulted in a weak response for each of the 14 strains and visualized the distribution with a histogram (Figure 4C). Based on this histogram we expect that classification will not be a trivial task because by it appears that the distribution follows either a uniform (Kolmogorov-Smirnov test: $p \approx .93$, t test: $p \approx .95$) or truncated normal distribution (Kolmogorov-Smirnov test: $p \approx .33$, t test: $p \approx .96$).

It has been previously observed that biosynthetically costly amino acids tend to promote stronger cooperative interactions than biosynthetically cheap amino acids (31). Given this observation we tested if biosynthetic costs would be useful for predicting whether a strain will have a strong or a weak response in a given interaction. We used the biosynthetic cost of the amino acid provided by the giver strain, needed by the receiver strain, or the difference of these costs as decision thresholds to classify response classes and compared the performance of the random forest

classifier to these benchmarks. Since the auxotrophic *E. coli* strains were all derived from the same ancestral strain, only the biosynthetic capabilities for each of the 14 amino acids were relevant predictors for machine learning. As a result, vectors with just 28 elements represented each pairwise *E. coli* interaction. Examination of the corresponding ROC curves shows that biosynthetic costs of amino acids are poor predictors for the qualitative outcome of these *E. coli* auxotroph experiments whereas random forest fares much better as indicated by its ROC curve and by area under the curve (Figure 4D inset). Random forest yielded a balanced accuracy of ~79.2%. Moreover, the trajectory of the learning curve most closely resembles the trajectory of the learning curve for *in silico* communities of 20 members (Figure 4E). Variable importance rankings show that in general, the identity of the receiver's needed amino acid is often more impactful on classification accuracy than the amino acid that the giver needs, suggesting that specificity of interaction is dominated by auxotrophies, whereas most mutants can in principle provide the missing amino acid (Figure 4F).

Since the pair of knocked-out amino acids determines the interaction for any two auxotrophs, we evaluated how often the corresponding features were the most influential for classification. For each sample, we ranked the feature contributions by the magnitude of their influence and then identified which amino acids were ranked first and second. The feature space is relatively simple and the interaction outcomes are a direct consequence of the relevant auxotrophies. Therefore it is reasonable to expect that the random forest be more strongly influenced by the absence of a predictor than the presence. Of all 182 observations, the knocked out amino acid from the receiver had the largest feature contribution 140 times and the amino acid from the giver was the largest contributor 40 times (Supplemental Table 2). We further found that the top two ranked positions

322 were occupied by the giver and receiver's amino acids, in any order, 132 of 182 times. Thus the
323 pair of most influential predictors tended to correspond to the underlying mechanism of the
324 interaction, even in instances where the predicted class was incorrect.
325
326 Scenarios where the presumed mechanisms are the strongest contributors yet result in
327 misclassification present opportunities to direct research toward interesting outliers in order to
328 understand why they diverge from our expectations. The response of the methionine auxotroph
329 (**Δ**Met) to co-culture with the cysteine auxotroph (**Δ**Cys) was one such case, which is described in
330 detail in Supplemental Figure 2.
331

332 Discussion

333 Exhaustive pairwise co-culture studies of microbial communities are becoming increasingly
334 common. While such pairwise interactions do not necessarily capture all possible
335 interdependencies in a community (33, 34), they have been shown to be a dominant factor (35),
336 making the reliable prediction and interpretation of predictive models matters of great importance.
337 In this study, we have described a conceptual framework for the representation of microbes and
338 their pairwise interactions in order to address both of these challenges. Our results indicate that
339 representing genome-derived traits of microbes as binary vectors is sufficient for building reliable
340 classifiers for microbial interactions. We also demonstrated two methods for utilizing feature
341 contributions from random forest models to aid in the development of testable hypotheses
342 regarding interaction mechanisms.
343

344 Qualitatively predicting the outcome of unobserved interactions is most valuable if those
345 predictions allow precious resources and time to be conserved. To this end the construction of
346 learning curves is an important step in identifying how much data is required in order to achieve
347 desired prediction accuracy from machine learning. Our results indicate that prediction of
348 interactions in communities of all sizes may benefit from machine learning. In situations where
349 the prediction of interaction outcomes is not necessary the properties of random forest makes it a
350 good choice for deciding which traits to study first. Because predictions will always benefit from
351 more data we suggest that if the experimental space is not prohibitively large that all pairwise
352 experiments be performed and feature contributions then be used to derive testable hypotheses.
353 Conversely, when the experimental space is large then one should perform a subset of
354 experiments and incorporate the results of any additional experiments into the training set in order
355 to build new classifiers recursively. The larger the community being studied is the smaller the
356 relative fraction of possible experiments that must be initially performed, these results are
357 promising both for future studies of natural communities, which can contain upward of a thousand
358 unique community members (36) and the assembly of synthetic communities from large libraries
359 of microbes.
360
361 Feature contributions provide a clear window into the decision-making process of a random forest
362 when the underlying mechanisms are straightforward, as was the case for the *E. coli* auxotrophs.
363 In cases such as this, interpretation of the model is clear and novel detailed hypotheses can be
364 readily developed. Even incorrect predictions are useful since instances that go against the
365 expectations of the classifier are likely to be worth closer scrutiny. In situations where many
366 factors contribute in a complex manner to an interaction we demonstrated the utility of feature

367 contributions for guiding exploratory experiments by enriching underlying competitive
368 mechanisms near the top of ranked lists. In our particular application of feature contributions to
369 an *in silico* community we used their net effects to identify metabolites that are competed for in
370 negative cases or are used in cross feeding in positive interactions. The particular method of
371 evaluating feature contributions, summing the net effects or sorting them by magnitude is a
372 decision that must be made on a case-by-case basis.
373
374 Although we concatenated the binary trait vectors of two organisms in order to form a new
375 composite trait representation, alternative representations of microbes and their interactions
376 should also be explored. Predictions may also benefit from incorporating more information into
377 the base representation of a microbe such as gene copy number or mean transcriptional levels. We
378 also encourage future researchers to acquire information on the same interactions in multiple
379 environmental conditions. Varying the environmental conditions will allow abiotic factors to be
380 incorporated into the representation and should yield additional insight into situations where
381 certain combinations of traits are most relevant.
382
383 Going forward we expect that the importance of machine learning in microbial ecology will
384 continue to grow. The need to identify microbial strains that interact in a desired way in any given
385 environment will be one of the most pressing issues for synthetic ecology in the future. Likewise,
386 it would not be surprising if many novel methods for interpreting machine learning algorithms
387 emerge in response to the challenges of understanding the properties of microbial communities.
388

389 **Data Availability.** Pointers to Datasets obtained from previous work, and used in our analysis are
390 reported in the Materials and Methods Section.
391 The code necessary to reproduce all our figures and analyses is hosted at:
392 https://github.com/ddimucci/MicrobialCommunities
393

394 # Materials and Methods

395 **Auxotrophic *E. coli* interaction network data**. We obtained the measured growth response of
396 individual *E.coli* strains and biosynthetic costs of amino acids from the supplemental files
397 provided by (31).
398

399 **Simulation of *in silico* pairwise interactions among gut microbes with dFBA in COMETS.**
400 Metabolic reconstructions of human gut associated microbes were obtained from Bauer et al (18).
401 At the time of this writing these models can be downloaded directly from the following URL:
402 https://wwwen.uni.lu/content/download/86230/1056013/file/Bauer_et_al_301_microbe_models.ra
403 r
404 Each metabolic reconstruction encompasses the stoichiometry of virtually all metabolic reactions
405 present in an organism, including uptake/secretion. Flux balance analysis (FBA) is a constraint-
406 based steady state approach that uses this stoichiometry to predict fluxes and growth capacity
407 under a given boundary condition of nutrient availability, and has been described in detail before
408 (19, 37–39). Dynamic flux balance analysis (DFBA) (16) extends classical FBA to perform
409 dynamic simulations in which intracellular metabolites are still assumed to be at steady states, but
410 total biomass and environmental metabolites are treated as time-dependent variables in a
411 discretized approximation. We performed DFBA simulations using our platform for Computation

412 of Microbial Ecosystems in Time and Space (COMETS), which has been previously used to
413 model microbial communities (17) . COMETS allows users to implement DFBA in a two-
414 dimensional simulated world populated with multiple metabolic models. We selected 100
415 metabolic models (18) and identified a common medium that would permit the growth of nearly
416 all models in a monoculture scenario. We then performed all pairwise co-culture simulations of
417 the 100 models using the common media set in a well-mixed batch culture scenario
418 (approximated by using COMETS without spatial structure). For each scenario we saved the
419 record of biomass accumulation and fluxes in order to calculate relative yield and identify
420 mechanisms of interaction, respectively.
421
422 **Relative Yield.** Relative yield quantifies the change in net growth of an organism in a new
423 environment relative to a reference scenario. For our purpose we compared the growth of models
424 in co-culture to their accumulated growth in monoculture. Relative yield was calculated according
425 to the equation:
426
427 $$RY = (Biomass_{coculture} - Biomass_{monoculture})/Biomass_{monoculture}$$
428
429 An $RY < 0$ indicates that the responder is detrimentally affected by its partner. Correspondingly,
430 an RY of 0 indicates no effect and an $RY > 0$ indicates a positive affect from co-culture with the
431 partner.
432

Jaccard Distance. A profile ($model_i$) encoding the presence (1)/absence (0) of different metabolic reactions in each model i was identified based on its stoichiometric matrix. The Jaccard distance (JD) between two metabolic models i and j was determined with the equation:

$$\mathrm{JD}(model_i, model_j) = 1 - | model_i \cap model_j | / | model_i \cup model_j |$$

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 j is encoded into the element X_{ij} of an interaction matrix $\boldsymbol{X}$. To define a set of trait vectors for each organism in C, we start by obtaining a list of λ features that can be assigned systematically across all organisms. These features could be the presence/absence of specific genes, functions, or any other trait. In the two case studies reported here, the feature vectors correspond to the presence/absence of specific metabolite transporters in stoichiometric models, and to amino acid auxotrophies, respectively. Thus, each organism i is associated with an λ-long vector V_i, whose element k is 0 or 1 depending on whether or not trait k is absent or present in organism i. Traits (i.e. elements of the feature vector) that are identical across all organisms are removed from the feature set. Each pair of organisms (i,j) is then associated with a co-culture feature vector, defined as the concatenation of vectors V_i and V_j, also indicated as $V_{i,j} = [V_i , V_j]$ (see also Fig. 1). Note that in this concatenated vector, order matters. For the purpose of feeding the data into the random forest algorithm, the complete information about how organism i responds in a co-culture with organism j, is encoded in the composite feature vector $V_{i,j}$ and the observed phenotypic response for that specific pair, i.e. X_{ij}.

Machine learning. R was used to implement K-nearest neighbor (Knn), Random Forest (RF), and support vector machine (SVM). Knn was implemented with the rknn package, with K set to 3. Library e1071 was used for SVM and we used a linear kernel and cost=0.1. Random forest was implemented with the randomForest library. All three algorithms had similar performance on the *E. coli* data set but RF was significantly better on the *in silico* data set (data not shown).

Random forests are ensemble classifiers that aggregate the results of many individual decision trees. Each tree in a 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. Classification accuracy of individual trees is assessed by running their out of bag samples down the tree and recording the results. See (19) for a full description of the algorithm.

Evaluation of classifiers. We used balanced accuracy to evaluate performance of classifiers. 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:

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

Learning curves. To construct learning curves we defined a set of fractions, fr = [.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 fr we randomly selected $\max(1/fr_i, 1/(1-fr_i))$ mutually exclusive subsets of the data to use as a training set or as a hold out test set if $fr_i > .5$. This process was repeated until at least 10 subsets of the data were selected for each fraction. The median balanced accuracy of the classifier was then calculated at each point along fr. In order to investigate the effect of the community size on the learning curve we defined a set of community sizes S = [10, 20, 30, 40, 50, 60, 70, 80, 90]. For each community size S_i we randomly selected five community sub-matrices, X_k, from the full *in silico* community matrix (X). Then for each X_k a learning curve was determined and the median learning curve for balanced accuracy of each community size was calculated

Feature contributions for binary classifications. Each tree in the random forest is presented with a bootstrapped subset of the provided data as a training set that is used to grow the tree. The training set is placed at the root node of the tree. The fraction of instances in the training set of class C_1 is Y^n_{mean}. Y^n_{mean} is the probability that an instance selected at random from the root node is assigned to class C_1.

There are two steps in the calculation of feature contributions for a new sample: First calculate the sum of all local increments in each tree, second determine the average contribution over the forest. A local increment for a feature f is calculated if the split was performed on f and is defined as the change in probability of C_1 in a parent node (p) and a child node (c):

$$LI^c_f = Y^c_{mean} - Y^p_{mean}$$

Local increments quantify the change in the probability of a sample being classified as C_1 between the child node and the parent node when f is the splitting criterion. The feature contribution, $FC^f_{i,t}$ of feature f in tree t for observation i is the sum of all LI_f along the path from the root node to the terminal node. FC^f_i is given by the equation:

$$FC^f_i = 1/T \sum(FC^f_{i,t})$$

Feature contributions were calculated with the forestFloor package and are calculated using only trees for which a sample was out of bag.

Application to positive samples. Positive samples constituted a small minority of the total samples (420/9900) and result in a heavily imbalanced data set when we try to classify these samples as separate from either neutral or negative responses. Imbalanced datasets often result in very high classification accuracies because the classifier simply predicts every instance to be of the majority class. This is an issue because if we are to have confidence in feature contributions an adequately performing classifier is desired. In order to develop a reliable classifier we under sampled the non-positive instances by randomly selecting a subset of 420 of them to pair with the 420 positive observations and then trained a random forest on the new data set. To estimate the efficacy of the random forest we repeated this process 100 times and note that there was a median balanced accuracy of ≈ 85%. We then used a single random forest model to calculate feature contributions in the context of identifying positive responses.

Acknowledgments

We are grateful to Dr. Jenny Bhatnagar and Dr. Carolyn Zeiner whose work with fungal interactions served as the initial inspiration for this study and to members of the Segrè lab for helpful discussions and for feedback on the manuscript. DS and DD acknowledge funding from the Defense Advanced Research Projects Agency (Purchase Request No. HR0011515303, Contract No. HR0011-15-C-0091), the U.S. Department of Energy (DE-SC0012627), the NIH (T32GM100842, 5R01DE024468, R01GM121950 and Sub_P30DK036836_P&F), the National Science Foundation (1457695 and NSFOCE-BSF grant 1635070), the Human Frontiers Science Program (RGP0020/2016), and the Boston University Interdisciplinary Biomedical Research Office.

References

1. Røder HL, Sørensen SJ, Burmølle M. 2016. Studying Bacterial Multispecies Biofilms: Where to Start? Trends Microbiol.
2. Stein RR, Bucci V, Toussaint NC, Buffie CG, Rätsch G, Pamer EG, Sander C, Xavier JB. 2013. Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota. PLoS Comput Biol 9.
3. Aziz FAA, Suzuki K, Ohtaki A, Sagegami K, Hirai H, Seno J, Mizuno N, Inuzuka Y, Saito Y, Tashiro Y, Hiraishi A, Futamata H. 2015. Interspecies interactions are an integral determinant of microbial community dynamics. Front Microbiol 6.
4. Johns NI, Blazejewski T, Gomes ALC, Wang HH. 2016. Principles for designing synthetic microbial communities. Curr Opin Microbiol.

547 5. Ling LL, Schneider T, Peoples AJ, Spoering AL, Engels I, Conlon BP, Mueller A,
548 Schäberle TF, Hughes DE, Epstein S, Jones M, Lazarides L, Steadman VA, Cohen DR,
549 Felix CR, Fetterman KA, Millett WP, Nitti AG, Zullo AM, Chen C, Lewis K. 2015. A
550 new antibiotic kills pathogens without detectable resistance. Nature 517:455–459.
551 6. Wolfe BE. 2018. Using Cultivated Microbial Communities To Dissect Microbiome
552 Assembly: Challenges, Limitations, and the Path Ahead. mSystems 3:e00161-17.
553 7. Momeni B, Xie L, Shou W. 2017. Lotka-Volterra pairwise modeling fails to capture
554 diverse pairwise microbial interactions. Elife 6.
555 8. Goers L, Freemont P, Polizzi KM. 2014. Co-culture systems and technologies: taking
556 synthetic biology to the next level. J R Soc Interface 11:20140065–20140065.
557 9. Lasken RS, McLean JS. 2014. Recent advances in genomic DNA sequencing of microbial
558 species from single cells. Nat Rev Genet 15:577–584.
559 10. Gawad C, Koh W, Quake SR. 2016. Single-cell genome sequencing: Current state of the
560 science. Nat Rev Genet.
561 11. Seemann T. 2014. Prokka: Rapid prokaryotic genome annotation. Bioinformatics
562 30:2068–2069.
563 12. Dam P, Olman V, Harris K, Su Z, Xu Y. 2007. Operon prediction using both genome-
564 specific and general genomic information. Nucleic Acids Res 35:288–298.
565 13. Rahman SF, Olm MR, Morowitz MJ, Banfield JF. 2018. Machine Learning Leveraging
566 Genomes from Metagenomes Identifies Influential Antibiotic Resistance Genes in the
567 Infant Gut Microbiome. mSystems 3:e00123-17.
568 14. Liaw a, Wiener M. 2002. Classification and Regression by randomForest. R news 2:18–
569 22.

570 15. Palczewska A, Palczewski J, Robinson RM, Neagu D. 2013. Interpreting random forest
571 classification models using a feature contribution method (extended). 2013 IEEE 14th Int
572 Conf Inf Reuse Integr 1–30.
573 16. Henson MA, Hanly TJ. 2014. Dynamic flux balance analysis for synthetic microbial
574 communities. IET Syst Biol 8:214–229.
575 17. Harcombe WR, Riehl WJ, Dukovski I, Granger BR, Betts A, Lang AH, Bonilla G, Kar A,
576 Leiby N, Mehta P, Marx CJ, Segrè D. 2014. Metabolic resource allocation in individual
577 microbes determines ecosystem interactions and spatial dynamics. Cell Rep 7:1104–1115.
578 18. Bauer E, Laczny CC, Magnusdottir S, Wilmes P, Thiele I. 2015. Phenotypic
579 differentiation of gastrointestinal microbes is reflected in their encoded metabolic
580 repertoires. Microbiome 3:55.
581 19. Orth JD, Thiele I, Palsson BØ. 2010. What is flux balance analysis? Nat Biotechnol
582 28:245–248.
583 20. Bach EM, Williams RJ, Hargreaves SK, Yang F, Hofmockel KS. 2018. Greatest soil
584 microbial diversity found in micro-habitats. Soil Biol Biochem 118:217–226.
585 21. Mainali KP, Bewick S, Thielen P, Mehoke T, Breitwieser FP, Paudel S, Adhikari A,
586 Wolfe J, Slud E V., Karig D, Fagan WF. 2017. Statistical analysis of co-occurrence
587 patterns in microbial presence-absence datasets. PLoS One 12.
588 22. Svetnik V, Liaw A, Tong C, Christopher Culberson J, Sheridan RP, Feuston BP. 2003.
589 Random Forest: A Classification and Regression Tool for Compound Classification and
590 QSAR Modeling. J Chem Inf Comput Sci 43:1947–1958.
591 23. Breiman L. 2001. Random forests. Mach Learn 45:5–32.
592 24. Kuz'min VE, Polishchuk PG, Artemenko AG, Andronati SA. 2011. Interpretation of

593 QSAR models based on random forest methods. Mol Inform 30:593–603.
594 25. Welling SH, Refsgaard HHF, Brockhoff PB, Clemmensen LH. 2016. Forest Floor
595 Visualizations of Random Forests. arXiv.
596 26. Di Luccia B, Crescenzo R, Mazzoli A, Cigliano L, Venditti P, Walser JC, Widmer A,
597 Baccigalupi L, Ricca E, Iossa S. 2015. Rescue of fructose-induced metabolic syndrome by
598 antibiotics or faecal transplantation in a rat model of obesity. PLoS One 10.
599 27. Khitan Z, Kim DH. 2013. Fructose: A key factor in the development of metabolic
600 syndrome and hypertension. J Nutr Metab.
601 28. Bantle JP. 2009. Dietary Fructose and Metabolic Syndrome and Diabetes. J Nutr
602 139:1263S–1268S.
603 29. Lambertz J, Weiskirchen S, Landert S, Weiskirchen R. 2017. Fructose: A dietary sugar in
604 crosstalk with microbiota contributing to the development and progression of non-
605 alcoholic liver disease. Front Immunol.
606 30. Payne AN, Chassard C, Lacroix C. 2012. Gut microbial adaptation to dietary consumption
607 of fructose, artificial sweeteners and sugar alcohols: Implications for host-microbe
608 interactions contributing to obesity. Obes Rev 13:799–809.
609 31. Mee MT, Collins JJ, Church GM, Wang HH. 2014. Syntrophic exchange in synthetic
610 microbial communities. Proc Natl Acad Sci 111:E2149–E2156.
611 32. POWELL EO. 1956. Growth Rate and Generation Time of Bacteria, with Special
612 Reference to Continuous Culture. J Gen Microbiol 15:492–511.
613 33. Bairey E, Kelsic ED, Kishony R. 2016. High-order species interactions shape ecosystem
614 diversity. Nat Commun 7.
615 34. Levine JM, Bascompte J, Adler PB, Allesina S. 2017. Beyond pairwise mechanisms of

616 species coexistence in complex communities. Nature.
617 35. Venturelli OS, Carr AC, Fisher G, Hsu RH, Lau R, Bowen BP, Northern T, Arkin AP.
618 2017. Deciphering microbial interactions in synthetic human gut microbiome
619 communities. bioArxiv.
620 36. Raynaud X, Nunan N. 2014. Spatial ecology of bacteria at the microscale in soil. PLoS
621 One 9.
622 37. Raman K, Chandra N. 2009. Flux balance analysis of biological systems: Applications
623 and challenges. Brief Bioinform.
624 38. Maarleveld TR, Khandelwal RA, Olivier BG, Teusink B, Bruggeman FJ. 2013. Basic
625 concepts and principles of stoichiometric modeling of metabolic networks. Biotechnol J.
626 39. O'Brien EJ, Monk JM, Palsson BO. 2015. Using genome-scale models to predict
627 biological capabilities. Cell 161:971–987.
628 40. Park S, Imlay JA. 2003. High levels of intracellular cysteine promote oxidative DNA
629 damage by driving the Fenton reaction. J Bacteriol 185:1942–1950.
630 41. Takumi K, Nonaka G. 2016. Bacterial cysteine-inducible cysteine resistance systems. J
631 Bacteriol 198:1384–1392.
632

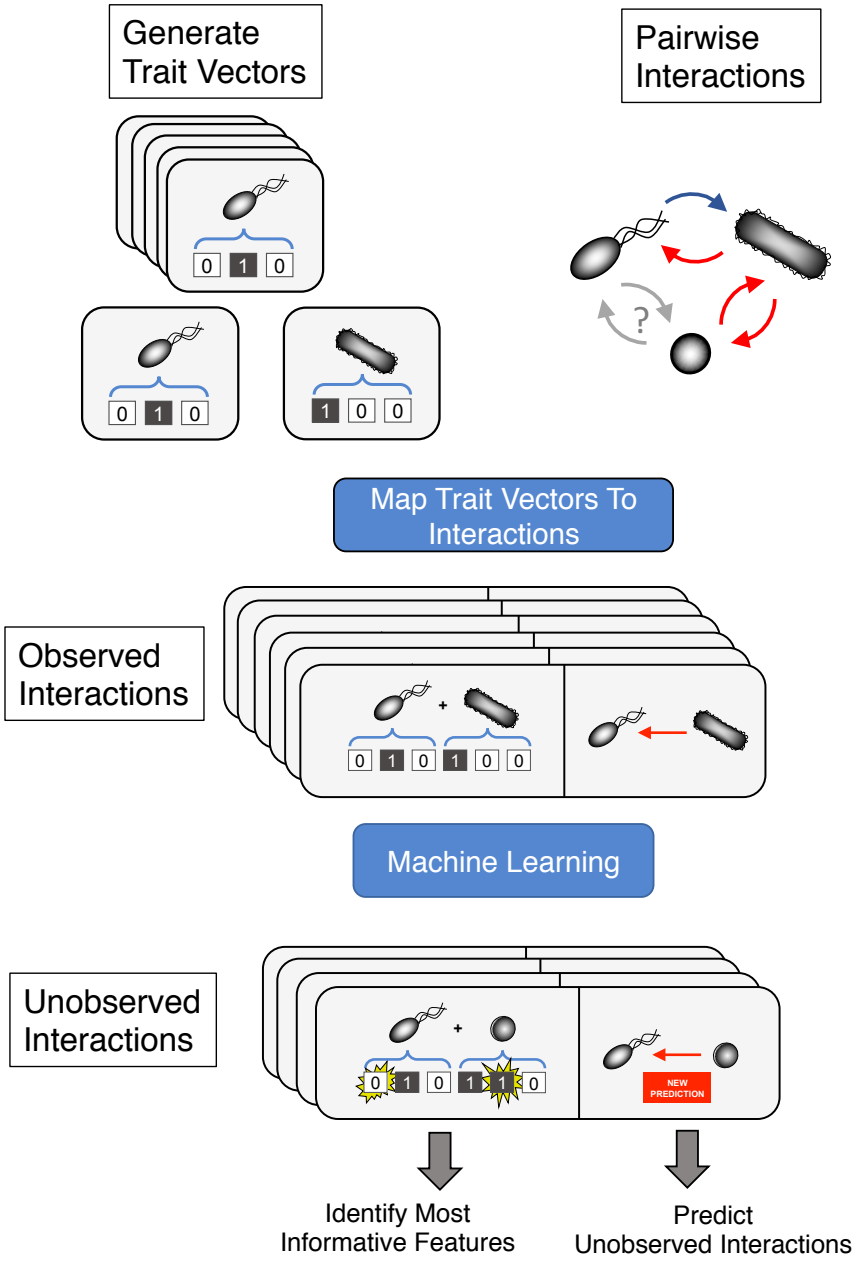


Figure 1. A schematic representation of our machine learning approach for inferring interactions among microbes. A trait vector captures the characteristics of each organism in the community of interest. The presence or absence of a trait in a given organism is encoded (as a binary number) in the corresponding 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.

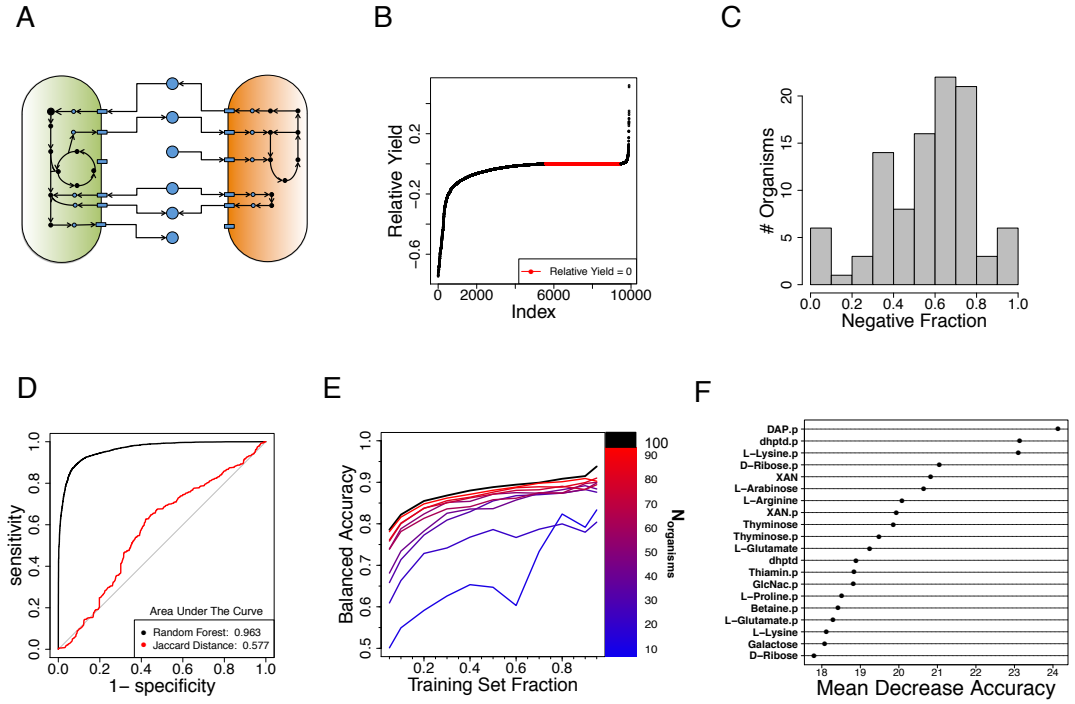
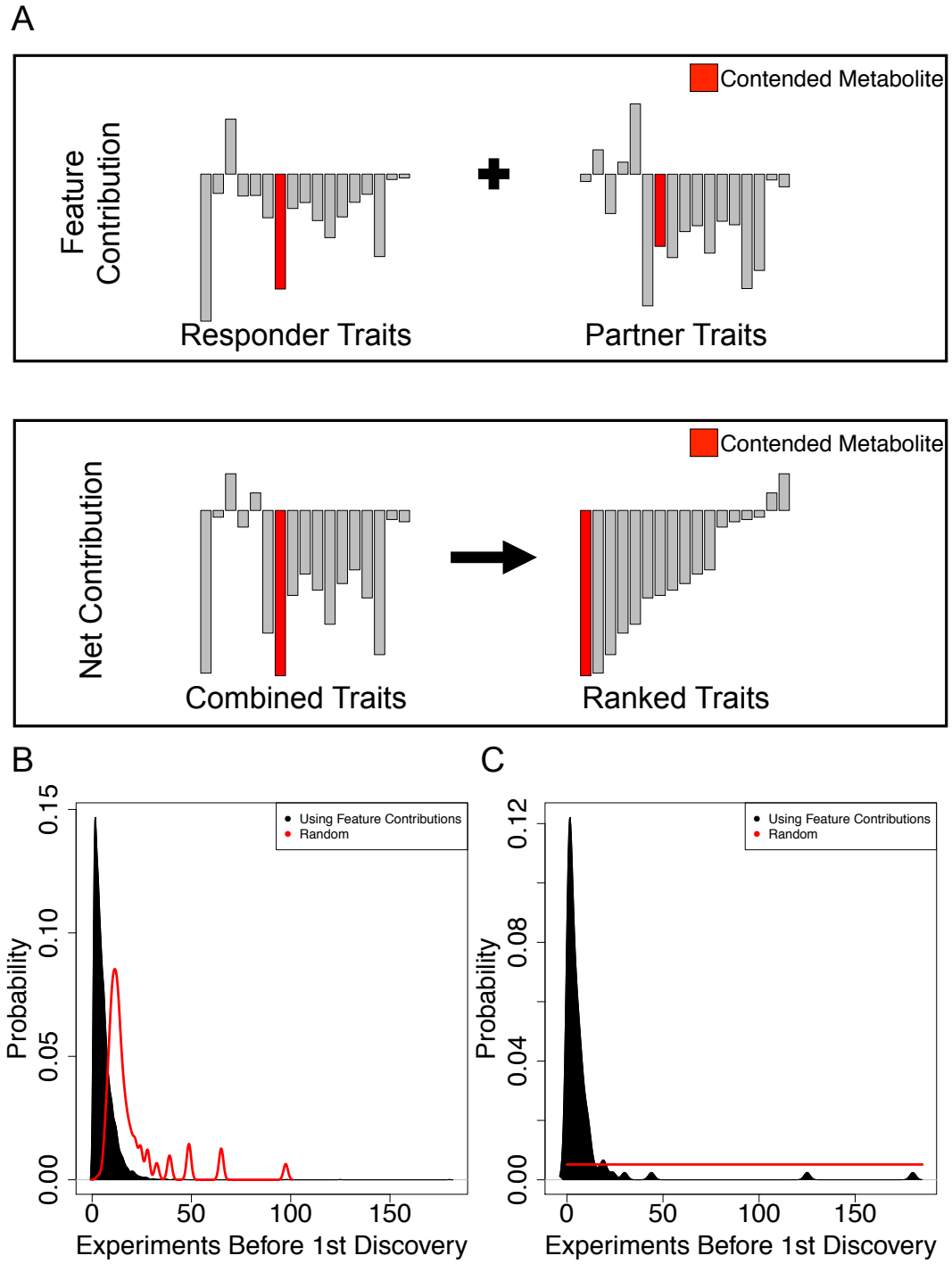


Figure 2. Classification of interactions in an *in silico* model of a community of human gut microbes. **A.** Organisms are represented *in silico* as large networks of metabolic reactions that take up metabolites (blue circles) from the environment (arrows leading to model) and release by-products (arrows 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

654 where one model consumes a by-product of the other. **B.** The distribution of relative yields is
655 sorted and plotted. Interactions resulted in a model producing a negative relative yield 5563 times.
656 Neutral interactions, a relative yield of zero, occurred 3917 times, and positive relative yield
657 happened 420 times. Samples were classified as negative or non-negative. **C.** Histogram of the
658 fraction of interactions that were the negative class for each organism. The distribution is
659 consistent with a truncated normal distribution centered on .562 (t test, $p \approx .67$) **D.** ROC curve of
660 a random forest classifier using 388 exchange reactions as predictors for 9900 *in silico*
661 observations compared to the ROC curve obtained from using Jaccard Distance as a threshold to
662 predict negative versus non-negative relative yield. **E.** Learning curves for sub-communities of
663 the full *in silico* community. These learning curves are the median learning curve for 5 sub-
664 communities selected at random for each value of $N_{organisms}$. **F.** 20 most influential predictors as
665 determined by mean decrease in accuracy. The '.p' suffix indicates that the predictor belongs to
666 the interaction partner. Some metabolite names are shortened: DAP = meso-2,6-Diaminopimelate,
667 dhptd = 4-5-dihydroxy-2-3-pentanedione, XAN = xanthine, GlcNac = N-Acetylglucosamine.
668



669

670 **Figure 3.** Using feature contributions to find a metabolite for which two organisms compete (also

671 referred to as a contended metabolite) **A.** The first half of the composite trait vector corresponds

672 to metabolite transporters belonging to the organism of interest while the second half corresponds
673 to metabolite transporters belonging to its interaction partner. We were interested in identifying a
674 metabolite that is associated with the negative relative yield for the organism of interest. To
675 establish a ranking of metabolites we took the summation of feature contributions from both
676 halves of the composite trait vector and then sorted the new vector according to the net
677 contribution. Proceeding from the negative end, the rank and identity of the first contended
678 metabolite encountered relative to the negative end of the new vector was recorded. **B.** The
679 probability distribution of the average rank at which the first contended metabolite would be
680 encountered by sampling metabolites randomly one at a time was calculated for each sample and
681 compared to the observed probability distribution. By chance the first metabolite would be
682 encountered after 13 queries. Feature contributions reduce the median number of queries to 4. **C.**
683 99 samples produced a negative relative through competition for exactly one metabolite.
684 Randomly investigating the 194 candidate metabolites one at a time results in an average of 97.5
685 experiments before discovering the contended metabolite. Using feature contributions to prioritize
686 the order in which to investigate metabolites instead would typically reveal the contended
687 metabolite on or before the fourth experiment (median = 4).
688

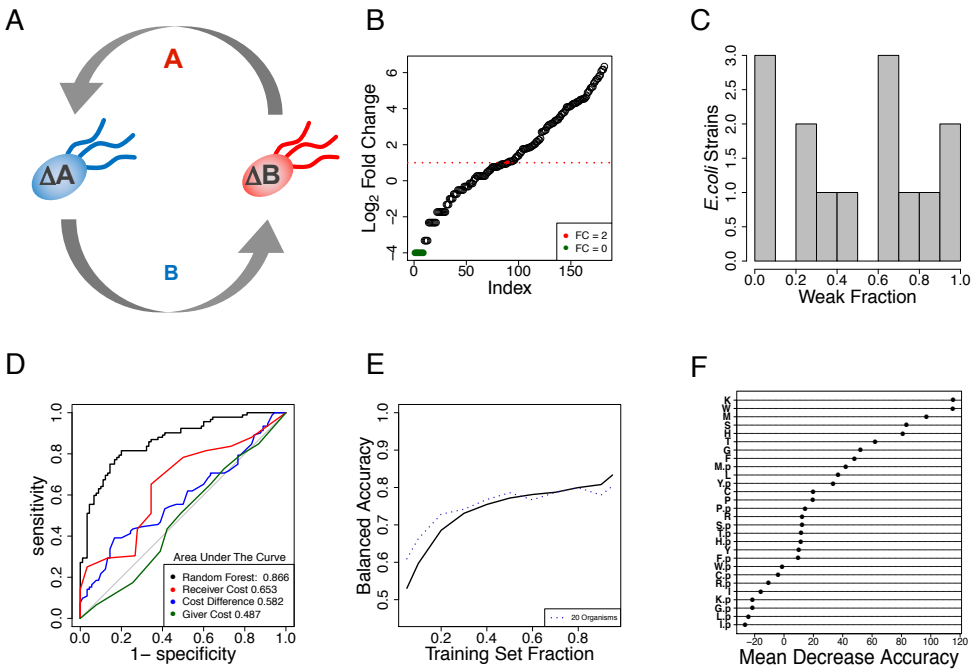


Figure 4. Data representation and results for the case study of a network of auxotrophic *E. coli* strains. **A.** In the original experiment, single gene knockout *E. coli* auxotrophs were co-cultured in a minimal medium. In order for ΔA to grow it must receive amino acid A from ΔB, which in turn must 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.** Sorted fold change for each auxotroph strain across all experiments. *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.** Histogram of weak responses as a fraction of each strains' total number of interactions. **D.** ROC curve for random forest classifier using 28 amino acids as predictors for 182 observations. Single value thresholds based on the biosynthetic costs of knocked out amino acids resulted in poorer performance than random forest. **E.** The trajectory of a learning curve built for the *E. coli* interactions closely

703 resembles that of the learning curve for *in silico* communities with 20 organisms. **F.** The 28
704 amino acids ranked according to their affects on prediction accuracy when randomly permuted.
705 Amino acids corresponding to the receiver strain are enriched near the top of the list. Amino acids
706 are represented by their single letter codes. The suffix '.p' indicates that the predictive feature
707 belongs to the giver strain.