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: Prediction of Pairwise Microbial Interaction Outcomes

[#] 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). 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 (2) and for the design of synthetic communities (3).

70 A powerful, and increasingly employed method for assessing intermicrobial interactions is
71 the direct measurement of phenotypes of microbial species grown in co-culture (4). A
72 fundamental challenge in this endeavor is the huge diversity of many natural communities,
73 which could count up to several hundred strains or species of microbes. Performing
74 experiments for all possible pairwise interactions constitutes a herculean, and likely
75 insurmountable task for even a moderately sized community. It is however, conceivable
76 that new computational approaches could systematically complement existing tools such
77 as high-throughput sequencing and genome annotation (5–9) to help extract as much
78 information as possible from interaction datasets, providing both insight on yet-to-be-
79 measured interactions, and on possible biological mechanisms mediating specific
80 partnerships.
81
82 Here we present a conceptual framework for the mathematical representation of microbial
83 interactions and subsequent use of supervised learning to build a classifier with high
84 predictive accuracy. While any algorithm may be used, we obtained our best results with
85 random forest (10). Random forests are ensembles of many decision trees that individually
86 are poor classifiers but can be democratically pooled to create a very good classifier.
87 Random forests have two attributes that we found particularly attractive for our purposes
88 here. First, they are non-parametric and thus require no *a priori* definition about
89 underlying relationships between predictive variables. Second, recent methodological
90 developments in the interpretation of random forests have been made that allow users to
91 query why a specific example was classified the way that it was through the calculation of
92 feature contributions (11). Feature contributions can be exploited to develop new

93 hypotheses about the mechanisms that mediate specific interactions. In order to
94 demonstrate a proof of principle for the classification of microbial interactions using
95 organism traits and the utility of feature contributions for developing insight into the
96 underlying mechanisms, we applied this approach to two communities where all pairwise
97 experiments had been performed and the mechanisms of interaction identified. The first is
98 an *in silico* community of 100 metabolic models of human gut associated bacteria and the
99 second are the experimental results of a study involving 14 amino acid auxotrophic strains
100 of *Escherichia coli* (*E. coli*). Our results show that the application of random forests and
101 feature contributions to the study of microbial communities can significantly increase the
102 capacity to estimate a large number of interactions based on a limited number of
103 experiments, and has the potential to enable the discovery of new interaction mechanisms.
104

105 # Results

106 **Representing Pairwise Interactions**

107 Our objective in this study was twofold; first, we sought to predict the qualitative outcome
108 of unobserved pairwise interactions in a microbial community; second, we wanted to
109 identify predictive variables that suggest potential mechanisms of interaction. In order to
110 achieve both of these goals it was important to establish a representation that enables an
111 algorithm to make good predictions and also be easily parsed for interpretation. Our
112 approach relies of the availability of a trait-level description available for each organism in
113 the community under consideration. Trait descriptions are used to construct trait vectors
114 for each organism. Specific interactions are represented as the concatenation of the
115 relevant trait vectors (Fig 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.

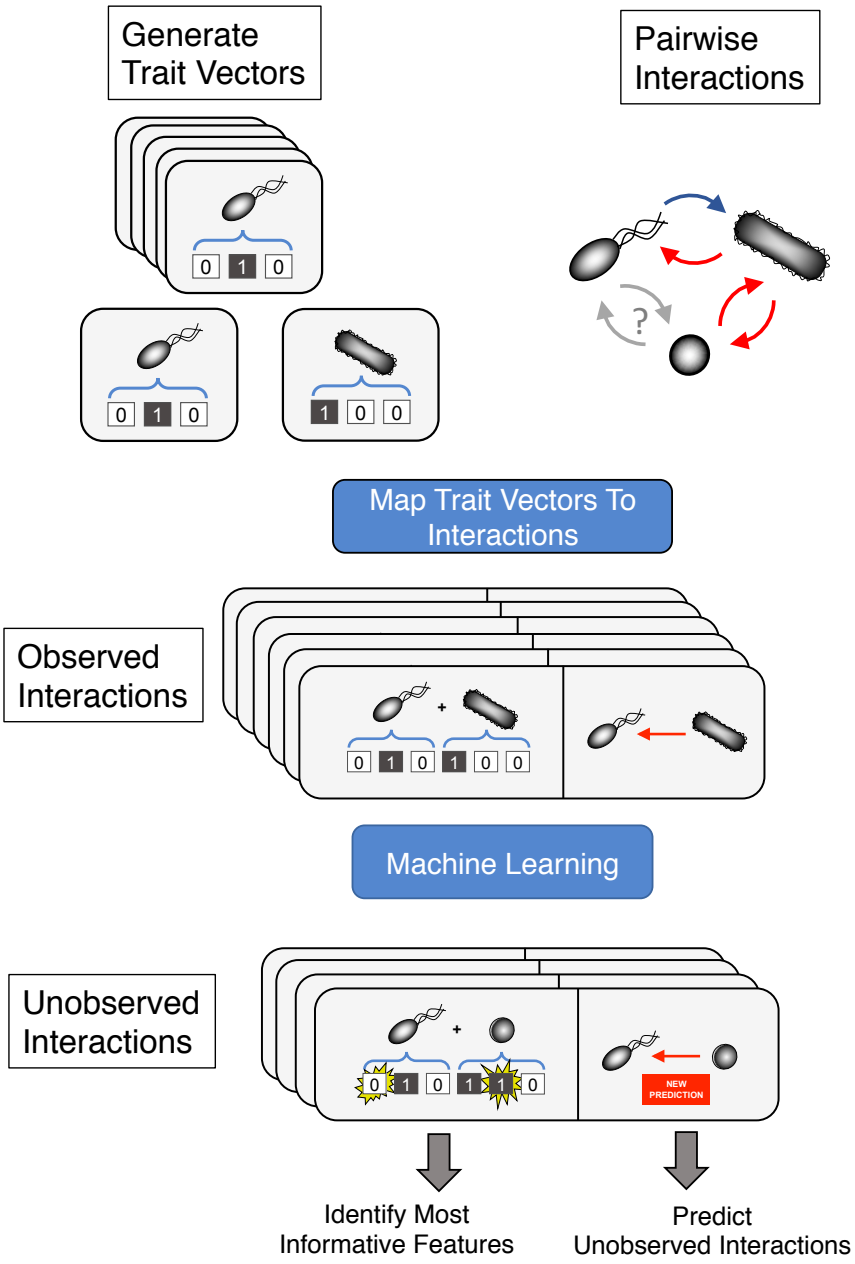


Figure 1. A schematic representation of our machine learning approach for inferring interactions among microbes. A trait vector captures the characteristics of each organism in the community of interest. The presence or absence of a trait in a given organism is

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.

These trait vectors, together with the known outcome of specific interactions, can be fed into a machine learning algorithm capable of absorbing existing patterns and extending predictions to unknown interactions. The approach we use here, a random forests classifier, is 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. Single trees tend to be poor classifiers; however, when their decisions are pooled, the collective predictive power is often impressive. In order to find potential mechanisms of interaction we took advantage of the binary nature of individual trees in order to identify which variables are the most influential for the classification of specific samples.

Application to computationally predicted interactions between human gut microbes

We first applied our approach to a large *in silico* data set that we generated by simulating with dynamic flux balance analysis (12) (using COMETS (13), see Methods) all pairwise co-culture interactions between 100 metabolic models of human-gut associated bacteria (14) under rich medium, in a well-mixed batch culture. Each metabolic model is a network of over a thousand interconnected metabolic reactions that fall into two categories: intracellular and exchange reactions (15). Intracellular reactions correspond to the various

pathways found within the cell, while exchange reactions represent the ability of an organism to transport a metabolite across the cell membrane. Among the 100 organisms, there were a total of 2083 unique metabolic reactions that could serve as potential predictors. Out of these, 194 were exchange reactions. Since interactions between organisms emerge through the exchange of or competition for one or more extracellular metabolites in the environment, we used only the exchange reactions as predictive features (Fig 2A).

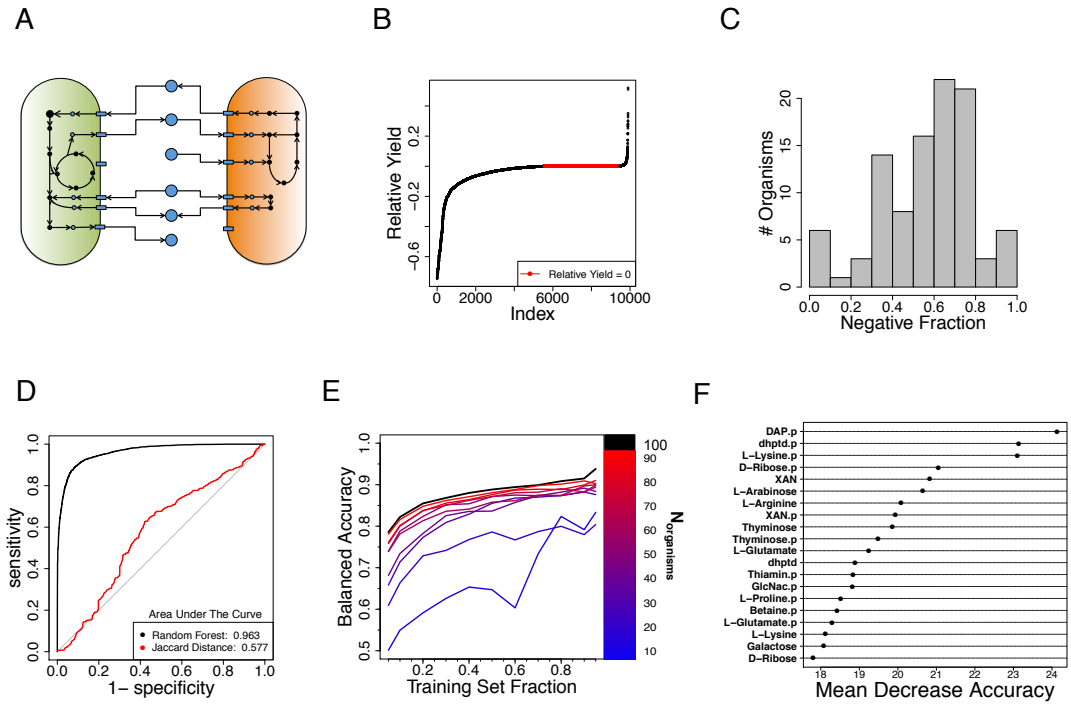


Figure 2. Classification of interactions in an *in silico* model of a community of human gut microbes. **A.** Organisms are represented *in silico* as large networks of metabolic reactions that take up metabolites (blue circles) from the environment (arrows leading to model) and release by-products (arrows 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 where one model consumes a by-product of the other. **B.** The distribution of relative yields is sorted and plotted. Interactions resulted in a model producing a negative relative yield 5563 times. 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.** Histogram of the fraction of interactions that were the negative class for each organism. The distribution is consistent with a truncated

171 normal distribution centered on .562 (t test, $p \approx .67$) **D.** ROC curve of a random forest
172 classifier using 388 exchange reactions as predictors for 9900 *in silico* observations
173 compared to the ROC curve obtained from using Jaccard Distance as a threshold to predict
174 negative versus non-negative relative yield. **E.** Learning curves for sub-communities of
175 the full *in silico* community. These learning curves are the median learning curve for 5
176 sub-communities selected at random for each value of $N_{organisms}$. **F.** 20 most influential
177 predictors as determined by mean decrease in accuracy. The '.p' suffix indicates that the
178 predictor belongs to the interaction partner. Some metabolite names are shortened: DAP =
179 meso-2,6-Diaminopimelate, dhptd = 4-5-dihydroxy-2-3-pentanedione, XAN = xanthine,
180 GlcNac = N-Acetylglucosamine.
181

182

183 Interactions in the network were computed by determining the influence of each organism
184 on each other organism in co-culture simulations. We compared the accumulated biomass
185 of each model in co-culture experiment to its accumulated biomass in monoculture and
186 expressed the phenotypic response as relative yield (Methods). A negative relative yield
187 indicates that a model produces less biomass in co-culture than it does when growing
188 alone, a relative yield of 0 indicates that its growth is unaffected by interaction with
189 another model, and a positive relative yield means that the model benefits from the
190 presence of another. For the purpose of binary classification, we divided observations into
191 two classes based on their relative yields. Samples with a negative relative yield were
192 classified as 'negative' responders (5563/9900). Samples with a positive relative yield
193 (420/9900) or a relative yield of zero (3917/9900) were classified as 'non-negative'
194 responders (Fig 2B). We refer to the model whose response is being predicted as the
195 'responder' and its interaction partner is the 'partner'. In order to develop an intuition for
196 how difficult this classification task might be we calculated what fraction of the 99
197 interactions each responder took part resulted in a negative relative yield. Classification
198 could become a trivial task if the distribution of these fractions were bimodal with peaks

199 far from each other, the most extreme scenario being one where responders experience
200 either only negative or non-negative relative yields in co-culture. On the other hand, a
201 normal or uniform distribution would indicate a more complicated reality where
202 classification would benefit from machine learning. This last option was indeed confirmed
203 to hold for our data, as the histogram of the negative fractions follows a truncated normal
204 distribution (Fig 2C).
205
206 In order to establish a baseline for comparing the effectiveness of the random forest as a
207 classifier, we calculated the Jaccard distance of each interaction pair and then used the
208 Jaccard distance as a decision threshold for predicting whether a model would have a
209 negative or non-negative relative yield in co-culture. Jaccard distance was chosen as a
210 baseline predictor because it is easy to calculate and is commonly used in statistical
211 analyses of microbial communities to infer interactions (16, 17). The resulting receiver
212 operator characteristic (ROC) curve was then compared to the ROC curve obtained from a
213 random forest classifier trained using the 194 exchange reactions as features to predict the
214 outcome the full set of 9900 observations. As a classifier, the Jaccard distance does not
215 perform much better than randomly guessing, while the random forest was surprisingly
216 good (Fig 2D). To evaluate the accuracy of the random forest classifier on the full data set
217 we examined the out of bag (OOB) error estimates. OOB estimates in random forests have
218 been shown to be roughly equivalent to five-fold cross-validation (18). We therefore took
219 the OOB error rate as an estimate for the overall test error rate. The random forest resulted
220 in a balanced accuracy of 90.4%.
221

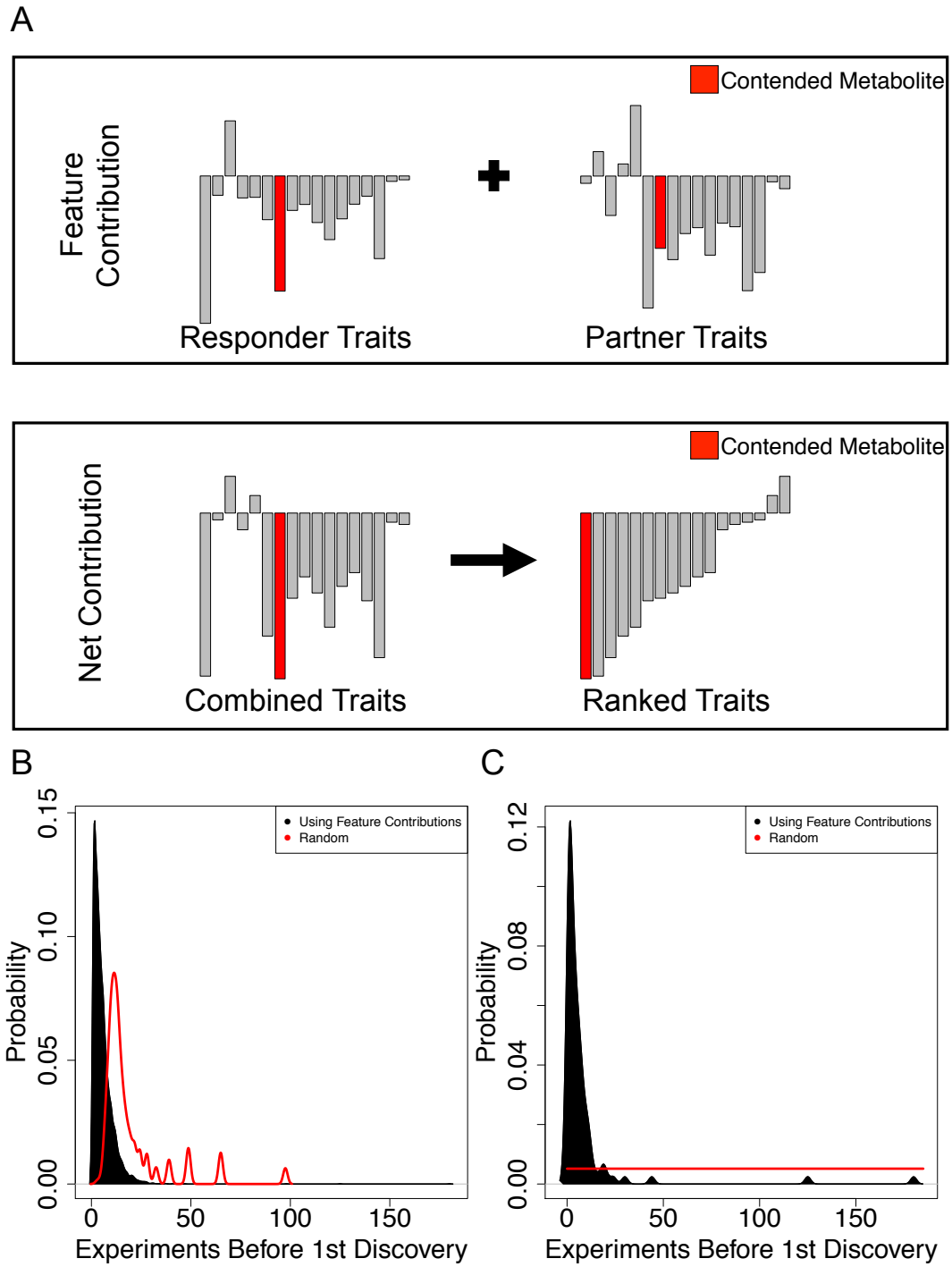
222 High predictive accuracies are encouraging but will be of little use if they can only be
223 achieved when the vast majority of the experiment outcomes are already known. Learning
224 curves visualize how the performance of a classifier behaves as the data available for
225 trainings increases. We constructed a series of learning curves to measure how the
226 balanced accuracy of the random forest classifier is affected by the size of the community
227 being studied and as the fraction of interaction outcomes known increases (Fig 2E). As is
228 typically seen with learning curves, classifier performance improves when more of the
229 data is made available. Interestingly we observe that for the smallest communities
230 ($N_{organisms}$ = 10) balanced accuracy was better than random when as little as 10% of the
231 data was available (2 experiments), but the trajectory with additional data improves at a
232 much slower rate than it does for larger communities. When $N_{organisms}$ is increased to 20,
233 significant improvements to balanced accuracy manifest immediately and have a stronger
234 benefit with the outcome of additional experiments. Encouragingly, these benefits appear
235 moderate when compared to the trajectory of much larger communities ($N_{organisms} \geq 30$).
236 The general trend indicates that the larger a community is, the smaller the relative fraction
237 of experiments needed to get a high accuracy. In fact, similar learning curves could be
238 used as guidelines to determine how many experiments should be performed to reach a
239 desired balanced accuracy.
240
241 Variable importance plots are commonly used with random forests to evaluate which
242 variables are the most important to the model by comparing their mean decrease in
243 accuracy scores. Mean decrease in accuracy for each variable is the change in the mean
244 accuracy of the forest predictions when the variable in question is randomly permuted

245 (19). Variable importance is then determined by computing the ratio between their mean
246 change in accuracy and the corresponding standard deviation of accuracy. The
247 randomForest package for R (10) performs this calculation, which we then used to rank
248 exchange reactions as predictors. Variable importance should be interpreted as a ranked
249 list of which variables are generally the most informative rather than a quantification of
250 their effects. These variable rankings are indicators of global predictive importance and
251 can be used to develop insights into why certain features are influential in a general sense.
252 Visualization of the top 20 predictors revealed that the most important predictors tend to
253 belong to the interaction partner (Fig 2F). Interestingly, features from both halves of the
254 interaction vector for the amino acids L-Lysine and L-Glutamate are highly ranked, as are
255 the features for the monosaccharides D-Ribose and Thyminose.
256
257 The tree-based approach of random forest to classification can be exploited to determine
258 why specific samples were classified the way they were by examining the feature
259 contributions of each predictor. A feature contribution (Methods) is the quantification of
260 how much a given variable influences the decision of the random forest when a single
261 sample is evaluated. Feature contributions were originally developed for analysis of
262 regression models (20) but have since been adapted for binary classification models (11).
263 In the context of binary classification, feature contributions can be interpreted as how
264 much a feature changes the probability that a given sample is classified as class 1 by the
265 random forest. In the case of our *in silico* data class 1 refers to the non-negative responder
266 type. For our purposes this convention means that a negative feature contribution increases
267 the probability of a sample being classified as a negative responder and a positive feature

268 contribution increases the probability of classified as a non-negative responder. We used
269 the forestFloor package in R (21) to efficiently calculate feature contributions for all 9900
270 samples.
271
272 If the choice of representation faithfully reflects the underlying nature of the system of
273 interest, then feature contributions may be used to gain insight into the underlying
274 mechanisms of an interaction. Were we investigating our *in silico* data as a novel
275 microbial community via pairwise interactions, we would wish to perform additional
276 experiments to identify a metabolite that mediates competition in each negative interaction
277 and then design further detailed experiments to describe molecular details. Absent any
278 particular insight into the nature of the community, metabolites mediating any particular
279 interaction would have to be identified by querying them randomly. In situations where
280 there are few metabolites mediating competition such a process would be a costly
281 endeavor. We wondered if the experimental load could be reduced if for each sample we
282 ranked the 194 metabolites by their net feature contributions and then used the rankings as
283 a guide for the order in which metabolites should be examined (Fig 3A).
284
285 Because the role of every extracellular metabolite in an interaction is known with
286 COMETS, we were able to evaluate the effectiveness of feature contributions as an
287 experimental guide. For each negative interaction, we identified which metabolites were
288 subject to competition and calculated how many metabolites we would expect to examine
289 at random before encountering any one of the contended metabolites. We then determined
290 the guided rate of discovery by proceeding along the ranked list as determined by net

291 feature contributions and recorded the position at which we first find one of the contended
292 metabolites. Using feature contributions to guide discovery compares favorably to a
293 random selection process. For this community, one would expect on average to perform
294 19.55 experiments before discovering a relevant metabolite through random selection.
295 Using feature contributions as a guide instead shifted the probability distribution to the
296 left, reducing the expected number of inquiries to 5.54 (Fig 3B). Early discovery of
297 metabolites is particularly valuable when there are very few metabolites mediating an
298 interaction. In particular we wanted to determine how effective feature contributions were
299 in the most challenging cases, those in which a single contended metabolite caused the
300 negative response. We found that for those 99 samples the contended metabolite was
301 found at median rank 4 and in 95 cases was encountered within the top 20 positions (Fig
302 3C).
303
304 Since random forest classifies based on patterns it finds in the data, we can expect that
305 metabolites that tend to be high-ranking predictors are also more likely to be broadly
306 significant to the community than metabolites encountered through random sampling.
307 Across all negative samples there were 109 metabolites that were subjected to competition
308 in at least one interaction. 82 of those metabolites were the first ones encountered using
309 feature contributions, but the majority of encounters were concentrated among eight of
310 them (Supplemental Table 1). D-Fructose was the most commonly found metabolite,
311 accounting for ~18.7% of all discovered metabolites. Interestingly, fructose has been
312 shown to be implicated in altering the gut microbiome in connection to a number of
313 diseases, including antibiotic treatable (22) metabolic syndrome (23, 24), liver disease

314 (25), and obesity (26).



315

316

317 **Figure 3.** Using feature contributions to find a metabolite for which two organisms
318 compete (also referred to as a contended metabolite) **A.** The first half of the composite
319 trait vector corresponds to metabolite transporters belonging to the organism of interest
320 while the second half corresponds to metabolite transporters belonging to its interaction
321 partner. We were interested in identifying a metabolite that is associated with the negative
322 relative yield for the organism of interest. To establish a ranking of metabolites we took
323 the summation of feature contributions from both halves of the composite trait vector and
324 then sorted the new vector according to the net contribution. Proceeding from the negative
325 end, the rank and identity of the first contended metabolite encountered relative to the
326 negative end of the new vector was recorded. **B.** The probability distribution of the
327 average rank at which the first contended metabolite would be encountered by sampling
328 metabolites randomly one at a time was calculated for each sample and compared to the
329 observed probability distribution. By chance the first metabolite would be encountered
330 after 13 queries. Feature contributions reduce the median number of queries to 4. **C.** 99
331 samples produced a negative relative through competition for exactly one metabolite.
332 Randomly investigating the 194 candidate metabolites one at a time results in an average
333 of 97.5 experiments before discovering the contended metabolite. Using feature
334 contributions to prioritize the order in which to investigate metabolites instead would
335 typically reveal the contended metabolite on or before the fourth experiment (median = 4).
336

337

338 Notably, our approach is also readily applicable for the discovery of metabolites that
339 mediate positive interactions, which comprise a small minority of all interactions in this
340 specific dataset (420/9900). Due to the scarcity of their occurrence and the dearth of
341 metabolites that mediate positive interactions, discovery of these mechanisms is more
342 challenging. Nevertheless, using ranked feature contributions to find facilitative
343 metabolites was a powerful improvement over a naive approach; requiring a median of 27
344 queries before discovery versus 65 when selecting metabolites randomly (Supplemental
345 Figure 1).

346

347 **Application to a Community of Auxotrophic *Escherichia coli* strains**

348 We analyzed the results of a study in which all-pairwise co-culture experiments for 14 *E.*
349 *coli* MG1655 auxotrophs were performed (27). In this study, auxotrophic *E. coli* strains

350 were generated by knocking out a single gene that was essential for the production of each
351 of 14 amino acids. Interactions between any given pair of *E. coli* strains are presumably
352 dependent on the direct exchange of the knocked out amino acids, or related precursors
353 (Fig 4A). The total growth of each strain in the 91 experiments was measured after 84
354 hours and reported as the net fold change relative to the initial inoculum, resulting in 182
355 total observations. We classified growth phenotypes based on the fold change response a
356 given *E. coli* auxotroph strain had in co-culture with another auxotrophic strain. We used a
357 fold change of 2 as a threshold in order to separate response classes. Fold change 2 is an
358 intuitive threshold since doubling is often used when discussing growth of bacterial
359 populations (28) and classifying all samples with a fold change ≤ 2 as a separate class
360 from those with a fold change greater than 2 results in a balanced data set (Fig 4B). We
361 refer to the 90 growth responses with fold change ≤ 2 as 'weak' responses and the 92
362 growth responses with fold change > 2 as 'strong' responses. When discussing a particular
363 *E. coli* interaction, the strain whose response is being predicted is referred to as the
364 'receiver' and its interaction partner is the 'giver'. We calculated the fraction of
365 interactions that resulted in a weak response for each of the 14 strains and visualized the
366 distribution with a histogram (Fig 4C). Based on this histogram we expect that
367 classification will not be a trivial task because by it appears that the distribution follows
368 either a uniform (Kolmogorov-Smirnov test: $p \approx .93$, t test: $p \approx .95$) or truncated normal
369 distribution (Kolmogorov-Smirnov test: $p \approx .33$, t test: $p \approx .96$).
370
371 It has been previously observed that biosynthetically costly amino acids tend to promote
372 stronger cooperative interactions than biosynthetically cheap amino acids (27). Given this

373 observation we tested if biosynthetic costs would be useful for predicting whether a strain
374 will have a strong or a weak response in a given interaction. We used the biosynthetic cost
375 of the amino acid provided by the giver strain, needed by the receiver strain, or the
376 difference of these costs as decision thresholds to classify response classes and compared
377 the performance of the random forest classifier to these benchmarks. Since the
378 auxotrophic *E. coli* strains were all derived from the same ancestral strain, only the
379 biosynthetic capabilities for each of the 14 amino acids were relevant predictors for
380 machine learning. As a result, vectors with just 28 elements represented each pairwise *E.*
381 *coli* interaction. Examination of the corresponding ROC curves shows that biosynthetic
382 costs of amino acids are poor predictors for the qualitative outcome of these *E. coli*
383 auxotroph experiments whereas random forest fares much better as indicated by its ROC
384 curve and by area under the curve (Fig 4D inset). Random forest yielded a balanced
385 accuracy of ~79.2%. Moreover, the trajectory of the learning curve most closely resembles
386 the trajectory of the learning curve for *in silico* communities of 20 members (Fig 4E).
387 Variable importance rankings show that in general, the identity of the receiver's needed
388 amino acid is often more impactful on classification accuracy than the amino acid that the
389 giver needs, suggesting that specificity of interaction is dominated by auxotrophies,
390 whereas most mutants can in principle provide the missing amino acid (Fig 4F).
391
392 Since the pair of knocked-out amino acids determines the interaction for any two
393 auxotrophs, we evaluated how often the corresponding features were the most influential
394 for classification. For each sample, we ranked the feature contributions by the magnitude
395 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 were occupied by the giver and receiver's amino acids, in any order, 132 of 182 times. 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.

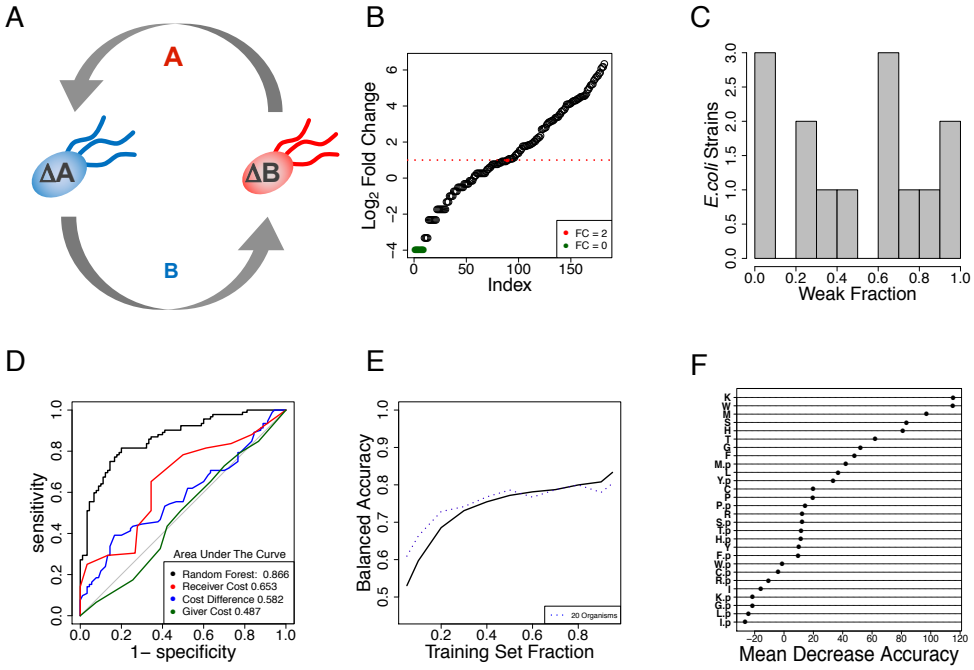


Fig 4. Data representation and results for the case study of a network of auxotrophic *E. coli* strains. **A.** In the original experiment, single gene knockout *E. coli* auxotrophs were co-cultured in a minimal medium. In order for ΔA to grow it must receive amino acid A from ΔB, which in turn must 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

414 to grow at all 9 times (green circles). In 92 instances the *E. coli* auxotroph population
415 more than doubled over the course of 84 hours. **C.** Histogram of weak responses as a
416 fraction of each strains' total number of interactions. **D.** ROC curve for random forest
417 classifier using 28 amino acids as predictors for 182 observations. Single value thresholds
418 based on the biosynthetic costs of knocked out amino acids resulted in poorer performance
419 than random forest. **E.** The trajectory of a learning curve built for the *E. coli* interactions
420 closely resembles that of the learning curve for *in silico* communities with 20 organisms.
421 **F.** The 28 amino acids ranked according to their affects on prediction accuracy when
422 randomly permuted. Amino acids corresponding to the receiver strain are enriched near
423 the top of the list. Amino acids are represented by their single letter codes. The suffix '.p'
424 indicates that the predictive feature belongs to the giver strain.
425

426 Scenarios where the presumed mechanisms are the strongest contributors yet result in
427 misclassification present opportunities to direct research toward interesting outliers in
428 order to understand why they diverge from our expectations. The response of the
429 methionine auxotroph (**Δ**Met) to co-culture with the cysteine auxotroph (**Δ**Cys) was one
430 such case, which is described in detail in Fig. S2.
431

432 Discussion

433 Exhaustive pairwise co-culture studies of microbial communities are becoming
434 increasingly common. While such pairwise interactions do not necessarily capture all
435 possible interdependencies in a community (29, 30), they have been shown to be a
436 dominant factor (31), making the reliable prediction and interpretation of predictive
437 models matters of great importance. In this study, we have described a conceptual
438 framework for the representation of microbes and their pairwise interactions in order to
439 address both of these challenges. Our results indicate that representing genome-derived
440 traits of microbes as binary vectors is sufficient for building reliable classifiers for
441 microbial interactions. We also demonstrated two methods for utilizing feature

442 contributions from random forest models to aid in the development of testable hypotheses
443 regarding interaction mechanisms.
444
445 Qualitatively predicting the outcome of unobserved interactions is most valuable if those
446 predictions allow precious resources and time to be conserved. To this end the
447 construction of learning curves is an important step in identifying how much data is
448 required in order to achieve desired prediction accuracy from machine learning. Our
449 results indicate that prediction of interactions in communities of all sizes may benefit from
450 machine learning. In situations where the prediction of interaction outcomes is not
451 necessary the properties of random forest makes it a good choice for deciding which traits
452 to study first. Because predictions will always benefit from more data we suggest that if
453 the experimental space is not prohibitively large that all pairwise experiments be
454 performed and feature contributions then be used to derive testable hypotheses.
455 Conversely, when the experimental space is large then one should perform a subset of
456 experiments and incorporate the results of any additional experiments into the training set
457 in order to build new classifiers recursively. The larger the community being studied is the
458 smaller the relative fraction of possible experiments that must be initially performed, these
459 results are promising both for future studies of natural communities, which can contain
460 upward of a thousand unique community members (32) and the assembly of synthetic
461 communities from large libraries of microbes.
462
463 Feature contributions provide a clear window into the decision-making process of a
464 random forest when the underlying mechanisms are straightforward, as was the case for

465 the *E. coli* auxotrophs. In cases such as this, interpretation of the model is clear and novel
466 detailed hypotheses can be readily developed. Even incorrect predictions are useful since
467 instances that go against the expectations of the classifier are likely to be worth closer
468 scrutiny. In situations where many factors contribute in a complex manner to an
469 interaction we demonstrated the utility of feature contributions for guiding exploratory
470 experiments by enriching underlying competitive mechanisms near the top of ranked lists.
471 In our particular application of feature contributions to an *in silico* community we used
472 their net effects to identify metabolites that are competed for in negative cases or are used
473 in cross feeding in positive interactions. The particular method of evaluating feature
474 contributions, summing the net effects or sorting them by magnitude is a decision that
475 must be made on a case-by-case basis.
476
477 Although we concatenated the binary trait vectors of two organisms in order to form a new
478 composite trait representation, alternative representations of microbes and their
479 interactions should also be explored. Predictions may also benefit from incorporating
480 more information into the base representation of a microbe such as gene copy number or
481 mean transcriptional levels. We also encourage future researchers to acquire information
482 on the same interactions in multiple environmental conditions. Varying the environmental
483 conditions will allow abiotic factors to be incorporated into the representation and should
484 yield additional insight into situations where certain combinations of traits are most
485 relevant.
486

487 Going forward we expect that the importance of machine learning in microbial ecology
488 will continue to grow. The need to identify microbial strains that interact in a desired way
489 in any given environment will be one of the most pressing issues for synthetic ecology in
490 the future. Likewise, it would not be surprising if many novel methods for interpreting
491 machine learning algorithms emerge in response to the challenges of understanding the
492 properties of microbial communities.
493
494 **Data Availability.** Pointers to Datasets obtained from previous work, and used in our
495 analysis are reported in the Materials and Methods Section.
496 The code necessary to reproduce all our figures and analyses is hosted at:
497 https://github.com/ddimucci/MicrobialCommunities
498

499 Materials and Methods

500 **Auxotrophic *E. coli* interaction network data**. We obtained the measured growth
501 response of individual *E.coli* strains and biosynthetic costs of amino acids from the
502 supplemental files provided by (27).
503
504 **Simulation of *in silico* pairwise interactions among gut microbes with dFBA in**
505 **COMETS.** Metabolic reconstructions of human gut associated microbes were obtained
506 from Bauer et al (14). At the time of this writing these models can be downloaded directly
507 from the following URL:
508 https://wwwen.uni.lu/content/download/86230/1056013/file/Bauer_et_al_301_microbe_m
509 odels.rar

510 Each metabolic reconstruction encompasses the stoichiometry of virtually all metabolic
511 reactions present in an organism, including uptake/secretion. Flux balance analysis (FBA)
512 is a constraint-based steady state approach that uses this stoichiometry to predict fluxes
513 and growth capacity under a given boundary condition of nutrient availability, and has
514 been described in detail before (15, 33–35). Dynamic flux balance analysis (DFBA) (12)
515 extends classical FBA to perform dynamic simulations in which intracellular metabolites
516 are still assumed to be at steady states, but total biomass and environmental metabolites
517 are treated as time-dependent variables in a discretized approximation. We performed
518 DFBA simulations using our platform for Computation of Microbial Ecosystems in Time
519 and Space (COMETS), which has been previously used to model microbial communities
520 (13) . COMETS allows users to implement DFBA in a two-dimensional simulated world
521 populated with multiple metabolic models. We selected 100 metabolic models (14) and
522 identified a common medium that would permit the growth of nearly all models in a
523 monoculture scenario. We then performed all pairwise co-culture simulations of the 100
524 models using the common media set in a well-mixed batch culture scenario (approximated
525 by using COMETS without spatial structure). For each scenario we saved the record of
526 biomass accumulation and fluxes in order to calculate relative yield and identify
527 mechanisms of interaction, respectively.
528
529 **Relative Yield.** Relative yield quantifies the change in net growth of an organism in a new
530 environment relative to a reference scenario. For our purpose we compared the growth of
531 models in co-culture to their accumulated growth in monoculture. Relative yield was
532 calculated according to the equation:

533

534 $$RY = (Biomass_{coculture} - Biomass_{monoculture})/Biomass_{monoculture}$$

535

536 An $RY < 0$ indicates that the responder is detrimentally affected by its partner.
537 Correspondingly, an RY of 0 indicates no effect and an $RY > 0$ indicates a positive affect
538 from co-culture with the partner.

539

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

544

545 $$\mathrm{JD}(model_i, model_j) = 1 - | model_i \cap model_j | / | model_i \cup model_j |$$

546

547 **Representation of interactions with trait-derived features.** For a given community C
548 the observed co-culture response of each species i in the presence of j is encoded into the
549 element X_{ij} of an interaction matrix $\boldsymbol{X}$. To define a set of trait vectors for each organism in
550 C, we start by obtaining a list of λ features that can be assigned systematically across all
551 organisms. These features could be the presence/absence of specific genes, functions, or
552 any other trait. In the two case studies reported here, the feature vectors correspond to the
553 presence/absence of specific metabolite transporters in stoichiometric models, and to
554 amino acid auxotrophies, respectively. Thus, each organism i is associated with an λ-long
555 vector V_i, whose element k is 0 or 1 depending on whether or not trait k is absent or

556 present in organism *i*. Traits (i.e. elements of the feature vector) that are identical across
557 all organisms are removed from the feature set. Each pair of organisms (*i*,*j*) is then
558 associated with a co-culture feature vector, defined as the concatenation of vectors V_i and
559 V_j, also indicated as $V_{i,j} = [V_i , V_j]$ (see also Fig. 1). Note that in this concatenated vector,
560 order matters. For the purpose of feeding the data into the random forest algorithm, the
561 complete information about how organism *i* responds in a co-culture with organism *j*, is
562 encoded in the composite feature vector $V_{i,j}$ and the observed phenotypic response for that
563 specific pair, i.e. X_{ij}.
564
565 **Machine learning**. R was used to implement K-nearest neighbor (Knn), Random Forest
566 (RF), and support vector machine (SVM). Knn was implemented with the rknn package,
567 with K set to 3. Library e1071 was used for SVM and we used a linear kernel and
568 cost=0.1. Random forest was implemented with the randomForest library. All three
569 algorithms had similar performance on the *E. coli* data set but RF was significantly better
570 on the *in silico* data set (data not shown).
571 Random forests are ensemble classifiers that aggregate the results of many individual
572 decision trees. Each tree in a random forest is assigned a synthetic data set that is of the
573 same size as the training set but generated through sampling with replacement. The result
574 is that the average tree is trained on approximately 2/3 of the observations; these
575 observations are referred to as in bag samples. The remaining 1/3 of observations not in
576 the synthetic data sets are referred to as out of bag samples. The new synthetic data set is
577 placed at the root node of a new tree; next a randomly selected subset of predictive
578 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

601 curve was determined and the median learning curve for balanced accuracy of each

602 community size was calculated

603

604 **Feature contributions for binary classifications.** Each tree in the random forest is

605 presented with a bootstrapped subset of the provided data as a training set that is used to

606 grow the tree. The training set is placed at the root node of the tree. The fraction of

607 instances in the training set of class C_1 is Y^n_{mean}. Y^n_{mean} is the probability that an instance

608 selected at random from the root node is assigned to class C_1.

609 There are two steps in the calculation of feature contributions for a new sample: First

610 calculate the sum of all local increments in each tree, second determine the average

611 contribution over the forest. A local increment for a feature *f* is calculated if the split was

612 performed on *f* and is defined as the change in probability of C_1 in a parent node (p) and a

613 child node (c):

614

615 $$\mathrm{LI}^{c}_{f} = Y^{c}_{mean} - Y^{p}_{mean}$$

616

617 Local increments quantify the change in the probability of a sample being classified as C_1

618 between the child node and the parent node when *f* is the splitting criterion. The feature

619 contribution, $\mathrm{FC}^{f}_{i,t}$ of feature *f* in tree *t* for observation *i* is the sum of all LI_{f} along the path

620 from the root node to the terminal node. FC^{f}_{i} is given by the equation:

621

622 $$\mathrm{FC}^{f}_{i} = 1/\mathrm{T} \sum(\mathrm{FC}^{f}_{i,t})$$

623

624 Feature contributions were calculated with the forestFloor package and are calculated
625 using only trees for which a sample was out of bag.
626
627 **Application to positive samples.** Positive samples constituted a small minority of the
628 total samples (420/9900) and result in a heavily imbalanced data set when we try to
629 classify these samples as separate from either neutral or negative responses. Imbalanced
630 datasets often result in very high classification accuracies because the classifier simply
631 predicts every instance to be of the majority class. This is an issue because if we are to
632 have confidence in feature contributions an adequately performing classifier is desired. In
633 order to develop a reliable classifier we under sampled the non-positive instances by
634 randomly selecting a subset of 420 of them to pair with the 420 positive observations and
635 then trained a random forest on the new data set. To estimate the efficacy of the random
636 forest we repeated this process 100 times and note that there was a median balanced
637 accuracy of ≈ 85%. We then used a single random forest model to calculate feature
638 contributions in the context of identifying positive responses.
639
640 **Acknowledgments**
641 We are grateful to Dr. Jenny Bhatnagar and Dr. Carolyn Zeiner whose work with fungal
642 interactions served as the initial inspiration for this study and to members of the Segrè lab
643 for helpful discussions and for feedback on the manuscript. DS and DD acknowledge
644 funding from the Defense Advanced Research Projects Agency (Purchase Request No.
645 HR0011515303, Contract No. HR0011-15-C-0091), the U.S. Department of Energy (DE-
646 SC0012627), the NIH (T32GM100842, 5R01DE024468, R01GM121950 and

647 Sub_P30DK036836_P&F), the National Science Foundation (1457695 and NSFOCE-BSF
648 grant 1635070), the Human Frontiers Science Program (RGP0020/2016), and the Boston
649 University Interdisciplinary Biomedical Research Office.

650

651 **References**

652 1. Røder HL, Sørensen SJ, Burmølle M. 2016. Studying Bacterial Multispecies
653 Biofilms: Where to Start? Trends Microbiol.
654 2. Aziz FAA, Suzuki K, Ohtaki A, Sagegami K, Hirai H, Seno J, Mizuno N, Inuzuka
655 Y, Saito Y, Tashiro Y, Hiraishi A, Futamata H. 2015. Interspecies interactions are
656 an integral determinant of microbial community dynamics. Front Microbiol 6.
657 3. Johns NI, Blazejewski T, Gomes ALC, Wang HH. 2016. Principles for designing
658 synthetic microbial communities. Curr Opin Microbiol.
659 4. Ling LL, Schneider T, Peoples AJ, Spoering AL, Engels I, Conlon BP, Mueller A,
660 Schäberle TF, Hughes DE, Epstein S, Jones M, Lazarides L, Steadman VA, Cohen
661 DR, Felix CR, Fetterman KA, Millett WP, Nitti AG, Zullo AM, Chen C, Lewis K.
662 2015. A new antibiotic kills pathogens without detectable resistance. Nature
663 517:455–459.
664 5. Goers L, Freemont P, Polizzi KM. 2014. Co-culture systems and technologies:
665 taking synthetic biology to the next level. J R Soc Interface 11:20140065–
666 20140065.
667 6. Lasken RS, McLean JS. 2014. Recent advances in genomic DNA sequencing of
668 microbial species from single cells. Nat Rev Genet 15:577–584.
669 7. Gawad C, Koh W, Quake SR. 2016. Single-cell genome sequencing: Current state
670 of the science. Nat Rev Genet.
671 8. Seemann T. 2014. Prokka: Rapid prokaryotic genome annotation. Bioinformatics
672 30:2068–2069.
673 9. Dam P, Olman V, Harris K, Su Z, Xu Y. 2007. Operon prediction using both
674 genome-specific and general genomic information. Nucleic Acids Res 35:288–
675 298.
676 10. Liaw a, Wiener M. 2002. Classification and Regression by randomForest. R news
677 2:18–22.
678 11. Palczewska A, Palczewski J, Robinson RM, Neagu D. 2013. Interpreting random
679 forest classification models using a feature contribution method (extended). 2013
680 IEEE 14th Int Conf Inf Reuse Integr 1–30.
681 12. Henson MA, Hanly TJ. 2014. Dynamic flux balance analysis for synthetic
682 microbial communities. IET Syst Biol 8:214–229.
683 13. Harcombe WR, Riehl WJ, Dukovski I, Granger BR, Betts A, Lang AH, Bonilla G,
684 Kar A, Leiby N, Mehta P, Marx CJ, Segrè D. 2014. Metabolic resource allocation
685 in individual microbes determines ecosystem interactions and spatial dynamics.
686 Cell Rep 7:1104–1115.
687 14. Bauer E, Laczny CC, Magnusdottir S, Wilmes P, Thiele I. 2015. Phenotypic

688 differentiation of gastrointestinal microbes is reflected in their encoded metabolic
689 repertoires. Microbiome 3:55.
690 15. Orth JD, Thiele I, Palsson BØ. 2010. What is flux balance analysis? Nat
691 Biotechnol 28:245–248.
692 16. Bach EM, Williams RJ, Hargreaves SK, Yang F, Hofmockel KS. 2018. Greatest
693 soil microbial diversity found in micro-habitats. Soil Biol Biochem 118:217–226.
694 17. Mainali KP, Bewick S, Thielen P, Mehoke T, Breitwieser FP, Paudel S, Adhikari
695 A, Wolfe J, Slud E V., Karig D, Fagan WF. 2017. Statistical analysis of co-
696 occurrence patterns in microbial presence-absence datasets. PLoS One 12.
697 18. Svetnik V, Liaw A, Tong C, Christopher Culberson J, Sheridan RP, Feuston BP.
698 2003. Random Forest: A Classification and Regression Tool for Compound
699 Classification and QSAR Modeling. J Chem Inf Comput Sci 43:1947–1958.
700 19. Breiman L. 2001. Random forests. Mach Learn 45:5–32.
701 20. Kuz'min VE, Polishchuk PG, Artemenko AG, Andronati SA. 2011. Interpretation
702 of QSAR models based on random forest methods. Mol Inform 30:593–603.
703 21. Welling SH, Refsgaard HHF, Brockhoff PB, Clemmensen LH. 2016. Forest Floor
704 Visualizations of Random Forests_Supplementrary. Supplementrary 1–13.
705 22. Di Luccia B, Crescenzo R, Mazzoli A, Cigliano L, Venditti P, Walser JC, Widmer
706 A, Baccigalupi L, Ricca E, Iossa S. 2015. Rescue of fructose-induced metabolic
707 syndrome by antibiotics or faecal transplantation in a rat model of obesity. PLoS
708 One 10.
709 23. Khitan Z, Kim DH. 2013. Fructose: A key factor in the development of metabolic
710 syndrome and hypertension. J Nutr Metab.
711 24. Bantle JP. 2009. Dietary Fructose and Metabolic Syndrome and Diabetes. J Nutr
712 139:1263S–1268S.
713 25. Lambertz J, Weiskirchen S, Landert S, Weiskirchen R. 2017. Fructose: A dietary
714 sugar in crosstalk with microbiota contributing to the development and progression
715 of non-alcoholic liver disease. Front Immunol.
716 26. Payne AN, Chassard C, Lacroix C. 2012. Gut microbial adaptation to dietary
717 consumption of fructose, artificial sweeteners and sugar alcohols: Implications for
718 host-microbe interactions contributing to obesity. Obes Rev 13:799–809.
719 27. Mee MT, Collins JJ, Church GM, Wang HH. 2014. Syntrophic exchange in
720 synthetic microbial communities. Proc Natl Acad Sci 111:E2149–E2156.
721 28. POWELL EO. 1956. Growth Rate and Generation Time of Bacteria, with Special
722 Reference to Continuous Culture. J Gen Microbiol 15:492–511.
723 29. Bairey E, Kelsic ED, Kishony R. 2016. High-order species interactions shape
724 ecosystem diversity. Nat Commun 7.
725 30. Levine JM, Bascompte J, Adler PB, Allesina S. 2017. Beyond pairwise
726 mechanisms of species coexistence in complex communities. Nature.
727 31. Venturelli OS, Carr AC, Fisher G, Hsu RH, Lau R, Bowen BP, Northern T, Arkin
728 AP. 2017. Deciphering microbial interactions in synthetic human gut microbiome
729 communities. bioArxiv.
730 32. Raynaud X, Nunan N. 2014. Spatial ecology of bacteria at the microscale in soil.
731 PLoS One 9.
732 33. Raman K, Chandra N. 2009. Flux balance analysis of biological systems:
733 Applications and challenges. Brief Bioinform.

734 34. Maarleveld TR, Khandelwal RA, Olivier BG, Teusink B, Bruggeman FJ. 2013.
735 Basic concepts and principles of stoichiometric modeling of metabolic networks.
736 Biotechnol J.
737 35. O'Brien EJ, Monk JM, Palsson BO. 2015. Using genome-scale models to predict
738 biological capabilities. Cell 161:971–987.
739 36. Park S, Imlay JA. 2003. High levels of intracellular cysteine promote oxidative
740 DNA damage by driving the Fenton reaction. J Bacteriol 185:1942–1950.
741 37. Takumi K, Nonaka G. 2016. Bacterial cysteine-inducible cysteine resistance
742 systems. J Bacteriol 198:1384–1392.
743