Prediction of representative phenotypes using Multi-Attribute Subset Selection

Konrad Herbst[1,2,*], Taiyao Wang[3,*], Elena J. Forchielli[2,4,*], Meghan Thommes[2,5,*], Ioannis Ch. Paschalidis[3,5,6,7,#], Daniel Segrè[1,2,4,5,6,#]

[1]Bioinformatics Program, Boston University, Boston, MA, USA
[2]Biological Design Center, Boston University, Boston, MA, USA
[3]Division of Systems Engineering, Boston University, Boston, MA, USA
[4]Department of Biology, Boston University, Boston, MA, USA
[5]Department of Biomedical Engineering, Boston University, Boston, MA, USA
[6]Faculty of Computing and Data Science, Boston University, Boston, MA, USA
[7]Department of Electrical and Computer Engineering, Boston University, Boston, MA, USA

* Equally contributing authors
# Corresponding authors

Abstract

The interpretation of complex biological datasets requires the identification of representative variables that describe the data without critical information loss. This is particularly important in the analysis of large phenotypic datasets ("phenomics"). We introduce Multi-Attribute Subset Selection (MASS), an algorithm which separates a matrix of phenotypes (e.g., yield across microbial species and environmental conditions) into predictor and response sets of conditions. Using mixed integer linear programming, MASS expresses the response conditions as a linear combination of the predictor conditions, while simultaneously searching for the optimally descriptive set of predictors. We applied the algorithm to three microbial datasets and identified environmental conditions that predict phenotypes under other conditions, providing biologically interpretable axes for strain discrimination. MASS could be used to reduce the number of experiments needed to identify species or to map their metabolic capabilities. The generality of the algorithm allows addressing subset selection problems in areas beyond biology.

Keywords

Phenotyping, Feature selection, Yeast, Marine bacteria, Microbial growth, Environmental conditions, Machine learning

Introduction

As the price of DNA sequencing keeps decreasing, reading genomes is no longer the limiting factor in understanding an organism: interpreting genomes is. In principle, genomes carry information on how an organism will behave in different environments, but in practice the interpretability of genomic information is severely limited by our lack of knowledge about the function of many individual genes (Roberts 2004; Furnham, de Beer, and Thornton 2012) and the interactions between genes and environments (Kishony and Leibler 2003; de Vos et al. 2013), as well as by a limited understanding of how genes work together to produce higher-level functions (Collado-Vides, Gaudet, and de Lorenzo 2022; Yu et al. 2016). The challenge of predicting organismal functions from genomes is particularly relevant in microbes, where systems biology approaches provide a potential avenue for building mechanistic models of metabolism at the genome-scale (Seaver et al. 2021; Passi et al. 2021). Despite constant improvements in these genome-to-phenotype inferences, there is rising interest in exploring a complementary avenue to characterizing and understanding organismal behavior starting from large-scale phenotypic data. Through this approach, also referred to as “phenomics”, insight and knowledge are sought from the measurement of phenotypes across many organisms (or genetic perturbations) and environmental conditions (Schilling, Edwards, and Palsson 1999; Sauer 2004; Jewett, Hofmann, and Nielsen 2006; Bochner 2009; Acin-Albiac et al. 2020). Phenomics approaches the challenge of understanding biological systems from a top-down data-driven perspective: phenotypes are measured, often in a high-throughput manner, and then analyzed in search for associations (between genotypes and phenotypes, or between different phenotypes) (Houle, Govindaraju, and Omholt 2010; Baran et al. 2013; Bowen et al. 2011; Baran, Bowen, and Northen 2011; Jim et al. 2004; Tamura and D’haeseleer 2008; Ohya et al. 2005; Forchielli, Sher, and Segrè 2022). In turn, these associations can be used to understand how the individual subsystem properties, give rise to cell-, organism- or community-level functions, including mutual interdependencies in microbial communities (Zelezniak et al. 2015; Zoccarato et al. 2022; DiMucci, Kon, and Segrè 2018). Furthermore, phenotypic data play an important role in gene annotation (Price et al. 2018), and mechanistic model testing and refinement (Bernstein et al. 2021).

The development and applicability of phenomics approaches is crucially dependent on the capacity to reduce measurement costs, and to appropriately analyze high-throughput phenotypic data of microorganisms grown in a variety of environments (Price et al. 2018; Barnett, Payne, and Yarrow 1990; Segrestin et al. 2021). A lot of work has been done to characterize the phenotypes of microbes for applications ranging from drug discovery to industrial fermentation (Demain and Sanchez 2009; Schmidt 2005). Numerous technologies have been developed to facilitate the generation of phenotypic data, in microbes and other biological systems (Yeung, Sahin, and Andrews 2022), including Biolog Phenotype MicroArrays (Hosmer et al. 2022; Bochner 2009), robotic screening tools (Barton et al. 2018), microfluidics devices (Cario et al. 2022; Behrendt et al. 2020; Kehe et al. 2021), imaging (Ohya et al. 2005; Kritikos et al. 2017; D’Orazio et al. 2022) and multiplexed bioreactors, such as the eVOLVER (Wong et al. 2018). However, compared to the rapid decrease in sequencing cost, measuring

phenotypes is still relatively cost and labor intensive; thus, prioritizing which conditions and organisms to measure is a crucial component of experimental design. Ideally one would want to narrow down a set of measurements to those organisms and conditions that are most informative for reconstructing larger phenotype matrices and provide the most predictive power regarding other measurable phenotypes. For biologists, focusing on those phenotypes of highest informational value would have two major advantages: 1) reduce the experimental burden of large screens, and 2) provide insight into mechanistic links between traits.

Numerous statistical approaches have been employed to help extract the most informative subsets of phenotypic datasets (Hastie, Tibshirani, and Friedman 2001; James et al. 2021; Asnicar et al. 2023). Classical methods, such as clustering and latent variable methods (including PCA), are often used as dimensionality reduction tools for data interpretation and visualization purposes (Mirza et al. 2019). However, those methods do not provide a clear identification of a subset of variables such that the whole dataset can be quantitatively expressed as a function of those selected variables. Conversely, regression analysis can be used to build predictive models based on a specific set of attributes (Blaise et al. 2021). However, regression typically relies on prior knowledge of which attributes should serve as predictors and which model is used to represent the relation between predictor and response attributes, limiting their utility for phenotype organization, where no a priori distinction between predictor and response variables exists.

We approach the identification of the most “informative” phenotypes by asking the following question: Which subset of phenotypes of a given subset size describes the remaining phenotypes best? We formulate a mathematical framework to answer this question, in which phenotypes are classified into sets of predictor and response attributes. At the same time generalized linear regressions of those sets are used to identify the most descriptive phenotype set. Exhaustive enumeration of all possible sets of predictor and response attributes gives rise to an extremely large number of possible solutions. Nevertheless, efficient training of the regression models can be achieved by using mixed integer linear programming (MILP). This algorithm, to which we refer as “multi-attribute subset selection” (MASS), constitutes a new approach to the phenotype selection problem as it finds solutions explicitly representing the most informative phenotype sets.

Over the last decade, dramatic improvements in MILP solvers in conjunction with the widespread availability of high-dimensional biological datasets have created attractive opportunities to expand the applications of MILP to questions of biological importance (Gusfield 2019). Great strides have been made in the analysis of DNA sequences (Lancia 2008), protein-protein interaction networks (Dittrich et al. 2008), and mass spectrometry data (DiMaggio et al. 2009) using linear programming methods. However, the preceding examples modeled a single response attribute vector as a linear combination of predictor attributes; methods capable of modeling multiple attributes as the combination of other phenotypes require a novel formulation and have to our knowledge not been attempted.

MASS is inspired by the application of MILP to solve the subset selection problem (Miller 1984; Bertsimas, King, and Mazumder 2016) and exploits several heuristics that can substantially speed up the computational time needed to find an optimal solution, similar to the process outlined in (Thommes et al. 2019). We applied MASS to three microbial phenotype datasets, and successfully identified sets of predictor attributes with the highest predictive power. MASS has the potential to enable experimentalists to minimize the number of experiments they need to perform while maintaining, with a certain degree of uncertainty, the most information regarding a microbe's phenotype. MASS and its possible extensions could find broad applicability in other data-rich scientific areas beyond biology.

Results

A method to separate growth phenotypes into predictors and responses

We designed MASS with the goal to explore large phenotypic datasets. Specifically, we want to choose the most informative subset of the environmental conditions and aim at using phenotypes under such conditions as predictors of phenotypes under all the remaining conditions. The challenge of this question is that it involves two steps that are usually performed in separate calculations. In a first step we would typically choose which phenotypes are predictors (independent variables), and which are the response (dependent variables). In a second step, we would perform a linear regression, i.e., find coefficients necessary to compute the response phenotypes as linear combinations of the predictor phenotypes. What makes our approach novel and mathematically challenging, is that we do not know or assume a priori which phenotypes will be predictors and which ones will be responses. The MASS algorithm makes it possible to pursue both steps concurrently. In other words, the algorithm explores the many possible choices of predictors and simultaneously identifies the ones such that regression done using those predictors gives the best estimate of the responses. This algorithm involves both integer variables, describing which phenotypes are chosen as predictors, and continuous variables which capture the regression coefficients, and is thus implemented using a Mixed Integer Linear Programming (MILP) approach.

The formulation of MASS as a combinatorial optimization problem is described in detail in the **Methods** and illustrated in **Figure 1**. Inputs to MASS include a phenotype matrix (**X**) of *n* organisms by *m* environmental conditions. For each possible number of predictors *p*, MASS selects p predictors from m environmental conditions to predict $m - p$ responses using linear regression. The outputs of MASS consist of a binary predictor vector (***z***) whose elements indicate whether an environment is a predictor or response. Note that as we increase the number of predictors p for a given dataset, a condition that is selected as a predictor for a given p, may be selected to be in the response set for a different p.

In our case the data consists of discrete growth phenotypes. While MASS provides optimal choices of predictors based on linear regression, we wanted to check whether those same predictors would also be useful as variables for a classification of the data based on an

alternative methodology (see **Methods**). For each dataset we trained random forest (RF) models (Breiman 2001; Ho 1995; Liaw, Wiener, and Others 2002) using the predictors determined by MASS. RF models can have multiple predictors and one response variable. Therefore, the number of RF models trained depends on the number of predictors to which MASS has been constrained. Consider for example a dataset containing 11 conditions. If the number of predictors *p* is set to 1, then there will be 11 - 1 = 10 responses, and 10 random forest models need to be trained, one for each response. Likewise, when *p* is set to 10 predictors, there will be 11 - 10 = 1 response, and only one random forest model has to be trained for the one response. Prior to training the models, we split the data into training and test sets. We then calculated a score for the test sets to estimate the performance of the random forest. We chose the Matthews Correlation Coefficient (MCC) as a performance score able to deal with potential imbalance in data categories (Chicco and Jurman 2020), (see **Methods**). While random forests are used as a practical approach in this study to validate our MASS analysis, we would like to highlight that MASS essentially is a feature selection technique and we envision that the resulting predictor sets can be agnostically used with any other supervised learning method (**Figure 1**).

We apply MASS to three distinct datasets. The first dataset (DATASET 1) is a matrix of growth phenotypes (optical density (OD)) for 65 marine heterotrophic bacteria on 11 different types of carbon sources (see (Forchielli, Sher, and Segrè 2022) and **Figure 2**). The second dataset (DATASET 2), considerably larger (637 bacteria for 46 carbon sources), is a subset of the comprehensive BacDive database of microbial phenotypes (Reimer et al. 2022) including fermentation phenotypes on different carbon sources. The third dataset (DATASET 3) is a compendium of discrete growth phenotypes collected in a practical guide for the identification of different yeast species based on their growth capacity under different conditions (Barnett, Payne, and Yarrow 1990), and includes more complex phenotypes (three growth phenotypes, for 462 yeast strains over 38 conditions).

Analysis of DATASET 1: Growth of marine heterotrophic bacterial on well-defined media

We first applied MASS to a dataset of limited size (DATASET 1), which resulted from a characterization of marine heterotrophic bacteria grown on the different carbon sources contained in a standard marine broth medium (**Figure 2A**) (Forchielli, Sher, and Segrè 2022). As illustrated in **Figure 2B**, MASS provides the predicted optimal choice of predictor media conditions as a function of the number of predictors. In this case, if only one predictor is allowed (p = 1), the MASS algorithm selects the “oligosaccharide” condition. For p = 2, the chosen representative conditions are “neutral sugars” and “peptides”, supplemented by “amino acids” when p = 3.

We noted that certain conditions were selected as predictors more often than others (**Figure 2A**) and hypothesized that the frequency a condition is selected as a predictor by MASS corresponds with its information content (**Figure 2C**). To test this hypothesis, we calculated the Shannon Entropy as a measure for the information content of individual conditions and noted

that it correlates with the number of times a condition is selected as a predictor by MASS (**Figure 2B, C**, **Supplementary Figure S1A**). However, conditions selected most often are not necessarily associated with the highest entropy. This is further confirmed by the random forest prediction based on conditions selected by MASS (**Figure 2D**) which shows that MASS performs better than a predictor condition set selected based on maximal entropy. As a control, we also trained random forest models with a random choice of predictor conditions. These models performed considerably worse than the MASS or maximum entropy predictor set selections highlighting the importance of careful predictor set selections (**Figure 2D, Supplementary Figure S3**). As shown later, these results still hold in principle for larger datasets.

In order to obtain further insight into the implications of the MASS results for the biological systems under study, we also inspected, for each response individually, how well that response phenotype could be predicted by random forest models trained with the predictor sets selected by the MASS algorithm (**Supplementary Figure S2**). Interestingly, “amino sugars” are well predicted even when considering only the top three predictors selected by MASS (p = 3) The respective predictor conditions contain “neutral sugars”, “peptides”, and “amino acids”. This suggests potential overlaps in metabolic pathways for the utilization of these carbon sources. In contrast, growth under the medium condition containing “organic acid” was difficult to predict, suggesting that this trait is not linked to any particular set of carbon sources tested in the experiment. Altogether, these findings are in line with the previously reported biological interpretation of this dataset, which demonstrated how the different bacteria can be broadly subdivided into major subgroups associated with sugars vs. amino acid/peptide vs. organic acid substrates respectively (Forchielli, Sher, and Segrè 2022; Gralka, Pollak, and Cordero 2023).

Analysis of DATASET 2: Bacterial fermentation on different carbon sources

To evaluate the scalability and performance of MASS on a larger dataset, we identified phenotypic matrices of larger size and complexity, and for which no prior analysis of this kind had been implemented. BacDive is a large database collecting a wide range of phenotypes and metadata for several thousand organisms deposited at the German Collection of Microorganisms and Cell Cultures (DSMZ) (Reimer et al. 2022). We focused here on a complete subset of this data (see also **Methods**) which amounts to a matrix of fermentation phenotypes for 637 microbial species on 46 different carbon sources as environmental conditions (**Figure 3A**).

The predictors selected by MASS with increasing numbers of predictors p provide a ranking by importance which might be informative about the underlying biological system. We tested this conjecture by dissecting the first couple of conditions which have been selected by MASS as most descriptive (**Figure 3E**). We mapped the fermented carbon sources to their respective monomers where applicable (**Figure 3F**) and highlighted their entries into glycolysis as part of the central carbon catabolism (**Figure 3G**). When one predictor was allowed (p = 1), MASS selected cellobiose, a disaccharide consisting of two glucose monomers. Glucose enters

glycolysis at the very beginning of the pathway. When two predictors were allowed (p = 2), in addition to cellobiose, MASS selected raffinose, a trisaccharide which consists of the monomers galactose, fructose and glucose. This opened a new entry point through fructose into glycolysis, while galactose enters glycolysis through glucose-6-phosphate as an alternative to glucose. Raffinose was substituted by lactose and arabinose when three predictors were allowed (p = 3), opening an entry point via fructose-1,6-diphosphate as an alternative to fructose. Allowing for four predictors (p = 4) apparently resulted in major changes in the selection of the most informative substrates by MASS but mapping the selected substrates to their respective monomers revealed that just one additional entry point into glycolysis was opened through arbutin. This pattern of incremental addition of entry points into glycolysis continued as more and more predictors were allowed to be selected by MASS (**Figure 3E-G**). In summary, we see that with an increasing number of predictors MASS selected carbon sources covering an increasing number and tiers of entry points into glycolysis (**Figure 3G**). This indicates that the predictor selection based on the fermentation phenotype data recapitulates potential topological constraints on central carbon catabolism. These results highlight how the application of MASS on suitable datasets has the potential to reveal more fundamental principles underlying the biological system under study.

Conditions which were selected by MASS more frequently were also more likely to have a higher Shannon Entropy, similarly as observed before for DATASET 1 (**Figure 3B,C**, **Supplementary Figure S1B**). This suggests that using high-entropy conditions would lead to models which are equally predictive as when trained using predictor conditions resulting from MASS. Indeed, as shown in the analysis of random forest performance (**Figure 3D**), entropy-based choice of attributes worked much better than when the predictor conditions were selected randomly. However, the choice suggested by MASS still outperforms an entropy-based selection. This is likely because MASS selects multiple attributes for p>1 that are jointly the most informative ones, while entropy is associated with individual conditions. Thus, MASS may recapitulate the complex structure of the dataset, and account for correlations across phenotypic vectors, going beyond the capabilities of an entropy criterion.

Of note, some conditions have very unbalanced fermentation phenotype frequencies reflecting either rare or ubiquitous phenotypes. Those phenotype vectors were generally selected last by MASS highlighting how unbalanced phenotype vectors are poor descriptors for other phenotypes. Generally, we observed that the mean performance of the random forest classifiers increased with the number of predictors, irrespective of the performance measure chosen to evaluate the classifiers (**Supplementary Figure S4**). This trend ends after 27 or 33 predictors for predictor selections based on Shannon entropy or MASS respectively coinciding with the transition from conditions of high to low informational value. We interpret this as an indication that the available information within a dataset is used by the MASS predictor selections most efficiently (**Figure 3D**).

Analysis of DATASET 3: Yeast carbon assimilation growth profiles

Following the application of MASS to two binary datasets, we were interested in evaluating its performance on a categorical dataset that included more than two possible response values. We used a subset of a vast phenotypic resource describing the aerobic growth of 462 yeast species on 44 different compounds as the sole carbon sources (Barnett, Payne, and Yarrow 1990). This reference manual was written mainly as a practical guide to identify yeasts using phenotype profiles. The idea is that given an unknown yeast species, one would grow it on a predefined series of conditions, gradually narrowing down the options for its taxonomy, and ultimately resulting in a unique identification. Our algorithm has the potential to produce a sequence of maximally informative conditions that would allow one to solve this problem in a general, unsupervised fashion.

Each carbon source and yeast species exhibit different profiles (**Figure 4A**). Some carbon sources (e.g., glucose) could be utilized by almost all species, while other carbon sources (e.g., methanol and inulin) could only be utilized by very few species (**Figure 4A**). Likewise, some species exhibited “generalist” tendencies, since they were able to grow on most carbon sources, while others were “specialists”, as they could only grow on a small portion of carbon sources (**Figure 4A**).

We used MASS to determine which carbon sources best predict yeast growth on other carbon sources. In case only one predictor was allowed ($p = 1$), D-xylose was selected as the most descriptive predictor. This compound is a major component of lignocellulosic material. As p increases (for all $p > 1$), D-gluconate, an acid frequently found in fruit, honey, and wine (Ramachandran et al. 2006), becomes a prominent and ubiquitous predictor (**Figure 4B**). The next two important predictors appearing after D-xylose and gluconate are maltose (a disaccharide degradable into glucose) and glycerol, which feeds into the middle of glycolysis. As observed for DATASET 1 and 2, the larger the entropy of a carbon source, the more often it was used as a predictor (**Figure 4B, C**; **Supplementary Figure S1C**). However, once again, upon implementing a random forest model based on the MASS choice of predictors at different values of p, the predictors chosen by MASS perform better than both random choices and predictors ranked by entropy (**Figure 4D**; **Supplementary Figure S5**).

We could observe that the performance generally increased with an increasing number of predictors p for up to 16 predictors and training with predictors selected through MASS resulted in the best-performing classifiers (**Figure 4D**). This indicates that MASS succeeds to identify the most relevant environmental conditions in phenotypic datasets and can handle categorical data beyond two classes.

Discussion

The problem of attribute or feature selection is not new to biology, and considerable work has been done to improve model performance on high-dimensional biological datasets by

separating the most relevant variables from those that may be uninformative, irrelevant, or redundant (Keleş, van der Laan, and Eisen 2002; Saeys, Inza, and Larrañaga 2007; Asnicar et al. 2023). One traditional response to this challenge can be to use dimensionality reduction using methods such as PCA, but these approaches tend to obscure interpretability by creating latent variables, complicating downstream analysis (Jolliffe and Cadima 2016). Feature selection procedures are usually employed in a supervised learning framework for which class labels must be known *a priori* (Hastie, Tibshirani, and Friedman 2001; Mirza et al. 2019; James et al. 2021; Asnicar et al. 2023). To the best of our knowledge, MASS is the first method that optimally selects a subset of predictor and a complementary set of response attributes from a dataset using a MILP framework. A key aspect of the MILP approach is that it simultaneously optimizes the regression coefficients which are continuous variables, and the integer variables that identify which features should be used as predictors. In contrast, for example, each principal component in PCA is a linear combination of all available attributes and it would be less straightforward to choose the most descriptive attributes out of those. As opposed to other methods that broadly fall into the category of learning problems (such as Multi-Modal Best Subset (MOSS) modeling, which is focused on product quality optimization based on sensors in additive manufacturing (Wang, Du, and Jin 2021)), our approach is specifically helpful for analyzing biological phenotype matrices as it is designed to find, given a certain number of attributes (p), a selection of predictor conditions (attributes) which optimally predicts all the remaining attributes.

It is important to note that while MASS selects the optimal subset of features to explain the remaining features, it does not optimize the predictability of any individual feature. In the future, it may be interesting to explore our formulation when using tree-based models for the regression rather than linear regression (Bertsimas and Dunn 2017). Furthermore, in this work MASS was applied solely to discretized fitness data; future developments could extend the current method to more general cases, including continuous variables. This is a particularly interesting prospect, as it would add a layer to the biological interpretation of microbial phenotypes by considering for example the magnitude of growth in addition to the ability to grow under various conditions.

Our results suggest that MASS has powerful applications for large-scale phenotypic screens by greatly reducing experimental burden. By identifying combinations of phenotypes that best predict other phenotypes within a dataset, MASS can help prioritize large-scale phenotyping efforts. After performing an initial phenotypic screen for a limited number of samples, MASS would allow one to select the subset of most descriptive conditions to be used for phenotyping a larger number of samples, based on available resources and desired level of accuracy. Here, we focused on microbial phenotypes in different environmental conditions, but it is not difficult to imagine other contexts in which MASS could be applied: examples include predicting antibiotic sensitivity of different bacterial strains, optimizing metabolite fluxes for biotechnological applications, and mapping inter-microbial interactions in a natural or synthetic microbial community. A particularly exciting potential application of MASS is in the selection of drug candidates for *in vivo* studies and *ex vivo* drug screening applications (Zheng, Thorne, and McKew 2013; Moffat, Rudolph, and Bailey 2014). Reducing the enormous experimental burden

required to execute these assays would represent an important outcome for many researchers, likely to advance the pace of biological discovery, and to facilitate the commercial feasibility of large-scaled screening assays.

From a biological perspective, the results of MASS could also help to interpret the resulting data. The fact that a given phenotype can be expressed as a linear combination of other phenotypes may suggest an underlying mechanistic relationship between these phenotypes. In the analysis of the marine dataset (DATASET 1), some phenotypes can be well predicted from other phenotypes, suggesting that these metabolic traits are linked for mechanistic or ecological reasons. For example, growth on amino sugars is predicted very well by using growth on neutral sugars, amino acids, and peptides as predictors. This suggests that overlapping metabolic pathways are utilized in the assimilation of these carbon sources. Conversely, growth on the organic acids was difficult to predict through other phenotypes, suggesting that this trait is not linked to any particular set of carbon sources tested in the experiment, (Forchielli, Sher, and Segrè 2022; Gralka, Pollak, and Cordero 2023). In another example (DATASET 2), the most descriptive conditions of carbon sources selected by MASS for fermentation seem to reveal a hierarchy of entry points into glycolysis hinting at fundamental constraints acting on the structure of central carbon catabolism.

An interesting finding of our analysis is the fact that, while predictability increases with the number of predictors (p) until a certain level (p>15), the performance score drops abruptly as p approaches the total number of variables (see **Figure 3D** and **4D**). This is because MASS is designed to select all the most descriptive attributes first, leaving for last the least descriptive ones. In other words, for large p, the n-p response attributes which MASS can still select cannot be easily predicted from the predictor attributes. While exploring very large p values is of little practical utility anyway, we thought it may be interesting to understand whether the specific threshold may relate to some biological properties of the datasets. One can note that both in DATASETS 2 and 3, numerous conditions resulted in positive phenotypes (i.e., observed growth or fermentation) across many of the species. We hypothesize that the substrates used in those conditions might belong to core metabolic pathways evolutionary conserved across species. In contrast, other conditions rarely resulted in positive phenotypes and therefore might relate to auxiliary metabolic pathways. Interestingly, the number of conditions resulting in ubiquitous phenotypes seems to coincide with the number of allowed attributes (p) beyond which the performance starts to drop for random forest models trained with attributes selected by MASS (**Figure 3D and 4D**). This suggests that the specific structure of our datasets, and specifically the nature of core vs. auxiliary metabolic pathways may contribute to the observed drop in performance score.

In summary, we present MASS as an algorithm to partition phenotype vectors into a set of most descriptive predictors and respective responses. Phenotypic information processed through MASS will allow researchers to prioritize efforts on high-value phenotypes or shift resource allocation to ensure the most important data is collected. Notably, our generic problem formulation and the MILP framework underlying MASS are not restricted to the particular type of

data explored in this study. It would be therefore valuable to explore the usability of this approach to other fields of research and practical applications.

Methods

Data pre-processing

DATASET 1

This dataset was previously acquired in our lab and published elsewhere (Forchielli, Sher, and Segrè 2022). The data was downloaded from the data analysis repository of the original study (https://github.com/segrelab/marine_heterotrophs/blob/main/data/growth_profiles.xlsx). The resulting growth phenotype matrix was discretized by assigning the integer 1 to non-zero (significant growth) values or the integer 0 otherwise.

DATASET 2

We accessed the BacDive (Reimer et al. 2022) API by using the R package *BacDive* (Vers. 0.8.0) to download for all available species data values of the subcategories "Enzymes", “Metabolite Utilization”, and “Metabolite Production” in the category “Physiology and metabolism". This resulted in a matrix of 5,289 attributes for 22,535 strains (as of 30th of November 2022). We noticed that the matrix is very sparse (99.3 % missing values) and decided to focus on a subset of complete measurements concerned with fermentation phenotypes on different carbon sources (identified by “builds acid from …”). The final matrix used for our MASS analysis encompasses fermentation of 637 species on 46 different carbon sources. Scripts used to mine the database and data itself are available on the project’s GitHub repository (https://github.com/segrelab/MASS).

DATASET 3

Values from the table in Chapter 6 of (Barnett, Payne, and Yarrow 1990) were digitized into a table with 590 yeast strains (samples, n) and 92 phenotypes (Supplementary Table S2). We focused in this study on the 44 phenotypes representing growth on specific carbon sources (attributes, m). Symbols for positive (“+”), negative (“–”), delayed (“D”), weak (“W”), variable (“V”), and unknown (“?”) growth were cast into categorical numbers. Negative growth was cast to 0; variable, weak, or delayed growth was cast to 1; positive growth was cast to 2; and unknown values were cast as null (“NaN”). Spearman correlations between each pair of features were calculated, and one of two highly correlated features (absolute Spearman correlation greater than 0.74) were dropped (Supplementary Table S3)**,** reducing the number of features to 38. Lastly, samples with at least one missing (null) value reduced the number of samples to 462.

For MASS analysis, growth on D-Glucose was excluded as almost all species exhibited a positive phenotype. Furthermore, the sample ‘*Saccharomycodes sinensis*’ was dropped as this species did not exhibit growth on any of the carbon sources. Hence, the phenotype matrix used for MASS consisted of 461 samples and 37 attributes. Categorical attributes were encoded as a

one-hot numerical array, where each categorical level (negative; variable, weak, or delayed; and positive) was separately encoded and indicated by a 1 when that type of growth was exhibited. However, we eliminated the negative (0) categorical level, thereby yielding 2 dummy variables out of the 3 categorical levels and doubling the number of features to 74. Zero values were cast to -1 so that the categorical one-hot arrays had values of -1 and 1, which is appropriate for hinge loss.

Linear programming approach to linear regression

Linear regression is a statistical method to model the linear relationship between the i'th response variable of n samples, y_i (a scalar), and the predictor variables of m attributes, $\boldsymbol{x_i}$, (an $m \times 1$ vector), by estimating the parameters, $\boldsymbol{\beta}$, (an $m \times 1$ vector), that provide the best explanation for the data:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i, \tag{1}$$

where prime denotes transpose. One approach to model fitting involves minimizing the difference between the actual and predicted responses (the error). If the ℓ_1-norm, $\sum_{i=1}^{n}|\hat{y}_i - y_i|$ is used as an error metric, then linear regression can be formulated as an optimization problem:

$$\begin{aligned} &minimize_{\beta} \sum_{i=1}^{n} |y_i - \boldsymbol{x}_i'\boldsymbol{\beta}| \\ &such\ that - M \leq \beta_j \leq M, \forall j, \end{aligned} \tag{2}$$

where M is a scalar value that bounds the estimated coefficients.

A method to determine predictors and responses

We use the convention that all vectors are column vectors. We first define the $n \times m$ matrix, $\boldsymbol{X} = (X_{i,j}) = (\boldsymbol{x}_1', \ldots, \boldsymbol{x}_n')'$, of observations with n samples and m attributes. We want to select a subset p from the attributes and use this subset as predictors for the remaining $m - p$ attributes (responses):

$$\boldsymbol{X} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{B_0} + \boldsymbol{E}, \qquad \Leftrightarrow \ \boldsymbol{x}_i' = \boldsymbol{x}_i'\boldsymbol{B} + \boldsymbol{\beta}_0' + \boldsymbol{\varepsilon}_i' \tag{3}$$

where $\boldsymbol{B} = (\boldsymbol{B}_{i,j}) = (\boldsymbol{\beta_1}, \ldots, \boldsymbol{\beta_m})$ is the $m \times m$ coefficient matrix, $\boldsymbol{B_0} = (\boldsymbol{\beta}_0', \ldots, \boldsymbol{\beta}_0')'$ is the $n \times m$ constant matrix, $\boldsymbol{\beta_0} = \left(\beta_{0,1}, \ldots, \beta_{0,m}\right)'$ is the constant vector and $\boldsymbol{E} = (\boldsymbol{\varepsilon}_1', \ldots, \boldsymbol{\varepsilon}_n')'$ is the noise matrix. Note that $(\boldsymbol{\beta_j})_k = B_{k,j}$ represents how attribute j is used to predict attribute k, and $x_{i,j}$ denotes the value of attribute j of sample i. Using the ℓ_1-penalty as a loss function and including a sparsity constraint, **Equation (3)** can be formulated as a mixed integer linear program:

$$
\begin{aligned}
&minimize_{\boldsymbol{B},\boldsymbol{\beta_0},\boldsymbol{w},\boldsymbol{z},\boldsymbol{t}} \ \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} w_{i,j} + \lambda \sum_{j=1}^{m} \|\boldsymbol{\beta_j}\|^1 \\
&such\ that \qquad t_{i,j} - \boldsymbol{x_i'}\boldsymbol{\beta_j} - \beta_{0,j} \le w_{i,j}, \forall i,j, \\
&\qquad -t_{i,j} + \boldsymbol{x_i'}\boldsymbol{\beta_j} + \beta_{0,j} \le w_{i,j}, \forall i,j, \\
&\qquad \sum_{j=1}^{m} z_j \le p, \\
&\qquad -Mz_k \le \left(\boldsymbol{\beta_j}\right)_k = B_{k,j} \le Mz_k, \forall j,k, \\
&\qquad -Mz_j \le t_{i,j} - X_{i,j} \le Mz_j, \forall i,j, \\
&\qquad w_{i,j} \ge 0, \forall i, \\
&\qquad z_j \in \{0,1\}, \forall j, \\
&\qquad B_{i,j} \in R, \beta_{0,j} \in R, \forall i,j,
\end{aligned}
\tag{4}
$$

where $\boldsymbol{w}$ is a dummy variable for the loss, reformulating the absolute loss using linear constraints; λ controls the sparsity or robustness constraint; $\boldsymbol{t}$ is a dummy variable for whether an attribute affects the loss; z is the indicator variable for whether an attribute is a predictor $(z_j = 1)$ or a response $(z_j = 0)$; $\boldsymbol{x_i}$ is the vector of the attributes of sample i; and $\boldsymbol{\beta_j}$ is the vector of coefficients to predict attribute j. If an attribute is a predictor, then coefficients are bounded between $\pm M$ and $t_{i,j}$ can be anything and will be set to $\boldsymbol{x_i'}\boldsymbol{\beta_j}$, thus not affecting the loss. However, if an attribute is a response, then coefficients are set to 0 and $t_{i,j}$ is set to $X_{i,j}$, therefore affecting the loss.

Since our datasets involved categorical data, **Problem** (**4)** was further reformulated to use hinge loss for $X_{i,j} \in \{-1,1\}$ (Dedieu, Hazimeh, and Mazumder 2021):

$$
\begin{aligned}
&minimize_{\boldsymbol{B},\boldsymbol{\beta_0},\boldsymbol{w},\boldsymbol{z}} \ \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} w_{i,j} + \lambda \sum_{j=1}^{m} \|\boldsymbol{\beta_j}\|^1 \\
&such\ that \qquad \sum_{j=1}^{m} z_j \le p, \\
&\qquad -Mz_k \le \left(\boldsymbol{\beta_j}\right)_k = B_{k,j} \le Mz_k, \forall j,k, \\
&\qquad X_{i,j}(\boldsymbol{x_i'}\boldsymbol{\beta_j} + \beta_{0,j}) \ge 1 - w_{i,j} - Mz_j, \forall i,j, \\
&\qquad w_{i,j} \ge 0, \forall i, \\
&\qquad z_j \in \{0,1\}, \forall j, \\
&\qquad z_j = z_k \ if\ (j,k)\ are\ in\ one\ group, \\
&\qquad B_{i,j} \in R, \beta_{0,j} \in R, \qquad \forall i,j.
\end{aligned}
\tag{5}
$$

MASS is implemented in *Python* using the MILP solver *GUROBI*.

Heuristics solutions to speed up solution time

The MILP **Problem (5)** is NP-complete. We observed that solving **Problem (5)** with a larger p (relatively close to the number of total attributes m) can be solved relatively quickly (on the order

of minutes). However, solving this problem for fewer predictors becomes very slow or impossible. Consequently, we used a similarity-based heuristic method first to obtain a near-optimal feasible solution and offer it to the solver. This reduced the time needed by the solver to reach an optimal solution.

We observed a similarity between the predictors selected at constraint bounds p and $p+1$, where $p+1$ is a larger constraint bound. Consequently, after solving **Problem (5)** at the constraint bound $p+1$, the selector indicator variable is $\boldsymbol{z_{p+1}}$. When solving at the constraint bound p, we added an additional constraint to ensure that the predictors at constraint bound p are selected from the predictors at constraint bound $p+1$:

$$\boldsymbol{z'_{p+1}z} \geq p, \qquad (\mathbf{6})$$

where $\boldsymbol{z}$ is the selector indicator at the constraint bound p. After solving **Problem (5)** with the additional constraint, we obtain a sub-optimal selection indicator vector $\boldsymbol{z_{greed1}}$. This first heuristic solution performs monotone selection since we remove a single attribute from $\boldsymbol{z_{p+1}}$.

In an alternative heuristic, we replaced **Constraint** (**6)** with the following relaxed version:

$$\boldsymbol{z'_{p+1}z} \geq p-1, \qquad (\mathbf{7})$$

using $\boldsymbol{z_{greed1}}$ as the starting point of the resulting MILP.

After solving **Problem (5)** with **Constraint (7)**, we obtain a sub-optimal selection indicator vector $\boldsymbol{z_{greed2}}$. In all experiments discussed in this paper we used $\boldsymbol{z_{greed2}}$.

Assessing the attribute selection by MASS using random forest models

We used random forest (RF) classifiers as implemented in the Python package *scikit-learn* to assess the performance of attribute sets selected through MASS. Prior to training random forest models, we randomly selected 50% of the samples in our dataset to form the training and validation set and retained the remaining 50% of the samples as the test set. Three-fold (for DATASET 3) and five-fold (for DATASET 1 and 2) cross-validation was used to tune parameters (*e.g.,* the number of variables randomly sampled as candidates at each split, the maximum depth of the tree, *etc.*). The number of RF trained is determined by the number of allowed predictor variables because this also determines the number of available response variables. For example, given that a dataset consists of 11 variables and 1 predictor is allowed (p = 1), MASS will identify (11-1) = 10 response variables. For each of these response variables one RF model is trained using the 1 predictor, which is identified by MASS. Similarly, if 10 predictors are allowed (p = 10), MASS will identify (11-10) = 1 response variable for which one RF model will be trained using all the 10 predictor variables selected by MASS. To assess the performance of the predictor selections by MASS we build additional RF models with predictor sets either selected randomly to generate baseline performance values or using a predictor set selected by

maximal entropy. The choice of the latter was motivated by the observation that MASS predictor selection preference correlated with Shannon entropy value for the respective attributes and served as a comparison of MASS to a more naïve predictor set selection approach.

We compared multiple performance metrics:

1) Accuracy,
2) Balanced Accuracy,
3) Cohen's kappa coefficient,
4) Weighted macro-averaged F1 score,
5) Jaccard distance, and
6) Matthews correlation coefficient (MCC).

These metrics were applied as implemented in the Python package *scikit-learn* and were calculated after combining results from all RF models for one specific p and selection method. For our main results we decided to use Matthews Correlation Coefficient because this metric has been shown to give robust estimates even for unbalanced datasets (Chicco and Jurman 2020). However, for comparison we also show the remaining set of performance metrics in the supplement (**Supplementary Figure S2, S3, and S4**).

Acknowledgments

IChP acknowledges funding by the NSF under grants CCF-2200052, DMS-1664644, and IIS-1914792, by the ONR under grant N00014-19-1-2571, by the DOE under grant DE-AC02-05CH11231, by the NIH under grant UL54 TR004130, and by Boston University. DS acknowledges funding by the BU Kilachand Multicellular Design Program, the U.S. Department of Energy, Office of Science, Office of Biological & Environmental Research through the Microbial Community Analysis and Functional Evaluation in Soils Science Focus Area Program (m-CAFEs) under contract number DE-AC02-05CH11231 to Lawrence Berkeley National Laboratory, the NIH National Institute on Aging award number UH2AG064704, the NSF Center for Chemical Currencies of a Microbial Planet (C-CoMP), NSF-BSF grant 2246707, and the Human Frontiers Science Program (grant number RGP0060/2021).

Data and Code Availability

All data generated and analyzed in this study and the corresponding codes are open source and available in the GitHub repository (https://github.com/segrelab/MASS).

List of Abbreviations Used

LP: Linear Programming; MASS: multi-attribute subset selection; MILP: Mixed Integer Linear Programming; RF: random forest

References

Acin-Albiac, Marta, Pasquale Filannino, Marco Gobbetti, and Raffaella Di Cagno. 2020. "Microbial High Throughput Phenomics: The Potential of an Irreplaceable Omics." *Computational and Structural Biotechnology Journal* 18 (August): 2290–99.

Asnicar, Francesco, Andrew Maltez Thomas, Andrea Passerini, Levi Waldron, and Nicola Segata. 2023. "Machine Learning for Microbiologists." *Nature Reviews. Microbiology*, November. https://doi.org/10.1038/s41579-023-00984-1.

Baran, Richard, Benjamin P. Bowen, and Trent R. Northen. 2011. "Untargeted Metabolic Footprinting Reveals a Surprising Breadth of Metabolite Uptake and Release by Synechococcus Sp. PCC 7002." *Molecular BioSystems* 7 (12): 3200–3206.

Baran, Richard, Benjamin P. Bowen, Morgan N. Price, Adam P. Arkin, Adam M. Deutschbauer, and Trent R. Northen. 2013. "Metabolic Footprinting of Mutant Libraries to Map Metabolite Utilization to Genotype." *ACS Chemical Biology*. https://doi.org/10.1021/cb300477w.

Barnett, J. A., R. W. Payne, and D. Yarrow. 1990. "Yeasts: Characteristics and Identification," 1012 pp.

Barton, David B. H., Danae Georghiou, Neelam Dave, Majed Alghamdi, Thomas A. Walsh, Edward J. Louis, and Steven S. Foster. 2018. "PHENOS: A High-Throughput and Flexible Tool for Microorganism Growth Phenotyping on Solid Media." *BMC Microbiology* 18 (1): 9.

Behrendt, Lars, M. Mehdi Salek, Erik L. Trampe, Vicente I. Fernandez, Kang Soo Lee, Michael Kühl, and Roman Stocker. 2020. "PhenoChip: A Single-Cell Phenomic Platform for High-Throughput Photophysiological Analyses of Microalgae." *Science Advances* 6 (36). https://doi.org/10.1126/sciadv.abb2754.

Bernstein, David B., Snorre Sulheim, Eivind Almaas, and Daniel Segrè. 2021. "Addressing Uncertainty in Genome-Scale Metabolic Model Reconstruction and Analysis." *Genome Biology* 22 (1): 64.

Bertsimas, Dimitris, and Jack Dunn. 2017. "Optimal Classification Trees." *Machine Learning* 106 (7): 1039–82.

Bertsimas, Dimitris, Angela King, and Rahul Mazumder. 2016. "Best Subset Selection via a Modern Optimization Lens." *Annals of Statistics* 44 (2): 813–52.

Blaise, Benjamin J., Gonçalo D. S. Correia, Gordon A. Haggart, Izabella Surowiec, Caroline Sands, Matthew R. Lewis, Jake T. M. Pearce, et al. 2021. "Statistical Analysis in Metabolic Phenotyping." *Nature Protocols* 16 (9): 4299–4326.

Bochner, Barry R. 2009. "Global Phenotypic Characterization of Bacteria." *FEMS Microbiology Reviews* 33 (1): 191–205.

Bowen, Benjamin P., Curt R. Fischer, Richard Baran, Jillian F. Banfield, and Trent Northen. 2011. "Improved Genome Annotation through Untargeted Detection of Pathway-Specific Metabolites." *BMC Genomics* 12 Suppl 1 (June): S6.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Cario, Anaïs, Marina Larzillière, Olivier Nguyen, Karine Alain, and Samuel Marre. 2022. "High-Pressure Microfluidics for Ultra-Fast Microbial Phenotyping." *Frontiers in Microbiology* 13 (May): 866681.

Chicco, Davide, and Giuseppe Jurman. 2020. “The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation.” *BMC Genomics* 21 (1): 6.

Collado-Vides, Julio, Pascale Gaudet, and Víctor de Lorenzo. 2022. “Missing Links Between Gene Function and Physiology in Genomics.” *Frontiers in Physiology* 13 (February): 815874.

Dedieu, Antoine, Hussein Hazimeh, and Rahul Mazumder. 2021. “Learning Sparse Classifiers: Continuous and Mixed Integer Optimization Perspectives.” *Journal of Machine Learning Research: JMLR*, 135, 22 (1): 6008–54.

Demain, Arnold L., and Sergio Sanchez. 2009. “Microbial Drug Discovery: 80 Years of Progress.” *The Journal of Antibiotics* 62 (1): 5–16.

DiMaggio, Peter A., Jr, Nicolas L. Young, Richard C. Baliban, Benjamin A. Garcia, and Christodoulos A. Floudas. 2009. “A Mixed Integer Linear Optimization Framework for the Identification and Quantification of Targeted Post-Translational Modifications of Highly Modified Proteins Using Multiplexed Electron Transfer Dissociation Tandem Mass Spectrometry.” *Molecular & Cellular Proteomics: MCP* 8 (11): 2527–43.

DiMucci, Demetrius, Mark Kon, and Daniel Segrè. 2018. “Machine Learning Reveals Missing Edges and Putative Interaction Mechanisms in Microbial Ecosystem Networks.” *MSystems* 3 (5). https://doi.org/10.1128/mSystems.00181-18.

Dittrich, Marcus T., Gunnar W. Klau, Andreas Rosenwald, Thomas Dandekar, and Tobias Müller. 2008. “Identifying Functional Modules in Protein-Protein Interaction Networks: An Integrated Exact Approach.” *Bioinformatics* 24 (13): i223-31.

D’Orazio, M., M. Murdocca, A. Mencattini, P. Casti, J. Filippi, G. Antonelli, D. Di Giuseppe, et al. 2022. “Machine Learning Phenomics (MLP) Combining Deep Learning with Time-Lapse-Microscopy for Monitoring Colorectal Adenocarcinoma Cells Gene Expression and Drug-Response.” *Scientific Reports* 12 (1): 8545.

Forchielli, Elena, Daniel Sher, and Daniel Segrè. 2022. “Metabolic Phenotyping of Marine Heterotrophs on Refactored Media Reveals Diverse Metabolic Adaptations and Lifestyle Strategies.” *MSystems* 7 (4): e0007022.

Furnham, Nicholas, Tjaart A. P. de Beer, and Janet M. Thornton. 2012. “Current Challenges in Genome Annotation through Structural Biology and Bioinformatics.” *Current Opinion in Structural Biology* 22 (5): 594–601.

Gralka, Matti, Shaul Pollak, and Otto X. Cordero. 2023. “Genome Content Predicts the Carbon Catabolic Preferences of Heterotrophic Bacteria.” *Nature Microbiology*, August. https://doi.org/10.1038/s41564-023-01458-z.

Gusfield, Dan. 2019. *Integer Linear Programming in Computational and Systems Biology: An Entry-Level Text and Course*. Cambridge University Press.

Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Ho, Tin Kam. 1995. “Random Decision Forests.” In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1:278–82 vol.1.

Hosmer, Jennifer, Marufa Nasreen, Rabeb Dhouib, Ama-Tawiah Essilfie, Horst Joachim

Schirra, Anna Henningham, Emmanuelle Fantino, Peter Sly, Alastair G. McEwan, and Ulrike Kappler. 2022. “Access to Highly Specialized Growth Substrates and Production of Epithelial Immunomodulatory Metabolites Determine Survival of Haemophilus Influenzae in Human Airway Epithelial Cells.” *PLoS Pathogens* 18 (1): e1010209.

Houle, David, Diddahally R. Govindaraju, and Stig Omholt. 2010. “Phenomics: The next Challenge.” *Nature Reviews. Genetics* 11 (12): 855–66.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning: With Applications in R*. Springer US.

Jewett, Michael C., Gerald Hofmann, and Jens Nielsen. 2006. “Fungal Metabolite Analysis in Genomics and Phenomics.” *Current Opinion in Biotechnology* 17 (2): 191–97.

Jim, Kam, Kush Parmar, Mona Singh, and Saeed Tavazoie. 2004. “A Cross-Genomic Approach for Systematic Mapping of Phenotypic Traits to Genes.” *Genome Research* 14 (1): 109–15.

Jolliffe, Ian T., and Jorge Cadima. 2016. “Principal Component Analysis: A Review and Recent Developments.” *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 374 (2065): 20150202.

Kehe, Jared, Anthony Ortiz, Anthony Kulesa, Jeff Gore, Paul C. Blainey, and Jonathan Friedman. 2021. “Positive Interactions Are Common among Culturable Bacteria.” *Science Advances* 7 (45): eabi7159.

Keleş, Sündüz, Mark van der Laan, and Michael B. Eisen. 2002. “Identification of Regulatory Elements Using a Feature Selection Method.” *Bioinformatics* 18 (9): 1167–75.

Kishony, Roy, and Stanislas Leibler. 2003. “Environmental Stresses Can Alleviate the Average Deleterious Effect of Mutations.” *Journal of Biology* 2 (2): 14.

Kritikos, George, Manuel Banzhaf, Lucia Herrera-Dominguez, Alexandra Koumoutsi, Morgane Wartel, Matylda Zietek, and Athanasios Typas. 2017. “A Tool Named Iris for Versatile High-Throughput Phenotyping in Microorganisms.” *Nature Microbiology* 2 (February): 17014.

Lancia, Giuseppe. 2008. “Mathematical Programming in Computational Biology: An Annotated Bibliography.” *Algorithms* 1 (2): 100–129.

Liaw, Andy, Matthew Wiener, and Others. 2002. “Classification and Regression by RandomForest.” *R News* 2 (3): 18–22.

Miller, Alan J. 1984. “Selection of Subsets of Regression Variables.” *Journal of the Royal Statistical Society. Series A* 147 (3): 389.

Mirza, Bilal, Wei Wang, Jie Wang, Howard Choi, Neo Christopher Chung, and Peipei Ping. 2019. “Machine Learning and Integrative Analysis of Biomedical Big Data.” *Genes* 10 (2). https://doi.org/10.3390/genes10020087.

Moffat, John G., Joachim Rudolph, and David Bailey. 2014. “Phenotypic Screening in Cancer Drug Discovery - Past, Present and Future.” *Nature Reviews. Drug Discovery* 13 (8): 588–602.

Ohya, Yoshikazu, Jun Sese, Masashi Yukawa, Fumi Sano, Yoichiro Nakatani, Taro L. Saito, Ayaka Saka, et al. 2005. “High-Dimensional and Large-Scale Phenotyping of Yeast Mutants.” *Proceedings of the National Academy of Sciences of the*

United States of America 102 (52): 19015–20.

Passi, Anurag, Juan D. Tibocha-Bonilla, Manish Kumar, Diego Tec-Campos, Karsten Zengler, and Cristal Zuniga. 2021. “Genome-Scale Metabolic Modeling Enables In-Depth Understanding of Big Data.” *Metabolites* 12 (1). https://doi.org/10.3390/metabo12010014.

Price, Morgan N., Kelly M. Wetmore, R. Jordan Waters, Mark Callaghan, Jayashree Ray, Hualan Liu, Jennifer V. Kuehl, et al. 2018. “Mutant Phenotypes for Thousands of Bacterial Genes of Unknown Function.” *Nature* 557 (7706): 503–9.

Ramachandran, S., P. Fontanille, A. Pandey, and C. Larroche. 2006. “Gluconic Acid: Properties, Applications and Microbial Production.” *Food Technology and Biotechnology*. https://www.semanticscholar.org/paper/cc25bd3a96f70da84e38a10924cea0448030a940.

Reimer, Lorenz Christian, Joaquim Sardà Carbasse, Julia Koblitz, Christian Ebeling, Adam Podstawka, and Jörg Overmann. 2022. “BacDive in 2022: The Knowledge Base for Standardized Bacterial and Archaeal Data.” *Nucleic Acids Research* 50 (D1): D741–46.

Roberts, Richard J. 2004. “Identifying Protein Function--a Call for Community Action.” *PLoS Biology* 2 (3): E42.

Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga. 2007. “A Review of Feature Selection Techniques in Bioinformatics.” *Bioinformatics* 23 (19): 2507–17.

Sauer, Uwe. 2004. “High-Throughput Phenomics: Experimental Methods for Mapping Fluxomes.” *Current Opinion in Biotechnology* 15 (1): 58–63.

Schilling, C. H., J. S. Edwards, and B. O. Palsson. 1999. “Toward Metabolic Phenomics: Analysis of Genomic Data Using Flux Balances.” *Biotechnology Progress* 15 (3): 288–95.

Schmidt, F. R. 2005. “Optimization and Scale up of Industrial Fermentation Processes.” *Applied Microbiology and Biotechnology* 68 (4): 425–35.

Seaver, Samuel M. D., Filipe Liu, Qizhi Zhang, James Jeffryes, José P. Faria, Janaka N. Edirisinghe, Michael Mundy, et al. 2021. “The ModelSEED Biochemistry Database for the Integration of Metabolic Annotations and the Reconstruction, Comparison and Analysis of Metabolic Models for Plants, Fungi and Microbes.” *Nucleic Acids Research* 49 (D1): D575–88.

Segrestin, Jules, Kevin Sartori, Marie-Laure Navas, Jens Kattge, Sandra Díaz, and Eric Garnier. 2021. “PhenoSpace: A Shiny Application to Visualize Trait Data in the Phenotypic Space of the Global Spectrum of Plant Form and Function.” *Ecology and Evolution* 11 (4): 1526–34.

Tamura, Makio, and Patrik D’haeseleer. 2008. “Microbial Genotype-Phenotype Mapping by Class Association Rule Mining.” *Bioinformatics* 24 (13): 1523–29.

Thommes, Meghan, Taiyao Wang, Qi Zhao, Ioannis C. Paschalidis, and Daniel Segrè. 2019. “Designing Metabolic Division of Labor in Microbial Communities.” *MSystems* 4 (2). https://doi.org/10.1128/mSystems.00263-18.

Vos, Marjon G. J. de, Frank J. Poelwijk, Nico Battich, Joseph D. T. Ndika, and Sander J. Tans. 2013. “Environmental Dependence of Genetic Constraint.” *PLoS Genetics* 9 (6): e1003580.

Wang, Lening, Pang Du, and Ran Jin. 2021. “MOSS-Multi-Modal Best Subset Modeling in Smart Manufacturing.” *Sensors* 21 (1). https://doi.org/10.3390/s21010243.

Wong, Brandon G., Christopher P. Mancuso, Szilvia Kiriakov, Caleb J. Bashor, and Ahmad S. Khalil. 2018. “Precise, Automated Control of Conditions for High-Throughput Growth of Yeast and Bacteria with EVOLVER.” *Nature Biotechnology* 36 (7): 614–23.

Yeung, Clarence Hue Lok, Nil Sahin, and Brenda Andrews. 2022. “Phenomics Approaches to Understand Genetic Networks and Gene Function in Yeast.” *Biochemical Society Transactions* 50 (2): 713–21.

Yu, Michael Ku, Michael Kramer, Janusz Dutkowski, Rohith Srivas, Katherine Licon, Jason Kreisberg, Cherie T. Ng, Nevan Krogan, Roded Sharan, and Trey Ideker. 2016. “Translation of Genotype to Phenotype by a Hierarchy of Cell Subsystems.” *Cell Systems* 2 (2): 77–88.

Zelezniak, Aleksej, Sergej Andrejev, Olga Ponomarova, Daniel R. Mende, Peer Bork, and Kiran Raosaheb Patil. 2015. “Metabolic Dependencies Drive Species Co-Occurrence in Diverse Microbial Communities.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (20): 6449–54.

Zheng, Wei, Natasha Thorne, and John C. McKew. 2013. “Phenotypic Screens as a Renewed Approach for Drug Discovery.” *Drug Discovery Today* 18 (21–22): 1067–73.

Zoccarato, Luca, Daniel Sher, Takeshi Miki, Daniel Segrè, and Hans-Peter Grossart. 2022. “A Comparative Whole-Genome Approach Identifies Bacterial Traits for Marine Microbial Interactions.” *Communications Biology* 5 (1): 276.

Main Figure Captions

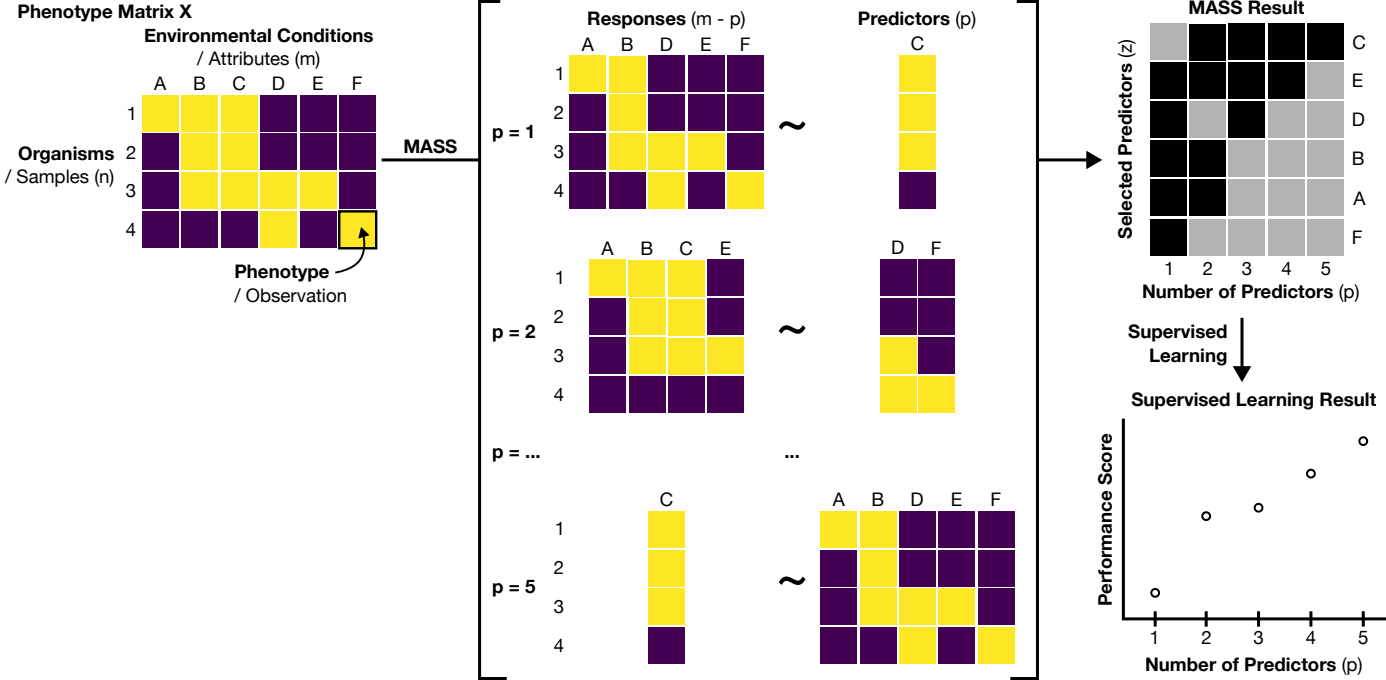
Figure 1. MASS takes as input a matrix (**X**) of observations of samples by attributes, in this case the phenotypes of **n** organisms under **m** different environmental conditions. For each fixed number of predictor variables, **p**, MASS provides as output a binary vector **z** indicating the predictor variables that predict the remaining response variables with the highest accuracy. Subsequently, the labeled data can be used to build models using supervised learning methods, such as random forest models, as done in this study.

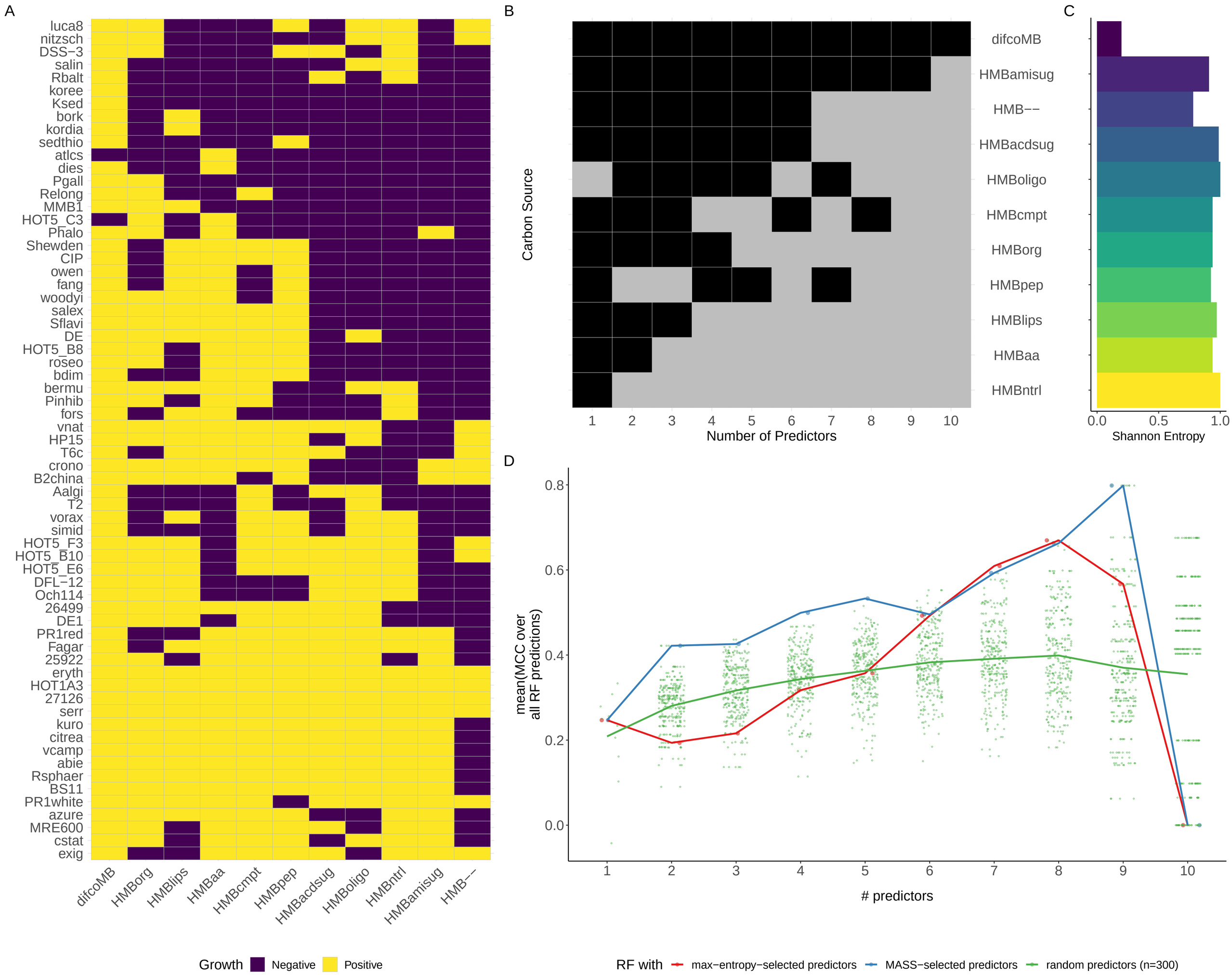
Figure 2. MASS successfully identifies predictor features for a small microbial growth dataset. (**A**) 65 different marine heterotrophic bacterial strains (rows) were grown individually on 11 different media (columns) (see (Forchielli, Sher, and Segrè 2022) for details) including Difco Marine Broth (difcoMB), eight engineered media with single classes of carbon sources (HMBpep = peptides; HMBaa = amino acids; HMBlips = lipids; HMBoligo = oligosaccharides; HMBorg = organic acids; HMBntrl = neutral sugars; HMBamisug = amino sugars; HMBacdsug = acidic sugars), a defined medium containing all 8 carbon classes (HMBcmpt), and a medium with no added carbon sources (HMB–). The name of the different strains are available in Supplementary Table S1. (**B**) Matrix showing which media were used as a predictor (gray) or as a response (black) as a function of the total number of predictors allowed (parameter *p*). (**C**) Shannon entropy of each medium. (**B** and **C**) Media are arranged in descending order of how frequently they were used as predictors. (**D**) Average Matthews correlation coefficient (MCC) of random forest classifiers for each number of predictors, *p*. The classifiers were trained either using the MASS selection of predictors (blue), or predictor sets selected based on maximum Shannon entropy (red) or 300 random draws of conditions used as predictors (green). Each point represents the mean MCC obtained via 5-fold cross-validation; the thick black line is the mean of those means across all MCC values for a respective *p*.

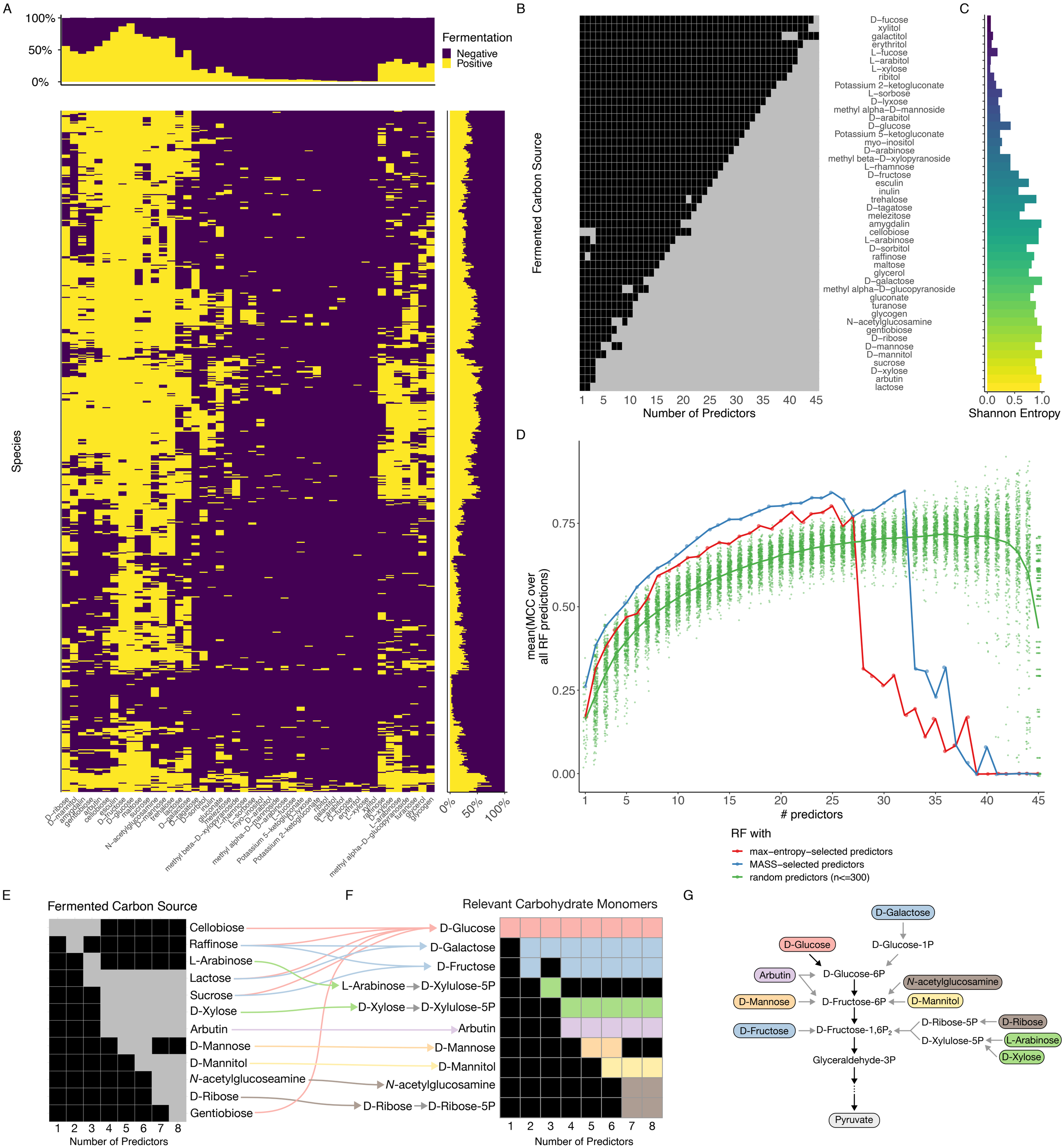
Figure 3. MASS is applicable for larger datasets and recapitulates biological understanding. (A) Fermentation phenotypes of 637 bacterial species (rows) grown on 46 different carbon sources (columns) downloaded from BacDive. Fermentation of a specific carbon source is indicated in yellow; and negative (no) fermentation is indicated in purple. Marginal bar charts summarize the phenotype frequency for each species (right) or each carbon source (top) respectively. (**B**) Matrix showing MASS result in which carbon sources were used as a predictor (gray) or as a response (black) as a function of the total number of predictors allowed (parameter *p*). (**C**) Shannon entropy of each carbon source. (**B** and **C**) Media are arranged in descending order of how frequently they were used as predictors. (**D**) Average Matthews correlation coefficient (MCC) of random forest classifiers for each number of predictors, *p*. The classifiers were trained either using the MASS selection of predictors (blue), or predictor sets selected based on maximum Shannon entropy (red) or 300 random draws of conditions

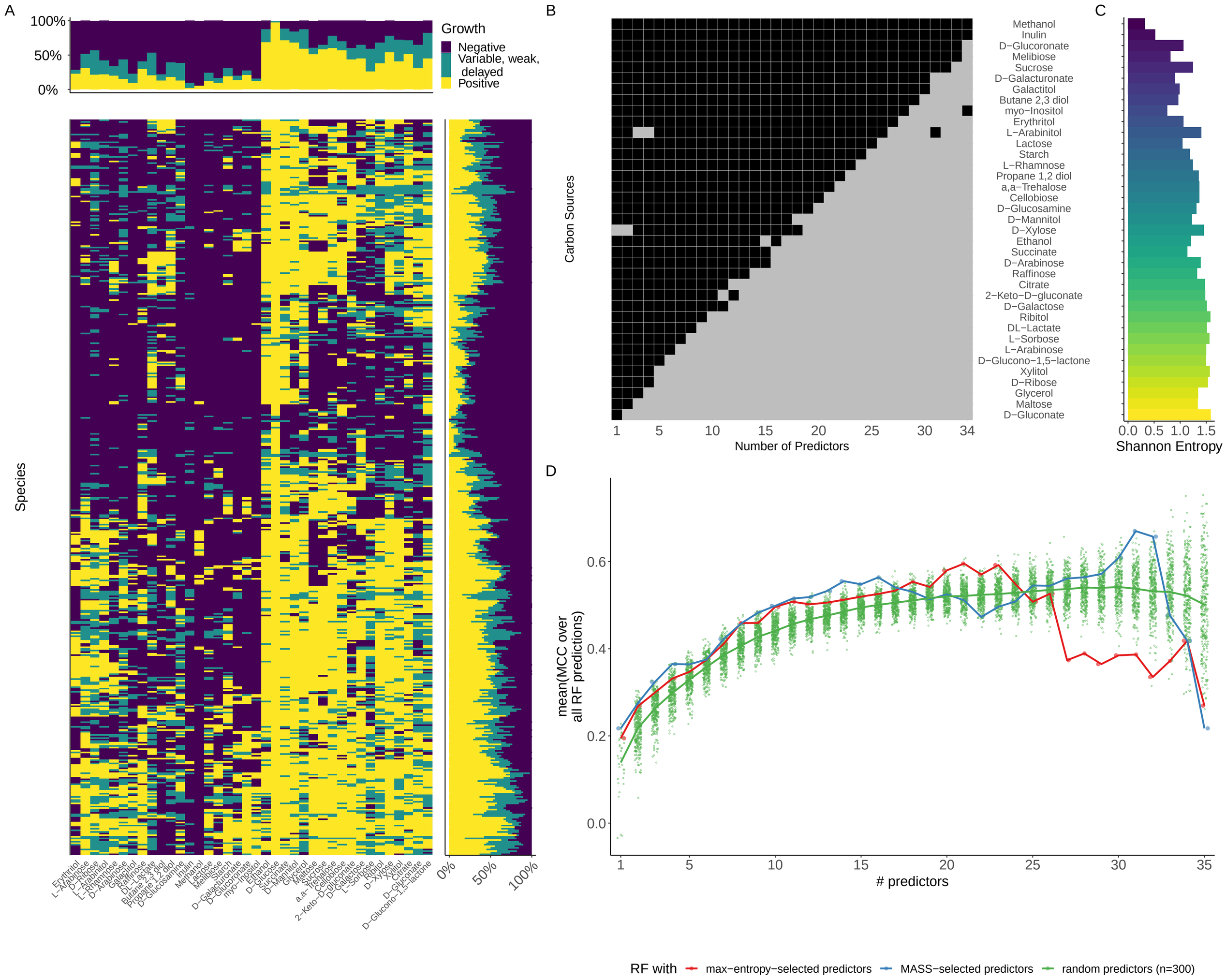
used as predictors (green). Each point represents the mean MCC obtained via 5-fold cross-validation; the thick black line is the mean of those means across all MCC values for a respective *p*. (**E**) Increasing the number of fermented carbon sources selected as predictors with MASS (focusing on up to *p* = 8 predictors) reveals a hierarchy of most descriptive carbohydrate monomers (**F**). (**G**) Those monomers enter central carbon metabolism via different routes. Only the relevant parts of glycolysis are shown, and reactions are differentiated into direct, single-step (black arrows) and indirect, multi-step reactions (gray arrow).

Figure 4. MASS is applicable for categorical data. (**A**) Growth phenotypes of 492 yeast species (rows) grown on 39 different carbon sources (columns) (Barnett, Payne, and Yarrow 1990). Positive growth on a specific carbon source is indicated in yellow; variable, weak, or delayed growth is indicated in teal; and negative (no) growth is indicated in purple. Marginal bar charts summarize the phenotype frequency for each species (right) or each carbon source (top) respectively. (**B**) Matrix showing MASS result in which carbon sources were used as a predictor (gray) or as a response (black) as a function of the total number of predictors allowed (parameter *p*). Growth on glucose was excluded from the MASS analysis. (**C**) Shannon entropy of each carbon source. (**B** and **C**) Media are arranged in descending order of how frequently they were used as predictors. (**D**) Average Matthews correlation coefficient (MCC) of random forest classifiers for each number of predictors, *p*. The classifiers were trained either using the MASS selection of predictors (blue), or predictor sets selected based on maximum Shannon entropy (red) or 300 random draws of conditions used as predictors (green). Each point represents the mean MCC obtained via 3-fold cross-validation; the thick black line is the mean of those means across all MCC values for a respective *p*.


Phenotype Matrix X
Environmental Conditions
/ Attributes (m)
A B C D E F
Organisms
/ Samples (n)
1 2 3 4
Phenotype
/ Observation
MASS
Responses (m - p)
Predictors (p)
p = 1
p = 2
p = ...
p = 5
MASS Result
Selected Predictors (z)
Number of Predictors (p)
Supervised
Learning
Supervised Learning Result
Performance Score
Number of Predictors (p)


A
luca8
nitzsch
DSS−3
salin
Rbalt
koree
Ksed
bork
kordia
sedthio
atlcs
dies
Pgall
Relong
MMB1
HOT5_C3
Phalo
Shewden
CIP
owen
fang
woodyi
salex
Sflavi
DE
HOT5_B8
roseo
bdim
bermu
Pinhib
fors
vnat
HP15
T6c
crono
B2china
Aalgi
T2
vorax
simid
HOT5_F3
HOT5_B10
HOT5_E6
DFL−12
Och114
26499
DE1
PR1red
Fagar
25922
eryth
HOT1A3
27126
serr
kuro
citrea
vcamp
abie
Rsphaer
BS11
PR1white
azure
MRE600
cstat
exig
difcoMB
HMBorg
HMBlips
HMBaa
HMBcmpt
HMBpep
HMBacdsug
HMBoligo
HMBntrl
HMBamisug
HMB−−
Growth
Negative
Positive
B
Carbon Source
difcoMB
HMBamisug
HMB−−
HMBacdsug
HMBoligo
HMBcmpt
HMBorg
HMBpep
HMBlips
HMBaa
HMBntrl
1
2
3
4
5
6
7
8
9
10
Number of Predictors
C
0.0
0.5
1.0
Shannon Entropy
D
mean(MCC over all RF predictions)
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
6
7
8
9
10
predictors
RF with
max−entropy−selected predictors
MASS−selected predictors
random predictors (n=300)


A
100%
50%
0%
Fermentation
Negative
Positive
Species
B
Fermented Carbon Source
Number of Predictors
C
Shannon Entropy
D
mean(MCC over all RF predictions)
# predictors
RF with
max−entropy−selected predictors
MASS−selected predictors
random predictors (n<=300)
E
Fermented Carbon Source
Cellobiose
Raffinose
L-Arabinose
Lactose
Sucrose
D-Xylose
Arbutin
D-Mannose
D-Mannitol
N-acetylglucoseamine
D-Ribose
Gentiobiose
Number of Predictors
F
Relevant Carbohydrate Monomers
D-Glucose
D-Galactose
D-Fructose
L-Arabinose
D-Xylulose-5P
D-Xylose
D-Xylulose-5P
Arbutin
D-Mannose
D-Mannitol
N-acetylglucosamine
D-Ribose
D-Ribose-5P
Number of Predictors
G
D-Galactose
D-Glucose
D-Glucose-1P
Arbutin
D-Glucose-6P
N-acetylglucosamine
D-Mannose
D-Fructose-6P
D-Mannitol
D-Fructose
D-Fructose-1,6P2
D-Ribose-5P
D-Ribose
D-Xylulose-5P
L-Arabinose
D-Xylose
Glyceraldehyde-3P
Pyruvate


A
100%
50%
0%
Growth
Negative
Variable, weak, delayed
Positive
Species
Erythritol
L−Arabinose
D−Ribose
L−Arabinitol
L−Rhamnose
D−Arabinose
Galactitol
Raffinose
D−Lactate
Butane 2,3 diol
Propane 1,2 diol
D−Glucosamine
Inulin
Methanol
Lactose
Melibiose
Starch
D−Galacturonate
D−Glucuronate
myo−Inositol
Ethanol
D−Glucose
Succinate
D−Mannitol
Glycerol
Maltose
Sucrose
a,a−Trehalose
Cellobiose
2−Keto−D−gluconate
D−Galactose
L−Sorbose
Ribitol
D−Xylose
Xylitol
Citrate
D−Gluconate
D−Glucono−1,5−lactone
0%
50%
100%
B
Carbon Sources
Number of Predictors
1
5
10
15
20
25
30
34
C
Methanol
Inulin
D−Glucoronate
Melibiose
Sucrose
D−Galacturonate
Galactitol
Butane 2,3 diol
myo−Inositol
Erythritol
L−Arabinitol
Lactose
Starch
L−Rhamnose
Propane 1,2 diol
a,a−Trehalose
Cellobiose
D−Glucosamine
D−Mannitol
D−Xylose
Ethanol
Succinate
D−Arabinose
Raffinose
Citrate
2−Keto−D−gluconate
D−Galactose
Ribitol
DL−Lactate
L−Sorbose
L−Arabinose
D−Glucono−1,5−lactone
Xylitol
D−Ribose
Glycerol
Maltose
D−Gluconate
0.0
0.5
1.0
1.5
Shannon Entropy
D
mean(MCC over all RF predictions)
0.0
0.2
0.4
0.6
1
5
10
15
20
25
30
35
predictors
RF with
max−entropy−selected predictors
MASS−selected predictors
random predictors (n=300)