Prediction of representative phenotypes using multi-output subset selection

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

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

* Equally contributing authors
# Corresponding authors

Abstract
Given an array of phenotypes (e.g., yield across strains and conditions), one can ask how to best choose subsets of conditions that are informative about the whole dataset, enabling efficient system identification and providing a basis vector in phenotype space. Here we introduce a mixed integer linear programming approach to choose explanatory and response variables for a phenotypic matrix. We applied the algorithm to a set of fitness measurements for 462 yeast strains under 38 carbon sources, and to the growth phenotypes of 65 marine bacteria on 11 media. The algorithm identifies environments that can be used as features to predict growth under other conditions, providing biologically interpretable metabolic axes for strain discrimination. Our approach could be used to reduce the number of experiments needed to identify a strain or to map its metabolic capabilities. The generality of the algorithm makes it appropriate for 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 has decreased, reading genomes is no longer the limiting factor in understanding microbes: interpreting genomes is. In principle, genomes carry information on how microbes 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, as well as by a limited understanding of how genes work together to produce higher-level functions. In many biological problems where the goal is to understand the ultimate behavior of an organism or system, the new frontier of “phenomics” has become increasingly valuable. Defined as the study of arrays of organismal behaviors (phenotypes) under multiple conditions or perturbations (Houle et al. 2010; Baran et al. 2013; Bowen et al. 2011; Baran et al. 2011), phenomics approaches the challenge of understanding biological systems from a top-down perspective: empirical phenotypes are measured, often in a high-throughput manner, and then analyzed in search for associations (between genotypes and phenotypes, or between different phenotypes). These associations can, in turn, be used to understand how the individual subsystem properties of diverse organisms give rise to cell-level or community-level functions, including mutual interdependencies in microbial communities. An important goal towards the development and applicability of these approaches is to understand and efficiently organize high-throughput phenotypic data of organisms grown in a variety of environments (Price et al. 2018; Barnett et al. 1990). This goal is the main focus of this paper.

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 these data, including Biolog Phenotype MicroArrays (Hosmer et al. 2022), robotic screening tools (Barton et al. 2018), and multiplexed bioreactors such as the eVolver (Wong et al. 2018). However, compared to the extreme speed in the decrease of 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. Narrowing in on “high value” phenotypes would have two major advantages for biologists: 1) reduce the experimental burden of large screens, and 2) provide insight into mechanistic links between traits.

The question of finding an optimal phenotype or set of phenotypes can be effectively modeled as a mathematical optimization problem that integer linear programming is well-suited to address. Integer linear programming (ILP) refers to linear optimization problems over integer variables and has been widely used to formulate problems where discrete choice decisions have to be made. In our setting, we seek to assign phenotype variables to either a set of predictors or response variables, giving rise to an extremely large number of possible solutions. ILP offers a methodology to search that solution space much more efficiently than exhaustively enumerating all possible solutions. Over the last decade, dramatic improvements in ILP solvers

and the widespread availability of high-dimensional biological datasets have created attractive opportunities to expand the applications of ILP 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 phenotype as a linear combination of phenotypes; methods capable of modeling multiple phenotypes as the combination of other phenotypes are still limited.

One of the questions specifically addressed in this study is whether it is possible to efficiently organize phenotypic datasets, based on the fact that some phenotypes may be predictable as combinations of other phenotypes. One challenge is that the search space can be very large and despite increasing capabilities, it is often very expensive to explore all possible combinations of variables. This work begins to address this challenge by describing the results of a mathematical framework that we call "multi-output subset selection" (MOSS), to understand which microbial phenotypes are most "informative." This method splits a dataset's features into responses and predictors and performs regularized linear regression using these features. The training of the regression model is performed by solving an ILP and exploiting 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. Although MOSS has its origins in subset selection problems (Miller 1984) and MILP (Bertsimas et al. 2016), in contrast to the earlier works it considers a setting with multiple response variables, which requires a novel formulation. We applied MOSS to two microbial phenotype datasets, and successfully identified sets of predictor variables with the highest predictive power. MOSS 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. MOSS 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

Given a dataset encoding phenotypic information (e.g., growth yield) for a variety of organisms under different environmental conditions, one can perform a regression to predict an organism's phenotype in environment *A* based on the linear combination of its phenotypes in environments *B* and *C*. However, it may be better (from an accuracy perspective) to use environments *A* and *B* to predict environment *C*, or some other combination. Multi-output subset selection (MOSS), formulated as a combinatorial optimization problem and described in detail in the **Methods**, selects p predictors from m environments in order to predict $m - p$ responses using linear regression. Accuracy is measured using the hinge loss function.

Inputs to MOSS (see **Figure 1** for an illustration) include a matrix (**X**) of organisms by phenotypes; the number of predictors (*p*); and the maximum and minimum allowable values on

the regression coefficients (*M*). Outputs of MOSS consist of a binary predictor vector (***z***) whose elements indicate whether an environment is a predictor or response, and the regression coefficients (**β**). Phenotypes (elements of ***z***) can alternate between being a predictor or response, irrespective of their state under different constraints.

MOSS uses regression to select predictors and responses, but a classification model would be better suited for our discrete growth data, which contains response variables mapping to one of three different categories (see **Methods**). Therefore, we elected to train random forest (RF) models (Breiman 2001) using the predictors determined by MOSS. A random forest is an ensemble learning model that can perform classification by constructing a large collection of decision trees and taking the majority vote of all the trees. Individually, decision trees are poor classifiers, but by democratically pooling their results, random forests can improve the predictive accuracy and control for overfitting (Ho 1995; Liaw et al. 2002). The number of RF models trained depend on the number of predictors to which MOSS has been constrained: for example, when applied to a dataset containing 11 phenotypes, if *p* was set to 1, 10 random forest models were trained, and when *p* was set to 10 predictors, only 1 random forest model was trained. The accuracy score was used as the metric to assess the random forest models (see **Methods**). Prior to training the models, we split the data into training and test sets (see **Methods**).

We designed MOSS with the goal of facilitating the prioritization of conditions under which a biological system should be characterized in order to maximize knowledge gained across a broader set of conditions. The question is the following: if we can choose only a subset of the environmental conditions, and aim at using phenotypes under such conditions as predictors of phenotypes under all other conditions, what is the ideal choice of conditions we should choose? Here we apply MOSS to two distinct datasets. The first dataset (DATASET 1) is a matrix of growth phenotypes (optical density (OD)) for 65 marine bacteria on 11 different media (see Forchielli et al. 2022 and **Figure 2**). This matrix is relatively small, and it pertains to a single set of experiments performed in our laboratory in order to map the metabolic capabilities of these heterotrophic bacteria on different carbon sources. The second dataset (DATASET 2) is a compendium of discrete growth phenotypes collected in a large book originally written as a practical guide for identification of different yeast strains based on their growth capacity under different conditions (Barnett et al. 1990).

Analysis of DATASET 1: Marine heterotrophic bacterial phenotypes on well-defined media

We systematically applied MOSS to DATASET 1 at different values of the number of predictors. For each number of predictors, MOSS yielded an optimal choice of media selected as predictors, which were then used as variables in a RF. We then built the RF (see **Methods**) and evaluated its performance in predicting growth on other media (i.e., those not selected as predictors), compared to a RF that uses all phenotypes except one as variables for predicting the phenotype left out.

MOSS labeled certain phenotypes as predictors more often than others (**Figure 3A**), potentially indicating that these frequently chosen phenotypes are more informative than phenotypes that were selected less frequently. To test this hypothesis, we used Shannon entropy (e) (see **Methods**) as a metric to quantify how informative each variable is (**Figure 3B**). In this dataset, difcoMB, which was never selected as a predictor, has the lowest entropy: almost all of the strains experience positive growth on this condition; thus, very few experiments must be conducted in order to "know" whether a strain grows on difcoMB. Despite this clear interpretation for the difcoMB condition, for the remaining media the relationship between Shannon entropy and W (the number of times a medium was selected) is poorly represented by a linear regression model (Adjusted R-squared = 0.056, p-value = 0.25) (**Figure S2A**). On the other hand, there is a mild negative relationship between the number of times a phenotype is selected as a predictor variable and the best accuracy score for that phenotype across all models (Adjusted R-squared = 0.32, p-value = 0.042). This suggests that MOSS does not necessarily select the most easily predicted phenotypes as response variables more often: when assigning class labels, MOSS considers a variable's ability to be predicted as well as its ability to predict other variables. Therefore, highly informative variables may not be selected as predictors if they are also difficult to predict. These selection patterns may hint at underlying relationships between the predictor variables, although it is also possible that the small size of the dataset and skewed distribution of entropy values are masking the relationship between entropy and selection.

Some phenotypes are more difficult to predict than others due to an apparent lack of relationship with other measured phenotypes, while others can be predicted more accurately due to associations with multiple phenotypes. No single medium predicts growth on any other medium well, but growth on certain media can be predicted almost perfectly with as few as 3 predictor variables (Accuracy = 0.95 for difcoMB). Model performance was maximized for 9 of the 11 media when p was less than 5. In fact, MOSS equals or outperforms RF alone for almost all phenotypes, which is to say that using all 10 remaining phenotypes as predictors results in poorer model performance compared to using a subset of phenotypes (**Figure 3C**). Only three phenotypes were predicted better using all the available data, but the differences in model performance were small (**Figure 3C**). One consequence of using all variables to predict one response is overfitting; using all of the available data may introduce too much noise to the model, and therefore negatively affect performance. Our observations suggest that selecting multiple predictor variables improves model performance, but the number and identity of the predictor variables is central to model performance.

Analysis of DATASET 2: Yeast carbon assimilation growth profiles

Following the application of MOSS to a binary dataset, we were interested in evaluating its performance on a categorical dataset that included more than two possible response values. For our main dataset, we used a subset of a vast phenotypic resource describing the aerobic growth of a variety of yeast strains on 44 different compounds as the sole carbon sources. This reference manual was initially used to identify yeasts using phenotypic tests, and therefore

describes the results of 91 physiological tests (Barnett et al. 1990). Descriptions of how these carbon assimilation tests were performed can be found in (Barnett et al. 1990).

Growth of each yeast strain was originally discretized into six categories: negative (no growth), variable, weak, delayed, positive, and unknown. Strains with missing (unknown) data were removed; one of two carbon sources that were highly linearly correlated were also removed; and the variable, weak, and delayed growth categories were combined into a single category (see **Methods**). Each carbon source and yeast strain exhibit different profiles (**Figure 4A, Figure 4C**). Some carbon sources (*e.g.,* glucose) could be utilized by almost all strains, while other carbon sources (*e.g.,* methanol and inulin) could only be utilized by very few strains (**Figure 4B**). Likewise, some strains exhibited "generalist" tendencies, since they were able to grow on most carbon sources, while others were "specialists," since they could only grow on a small portion of carbon sources (**Figure 4C**).

In order to determine which carbon sources best predict yeast growth on other carbon sources, we applied MOSS to the dataset of 462 yeast strains grown on 38 carbon sources (**Figure 4A**). The number of predictors, p, varied between using 1 and 37 carbon sources. D-gluconate was used as a predictor in almost all runs, except when there was only 1 predictor (**Figure 5A**), in which D-xylose was used instead. Inulin, methanol, and D-glucose were never used as predictors (**Figure 5A**), since the yeast strains either always grew or never grew on these carbon sources (**Figure 4A** and **Figure 5A**). In other words, inulin, methanol, and D-glucose contained less information (entropy) than the other carbon sources (**Figure 5B**). Generally, the larger a carbon source's entropy, the more often it was used as a predictor (**Supplementary Figure S2B**). This contrasts with the observations for DATASET 1; however, this is easily explained by the limited size of DATASET 1 in comparison to DATASET 2 and the skewed distribution of values, which could easily obscure the relationship between entropy and the number of times a phenotype is selected as a predictor.

We trained random forest models using 1 to 34 predictors, instead of 37 predictors, for each response carbon source, excluding D-glucose. Models for D-glucose were not trained because almost all strains exhibited positive growth. Overall, the accuracy increases as the number of predictors increases (**Figure 5C**), although not every carbon source is predicted equally well. When D-xylose is the only predictor, xylitol has the best accuracy (Accuracy = 0.70) and inulin has the worst accuracy (Accuracy = 0.26). However, inulin prediction accuracy increases as there are more predictors, while xylitol prediction accuracy decreases until it becomes a predictor (when $p = 5$). Furthermore, methanol often has the best prediction accuracy, followed by inulin, sucrose, and *myo*inositol, while DL-lactate often has the worst prediction accuracy. Overall, the yeast phenotypes are harder to predict for low values of p compared to the bacterial growth phenotypes. This could result from the increased difficulty of predicting categorical growth phenotypes compared to binary positive/negative growth. In the future, this issue could be better addressed by comparing the prediction accuracy for binary and categorical– or even continuous– growth using the same or more comparable datasets.

Discussion

The rise of large-scale phenotyping in biology is motivated by the enormous landscape of possible functions that cells can perform, and the value of being able to screen for those with relevance to specific applications. The development of high-throughput technology enables the production of large multidimensional datasets (for multiple phenotypes across multiple conditions) that can be extremely valuable for academic and industrial research, but that can be costly to generate and difficult to interpret. 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 screening assays. MOSS helps achieve this goal by identifying combinations of phenotypes that best predict other phenotypes within a dataset. Here, we focused on microbial growth in different environments, but it is not difficult to imagine other contexts in which MOSS could be applied: examples include predicting antibiotic sensitivity, metabolic engineering production rates, and microbial interactions.

Our results suggest that MOSS has powerful applications for large-scale phenotypic screens by greatly reducing experimental burden. Following an initial screen, MOSS allows experimentalists to select a subset of conditions to test depending on available resources, desired level of accuracy, and the phenotypes of interest. The MOSS/RF model provides a clear picture of the effort needed to achieve a certain level of results, which can vary greatly depending on the experimental stakes. For example, selecting metabolically compatible bacteria for engineered microbial communities may allow a higher margin of error because the consequences of incorrect predictions are less severe compared to applications that involve selecting drug candidates for *in vivo* studies. Furthermore, our results indicate that the quality of predictions varies by phenotype, possibly due to the inherent biology of the system and associated data structure, or to experimental design. In the future, it may be interesting to ask if specific types of data and organism/condition pairs are better captured by MOSS, and whether this can teach us anything about the underlying structure of the system itself. Phenotypic information processed through MOSS will allow researchers to prioritize efforts on high-value phenotypes or shift resource allocation to ensure the most important data is collected.

We should point out that the problem of feature selection is not new to biology, and considerable work has been done to improve model performance on high-dimensional biological datasets by extracting the most relevant variables from those that may be uninformative, irrelevant, or redundant (Saeys et al. 2007). The traditional response to this challenge has been dimensionality reduction using projection-based methods such as PCA, but these approaches do not preserve the original semantics of the data, complicating downstream analysis. On the other hand, feature selection techniques have focused mainly on supervised methods that require the *a priori* application of class labels (Varshavsky et al. 2006); without specific domain knowledge, few practical tools exist for the optimal subset selection of biological data. In addition, we are not aware of any methods that enable biologists to organize phenotypic datasets in such a way that highlights the best predicted features, although MILP is a well-

known method in other disciplines (Floudas 1995). To the best of our knowledge, MOSS is the first method that simultaneously solves a number of prediction problems by optimally selecting a subset of predictors and a complementary set of response variables.

From a biological perspective, the results of MOSS, in addition to helping prioritize and plan larger phenotypic screens from pilot ones, could also help 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. For example, in the analysis of the marine dataset, phenotypes are easy to predict because their growth is positively/negatively associated with growth on specific subsets of other media. This suggests that these metabolic traits are linked by mechanistic or ecological reasons. For example, growth on amino sugars is predicted with a high level of accuracy (Accuracy = .923) when growth on neutral sugars, amino acids, and peptides are used as predictors, which suggests overlapping metabolic pathways are utilized in the assimilation of these carbon sources. On the other hand, on the rare occasions that growth on amino acids was selected as a response variable, the accuracy was relatively low compared to the other media (Accuracy = .571 and .708 for $p = 1$ and 2, respectively). A possible explanation could be that the organisms grow on different subsets of the same medium (HMBaa contains all 20 standard amino acids) and the particular patterns between individual components or subsets cannot be detected due to the experimental design. Growth on the organic acid medium was also difficult to predict, suggesting that this trait is not linked to any particular set of carbon sources tested in the experiment, which could suggest that these particular organisms possess multiple different metabolic pathways for the incorporation of organic acids into biomass.

It is important to note the limitations of the algorithm and their implications for the biological interpretation of results. MOSS selects the optimal subset of features to explain the remaining features, but it does not optimize the predictability of any individual feature. In the future, it may be interesting to explore our formulation with tree-based models. Furthermore, in this work MOSS was applied to discretized fitness data; future developments could extend the current methods to be more comprehensive. We envisage that it would work well on continuous variables because the objective function can be easily substituted with l1-norm or l2-norm. This is a particularly interesting prospect, as it would add a layer to the biological interpretation of microbial phenotypes by considering the magnitude of growth in addition to the ability to grow under various conditions. We expect that there will be many more datasets similar to those analyzed in this work; our code is freely available, open source, and could be further developed for usage in myriad scientific disciplines.

Methods

Data pre-processing

Values from the table in Chapter 6 of (Barnett et al. 1990) were digitized into a table with 590 yeast strains (samples, n) and 44 carbon sources (features, m) (Supplementary Table S2). 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 were also dropped, reducing the number of samples to 462.

Categorical features 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 76. 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 response variables, $\boldsymbol{y}$, and predictor variables, $\boldsymbol{x}$, by estimating the parameters, $\boldsymbol{\beta}$, that provide the best explanation for the data:

$$\boldsymbol{y} = \boldsymbol{x}'\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad (\mathbf{1})$$

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:

$$\begin{aligned} &minimize_{\beta} \sum_{i=1}^{n} \left| y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \right| \\ such\ that \quad &-M \leq \hat{\beta}_{\boldsymbol{j}} \leq M, \forall j, \end{aligned} \qquad (\mathbf{2})$$

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, $\boldsymbol{X} = (X_{i,j}) = (\boldsymbol{x_1'}, \dots, \boldsymbol{x_n'})'$, 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):

$$\boldsymbol{X} = \boldsymbol{XB} + \boldsymbol{B}_0 + \boldsymbol{E}, \qquad \Leftrightarrow \ \boldsymbol{x}_i' = x_i'\boldsymbol{B} + \beta_0' + \varepsilon_i' \tag{3}$$

where $\boldsymbol{B} = (B_{i,j}) = (\beta_{\boldsymbol{1}}, \dots, \boldsymbol{\beta_m})$ is the $m \times m$ coefficient matrix, $\boldsymbol{B_0} = (\beta_0', \dots, \beta_0')'$ is the $n \times m$ constant matrix, $\beta_0 = (\beta_{0,1}, \dots, \beta_{0,m})'$ is the constant vector and $\boldsymbol{E} = (\varepsilon_1', \dots, \varepsilon_n')'$ is the noise matrix. Note that $\beta_{j,k}$-$(\beta_j)_k = B_{k,j}$ 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 and including a sparsity constraint, **Equation (3)** can be formulated as a mixed integer linear program:

$$\begin{aligned}
&minimize_{\boldsymbol{B},\boldsymbol{\beta_0},\boldsymbol{w},\boldsymbol{z},t} \ \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} w_{i,j} + \lambda \sum_{j=1}^{m} \|\beta_j\|_1 \\
&such\ that \qquad t_{i,j} - x_i'\beta_j - \beta_{0,j} \le w_{i,j}\ \forall i,j, \\
&\qquad -t_{i,j} + x_i'\beta_j + \beta_{0,j} \le w_{i,j}\ \forall i,j, \\
&\qquad \sum_{j=1}^{m} z_j \le p, \\
&\qquad -Mz_k \le (\beta_j)_k = B_{k,j} \le Mz_k\ \forall j,k, \\
&\qquad -Mz_j \le t_{i,j} - X_{i,j} \le Mz_j\ \forall i,j, \\
&\qquad w_{i,j} \ge 0, \forall i, \\
&\qquad z_j \in \{0,1\}, \forall j, \\
&\qquad B_{i,j} \in R, \beta_{0,j} \in R, \forall i,j,
\end{aligned} \tag{4}$$

where w is a dummy variable for the loss, reformulating the absolute loss using linear constraints; λ controls the sparsity or robustness constraint; 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 $(z_j = 0)$; x_i is the vector of the attributes of sample i; and β_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 $x_i'\beta_j$, 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 this dataset [9] involved categorical data, **Problem** (**4)** was further reformulated to use hinge loss for $X_{i,j} \in \{-1,1\}$:

$$
\begin{aligned}
&minimize_{B,\beta_0,w,z}\ \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} w_{i,j} + \lambda \sum_{j=1}^{m} \|\beta_j\|_1 \\
&such\ that \quad \sum_{j=1}^{m} z_j \le p, \\
&-Mz_k \le (\beta_j)_k = B_{k,j} \le Mz_k\ \forall j,k, \\
&X_{i,j}(x_i'\beta_j + \beta_{0,j}) \ge 1 - w_{i,j} - Mz_j, \forall i,j, \\
&w_{i,j} \ge 0, \forall i, \\
&z_j \in \{0,1\}, \forall j, \\
&z_j = z_k\ if\ (j,k)\ are\ in\ one\ group, \\
&B_{i,j} \in R, \beta_{0,j} \in R, \qquad \forall i,j.
\end{aligned}
\tag{5}
$$

Heuristics solutions to speed up solution time

The MILP **Problem (5)** is NP-complete. We observed that solving **Problem (5)** with larger p (*e.g.,* when the number of predictors is 34 or more) 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 in order 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 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 $p+1$:

$$z_{p+1}'\boldsymbol{z} \ge p, \tag{6}$$

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

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

$$z_{p+1}'\boldsymbol{z} \ge p-1, \tag{7}$$

using 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 z_{greed2}. In all experiments discussed in this paper we used z_{greed2}.

Using the predictors in random forest models

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. Two-fold 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.*). Models were built using 1 to 34 features as predictors, out of the 37 total features. When 1 feature was used as a predictor, 35 models were constructed, and when 34 features were used as predictors, 3 models were constructed, one model for each response feature, excluding D-glucose. Accuracy was used to evaluate model quality:

$$\text{Accuracy Score} = \frac{(\text{True Positives} + \text{True Negatives})}{(\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives})}.$$

Acknowledgments

IChP acknowledges funding by the NSF under grants IIS-1914792, DMS-1664644, and CNS-1645681, by the ONR under grant N00014-19-1-2571, by the DOE under grant DE-EE0009696, and by the NIH under grants R01 GM135930 and UL54 TR004130. 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) and the Human Frontiers Science Program (grant number RGP0060/2021).

Data availability

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

List of Abbreviations Used

LP: Linear Programming; MILP: Mixed Integer Linear Programming

References

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.
2. 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.
3. Barnett, J. A., R. W. Payne, and D. Yarrow. 1990. “Yeasts: Characteristics and Identification,” 1012 pp.
4. 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.
5. Bertsimas, Dimitris, Angela King, and Rahul Mazumder. 2016. “Best Subset Selection via a Modern Optimization Lens.” Annals of Statistics 44 (2): 813–52.
6. 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.
7. Breiman, Leo. 2001. “Random Forests.” Machine Learning 45 (1): 5–32.
8. Demain, Arnold L., and Sergio Sanchez. 2009. “Microbial Drug Discovery: 80 Years of Progress.” The Journal of Antibiotics 62 (1): 5–16.
9. 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.
10. 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.
11. Floudas, Christodoulos A. 1995. Nonlinear and Mixed-Integer Optimization: Fundamentals and Applications. Oxford University Press.
12. Forchielli, Elena, Daniel Sher, and Daniel Segrè. 2022. “Metabolic Phenotyping of Marine Heterotrophs on Refactored Media Reveals Diverse Metabolic Adaptations and Lifestyle Strategies.” bioRxiv. https://doi.org/10.1101/2022.01.07.475430.
13. Gusfield, Dan. 2019. Integer Linear Programming in Computational and Systems Biology: An Entry-Level Text and Course. Cambridge University Press.
14. 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.

15. Ho, Tin Kam. 1995. “Random Decision Forests.” In Proceedings of 3rd International Conference on Document Analysis and Recognition, 1:278–82 vol.1.
16. Houle, David, Diddahally R. Govindaraju, and Stig Omholt. 2010. “Phenomics: The next Challenge.” Nature Reviews. Genetics 11 (12): 855–66.
17. Lancia, Giuseppe. 2008. “Mathematical Programming in Computational Biology: An Annotated Bibliography.” Algorithms 1 (2): 100–129.
18. Liaw, Andy, Matthew Wiener, and Others. 2002. “Classification and Regression by randomForest.” R News 2 (3): 18–22.
19. Miller, Alan J. 1984. “Selection of Subsets of Regression Variables.” Journal of the Royal Statistical Society. Series A 147 (3): 389.
20. 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.
21. Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga. 2007. “A Review of Feature Selection Techniques in Bioinformatics.” Bioinformatics 23 (19): 2507–17.
22. Schmidt, F. R. 2005. “Optimization and Scale up of Industrial Fermentation Processes.” Applied Microbiology and Biotechnology 68 (4): 425–35.
23. 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.
24. Varshavsky, Roy, Assaf Gottlieb, Michal Linial, and David Horn. 2006. “Novel Unsupervised Feature Filtering of Biological Data.” Bioinformatics 22 (14): e507–13.
25. 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.

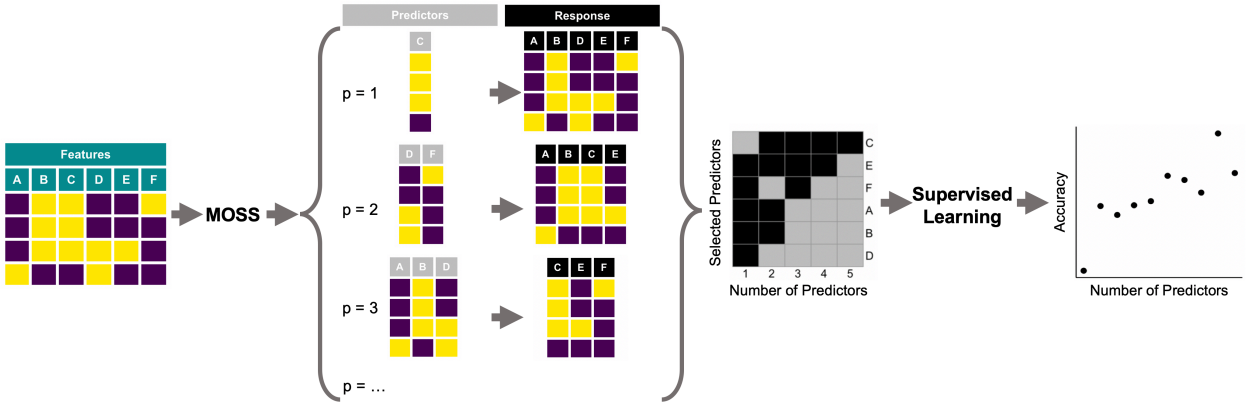

Figure 1. MOSS takes as input a matrix (**X**) of observations by variables, in this case organisms by features. For each fixed number of predictor variables, *p*, MOSS provides as output the predictor variable(s) that predict the remaining response variables with the highest accuracy. Subsequently, the labeled data can be used to build models using supervised learning methods, such as Random Forest.

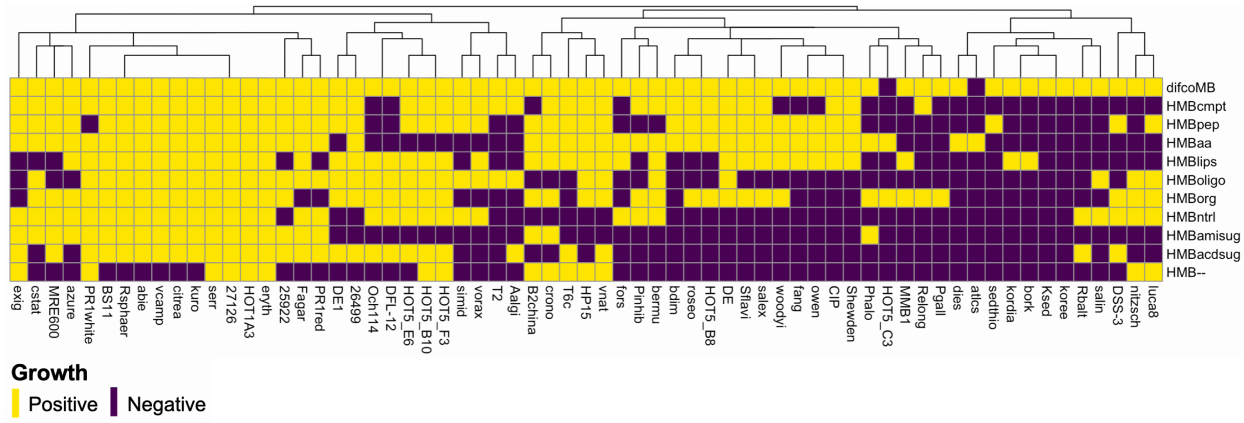

Figure 2. Growth profiles of bacteria grown on various media. 65 different marine heterotrophic bacterial strains (columns) were grown individually on 11 different media (rows) (see Forchielli et al. 2022 for details) including Difco Marine Broth (difcoMB), eight engineered media with single classes 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. The value of each phenotype represents final optical density (OD600) (normalized to time zero), with yellow indicating that the growth was significantly greater than the change in OD for the negative control without added bacteria (see details of criteria for significant growth in Forchielli et al. 2022). The data were clustered by strains just for visualization purposes. This heatmap corresponds to matrix **X** for dataset 1.

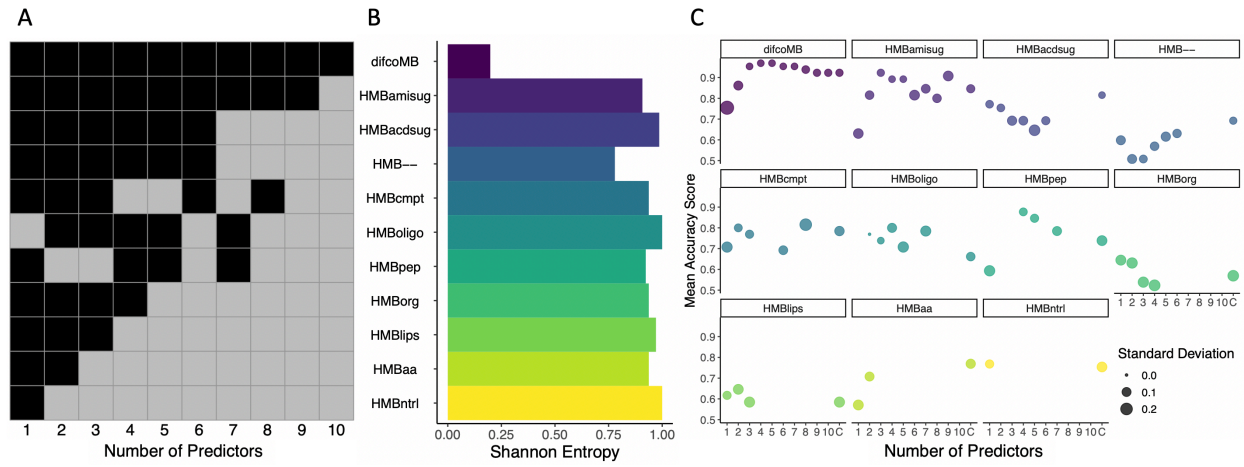

Figure 3. Accuracy of predictions based on MOSS-defined predictors. Outcome of MOSS predictions for the marine heterotrophic library dataset described in Fig. 2. (**A**) Matrix showing which media were used as a predictor (white) or as a response (black) as a function of the total number of predictors allowed. (**B**) Shannon entropy of each medium. (**A** and **B**) Media are arranged in descending order of how frequently they were used as predictors. (**C**) Accuracy of each random forest model for each number of predictors, *p*. Dot size represents the standard deviation of accuracy scores for values obtained via 5-fold cross-validation; column C is the RF control without MOSS; if no dot is displayed for a medium, the medium was selected as predictor.

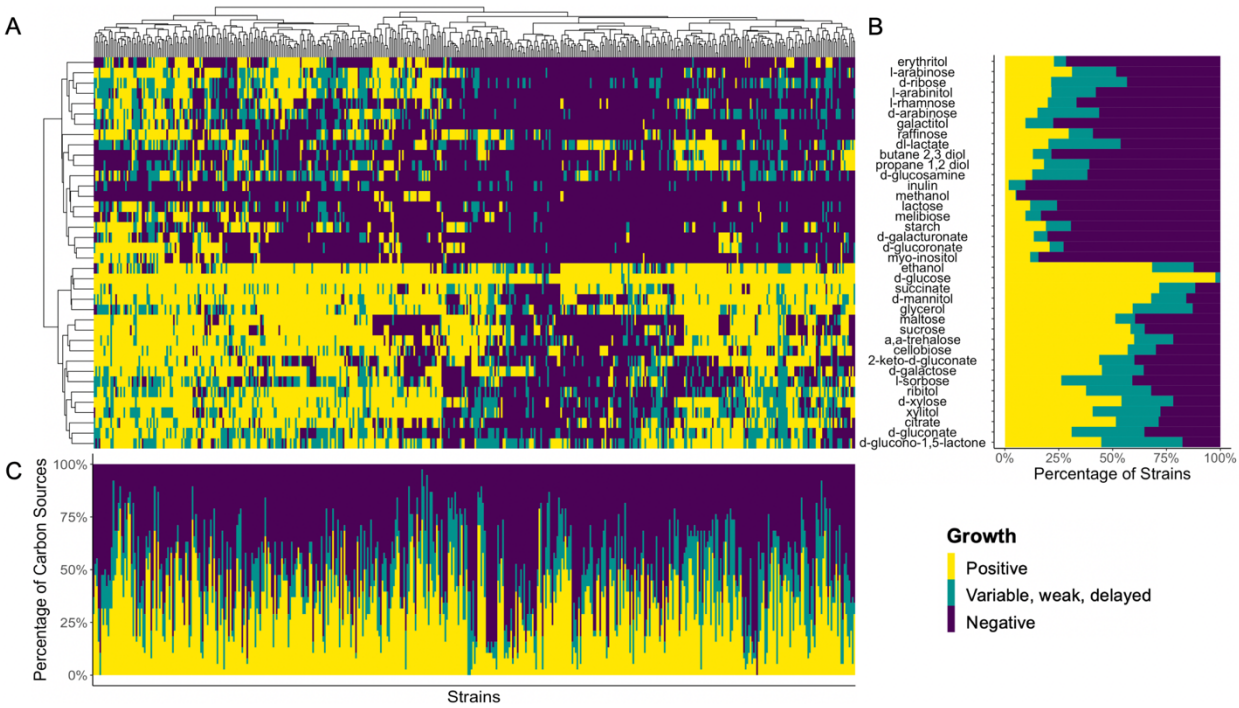

Figure 4. Growth profiles of yeast strains grown aerobically on various carbon sources. (**A**) Two-dimensional hierarchical clustering of the experimental phenotypes of 462 yeast strains grown aerobically on 38 different carbon sources from (Barnett et al. 1990). Each entry represents the phenotype of a yeast under a specific condition, where the columns are different yeast strains and the rows are the carbon sources. (**B** and **C**) Stacked bar charts of the growth profile of each carbon source (**B**) and each yeast strain (**C**). Positive growth is indicated in yellow; variable, weak, or delayed growth is indicated in teal; and negative (no) growth is indicated in purple.

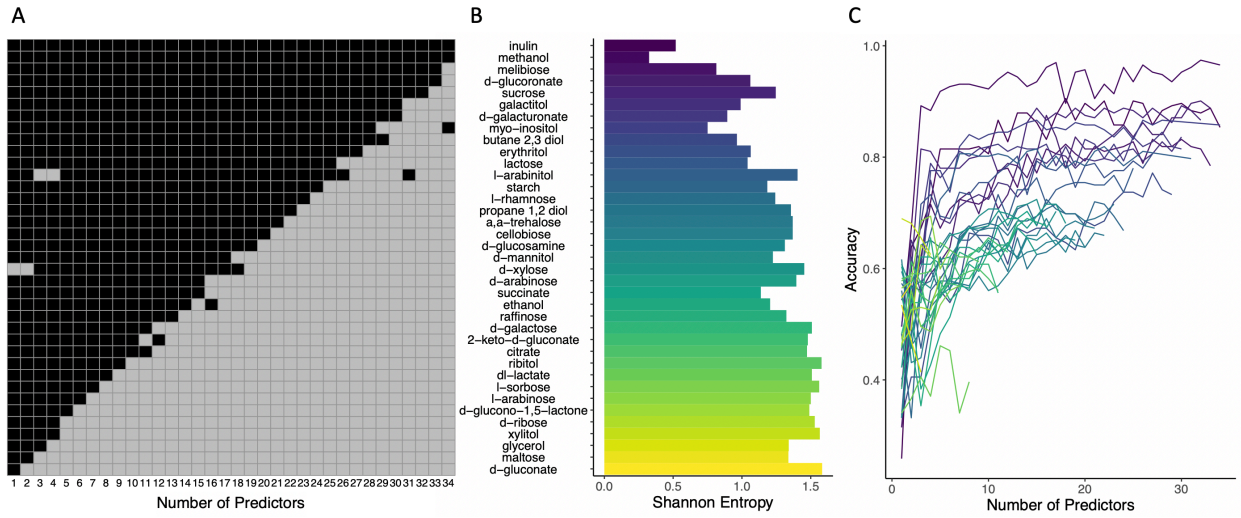

Figure 5. Carbon sources used as predictors. (**A**) Heat map of when a carbon source was used as a predictor (while) or as a response (black) as a function of the total number of predictors allowed. (**B**) Entropy of each carbon source. (**C**) Accuracy of each random forest model. Each line represents the carbon source that is being predicted.