

Computational Identification and Annotation of Key-reactions in Metabolic Pathways of *E. coli* that Discriminate Different Growth Conditions.

Viswanadham Sridhara¹, Austin G. Meyer^{1,2,5}, Piyush Rai³, Jeffrey E. Barrick^{1,2},
Pradeep Ravikumar³, Daniel Segre⁴, Claus O. Wilke^{1,5,*}

1 Center for Computational Biology and Bioinformatics, The University of Texas at Austin, Austin, TX, USA

2 Department of Chemistry and Biochemistry, The University of Texas at Austin, Austin, TX, USA

3 Department of Computer Science, The University of Texas at Austin, Austin, TX, USA

4 Department of Biology, Boston University, Boston, MA, USA

5 Section of Integrative Biology, The University of Texas at Austin, Austin, TX, USA

* E-mail: wilke@austin.utexas.edu

Abstract

Currently, predicting bacterial growth conditions without prior knowledge of the medium is an unresolved problem. By contrast, computing bacterial metabolic output, given a set of starting conditions has become a comparatively routine task via flux balance analysis (FBA). To compute metabolic output, one specifies a set of starting conditions within the context of a complete bacterial metabolic model. Here, we selected 7 carbon and 7 nitrogen sources along with 4 more commonly used experimental media. For these 18 growth conditions, we generated metabolic flux data using FBA with iAF1260 model for *Escherichia coli* MG1655 strain. We then used a model selection algorithm, LASSO, to identify the key reactions that discriminate among the tested growth conditions. These models on average require assaying the fluxes through 10.28 reactions to accurately predict the correct carbon and nitrogen source used during growth. For each source metabolite, we mapped its predictive reactions onto the *E. coli* central carbon metabolism to highlight important metabolic regions. Our analysis provides several important physiological and statistical insights. First, by analyzing metabolic end products, we can consistently predict growth conditions. Second, despite its heterogeneity, the experimental media appears to be similarly predictable to the homogeneous media. Third, predictive reactions seem to frequently lie near the initial entry point into central metabolism for the metabolite being predicted. Finally, we found that separately predicting the carbon and nitrogen sources is better than making joint predictions. In addition, the fact that separate prediction performs better than a more sophisticated joint prediction scheme, generates several potentially interesting hypotheses regarding bacterial physiology.

Introduction

Predicting bacterial growth conditions from its cellular composition is a problem of increasing interest. For example, this prediction can help identify the source of the pathogen

i.e., to know if the pathogen is a natural outbreak or a deliberate one, such as anthrax mailings. If we find the association between the growth conditions and the cellular composition, we can also engineer bacteria to produce useful vaccines or increase the production of chemicals of interest, such as, ethanol as bio-fuel. Here, we used models of metabolic pathways to understand this association. These models, also called metabolic or biochemical networks, contain hundreds of metabolites and reactions.

For some organisms, such as *E. coli*, these networks are well catalogued as there is experimental evidence for most of the genes that catalyze these metabolic reactions. These networks can then be used to understand the cell's metabolism under given conditions. Since these networks are large, mathematical models are used to understand them. A biochemical network is often represented as stoichiometric matrix, with dimensions of the matrix equal to the number of reactions and metabolites. Then, flux balance analysis (FBA) [1] is used on this network to analyze the flow of metabolites for a particular growth condition. To this end, we used *E. coli* iAF1260 genome-scale metabolic model (GEM) [2–5] to simulate metabolic fluxes. We then used sophisticated machine learning techniques to not only predict the growth conditions, but also identify the key-reactions that discriminate these growth conditions from others. Since the number of regressors, i.e., fluxes of all reactions for a particular growth condition is large, the nature of this prediction problem is high-dimensional. Moreover the available observations are generally smaller than the features (reaction flux). So, commonly used linear models for prediction cannot be used without reducing the number of features (some also referred to as covariates in machine learning and statistics). In such cases, prediction algorithms such as LASSO [6] work well, as they use regularization techniques to reduce the dimensions of the problem. LASSO assumes that only those coefficients in the model are nonzero that correspond to the relevant features in the data. Therefore, the LASSO also gives the model interpretability: by simply inspecting the model coefficients, we can infer which features are relevant for prediction.

In this study, we used LASSO [7] along with FBA to answer the following questions: 1. Can we predict the nutrient source on which the microbe is grown, given the simulated flux data? 2. In a mixture of different growth sources, can we predict each growth source separately, or whether simultaneously learning the prediction models for all sources more useful? 3. Can we make any mechanistic insights into features (here, reactions in metabolic pathways) that are the most relevant/informative for predicting the growth sources?

Results

Predicting growth conditions from simulated flux in *E. coli*

We wanted to know to what extent bacterial physiology reflects specifics about the growth environment. In other words, if we have measured the physiological state of a bacterium, can we deduce under what conditions it was grown? Here, we addressed this question in a simulation framework, using flux-balance analysis (FBA) as our model for bacterial physiology. Our overall strategy was as follows: (i) simulate metabolic fluxes under a variety of different growth conditions (primarily distinct carbon and nitrogen sources); (ii) develop regression models that regress the growth conditions against the calculated metabolic fluxes; (iii) evaluate how accurately the regression model can predict growth conditions from fluxes.

One inherent challenge with our approach is that flux-balance models do not allow for promiscuous reactions. Each reaction in the model has a small and unique set of reactants and a similarly minimal set of products. A biochemically similar reaction on a different substrate is represented as a separate reaction in the model. Further, substrates are brought into the cell and transported among different compartments in the cell via *transport reactions*, which simply take up a molecule of a specific metabolite in one compartment and release that same molecule in another compartment of the cell. Thus, any

metabolic flux model contains a substantial number of transport reactions whose sole purpose it is to get specific metabolites into the cell. Clearly, predicting environmental growth conditions from fluxes through these transport reactions would be trivial, and it would not be a reflection of what information the internal metabolic state of the cell holds about the external environment. To address this issue, we discarded all transport and exchange reactions in our regression analysis. In our model (iAF1260 metabolic model of the *E. coli* K-12 MG1655 strain [8]), this amounted to 939 reactions among a total of 2382 distinct reactions. We also discarded biomass composition reaction that resulted in total of 1442 reactions for regression analyses.

Further, to make the task of predicting growth conditions from fluxes more difficult and more realistic, we introduced background contamination in all simulated environments. Each environment consisted of a set of primary metabolites (usually one carbon and one nitrogen source) plus a small quantity of randomly chosen other metabolites. We varied the number of contaminant metabolites to evaluate how sensitive the regression model was to the amount of background contamination. Contaminant sources were selected at random from a set of 174 carbon and 78 nitrogen sources used previously with the *E. coli* model [4]. A different set of random contaminants was chosen for each individual FBA calculation. **Whenever we randomly picked 1 carbon source as contamination, we also picked an accompanying nitrogen source and we denoted this contamination pair as 1 C/N contamination in our analyses.**

We first wanted to test how well prediction might perform in a best-case scenario. To this end, we selected seven carbon and seven nitrogen sources (Table 1) that generated substantially distinct flux patterns in the absence of contaminants. We assessed the similarity of flux patterns by *k*-means clustering of fluxes obtained for all 174 carbon and 78 nitrogen sources (data not shown). We then generated fluxes for environments with contaminants for all pairwise combinations of the seven carbon and seven nitrogen sources. We generated up to 100 replicates of each pairwise combination, for a total of

4900 sets of flux values. We discarded solutions that we considered to be non-viable. We subdivided the remaining sets of flux values into two groups, a training data set and a test data set. We then fit a regularized regression model to the training data set and subsequently evaluated how well the model could predict growth conditions from fluxes on the test data set.

We considered two alternative approaches to prediction, joint prediction and separate prediction. Under joint prediction, we considered all 49 pairwise combinations of the seven carbon and seven nitrogen sources as distinct outcomes, and we trained a single model to predict one of those 49 possibilities. Under separate prediction, we trained two separate models, one for the seven carbon sources and one for the seven nitrogen sources. Overall, both prediction approaches worked quite well. Even at relatively high numbers of contaminants, we could correctly identify the main carbon and nitrogen sources in over 80% of the cases (Figure 2). And for very few contaminants i.e., 1C/N contamination level (randomly picked 1C and 1N source), the misclassification rate fell below 5%. Note that by random chance, we would expect a correct prediction only one time out of 49, i.e., by random chance the misclassification rate would be 98%.

To understand where the misclassifications are coming from, we plotted heatmaps that show the actual growth sources on x-axis and the predicted sources on y-axis at 2 contaminant levels (1 C/N and 10 C/N). At 10 C/N, some of the carbon sources are getting predicted as either acetate or pyruvate (Figure 3). A closer look at the key-reactions unique to these sources revealed that the reactions either are near the site of entry into the TCA cycle or in the TCA cycle. The role of TCA cycle is to generate energy, useful amino acids and also other cofactors such as NADH. So, clearly TCA cycle is of primary interest to almost all the organisms. This also means that given any carbon source, there should be some flux in the reactions that enter TCA cycle. So, as the contaminants increase, the flux resulting from these contaminants is seen through these reactions and hence some observations get mispredicted as acetate or pyruvate.

In a direct comparison, however, the separate prediction always outperformed the joint prediction (Figure 2). The performance gap was virtually independent of the amount of contaminants, but it did depend more strongly on the size of the training data set. In particular for smaller training-set sizes, independent prediction performed much better. We assume that the advantage at small sizes of training data sets arose because the independent prediction had effectively seven times more data to train than the joint prediction. For example, if the training data set was so small that it contained only one observation for each of the 49 joint conditions, it couldn't be used at all to train the joint model. However, two independent models (either carbon sources only or nitrogen sources only), there would be seven observations for each of the seven carbon or nitrogen sources.

Next, we looked into understanding the role of resource limitation on prediction results. In the current study, we used an uptake amount of $20 \text{ mmol gDW}^{-1} \text{ hr}^{-1}$ that is generally used for carbon and nitrogen sources in FBA studies. If these uptake amounts are not limiting to the growth, then the individual prediction might perform well than joint prediction. To make sure that the results still hold for resource limitation, we changed the lower bounds (uptake amounts) of carbon source, keeping the nitrogen source at initial level and vice versa. We then run the FBA analyses and we plotted the uptake amounts for different upper bounds as scatter plots in Supplementary info (Supplementary Figure 1 and Supplementary Figure 2). Regressing the growth conditions again gave similar results with these conditions too i.e., individual prediction performed better than joint prediction (data not shown).

Since individual prediction seemed to work well, we next tested whether we could use this approach to predict growth conditions chosen from the comprehensive list of 174 carbon and 78 nitrogen sources. Joint prediction in this case was infeasible, since we would have had to train a model to distinguish between $174 \times 78 = 13572$ distinct conditions. To test independent prediction in this case, we generated simulated fluxes for all pair-wise combinations of carbon/nitrogen sources for two replicates. We used one replicate to train

the regression model and we used the second replicate to evaluate the prediction accuracy of the model. We found that the misclassification rate for carbon sources was 86.3% and the misclassification rate for nitrogen sources was 37.2%. In combination, the two models identified the correct carbon/nitrogen combination 8.7% of the time. By random chance, we would have expected $1/13572 = 0.007\%$.

Identifying the predictive fluxes

The previous subsection has shown that a regularized regression model is capable of predicting the primary carbon and nitrogen sources used from steady-state metabolic fluxes. We next wanted to investigate how exactly the regularized regression model carries out this task. For each flux-balance simulation, the resulting flux data set contains 1443 flux values, corresponding to 1443 reactions that are not transport reactions. Thus, we have 1443 predictor variables that we feed into the regression model. In this situation, a standard regression model would have to determine 1444 regression coefficients, one per reaction plus an intercept. By contrast, the regularized regression model we employed sets most regression coefficients to zero and retains only a small number of non-zero coefficients. (The exact number of non-zero coefficients is determined through the choice of a tuning parameter, which is selected by cross-validation. See Methods for details.) Thus, we can consider the fluxes with non-zero regression coefficients as *predictive fluxes*. Those are the fluxes whose state is actually used for prediction.

At 1C/N contaminants level and using the largest training data size (see Figure 2), there were 72 key reactions discriminating 7 carbon sources and 72 key reactions discriminating 7 nitrogen sources. So, in total there are 144 key reactions, which average out to 10.28 per source. Out of these 144 reactions, 104 reaction IDs are unique. (There are few reaction IDs that have the same reaction name.) Excel table showing these growth source-reaction pairs is provided in Supplementary info for both carbon and nitrogen sources (Supplementary Table 1 and Supplementary Table 2).

To gain mechanistic insight into predictive reactions, we mapped them onto the *E. coli* central metabolism (Figures 4 and 5, Supplementary Table 1 and Supplementary Table 2). Note that each of the metabolic maps is meant to highlight only the central carbon metabolism in the *E. coli*. We found that each carbon or nitrogen source had a few reactions that were predictive to that growth source, and these reactions generally made physiologic sense. For example, using acetate as the carbon source unsurprisingly isolates TCA cycle entry as a predictive reaction (Figure 4). The key-reactions identified for D-glucose source were glucose 6-phosphate dehydrogenase (G6PDH), glucose-6-phosphate isomerase (PGI) and 6 phosphogluconolactonase (PGL). A literature search to look into the glucose metabolism of *E. coli* revealed that these 3 genes, when knocked off either separately or jointly resulted in diminished growth rate [?]. This agrees very well with our predictions on simulated fluxes generated with biomass maximization as an optimization function. Similarly, sorbitol (the singly reduced alcohol of D-glucose), and fructose each possess predictive reactions in the relative vicinity of the glycolytic pathway (Figure 4). Mapping nitrogen sources to central metabolism reveals a similar trend. For example, L-alanine as a growth source has predictive reactions near its site of entry into the three and four carbon metabolism of the TCA cycle (Figure 5).

We also analyzed how the regression model performed when some of the key predictive reactions were removed. As mentioned above, there were 104 unique reaction IDs for individual prediction of carbon and nitrogen sources at lowest contamination level and with the largest training data set analyzed. We knocked-off each of these 104 reactions at a time, trained a new model separately for both the Carbon and Nitrogen sources, and calculated the prediction accuracy. We combined the results of individual predictions to calculate the prediction accuracy of the combination of the sources. With the exception of knocking-off 'glucose 6-phosphate isomerase' (PGI), the misclassification rate remained unchanged when we knocked-off the other 103 reactions before model fitting. PGI catalyzes reaction that produces fructose-6-P from glucose-6-P. This is a reversible reaction,

although the forward reaction is catalyzed most of the times unless concentrations of fructose-6-P change significantly. A literature search to see the experimental evidence of the importance of knocking this PGI gene pointed out to a study that showed diminished growth rate [9]. This also shows the importance of simulations to first predict the essential genes before conducting time consuming gene knockout studies. Thus, even though the regression model needed on average 10.28 reactions to make a prediction, the specific set of reactions used for successful prediction is not unique.

Predicting specific media or novel metabolites

In order to generalize our simulations to more experimentally relevant test conditions, we performed similar analyses for several media that are more commonly used in experimental microbiology. Specifically, we tested autoinducer bioassay (AB) minimal media, proprietary media from the company ATCC, Davis Mingioli (DM) media and Bochner defined minimal media. Table 2 shows the composition used in the analysis for these growth media. We were able to classify these at higher accuracy even at higher noise levels (20 contaminants) and lower number of replicates (50). This may be due to the small (4) number of growth conditions used in the study. Our misclassification rate was less than 1% for noise levels up to 20 C/N contaminants.

Finally, we wanted to determine how the prediction would perform on previously unseen carbon or nitrogen sources. We first obtained simulated flux measurements using maltose as carbon source and using either of the seven nitrogen sources used earlier. We generated simulated flux data for 100 replicates and at 1C/N and 20C/N contaminant level. This resulted in 700 observations and after using a threshold, there were 695 viable flux measurements at 1C/N contaminant level. We used all these observations for testing. Note that we trained the model using the carbon/nitrogen sources in Table 1. When we tested individual prediction of either carbon or nitrogen sources, we found that maltose was classified as glucose over 85% of the time. Since maltose is a disaccharide formed

from two units of glucose, this prediction is reasonable. At the same time, the seven nitrogen sources were predicted correctly over 95% of the time (**how does this compare to the previous result when no novel C source was used? Has the choice of C source an effect on the predictive power for N sources?**). **Claus, I did not see any effect of C source on the predictive power for N sources. Likewise, when I compare to the study where I used unseen nitrogen source (please see below), I did not see any effect of N source on the predictive power for C sources. But I did this study for 1C/N contaminants.** However, when we tried to predict using the joint model, we found that using an unknown carbon source had a substantial effect on the model's ability to predict nitrogen sources. 33% of the growth conditions were predicted to be sorbitol/putrescine. While sorbitol is a reasonable choice considering the model had never seen maltose (sorbitol is the singly reduced alcohol of glucose), putrescine is not a good prediction for nitrogen sources the model has been trained on.

For 20C/N level, there were 699 viable flux measurements. At this contamination level, maltose was predicted as glucose 68% of the times, while the correct nitrogen source was predicted 81% of the times. Another interesting result is that 42% of the times, the observations were predicted as D-glucose/adenine. For the 2 different contaminant levels, individual prediction seem to outperform joint prediction and would help in separately predicting all the known growth sources, while predicting the unknown ones to its nearest compound.

Next, we did simulations to test how unseen nitrogen source gets predicted with the above developed models. For this, we used cytosine as a nitrogen source and either of the 7 carbon sources used earlier. Note that cytosine is one of the 4 bases found in DNA and RNA. We used 2 contamination levels i.e., 1C/N and 20C/N. Like earlier, we generated 100 replicates for testing. At 1C/N, there were 602 viable flux measurements i.e., for these measurements, biomass was greater than the threshold used in this study. Interestingly, all 98 non-viable flux measurements were for Cytosine + Acetate sources. For the viable flux

measurements, only 5 carbon sources were wrongly predicted (0.01% misclassification). Interestingly, in all the 100% cases, cytosine nitrogen source was predicted as ammonia. This can be due to the amine group at position 4.

At 20C/N contamination level, all the 700 flux measurements are viable (biomass greather than threshold). It looks like the background carbon and nitrogen sources are contributing to the (cytosine+acetate) sources to make the biomass larger than the threshold compared with 1C/N level. 27 carbon sources were wrongly predicted (0.04% misclassification) at this contamination level. The nitrogen source cytosine is predicted as ammonia 78.8% of the times, while it is predicted as adenine rest of the times.

Discussion

We have developed a method for making predictions regarding growth media from known metabolic flux data. We generated fluxes by simulating the complete E. coli metabolic network for 7 carbon, 7 nitrogen, and 4 experimental mixed media types. Then, we divided the data and employed machine learning with a generalized linear framework to train a model to predict growth conditions. We found that even at high noise levels, we could make reliable predictions regarding growth media for all of the sources we tested. Also, extending our prediction algorithm to more experimentally relevant growth media, our scheme gave comparable accuracy. Of note, our data indicates separately predicting carbon and nitrogen sources always performed better than joint prediction as paired input. Although this result is to some extent influenced by the volume of training data, it very likely says something important about rate-limiting reactions in the E. coli metabolic network. In addition, our results indicate that for most input metabolites at least one predictive reaction commonly occurs near its entry point to central metabolism. Finally, we found that the number of reaction fluxes required to make accurate predictions is relatively small and can probably be reduced further with few trade-offs. Thus, we have shown that predicting growth conditions from metabolic flux data is an experimentally

tractable problem.

We have shown that given simulated metabolic flux data, growth conditions can be accurately predicted via machine learning. Although the background contamination can have a dramatic affect on model accuracy, the misclassification rate remains acceptably low even with 10C/N or 20C/N randomly picked contaminants. The addition of this contamination revealed one interesting and unexpected physiological hypothesis about *E. coli* metabolism. Namely, as the contamination increases from 1C/N to 20C/N, our model increasingly predicts acetate as the default carbon source and ammonia as the default nitrogen source. Due to its centrality in terms of energy production, for any input growth source the reactions that lead to the TCA cycle should have some reasonable amount of reaction flux. In other words, acetate and ammonia as default nutrient sources may not be so surprising when one considers acetate’s central role in the TCA cycle—it is essentially the center of bacterial metabolism. Further, ammonia is one of the few, if not the only, source of nitrogen without any associated carbon atoms; as a result, it is unique among the input nutrient sources we tested. To be sure that the observed default carbon source misclassification was not an artifact of nutrient limitation (carbon versus nitrogen), we changed the uptake rates of carbon and nitrogen sources so that there was no limiting factor and re-analyzed by training a new model. Reversing nutrient limitations appears to have no effect on the default behavior of our trained model.

It was surprising to us that given the same number of observations in the training set, separate prediction of starting materials always performed better than joint prediction. There are two likely explanations for this result. First, making joint predictions requires discriminating between 49 different pairwise combinations. By contrast making individual predictions only requires discriminating 7 different conditions in two different sets. Thus, one possible explanation for the lack of predictive power is that we simply did not have the appropriate level of training data. Indeed adjusting the amount of training data appears to have a dramatic effect on joint prediction in particular (Figure 2). On

the other hand, such an issue represents an important experimental concern. Often the size of the training set, being experimentally determined, is just as limiting as the size of the testing set. As a result, our analysis indicates employing a separate prediction strategy will generally be more useful for experimental application. Second, although the mechanism is not completely clear to us, separate prediction may gain additional power due to the physiology of the organism. For example, if the initial, metabolite unique, steps of metabolic entry are often predictive (as they appear to be), running independent predictions would be expected to perform better per amount of data; in essence such a prediction strategy makes the assumption that pathways for the various metabolites are largely disconnected. By contrast, if one were using a single metabolite as a combined carbon and nitrogen source, we may expect an independent prediction strategy to perform relatively poorly as the independence assumption is not satisfied.

In this study, we used a relatively common machine learning technique called LASSO to prevent over fitting during feature selection in the regression model. To our knowledge, LASSO methods have not previously been applied to analyze metabolic pathways. By contrast, there are studies applying LASSO to other biochemical networks (such as gene regulatory networks) [10] or identifying SNPs in GWAS studies [11]. We would like to point out that beyond LASSO there are a number of other commonly used regularization techniques. For example, graphical Lasso [12] and Ising Markov Random Field models [13] can also be used to study biological networks. We chose LASSO because it provides a relatively simple and particularly robust framework for feature reduction. Thus, considering the large size of our simulation model, we were able to achieve a remarkably small number of source-predicting reactions.

Finally, we have shown that there is no obvious experimental restriction for applying FBA and machine learning to predict initial growth media from final metabolic flux data. Nine reaction fluxes on average provided the optimal solution to our regression model; however it is evidently not a unique solution. There are very likely other possible

alternative solutions that may garner similar predictive power. By individually eliminating reactions and retraining the model, it appears the minimum number of critically important reactions is three for *E. coli* MG1655. With such a small number of necessary reactions for gaining predictive power in reverse flux balance analysis, it should be possible to immediately apply this technique to experimental prediction.

Materials and Methods

We used MATLAB and R for this study. For flux balance analysis, we used COBRA toolbox [14] with MATLAB and for multinomial regression, we used GLMNET package [7] with R. The methods are described in detail below.

Flux Balance Analysis

We carried out flux balance analysis using the COBRA toolbox [14] for MATLAB. We used the iAF1260 model from the BiGG database [8]. In the current iAF1260 model, there are 2382 reactions involving 1668 metabolites. The biomass composition reaction is also included in the model. Except for the input growth sources (carbon and nitrogen sources), we left all parameter settings at their default for this model. The upper bounds on 2377 reactions is set to $1000 \text{ mmol gDW}^{-1} \text{ hr}^{-1}$, i.e., there is virtually no limit on the production of metabolites involved in these reactions. But for 5 reactions, i.e., ATPM, CAT, FHL, SPODM, SPODMpp, the upper bound is set to $50 \text{ mmol gDW}^{-1} \text{ hr}^{-1}$. The lower bound for the majority of reactions (nearly 1800) is set to $0 \text{ mmol gDW}^{-1} \text{ hr}^{-1}$, meaning that these reactions cannot uptake any metabolites from the media. A set non-growth associated maintenance (NGAM) of $8.39 \text{ mmol gDW}^{-1} \text{ hr}^{-1}$ is used for ATPM reaction. ATPM is ATP maintenance requirement value. The lower bounds of some exchange reactions was set to non-zero values i.e., these reactions by default were meant to uptake some ions, carbon and nitrogen sources etc. We did not change the default oxygen uptake rate (-18.5

mmol gDW⁻¹ hr⁻¹) in the iAF1260 model, but we set the lower bounds of glucose and ammonia to zero. If we used glucose/ammonia as growth sources, we then set the lower bounds of these sources accordingly.

Growth conditions

We used k -means clustering to pick the growth conditions using data from [4]. In our study, we used pairwise combinations of 7 carbon sources, and 7 nitrogen sources. 7 carbon sources when used alone did not result in any growth. On the other hand, the nitrogen sources except ammonia contributed to E. coli biomass composition that is non-zero. These carbon and nitrogen sources are listed in Table I. Depending on the input growth, we set the lower bound of that particular exchange reaction to -20 mmol gDW⁻¹ hr⁻¹. This lower bound of -20 mmol gDW⁻¹ hr⁻¹ is previously used as reasonable uptake amount in many studies [4]. For 49 pair-wise combinations of the sources, we generated 100 replicates of data. Apart from these growth conditions, we also used 4 growth media, generally used in E. coli K-12 MG1655 experiments as cited in EcoCyc database [Cite URL of EcoCyc].

Next, we varied the uptake amounts of carbon sources, keeping nitrogen at the same uptake value used previously. This means we changed the lower bounds (uptake rate) of nitrogen sources to -1000 mmol gDW⁻¹ hr⁻¹, keeping the lower bounds of carbon sources to -20 mmol gDW⁻¹ hr⁻¹ to study carbon limiting case. We repeated the analysis keeping the carbon source lower bounds (uptake rate) to -1000 mmol gDW⁻¹ hr⁻¹ and keeping the lower bound of nitrogen sources to -20 mmol gDW⁻¹ hr⁻¹. This change in uptake values helps validate the results of joint prediction and separate prediction in a more robust way.

Contaminants

To make the simulation scenario more challenging and more realistic, we incorporated different numbers of contaminants to the simulated growth media. For this, we used a

subset of the 174 carbon and 78 nitrogen sources, previously used in Feist et. al [4]. We used different background contamination levels, ranging from 1C/N to 20C/N sources. For example, if we want to set 5C/N contamination level, we randomly picked 5 Carbon and 5 nitrogen sources and set their lower bounds to $-0.2 \text{ mmol gDW}^{-1} \text{ hr}^{-1}$. Please note that we generated the flux data for a pairwise combination of 1 carbon and 1 nitrogen source along with the background noise as described above.

For all the results described above, we used a biomass threshold to filter out non-viable flux measurements. We calculated this threshold value using biomass measurements at the lowest contamination level case (1C/N contaminant pair) and with the largest training dataset size (2450 replicates). Result for (mean - 3*standard deviation) of all the biomass values is used as threshold value, which came out to 0.558.

Regularized regression

We predicted growth conditions using regularized multinomial logistic regression, as implemented in the GLMNET package [7] for R. Other parameters used were standard settings that are generally used in GLMNET package.

After filtering for biomass, for each number of contaminants, we used half of the dataset as test set. We used subsets of the remaining half as training (i.e, 245, 490, 2450 observations). On the training sets we did 3-fold cross validation. We used cross-validation in GLMNET package for model selection. Model selection means picking the regression coefficients at the lambda value that had the lowest misclassification rate with 3-fold cross-validation. We then used this model to calculate the misclassification rate on the test set. We repeated this step to calculate the misclassification rates at different numbers of contaminants (1 C/N, 5 C/N, 10 C/N) and different training data sizes (245, 490, 2450 observations).

Acknowledgments

This project is funded by ARO Grant W911NF-12-1-0390. We would like to thank Segré lab members at Boston University for useful discussions on flux balance analysis. One of the authors (VS) would like to thank Wilke lab members for critical reading of the manuscript. We would also like to thank BCG and TACC at UT for computing resources.

Author Contributions

Conceived and designed the experiments: V.S. and C.O.W. Performed the experiments: V.S. Analyzed the data: V.S, A.G.M and C.O.W. Wrote the paper: V.S, A.G.M, P.R, J.E.B, P.R, D.S. and C.O.W.

References

1. Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? *Nat Biotechnol* 28: 245-8.
2. Edwards JS, Palsson BO (2000) The escherichia coli mg1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci U S A* 97: 5528-33.
3. Reed JL, Vo TD, Schilling CH, Palsson BO (2003) An expanded genome-scale model of escherichia coli k-12 (ijr904 gsm/gpr). *Genome Biol* 4: R54.
4. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, et al. (2007) A genome-scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for 1260 orfs and thermodynamic information. *Mol Syst Biol* 3: 121.

5. Orth JD, Palsson B (2012) Gap-filling analysis of the ijo1366 escherichia coli metabolic network reconstruction for discovery of metabolic functions. *BMC Syst Biol* 6: 30.
6. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B-Methodological* 58: 267-288.
7. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33: 1-22.
8. Schellenberger J, Park JO, Conrad TM, Palsson BO (2010) Bigg: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *Bmc Bioinformatics* 11.
9. Canonaco F, Hess TA, Heri S, Wang T, Szyperski T, et al. (2001) Metabolic flux response to phosphoglucose isomerase knock-out in escherichia coli and impact of overexpression of the soluble transhydrogenase udha. *FEMS Microbiol Lett* 204: 247-52.
10. Menendez P, Kourmpetis YA, ter Braak CJ, van Eeuwijk FA (2010) Gene regulatory networks from multifactorial perturbations using graphical lasso: application to the dream4 challenge. *PLoS One* 5: e14147.
11. Wu TT, Chen YF, Hastie T, Sobel E, Lange K (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* 25: 714-21.
12. Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9: 432-441.
13. Ravikumar P, Wainwright MJ, Lafferty JD (2010) High-dimensional ising model selection using $l(1)$ -regularized logistic regression. *Annals of Statistics* 38: 1287-1319.

14. Schellenberger J, Que R, Fleming RM, Thiele I, Orth JD, et al. (2011) Quantitative prediction of cellular metabolism with constraint-based models: the cobra toolbox v2.0. Nat Protoc 6: 1290-307.

Tables

Table 1: **Carbon and nitrogen sources used in the study.**

Carbon Sources	Nitrogen Sources
D-glucose	Ammonia
Pyruvate	Adenine
Glycerol	Cytidine
Acetate	Putrescine
D-ribose	L-glycine
D-fructose	L-alanine
D-sorbitol	L-glutamine

7 carbon and 7 nitrogen sources are used in this study.

Table 2: **Growth media for K-12 MG1655.**

Other media	Composition
AB medium	D-glucose/Ammonia
ATCC medium 57	Glycerol/L-lysine/Ammonia
Bochner medium	Pyruvate/Ammonia
Davis Mingioli medium	D-glucose/Citrate/Ammonia

4 commonly used growth media for E. coli MG1655 strain that were picked from EcoCyc website were also used in this study.

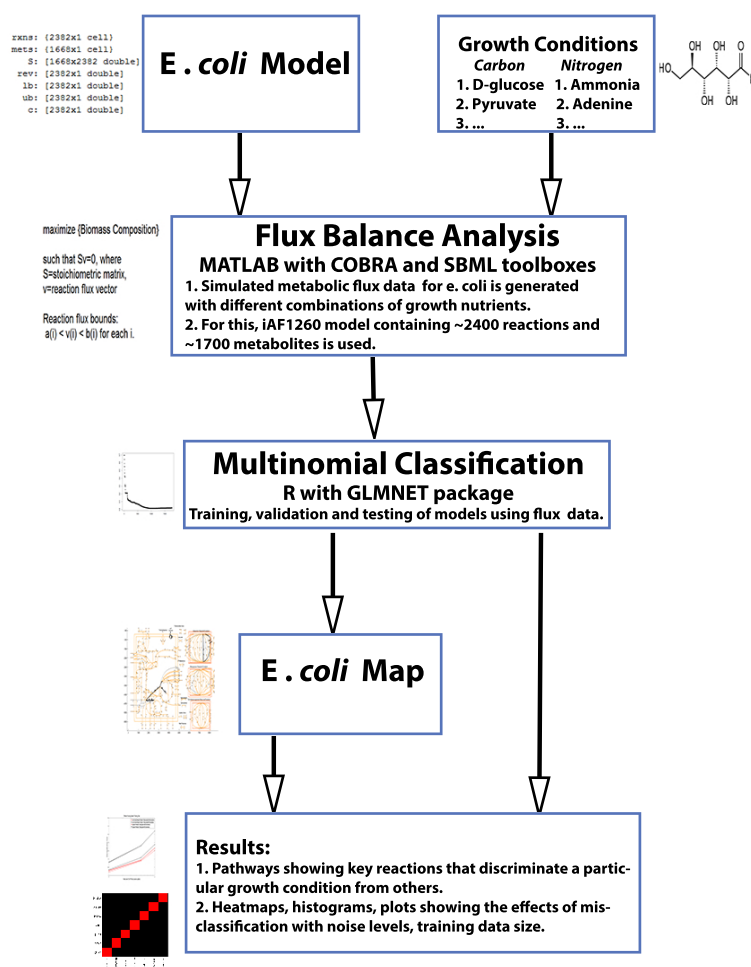


Figure 1: **Flowchart** describing methodology used in this study. We obtained E. coli model and map from BiGG database. The key steps involved are Flux Balance Analysis and multinomial classification routines.

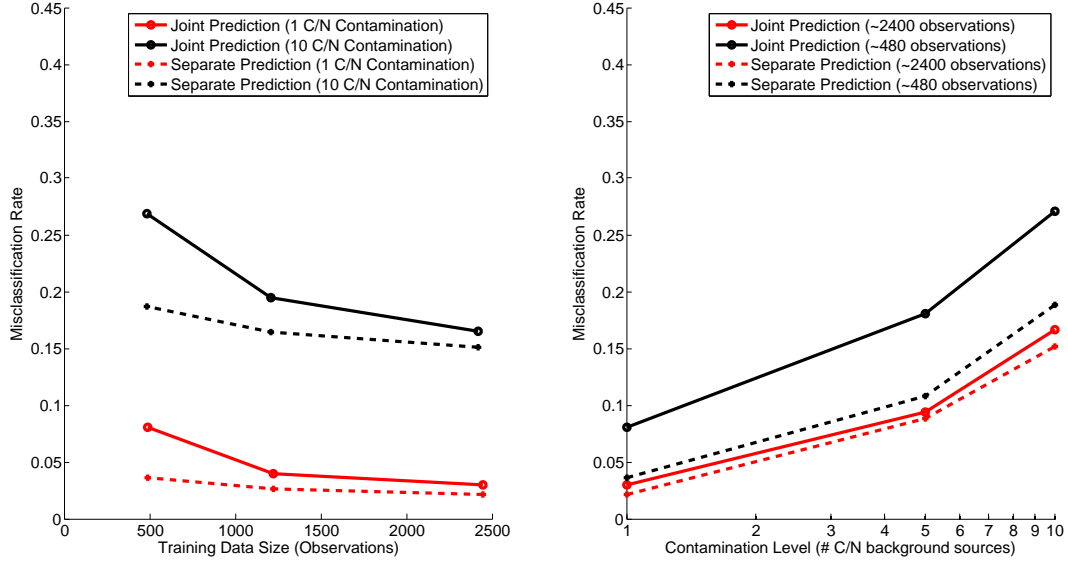


Figure 2: **Misclassification rate versus number of contaminants and amount of training data.** (A) The misclassification rate increases as the number of contaminants increases. (B) The misclassification rate decreases as the size of the available training data decreases. In all cases, separate prediction out-performs joint prediction.

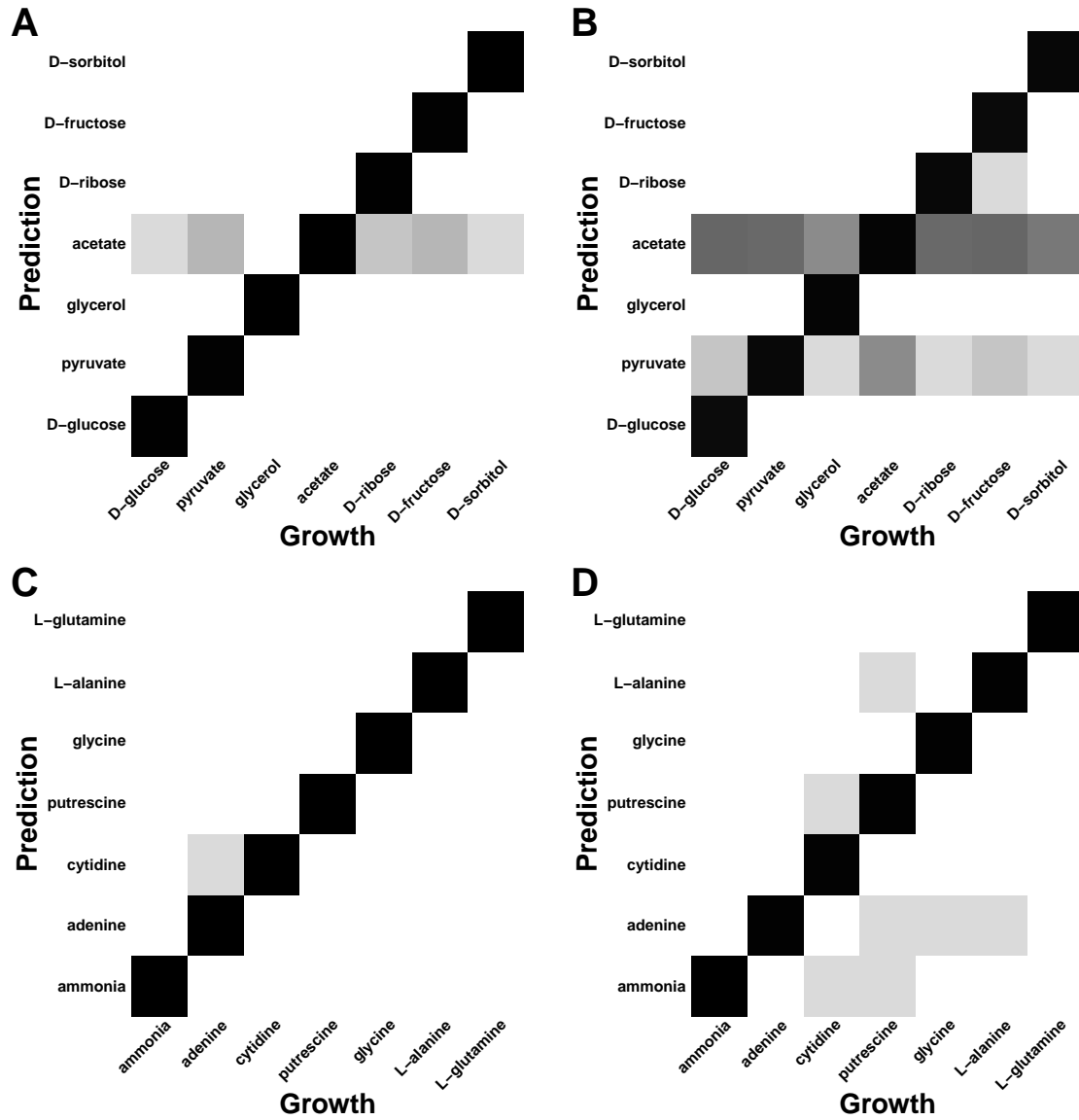


Figure 3: **Heat maps with actual sources as columns and predicted ones in rows.** At 10% noise, most of the misclassified C sources are predicted wrongly to be acetate and few gets wrongly predicted as pyruvate.

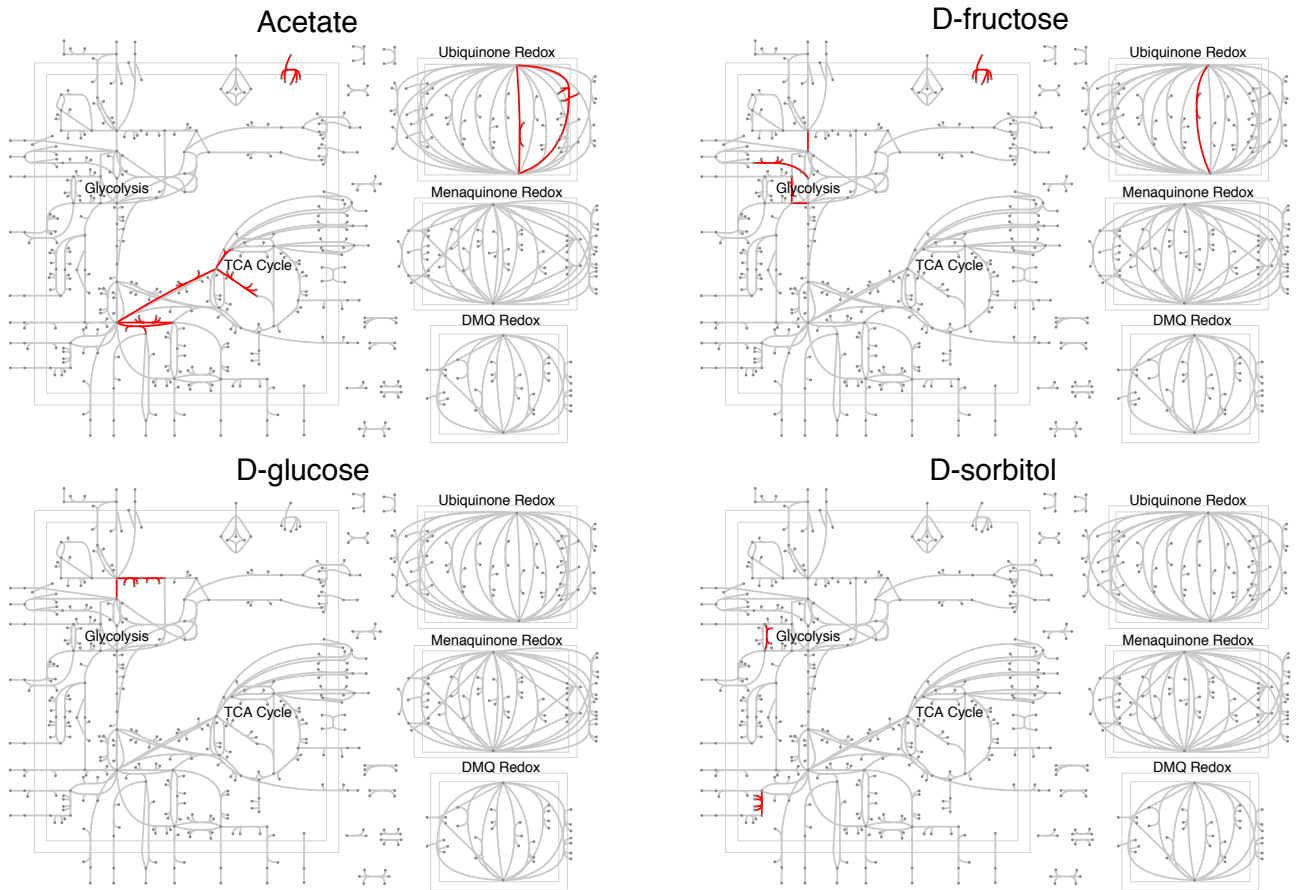


Figure 4: **Discriminatory carbon sources** The key-reactions identified by GLMNET package were mapped onto *E. coli* central metabolism to visually show the differences between different growth conditions. Out of 7 carbon sources, here we show 4 carbon sources and the key-reactions.

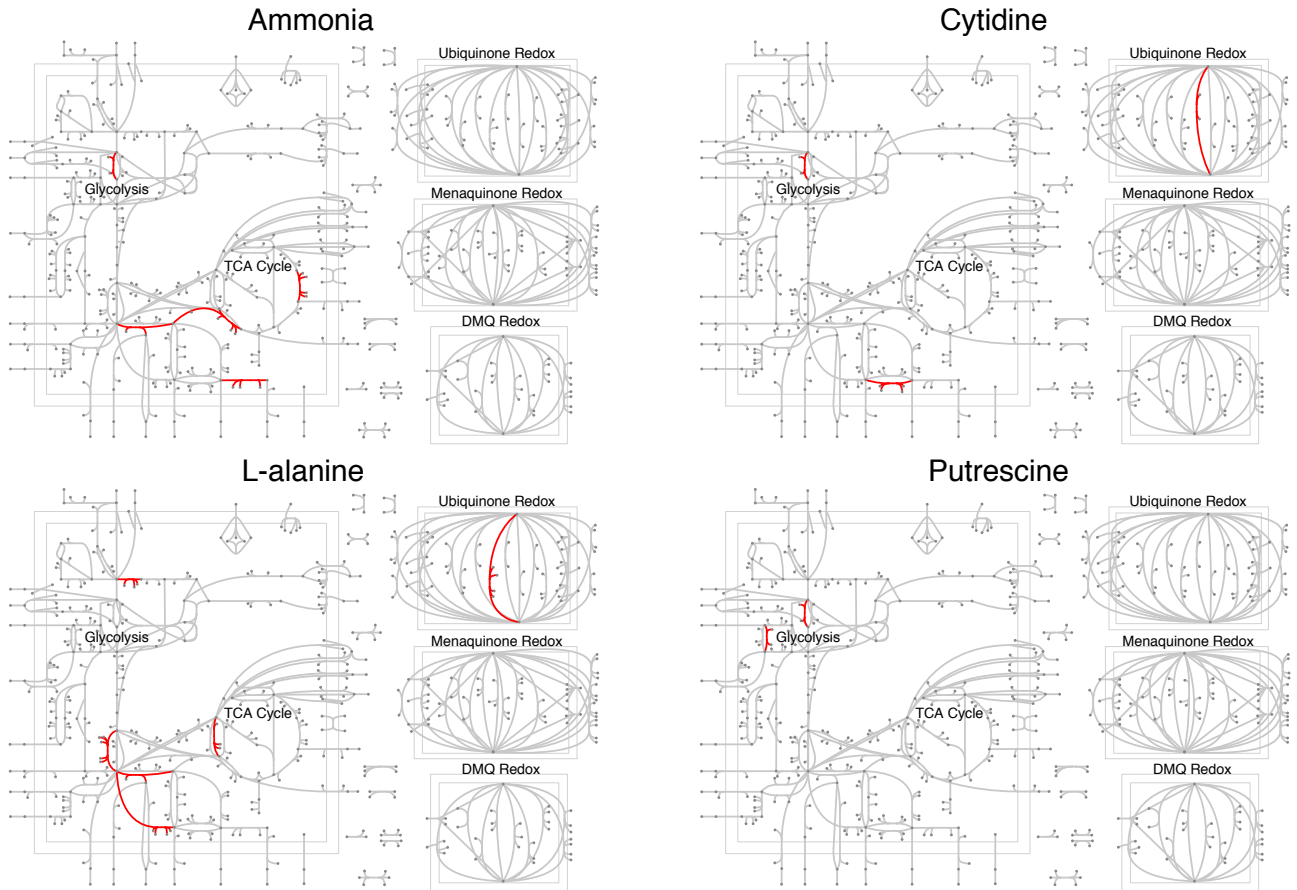
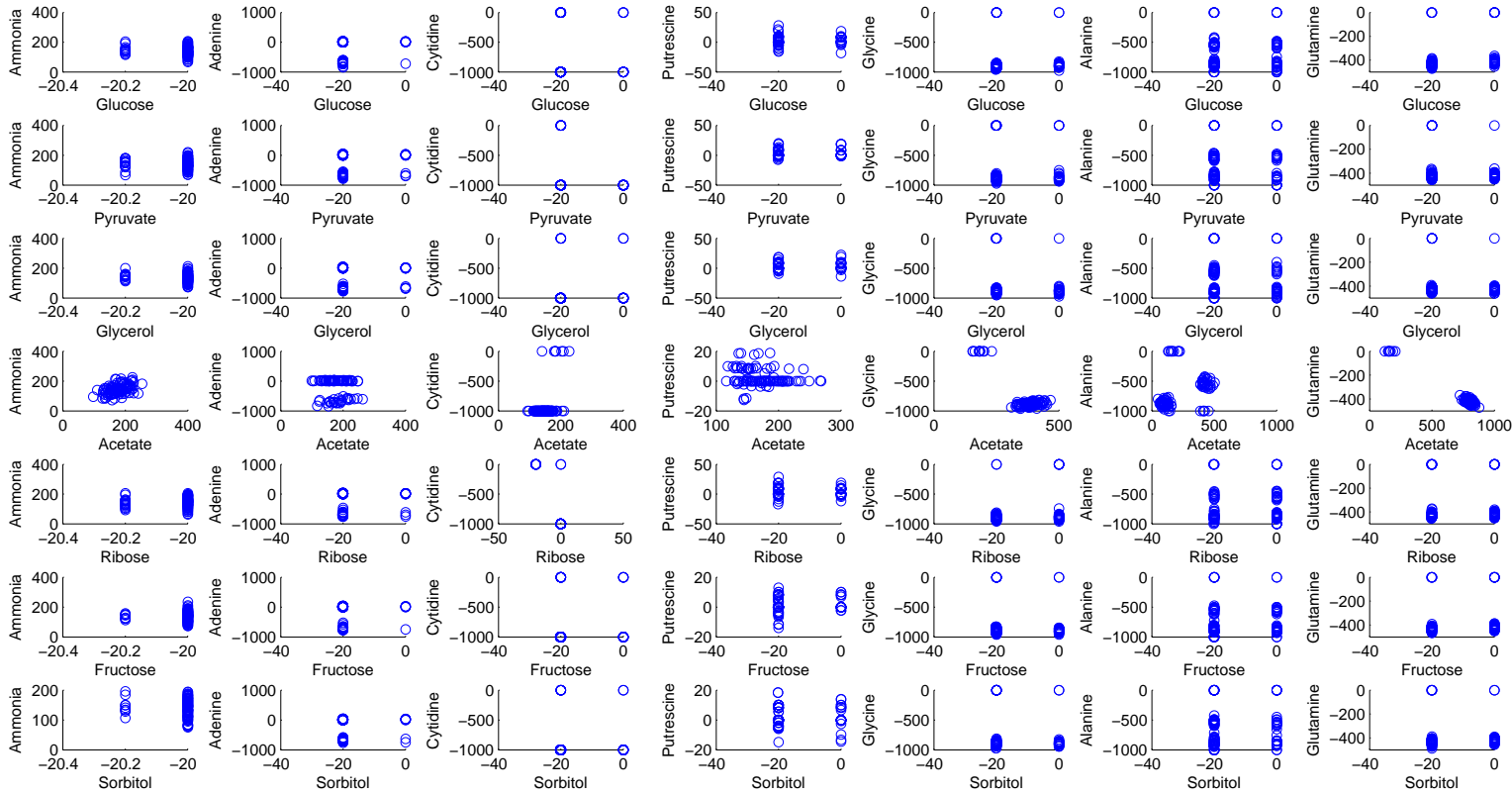
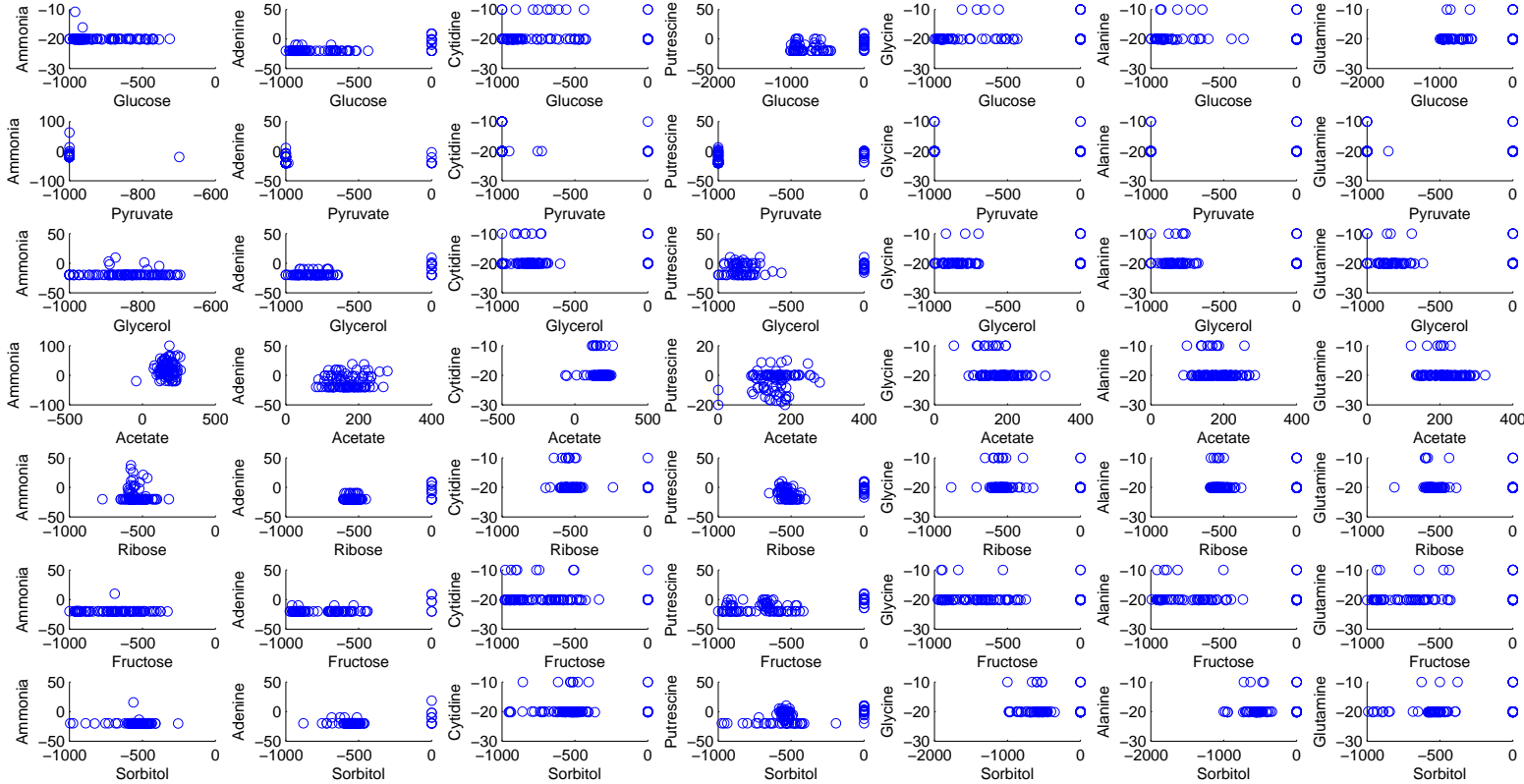


Figure 5: **Discriminatory nitrogen sources** The key-reactions identified by GLMNET package were mapped onto *E. coli* central metabolism to visually show the differences between different growth conditions. Here, the growth medium used are generally used for K-12 MG1655 strain.



Supplementary Figure 1: **Scatter plot showing varying uptake amounts of C/N sources when carbon source is limited.** We increased the upper bounds of the nitrogen sources and plotted the uptake amounts of carbon and nitrogen sources.



Supplementary Figure 2: **Scatter plot showing varying uptake amounts of C/N sources when nitrogen source is limited.** We increased the upper bounds of the carbon sources and plotted the uptake amounts of carbon and nitrogen sources.

Supplementary Table 1: **Carbon sources and key reactions identified to discriminate these sources.**

'Carbon Source is D-Glucose'

'glucose 6-phosphate dehydrogenase'

'glucose-6-phosphate isomerase'

'6-phosphogluconolactonase'

'Carbon Source is Pyruvate'

'4-aminobutyrate transaminase'

'acyl-CoA dehydrogenase (decanoyl-CoA)'

'acyl-CoA dehydrogenase (dodecanoyl-CoA)'

'L-aspartase'

'aspartate transaminase'

'citrate synthase'

'Cysteine Desulfhydrase'

'cytochrome oxidase bo3 (ubiquinol-8: 4 protons) (periplasmic)'

'3-hydroxyacyl-CoA dehydratase (3-hydroxydecanoyl-CoA)'

'3-hydroxyacyl-CoA dehydratase (3-hydroxydodecanoyl-CoA)'

'fructose-bisphosphatase'

'glucose 6-phosphate dehydrogenase'

'glutamate dehydrogenase (NADP)'

'Glycolate oxidase'

'3-hydroxyacyl-CoA dehydrogenase (3-oxodecanoyl-CoA)'

'3-hydroxyacyl-CoA dehydrogenase (3-oxododecanoyl-CoA)'

'hypoxanthine dehydrogenase'

'3-ketoacyl-CoA thiolase'

'3-ketoacyl-CoA thiolase'

'pyruvate dehydrogenase'

'6-phosphogluconolactonase'

'phosphoenolpyruvate synthase'

'triose-phosphate isomerase'

'Carbon Source is Glycerol'

'glycerol-3-phosphate dehydrogenase (NADP)'

'Glycerol dehydrogenase'

'glycerol kinase'

'Carbon Source is Acetate'

' 2',3'-cyclic-nucleotide phosphodiesterase (UMP) (periplasm)'

'acyl-[acyl-carrier-protein] synthetase (n-C14:1)'

'ATP synthase (four protons for one ATP) (periplasm)'

'Carbamate kinase'

'CTP synthase (glutamine)'

'cytochrome oxidase bo3 (ubiquinol-8: 4 protons) (periplasm)'

'cytidine deaminase'

'fatty-acid-CoA thioesterase (tetradecenoate)'

'fumarase'

'glycerate kinase'

'Glycolate oxidase'

'malate synthase'

'malic enzyme (NAD)'

'pyruvate dehydrogenase'

'pyruvate formate lyase'

'pyrimidine-nucleoside phosphorylase (uracil)'

'L-serine via sodium symport (periplasm)'

'succinate-semialdehyde dehydrogenase (NAD)'

'Carbon Source is D-ribose'

'acyl-CoA dehydrogenase (butanoyl-CoA)'

'acyl-CoA dehydrogenase (hexanoyl-CoA)'

'acyl-CoA dehydrogenase (octanoyl-CoA)'

' 3-hydroxyacyl-CoA dehydratase (3-hydroxybutanoyl-CoA)'

' 3-hydroxyacyl-CoA dehydratase (3-hydroxyhexanoyl-CoA)'

' 3-hydroxyacyl-CoA dehydratase (3-hydroxyoctanoyl-CoA)'

' 3-hydroxyacyl-CoA dehydrogenase (acetoacetyl-CoA)'

' 3-hydroxyacyl-CoA dehydrogenase (3-oxohexanoyl-CoA)'

'ribokinase'

'Carbon Source is D-fructose'

'acetylglutamate kinase'

'N-acetylglutamate synthase'

'N-acetyl-g-glutamyl-phosphate reductase'

'ATP synthase (four protons for one ATP) (periplasm)'

'Carbamate kinase'

'CTP synthase (glutamine)'

'Dihydroxyacetone phosphotransferase'

'fructose-1-phosphate kinase'

'glycerol-3-phosphate dehydrogenase (ubiquinone-8)'

'glycine C-acetyltransferase'

'glucose-6-phosphate isomerase'

'triose-phosphate isomerase'

'Carbon Source is D-sorbitol'

'glycerol-3-phosphatase'

'lactaldehyde dehydrogenase'

'L-Lactate dehydrogenase (ubiquinone)'

'sorbitol-6-phosphate dehydrogenase'

Supplementary Table 2: **Nitrogen sources and key reactions identified to discriminate these sources.**

'Nitrogen Source is Ammonia'	'Nitrogen Source is Adenine'
'adenine phosphoribosyltransferase'	'adenine deaminase'
'2-Oxoglutarate dehydrogenase'	'adenine phosphoribosyltransferase'
'alcohol dehydrogenase (ethanol)'	'adenylsuccinate lyase'
'L-aspartase'	'adenylosuccinate synthase'
'aspartate transaminase'	'L-aspartase'
'L-allo-threonine dehydrogenase'	'formyltetrahydrofolate deformylase'
'citrate synthase'	'glutamate dehydrogenase (NADP)'
'fructose-bisphosphatase'	'hypoxanthine dehydrogenase'
'glutamate dehydrogenase (NADP)'	'5'-nucleotidase (dAMP)'
'hydroxypyruvate isomerase'	'purine-nucleoside phosphorylase (Adenosine)'
'Hydroxypyruvate reductase (NADPH)'	'purine-nucleoside phosphorylase (Deoxyadenosine)'
'sodium proton antiporter (H:NA is 1.5) (periplasm)'	
'pyruvate formate lyase'	

'Nitrogen Source is Cytidine'

'adenosine kinase'

'aldehyde dehydrogenase (acetaldehyde, NADP)'
'AMP nucleosidase'
'aspartate carbamoyltransferase'
'L-allo-threonine dehydrogenase'
'CTP synthase (glutamine)'
'Cysteine Desulfhydrase'
'cytidine deaminase'
'fructose-bisphosphatase'
'formate dehydrogenase (quinone-8) (periplasm)'
'glycerol-3-phosphate dehydrogenase (ubiquinone-8)'
'hydroxypyruvate isomerase'
'Hydroxypyruvate reductase (NADPH)'
'sodium proton antiporter (H:NA is 1.5) (periplasm)'
' 5'-nucleotidase (AMP)'
' 2-Octaprenylphenol hydroxylase (anaerobic)'
'pyrimidine-nucleoside phosphorylase (uracil)'
'tartronate semialdehyde reductase'

'Nitrogen Source is Putrescine'

' 4-aminobutyrate transaminase'

'Aminobutyraldehyde dehydrogenase'
'fructose-bisphosphatase'
'glycerol-3-phosphatase'
'sodium proton antiporter (H:NA is 1.5) (periplasm)'
'Putrescine Transaminase'

'Nitrogen Source is Glycine'

'glycine hydroxymethyltransferase, reversible'

'Glycine Cleavage System'

'Nitrogen Source is L-Alanine'

'5'-deoxyadenosine nucleosidase'

'acetylglutamate kinase'

'N-acetylglutamate synthase'

'aspartate transaminase'

'glucose 6-phosphate dehydrogenase'

'glutamine synthetase'

'glutamate dehydrogenase (NADP)'

'malate dehydrogenase'

'pyruvate formate lyase'

'pyruvate oxidase'

'phosphoenolpyruvate synthase'

'Nitrogen Source is L-Glutamine'

'N-Acetyl-D-glucosamine(anhydrous)N-Acetylmuramyl-tripeptide
amidase'

'L-allo-threonine dehydrogenase'

'Cysteine Desulfhydrase'

'glucosamine-1-phosphate N-acetyltransferase'

'glutaminase'

'glycine C-acetyltransferase'

'murein polymerizing transglycosylase'

'phospho-N-acetylmuramoyl-pentapeptide-transferase (meso-2,6-
diaminopimelate)'

'phosphoglucosamine mutase'

'L-allo-Threonine Aldolase'

'UDP-N-acetylglucosamine diphosphorylase'