Predicting bacterial growth conditions from bacterial physiology

Viswanadham Sridhara[1], Austin G. Meyer[1,3], Piyush Rai[5], Jeffrey E. Barrick[2,3,4], Pradeep Ravikumar[5], Daniel Segrè[6], Claus O. Wilke[1,2,3,7,*]

[1]Center for Computational Biology and Bioinformatics, The University of Texas at Austin, Austin, TX, USA
[2]Center for Systems and Synthetic Biology, The University of Texas at Austin, Austin, TX, USA
[3]Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin, TX, USA
[4]Department of Molecular Biosciences, The University of Texas at Austin, Austin, TX, USA
[5]Department of Computer Science, The University of Texas at Austin, Austin, TX, USA
[6]Department of Biology, Boston University, Boston, MA, USA
[7]Department of Integrative Biology, The University of Texas at Austin, Austin, TX, USA

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

Abstract

Bacterial physiology reflects the environmental conditions a bacterium grows on. A widely studied problem in systems biology is to predict bacterial phenotype from growth conditions, using mechanistic models such as flux-balance analysis (FBA). However, the inverse prediction of growth conditions from phenotype is rarely considered. Here we develop a computational framework to carry out this inverse prediction on a computational model of bacterial physiology. We use FBA to calculate bacterial phenotypes from growth conditions in *E. coli*, and then we assess how accurately we can predict the original growth conditions from the phenotypes. Prediction is carried out via regularized multinomial regression. Our analysis provides several important physiological and statistical insights. First, we show that by analyzing metabolic end products we can consistently predict growth conditions. Second, prediction is reliable even in the presence of small amounts of contaminants. Third, a relatively small number of reactions per growth source (~ 10) is sufficient for accurate prediction. Fourth, combining the predictions from two separate models, one trained only on carbon sources and one only on nitrogen sources, performs better than models trained to perform joint prediction. Finally, that separate prediction performs better than a more sophisticated joint prediction scheme suggests that carbon and nitrogen metabolisms are largely decoupled in bacterial physiology.

Introduction

Research into metabolism and physiology generally tries to uncover how an organism's physiology is determined by the environment to which the organism is exposed. For example, one might ask which genes are up or downregulated as microbes are grown on different nutrient sources [1–3]. Similarly, one might ask how changes in nutrient availability and gene expression alter flux of metabolites through the organism's metabolic network [4–7]. One can also pose the inverse question, however: If we know an organism's physiology, can we infer the environment in which the organism is living or was grown? Is physiology predictive of the current environment? On the one hand, one could envision a scenario where an organism can only assume a small number of distinct physiological states, and many diverse environments elicit the same physiological response. In other words, the mapping from environment to physiology is many-to-one. Under this scenario, physiology would not be particularly predictive of environment. On the other hand, each environment might elicit an entirely different physiological response, i.e., the mapping from environment to physiology is one-to-one. Under this scenario, organismal physiology can be considered an accurate reflection of the specific environment the organism resides in, and the environment can be predicted accurately from the physiology. In reality, we can expect the mapping between environment and physiology to fall somewhere between these two extremes. While there are probably many different physiological states an organism can assume, there will also be distinct environments that create a similar physiological response.

Modeling approaches to metabolism generally ask the forward question, i.e., how can we calculate the metabolic state as a function of the environment. For example, flux-balance approaches calculate the metabolic fluxes in an organism as a function of input fluxes and the organism's metabolic network [8–11]. More sophisticated approaches might use kinetic or dynamic models [12, 13]. All these modeling approaches are mechanistic approachs that mimmick the chain of causality in physiology: environment and genetic architecture are given, and physiological state follows from the laws of physics and biochemistry. To ask the inverse question, which environment corresponds to which physiology, we have to go against the chain of causality. Therefore, a mechanistic model is not the most appropriate to ask this question. Instead, it makes more sense to use a statistical approach to search for associations between physiological states and environmental conditions.

Here, we asked how predictive physiology is of environment for a simulation model of the bacterium *E. coli*. We first calculated physiological states for a variety of environmental growth conditions, using flux-balance analysis as our model for bacterial physiology. We then developed statistical models to predict the particular growth conditions for a particular solution of the flux-balance equation. We found that prediction is possible with surprisingly low error rates, even if a moderate amount of contamination is present in the simulated growth media. We further found that carbon and nitrogen metabolisms seem

to be largely decoupled (prediction accuracy of a given carbon source does not strongly depend on the presence of particular nitrogen sources and vice versa) and that most carbon and nitrogen sources can be indentified reliably from a small number of predictive fluxes.

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.

A biochemical network can be treated as something that takes up the high energy nutrients from the environment and converts them into useful metabolic precursors such as amino acids, nucleotides etc. These environmental nutrients 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 [14]), this amounted to 939 reactions among a total of 2382 distinct reactions. We also discarded the biomass composition reaction, and thus were left with a 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 quantitiy 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 [9]. A different set of random contaminants was chosen for each individual FBA

calculation. We tuned the amount of contamination in our simulations by changing the number of contaminants. Throughout this work, we refer to a contamination level of n carbon-source and n nitrogen-source contaminants as n C/N. (We always chose an equal number of nitrogen and carbon sources as contaminants.)

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 simulated fluxes for environments containing all pairwise combinations of the seven carbon and seven nitrogen sources, plus contaminants. For each contamination level and training data-set size, we generated 100 replicates of each pairwise combination, for a total of 4900 sets of flux values (see Table 2 for an overview of all simulations performed). We discarded solutions that we considered to be non-viable (see Methods). 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. 1 C/N, 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 and the predicted sources at two different 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 or nitrogen 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 outperfomed 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. Above, we used the conventional maximum uptake rate of 20 mmol gDW^{-1} hr^{-1} that is generally used for carbon and nitrogen sources in FBA studies. To determine to what extent our results depended on this choice, we artificially increased the uptake rates of the carbon source to a maximum of 1000 mmol gDW^{-1} hr^{-1} while keeping the nitrogen source at the normal rate, and vice versa. These simulations can be considered nitrogen limited (when carbon is artificially increased) or carbon limited (when nitrogen is artificially increased). Figures 6 and 7 show the realized uptake rates for these simulations, demonstrating that indeed nitrogen and carbon are limiting in most of these scenarios. (Realized nitrogen uptake corresponds to the maximum possible uptake while carbon does not or vice versa.) When predicting growth conditions from the final fluxes, we obtained similar results to before, i.e., individual prediction performed better than joint prediction. For an artificially high uptake rate for nitrogen but with a normal uptake rate for carbon sources, the misclassification rate with separate prediction was 10.31%, while the misclassification rate with joint prediction is 26.24%, at a contamination level of 1 C/N and with a training data size of ~ 2400 replicates. Separately predicting carbon resulted in 151 mispredictions compared to 109 mispredictions for nitrogen. In combination, there were 250 mispredictions using separately trained models. Joint prediction resulted in 638 mispredictions. Similarly, for artificially high uptake rates for carbon sources and normal uptake rates for nitrogen sources, the misclassification rate under separate prediction was 3.86% while the misclassification rate under joint prediction was 13.82%. Joint prediction resulted in 324 mispredictions. Separate prediction resulted in a combined misprediction from C and N sources of 94 mispredictions (64 C and 32 N, respectively). Clearly prediction rates were better for separate prediction compared to joint prediction even with larger training data sizes, which was not the case for *normal* uptake rates for both sources. It seems that as one source becomes limiting, separate prediction increasingly outperforms joint prediction.

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 pairwise combinations of carbon/nitrogen sources for two replicates. We used one replicate (78 carbon observations and 174 nitrogen observations respectively for each of carbon

and nitrogen sources) 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.

To gain mechanistic insight into predictive reactions, we mapped them onto the *E. coli* central metabolism (Figures 4 and 5, Tables 4 and 5). Note that each of the metabolic maps is meant to highlight only the central carbon metabolism in the *E. coli* metabolic network. 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 isolated 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), required enzymes in the glycolytic and pentose phosphate pathways [15]. Similarly, sorbitol (the singly reduced alcohol of D-glucose) and fructose each possessed predictive reactions in the relative vicinity of the glycolytic pathway (Figure 4). Mapping nitrogen sources to the central metabolism revealed a similar trend. For example, L-alanine as a growth source had predictive reactions near its site of entry into the three and four carbon metabolism of the TCA cycle (Figure 5).

At a contaminant level of 1 C/N 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 were 144 key predictive reactions, or on average 10.3 predictive reactions per carbon or nitrogen source. Among the 144 predictive reactions, we found 100 unique reaction IDs. Lists of the predicitve reactions for each growth source are provided in Tables 4 and 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 the lowest contamination level and with the largest training data set analyzed. We eliminated each of these 104 reactions at a time as predictors in the regression model, 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 the reaction "glucose 6-phosphate isomerase" (PGI), the misclassification rate remained unchanged when we eliminated any of the other 103 reactions before model fitting. PGI catalyzes a reaction that produces fructose-6-P from glucose-6-P, and knock-out of the PGI gene causes diminished growth rate [16]. The reaction PGI catalyzes is reversible, but the forward reaction occurs most of the time unless concentrations of fructose-6-P are very high. Thus, PGI seems to be critical to distinguish glucose from fructose as growth source. However, more generally, the specific set of reactions used for successful prediction was not unique.

Predicting specific media or novel metabolites

To generalize our simulations to more experimentally relevant test conditions, we performed similar analyses for several media that are 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 3 shows the composition used in the analysis for these growth media. We were able to classify these at higher accuracy even at higher contamination levels and lower number of replicates. This may be due to the small (4) number of growth conditions used in this comparison. Our misclassification rate was consistently 0 for all cases of at least 10 replicates and up to 20 C/N contaminants. Only for fewer replicates did we see misclassification rates on the order of 10–20%. (Full results are provided as part of the accompanying complete data sets and scripts.)

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 contamination levels of 1 C/N and 20 C/N. This resulted in 700 observations and after using a threshold, there were 695 viable flux measurements at 1 C/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. 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 (231 cases) were predicted to be sorbitol/putrescine. Sorbitol is a reasonable choice considering the model had never seen maltose. (Sorbitol is the singly reduced alcohol of glucose.) However, of the 231 cases, only 98 had actually been grown on putrescine, so the joint model's prediction of the nitrogen source was misled by the unknown carbon source.

For a contamination of 20 C/N, 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 founds in DNA and RNA. We used 2 contamination levels, 1 C/N and 20 C/N, and we generated 100 replicates for each contamination level for testing. At 1 C/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 ($\sim$ 0.01% misclassification). Interestingly, in all cases, nitrogen source cytosine was predicted as ammonia. This result may be due to the amine group in position 4 of cytosine.

At a contamination level of 20 C/N, all the 700 flux measurements were viable (biomass greather than threshold). In this case, 27 carbon sources were wrongly predicted ($\sim$ 0.04% misclassification rate). The nitrogen source cytosine is predicted as ammonia in 78.8% of cases and as adenine in all other cases.

Discussion

We have developed a method for making predictions regarding bacterial growth conditions from known simulated metabolic fluxes. We generated fluxes using the complete *E. coli* metabolic network model for 7 carbon and 7 nitrogen sources. Then, we divided the data into training and test sets and employed machine learning with a generalized linear framework to train a model to predict growth conditions. We found that even at high contamination levels, we could make reliable predictions regarding growth media for all of the sources we tested. Of note, our data indicates separately predicting carbon and nitrogen sources performs 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 metabolic 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. Also, extending our prediction algorithm to more experimentally relevant growth media, our scheme gave comparable accuracy. Thus, we have shown that predicting growth conditions from simulated metabolic flux data is a computationally tractable problem.

Although the background contamination can have a dramatic affect on model accuracy, the misclassification rate remains acceptably low even with 10 randomly picked C/N contaminants. The addition of this contamination revealed one interesting and unexpected physiological hypothesis about *E. coli* metabolism. Namely, as the contamination increases from 1 C/N to 10 C/N, our model increasingly predicts acetate and pyruvate as the default carbon sources. Due to its centrality in terms of energy production, for any input growth source the reactions that lead to the TCA cycle or the reactions within the TCA cycle should have some reasonable amount of reaction flux. In other words, acetate and pyruvate as default nutrient sources may not be so surprising when one considers acetate's and pyruvate's central role in the TCA cycle—it is essentially the center of bacterial metabolism. To be sure that the observed default carbon source misclassification was not an artifact of nutrient limitation (carbon versus nitrogen), we increased the uptake rates of carbon source artificially high while keeping nitrogen source at normal uptake and vice versa. This ensures limiting conditions for one source and non-limiting for the other. These simulations did not alter our earlier conclusion that separate prediction performs better than joint prediction.

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) [17] or identifying SNPs in GWAS studies [18]. We would like to point out that beyond LASSO there are a number of other commonly used regularization techniques. For example, graphical Lasso [19] and Ising Markov Random Field models [20] 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 remarkbly small number of source-predicting reactions.

Throughout this work, we used basic flux-balance analysis to predict the bacterial phenotype. In future work more realistic models could be used that integrate regulatory information and/or signalling-pathway information with flux-balance analysis techniques [12, 21]. *I don't understand this sentence:* Moreover we would like to integrate the experimental data [22] to refine lower and upper bound constraints for uptake and secretion rate respectively to improve the model's predictions.

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. Ten 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 only one for this particular set of carbon-nitrogen pairwise sources. 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

Flux Balance Analysis

We carried out flux balance analysis using the COBRA toolbox [23] for MATLAB. We used the iAF1260 model from the BiGG database [14]. 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 mmol gDW^{-1} hr^{-1}. But for 5 reactions, i.e., ATPM, CAT, FHL, SPODM, SPODMpp, the upper bound is set to 50 mmol gDW^{-1} hr^{-1}. The lower bound for the majority of reactions (nearly 1800) is set to 0 mmol gDW^{-1} hr^{-1}, which means that the flux cannot flow in the direction opposite to that specified in the model (irreversibility constraint). A set non-growth associated maintenance (NGAM) of 8.39 mmol gDW^{-1} hr^{-1} is used for the ATPM reaction (ATP maintenance). The lower bounds of some exchange reactions is set to non-zero values, i.e., these reactions by default are meant to uptake ions, carbon and nitrogen sources, and so on. We used the default oxygen uptake rate (-18.5 mmol gDW^{-1} hr^{-1}) in the iAF1260 model, but we changed the lower bounds of glucose and ammonia to zero, expect when we used these compounds as carbon or nitrogen sources.

Need a few sentences on maximizing fluxes here. How exactly was FBA carried out? We considered all solutions with a biomass flux below a threshold of 0.558 to be inviable. The origin of this threshold value is explained in subsection "Contaminants" below.

Growth conditions

We initially carried out simulations on 49 growth conditions consisiting of all pairwise combinations of 7 carbon and 7 nitrogen sources (Table 1). The compounds were selected from among the 174 carbon and 78 nitrogen sources previously used in Feist et. al [9], and they were chosen to yield distinct flux profiles, as assessed by k-means clustering of steady-state fluxes obtained using all pair-wise combinations of the 174 carbon and

78 nitrogen sources. Of the 7 carbon sources chosen, none resulted in any growth when used as a growth condition in the absence of a nitrogen source. By contrast, all nitrogen sources except ammonia yielded growth when supplied in the absence of a carbon source.

For any given growth environment we simulated, we set the lower bound of the exchange reactions corresponding to the carbon and nitrogen sources present to -20 mmol gDW^{-1} hr^{-1}. This lower bound is commonly used in many studies [9]. We set the lower bound of all other exchange reactions for carbon and nitrogen sources to 0, except for contaminants (see next subsection for details). For each of the 49 pair-wise combinations of the 7 carbon and 7 nitrogen sources, we generated 100 replicates, each with different, randomly chosen contaminants. Apart from these 49 growth conditions, we also used 4 growth media that are generally used in *E. coli* K-12 MG1655 experiments, as cited in the EcoCyc database (http://ecocyc.org/). A complete overview of all simulation conditions used is given in Table 2.

For carbon or nitrogen limited conditions, we increased the maximum update rate of one source while keeping the other one fixed. Thus, we changed the lower bounds (uptake rate) of carbon sources to -1000 mmol gDW^{-1} hr^{-1} while keeping the lower bounds of nitrogen sources at -20 mmol gDW^{-1} hr^{-1} and vice versa.

Finally, we also carried out simulations on all pairwise combinations of all 174 carbon and 78 nitrogen sources previously used in Feist et. al [9], again with contaminants as explained below. For this analysis, we only used two replicates for each pairwise combination of carbon and nitrogen sources, one for training and one for model evaluation.

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 [9]. We used different background contamination levels, ranging from 1 C/N to 10 C/N sources. For example, if we want to set 5 C/N contamination level, we randomly picked 5 carbon and 5 nitrogen sources and set their lower bounds to -0.2 mmol gDW^{-1} hr^{-1}. Note that we generated the flux data for a pairwise combination of 1 carbon and 1 nitrogen source along with the background contamination 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 level of contamination (1 C/N), using all pairwise combinations of the 7 carbon and 7 nitrogen sources chosen for the first analysis, and with the largest training dataset size ($\sim$ 2450 replicates). We recorded the biomass values for all these simulations, and used as lower threshold of viability three standard deviations below the mean, which came out to 0.558.

Regularized regression

We predicted growth conditions using regularized multinomial logistic regression, as implemented in the GLMNET package [24] for R. Unless specified otherwise below, we used the standard settings of the GLMNET package.

Some details regarding lasso seem to be missing here. E.g., which parameters were changed from default? Also, list R functions from the GLMNET package where appropriate, e.g., which function was used for cross-validation?

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).

To guarantee that the LASSO model would converge, we imposed a threshold of 10^{-6} on all flux values. Flux values below the threshold were rounded down to zero before fitting the LASSO model.

Raw data and analysis scripts

All raw data and analysis scripts are available online in the form of a git repository at https://github.com/clauswilke/Ecoli_FBA_input_prediction.

Acknowledgments

This project is funded by ARO Grant W911NF-12-1-0390. We would like to thank Segrè lab members at Boston University for useful discussions on flux balance analysis, and members of the Wilke lab 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. Tao H, Bausch C, Richmond C, Blattner FR, Conway T (1999) Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media. J Bacteriol 181: 6425–6440.

2. Hua Q, Yang C, Oshima T, Mori H, Shimizu K (2004) Analysis of gene expression in *Escherichia coli* in response to changes of growth-limiting nutrient in chemostat cultures. Appl Environ Microbiol 70: 2354–2366.

3. Wu J, Zhang N, Hayes A, Panoutsopoulou K, Oliver SG (2004) Global analysis of nutrient control of gene expression in *Saccharomyces cerevisiae* during growth and starvation. Proc Natl Acad Sci USA 101: 3148–3153.

4. Fischer E, Sauer U (2003) Metabolic flux profiling of *Escherichia coli* mutants in central carbon metabolism using GC-MS. Eur J Biochem 270: 880–891.

5. Lee MC, Chou HH, Marx CJ (2009) Asymmetric, bimodal trade-offs during adaptation of methylobacterium to distinct growth substrates. Evolution 63: 2816–2830.

6. Boer VM, Crutchfield CA, Bradley PH, Botstein D, Rabinowitz JD (2010) Growth-limiting intracellular metabolites in yeast growing under diverse nutrient limitations. Mol Biol Cell 21: 198–211.

7. Haverkorn van Rijsewijk BR, Nanchen A, Nallet S, Kleijn RJ, Sauer U (2011) Large-scale 13C-flux analysis reveals distinct transcriptional control of respiratory and fermentative metabolism in *Escherichia coli*. Mol Syst Biol 7: 477.

8. Segrè D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci USA 99: 15112–15117.

9. 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.

10. Snitkin ES, Segrè D (2008) Optimality criteria for the prediction of metabolic fluxes in yeast mutants. Genome Informatics 20: 123–134.

11. Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? Nat Biotechnol 28: 245–248.

12. Covert MW, Xiao N, Chen TJ, Karr JR (2008) Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. Bioinformatics 24: 2044-50.

13. Adadi R, Volkmer B, Milo R, Heinemann M, Shlomi T (2012) Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. PLoS Comput Biol 8: e1002575.

14. Schellenberger J, Park JO, Conrad TM, Palsson BØ (2010) BiGG: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinformatics 11: 213.

15. Kupor SR, Fraenkel DG (1972) Glucose metabolism in 6 phosphogluconolactonase mutants of escherichia coli. J Biol Chem 247: 1904–1910.

16. 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: 247–252.

17. 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.

18. Wu TT, Chen YF, Hastie T, Sobel E, Lange K (2009) Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25: 714–721.

19. Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9: 432–441.

20. Ravikumar P, Wainwright MJ, Lafferty JD (2010) High-dimensional ising model selection using l(1)-regularized logistic regression. Annals of Statistics 38: 1287–1319.

21. Covert MW, Palsson BØ (2002) Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. J Biol Chem 277: 28058–28064.

22. Brandes A, Lun DS, Ip K, Zucker J, Colijn C, et al. (2012) Inferring carbon sources from gene expression profiles using metabolic flux models. PLoS One 7: e36947.

23. 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: 1290–1307.

24. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33: 1–22.

Tables

Table 1: **Compounds used as primary carbon and nitrogen sources.**

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

Table 2: **Summary of all simulations performed.**

Growth condition	Replicates	Total observations	Contamination	Viable observations[a]	Training data size	Test data size
7C, 7N	100	4900	1 C/N	4893	489	2447
			5 C/N	4860	486	2430
			10 C/N	4836	483	2418
			1 C/N	4893	1223	2447
			5 C/N	4860	1215	2430
			10 C/N	4836	1209	2418
			1 C/N	4893	2446	2447
			5 C/N	4860	2430	2430
			10 C/N	4836	2418	2418
Maltose, 7N	100	700	1 C/N	695	NA[b]	695
			20 C/N	699	NA[b]	699
Cytosine, 7C	100	700	1 C/N	602	NA[b]	602
			20 C/N	700	NA[b]	700
High N, normal C (7N, 7C)	100	4900	1 C/N	4865	2432	2433
High C, normal N (7C, 7N)	100	4900	1 C/N	4848	2424	2424
174C, 78N	2	27144	1 C/N	25140	12596	12544

[a] Viable observations include only those observations with a biomass value above the viability threshold of 0.558.

[b] Models were trained on the 7C, 7N data set with 1 C/N contamination, 2446 data points.

Table 3: **Composition of commonly used growth media for *E. coli*.**

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

Figures

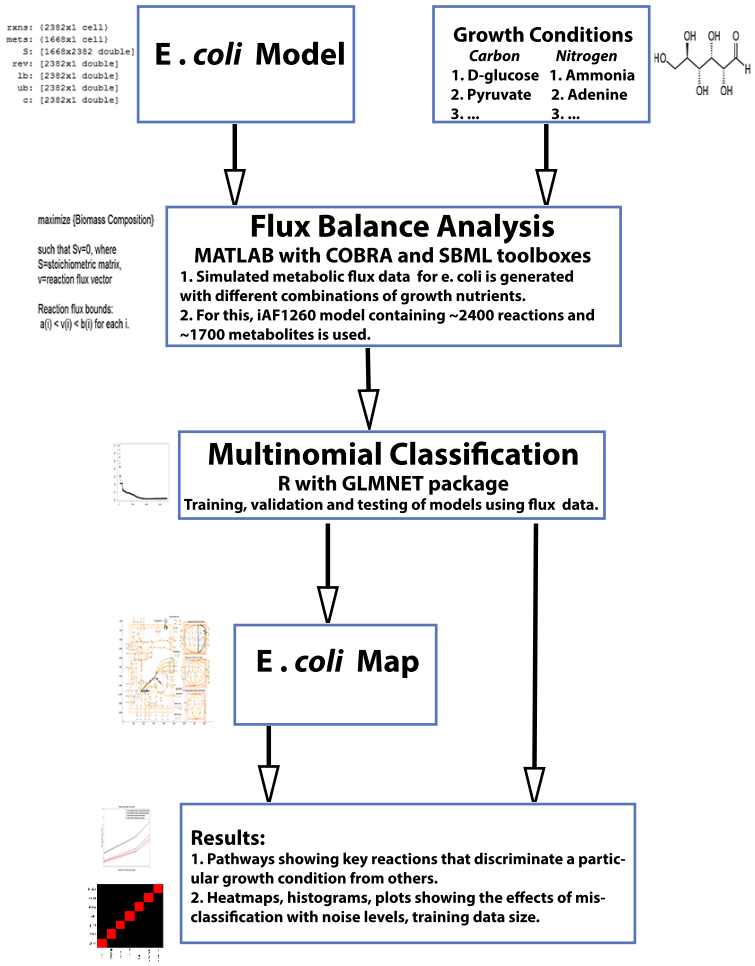

Figure 1: **Flowchart describing methodology used in this study.** Is this the latest version of the figure? It doesn't look correct.

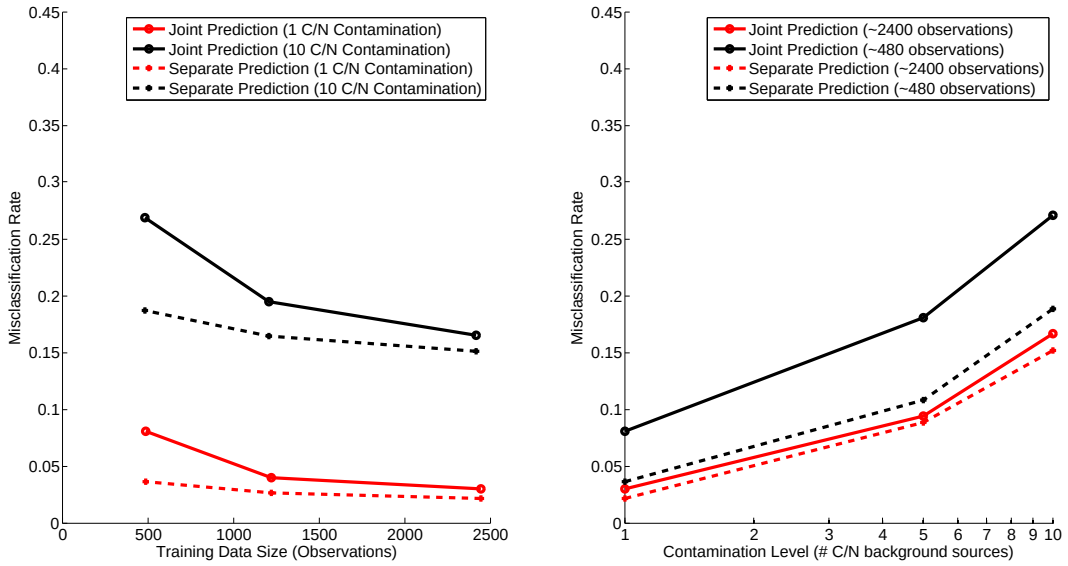

Figure 2: **Misclassification rate versus number of contaminants and amount of training data.** For joint prediction, each data point corresponds to training/testing a new regression model. Similarly for separate prediction, each data point corresponds to training/testing two separate new regression models. (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 outperforms joint prediction.

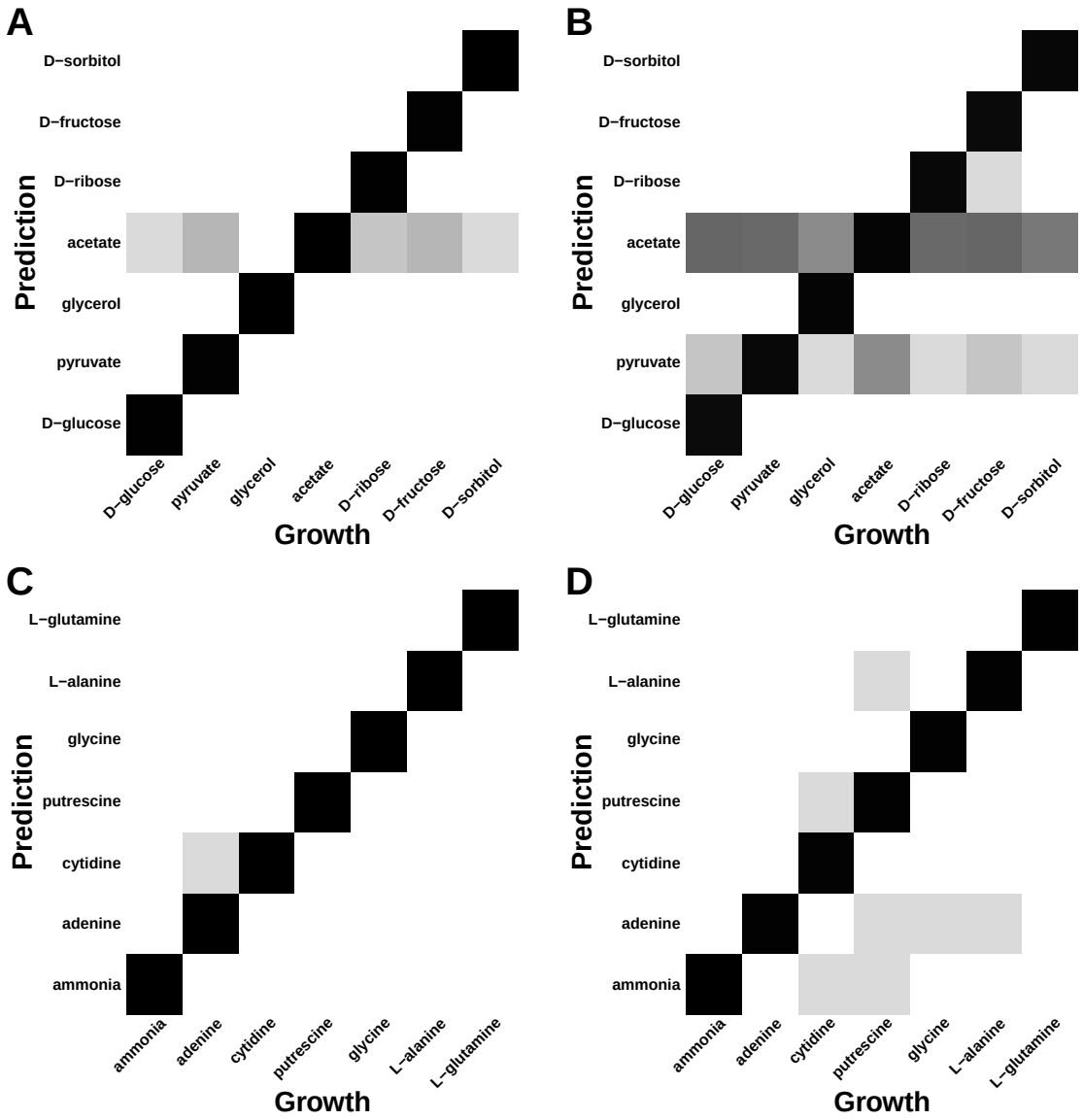

Figure 3: **Probability of misclassification of C and N sources.** For each heat map, the actual C or N source is plotted along the x-axis and the predicted one is plotted along the y-axis. The gray level of squares indicates the fraction of times a given C or N source was predicted, with white corresponding to 0% and black corresponding to 100%. (A,B) C sources predicted from models with 1 and 10 C/N contaminants, respectively. (C,D) N sources predicted from models with 1 and 10 C/N contaminants, respectively. In most cases, prediction is accurate (near-black squares along the diagonal). Prediction accuracy declines with the number of C/N contaminants, as expected. For C sources, most misclassifications lead to a prediction of acetate, or, to a lesser degree, pyruvate. A similar pattern does not exist for N sources.

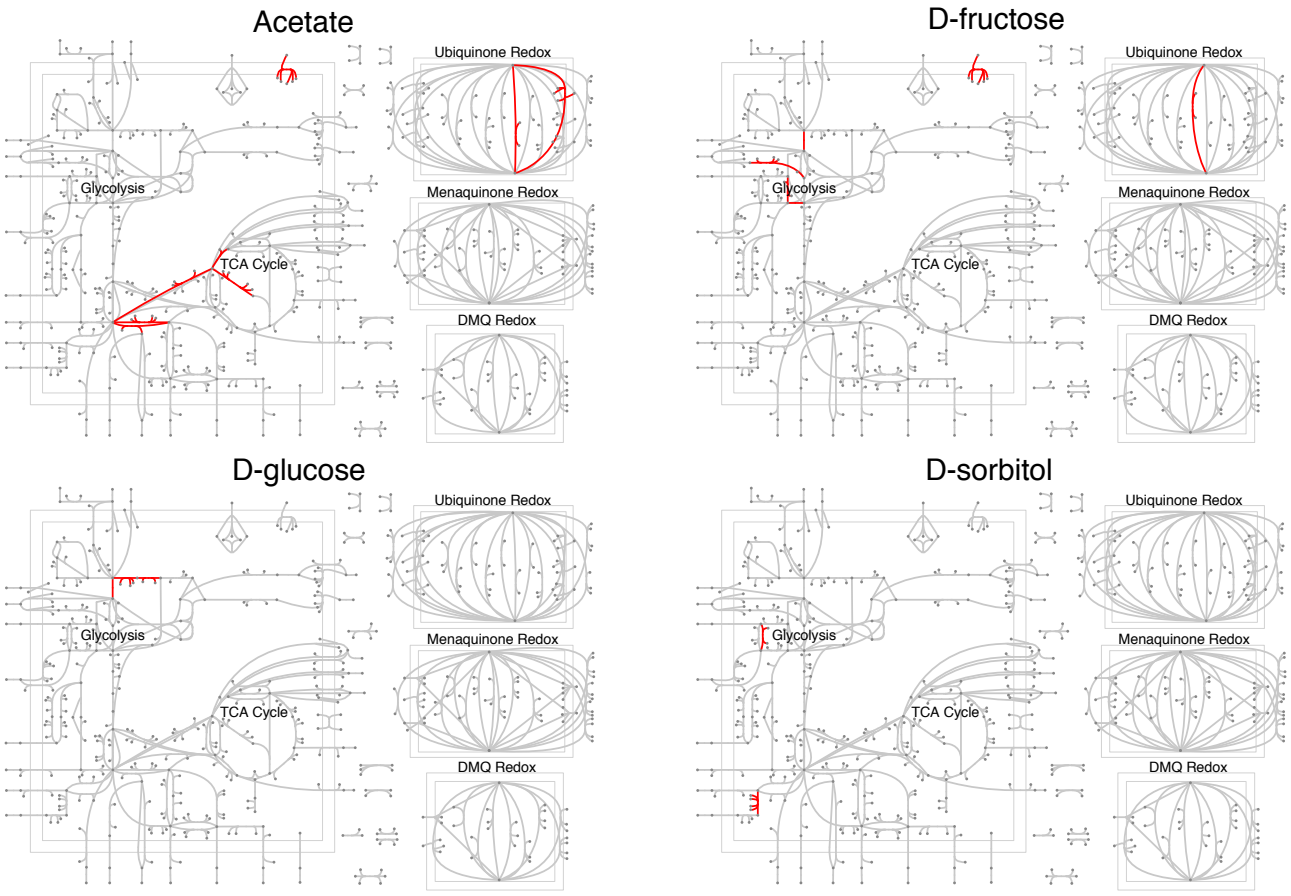

Figure 4: **Predictive reactions for four carbon sources, mapped onto the *E. coli* central metabolism.** Reactions in distinct parts of the metabolism are predictive for different carbon sources. A list of the predictive reactions can be found in Table 4.

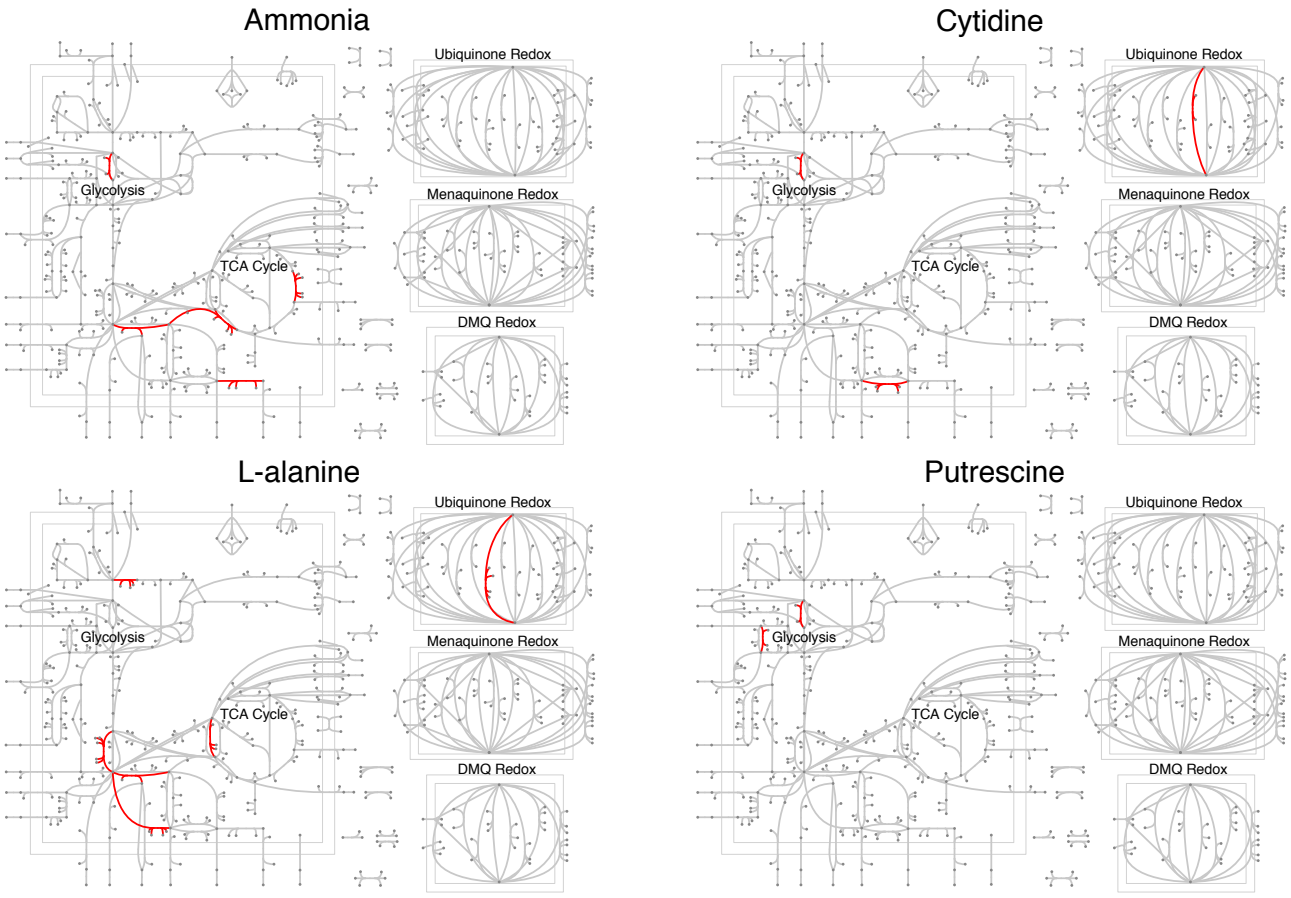

Figure 5: **Predictive reactions for four nitrogen sources, mapped onto the *E. coli* central metabolism.** Reactions in distinct parts of the metabolism are predictive for different nitrogen sources. A list of the predictive reactions can be found in Table 5.

Supplementary Tables

Table 4: **Key reactions identified to discriminate carbon sources.** These reactions were identified from a model trained to only predict carbon sources (separate prediction). The model was trained on a data set obtained with contamination of 1 C/N and with training data-set size of $\sim$ 2450 observations.

File: Figures/CarbonSources.xlsx

Table 5: **Key reactions identified to discriminate nitrogen sources.** These reactions were identified from a model trained to only predict carbon sources (separate prediction). The model was trained on a data set obtained with contamination of 1 C/N and with training data-set size of $\sim$ 2450 observations.

File: Figures/NitrogenSources.xlsx

Supplementary Figures

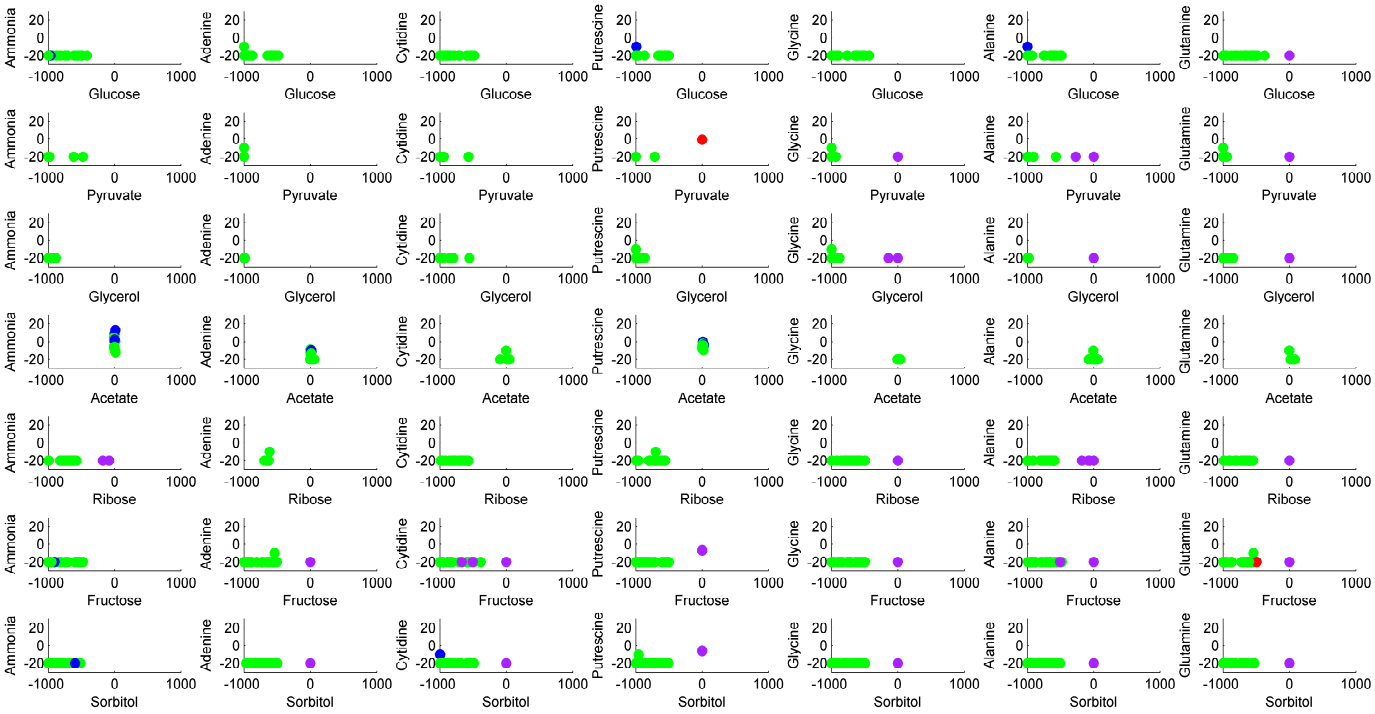

Figure 6: **Scatter plot showing varying uptake amounts of C/N sources when nitrogen is limited.** We increased the upper bounds of the carbon sources and plotted the uptake amounts of carbon and nitrogen sources. The color coding reflects whether separate prediction could predict the carbon and nitrogen sources correctly for each replicate. Green: both C and N source are correctly predicted; blue: only N source is correctly predicted; purple: only C source is correctly predicted; red: neither source is correctly predicted.

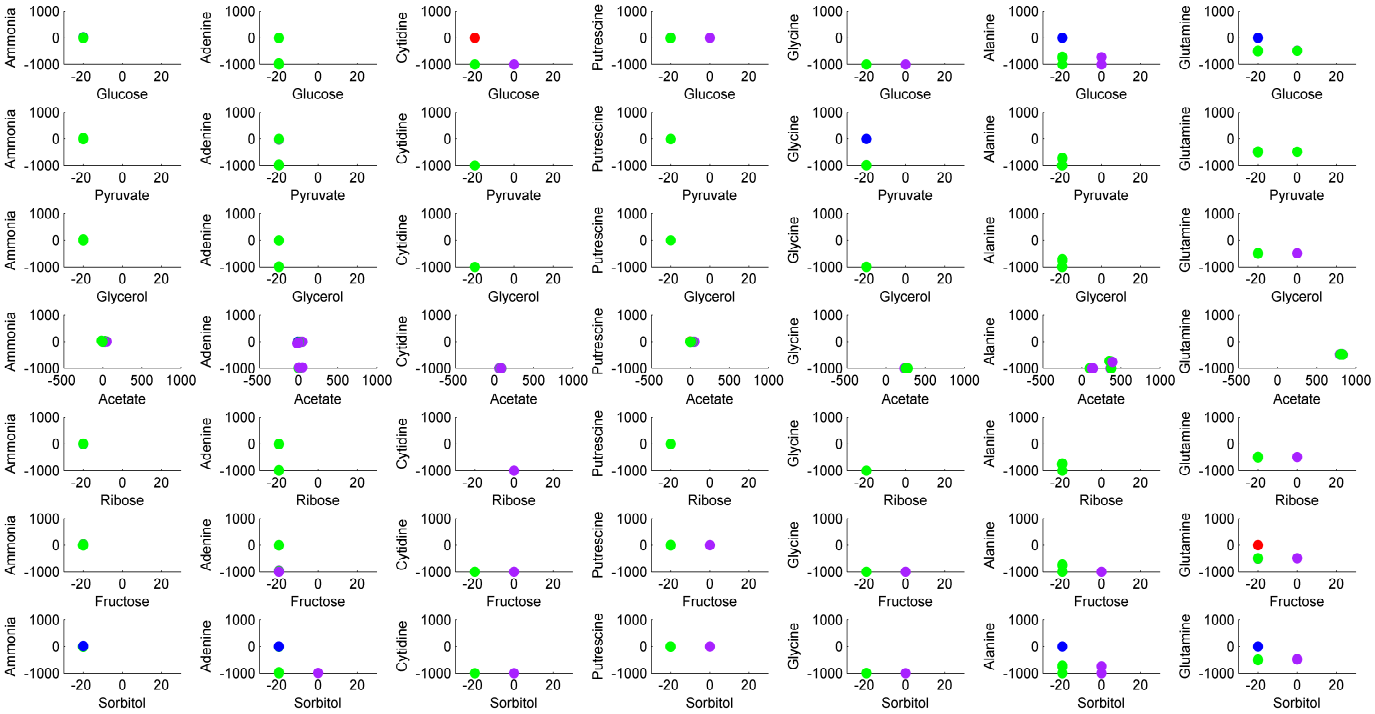

Figure 7: **Scatter plot showing varying uptake amounts of C/N sources when carbon is limited.** We increased the upper bounds of the nitrogen sources and plotted the uptake amounts of carbon and nitrogen sources. The color coding reflects whether separate prediction could predict the carbon and nitrogen sources correctly for each replicate. Green: both C and N source are correctly predicted; blue: only N source is correctly predicted; purple: only C source is correctly predicted; red: neither source is correctly predicted.