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

Viswanadham Sridhara[1], Austin G. Meyer[1,2,5], Piyush Rai[3], Jeffrey E. Barrick[1,2], Pradeep Ravikumar[3], Daniel Segre[4], 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. We generated metabolic flux data using FBA in E. coli MG1655 for the 18 specified growth conditions. We then used a model selection algorithm to identify the key reactions that discriminate among the tested growth conditions. These models on average require assaying the fluxes through 9 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.

Author Summary

Introduction

Microbial systems biology is developing at a rapid pace, with advances in sequencing technologies. Genome sequence along with the available annotation can be used to build a biochemical network of metabolic pathways. These metabolic pathways, when represented by mathematical models provide relationship of phenotype to its genotype. Such phenotype prediction given the growth nutrients or the mutant type (gene deletion etc) is generally carried out using Flux Balance Analysis (FBA) [1] on genome scale metabolic models (GEM) [2–4]. So, given the growth condition, reaction flux vector (phenotype) could be predicted. But can we solve the inverse problem, i.e.,predict the growth condition, given the simulated flux data? This is the question we are interested in answering as it has many applications. For example, microbes can be engineered in lab for useful purposes as well as deliberate attacks as shown in anthrax mailings. To identify the source of pathogen, mathematical models that can predict the growth condition given the cellular composition of the pathogen would be useful. In addition, another problem of significant interest that we address in this work is identifying the specific flux reactions that are the most relevant for discriminating the input growth conditions. Automatic identification of relevant flux reactions thus identified is also expected to lead to better prediction models for the growth conditions.

Here, we used e.coli iAF1260 GEM model [5] that is comprehensively and qualitatively well annotated by 2 research groups for almost 2 decades [2,5–7]. For the current analysis, we used 7 each of carbon and nitrogen growth sources, to generate simulated flux data. These sources were previously shown to result in growth with this iAF1260 model [5]. Once we have the flux data, we then used machine learning techniques to predict the growth condition for each source. To our knowledge, there were no studies that identified

growth growth sources from flux data.

Machine learning algorithms are becoming increasingly popular with computational biology , given the high-dimensional nature of the problems in computational biology. In the high-throughput experimental data, generally the number of samples is considerably smaller than the number of features. 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). For example, differential expression of tens of thousands of genes in microarray studies or identifying the SNPs in GWAS studies is a routine task in biomedical research now-a-days. Even though the sequencing technologies are becoming cheaper day-by-day, the number of samples sequenced is considerably smaller, compared to the number of features. In such cases, prediction algorithms such as LASSO [8] that simultaneously learn the relevant features tend to perform well. 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 [9] 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 seperately , 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 fluxes in our regression analysis. In our model (iAF1260 metabolic model of the *E. coli* K-12 MG1655 strain [10]), this amounted to **938 reactions** please insert number among a total of 2382 distinct reactions. **We also discarded biomass composition reaction that**

resulted in total of 1443 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 [5]. A different set of random contaminants was chosen for each individual FBA calculation.

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 and 1N source** (Specify precisely: is the lowest level of contaminants 1 C and 1 N 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%.

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.

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 Z% of the time. By random chance,

we would have expected $1/13572 = 0.007\%$.

Add a brief paragraph about carbon-limited vs. nitrogen-limited conditions. Add supplementary figures that show which conditions are which. **An uptake amount of 20mmol/g/DW/hr is generally used for carbon and nitrogen sources in FBA studies. If these uptake amounts are sufficient to each other i.e., none of the source is limited, then individual prediction performing better than joint prediction is trivial. However, if we make one source limiting, and if the individual prediction still performs better than joint prediction, that would be an interesting result. So, we changed the upper bounds of carbon to higher levels, while keeping the nitrogen at the same uptake amount and vice versa. We showed the uptake amounts of each source as a scatter plot and is available in Supplementary Info (S1 and S2). Regression analyses on these reaction fluxes was similar to the earlier result i.e., individual prediction performed better than joint prediction (data not shown).**

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.

For 1C and 1N contaminants, individual prediction with largest training data (See Figure 2), we found that our regression model required an average of 10.28 reactions per source (72 reactions each for 7 carbon and 7 nitrogen sources and 62 unique reactions separately for each source type). Out of these 144 reactions, 134 reactions are unique. Excel table showing these reactions is provided in Supplementary info (S3). Is this 9 reactions per C or N source? If yes, please state so. for accurate prediction. Viswanadham, can you add a little more detail here? This must depend on separate vs. joint prediction, number of contaminants, and so on. 3-4 additional sentences would be good.

To gain mechanistic insight into predictive reactions, we mapped them onto the *E. coli* central metabolism (Figures 4 and 5, Table X ← We need this table on top of the figures.). Note that each of the metabolic maps is meant to highlight generally important areas in the *E. coli* central metabolism; we had to limit the number of reactions shown to prevent the figure from becoming unwieldy. 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). Similarly, D-glucose, 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 is this entry or exit from TCA cycle? 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 134 unique reactions for individual prediction of Carbon and Nitrogen sources at

lowest contamination level and with the largest training data set analyzed. This needs some more detail. Where are these 126 reactions? Which specific scenario are we talking about? And how does 126 compare to the average of 9? that were assigned a non-zero coefficient in our trained model. For each of those 134 reactions, we eliminated one at a time, trained a new model for both the Carbon and Nitrogen sources, and calculated the prediction accuracy. With the exception of 3 reactions List which ones they are, we found that the misclassification rate remained unchanged when we removed any one of the other 131 reactions before model fitting. 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 for these growth media in FBA model.** ← There needs to be a table explaining the composition of these media.. **Since the number of experimental media used in the analysis is small (4), we were able to classify these at higher accuracy even at higher noise levels (20 contaminants) and lower number of replicates (50).**Due to the relatively small number of available starting conditions in the training set ←I don't understand this sentence. Also, give more details. How many replicates used? Were there contaminants?, we were able to classify our experimentally relevant media choices very accurately. Our misclassification rate was less than 1% for noise levels up to 20 contaminants. XXX 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. Please specify: # of replicates; contaminants; total number of observations. 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?). 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 please quantify "most" 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. I don't know why adenine gets predicted often?**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.**

can we add a result for a previously unseen nitrogen source? **I started Cytosine run and should be done sometime soon.**

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 fractional noise can have a dramatic affect on model accuracy, the misclassification rate remains acceptably low even with 10% or 20% noise. The addition of noise revealed one interesting and unexpected physiological hypothesis about E. coli metabolism. Namely, as noise increases from 1% to 20%, 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 order 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. **(Maybe another sentence to connect thoughts)**. 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 **??**). 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) [11] or identifying SNPs in GWAS studies [12]. **(Viswanadham, there needs to be something else to make the sentence structure flow well)**. We would like to point out that beyond LASSO there are a number of other commonly used regularization techniques. For example, graphical Lasso [13] and Ising Markov Random Field models [14] can also be used to study biological networks. **(...do something that I'm not really sure about)**. **(Viswanadham, I'm not sure if this is accurate... if not, we should say why we chose lasso.)** 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.

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 [15] with MATLAB and for multinomial regression, we used GLMNET package [9] with R. The methods are described in detail below.

Flux Balance Analysis

We carried out flux balance analysis using the COBRA toolbox [15] for MATLAB. We used the iAF1260 model from the BiGG database [10] (reference?). 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 1000mmol/gDWhr, 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 50mmol/gDWhr. The lower bound for the majority of reactions (nearly 1800) is set to 0mmol/gDWhr, meaning that these reactions cannot uptake any metabolites from the media. **A set non-growth associated maintenance of** 8.39**mmol/gDWhr is used for ATPM reaction. ATPM is ATP maintenance requirement value.** For the ATPM reaction could you specify what ATPM does?, the lower bound is set to be the same as upper bound at 8.39mmol/gDWhr. For the remaining reactions, all transport reactions that model metabolite uptake, the lower bounds were set to -1000mmol/gDWhr, except for glucose, nitrogen, and oxygen. I don't understand this. Shouldn't all the uptake reactions have a lower bound of 0? We did not change the oxygen uptake rate (-18.5mmol/gDWhr), 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 picked the growth conditions manually from [5] that seemed more common in experiments. 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/gDWhr. This lower bound of -20 mmol/gDWhr is previously used as reasonable uptake amount in many studies [5]. 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].

We changed the uptake rate of nitrogen source to -1000 mmol/gDW/hr, keeping the carbon source at -20mmol/gDW/hr to make sure carbon sources don't lack nitrogen source. We also repeated the analysis keeping the carbon source uptake at -1000 mmol/gDW/hr and nitrogen source uptake at -20mmol/g/DW/hr.

Contaminants

To make the simulation scenario more challenging and more realistic, we incorporated different numbers of contaminants to the simulated growth media. background noise levels. For this, we used a subset of the 174 carbon and 78 nitrogen sources, previously used in Feist et. al [5]. We used different background noise levels, ranging from 1% to 20%. For example, if we want to set 5% noise level, we randomly picked 5 Carbon and 5 nitrogen sources and set their lower bounds to -0.2mmol/gDW. 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 here.

Once we obtained the flux data, we then used a biomass threshold to filter out non viable growth measurements. We used a threshold value that is equal to the (mean - 3*standard deviation) of the biomass value across all the replicates. Explain how viable solutions are identified. What is your biomass threshold? How did you arrive at that number?

Regularized regression

We predicted growth conditions using regularized multinomial logistic regression, as implemented in the GLMNET package [9] for R. As you say below, λ is chosen by cross validation. Were there any other parameter choices that you had to make?**Other parameters used were standard settings that are used in GLMNET package.**

After filtering for biomass, for each number of contaminants, we used half of the dataset as test set. After filtering for biomass? We used subsets of the remaining half as training (i.e, 240, 480, 2400 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 (2, 10, 20) and different training data sizes (240, 480, 2400 observations).

Acknowledgments

Need to acknowledge ARO grant and any other relevant grants.**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. We would also like to

thank BCG and TACC at UT for computing resources.

Author Contributions

Need to update to include all authors. **Updated** 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. Karp PD, Riley M, Paley SM, Pelligrini-Toole A (1996) Ecocyc: an encyclopedia of escherichia coli genes and metabolism. Nucleic Acids Res 24: 32-9.

4. Ouzounis CA, Karp PD (2000) Global properties of the metabolic map of escherichia coli. Genome Res 10: 568-76.

5. 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.

6. 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.

7. 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.

8. Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B-Methodological 58: 267-288.

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

10. Schellenberger J, Park JO, Conrad TM, Palsson BO (2010) Bigg: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. Bmc Bioinformatics 11.

11. 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.

12. Wu TT, Chen YF, Hastie T, Sobel E, Lange K (2009) Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25: 714-21.

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

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

15. 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.

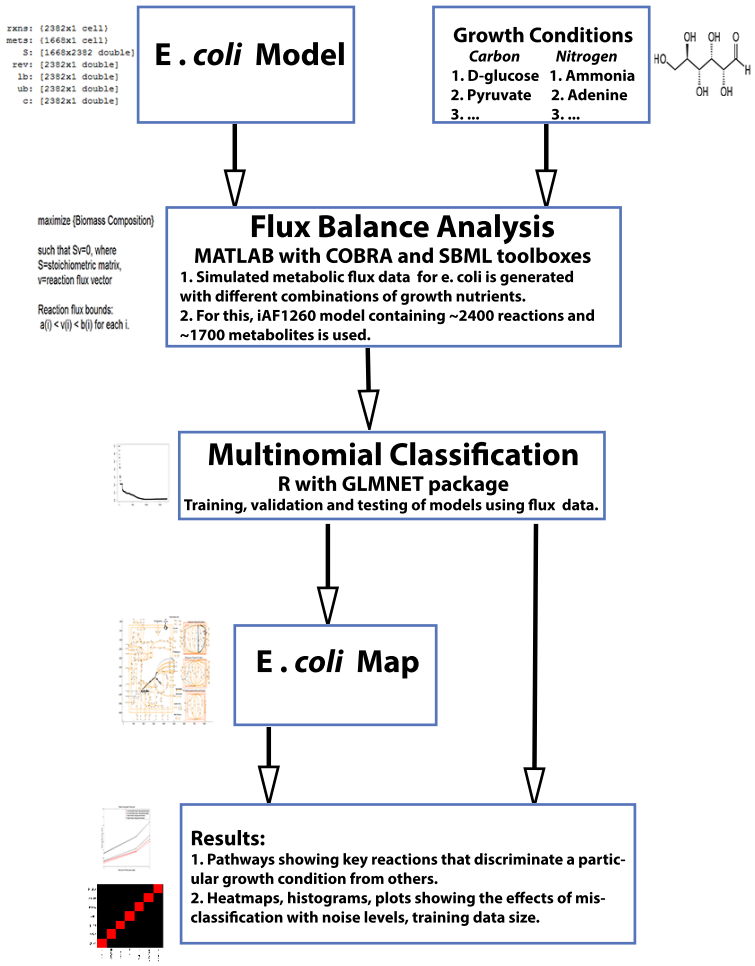

Figure 1: **Flowchart describing methodology used in this study.** We obtained E. coli model and map from BiGG database. The key steps involved are Flux Balance Analysis and multinomial classification routines.

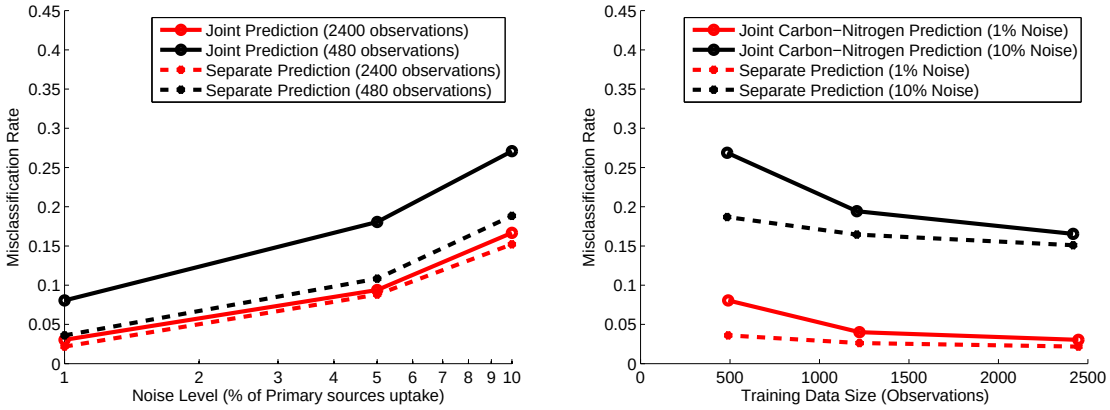

Figure 2: **Misclassification rate versus number of contaminants and amount of training data.** (left) The misclassification rate increases as the number of contaminants increases. (right) The misclassification rate decreases as the size of the available training data decreases. In all cases, separate prediction out-performs joint prediction. The two figures need to be combined into one, with labels A and B. Instead of "noise" write "n contaminants" where n is the appropriate number.

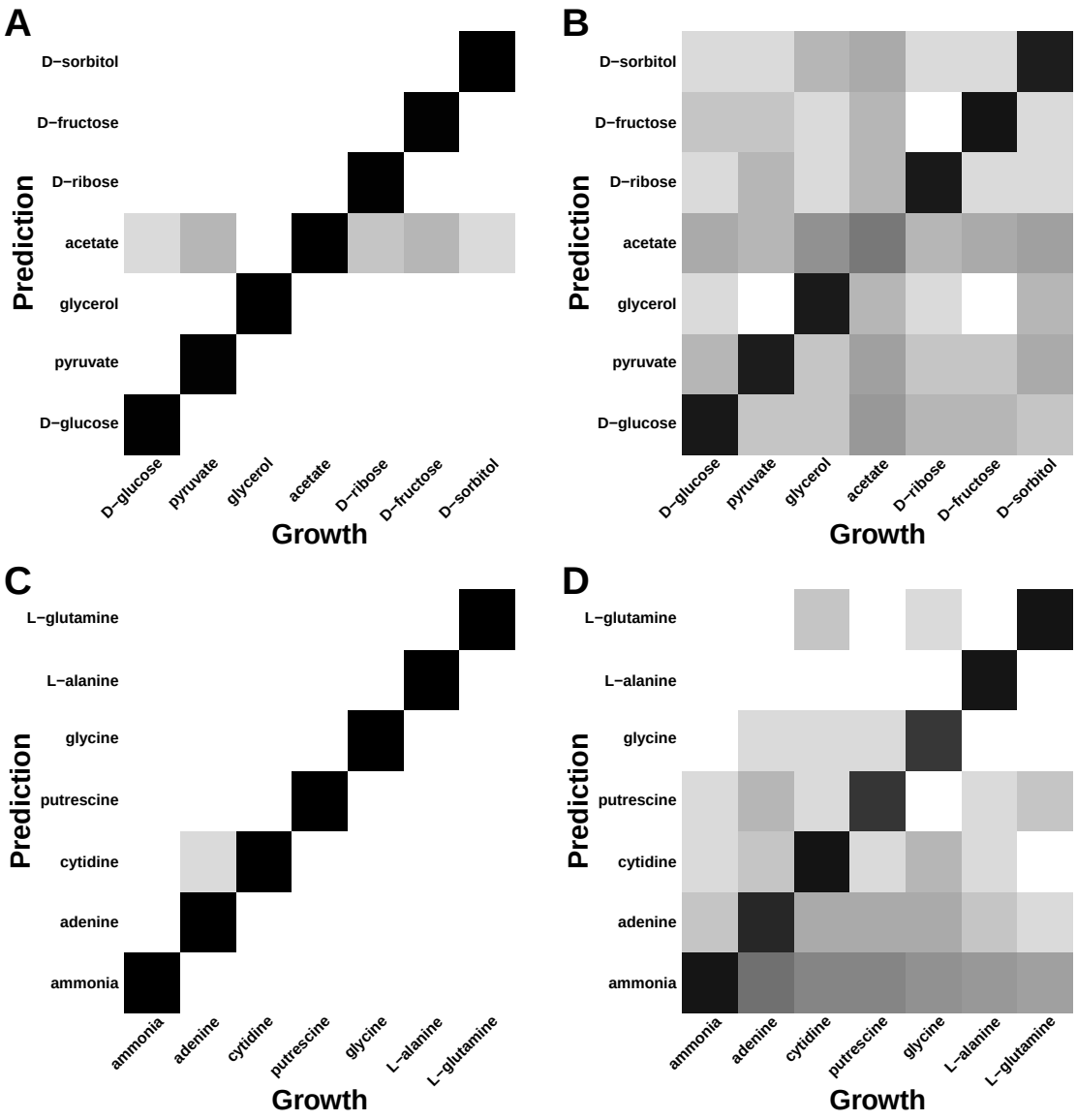

Figure 3: **Heat maps with actual sources as columns and predicted ones in rows.** At 20% noise, most C sources are predicted to be acetate and most N sources are predicted to be ammonia.

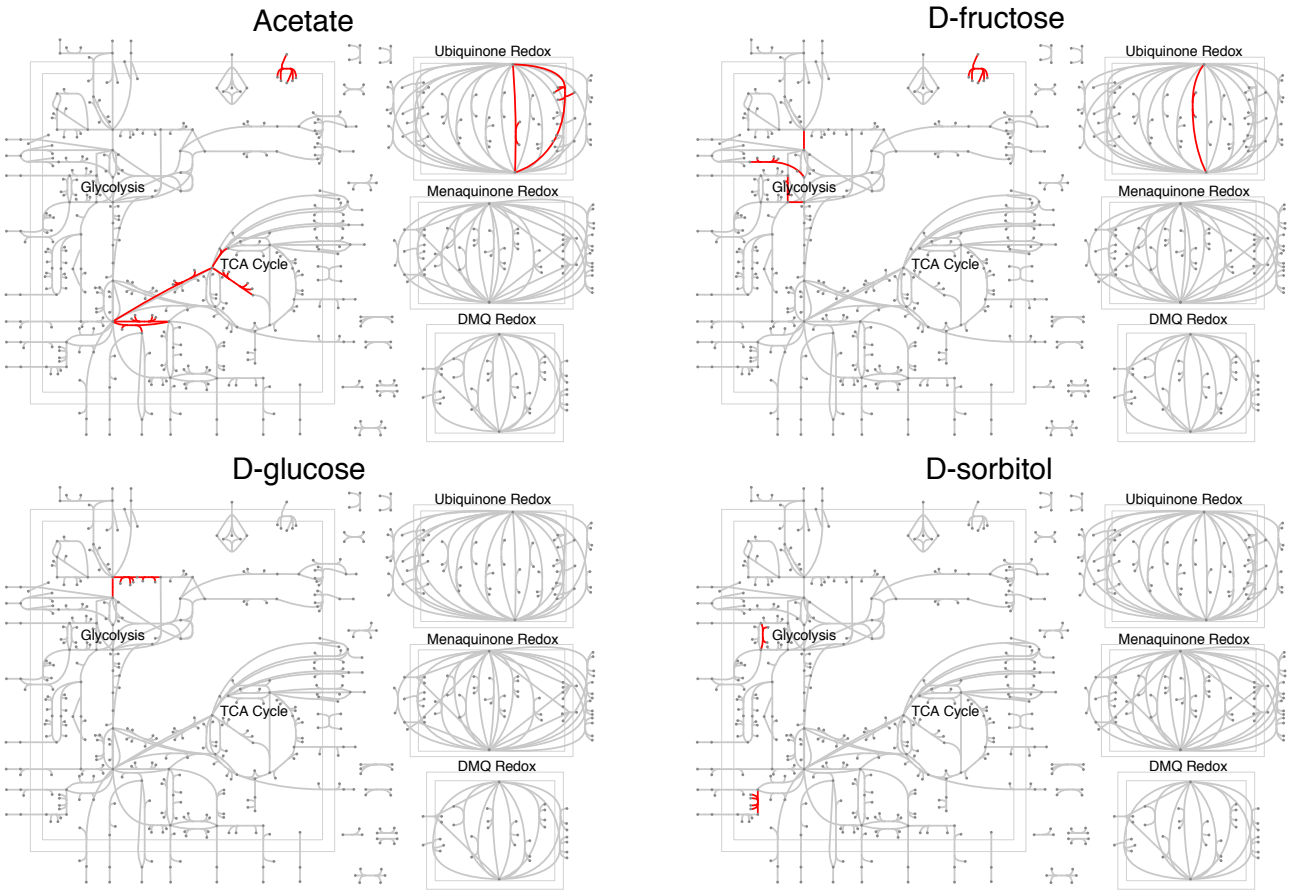

Figure 4: **Discriminatory carbon sources** The key-reactions identified by GLMNET package were mapped onto E. coli central metabolism to visually show the differences between different growth conditions. Out of 7 carbon sources, here we show 4 carbon sources and the key-reactions.

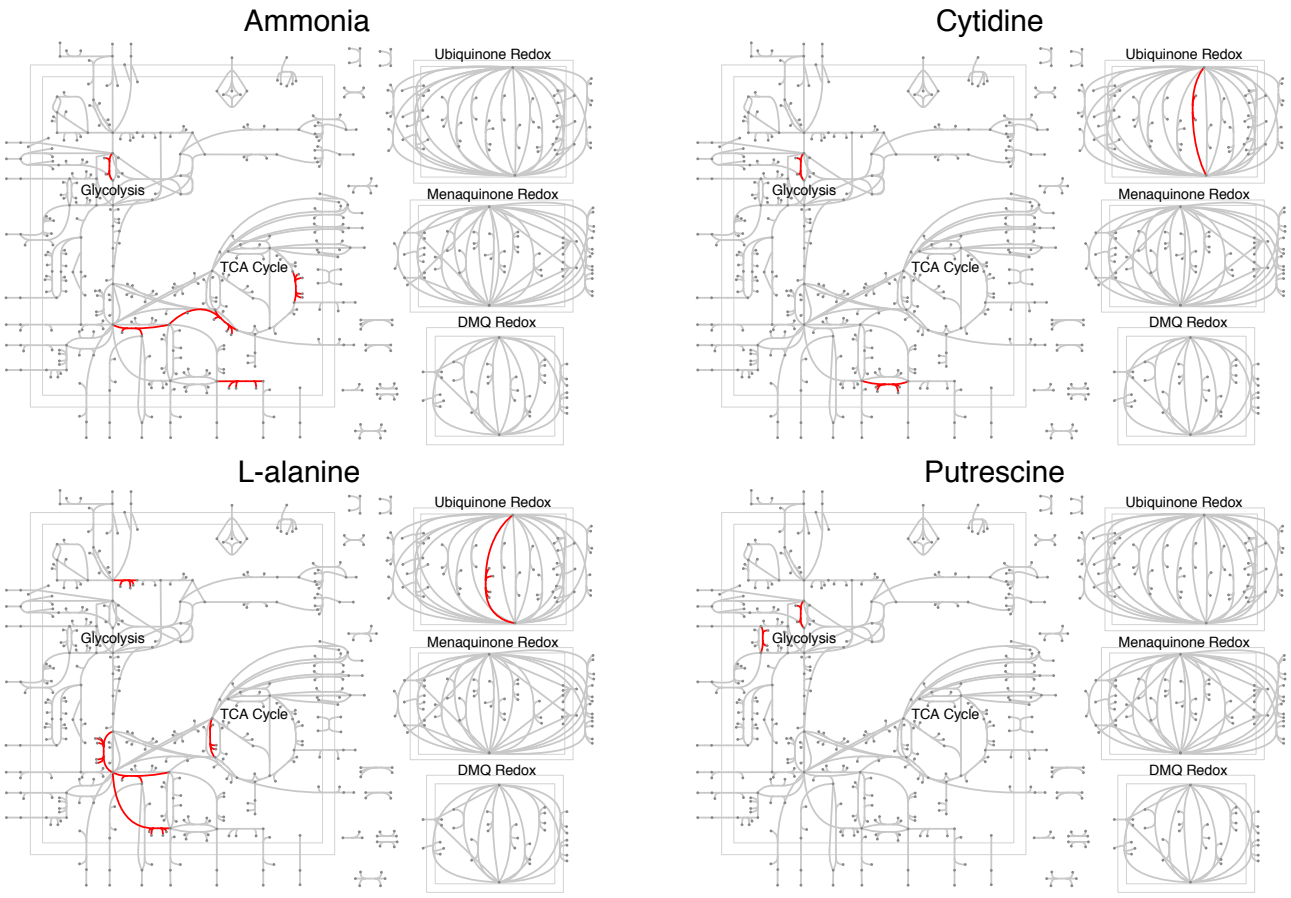

Figure 5: **Discriminatory nitrogen sources** The key-reactions identified by GLMNET package were mapped onto E. coli central metabolism to visually show the differences between different growth conditions. Here, the growth medium used are generally used for K-12 MG1655 strain.