COMPUTATIONAL MODELING OF METABOLIC PATHWAYS TOWARD PREDICTING DYNAMIC PHENOTYPES

A Dissertation
Presented to
The Academic Faculty

By

Justin Y. Lee

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in
Bioengineering

Georgia Institute of Technology

August 2021

COMPUTATIONAL MODELING OF METABOLIC PATHWAYS TOWARD PREDICTING DYNAMIC PHENOTYPES

Approved by:

Dr. Mark P. Styczynski, Advisor
School of Chemical & Biomolecular Engineering
Georgia Institute of Technology

Dr. Andrew J. Medford
School of Chemical & Biomolecular Engineering
Georgia Institute of Technology

Dr. Fani Boukouvala
School of Chemical & Biomolecular Engineering
Georgia Institute of Technology

Dr. Eberhard O. Voit
School of Biomedical Engineering
Georgia Institute of Technology

Dr. Melissa L. Kemp
School of Biomedical Engineering
Georgia Institute of Technology

Date Approved: May 24, 2021

ACKNOWLEDGEMENTS

The past five years have flown by faster than I could have ever imagined, and there are many people to thank that have made this journey possible.

First, I would like to acknowledge my advisor, Dr. Mark Styczynski, for his guidance, patience, and bad jokes that never fail to put a smile on the face of at least someone in our group. Without his confidence in my abilities, unwavering support both inside and outside the lab, and innate talent for understanding my occasionally scrambled thoughts when explaining a research problem, I would have been constantly stuck, spinning my wheels in the mud throughout this process. Thank you for making me a better scientist and I will undoubtedly continue to learn from you in the next stage of my career.

I would also like to thank my committee members, Dr. Fani Boukouvala, Dr. Melissa Kemp, Dr. Andrew Medford, and Dr. Eberhard Voit. Each has provided me with invaluable advice and critiques that have helped me solve some of my most difficult challenges and identify new potential obstacles before they even became apparent to me. Their feedback during my proposal and research updates has contributed significantly to realizing this thesis.

To my labmates in the Styczynski group, I would like to thank you all for your support and helpful questions during my group presentations, despite my research encompassing a vastly different scope than most of your projects. Though our group bonding activities were cut short this past year, I will never forget our trips to Jeni's, yurting adventures, and of course, encounters with bed bugs.

Outside of the lab, I would like to thank all my friends that I have made during my time in Atlanta, whether it be through Georgia Tech, Roche, kickball, or bar games. From countless brewery outings, to eating Cook Out absurdly late, to visiting Yosemite, to Tacosgiving – these get-togethers always kept me going.

And to my friends from Boston and Penn. I am extremely lucky to have formed such incredible friendships with you all before moving down to Atlanta. Though I sometimes felt like I was alone on an island in the South, you all have always kept in touch and made me feel like I was never too far away. Like Philadelphia, New York, Nashville, Bend, Lake Chelan, and other trips before, I look forward to many more amazing adventures with you all in the future.

Finally, and most importantly, I would like to thank my parents and my sister, Phyllis. You have always pushed me to achieve more than I ever thought possible. You have helped me through my toughest setbacks and have cheered me on for my smallest accomplishments. This dissertation would not have been possible without your endless support. For that, I am forever grateful.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ..... **iii**

LIST OF TABLES ..... **viii**

LIST OF FIGURES ..... **ix**

LIST OF ABBREVIATIONS ..... **xii**

SUMMARY ..... **xiv**

CHAPTER 1: Introduction **1**

1.1 Metabolic Systems 1
1.2 Metabolomics 2
1.2.1 Applications of metabolomics 3
1.2.2 Methods for measuring metabolomics data 4
1.3 Metabolic Modeling Approaches 6
1.3.1 Constraint-based modeling 6
1.3.2 Examples of constraint-based models 8
1.3.3 Ordinary differential equation-based modeling 9
1.3.4 Examples of ODE-based models 10
1.3.5 Linear Kinetics-Dynamic Flux Balance Analysis 11
1.4 Regulation in Metabolic Systems 12
1.4.1 Transcriptional regulation 12
1.4.2 Allosteric regulation 13
1.4.3 Constraint-based modeling frameworks that integrate regulation 14
1.4.4 Determining the topology of allosteric regulation 15
1.5 Thesis overview 16

CHAPTER 2: A Stepwise Machine Learning Framework for Predicting Metabolite-dependent Regulatory Interactions **18**

2.1 Background 18
2.2 Methods for Predicting Regulatory Interactions 21
2.2.1 Synthetic model networks 21
2.2.2 Biological models 22
2.2.3 Autogenerated training data 27
2.2.4 Noise-added data 29
2.2.5 Data pre-processing 30
2.2.6 Features 30
2.2.7 Scaling of feature matrices 34
2.2.8 Machine learning stacking 35
2.2.9 Machine learning algorithms 36
2.2.10 Stepwise approach 37
2.2.11 Framework performance metrics 40
2.3 Results 41

2.3.1 Performance on noiseless data 41
2.3.2 Performance on noisy data 42
2.4 Discussion 45
2.5 Conclusions 55

CHAPTER 3: Inferring Absolute Concentrations from Relative Abundances in Metabolomics Data 56

3.1 Background 56
3.2 Methods for Inferring Absolute Concentrations 59
3.2.1 Synthetic models 59
3.2.2 Biological models 60
3.2.3 Response factors 61
3.2.4 Kinetic equations approach 62
3.2.5 Optimization approach 63
3.2.6 Combining the kinetic equations and optimization approaches 65
3.2.7 Solving for flux distributions in the optimization approach 67
3.2.8 Noise-added data 68
3.2.9 Evaluation metrics and comparing to baseline methods 69
3.3 Results 70
3.3.1 MetaboPAC performance on noiseless data 70
3.3.2 MetaboPAC performance on noisy data 74
3.4 Discussion 83
3.5 Conclusions 88

CHAPTER 4: Improved Kinetics Constraints Increase the Predictivity and Applicability of a Linear Programming-based Dynamic Metabolic Modeling Framework 89

4.1 Background 89
4.2 Linear Kinetics Dynamic Flux Balance Analysis (LK-DFBA) 92
4.3 Methods for Improving LK-DFBA 93
4.3.1 Constraint approaches 93
4.3.1.1 LK-DFBA (LR) 93
4.3.1.2 LK-DFBA (NLR) 94
4.3.1.3 LK-DFBA (DR) 94
4.3.1.4 LK-DFBA (HP) 95
4.3.2 Translating constraints to contain training data 96
4.3.3 Test models 96
4.3.3.1 Synthetic model 97
4.3.3.2 Lactococcus lactis model 98
4.3.3.3 Escherichia coli model 98
4.3.4 Kinetic parameters in E. coli models 99
4.3.5 LK-DFBA objective functions 99
4.3.6 Pathway perturbations 100
4.3.7 Generating noisy data 101
4.3.8 Error calculation 101
4.3.9 Pearson correlation calculation 102
4.4 Results 103

4.4.1 Fitting and predicting phenotypes in synthetic models 103
4.4.2 Fitting and predicting phenotypes in L. lactis and E. coli models 108
4.4.3 Improved LK-DFBA predictions yield qualitative consistency with experimental L. lactis metabolite concentration data 114
4.4.4 Changes in LK-DFBA flux profiles due to gene knockouts are correlated with experimental E. coli steady-state flux data 118
4.5 Discussion 121
4.6 Conclusions 125

CHAPTER 5: A Workflow Toward Modeling Dynamics in Metabolic Systems 127

5.1 Introduction 127
5.2 Determined system with regulation 128
5.3 Metabolic modeling workflow 129
5.3.1 Method for imputing missing metabolomics data 129
5.3.2 Identifying metabolic regulation using relative abundances 130
5.3.3 Inferring absolute concentrations from relative abundances and identified regulation 132
5.3.4 Creating a metabolic model using inferred absolute concentrations and identified regulation 134
5.4 Conclusions 141

CHAPTER 6: Future Directions 142

6.1 Thesis contributions 142
6.2 Improvements to SCOUR 144
6.2.1 Different machine learning algorithms 145
6.2.2 Including more biologically relevant interactions in the training data 147
6.2.3 Modification of autogeneration methods to include topological information 148
6.2.4 Using SCOUR on experimental data 150
6.3 Improvements to MetaboPAC 150
6.3.1 Non-linear relationships 150
6.3.2 General improvements to the optimization approach 152
6.3.3 Developing a platform for predicting confidence in inferred response factors 153
6.4 Further improvements to LK-DFBA 154
6.4.1 Existing methods for optimizing objective functions 156
6.4.2 Optimizing the objective function in LK-DFBA 157
6.4.3 Improving predictions of cofactor metabolite concentrations 159
6.4.4 Using LK-DFBA in metabolic engineering 160
6.5 Closing remarks 161

APPENDIX A 163

REFERENCES 167

LIST OF TABLES

Table 1: List of controller metabolites and target fluxes in *S. cerevisiae* model. 23
Table 2: List of controller metabolites and target fluxes in *E. coli* model..... 24
Table 3: The number of *n*-controller metabolite interactions that exist in or are possible for each model. 26
Table 4: List of features for each step of the framework 31
Table 5: Penalties used in the optimization approach of MetaboPAC 64
Table 6: Kinetic parameters in synthetic models. 98
Table 7: Normalized root mean square error of the median concentration profile predictions by each LK-DFBA kinetics constraint approach compared to the ODE data..... 141

LIST OF FIGURES

Figure 1: Example of a chromatogram. 6
Figure 2: Example of metabolite-dependent allosteric regulation. 14
Figure 3: Synthetic systems tested with SCOUR. 22
Figure 4: Workflow of stacking process. 36
Figure 5: Workflow of stepwise machine learning framework for identifying one-, two-, and three-controller metabolite interactions. 39
Figure 6: SCOUR performance on synthetic and biological models using noiseless training and test data. 42
Figure 7: SCOUR performance on synthetic and biological models using noisy and low sampling frequency training and test data. 45
Figure 8: SCOUR's PPV for 3-controller metabolite interaction predictions is significantly greater than a random classifier. 47
Figure 9: The most common false negative two-controller metabolite interactions in the *E. coli* model. 49
Figure 10: F1 scores for synthetic and biological models using noisy and low sampling frequency training and test data. 50
Figure 11: SCOUR performance when training on autogenerated data and testing on autogenerated testing data. 52
Figure 12: Four-controller and higher-order metabolite regulatory interactions. 54
Figure 13: Synthetic systems tested with MetaboPAC. 60
Figure 14: MetaboPAC workflow for inferring absolute concentrations from relative abundances in metabolomics datasets. 67
Figure 15: MetaboPAC performance on noiseless data for synthetic systems. 71
Figure 16: Percent of response factors predicted by the kinetic equation and optimization approaches within each $\log_2$ error range for the synthetic systems when using noiseless data. 71
Figure 17: Percentage of response factors predicted by the kinetic equation and optimization approaches for the synthetic systems. 72
Figure 18: MetaboPAC performance on noiseless data for biological systems. 73
Figure 19: Percent of response factors predicted by the kinetic equation and optimization approaches within each $\log_2$ error range for the biological systems when using noiseless data. 73
Figure 20: Percentage of response factors predicted by the kinetic equation and optimization approaches for the biological systems. 74
Figure 21: MetaboPAC performance on all conditions of noisy data for the determined system. 75
Figure 22: MetaboPAC performance on all conditions of noisy data for the underdetermined system with regulation. 76
Figure 23: Percent of response factors predicted by the kinetic equation and optimization approaches within each $\log_2$ error range for the determined system when using noisy data 77
Figure 24: Percent of response factors predicted by the kinetic equation and optimization

approaches within each $\log_2$ error range for the underdetermined system with regulation when using noisy data. 78
Figure 25: MetaboPAC performance on all conditions of noisy data for the *S. cerevisiae* system. 80
Figure 26: MetaboPAC performance on all conditions of noisy data for the *E. coli* system. 81
Figure 27: Percent of response factors predicted by the kinetic equation and optimization approaches within each $\log_2$ error range for the *S. cerevisiae* system when using noisy data. 82
Figure 28: Percent of response factors predicted by the kinetic equation and optimization approaches within each $\log_2$ error range for the *E. coli* system when using noisy data. 83
Figure 29: MetaboPAC Performance when true response factors are sampled from a log uniform distribution. 87
Figure 30: Synthetic model. 97
Figure 31: NRMSE heatmap of LK-DFBA approaches on different synthetic models using noiseless data. 105
Figure 32: NRMSE heatmap of LK-DFBA approaches on different synthetic models using noise-added data (nT = 15, CoV = 0.15). 106
Figure 33: LK-DFBA performance on noisy synthetic model data after smoothing. 108
Figure 34: NRMSE heatmaps of constraint approaches on model of *L. lactis* metabolism. 110
Figure 35: NRMSE heatmaps of constraint approaches on model of *E. coli* metabolism. 111
Figure 36: LK-DFBA performance on different models with the same stoichiometric topology as the *E. coli* model. 113
Figure 37: Comparison of LK-DFBA metabolite concentration predictions when fitted to noiseless ODE data against noiseless ODE data and all available *L. lactis* experimental data. 115
Figure 38: Comparison of LK-DFBA metabolite concentration predictions when fitted to noisy data against ODE and all available *L. lactis* experimental data. 117
Figure 39: Comparison of LK-DFBA (LR) metabolite concentration predictions against ODE data and *L. lactis* experimental data when fitted to noisy ODE data. 118
Figure 40: Pearson correlation coefficients of LK-DFBA and ODE model flux predictions with *E. coli* experimental data. 120
Figure 41: Workflow for selecting the best constraint approach for LK-DFBA when modeling metabolic systems. 123
Figure 42: Workflow toward modeling dynamics in metabolic systems. 128
Figure 43: Determined system with regulation used to evaluate the metabolic modeling workflow. 129
Figure 44: SCOUR performance using relative abundances from the small synthetic model with unknown regulation. 132
Figure 45: MetaboPAC performance on determined system with regulation. 133
Figure 46: Comparison of LK-DFBA kinetic constraints on the determined system with regulation using inferred absolute concentrations from the MetaboPAC optimization approach. 136

Figure 47: Comparison of LK-DFBA kinetic constraints on the determined system with regulation using inferred absolute concentrations from the MetaboPAC kinetic equations approach. 137
Figure 48: Comparison of LK-DFBA kinetic constraints on the determined system with regulation using the original ODE data and assuming all regulatory interactions were known. 138
Figure 49: Comparison of LK-DFBA kinetic constraints on the determined system with regulation using the relative abundance data and assuming all regulatory interactions were unknown..... 140
Figure 50: LK-DFBA performance on noisy synthetic model data, nT = 50, CoV = 0.05. 164
Figure 51: LK-DFBA performance on noisy synthetic model data, nT = 50, CoV = 0.15. 165
Figure 52: LK-DFBA performance on noisy synthetic model data, nT = 15, CoV = 0.05. 166

LIST OF ABBREVIATIONS

arFBA	Allosteric Regulation Flux Balance Analysis
BOSS	Biological Objective Solution Search
BST	Biochemical Systems Theory
CBM	Constraint-Based Model
CoV	Coefficient of Variation
DFBA	Dynamic Flux Balance Analysis
DFE	Dynamic Flux Estimation
DOA	Dynamic Optimization Approach
DR	Dimension Reduction
FBA	Flux Balance Analysis
GC-MS	Gas Chromatography-Mass Spectrometry
HP	Hyperplane
iFBA	Integrated Flux Balance Analysis
kNN	k-Nearest Neighbors
LK-DFBA	Linear Kinetics-Dynamic Flux Balance Analysis
LC-MS	Liquid Chromatography-Mass Spectrometry
LP	Linear Program
LR	Linear Regression
LR+	Linear Regression Plus
MCAR	Missing Completely At Random
MetaboPAC	Metabolomics Prediction of Absolute Concentrations
MNAR	Missing Not At Random

MoMA	Minimization of Metabolic Adjustment
NET Analysis	Network-embedded Thermodynamic Analysis
NLR	Non-linear Regression
NMR	Nuclear Magnetic Resonance
NRMSE	Normalized Root Mean Square Error
NS-kNN	No Skip k-Nearest Neighbors
nT	Number of Timepoints
ODE	Ordinary Differential Equation
PCA	Principal Component Analysis
PPV	Positive Predictive Value
pFBA	Parsimonious Flux Balance Analysis
QCP	Quadratically Constrained Program
RA	Relative Abundance
RF	Response Factor
rFBA	Regulatory Flux Balance Analysis
SCOUR	Stepwise Classification Of Unknown Regulation
SIMMER	Systematic Identification of Meaningful Metabolic Enzyme Regulation
SOA	Static Optimization Approach
SR-FBA	Steady-state Regulatory Flux Balance Analysis
SVM	Support Vector Machines
TMFA	Thermodynamic Metabolic Flux Analysis
uFBA	Unsteady-state Flux Balance Analysis
WT	Wildtype

SUMMARY

Metabolic systems are important to a wide variety of applications, including therapeutic development, agricultural crop production, and manufacturing of industrial chemicals. Developing metabolic models is one of the best approaches to study metabolism, as computational experiments are generally cheaper and faster to perform than experiments in a laboratory. While there are computational frameworks that can model large metabolic systems at steady state or the metabolite dynamics of a small number of key metabolic pathways, it is substantially more difficult to model the dynamics of metabolism at the genome scale. In this thesis dissertation, I present three computational platforms that address several of the challenges in developing dynamic genome-scale metabolic models. First, I devised a stepwise machine learning strategy for identifying the regulatory topology within metabolic systems, which can be used to construct more accurate metabolic models. I then developed a framework for inferring absolute concentrations from relative abundances in metabolomics data, which will allow metabolomics (the systems-scale study of metabolites) to be more easily used with metabolic modeling tools. Finally, I implemented new constraints within a linear programming dynamic modeling framework that increase its ability to model a wider variety of metabolic systems. Together, these three platforms create a cohesive workflow for modeling the dynamics of metabolism at any scale.

CHAPTER 1: Introduction

1.1 Metabolic Systems

As the set of chemical reactions that are necessary to sustain life, metabolism is one of the most critical processes in all organisms, from the smallest bacterium to the largest mammals. Metabolism generates energy for our bodies after we eat, directs chemical resources to our muscles after a workout, and even breaks down unnecessary chemicals into waste products as we sleep. Whether we are aware of it or not, metabolic processes are constantly working behind the scenes to ensure our bodies are functioning properly. When speaking of metabolism, one may initially think of its importance in humans and how it pertains to topics such as weight-loss and aging. What some people may not realize is that metabolism also plays a key role in various diseases[1], drug discovery[2], and developing personalized medical treatment for patients[3].

Though understanding human metabolism clearly has biomedical relevance, there is also great interest in studying the metabolism of other species, including the model organisms *Escherichia coli* and *Saccharomyces cerevisiae*. Metabolism in these two unicellular organisms, and others like them, is less complex than the organism-scale metabolism found in humans, but it is still incredibly valuable to science. While the most important pathways, such as those in central carbon metabolism, are topologically well-conserved across different species[4], many organisms have unique metabolic reactions or are able to produce certain chemicals in much higher quantities than other species due to corresponding fitness advantages that have been selected for over the course of evolution. An example of a microorganism with unique metabolite production capabilities is *S.*

cerevisiae, a popular yeast species in the culinary arts, brewing, and bioresearch because of its innate ability to produce ethanol and CO_2 during fermentation[5]. Studying the differences in metabolism across various organisms can provide meaningful insight about the importance of certain metabolic pathways.

Besides just studying the metabolism of microorganisms, scientists also engineer metabolism, using genetic modifications to reroute metabolic resources and produce high quantities of molecules of interest. Heterologous pathways can even be introduced into the system to allow the cell to synthesize products it otherwise could not[6]. This practice of designing and engineering pathways is known as metabolic engineering[7]. The field of metabolic engineering has rapidly expanded over the last few decades[8] as researchers have harnessed its potential to engineer organisms to produce valuable chemicals that would otherwise be too expensive or too difficult to manufacture via chemical synthesis. Metabolic engineering has proven useful in a wide variety of applications including therapeutic development[9], agricultural crop production[10], and renewable energies[11]. In the next sections, I discuss the data used to understand metabolic systems.

1.2 Metabolomics

To study metabolism in a range of organisms, metabolomics has emerged as a valuable -omics field, following in the footsteps of transcriptomics, proteomics, and genomics. Metabolomics is defined as the systems-scale study of metabolites, the chemical intermediates used to sustain life[12]. Whereas transcriptomics, proteomics, and genomics provide more upstream views of cellular functions, metabolomics is a direct readout of biochemical activity and the metabolic state of a cell. Since its inception,

metabolomics has been used in a wide variety of applications.

1.2.1 Applications of metabolomics

Perhaps the most immediate and widely-known uses of metabolomics relate to human metabolism. One of the most important applications of metabolomics is its use in medicine as a tool to identify disease biomarkers[1]. By comparing the metabolite profiles of cancer patients to those of healthy individuals, metabolomics has elucidated key metabolite biomarkers that have been helpful in disease prognosis or monitoring disease progression[13]. Metabolomics has also provided insight for better understanding and preventing diseases in the first place[14]. Many common diseases[15], such as heart disease, diabetes, Parkinson's, and various cancers, are known to have clear connections with metabolism, which has led to great interest in using metabolomics to develop and screen new drugs. When testing different drug candidates, metabolomics can help determine which candidates affect specific metabolic pathways based on the metabolic changes they induce[16, 17]. This is especially important in determining if a drug candidate could lead to toxic levels of certain metabolites within the body. Recently, the idea of using metabolomics in personalized medicine[3], where medical treatment is tailored specifically to a patient's metabolome, has also gained popularity.

Metabolomics has not only been used to study metabolism in people, but it has also had a significant impact on plant and microbial research. In plant sciences, metabolomics has generated insight about plant response to changing environments[18], antimicrobial resistance in agricultural crops[19], differences in metabolic profiles between transgenic plants[20], and gene annotation[21]. Microbial metabolomics research has been

incredibly beneficial for disease research. In particular, yeast metabolomics has been used to study cancer[22] and bacterial metabolomics has the potential to reveal quorum sensing metabolites that could mitigate the effects of bacterial infections[23]. Additionally, the link between metabolomics and the gut microbiota has helped researchers understand how our bodies breakdown food or foreign substances, such as pharmaceutical drugs[24].

Metabolomics has also garnered interest as a tool for building metabolic models[25], as it can contribute a large amount of metabolic information. Despite providing direct insight on how metabolic resources are being consumed or produced, one area in which metabolomics has been used surprisingly little is metabolic engineering[26]. One of the most challenging obstacles to using metabolomics in metabolic engineering and other analytical tools is how limited raw metabolomics data can be without using standards for each metabolite. Because the properties of different chemicals cause metabolites to be measured relatively instead of absolutely, it is not possible to compare the quantities of different metabolites to each other. Below, I discuss the different analytical instruments used to measure metabolomics and further expand on the current limitations of using metabolomics data with metabolic tools.

1.2.2 Methods for measuring metabolomics data

There are three common analytical techniques used to measure metabolomics data. Nuclear magnetic resonance (NMR) spectroscopy uses strong magnetic fields to measure chemicals that behave differently based on the nuclei in their atoms. One key advantage of using NMR is that it is a nondestructive method, meaning samples can be reused[27]. However, the biggest disadvantage of NMR is its low sensitivity. To measure

hundreds or thousands of metabolites at low concentrations, researchers typically turn to gas chromatography or liquid chromatography coupled to mass spectrometry (GC-MS and LC-MS, respectively). GC-MS primarily separates chemicals in the gas column based on their boiling points and their affinity for the chromatography column, and then the molecules in the sample are ionized and further separated via a mass spectrometer. Prior to injection into a GC-MS instrument, samples must be derivatized because many chemicals are not volatile enough to be vaporized effectively in their native states[28]. The advantage of LC-MS is that no derivatization is necessary because chemicals are separated in a liquid phase based on their interactions with a stationary phase in the column instead of in the gas phase.

Both GC-MS and LC-MS have become increasingly popular[29] and arguably the preferred analytical tools to measure metabolites, but it is difficult to quantify absolute concentrations of metabolites using either method. The data that these two mass spectrometry approaches yield are relative, rather than absolute, abundances. While the relative abundances of a single metabolite can be compared across different timepoints or different samples in an experiment, comparing the relative abundances of different metabolites has little quantifiable meaning (Figure 1), making it difficult to integrate metabolomics data into many analytical tools, including computational modeling frameworks. When quantification of only a few metabolites is required, researchers can use chemical standards to measure absolute concentrations. Unfortunately, chemical standards can be expensive, time-consuming to run, and unavailable for many metabolites[30-32]. Developing a platform for inferring absolute concentrations without the use of chemical standards would be impactful for incorporating metabolomics data into

both present and future metabolic tools.

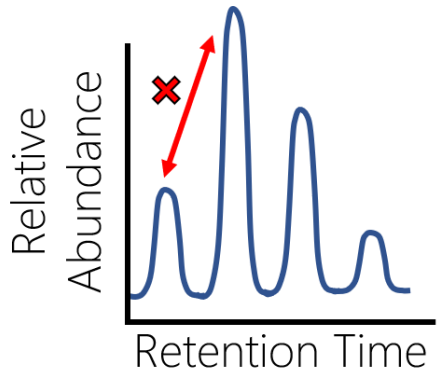

Figure 1: Example of a chromatogram.
Peaks in a chromatogram ideally represent different metabolites. Because metabolites are often measured in terms of relative abundances, the quantities of different metabolites cannot be directly compared.

1.3 Metabolic Modeling Approaches

There are two fundamental class of approaches researchers use to model metabolism[33]. Constraint-based models use linear programs to efficiently model metabolic systems, while ordinary differential equation-based models use detailed kinetic equations to accurately model reactions. The advantages and disadvantages of these approaches are discussed in detail in this section.

1.3.1 Constraint-based modeling

Constraint-based models (CBMs) are arguably the most popular metabolic modeling approach because they can be easily developed for any system in which the stoichiometry of reactions is known. The quintessential CBM method in metabolic modeling is flux balance analysis (FBA)[34]. FBA assumes that the studied system is at

steady-state and therefore the metabolite concentrations do not change over time. This assumption forces the influxes and effluxes of each metabolite to cancel out, which leads to a system of linear mass balance equations based on the stoichiometry of the system. The fluxes within this system of equations can be easily calculated using linear algebra tools without needing to estimate any kinetic parameters, one of the most appealing aspects of FBA. In biological systems, these systems of stoichiometric equations are generally underdetermined (i.e. there are more unknown fluxes than metabolites), leading to an infinite number of possible flux solutions. To overcome this obstacle, FBA implements constraints on the fluxes and most importantly, an objective function. The objective function is typically some hypothesized biological goal of the organism (i.e. maximizing biomass or ATP production), reflecting some evolutionary pressure for cell survival[35]. Together, the system of mass balance equations, flux constraints, and objective function create a linear program (LP), a mathematical optimization problem that can be solved efficiently even for large-scale problems, which is a key reason why CBMs are popular for modeling large systems. The formulation for FBA can be written as:

$$\max c^T v$$
$$s.t. \quad S \cdot v = 0 \qquad \text{(Equation 1)}$$
$$v_{LB} < v < v_{UB}$$

where *c* is a vector of weights that determine the objective function, *v* is the flux vector, *S* is the stoichiometric matrix that describes the inflow and outflow of fluxes for each

metabolite, and v_{LB} and v_{UB} are vectors of the lower and upper bounds for each flux, respectively.

While FBA models are easy to develop for many biological systems due to their LP structure, their assumption of steady state precludes capturing any metabolite dynamics. The steady-state assumption is useful for understanding the metabolism of an organism under static conditions, but cells are generally not at steady-state due to constant changes in the extracellular environment. Nevertheless, FBA is one of the best modeling tools to efficiently study metabolism at any scale and predict how metabolic resources are generally being produced and used.

1.3.2 Examples of constraint-based models

As the most common CBM approach in metabolic modeling, FBA has led to the development of numerous frameworks that are heavily influenced by or are direct extensions of the original FBA. Each of these new modeling approaches aim to supplement the relatively simple structure and assumptions of FBA, often integrating new biological information that improves modeling accuracy. Energy balance analysis (EBA)[36], thermodynamic metabolic flux analysis (TMFA)[37], and network-embedded thermodynamic analysis (NET analysis)[38] have incorporated thermodynamic constraints into their frameworks to eliminate flux distributions that are not energetically feasible[39]. Other methods, such as Minimization of Metabolic Adjustment (MoMA)[40] and parsimonious flux balance analysis (pFBA)[41] find flux distributions that are least taxing to the cell. MoMA predicts flux distributions in the feasible search space that are closest to the wildtype flux profile after a gene knockout and pFBA optimizes the system's

objective function while concurrently minimizing the total flux.

There have also been several extensions of FBA that attempt to track dynamic changes in metabolic systems. Recently, unsteady-state flux balance analysis (uFBA)[42] demonstrated it could predict dynamic metabolic flux states in several systems by using a novel piecewise simulation method that determines if a system is at steady-state based on the rates of change of its metabolites. Dynamic Flux Balance Analysis (DFBA)[43] in particular has garnered a substantial amount of attention as one of the first FBA-based methods to attempt to capture metabolite dynamics. DFBA uses two formulations to track dynamics: the Dynamic Optimization Approach (DOA) and the Static Optimization Approach (SOA). Although the DOA method is more accurate, it contains many non-linear constraints that make it more computationally taxing to use. The SOA formulation is much simpler and can still model metabolite dynamics, but it is unable to incorporate regulatory information like DOA. DFBA was a significant step in bridging the gap between CBM and ODE-based models, but most researchers still prefer kinetic models when accurate representations of metabolic systems are required. Overcoming the limitations of DFBA will be critical toward efficiently developing dynamic metabolic models at the genome scale.

1.3.3 Ordinary differential equation-based modeling

While CBMs excel at developing relatively simple models for systems of all sizes, ODE-based models use detailed kinetic equations to accurately model biological reactions. These kinetic equations allow for ODE-based models to track changes in metabolite concentrations and fluxes, but it comes at a price. Each of these kinetic

equations typically include several kinetic parameters that need to be estimated if they are not known *a priori,* which is often the case. While parameter estimation is achievable when there are only tens of reactions, it quickly becomes infeasible as the size of the modeled system increases due to the ample amount of data necessary to identify parameters accurately. This scaling problem makes it difficult to create kinetic models at the genome scale and is why most large models consisting of hundreds or thousands of metabolites and fluxes are developed using CBMs.

1.3.4 Examples of ODE-based models

As one of the most studied organisms, there are several kinetic models that have been developed for *E. coli*[44-46]. These ODE-based models have been used to predict different growth rates due to genetic perturbations and the increased biosynthesis of various target molecules, which make these models attractive for metabolic engineering applications. There have also been numerous ODE-based models developed for *S. cerevisiae*[47, 48], as it is one of the most important eukaryotic species in metabolism research. Like *E. coli*, kinetic models of *S. cerevisiae* have been vital to metabolic engineering, especially in the study of ethanol production[49]. While *E. coli* and *S. cerevisiae* are certainly two of the most widely-studied and kinetically modeled microorganisms, other scientifically relevant systems, such as *Lactococcus lactis*[50] and *Pseudomonas putida*[51], have also been modeled using ODEs.

Many of these kinetic models focus on central carbon metabolism, as it arguably contains the most important pathways in metabolism and is small enough that parameter estimation is still feasible. While most ODE-based models are restricted to modest

sections of metabolism, there have been a few attempts at constructing genome-scale kinetic models. An ODE-based model of *E. coli* metabolism was recently developed using Michaelis-Menten based kinetics with ensemble modeling and genetic algorithms for parameter estimation and was shown to predict several hundred engineered strains more accurately than other modeling approaches[52]. There have also been efforts to create more generalized approaches for developing genome-scale kinetic models[53-55], but most still require lengthy parameter estimation steps, known kinetic constants, or are only useful near the reference state of the modeled system. Despite these attempts, there remains a divide between efficient modeling frameworks that are scalable and modeling approaches that capture metabolite dynamics.

1.3.5 Linear Kinetics-Dynamic Flux Balance Analysis

To bridge the gap between CBMs and ODE-based models, our group developed Linear Kinetics-Dynamic Flux Balance Analysis (LK-DFBA)[56]. LK-DFBA is a novel modeling framework that can track metabolite dynamics while maintaining an LP structure, which is a significant step toward efficiently modeling biological systems at the genome scale. LK-DFBA is inspired by the work presented in DFBA and attempts to combine the strengths of the DOA and SOA formulations. LK-DFBA introduces novel linear kinetics constraints that model the interaction between a controller metabolite and the target flux that it regulates. These constraints are the driving force behind metabolite accumulation and depletion by constraining the maximum reaction rates of fluxes based on the metabolite concentration at a given time. Both mass action and allosteric regulatory interactions are modeled by these linear kinetics constraints.

We have previously demonstrated that LK-DFBA can recapitulate the training data of a synthetic system and a kinetic model of *E. coli*[44]. However, LK-DFBA has not been used to predict phenotypes when using different initial metabolite concentrations or introducing perturbations to pathways in the system. Before LK-DFBA can become a prominent metabolic modeling framework, it needs to at least be validated on data that is not used to train the actual model. Furthermore, one key area that could be improved in LK-DFBA is the construction of the kinetics constraints. In the original framework, the kinetics constraints are crude linear approximations of interactions between metabolites and fluxes. New kinetics constraints that better capture biological phenomena could lead to an increase in modeling accuracy.

1.4 Regulation in Metabolic Systems

The only information classical FBA requires is the stoichiometric topology of the system, an objective function, and flux constraints to build a metabolic model. However, in many cases this is not enough to construct an accurate representation of the system. One important feature of metabolism that FBA does not account for is regulation within an organism. Regulation can significantly impact the rate of reactions and must be incorporated when developing accurate metabolic models.

1.4.1 Transcriptional regulation

Transcriptional regulation is one of the most well-known and widely-studied forms of regulation in biological systems. In transcriptional regulation, transcription of DNA to RNA is controlled by regulators, such as transcription factors, that can increase

or decrease the amount of gene expression. Because gene expression ultimately leads to enzyme production (or lack thereof), transcriptional regulation can be an important factor in determining the rate of some metabolic reactions. However, because these transcription factors or other transcriptional regulatory proteins do not directly modulate enzyme activity, the timescale of these regulatory effects may not be immediately obvious[57, 58].

1.4.2 Allosteric regulation

While transcriptional regulation occurs at the DNA level, allosteric regulation takes place at the metabolite level. Allosteric regulation results from a regulator, often a small molecule, interacting with a protein, such as an enzyme, at a location other than its active site[59]. Like transcription factors in transcriptional regulation, allosteric regulators can inhibit or induce their target. Besides acting as the primary substrate in many different metabolic reactions, metabolites are also often allosteric regulators in other reactions (Figure 2). Because allosteric regulation occurs at the metabolite level, it directly impacts the metabolic state of the system and occurs on timescales much faster than transcriptional regulation[58], which makes it especially critical in understanding the metabolism of different organisms.

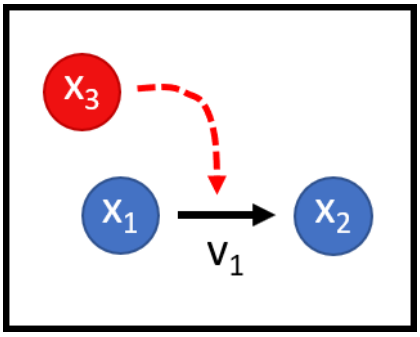

Figure 2: Example of metabolite-dependent allosteric regulation.
Metabolite x_1 is the substrate of the reaction flux v_1 and metabolite x_2 is the product. Unlike x_1, metabolite x_3 is not consumed by the reaction but instead acts as a regulator that inhibits the reaction rate.

1.4.3 Constraint-based modeling frameworks that integrate regulation

There have been a few iterations of FBA that try to account for regulation in their frameworks. Regulatory flux balance analysis (rFBA)[60], integrated flux balance analysis (iFBA)[61], and steady-state regulatory flux balance analysis (SR-FBA)[62] are three methods created to incorporate transcriptional regulation into the original FBA formulation. Each uses Boolean notation to designate whether a gene is active or not. These frameworks were found to more accurately predict different phenotypes and better understand how transcriptional regulation affects metabolism. When implementing transcriptional regulation into such frameworks, pseudo "time delays" are often used to demonstrate the indirect impact of transcriptional regulation on metabolic activity.

Despite its importance, allosteric regulation has not been integrated into CBMs as often. In ODE-based models, allosteric regulation can be readily implemented in a reaction by adding extra parameters and variables to its kinetic equation. There have only been a few attempts in the literature to integrate allosteric regulation with FBA. Allosteric Regulation FBA (arFBA)[63] is an extension of pFBA that modifies the objective function to implement allosteric regulation. This new objective function includes terms that set the

flux ratio equal to the turnover rate (of metabolites that regulate the flux) ratio. One downside of using this method is that it assumes that the ratios of the flux and turnover rate are exactly equal to each other, providing no flexibility. In LK-DFBA we have implemented allosteric regulation within the linear kinetics constraints.

1.4.4 Determining the topology of allosteric regulation

Although allosteric regulation is prevalent in most biological systems, the regulatory topology of different metabolic systems is often unknown. In contrast to the stoichiometry of metabolic reactions, which is relatively well-conserved across species[64], regulation can vary greatly. It can be difficult to experimentally identify the regulatory structure of a system because of the vast number of metabolites that could act as regulators for each reaction in metabolism. Without knowing how metabolic reactions are regulated, it is challenging to accurately model metabolism. As transcriptomics and genomics are currently much more mature fields than metabolomics and fluxomics, there have been many computational frameworks developed to map out transcriptional regulation in cells[65-72].

Computational methods for identifying allosteric regulation have been less common, though there are a few approaches that have been developed. Link et al. used ensemble modeling to establish the most likely regulatory structures of reactions[58]. In another approach, systematic identification of meaningful metabolic enzyme regulation (SIMMER)[73] uses non-linear optimization to fit data to simple Michaelis-Menten kinetic equations and determine if additional regulatory elements are required to sufficiently explain the data. Both of these methods require estimation of parameters, which can be

difficult to obtain for lesser-studied organisms. New methods for determining the topology of allosteric regulation are highly desirable.

Machine learning has become a popular approach for finding patterns within large datasets and creating predictive models using artificial intelligence. By training machine learning algorithms on "training data," models have been developed to predict protein structures, gene function, and gene regulation[74]. In the context of metabolomics, machine learning has been previously used to impute missing values[75] and discover new biomarkers[76], but it has not been used to identify the regulatory topology of metabolic systems. Applying machine learning to discover new regulatory interactions could enable development of accurate metabolic representations of systems where regulation has been poorly characterized.

1.5 Thesis overview

With applications in medicine, agriculture, energy, and more, the importance of studying metabolic systems is evident. Researchers have demonstrated that combining metabolomics with computational modeling has led to new insight into metabolic systems and it will continue to be a critical step toward understanding metabolism at the genome scale. Here, I have discussed several key areas in metabolic modeling that need to improve to create a streamlined process for modeling metabolic systems given only raw metabolomics data and the stoichiometry of the system. First, allosteric regulation in organisms that are not well-studied is often unknown, which can make it difficult to model metabolite dynamics. Second, while metabolomics is a direct readout of a system's metabolic state and therefore has promise in providing a plethora of data for

computational tools, any quantification of metabolites is often in terms of relative abundances, which makes it difficult to compare different metabolites to each other. Finally, although CBMs are more efficient and scalable in modeling different biological systems compared to ODE-based frameworks, they lack the ability to capture metabolite dynamics. LK-DFBA is a new framework that combines the advantages of both CBMs and ODE-based models, but there are still several areas that must be improved before it can become an invaluable tool in the metabolic modeling community. In this thesis, I aim to address each of these concerns and create a cohesive workflow for developing predictive metabolic models.

CHAPTER 2: A Stepwise Machine Learning Framework for Predicting Metabolite-dependent Regulatory Interactions

2.1 Background

Biochemists have amassed a large amount of knowledge about the topology of the chemical reaction network that cells use to transform nutrients into energy and the building blocks for more cells, collectively known as "metabolism". The substrates, products, and cofactors for hundreds of reactions have been elucidated, from the most central pathways like glycolysis to more distant pathways for the biosynthesis of uncommon metabolites. Many of these pathways are extremely well-conserved across the tree of life[64], with the basics of central carbon metabolism being quite similar from bacteria to humans. What varies much more greatly across species, and what allows such diverse metabolic phenotypes to arise from such otherwise similar reaction networks, is the regulation and utilization of the reactions in those networks. However, this regulation, despite its major importance in the function and diversity of life, is nowhere near as well understood as the topology of the metabolic network[46]. This is especially true for the direct regulation of reactions by metabolites, which is particularly poorly characterized compared to some other levels of regulation like transcriptional regulation. This is in large part due to the difficulty in experimental characterization of direct regulation by metabolites.

One critical form of direct regulation of metabolic reactions (and arguably the most common) is allosteric regulation, where a regulator and a protein (in this case an enzyme) interact at a location other than the active site[77]. In this mechanism, a metabolite

that is not the primary substrate of an enzyme binds to that enzyme and inhibits or promotes the reaction rate, most typically via an induced change in protein conformation. While metabolite levels can affect processes on the genome, transcriptome, and proteome levels[78], metabolite-dependent regulation of enzyme reaction rates is extremely important because it results in the control of reactions on a short timescale (less than 30 seconds) due to the direct interaction between metabolite and enzyme rather than requiring intermediate steps like transcription to effect changes[58, 63]. Their prevalence in metabolic systems makes it vital to account for these regulatory interactions to create accurate metabolic models.

Metabolic models that use only the known stoichiometry of the system and exclude metabolite-dependent regulation often have extremely limited accuracy. Machado et al. showed that including allosteric regulation in a model of *E. coli* is vital for predicting flux dynamics and can reveal "metabolic hubs," where a metabolite is connected to many reactions instead of only the few found in the stoichiometric topology[63]. Despite its prevalence and importance, the exact structure of this regulatory network (which metabolites regulate which fluxes) is typically unknown in all but the best-studied metabolic pathways in the best-studied organisms. With hundreds of metabolites and hundreds of fluxes in any given metabolic network and no effective high-throughput methods for finding metabolite-protein interactions (compared to, for example, protein-protein interactions[79, 80]), the space of possible regulatory interactions is too vast to experimentally explore[81].

As discussed at the beginning of this thesis, there have only been a few computational approaches to identify metabolite-dependent regulatory interactions of

enzymes in metabolic systems. Link et al.[58] used dynamic metabolite data to fit an ensemble of kinetic models with different putative regulatory interactions to rank which interactions contributed the most to fitting accuracy. Another approach by Hackett et al., named SIMMER[73], estimated kinetic parameters using non-linear optimization to establish if all reactions in a system could be sufficiently explained by Michaelis-Menten kinetics or if additional allosteric parameters were required. While these computational approaches are invaluable in saving time and costs for laboratory experiments, both methods rely on sampling[58] or estimating[73] kinetic parameters, which can be computationally taxing. An approach for identifying metabolite-dependent regulatory interactions without requiring kinetic parameters would be extremely useful for systems biology modeling. Although approaches using protein docking, such as AlloFinder[82], are promising for future *ab initio* prediction of regulatory interactions, current limitations in the accuracy of molecular simulations make systems-scale exploration of allosteric interaction space challenging and motivate a desire for approaches that can exploit increasingly widely available experimental datasets for these purposes.

Here, we present a new machine learning approach for Stepwise Classification Of Unknown Regulation (SCOUR) that leverages metabolomics and fluxomics data to predict likely metabolite-dependent regulatory interactions. SCOUR uses a stepwise process that focuses on identifying reactions controlled by one, two, or three metabolites. While SCOUR benefits from stepwise, serial inference of these increasingly complex interactions, each step is independent, uses different classification features, and can be performed without the others. Importantly, the classification task that SCOUR looks to address typically has insufficient training data available to generate useful models, so we

devise a strategy we refer to as "autogeneration" that we use to create sufficient data to train the models. We test our framework on two synthetic model networks, as well as on models of *S. cerevisiae* and *E. coli* metabolism, to show that SCOUR can be used on a variety of systems. Applying SCOUR to poorly-studied organisms has the potential to enable discovery of previously unknown regulatory interactions that are key to developing accurate and predictive metabolic models.

2.2 Methods for Predicting Regulatory Interactions

In this work, we examined four metabolic networks of varying size and complexity: two synthetic model networks and two biological systems. We simulated each metabolic network with fifteen sets of randomly generated metabolite concentration initial conditions (except for the first set of initial conditions for the biological systems, which were kept at their original values) to produce fifteen sets of metabolite concentration and flux data used in the testing sets of SCOUR. Each metabolic network is described in detail below.

2.2.1 Synthetic model networks

To initially test and evaluate SCOUR, we created two small synthetic model networks. The Smaller Synthetic Model (Figure 3A) contains six metabolites and six reactions, while the Bigger Synthetic Model (Figure 3B) contains ten metabolites and ten fluxes. Synthetic systems of these sizes are small and simple enough to easily assess the performance of SCOUR while developing the framework, but large enough to emulate behavior of metabolite-flux interactions in biological systems. In both models, the influx

(v_1) is a constant flux that is not controlled by any metabolites and is not considered when using SCOUR. Both models contain reactions controlled by one, two, or three metabolites, including both positive and negative regulatory interactions. Table 3 summarizes the number of each type of interaction in each model. The network dynamics were defined using Biochemical Systems Theory (BST) equations using power law kinetics for reaction rates[83], with mass action parameters randomly assigned between 0.1 and 1 and regulation parameters randomly assigned between 0.1 and 1 for positive regulatory interactions and between -1 and -0.1 for negative regulatory interactions. Each model was simulated for 10 seconds to generate synthetic data.

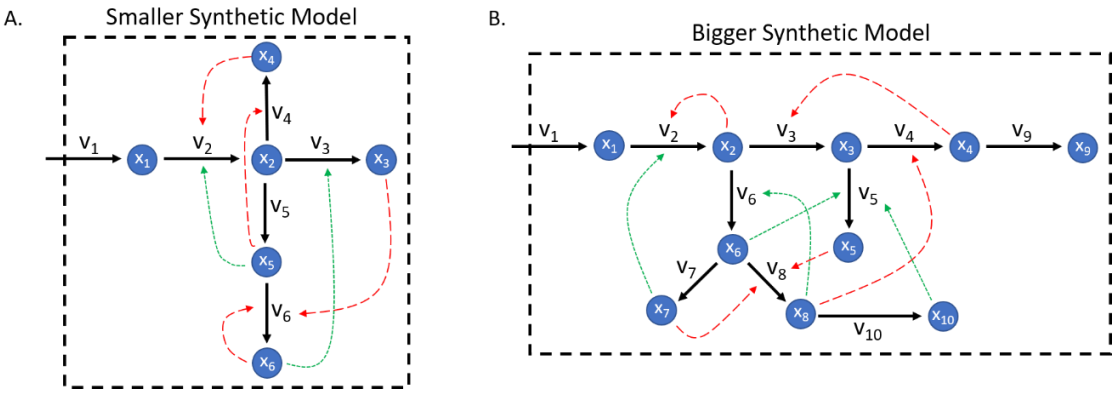

Figure 3: **Synthetic systems tested with SCOUR.**
Two synthetic model networks created using BST frameworks to generate *in silico* metabolomics and fluxomics data. x_i represent metabolites, v_i represent reaction fluxes (solid black lines), long-dashed red lines represent regulatory behavior that causes inhibition, and short-dashed green lines represent regulatory behavior that increases activity.

2.2.2 Biological models

To test SCOUR on more biologically relevant systems, we examined a model of glycolysis in *Saccharomyces cerevisiae*[47] and a model of central carbon metabolism in

Escherichia coli[44]. The *S. cerevisiae* model contains 22 metabolites and 24 reactions, while the *E. coli* model contains 18 metabolites and 48 reactions. We used the previously published kinetic equations and parameters for these systems; in both cases, the mathematical forms of the rate expressions include Michaelis-Menten, Hill, and mass action kinetics. Data for both biological systems were produced by reconstructing the ODE models in MATLAB and simulating the *S. cerevisiae* and *E. coli* models over 60 seconds and 10 seconds, respectively.

In the *S. cerevisiae* model, the fluxes for glucose mixed flow to extracellular medium and cyanide flow are constant and not controlled by any metabolites (Table 1). Likewise, in the *E. coli* model, the fluxes for glucose kinetics, murein synthesis, tryptophan synthesis, and methionine synthesis are constant and not controlled by any metabolites (Table 2). As in the synthetic models, both the *S. cerevisiae* and *E. coli* models include reactions controlled by one, two, or three metabolites, although they also have reactions controlled by four metabolites. Table 3 summarizes the number of each type of interaction in each of the four systems. Because both biological models have significantly more metabolites and reactions than the synthetic models, the number of possible interactions that need to be considered is substantially greater.

Table 1: List of controller metabolites and target fluxes in *S. cerevisiae* model.

Controller metabolite(s)	**Target flux**
N/A	Glucose mixed flow to extracellular medium
Extracellular glucose	Glucose uptake
Cytosolic glucose, ATP	Hexokinase
Glucose-6-phosphate, fructose-6-phosphate	Phosphoglucoisomerase
ATP, fructose-6-phosphate, AMP	Phosphofructokinase

Table 1 (continued)

Fructose 1,6-bisphosphate, glyceraldehyde 3-phosphate, dihydroxyacetone phosphate	Aldolase
Dihydroxyacetone phosphate, glyceraldehyde 3-phosphate	Triosephosphate isomerase
Glucose-6-Phosphate, 1,3-bisphosphoglycerate, NADH	Glyceraldehyde 3-phosphate dehydrogenase
ADP, 1,3-bisphosphoglycerate, phosphoenolpyruvate, ATP	Phosphoenolpyruvate synthesis
ADP, phosphoenolpyruvate	Pyruvate kinase
Pyruvate	Pyruvate decarboxylase
NADH, acetaldehyde	Alcohol dehydrogenase
Ethanol, extracellular ethanol	Ethanol out
Extracellular ethanol	Ethanol flow
Dihydroxyacetone phosphate, NADH, NAD	Glycerol synthesis
Glycerol	Glycerol out
Glycerol, extracellular glycerol	Glycerol flow
Acetaldehyde, extracellular acetaldehyde	Acetaldehyde out
Extracellular acetaldehyde	Acetaldehyde flow
Extracellular acetaldehyde, extracellular cyanide	Cyanide-acetaldehyde flow
N/A	Cyanide flow
ATP, glucose-6-phosphate	Storage
ATP	ATP consumption
AMP, ADP, ATP	Adenylate kinase

Table 2: List of controller metabolites and target fluxes in *E. coli* model.

Controller metabolite(s)	**Target flux**
Extracellular glucose	Extracellular glucose kinetics
Glucose-6-phosphate, pyruvate, extracellular glucose, phosphoenolpyruvate	Phosphotransferase system
Fructose-6-phosphate, 6-phosphogluconate, glucose-6-phosphate	Glucose-6-phosphate isomerase
Glucose-1-phosphate, glucose-6-phosphate	Phosphoglucomutase
Glucose-6-phosphate	Glucose-6-phosphate dehydrogenase
Phosphoenolpyruvate, fructose-6-phosphate	Phosphofructokinase
Erythrose-4-phosphate, fructose-6-phosphate, glyceraldehyde-3-phosphate, sedoheptulose-7-phosphate	Transaldolase

Table 2 (continued)

Glyceraldehyde-3-phosphate, sedoheptulose-7-phosphate, ribose-5-phosphate, xylulose-5-phosphate	Transketolase a
Fructose-6-phosphate, glyceraldehyde-3-phosphate, erythrose-4-phosphate, xylulose-5-phosphate	Transketolase b
N/A	Mureine synthesis
Dihydroxyacetonephosphate, glyceraldehyde-3-phosphate, fructose-1,6-bisphosphate	Aldolase
1,3-diphosphophoglycerate, glyceraldehyde-3-phosphate	Glyceraldehyde-3-phosphate dehydrogenase
Glyceraldehyde-3-phosphate, dihydroxyacetonephosphate	Triosephosphate isomerase
N/A	Tryptophan synthesis
Dihydroxyacetonephosphate	Glycerol-3-phosphate dehydrogenase
3-Phosphoglycerate, 1,3-diphosphophoglycerate	Phosphoglycerate kinase
3-Phosphoglycerate	Serine synthesis
2-Phosphoglycerate, 3-phosphoglycerate	Phosphoglycerate mutase
Phosphoenolpyruvate , 2-phosphoglycerate	Enolase
Fructose-1,6-bisphosphate, phosphoenolpyruvate	Pyruvate kinase
Fructose-1,6-bisphosphate, phosphoenolpyruvate	PEP carboxylase
Phosphoenolpyruvate	Synthesis 1
Pyruvate	Synthesis 2
Erythrose-4-phosphate, phosphoenolpyruvate	DAHP synthesis
Pyruvate	Pyruvate dehydrogenase
N/A	Methionine synthesis
6-Phosphogluconate	6-Phosphogluconate dehydrogenase
Ribose-5-phosphate, ribulose-5-phosphate	Ribose-phosphate isomerase
Xylulose-5-phosphate, ribulose-5-phosphate	Ribulose-phosphate epimerase
Ribose-5-phosphate	Ribose phosphate pyrophosphokinase
Fructose-1,6-bisphosphate, glucose-1-phosphate	Glucose-1-phosphate adenyltransferase
Glucose-6-phosphate	G6P degradation
Fructose-6-phosphate	F6P degradation
Fructose-1,6-bisphosphate	FDP degradation
Glyceraldehyde-3-phosphate	GAP degradation
Dihydroxyacetonephosphate	DHAP degradation

Table 2 (continued)

1,3-diphosphosphoglycerate	PGP degradation
3-phosphoglycerate	PG3 degradation
2-phosphoglycerate	PG2 degradation
Phosphoenolpyruvate	PEP degradation
Pyruvate	Pyruvate dilution
6-phosphogluconate	PG dilution
Ribulose-5-phosphate	Ribu5P dilution
Xylulose-5-phosphate	XYL5P dilution
Sedoheptulose-7-phosphate	SED7P dilution
Ribose-5-phosphate	Rib5P dilution
Erythrose-4-phosphate	E4P dilution
Glucose-1-phosphate	G1P dilution

Table 3: The number of *n*-controller metabolite interactions that exist in or are possible for each model.

For possible interactions, the first number assumes that regulatory interactions are correctly identified at each step of the framework and are removed from consideration for higher-order interactions. The number in parentheses is the total number of possible interactions if a stepwise framework were not used, illustrating the significant decrease in the number of interactions to be assessed in a stepwise framework.

	Smaller Synthetic Model	Bigger Synthetic Model	*S. cerevisiae*	*E. coli*
# of 1-controller interactions	1	3	5	25
# of possible 1-controller interactions	5	9	12	39
# of 2-controller interactions	2	3	10	13
# of possible 2-controller interactions	20 (25)	54 (81)	157 (262)	243 (668)
# of 3-controller interactions	2	3	5	2
# of possible 3-controller interactions	20 (50)	108 (324)	520 (2720)	336 (5384)
# of 4-controller interactions	N/A	N/A	2	4
# of possible 4-controller interactions	N/A	N/A	380 (17860)	480 (27120)

2.2.3 Autogenerated training data

Machine learning models must be trained using data that are broadly representative of the input data they are likely to encounter, which often entails using datasets that are as large as possible. In metabolism, there is a wide variety of metabolic reactions with disparate mechanisms and functional behaviors (e.g. bi-bi sequential reactions vs. ping-pong reactions)[84] or that are controlled by a different number of metabolites. However, appropriate training data for many of these possible situations are sometimes not available at all, let alone in sufficient quantity to enable machine learning model training. Accordingly, we chose to generate hundreds of artificial interactions to use as training data in an approach we refer to as "autogeneration". While the practice of creating artificial training data has been used in other machine learning contexts before[85-88], to the best of our knowledge it has not been used in producing metabolite-dependent regulatory interaction data.

For the training datasets in each step (meaning for the 1-controller, 2-controller, and 3-controller metabolite interactions), we created 300 autogenerated interactions, each with 15 different initial conditions. The 300 interactions included a mixture of samples that resembled true positive and true negative interactions. To create the time course data for each controller metabolite, concentration profiles were created from damped sine wave functions with randomized parameters (Equation 2). x_i is the concentration of controller metabolite i, t is the simulation time, and A, λ, ω, φ, are the amplitude, decay constant, angular frequency, and phase angle of a damped sine wave:

$$x_i = A_i e^{-\lambda_i t}(\cos(\omega_i t + \varphi_i) + \sin(\omega_i t + \varphi_i)) \qquad \text{(Equation 2)}$$

These concentration profiles were then used as input into BST equations to calculate dynamic flux profiles (Equation 3). BST is an ordinary differential equation-based modeling framework for metabolic systems that uses power-law kinetics and is generalizable to many types of metabolic reactions[83]. Each BST equation also was assigned randomized parameters. v represents the target reaction flux (rate), x_i is the concentration of controller metabolite i, n is the number of controller metabolites that regulate the target flux, and α and β are the randomly assigned BST parameters.

$$v = \alpha \prod_{i=1}^{n} {x_i}^{\beta_i} \qquad \text{(Equation 3)}$$

After generating both controller metabolite and target flux data, interactions were randomly assigned as either true positives or true negatives. If the interaction was labeled as a true positive, the correct sets of simulated controller metabolite profiles and corresponding calculated target flux profiles were used when calculating machine learning model features. However, if the interaction was labeled as a true negative, another set of metabolite concentration time course profiles would be generated from new damped sine wave functions and be used with the original target flux data to calculate features. Because these new pseudo-controller metabolites were not used in the calculation of the target flux data, there should be minimal relationship between the metabolites and the target flux, yielding a "true negative" data point.

To emulate the percentage of true positive interactions in the models tested in this work, of the 300 interactions in each step, 40%, 5%, and 5% were randomly assigned as existing interactions in the one-, two-, and three-controller interaction inference steps,

respectively. Changes to these percentages are expected to shift the sensitivity and specificity of the framework, so it is important to base the training percentages on what is expected to be seen in the testing data based on existing biochemical knowledge.

This approach for autogenerating training data aims to circumvent the requirement for large dynamic metabolomics and fluxomics datasets to train the machine learning framework, which are currently not widely available on the scale that would be required. Because this autogeneration approach is independent of SCOUR, it can possibly be used for other computational methods that require an abundance of metabolic data.

2.2.4 Noise-added data

To generate noisy data that are more representative of what is expected to be acquired experimentally, we used two different sampling frequencies and two coefficients of variation (CoV) for randomly-added noise, for a total of four conditions. Sampling frequencies of 50 and 15 timepoints (nT) and CoVs of 0.05 and 0.15 were used, where a higher CoV represents more noise (experimental error). The number of timepoints and amount of added noise are reasonable values for what one could possibly expect from mass spectrometry data for metabolomics or fluxomics. Starting with noiseless data, each metabolite and flux value in each time course was replaced with a random value drawn from $N_{i,k} \sim (y_i(t_k), CoV \cdot y_i(t_k))$, where $y_i(t_k)$ is the value of species (metabolite or flux) i at timepoint k. For each timepoint, three noisy data values were generated to resemble triplicate samples, which is a common practice in metabolomics and fluxomics experiments.

2.2.5 Data pre-processing

For noisy data, we applied two different pre-processing steps to the data. For the one-controller metabolite interaction inference step of the framework, we used the median sample of the triplicate noisy data to calculate the features in the training and testing sets. For the two- and three-controller metabolite interaction inference steps, instead of using the medians, a moving Gaussian filter was applied to smooth the triplicate noisy data before calculating their features. The window size of the filter was chosen to be ¼ of the total simulation time, which was found to smooth the data without overfitting to the noise itself. While smoothing the noisy data for two- and three-controller metabolite interactions led to an increase in SCOUR's performance, it was detrimental for one-controller metabolite interactions. We found that a few of the one-controller metabolite interaction features were more sensitive to the smoothed data than the noisy median data and would cause greater variability in SCOUR's performance across repetitions.

2.2.6 Features

Each step of the framework contains different "features" (Table 4) used to predict whether a particular interaction is likely to be correct. These "features" are scalar-valued outputs of functions that quantify characteristics of concentration and flux profiles and the relationships between different profiles, which may thus indicate whether a given metabolite or set of metabolites regulates a given flux. Different features were used for the prediction of interactions controlled by different numbers of metabolites (i.e., one-controller vs. two-controller vs. three-controller metabolite interactions). This allows the

features to be customized to specific interaction types (and avoids the requirement that they must be valid or useful for all interaction types), which is expected to increase SCOUR's overall accuracy compared to using the same features for all steps. Features were designed using biochemical insight into how metabolites are known to interact with enzymes and how these interactions would manifest in concentration and flux profile data. For example, for the one-controller metabolite interaction step, the Spearman correlation between fluxes and metabolites was used as a feature, as reaction fluxes are expected to be highly (though not necessarily linearly) correlated with the metabolites that control them. Additionally, features were created based on the expectation that for every set of metabolites that completely defines an output flux, each possible set of metabolite input concentrations can only yield one single output of flux reaction rate. A list and description of all features used in SCOUR can be found in Table 4.

Table 4: **List of features for each step of the framework**

1-controller metabolite interaction features		
Feature name	**Description**	**Reasoning for feature**
Correlation	Spearman correlation between controller metabolite and target flux.	If a reaction flux is controlled by a single metabolite (most likely a mass action interaction), the Spearman correlation should be close to +1.
Curve fit	A second-degree polynomial is fit to the controller metabolite vs. target flux data and the adjusted R^2 value (adjusted for the number of coefficients in the polynomial model) is calculated.	A simple polynomial curve should fit the data reasonably well if a reaction is controlled by a single metabolite (e.g. if the data exhibits a Michaelis-Menten saturation curve).
Flux prediction	2/3 of the available controller metabolite data and target flux data are randomly selected to train a kNN regression model, with the flux data acting as the dependent variable. The remaining 1/3 of controller metabolite data is used with the kNN model to predict the remaining flux values and the prediction error is calculated. This process is repeated 3 times and the mean of the prediction errors is taken.	It is likely easier to make predictions (i.e. lower prediction error) if the controller metabolite and target flux belong to an existing interaction in the system and no other metabolites truly regulate the target flux.

Table 4 (continued)

CoV of data	The average CoV of the target flux is calculated at 10 evenly-spaced individual concentrations within the range of the controller metabolite. Because flux data may not be sampled at these evenly-spaced concentrations, flux data are linearly interpolated at these concentrations using the closest higher and lower concentrations with sampled flux data.	The CoV of the target flux should be low at each metabolite concentration for a one-controller metabolite interaction because there should be a single flux value for every concentration value.
2-controller metabolite interaction features		
Feature name	**Description**	**Reasoning for feature**
Functionality	Plot the two putative controller metabolites against each other for all 15 datasets produced from different initial conditions. Identify where the two controller metabolite concentrations are approximately equal to each other in two of the datasets by finding where the two datasets intersect with each other on the plot. This is accomplished by using the InterX function in MATLAB[89] that uses vectorization to determine intersection points between datasets. At these intersection points, linearly interpolate the flux data for each of the two datasets (using the scatteredInterpolant function in MATLAB for 3-D interpolation), calculate the difference of these two interpolated target flux values, and divide by the mean of the flux values to normalize. For all intersection points found, take the mean of all normalized differences between interpolated target fluxes.	For every input of controller metabolites, there should be a single output for the target flux (this is the definition of a mathematical function) if those metabolites are the only variables that interact with the reaction.
Surface fit	Fit a plane surface to the data of the two controller metabolite concentrations and the target flux and calculate the root mean square error of the fit against the data.	Because an existing interaction must maintain "functionality," the controller metabolite and target flux data should form some sort of surface. There will likely be a better fit when fitting a plane to data from an existing interaction than data from a non-existing interaction.
Flux prediction	Same as flux prediction feature for 1-controller metabolite interactions, except two metabolites are used to train and test the kNN model.	Same as flux prediction feature for 1-controller metabolite interactions.

Table 4 (continued)

Correlation with one metabolite constant	Plot one of the putative controller metabolites (x-axis) against the target flux (y-axis) for each of the fifteen datasets. Next, plot ten vertical lines that are evenly-spaced within the range of the controller metabolite that represent ten constant concentrations. For one vertical line, identify if and where the line intersects with the fifteen data sets using the InterX function and linearly interpolate flux data at these intersection points using the closest higher and lower concentrations with sampled flux data. Calculate the Spearman correlation between the second controller metabolite and interpolated target flux at these intersection points where the first controller metabolite is constant. Repeat for each of the ten vertical lines and calculate the mean of all correlations. Switch which metabolite is held constant and repeat the process. Take the lesser of the absolute values of the two mean correlations.	The correlation between one controller metabolite and target flux should be consistently close to +1 (activation) or -1 (inhibition) for any constant concentration value for the second metabolite. This assumes no high concentration effects, such as substrate inhibition. The lesser of the two absolute mean correlations is taken as it is the worst performing.
Curve fit with one metabolite constant	Plot one of the putative controller metabolites (x-axis) against the target flux (y-axis) for each of the fifteen datasets. Next, plot ten vertical lines that are evenly-spaced within the range of the controller metabolite that represent ten constant concentrations. For one vertical line, identify if and where the line intersects with the fifteen data sets using the InterX function and linearly interpolate flux data at these intersection points using the closest higher and lower concentrations with sampled flux data. Fit a second-order polynomial to the second controller metabolite and target flux data at these intersection points and calculate the root mean square error between the fit and the data. Repeat for each of the ten vertical lines and calculate the mean of all errors. Switch which metabolite is held constant and repeat the process. Take the greater of the absolute values of the two mean errors.	If one controller metabolite is constant, a simple polynomial on the second controller metabolite and target flux should fit well, similar to the curve fit feature for 1-controller metabolite interactions. The greater of the two absolute mean errors is taken as it is the worst performing.

Table 4 (continued)

3-controller metabolite interaction features		
Feature name	**Description**	**Reasoning for feature**
Hyperplane fit	Fit a hyperplane to the data of the three controller metabolite concentrations and the target flux and calculate the root mean square error of the fit against the data.	Same as surface fit feature for 2-controller metabolite interactions.
Functionality (percentage method)	For any two datasets generated from different initial conditions, find concentrations where the three putative controller metabolites are within 5% (noiseless) or 10% (noisy) across datasets. Calculate the CoV of the target flux for the two datasets at these points.	Same as functionality feature for 2-controller metabolite interactions.
Flux prediction	Same as flux prediction feature for 1-controller metabolite interactions, except three metabolites are used to train and test the kNN model.	Same as flux prediction feature for 1-controller metabolite interactions.
Functionality (rounding method)	For any two datasets generated from different initial conditions, find concentrations where the three putative controller metabolites are equal across datasets after rounding to the second decimal place. Calculate the percent of target flux values that are equal (within 0.002 error). The majority of metabolites in this work had mean concentrations on the order of 10^{-1} to 10^{1}, and the majority of fluxes had mean rates on the order of 10^{-2} to 10^{2}, making the chosen rounding precision and error threshold reasonable for this feature.	Same as functionality feature for 2-controller metabolite interactions. In this feature, the parameters used to determine equivalence (i.e. rounding precision and error threshold) are fixed and are not proportional to the concentration or flux data used, unlike in the percentage method. This difference allows the rounding method to be more sensitive when determining equivalence for concentration or flux data that have larger orders of magnitude (i.e. greater concentrations and fluxes will be considered equal in fewer cases than in the percentage method; if they are considered equal, it will be with higher confidence). The number of decimal places and error bounds can be adjusted depending on the user's preference for sensitivity.
Correlation with two metabolites constant	Same as correlation with one metabolite constant, except two metabolites are held constant and the Spearman correlation between the third metabolite and target flux is calculated.	Same as correlation with one metabolite constant.

2.2.7 Scaling of feature matrices

When assessing SCOUR's performance on noiseless data, we found that the feature matrices did not require any scaling because the range of values across features was relatively consistent due to the lack of outliers caused by noise. However, scaling the

training and testing feature matrices significantly improved SCOUR's performance on noisy datasets. Each feature in the feature matrices for two- and three-controller metabolites were scaled between 0 and 1. When scaling data in the one-controller metabolite feature matrices between 0 and 1, we found poor performance due to a high sensitivity to outliers in a few of the one-controller metabolite features. To solve this problem, we scaled the feature matrices for one-controller metabolite interactions so that the 20^{th} and 80^{th} percentiles of the data were scaled between 0 and 1, which diminished the effect of outliers on the machine learning algorithms. This technique is called robust scaling[90].

2.2.8 Machine learning stacking

Stacking is a technique used in machine learning to aggregate predictions made by multiple classification or regression algorithms[91]. The idea behind stacking is that some algorithms will be able to classify certain samples better than others, such that by combining information from multiple algorithms one can more accurately classify samples overall. In SCOUR, we used four machine learning algorithms in a stacking model.

To train the four algorithms in the first layer of the stacking process, an initial set of autogenerated data with known training labels was used for each algorithm. To train the metamodel in the second layer of the stacking process, a second set of autogenerated data was passed through the previously trained first layer models and the prediction outputs from the four original machine learning algorithms were used as inputs to train the metamodel, along with the known training labels of the second set of autogenerated

data. For this work, we chose to use a discriminant analysis classifier as the metamodel, as it was shown to perform well in consolidating information from the four algorithms in the first layer. A workflow of the stacking process is shown in Figure 4.

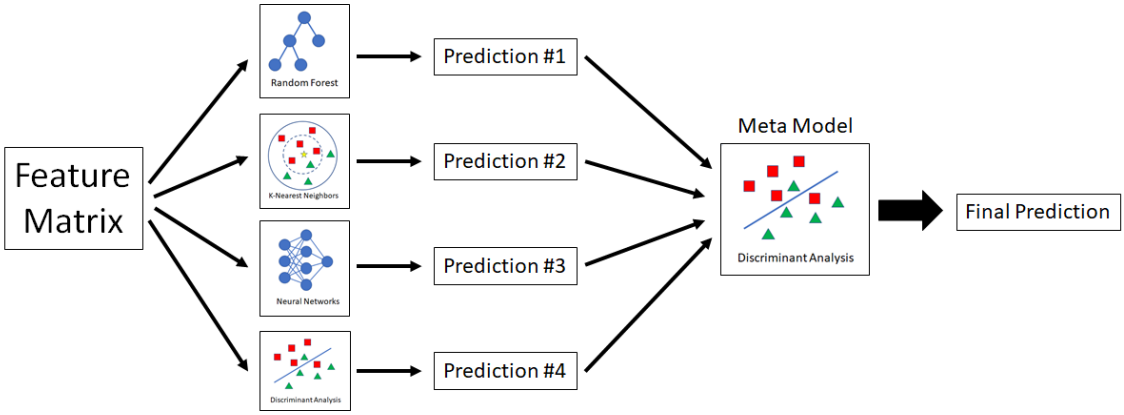

Figure 4: **Workflow of stacking process.**
After the feature matrix is calculated for all possible regulatory interactions, it is used on the first level of the stacking process as input for the four machine learning algorithms. The four resulting predictions are then used in the second level metamodel to produce a final prediction output for the framework.

2.2.9 Machine learning algorithms

The four machine learning algorithms used in the stacking model are random forest[92], k-nearest neighbors (kNN)[93, 94], shallow neural networks[95], and discriminant analysis[96]. Each of these algorithms are some of the most robust and commonly used machine learning approaches, but they are all fundamentally very different from one another. In SCOUR, we use kNN and discriminant analysis as binary classifiers: algorithms that can predict only two discrete labels. Random forest and neural networks are used as regression algorithms, where both predict continuous values. While most of these machine learning methods can be used as either discrete classifiers or regression models, we decided to have a mixture of these two types of algorithms because we believe that using a more diverse class of algorithms will enhance the stacking model and

prevent any potential bias toward algorithms that are very similar to one another. In the stacking process, the predictions from the four models were used as the input for a secondary metamodel (another discriminant analysis classifier) to give a final classification output for each regulatory interaction that was tested.

2.2.10 Stepwise approach

SCOUR uses a stepwise approach to identify different types of regulatory interactions at each step, beginning with the identification of one-controller metabolite interactions. First, two training datasets that consist of true positive (controlled by a single metabolite) and true negative (controlled by multiple metabolites) interactions are autogenerated for the two levels of the stacking model. The features described in Table 4 are calculated for each interaction in the first autogenerated dataset, which are used to train the first level of the stacking model. Next, the second level of the stacking model is trained using the feature matrix calculated from the second autogenerated dataset. Finally, the completely trained stacking model predicts whether or not each interaction in the testing dataset (comprised of the possible one-controller metabolite interactions in the system of interest) is controlled by a single metabolite. This process is repeated for predicting two- and three-controller metabolite interactions.

This stepwise approach has two key advantages. First, it allows for completely independent classification models and features that can be crafted for specifically identifying reactions that are controlled by one, two, or three metabolites. We found that developing a one-step platform for predicting multiple classes (i.e. reactions controlled by different numbers of metabolites) at once led to worse performance even when using

machine learning algorithms, such as random forest and neural networks, that can be tailored toward multiclass classification. Multiclass classification may have performed well if SCOUR was only classifying if a reaction is controlled by one, two, or three metabolites. However, because SCOUR is also trying to predict the exact controller metabolites that interact with a reaction flux, there is an additional layer of complexity that is easier to address with multiple binary classification models. The second advantage of using a stepwise approach is that after each step, fluxes whose regulatory status has already been identified are removed from consideration in the next step so that there are fewer interactions to be tested by the machine learning algorithms. This reduces the computation time of the entire stepwise process, reduces the chances of false positives, and allows subsequent steps and features to be more simply designed under the assumption that lower-order regulatory interactions will not be present in later steps. However, these advantages are at the risk of removing fluxes at a step earlier than when their true regulatory status could be identified. A comparison of the number of interactions that need to be tested whether or not the stepwise framework is used for each of the evaluated models can be found in Table 3. A schematic of SCOUR's workflow is shown in Figure 5.

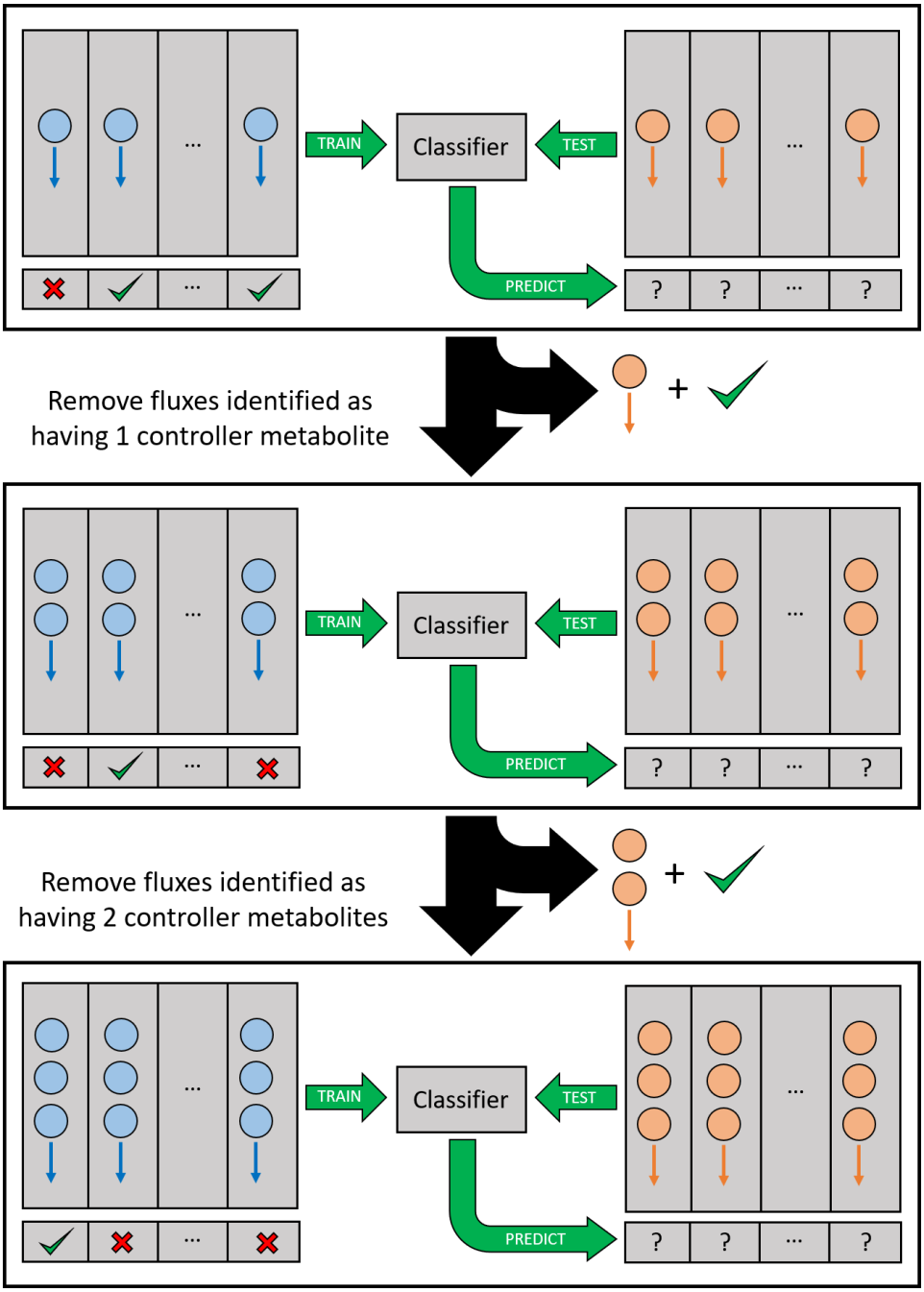

Figure 5: **Workflow of stepwise machine learning framework for identifying one-, two-, and three-controller metabolite interactions.**
Blue circles and arrows represent metabolites and the fluxes they might interact with, respectively, in the training set. Orange circles and arrows represent metabolites and the fluxes they might interact with, respectively, in the testing set. In each step, the training set is used to train the machine learning classifier for fluxes with a specific number of metabolite controllers, which is then applied to the testing set to predict which fluxes are in that category. Between each step of the workflow, fluxes that have been positively identified are removed from further consideration in the test set.

2.2.11 Framework performance metrics

To assess the performance of our framework in identifying different types of regulatory interactions, we evaluated four different metrics: accuracy, sensitivity, specificity, and positive predictive value (PPV). Accuracy is the percent of candidate regulatory interactions that are identified correctly as existing or not existing in a model. While accuracy can be a good metric if the classes (i.e. candidate interactions that are truly in the model vs. candidate interactions that are not in the model) are well-balanced, this is not the case for combinatorial consideration of potential metabolic regulatory interactions: there are many more candidate metabolite and reaction flux combinations than there are actual regulatory interactions in a given biological system. Sensitivity and specificity separate accuracy into two metrics that measure, respectively, the percent of positives (i.e. true regulatory interactions) that are identified correctly and percent of negatives (i.e. candidate regulatory interactions that are not actually in the model) that are identified correctly. PPV is the percentage of interactions predicted by the model that are true positives, an important metric to consider when one plans on experimental validation of predictions because it indicates how much effort is typically required for the validation of every newly discovered interaction. Exceedingly low PPVs are undesirable for predictions that are difficult to experimentally test, including metabolite-dependent regulation of reaction rates, because they signify that a large number of predicted interactions must be tested with these difficult experimental methods in order to find any true, validated interactions.

2.3 Results

2.3.1 Performance on noiseless data

When evaluating SCOUR on noiseless data, we found good overall predictive accuracy for both synthetic models (Figure 6A and Figure 6B). We trained SCOUR on 30 independent sets of noiseless autogenerated data to assess the sensitivity of the framework to different sets of autogenerated training data. The average sensitivities and specificities for all steps in SCOUR were above 88% for both models. PPVs were above 77% for predicted one- and two-controller metabolite interactions, and above 58% for predicted three-controller metabolite interactions.

We found similar results when testing SCOUR on noiseless data simulated from the *E. coli* and *S. cerevisiae* systems (Figure 6C and Figure 6D). As in the synthetic models, the PPV for both of these biological models decreased as the number of controller metabolites increased, though in a steeper fashion likely due to the increased complexity of these systems. The accuracy, sensitivity, and specificity in both biological models were still above 71% for all steps, and the PPVs for one- and two-controller metabolite interactions were all above 32%, despite the increase in model complexity. The low PPV for identification of three-controller metabolite interactions for both biological models (< 8%) despite high specificity (> 85%) was attributable to the highly imbalanced nature of the testing data. Out of the large number of candidate three-controller metabolite regulatory interactions that must be classified (Table 3), only a few are true positives and consequently there is an increased likelihood for false positive predictions. We note the large standard error of the mean for sensitivity in the *E. coli* model when identifying three-controller metabolite interactions. This is due to SCOUR

removing the fluxes of the two true positive interactions in a previous step, which leads to the sensitivity not being calculated in several of the repetitions (due to the absence of any true positives or false negatives).

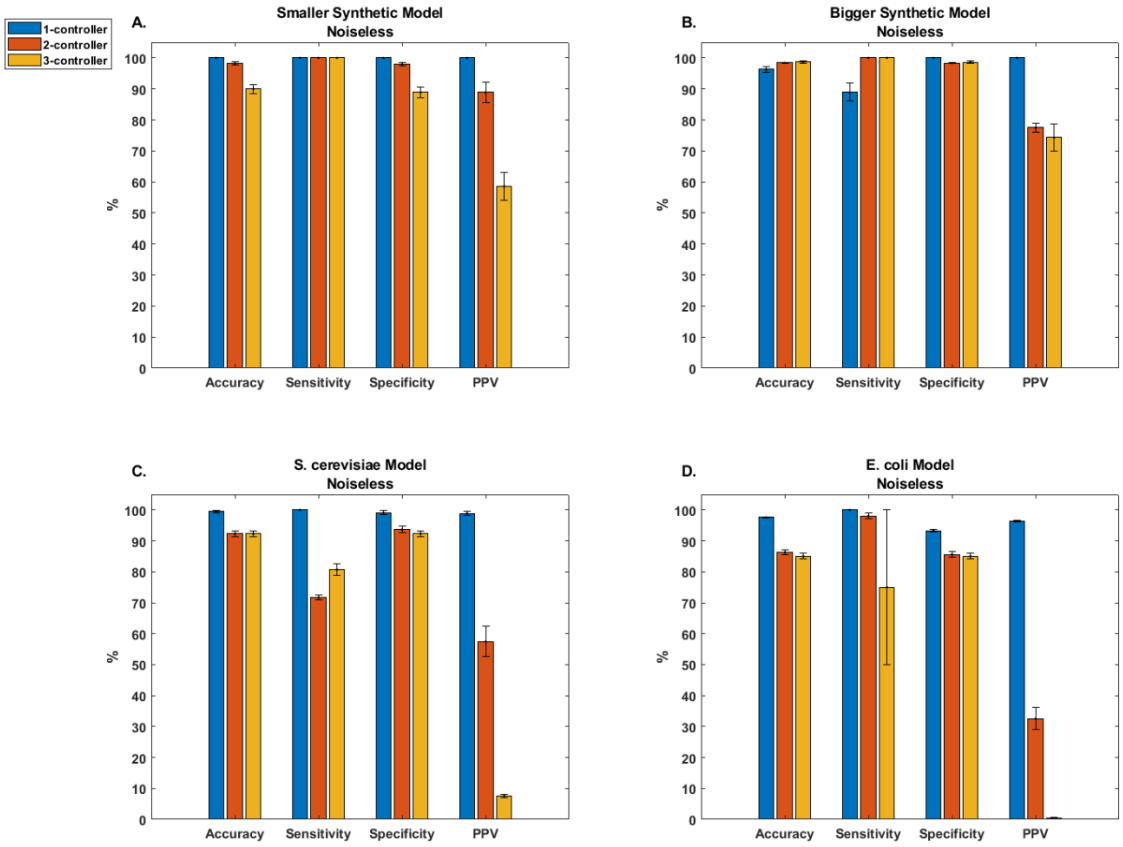

Figure 6: **SCOUR performance on synthetic and biological models using noiseless training and test data.**
Bar graphs for accuracy, sensitivity, specificity, and PPV for each step of SCOUR in each tested model. Error bars represent the standard error of the mean (n = 30 from independent autogenerated training replicates).

2.3.2 Performance on noisy data

While using noiseless data gives a sense for the framework's performance under ideal conditions, the realities of metabolomic and fluxomic experimental limitations can

lead to significant deviation from these idealized assumptions. To assess SCOUR's performance under more biologically relevant conditions, we examined two factors that need to be considered when using real metabolomics and fluxomics data: decreased experimental sampling frequency (and thus less information content to enable identification of true regulatory interactions) and increased experimental measurement noise. To give SCOUR a baseline performance level to compare to, we also created a classifier that randomly predicted whether a metabolite-flux interaction was a true positive or true negative interaction and used this to calculate a PPV at each step. Each interaction had a 50% chance of being classified as either a true positive or true negative in each step of the framework. For this random predictor, we assumed that the correct reaction fluxes were removed at each step, giving this classifier an advantage over our framework by greatly reducing the number of possible false positive interactions.

Assessment of SCOUR's performance on noisy data from the synthetic models (Figure 7A through Figure 7F) yielded similar trends to the results from noiseless data (Figure 6A and Figure 6B). For both decreased sampling frequency and increased experimental noise, SCOUR's overall accuracy unsurprisingly decreased, but still allowed for effective identification of many regulatory interactions in each model. In both synthetic models, there was an expected decrease in sensitivity and PPV with decreasing sampling frequency or increasing noise. As in the noiseless case, the PPV decreased for fluxes with more controller metabolites due to the increase in candidate regulatory interactions (and thus, an increase in possible false positive predictions) tested at each stage. In the most experimentally realistic scenario (nT = 15, CoV = 0.15), SCOUR still yielded PPVs that were acceptable for lab validation when classifying one- and two-

controller metabolite interactions (> 59% and > 18%, respectively). The mean PPVs for one- and two-controller metabolite interactions were also better than the random predictor in both synthetic models across all conditions and SCOUR outperformed the random predictor in most cases when classifying three-controller metabolite interactions.

The results from testing on biological models with noisy data (Figure 7G through Figure 7L) were similar to those from the synthetic models (Figure 7A through Figure 7F). For both the *S. cerevisiae* and *E. coli* models, the PPV was fairly consistent (with slight decreases) for any given interaction type across the increasingly challenging noisy conditions, while accuracy, sensitivity, and specificity sometimes exhibited slightly more variability across those conditions. For the *S. cerevisiae* model, the PPV remained high (> 69% on average) in all conditions for identification of one-controller metabolite interactions and was above 25% for identification of two-controller metabolite interactions. These PPVs for one- and two-controller metabolite interactions are sufficiently high enough if one wanted to experimentally validate these predictions to identify previously unknown interactions. For three-controller metabolite interactions, the PPV was below 12% for all conditions, which is not ideal from the standpoint of experimental practicality. In the *E. coli* model, the accuracy, sensitivity, specificity, and PPV were high for one-controller metabolite interactions in all conditions (> 92%), but the PPV dropped to less than 13% for two-controller metabolite interactions and was essentially 0% for three-controller metabolites (and a large standard error of the mean for sensitivity was observed, as in the noiseless condition in Figure 6D). This would make it challenging to experimentally validate the *E. coli* predictions for two- and three-controller metabolite interactions without a guided high-throughput approach.

Nevertheless, the PPVs for both biological systems when using SCOUR were on average better than the PPVs of the random predictor for one- and two-controller metabolite interactions for all conditions.

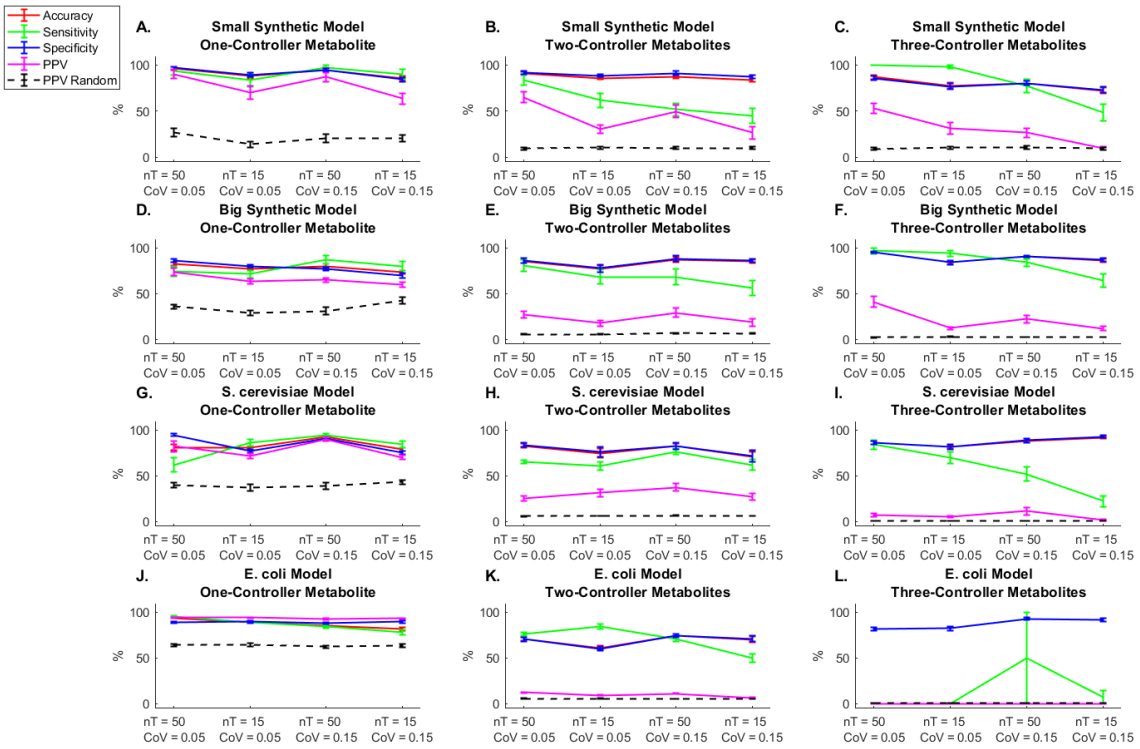

Figure 7: **SCOUR performance on synthetic and biological models using noisy and low sampling frequency training and test data.**
Solid lines represent accuracy, sensitivity, specificty, and PPV performance of SCOUR on each model for each step of the framework. Dashed lines represent the PPV if interactions were randomly classified. Error bars represent the standard error of the mean. (n = 30 from independent autogenerated training replicates).

2.4 Discussion

Our results indicate that SCOUR is a promising route for *in silico* prediction of metabolite-dependent regulation of metabolic fluxes using metabolomics and fluxomics data. On noiseless data, SCOUR predicts one- and two-controller regulatory interactions with high PPV in both synthetic and biological models, with three-controller interactions

also predicted extremely well in some systems. While the use of noisy data leads to an expected drop in performance, SCOUR still provides extremely high PPV for one-controller interactions in all systems and high (experimentally useful) PPV for the synthetic models and the *S. cerevisiae* model when predicting two-controller metabolite interactions. SCOUR's PPVs for these two steps greatly outperformed the PPVs of a random classifier in almost all cases. PPVs for three-controller metabolite interactions remained useful for the synthetic models but pushed the bounds of practical utility in the *S. cerevisiae* and *E. coli* models, likely attributable in large part to the combinatorial growth of the number of candidate interactions that must be tested and thus the concomitant growth in the number of false positives. Regardless of whether the three-controller interaction predictions are sufficient for experimental validation, the PPV in the *S. cerevisiae* model is significantly greater than the PPV for random classification for all noisy conditions except for the lowest sampling frequency and highest noise case (Figure 8). This suggests that SCOUR would still be helpful for identifying these types of interactions compared to indiscriminately testing all combinations of interactions as high-throughput guided experimental approaches are developed.

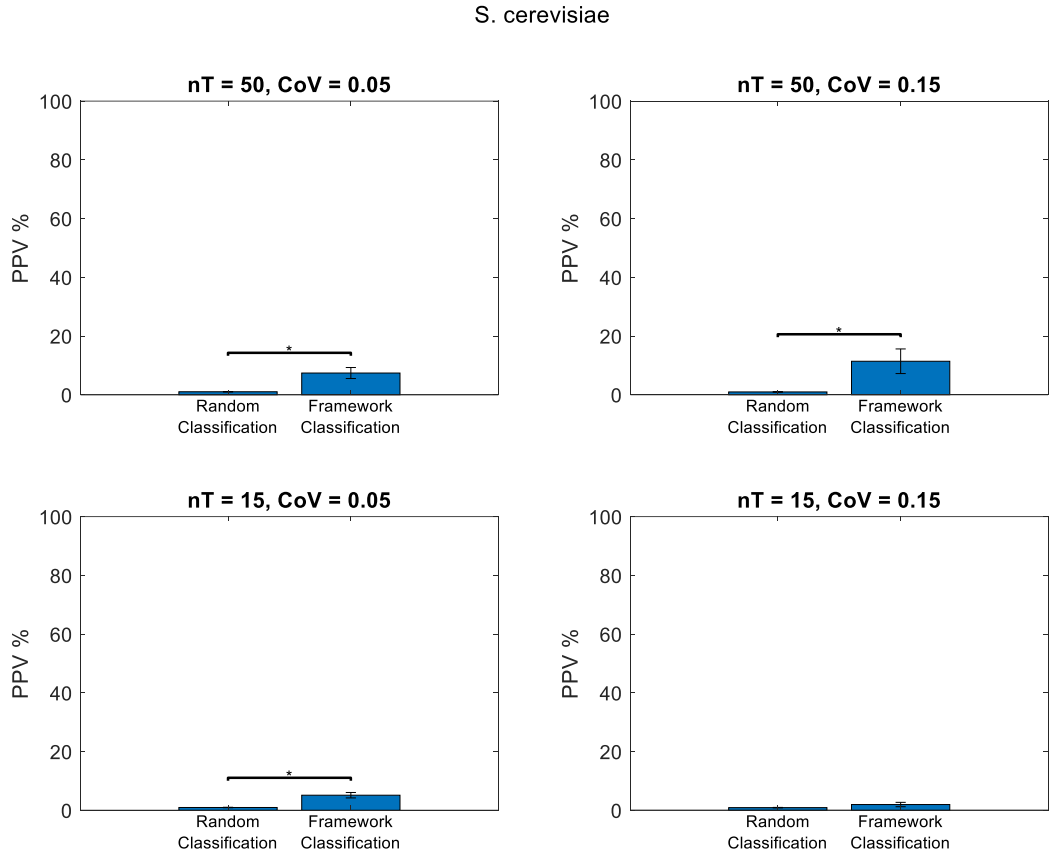

Figure 8: **SCOUR's PPV for 3-controller metabolite interaction predictions is significantly greater than a random classifier.**
While SCOUR yields low PPVs in the *S. cerevisiae* model for 3-controller metabolite interactions, these results are significantly better than random classification of 3-controller metabolite interactions for all conditions except for the case with the fewest timepoints and most noise (nT = 15, CoV = 0.15). We used a Wilcoxon rank-sum test (alpha = 0.05) to assess significance, as we found the distributions of the PPVs from SCOUR were not normal when using a Kolmogorov-Smirnov test (alpha = 0.05), except for the nT = 50, CoV = 0.05 condition. With appropriate guided high-throughput methods, SCOUR's predictions could still be useful for identifying these types of reactions.

We believe the unusually sharp decrease in PPV for the *E. coli* model between the one- and two-controller interaction predictions is largely attributable to two reasons. First, the *E. coli* model contains three two-controller metabolite interactions where the two controller metabolites are highly correlated with each other (> 99% correlation;

Figure 9). This presents an identifiability problem, with it being extremely difficult to decouple the effects of the two metabolites once even a small amount of noise is added to their data. This in turn affected the utility of several features in our machine learning models, which led to these three interactions rarely being identified by SCOUR and thus also led to lower sensitivities and PPVs. Second, as previously discussed, the large size of the *E. coli* model necessitates testing many candidate regulatory interactions. Even with relatively high specificity, the resulting false positives from these tests can suppress the PPV. This is a common problem found in other efforts to determine regulatory activity (or any work with imbalanced datasets), where one class (e.g. true negative interactions) significantly outnumbers the other class (e.g. true positive interactions)[73].

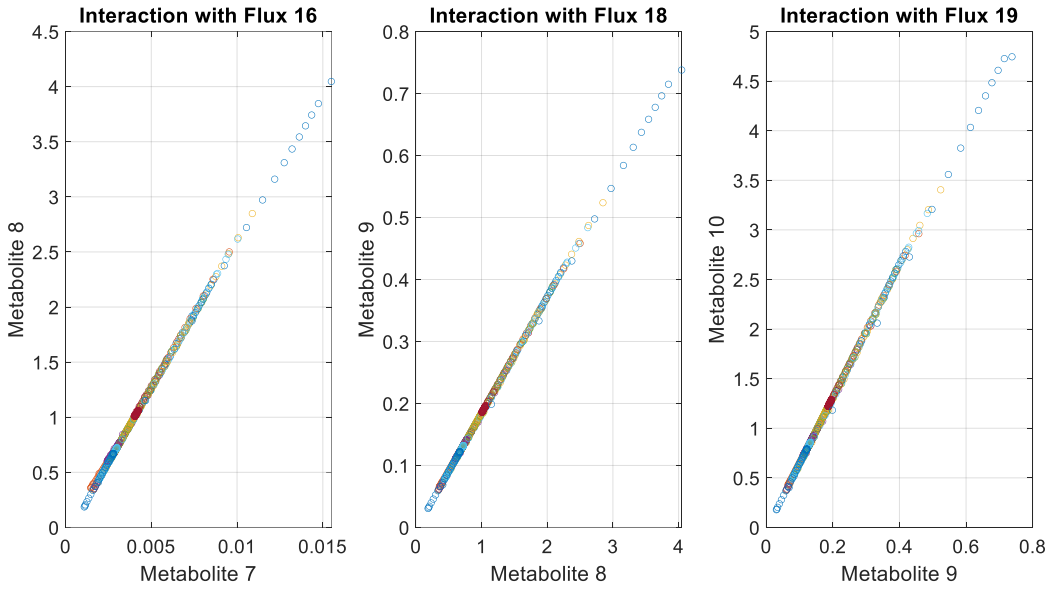

Figure 9: **The most common false negative two-controller metabolite interactions in the *E. coli* model.**
Three two-controller metabolite interactions were not correctly predicted across many conditions and replicates. The metabolites involved in these three interactions were highly correlated with each other, as indicated by these plots of noiseless concentration data for each of the pairs of metabolites. Each of the fifteen datasets with different initial conditions is represented by a different group of colored data. The scatter plots indicate extremely high correlation, which mitigated the utility of some of the features provided to the machine learning models (e.g. functionality and flux prediction). Inclusion of such interactions with highly correlated controller metabolites in the autogenerated data did not improve SCOUR's ability to identify these metabolites, especially once noise was introduced into the measurements.

Throughout the evaluation of SCOUR, we have relied on PPV as a performance metric because it is a valuable indicator for whether or not the predictions by SCOUR are worth experimentally validating. F1 score is another performance metric that is calculated from PPV and sensitivity and it is often used for imbalanced datasets, such as those found in this work. When using F1 score, we found that SCOUR still outperformed random classification of one- and two-controller metabolite interactions in all models under all noisy conditions evaluated (Figure 10). For three-controller metabolite interactions, the difference in performance between SCOUR and random classification is less clear and

we would once again conclude that it would be difficult to recommend lab validation of the predictions for these types of interactions without high-throughput guided methods. While F1 score is an important evaluation metric and still verifies that SCOUR is a useful platform for identifying one- and two-controller metabolite interactions, we argue that PPV is a more important criterion in the context of finding new regulatory interactions because it indicates how many undiscovered interactions could be identified out of those predicted by SCOUR, regardless of how many true positive interactions exist.

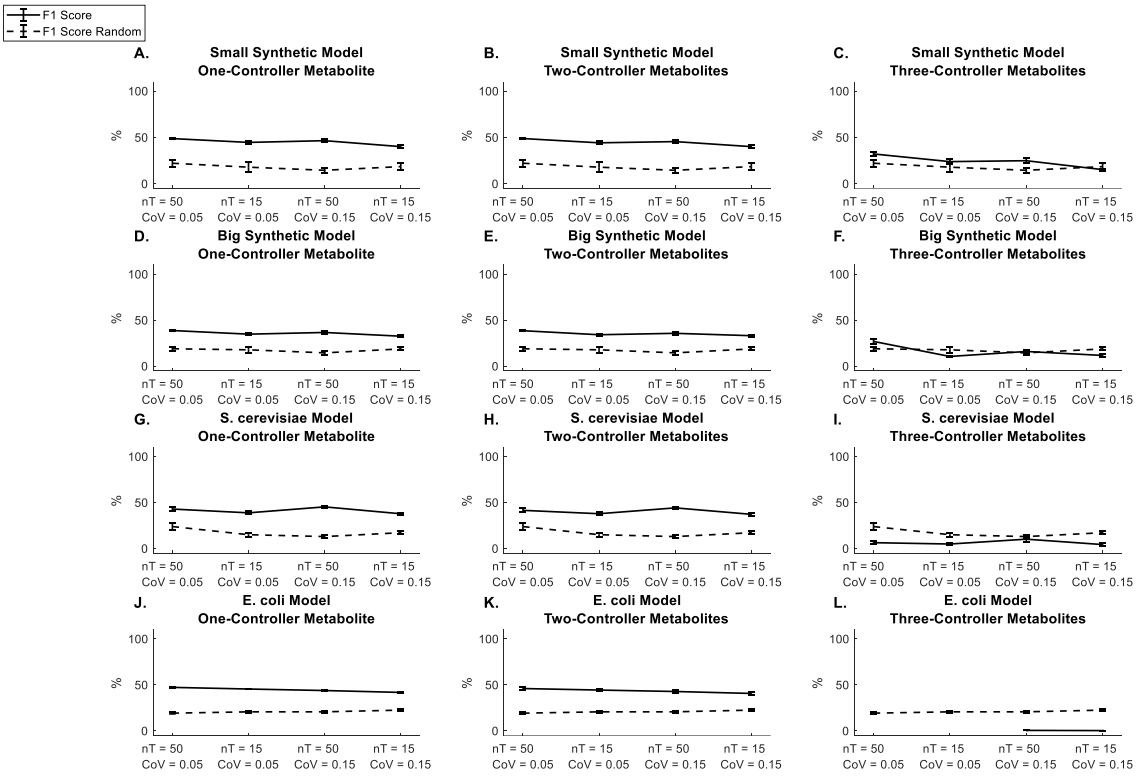

Figure 10: **F1 scores for synthetic and biological models using noisy and low sampling frequency training and test data.**
Bold lines represent the average F1 scores of SCOUR and dashed lines represent the average F1 scores when randomly classifying interactions (n = 30 from independent autogenerated training replicates). Note that the F1 score does not exist when identifying three-controller metabolites in the *E. coli* model for two conditions (nT = 50, CoV = 0.05 and nT = 15, CoV = 0.05) because SCOUR had removed both of the true positive three-controller metabolite interactions in a previous step of the framework for all repetitions, meaning sensitivity (and therefore F1 score) could not be calculated.

Perhaps the most striking feature of SCOUR is its use of the autogeneration of synthetic interactions for training data. Because machine learning models generally require large amounts of data for training, and because this scale of data is typically not available for metabolomics and fluxomics data, we created a method to automatically generate training data that are in some way representative of a wide variety of real biological interactions. While these autogenerated "interactions" may not perfectly recapitulate the data that result from real reactions, SCOUR's success shows that this autogeneration method can sufficiently train machine learning algorithms to identify regulatory interactions in many different systems. Because dynamic metabolomics and especially fluxomics data are so expensive and difficult to acquire with current analytical tools, this autogeneration method may prove useful for other tasks that require large amounts of these types of data.

Although this proof-of-principle framework has demonstrated significant potential for identification of many different regulatory interactions, there are several potential future avenues to improve overall performance. We note that both training and testing on autogenerated data produces higher PPVs (Figure 11), which indicates that the autogenerated data do not perfectly capture biological interactions. Autogeneration of training data using Michaelis-Menten or other kinetics equations instead of BST equations could improve machine learning performance by generating training data that are more representative of the types of kinetics encountered in biological systems. While we chose to initially use BST equations for autogeneration based on their simplicity and utility in modeling many different systems[97], there are undoubtedly limitations to their generalization.

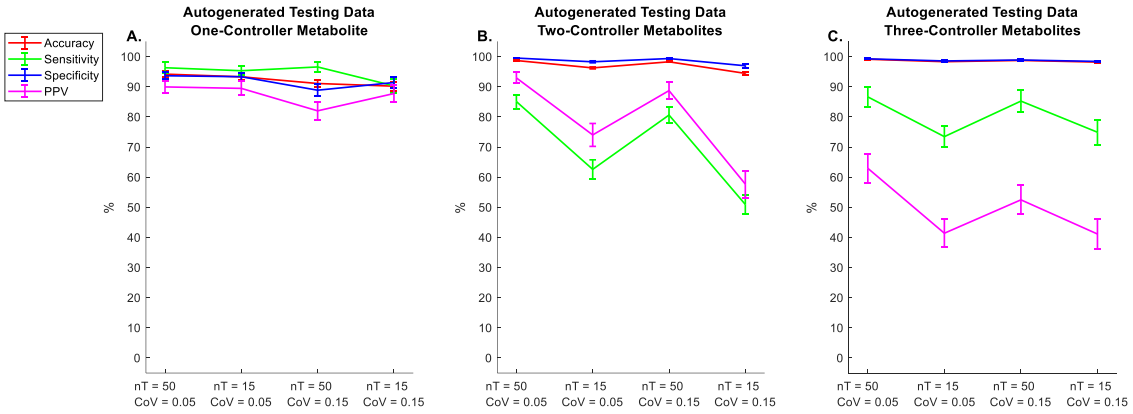

Figure 11: **SCOUR performance when training on autogenerated data and testing on autogenerated testing data.**

Testing data was autogenerated in a similar manner as the training data, but the percentage of true positive three-controller metabolite interactions in the testing set was set to 1% to more closely resemble the percentages found in the synthetic and biological models assessed (Note: training SCOUR with 1% instead of 5% true positives did not seem to help performance). The results when testing on autogenerated data were better than when testing on the synthetic or biological models, which is likely due to the testing data more closely resembling the training data because both datasets were autogenerated. Because we observe a similar decrease in PPV (when testing on the synthetic and biological systems) as the number of controller metabolites increases, this increase in performance when testing on autogenerated data is not likely completely due to overfitting, or else we would expect to see consistently high performance across all steps of the framework. The autogenerated training data has shown to be effective in classifying many interactions in the synthetic and biological models, and improvements to this autogeneration method or the integration of biological data will only lead to better performance with SCOUR.

Second, as in all machine learning approaches, there is room to improve the features used to help predict true interactions. We designed knowledge-driven features based on how metabolites interact with the reaction fluxes they control. Data-driven features derived from raw metabolomics and fluxomics data could be beneficial if there are sufficient data to drive the derivation of these features, including the underlying "ground truth" about whether a given interaction truly exists in a system. Such features could include graph theoretical characterization of the network topology of how

metabolites and fluxes are connected to each other, which has previously been used in metabolic contexts[98, 99]. However, an outstanding challenge will be how to include autogenerated data that are representative of these topological trends and capture the biological intricacies of metabolic systems, given that the autogenerated data are by definition synthetic and at least partly non-biological in nature. Additionally, machine learning algorithms for input to the stacking model beyond those tested here could also improve SCOUR's performance.

Finally, the preprocessing of noisy experimental data undoubtedly can impact downstream analytical performance. While we settled on the median sample of the triplicate data when calculating features for one-controller metabolite interactions, and a Gaussian moving filter to smooth the data for two- and three-controller metabolite interactions, we also tried an average moving filter as well as an in-house smoothing approach [100]. Because the framework mostly produces extremely accurate results on noiseless data for all models tested, an improved data pre-processing approach (e.g., filtering, normalization, scaling, or other smoothing methods[101-103]) could significantly increase classification performance.

Notwithstanding these potential avenues for improvement, SCOUR is already a useful tool. At the very least, SCOUR can determine with high confidence reaction fluxes that are only controlled by a single metabolite, eliminating swaths of the metabolic network where metabolite-dependent regulation is unlikely to occur. However, SCOUR can also identify many more complex interactions, including possibly pointing towards reaction fluxes controlled by four or more metabolites (Figure 12).

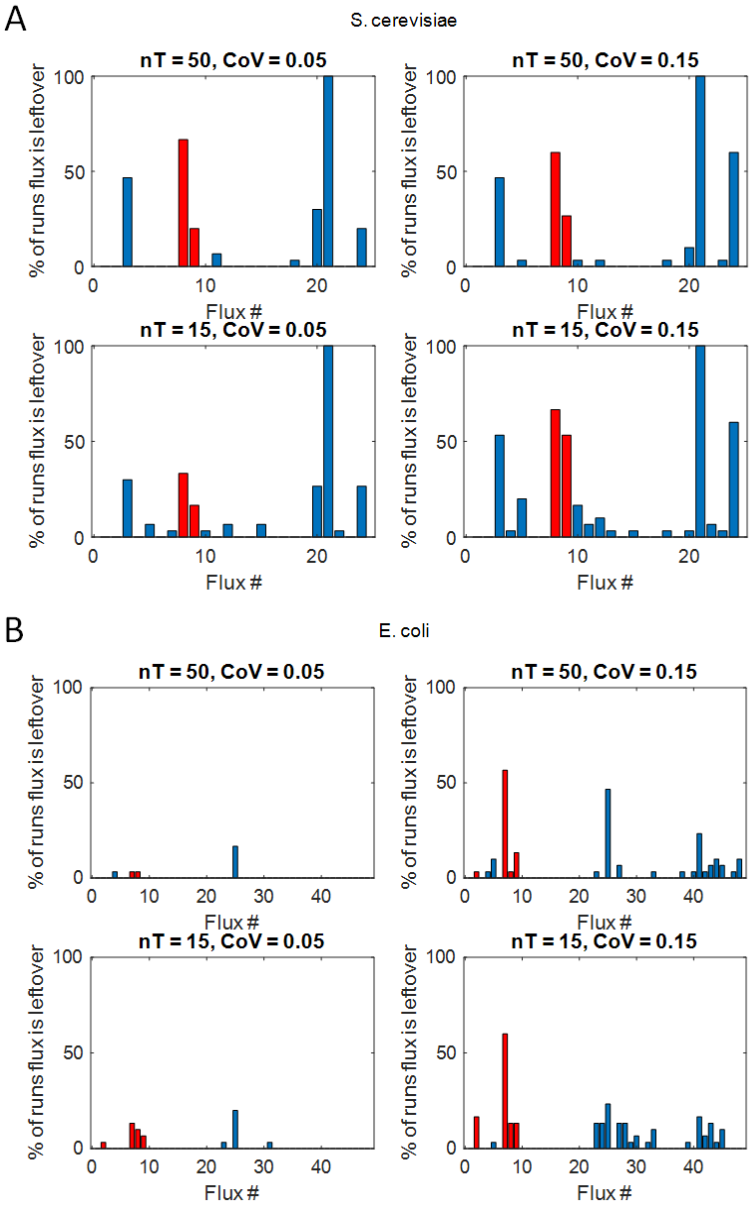

Figure 12: **Four-controller and higher-order metabolite regulatory interactions.** While we only focus on classifying one-, two-, and three-controller metabolite interactions, SCOUR may still provide information about four-controller or higher-order metabolite interactions. After removing the fluxes with predicted regulatory relationships in the three described steps of the framework, we have found that there is some evidence supporting the identification of leftover fluxes in the A) *S. cerevisiae* and B) *E. coli* models as being controlled by four metabolites. The two models contain 2 and 4 four-controller metabolite interactions (red bars), respectively. While the framework is unable to identify the controller metabolites that interact with these fluxes, being able to predict which fluxes are controlled by more than three metabolites may be useful for understanding reaction mechanisms. Further improvements to the accuracy in each of the steps of the framework would be required to ensure that the majority of fluxes that remain are truly controlled by four or more metabolites.

2.5 Conclusions

SCOUR is a proof-of-principle for how metabolomics and fluxomics data can be leveraged with machine learning to find metabolite-dependent regulatory interactions; to our knowledge, this is the first reported example of such an approach. The identification of metabolite-dependent regulatory interactions has to date been critically hampered by experimental limitations in measuring and validating these interactions, making SCOUR's predictions and triaging particularly valuable for such labor-intensive endeavors. Enabled by a method for autogenerating training data that reasonably mimic data from real biological systems, SCOUR circumvents the requirement for massive training sets that is typically associated with machine learning approaches. While metabolomics and fluxomics data are often collected at putative steady states, it is quite feasible to collect these data dynamically to leverage SCOUR's potential for biological discovery. This means that as analytical methods for measuring metabolomics and fluxomics become cheaper and easier, and more data are available for analysis, SCOUR will be ready to take full advantage of these new datasets to discover biochemical regulatory interactions.

CHAPTER 3: Inferring Absolute Concentrations from Relative Abundances in Metabolomics Data

3.1 Background

Since its inception, metabolomics has been used in a wide variety of applications, including identification of disease biomarkers, disease diagnosis, and even drug development[15, 104]. While genomics, proteomics, and transcriptomics provide an upstream view of cellular function and information about what may occur in a system, metabolomics is a direct readout of a system's current metabolic state and has emerged as an important area of study for understanding what is actually occurring in a system [105]. As the systems-scale study of metabolites, metabolomics has the potential to be integrated into metabolic modeling frameworks to better understand how cellular systems function and react to endogenous or exogenous perturbations[25]. In order to develop accurate metabolic models, an ample amount of metabolomics data will be necessary.

As described at the beginning of this thesis, researchers commonly use three analytical techniques to measure metabolomics: nuclear magnetic resonance (NMR) spectroscopy and gas and liquid chromatography-mass spectrometry (GC-MS and LC-MS, respectively). Because NMR is significantly less sensitive than GC-MS and LC-MS, many researchers turn to mass spectrometry when measuring hundreds or thousands of metabolites at low concentrations in a single sample[29]. LC-MS is able to detect more metabolites than GC-MS and does not require any sample derivatization[3, 27], but GC-MS still remains popular among researchers due to its relative inexpensiveness compared to other methods[27, 106].

The greatest weakness of these mass spectrometry approaches is quantification[107], as the metabolomics data resulting from using these instruments are typically relative abundances and not absolute concentrations. These relative abundances still allow for some types of analysis, including principal component analysis (PCA)[108] and t-tests, as relative abundance measurements of the same analyte can be compared from sample to sample. However, comparing the relative abundances of different metabolites has no quantifiable meaning[109]. Even if two metabolites have similar absolute concentrations, their peaks on a chromatogram and therefore their relative abundances can be radically different because of how chemicals with different structures and properties are derivatized, ionized, or fragmented[32, 110]. Similarly, peaks with comparable intensities do not necessarily imply equal absolute concentrations. This precludes the use of raw metabolomics data in many computational tools used to study metabolism. For example, several metabolic modeling platforms, such as MetDFBA, TMFA, and LK-DFBA[56, 111, 112] can directly integrate metabolite data into their frameworks, but they all require absolute concentrations.

Many researchers use chemical standards in mass spectrometry to quantify metabolites. However, these standards can be costly, time consuming to use, and unavailable for certain metabolites[30-32]. Using standards may only be feasible for quantifying a few metabolites, but for the purposes of untargeted metabolomics, where one attempts to measure all metabolites[113], it quickly becomes infeasible. Untargeted metabolomics data is restricted to being used with only the most exploratory computational tools, like PCA, for semi-quantitative analysis[109]. A method for determining absolute concentrations without the use of chemical standards would expand

the usability of metabolomics data in computational tools and would be incredibly beneficial to the metabolomics community.

As one of the most critical challenges preventing metabolomics from being more readily used in wider applications, there have previously been efforts attempting to quantify metabolomics data without chemical standards. Much of this work has focused on predicting ionization efficiencies of different chemicals, which is directly linked to the relative abundance output of a chromatogram in mass spectrometry. It has been shown that intrinsic thermodynamic properties, electrokinetic properties, structural properties, and solvent factors are all key factors that contribute to the prediction of ionization efficiencies[114, 115]. Recently, Liigand et al. developed a method for predicting ionization efficiencies using random forest machine learning[116]. Another recent approach is MetabQ, a calibration curve-free method for quantification of polar metabolites[117]. While MetabQ still requires chemical standards, they only need to be used once in the lifetime of an instrument to determine the relationship between relative abundances and absolute concentrations.

In this work, we have developed a new computational framework for inferring the most likely absolute concentrations from relative abundance metabolomics data for cellular metabolism, which we have named Metabolomics Prediction of Absolute Concentrations (MetaboPAC). MetaboPAC attempts to avoid the need for chemical standards by leveraging the mass balances of a metabolic system and determining the most biologically likely metabolic profiles. To the best of our knowledge, this is the first computational platform for standard-free inference of absolute concentrations using metabolic mass balances. MetaboPAC could play a significant role in improving the

ability to readily integrate metabolomics data with metabolic modeling and other metabolic analysis tools in the future.

3.2 Methods for Inferring Absolute Concentrations

3.2.1 Synthetic models

To assess MetaboPAC on different types of possible metabolic systems, we created two synthetic models. The first synthetic model (Figure 13A) contains four metabolites and five fluxes, where the initial influx is a known, constant reaction rate. With four metabolites and four unknown fluxes, the system is determined, which allows for the fluxes to be trivially solved using Equation 4, where $\frac{d\vec{x}}{dt}$ is the vector of the change in concentration over time for each metabolite, *S* is the stoichiometric matrix of the system, and $\vec{v}$ is the vector of fluxes. Since a determined system typically has a single unique flux profile solution, this simplification makes it a prime candidate for initial testing.

$$\frac{d\vec{x}}{dt} = S \cdot \vec{v} \qquad \text{(Equation 4)}$$

The second synthetic model (Figure 13B) contains four metabolites and eight fluxes. Once again, the influx is assumed to have a known, constant reaction rate. Unlike the determined model, the second model contains more unknown fluxes than metabolites and is therefore underdetermined. Furthermore, we have included two allosteric regulatory interactions, inhibiting flux v_3 and promoting flux v_8. Most biological systems are underdetermined and include metabolite-dependent , making this system a more

complex and more relevant test for MetaboPAC. Both synthetic systems were constructed using Michaelis-Menten kinetics to model each reaction and each system was simulated for 10 seconds to generate concentration and flux data.

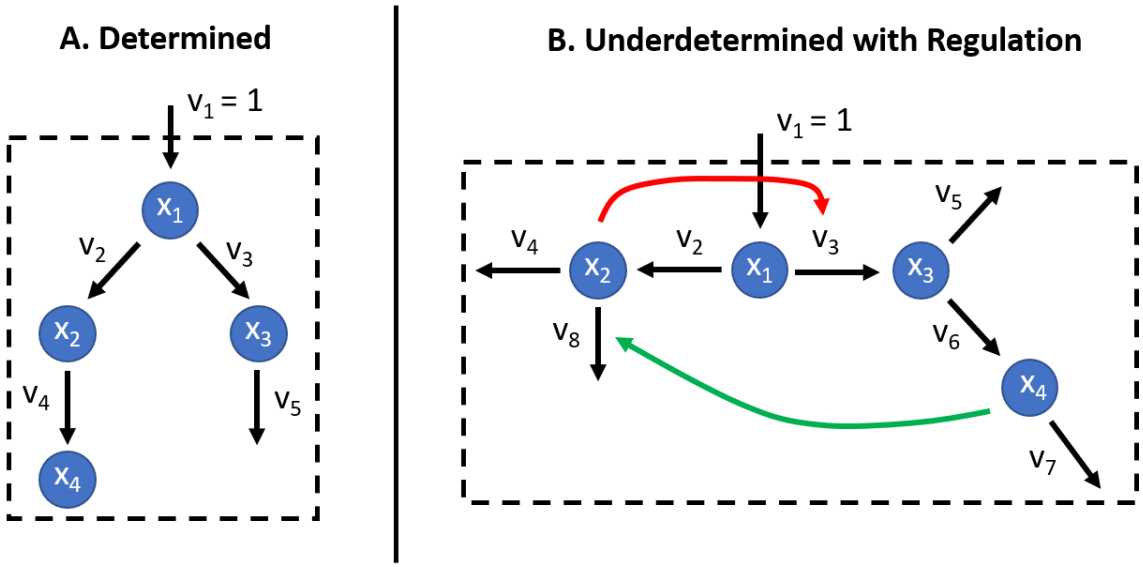

Figure 13: **Synthetic systems tested with MetaboPAC.**
We built one determined synthetic system and one underdetermined synthetic system with regulation using Michaelis-Menten kinetics for each reaction. x_i represents the ith metabolite and v_j represents the jth flux. In both systems, flux v_1 is assumed to be constant and known.

3.2.2 Biological models

While synthetic models are pragmatic for initially developing and testing MetaboPAC, they do not sufficiently resemble biological models to allow generalization of initial results to later applications. To further evaluate the robustness of MetaboPAC, we examined models of *Escherichia coli*[44] and *Saccharomyces cerevisiae*[47] metabolism. Both of these systems are underdetermined and include numerous allosteric regulatory interactions, with the *E. coli* model containing 18 metabolites and 48 fluxes, and the *S. cerevisiae* model containing 22 metabolites and 24 fluxes. The kinetic reaction equations

for both models include a mixture of Michaelis-Menten, Hill, and mass action kinetics. Data for both biological systems were produced by reconstructing the ODE models in MATLAB and simulating the *E. coli* and *S. cerevisiae* models over 10 seconds and 300 seconds, respectively.

3.2.3 Response factors

To emulate relative abundance data, we generated 20 sets of response factors for each metabolite found in the four systems evaluated in this work. Response factors describe the relationship between relative abundances and their absolute concentrations. Each response factor was randomly selected from a uniform distribution between 1 and 1000. These sets of response factors (RF_T) were multiplied by the true absolute concentration values simulated by the kinetic models to calculate the relative abundances, assuming there is a direct linear relationship between the two (Equation 5). While this relationship is not always linear, calibration curves using chemical standards are often calculated using the slope of the curve and cover the linear dynamic range of an instrument for a particular analyte[118]. Additionally, metabolite responses are typically linear over two to four orders of magnitude[32, 119], making it reasonable to assume a linear relationship between relative abundances and absolute concentrations. To infer absolute concentrations, the relative abundances are divided by the response factors predicted by MetaboPAC. The absolute concentrations for the systems used in this work ranged from 1e-4 mM to 20 mM.

$$Relative\ Abundance = Absolute\ Concentration \times RF_T \qquad \text{(Equation 5)}$$

3.2.4 Kinetic equations approach

If the kinetic rate law of each reaction in the system has been previously determined, the mass balances of the system and dynamic nature of time course metabolomics data can be leveraged to identify the response factors necessary to infer absolute concentrations. Based on the mass balances, the rate of change in concentration of a metabolite must always equal the sum of stoichiometrically balanced influxes and effluxes of the metabolite (Equation 4). When the kinetics of the reaction fluxes are known, each influx and efflux can be represented by a mathematical term containing kinetic parameters and the concentration of the metabolite(s) that participate in the reaction, either as a substrate or an allosteric regulator. Because only relative abundances and not absolute concentrations are available, the metabolite concentrations in these kinetic equations are replaced by their respective relative abundances divided by a response factor. The rate of change can be determined by calculating the difference in relative abundance at two subsequent timepoints and dividing by the change in time. Once again, the relative abundances in these rate of change calculations are divided by a response factor to infer absolute concentrations.

Across different timepoints in the metabolomics dataset, the response factors should remain constant for each metabolite, as they are not expected to change throughout an experiment. Together, the mass balances at each timepoint create a system of non-linear equations. A non-linear least-squares solver can determine the set of response factors that minimizes mass balance violations. These systems of equations must be determined or overdetermined to use the non-linear least-squares solver; as the number of timepoints in the data increases, the chance of having an underdetermined

system of equations decreases. In the kinetic equations approach, this system of non-linear equations is solved 48 times (chosen based on the maximum number of local workers (12 workers, each used 4 times) when performing parallel computations) with different initial seeds selected from a uniform distribution and the medians of the predicted response factors are calculated at the conclusion of all the runs as the most likely set of response factors.

3.2.5 Optimization approach

It is not uncommon for the kinetic equations of a reaction to be unknown, especially if the reaction is not in the most studied pathways of metabolism, such as central carbon metabolism. Instead of relying completely on the mass balances of the system to determine response factors, the optimization approach creates a minimization problem to predict the most likely set of response factors. In addition to minimizing mass balance violations in Equation 4 (without the known kinetic equations of the fluxes), there are also several penalties that can be added to the objective function to help identify sets of response factors that are biologically likely (Table 5). These penalties eliminate sets of response factors that lead to absolute concentrations that are biologically infeasible. For example, if a metabolite is the sole substrate of an enzyme, we expect the reaction rate to increase as the concentration of the metabolite increases. If this interaction is not observed between the inferred absolute concentration of the metabolite and the flux, the set of response factors would be heavily penalized. As in the kinetic equations approach, the optimization approach is performed 48 times with different initial

seeds and the medians of the predicted response factors from all the runs is calculated to determine the most likely set of response factors.

Table 5: Penalties used in the optimization approach of MetaboPAC

Penalty	Description	Reasoning
Mass balance	Calculate the sum of squared residuals between the inferred change in absolute concentration over time calculated from the raw relative abundance data (i.e. the change in relative abundance over time divided by the predicted response factor) and the inferred change in absolute concentration over time calculated from the stoichiometry of the system and inferred fluxes (i.e. Equation 4).	If the change in absolute concentration over time is vastly different between the two calculations (i.e. the sum of squared residuals is greater than zero), the predicted response factors have failed to produce inferred absolute concentration and flux profiles that do not violate any mass balances in the system.
Maximum concentration	If the inferred absolute concentration for any metabolite is above 5 mM or 50 mM for synthetic and biological systems, respectively, add a penalty equal to the maximum value of all inferred concentrations.	It is reasonable to assume that for many metabolites, there can be a general estimate for a maximum concentration that is biologically feasible, either due to limits in production or cell toxicity. Here, we use a single threshold for all metabolites, but imposing individual maximum thresholds would lead to better response factor predictions.
Correlation for mass action reaction with a single substrate	Calculate the correlation between the controller metabolite and inferred target flux. The correlation is expected to be positive (because metabolites induce mass action reactions), the penalty for each one-controller metabolite reaction equals the calculated correlation minus one.	If a reaction is only controlled by a single metabolite, the reaction rate should either increase or decrease as the concentration of the metabolite increases (assuming the kinetics of the reaction do not exhibit any behavior similar to substrate inhibition).
Curve fit for mass action reaction with a single substrate	Calculate the fit of a second-order polynomial to the controller metabolite and target flux data. The penalty for each one-controller metabolite reaction equals one minus the adjusted R^2 of the fit (adjusted for the number of coefficients).	A second-order polynomial should fit the data reasonably well if a reaction is controlled by a single metabolite (e.g. if the data is well-modeled by a Michaelis-Menten saturation curve).

Table 5 (continued)

Correlation for reactions regulated by two controller metabolites	Plot the data of one of the controller metabolites (x-axis) against the data of the inferred target flux (y-axis). Next, plot 23 vertical lines that are evenly spaced within the range of the controller metabolite that represent 23 constant concentrations. For one vertical line, identify if and where the line intersects with the data using the InterX function and linearly interpolate flux data at these intersection points using the closest higher and lower concentrations with sampled flux data. Calculate the Spearman correlation between the second controller metabolite and interpolated target flux at these intersection points where the first controller metabolite is constant. Repeat for each of the 23 vertical lines and then calculate the mean correlation. To calculate the penalty, take the mean correlation and subtract (for metabolites expected to induce the reaction) or add (for metabolites expected to inhibit the reaction) one. Switch which metabolite is held constant and repeat the process. Finally, sum all penalties together.	The correlation between one controller metabolite and its target flux should be consistently close to +1 (activation) or -1 (inhibition) for any constant concentration value for the second metabolite. This assumes that a controller metabolite cannot switch from inducing to inhibiting a reaction (or vice versa) at high concentrations, such as in substrate inhibition.
Fit to BST kinetic equations	For each reaction in a system, fit the inferred absolute concentration and flux data to a BST equation[97] representing the reaction rate. Calculate the sum of squared residuals of the fit.	A generic BST kinetic equation should fit reasonably well to correctly inferred absolute concentration and flux data.

3.2.6 Combining the kinetic equations and optimization approaches

In many cases, the kinetic equations of the reactions in a system are only partially known. In this scenario, the kinetic equations and optimization approaches can be used in serial. First, the kinetic equations approach is used for the metabolite mass balance equations where all the kinetic equations of the influxes and effluxes are known. If only a

few of the kinetic equations of the fluxes in a mass balance are known, it cannot be used in the kinetic equations approach; this can be a common occurrence when only a small percentage of the kinetic structure of the system is known. Only the response factors associated with the metabolites present in the useable mass balances can be identified in this step. After predicting all the possible response factors using the kinetic equations approach, the optimization approach proceeds as described above, except the response factors that have already been identified are fixed within the optimization problem and the remaining response factors are predicted. A workflow for this process is presented in Figure 14.

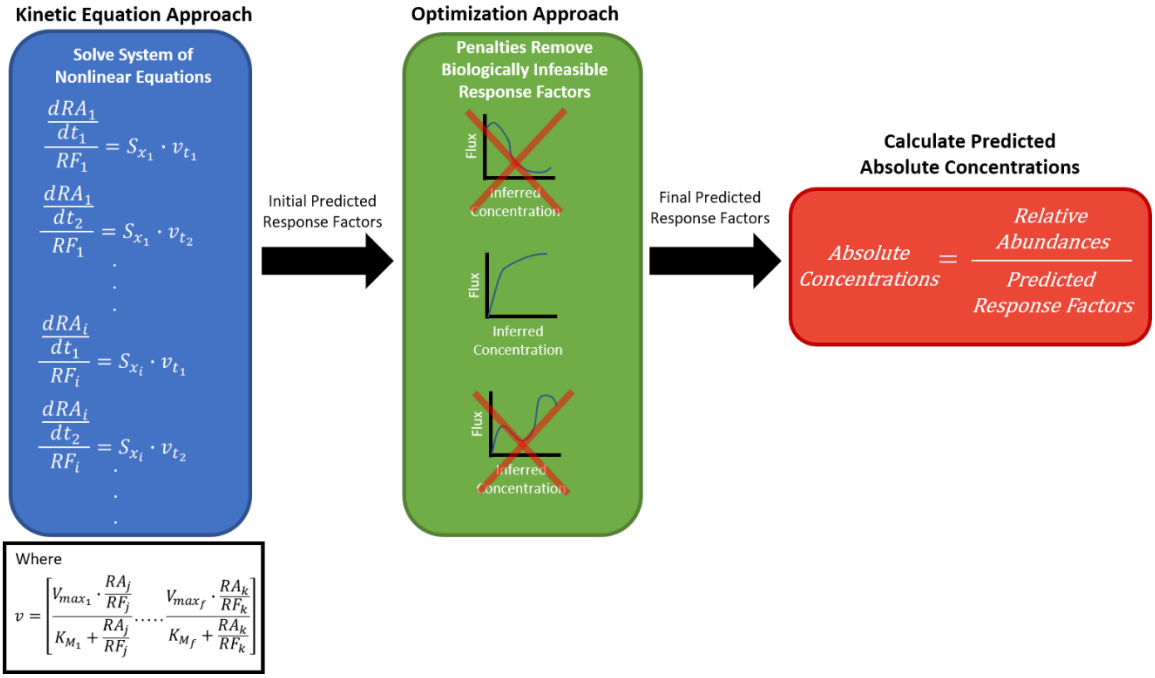

Figure 14: **MetaboPAC workflow for inferring absolute concentrations from relative abundances in metabolomics datasets.**
In the kinetic equations approach, the mass balances at each timepoint are used to create a system of non-linear equations where the response factors in the useable mass balances are predicted. These initial predicted response factors are transferred and fixed in the optimization approach, where penalties are used to eliminate possible sets of the remaining response factors. The final predicted response factors are used to infer the absolute concentrations of the data. RA_i is the relative abundance and RF_i is the unknown response factor of the *i*th metabolite, t_n is a particular timepoint in the data, S_{xi} is the stoichiometric mass balance coefficients of the *i*th metabolite, and v_{tn} is a vector of the fluxes at timepoint t_n. The kinetic equations (if known) of v_{tn} also contain relative abundances and response factors and are as shown in the inset (Michaelis-Menten kinetics are used as an example, where V_{maxf} and K_{Mf} are kinetic parameters of the *f*th flux).

3.2.7 Solving for flux distributions in the optimization approach

The only information that MetaboPAC assumes is known is the stoichiometry and metabolite-dependent allosteric regulation of the system, the kinetic structure of the system (if the kinetic equations approach is used), and the relative abundances of the data. In the optimization approach, the flux profiles of the reactions in the system are

used to calculate some of the penalties that describe the relationship between inferred absolute concentrations and the reactions they control. Because fluxomics data are not assumed to be available, the fluxes must be inferred by solving Equation 4. As in the kinetic equations approach, the rate of change is determined by calculating the difference in relative abundance between two timepoints divided by the time difference. This rate of change is divided by the corresponding response factor to infer the rate of change of absolute concentration for each metabolite. While the fluxes of a determined system can be trivially calculated, underdetermined systems have an infinite number of flux solutions. To choose a single solution, the optimization approach uses the Moore-Penrose pseudoinverse, which minimizes the norm of the flux solution[120]. If the kinetic equations of some of the fluxes are known, they can be used to create a less underdetermined system that could possibly be determined or even overdetermined, which would allow a unique flux solution to be found.

3.2.8 Noise-added data

To generate noisy data that more closely represent experimental metabolomics data, we used two sampling frequencies and two coefficients of variation (CoV) for randomly-added noise, for a total of four conditions. Sampling frequencies of 50 and 15 timepoints (nT) and CoVs of 0.05 and 0.15 were tested, where a higher CoV represents more noise (experimental error). Starting from the data generated by the ODEs defining the systems, each concentration value in each metabolomics dataset was replaced with a random value drawn from $N_{i,k} \sim (y_i(t_k), CoV \cdot y_i(t_k))$, where $y_i(t_k)$ is the value of metabolite i at timepoint k. Noisy data was smoothed using a Gaussian filter with a

window of one-fourth the length of the simulated time interval.

3.2.9 Evaluation metrics and comparing to baseline methods

To measure the performance of MetaboPAC, we calculated the relative difference between the true and predicted values of the response factors using a logarithmic scale and determined if it was within a range of $\log_2(1.1)$, $\log_2(1.3)$, and $\log_2(1.5)$ error, as shown in Equation 6. RF_T is the true response factor, RF_P is the predicted response factor, and x is the value that determine the $\log_2$ error range (i.e. 1.1, 1.3, or 1.5). We found that using absolute percent error instead of a logarithmic scale could lead to large error ranges that would make the evaluation metric less meaningful. For example, 100% error for a response factor of 500 would cover a range from 0 to 1000 (i.e. the entire search space of response factors for this work), whereas using a logarithmic scale would cover a range from only 250 to 1000. The percentages of predicted response factors that were within each $\log_2$ error range were compared among the methods assessed.

$$|\log_2 RF_T - \log_2 RF_P| < \log_2 x \qquad \text{(Equation 6)}$$

To provide a baseline performance for predicting response factors, we examined two other methods for predicting response factors that were compared to MetaboPAC. The first method randomly predicts response factors using a uniform distribution between 1 and 1000 for each metabolite. The second method uses a response factor of 500 for each metabolite, as predicted response factors close to the middle of the search space will have the greatest chance of being contained within the error range of the true response

factor if the response factors are chosen from a uniform distribution.

3.3 Results

3.3.1 MetaboPAC performance on noiseless data

When initially assessing the performance of MetaboPAC on the two synthetic systems, we found the framework to perform exceptionally well on noiseless data (Figure 15). For all percentages of known kinetic equations, MetaboPAC performed significantly better than the random response factors and response factors of 500 for each of the $\log_2$ error ranges examined. For the underdetermined system with regulation, MetaboPAC performed significantly better than the other two methods when at least 60% of the kinetic equations were known for the $\log_2(1.1)$ error range and when at least 40% of the kinetic equations were known for the $\log_2(1.3)$ and $\log_2(1.5)$ error ranges. Unsurprisingly, as the percentage of known kinetic equations increased, the accuracy of predicted response factors also generally increased for MetaboPAC, with 100% of the response factors within the $\log_2(1.1)$ error range for both systems when 100% of the kinetic equations were known. As expected, the response factors predicted when using the kinetic equations approach were more accurate than the predictions by the optimization approach (Figure 16). Figure 17 shows the mean percentage of response factors predicted by either the kinetic equations approach or optimization approach across different percentages of known kinetic equations.

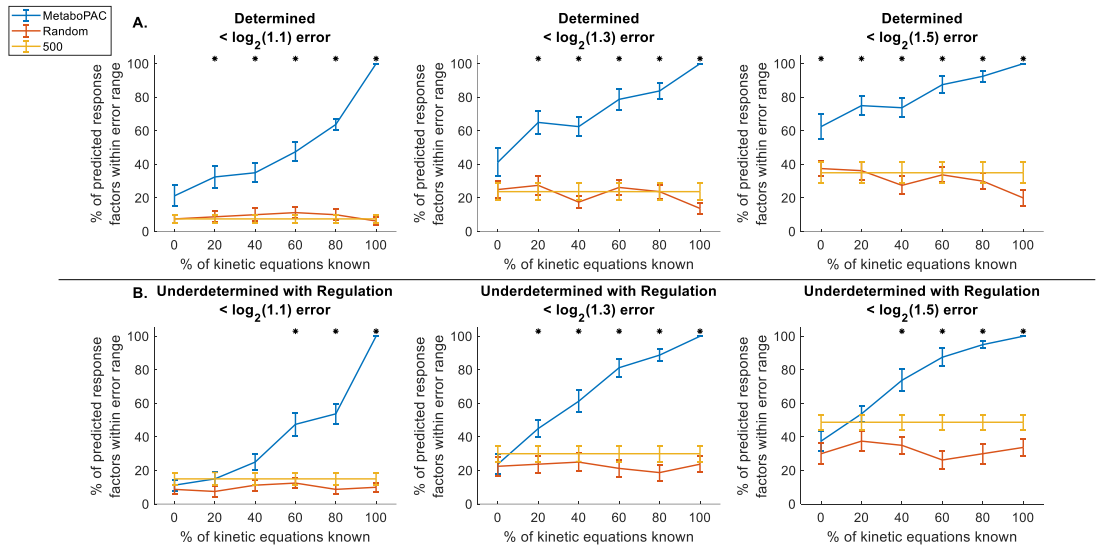

Figure 15: **MetaboPAC performance on noiseless data for synthetic systems.** MetaboPAC compared to random response factors and response factors of 500 for the A. determined and B. underdetermined with regulation systems using error ranges of $\log_2(1.1)$, $\log_2(1.3)$, and $\log_2(1.5)$. Lines represent the mean percent of predicted response factors within the error ranges for each method. Error bars represent the standard error of the mean (n = 20 for different sets of true response factors). Asterisks denote when MetaboPAC performed significantly better at predicting response factors than both of the other two methods (two-sample t-test with $\alpha = 0.05$).

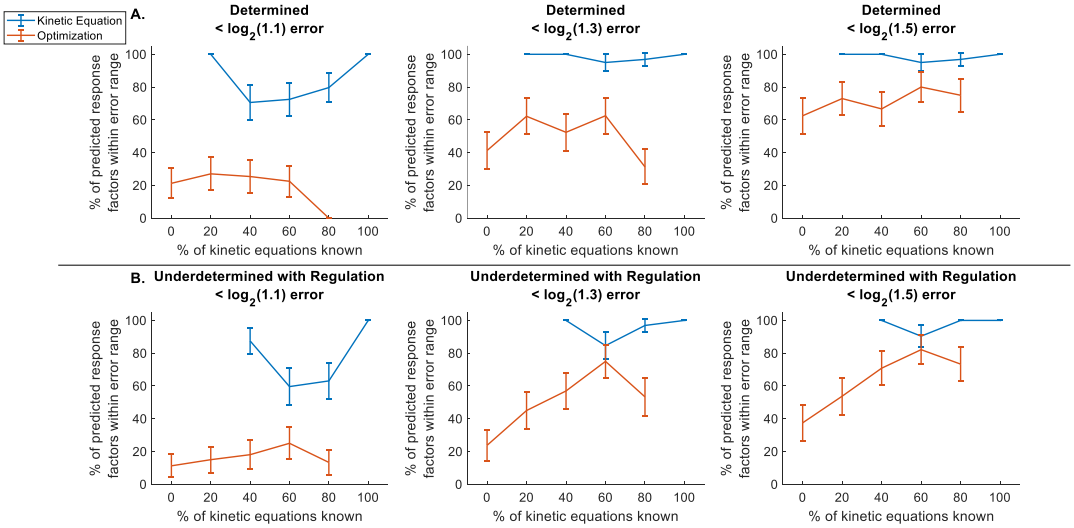

Figure 16: **Percent of response factors predicted by the kinetic equation and optimization approaches within each $\log_2$ error range for the synthetic systems when using noiseless data.**
The kinetic equations approach generally predicted more accurate response factors than the optimization approach. Error bars represent the standard error of the mean (number of samples varies based on the percentage of kinetic equations known (Figure 17))**.**

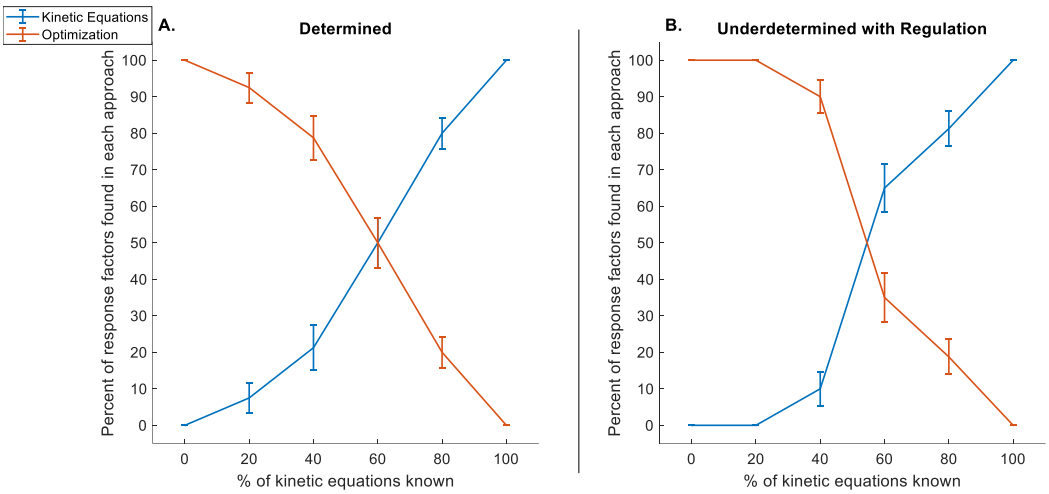

Figure 17: **Percentage of response factors predicted by the kinetic equation and optimization approaches for the synthetic systems.**
As the percentage of known kinetic equations increases, it is more likely for response factors to be solvable using the kinetic equations approach. Error bars represent the standard error of the mean (number of samples varies based on the percentage of kinetic equations known).

Testing MetaboPAC on the two biological systems with noiseless data yielded similar results (Figure 18). For the *S. cerevisiae* system, MetaboPAC performed significantly better than the other two methods across all $\log_2$ error ranges when at least 40% of the kinetic equations were known, except for one case in the $\log_2(1.5)$ error range. In the *E. coli* system, 60% of kinetic equations were required to be known for MetaboPAC to perform significantly better than the other two methods across all $\log_2$ error ranges. Once again, the kinetic equations approach typically outperformed the optimization approach (Figure 19). While the performance of MetaboPAC on the biological systems was not as high as the performance on the synthetic systems, it was still able to predict at least 58.9% of the response factors within $\log_2(1.1)$ error and at least 83% of the response factors within $\log_2(1.3)$ error in both systems when 100% of the kinetic equations are known.

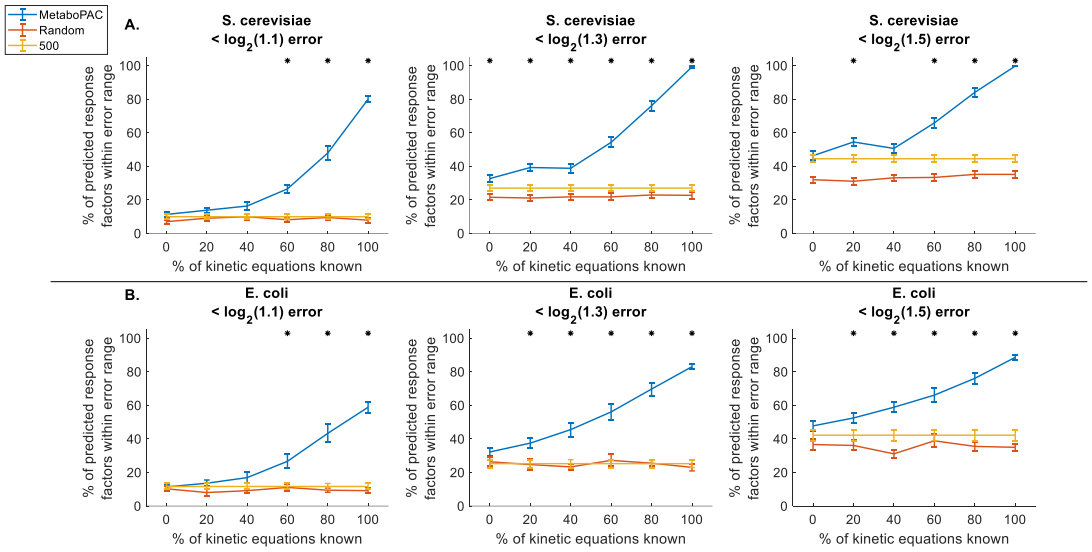

Figure 18: **MetaboPAC performance on noiseless data for biological systems.** MetaboPAC compared to random response factors and response factors of 500 for the A. *S. cerevisiae* and B. *E. coli* systems using error ranges of $\log_2(1.1)$, $\log_2(1.3)$, and $\log_2(1.5)$. Lines represent the mean percent of predicted response factors within the error ranges for each method. Error bars represent the standard error of the mean (n = 20 for different sets of true response factors). Asterisks denote when MetaboPAC performed significantly better at predicting response factors than both of the other two methods (two-sample t-test with $\alpha = 0.05$).

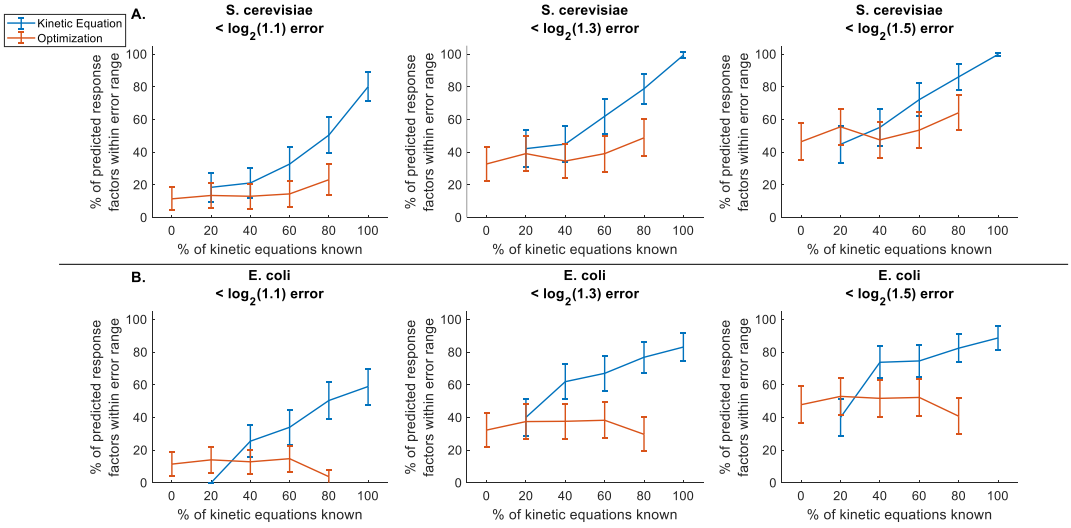

Figure 19: **Percent of response factors predicted by the kinetic equation and optimization approaches within each $\log_2$ error range for the biological systems when using noiseless data.**
The kinetic equations approach generally predicted more accurate response factors than the optimization approach. Error bars represent the standard error of the mean (number of samples varies based on the percentage of kinetic equations known (Figure 20))**.**

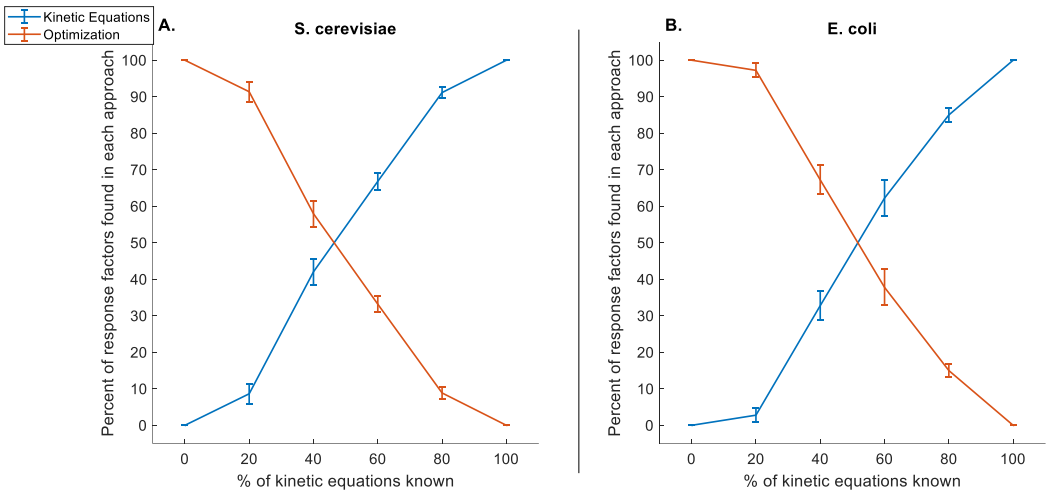

Figure 20: **Percentage of response factors predicted by the kinetic equation and optimization approaches for the biological systems.**
As the percentage of known kinetic equations increases, it is more likely for response factors to be solved using the kinetic equations approach. Error bars represent the standard error of the mean (number of samples varies based on the percentage of kinetic equations known).

3.3.2 MetaboPAC performance on noisy data

While noiseless data provides a good benchmark for the performance of MetaboPAC under ideal conditions, real experimental metabolomics data will have some degree of noise. To test the robustness of MetaboPAC under more realistic conditions, we assessed MetaboPAC on datasets with different sampling frequencies (nT = 50 or 15) and different amounts of added noise (CoV = 0.05 or 0.15). In the synthetic systems, MetaboPAC was significantly better than both the random and 500 response factor approaches for almost all $\log_2$ error ranges (Figure 21 and Figure 22) when 100% of kinetic equations were known under the low sampling frequency and high noise condition (nT = 15, CoV = 015). We also determined that MetaboPAC generally performed the best in the conditions with low amounts of noise (CoV = 0.05) and often only required 60% or 80% of the kinetics to be known to outperform the other methods. As found in the noiseless condition, the accuracy of the kinetic equations approach was higher than the

accuracy of the optimization approach in most cases (Figure 23 and Figure 24).

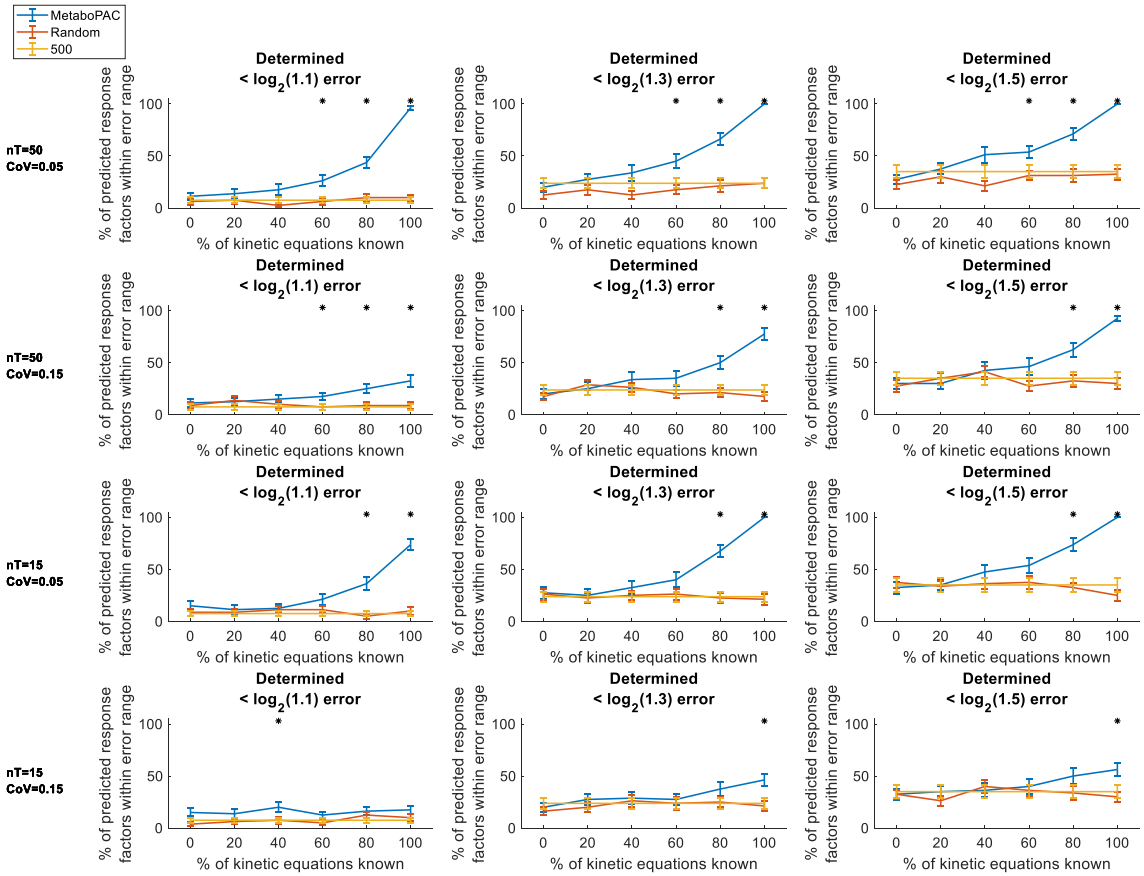

Figure 21: **MetaboPAC performance on all conditions of noisy data for the determined system.**
MetaboPAC compared to random response factors and response factors of 500 for the determined system using error ranges of $\log_2(1.1)$, $\log_2(1.3)$, and $\log_2(1.5)$ on data with different sampling frequencies (nT = 50 or 15) and noise added (CoV = 0.05 or 0.15). Lines represent the mean percent of predicted response factors within the error ranges for each method. Error bars represent the standard error of the mean (n = 20 for different sets of true response factors). Asterisks denote when MetaboPAC performed significantly better at predicting response factors than both of the other two methods (two-sample t-test with $\alpha = 0.05$).

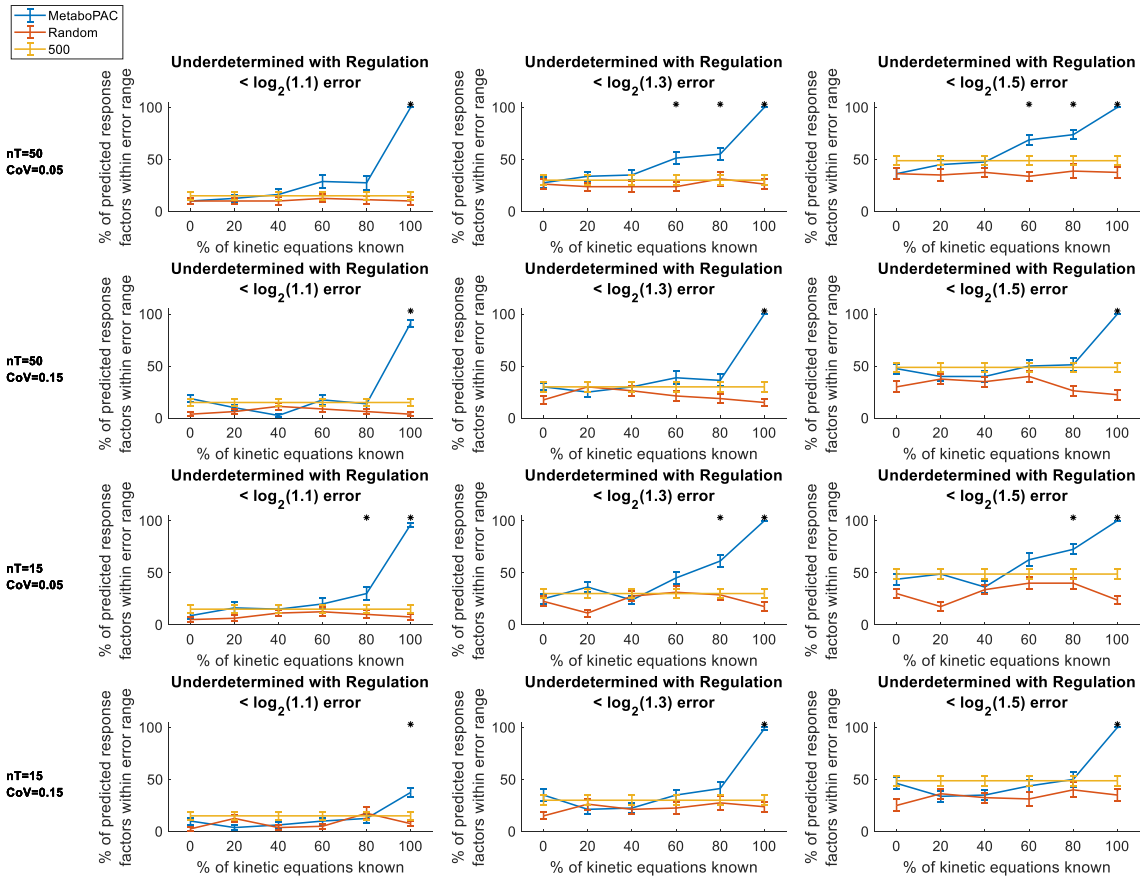

Figure 22: **MetaboPAC performance on all conditions of noisy data for the underdetermined system with regulation.**
MetaboPAC compared to random response factors and response factors of 500 for the underdetermined system with regulation using error ranges of $\log_2(1.1)$, $\log_2(1.3)$, and $\log_2(1.5)$ on data with different sampling frequencies (nT = 50 or 15) and noise added (CoV = 0.05 or 0.15). Lines represent the mean percent of predicted response factors within the error ranges for each method. Error bars represent the standard error of the mean (n = 20 for different sets of true response factors). Asterisks denote when MetaboPAC performed significantly better at predicting response factors than both of the other two methods (two-sample t-test with $\alpha = 0.05$).

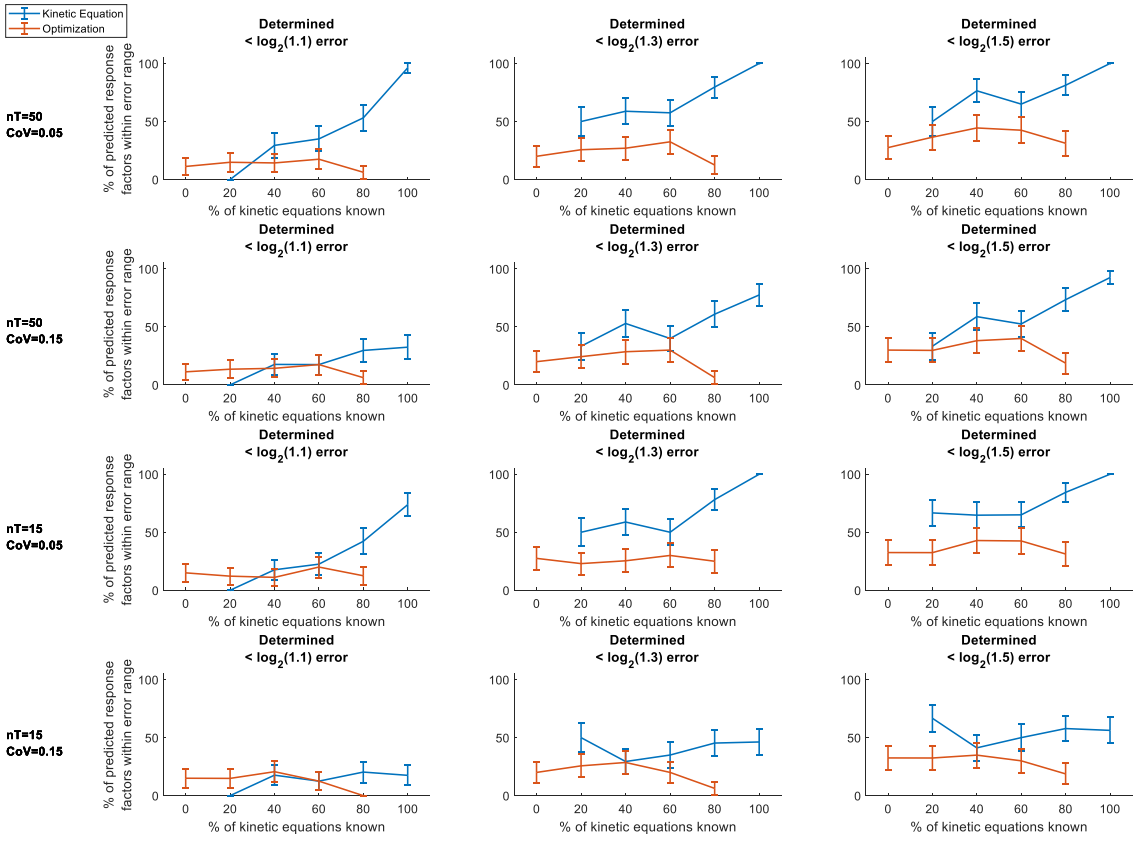

Figure 23: **Percent of response factors predicted by the kinetic equation and optimization approaches within each $\log_2$ error range for the determined system when using noisy data.**
The kinetic equations approach generally predicted more accurate response factors than the optimization approach. Error bars represent the standard error of the mean (number of samples varies based on the percentage of kinetic equations known (Figure 17))**.**

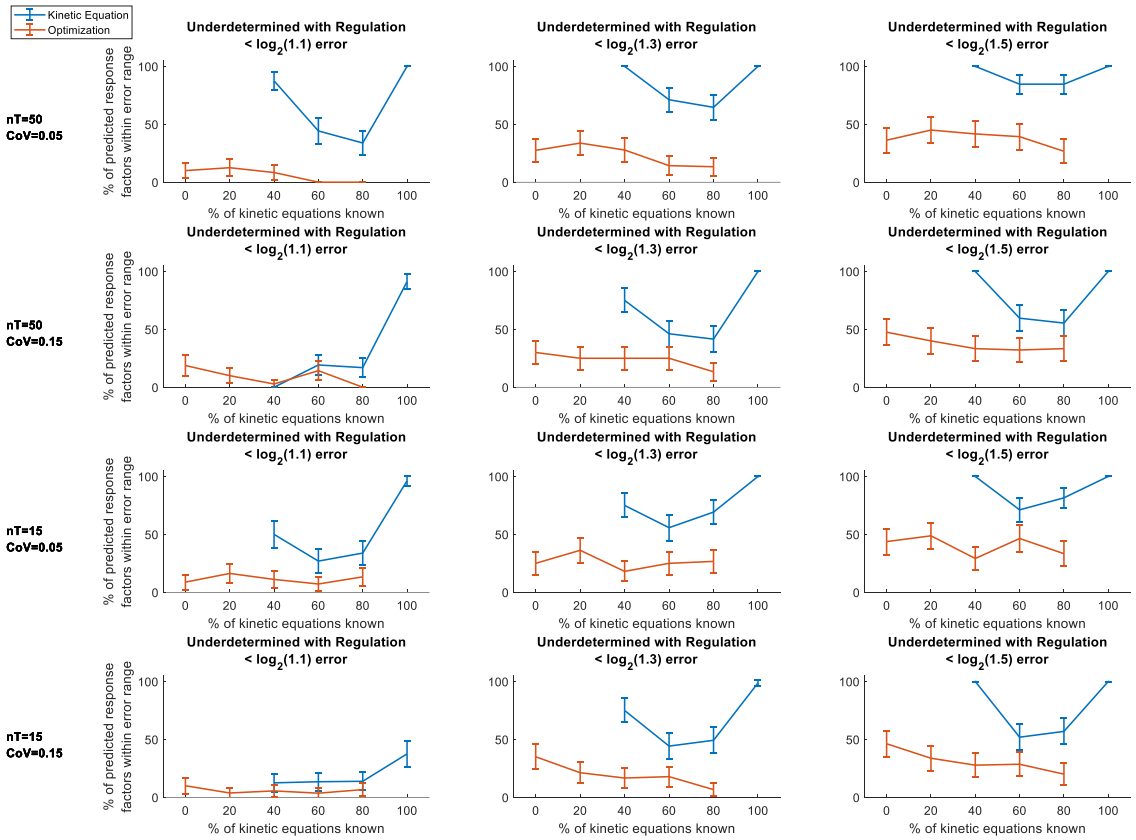

Figure 24: **Percent of response factors predicted by the kinetic equation and optimization approaches within each $\log_2$ error range for the underdetermined system with regulation when using noisy data.**
The kinetic equations approach generally predicted more accurate response factors than the optimization approach. Error bars represent the standard error of the mean (number of samples varies based on the percentage of kinetic equations known (Figure 17))**.**

While there was a predictable decrease in overall performance compared to the results when using noiseless data, MetaboPAC was still able to predict 56.3% and 100% of response factors within $\log_2(1.5)$ error for the determined and underdetermined system with regulation, respectively, when 100% of the kinetic equations were known. Surprisingly, MetaboPAC seems to perform better on the underdetermined system with regulation compared to the determined system at 100% known kinetic equations, despite the increase in complexity. The simplicity of the mass balance equations in the

determined system (with fewer reaction kinetics and no regulation, and therefore fewer instances of response factors within the mass balance equations) may actually hinder identification of accurate response factors in this instance.

For the two biological systems, MetaboPAC was still found to significantly predict response factors more accurately than random response factors or response factors of 500 for most of the $\log_2$ error ranges when 100% of the kinetic equations were known under low sampling and high noise conditions (Figure 25 and Figure 26). Once again, MetaboPAC showed some improved performance when under conditions where there was a high sampling frequency or low noise. Interestingly, the optimization approach often performed better than the kinetic equations approach when a low percentage of kinetic equations was known (Figure 27 and Figure 28), which was less common in the synthetic systems. In some cases, the optimization approach alone (0% known kinetic equations) was even significantly better than randomly predicting response factors or response factors of 500. This observation illustrates that both the kinetic equations and optimization approaches are important to the framework. When a low percentage of kinetic equations were known (20% to 60%), the performance of MetaboPAC was sometimes worse than its performance when using only the optimization approach. We found that these low percentages led to systems of non-linear equations in the kinetic equations approach that did not contain a large enough number of equations to reliably predict accurate response factors when using noise-added data. Instead, poor response factors were predicted by the kinetic equations approach, which led to the optimization approach underperforming when predicting the remaining response factors. If only a low percentage of kinetic equations are known, opting to use the optimization approach may

produce more accurate results.

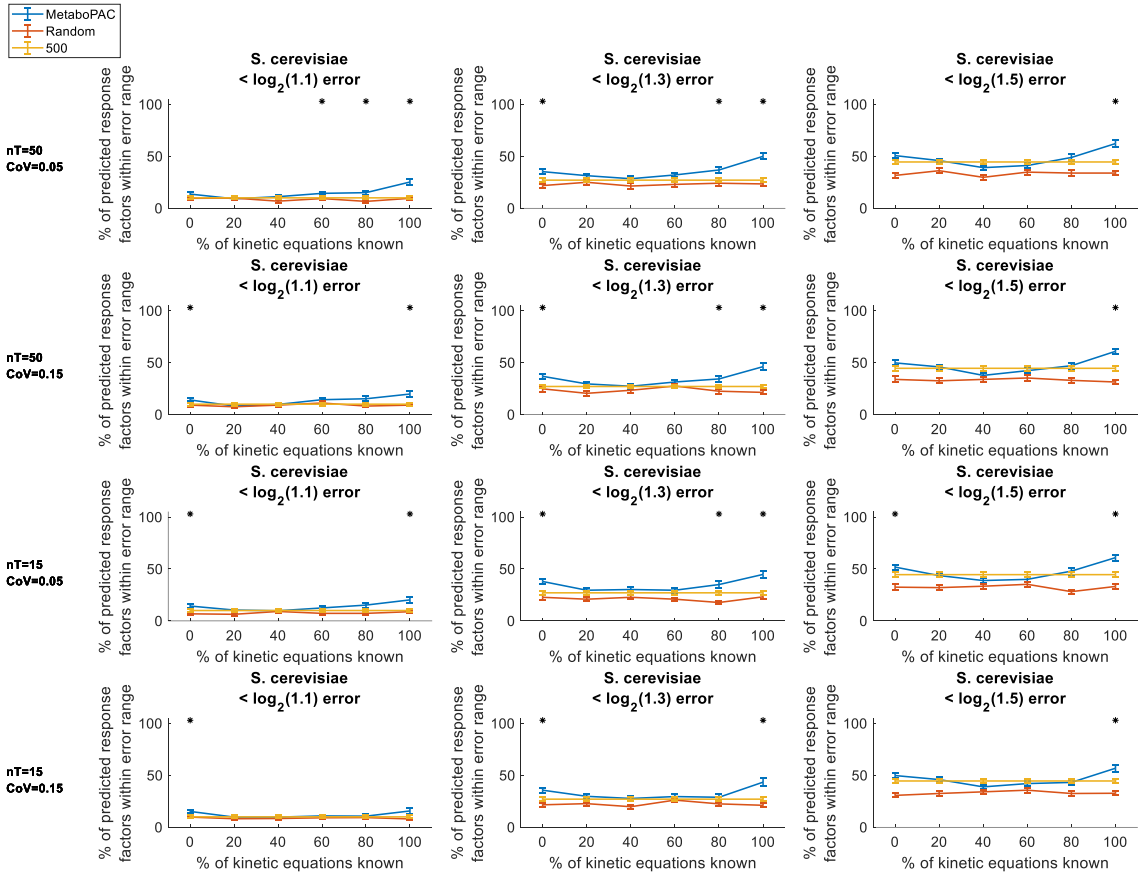

Figure 25: **MetaboPAC performance on all conditions of noisy data for the *S. cerevisiae* system.**
MetaboPAC compared to random response factors and response factors of 500 for the *S. cerevisiae* system using error ranges of $\log_2(1.1)$, $\log_2(1.3)$, and $\log_2(1.5)$ on data with different sampling frequencies (nT = 50 or 15) and noise added (CoV = 0.05 or 0.15). Lines represent the mean percent of predicted response factors within the error ranges for each method. Error bars represent the standard error of the mean (n = 20 for different sets of true response factors). Asterisks denote when MetaboPAC performed significantly better at predicting response factors than both of the other two methods (two-sample t-test with $\alpha = 0.05$).

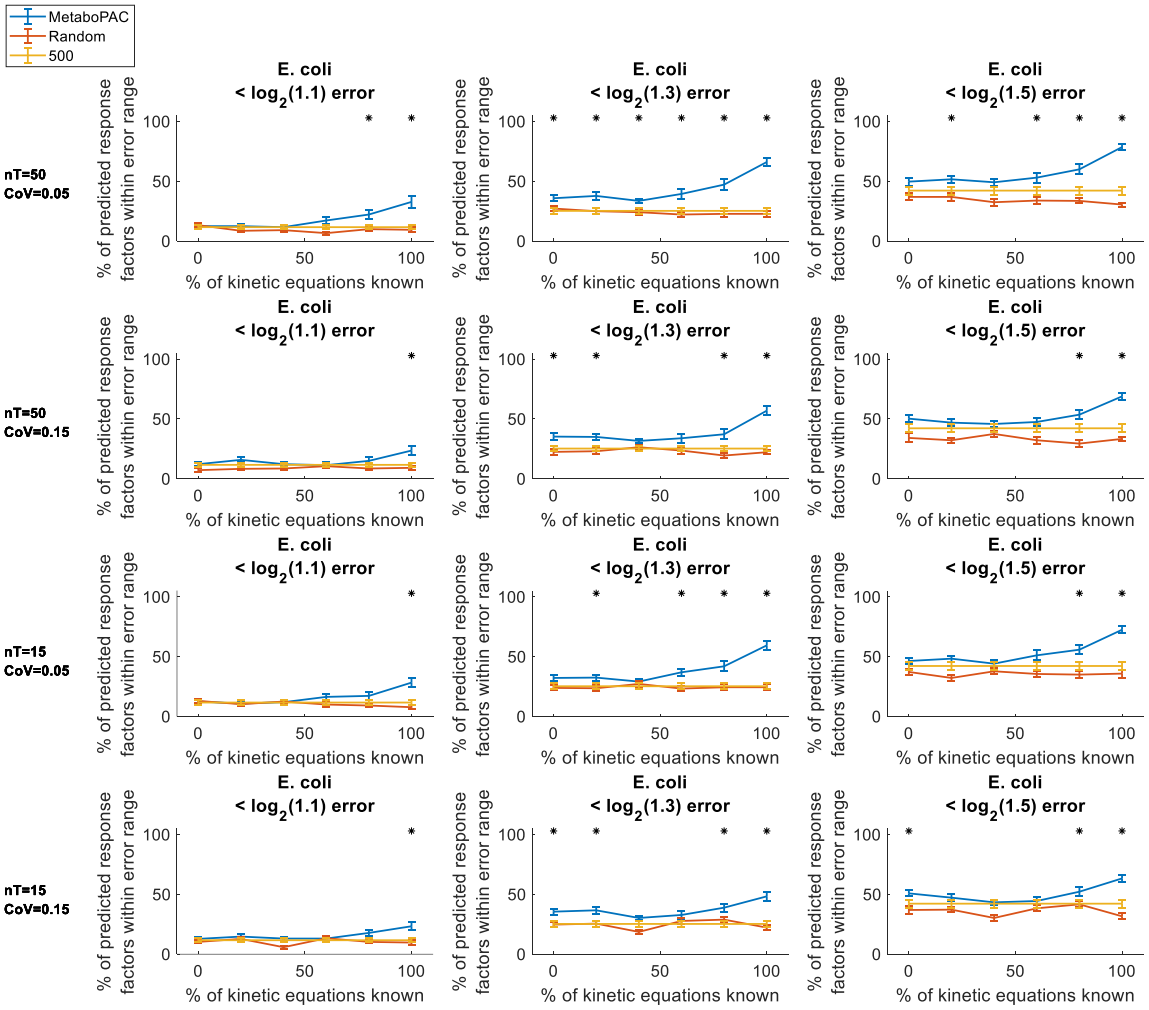

Figure 26: **MetaboPAC performance on all conditions of noisy data for the *E. coli* system.**
MetaboPAC compared to random response factors and response factors of 500 for the *E. coli* system using error ranges of $\log_2(1.1)$, $\log_2(1.3)$, and $\log_2(1.5)$ on data with different sampling frequencies (nT = 50 or 15) and noise added (CoV = 0.05 or 0.15). Lines represent the mean percent of predicted response factors within the error ranges for each method. Error bars represent the standard error of the mean (n = 20 for different sets of true response factors). Asterisks denote when MetaboPAC performed significantly better at predicting response factors than both of the other two methods (two-sample t-test with $\alpha = 0.05$).

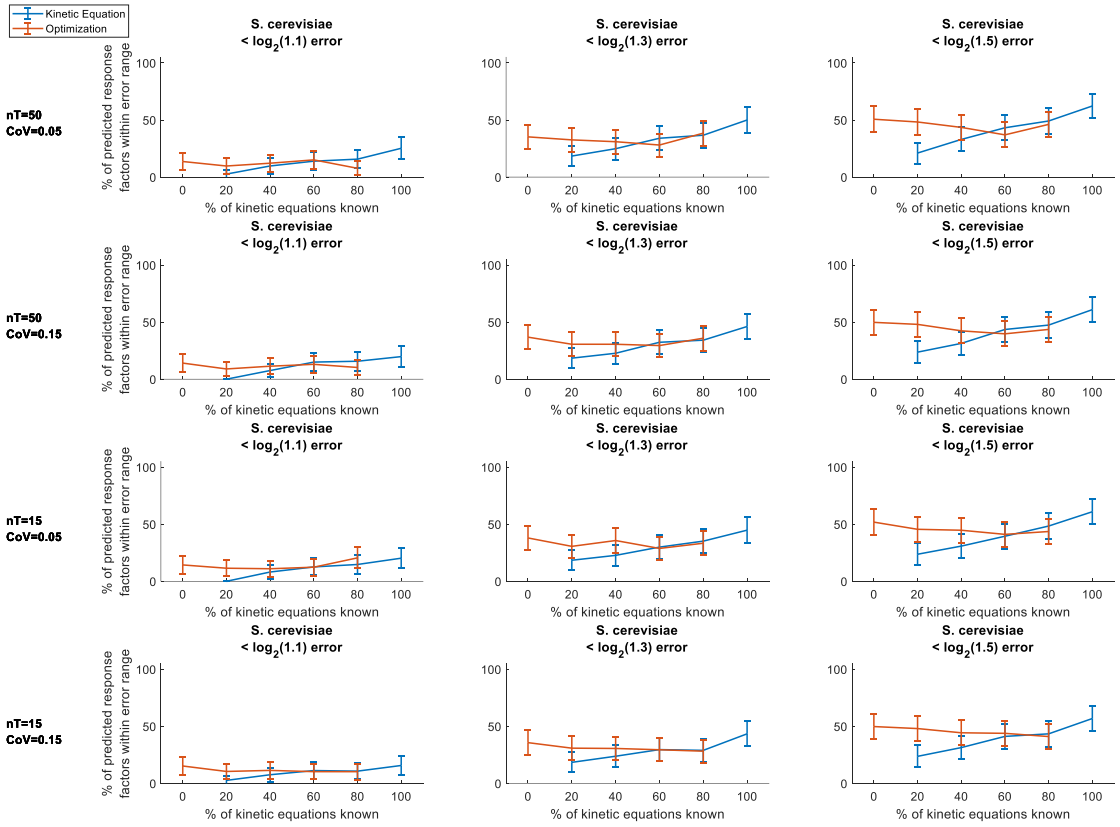

Figure 27: **Percent of response factors predicted by the kinetic equation and optimization approaches within each $\log_2$ error range for the *S. cerevisiae* system when using noisy data.**
At low percentages of known kinetic equations, the optimization approach often performed better than the kinetic equations approach. At around 80% known kinetic equations, the kinetic equations approach began to have improved performance over the optimization approach. Error bars represent the standard error of the mean (number of samples varies based on the percentage of kinetic equations known (Figure 20)).

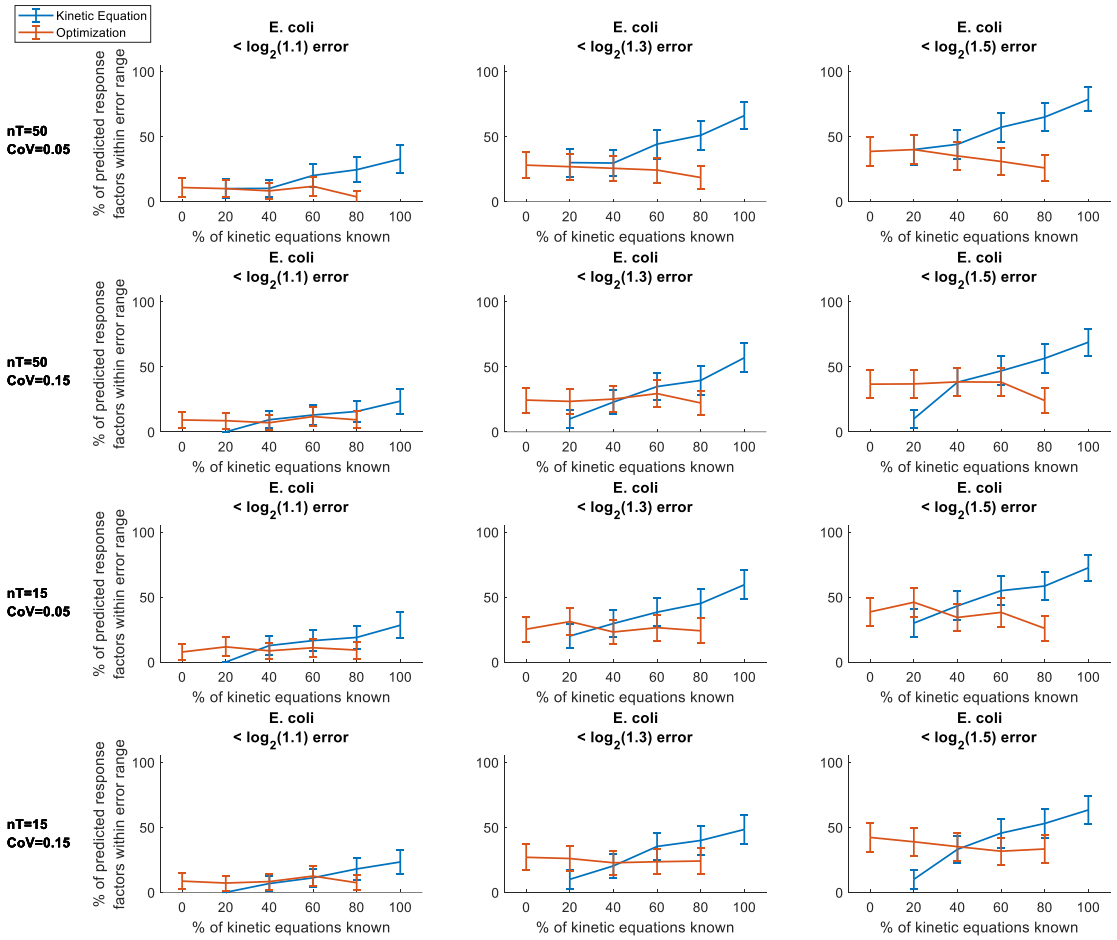

Figure 28: **Percent of response factors predicted by the kinetic equation and optimization approaches within each $\log_2$ error range for the *E. coli* system when using noisy data.**
At low percentages of known kinetic equations, the optimization approach often performed better than the kinetic equations approach. At around 60% known kinetic equations, the kinetic equations approach began to have improved performance over the optimization approach. Error bars represent the standard error of the mean (number of samples varies based on the percentage of kinetic equations known (Figure 20)).

3.4 Discussion

MetaboPAC has shown substantial potential to provide accurate absolute concentrations for metabolites in well-studied metabolic pathways (e.g. central carbon metabolism) whose kinetics have been previously determined for various biological

systems. Metabolomics research in common microorganisms, such as *E. coli* and *S. cerevisiae*, could benefit significantly from MetaboPAC, as it will allow metabolomics data to be more seamlessly integrated with metabolic modeling frameworks and data analysis methods that require absolute concentrations. The key component of MetaboPAC is the use of mass balances within a system with known stoichiometry. Previously, mass balances have been used to determine quenching leakage in metabolomics[121], but to the best of our knowledge, this is the first time mass balances have been used in the context of inferring absolute concentrations. Because MetaboPAC leverages the mass balances of a system to predict its response factors, it is unsurprising that the performance of MetaboPAC is hindered under conditions with high noise, as the mass balances can be affected. Nevertheless, MetaboPAC still significantly outperformed the other two methods assessed when all kinetic equations are known, suggesting that systems with known kinetic structures would benefit from MetaboPAC.

One of the strengths of MetaboPAC is that additional information can be easily integrated into the framework to reduce the number of possible sets of response factors. If the minimum or maximum possible or predicted concentrations of each (or a few) metabolites are known, this can greatly reduce the search space of possible sets of response factors. We found that constraining the range of possible response factors of one metabolite would often lead to the range of possible response factors of other metabolites also being constrained, especially metabolites nearby in the metabolic pathway. This is likely due to their mass balances sharing some of the same reaction fluxes. Along those lines, chemical standards could be used for some metabolites, which would decrease the number of response factors MetaboPAC would need to predict and would once again lead

to more constrained ranges of possible response factors for some metabolites.

Along with incorporating additional information directly into the framework, MetaboPAC could also be used concurrently with other metabolomics methods to improve the accuracy of response factor predictions. The use of methods focused on predicting ionization efficiencies[114-116] in conjunction with MetaboPAC could provide further insight about which response factors predicted by MetaboPAC are most likely to be accurate if the inferred concentrations of MetaboPAC and the ionization efficiency-based platforms are similar. When working with underdetermined systems, methods such as dynamic flux estimation[122, 123] or flux balance analysis[34] could be applied to determine more likely flux distributions than the Moore-Penrose pseudoinverse approach used in MetaboPAC, which could lead to improved predictions of response factors when using the optimization approach. Alternatively, if minimum or maximum values of individual fluxes are known, this could also reduce the number of possible flux solutions and benefit the calculation of penalties in the optimization approach.

In this proof-of-principle work, there are two key assumptions we have used to initially assess MetaboPAC. These assumptions are reasonable to assume for this work, but they will need to be adjusted in the future for MetaboPAC to be more widely applicable. First, we assumed the relationship between relative abundances and their absolute concentrations is linear. These relationships are not always linear in experimental data and non-linear relationships will need to be considered in the future . Both the kinetic equations and optimization approaches within MetaboPAC can be easily adjusted to account for non-linear relationships, but determining which metabolites have non-linear relationships is a more difficult problem and will need to be explored if this

information is not known *a priori*.

MetaboPAC also assumes the true response factors are sampled from a uniform distribution. Under the most realistic conditions, this may not be the case. To further assess the robustness of MetaboPAC, we tested the framework on noiseless relative abundance data from the two biological systems (assuming all kinetic equations are known) where the true response factors were drawn from a log uniform distribution (Figure 29). While there is a decrease in performance when using MetaboPAC on both biological systems compared to the results in Figure 18, this drop in performance is also seen in both the random and 500 response factor methods. MetaboPAC is still significantly better than the other two methods by a wide margin in all $\log_2$ error ranges examined, which indicates that MetaboPAC is still suitable even if response factors are not uniformly distributed.

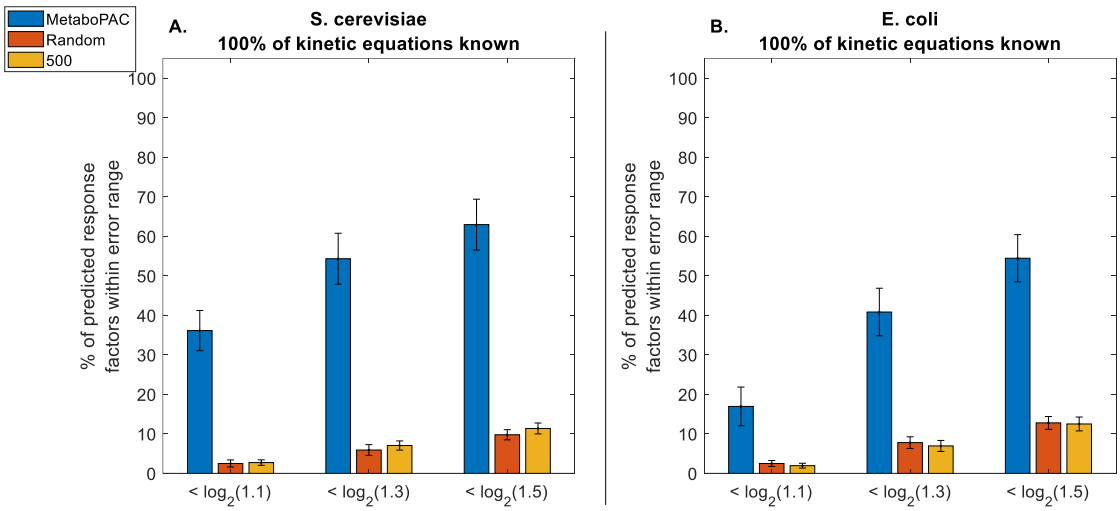

Figure 29: **MetaboPAC Performance when true response factors are sampled from a log uniform distribution.**
To further test the robustness of MetaboPAC, the true response factors were drawn from a log uniform distribution instead of a uniform distribution for the two biological systems with 100% known kinetic equations and noiseless data. MetaboPAC still outperformed the other two methods across all $\log_2$ error ranges in both the *S. cerevisiae* and *E. coli* systems. Bars represent the mean percent of predicted response factors within the error ranges for each method. Error bars represent the standard error of the mean (n = 20 for different sets of true response factors).

While the results presented here are promising, MetaboPAC currently has some limitations. First, because MetaboPAC leverages the stoichiometric mass balance of a biological system to identify response factors, it can only be used in the context of cellular metabolism in its current form. For example, MetaboPAC could not infer the absolute concentrations of metabolites in a blood sample because blood metabolite profiles are determined from metabolic contributions from across organs or systems in an organism with no stoichiometric basis to connect concentrations[124].

Perhaps the greatest obstacle for MetaboPAC is noisy data. In our results, we have determined that MetaboPAC performs particularly well on datasets with little or no noise, even when the kinetics of the system are not fully known. Under the noisiest condition (CoV = 0.15), we still generally find MetaboPAC to perform significantly

better than other methods for predicting response factors when all kinetic equations were known, but there is a noticeable decrease in accuracy. Leveraging the mass balances of a metabolic system is one of the critical ideas behind MetaboPAC and it is not surprising that data with high noise affect these calculations. Here, we have used a Gaussian filter to smooth the noisy data, which has proven to be effective, but other venues for mitigating the effect of noise should be considered. Exploring other options to reduce noise, such as filtering, normalization, scaling, other smoothing methods[102, 103, 125], or using triplicate samples as is common when collecting metabolomics data, could prove useful and lead to an increase in performance.

3.5 Conclusions

The need for chemical standards in mass spectrometry methods to absolutely quantify metabolomics data has been a challenging obstacle that has prevented the direct use of metabolomics in many metabolic modeling tools. MetaboPAC is a critical step toward preprocessing metabolomics data so that it can be readily used with metabolic modeling and other computational platforms that require absolute concentrations of metabolites. For well-studied systems, where the entire kinetic structure is known, MetaboPAC can infer absolute concentrations with high accuracy. Under conditions where the amount of noise in the data is minimal, MetaboPAC can still provide valuable information if the kinetic equations are only partially known. As research in metabolomics continues to grow and more computational frameworks aim to harness all the information that metabolomics data has to offer, MetaboPAC has the potential to become a powerful tool for absolute quantification in metabolomics.

CHAPTER 4: Improved Kinetics Constraints Increase the Predictivity and Applicability of a Linear Programming-based Dynamic Metabolic Modeling Framework

4.1 Background

Mathematical and computational models are often used to study metabolism, the set of reactions that supply the chemical precursors necessary for almost all cellular processes. These metabolic models are significantly cheaper and faster to run than laboratory experiments, meaning that they can be of tremendous value when they are able to predict how changes in or to a metabolic system can affect its state. While a few pathways and sections of metabolism (e.g., glycolysis and central carbon metabolism) have been modeled and characterized quite well in a few organisms (e.g., *Saccharomyces cerevisiae* and *Escherichia coli*)[44, 50], genome-scale models that capture metabolism at a systems scale have been more difficult to develop. Metabolism involves many interconnected reactions and pathways, making it critical to include as much of metabolism as possible in metabolic models to better represent the system and generate accurate predictions. Metabolomics has great potential to provide the information necessary to drive systems-scale metabolic models. However, creating genome-scale metabolic models that capture critical system behaviors like metabolic dynamics remains an outstanding challenge in the field, which has prevented the value of metabolomics data in this context from being fully realized.

To address this challenge, we recently developed Linear Kinetics-Dynamic Flux Balance Analysis (LK-DFBA), a modeling strategy to efficiently track metabolite dynamics[56]. LK-DFBA combines advantages of both constraint-based and ODE models,

unrolling the temporal aspect of the system into a larger stoichiometric matrix that captures metabolite dynamics while retaining a LP structure. The most critical element to accomplishing this goal is the addition of linear kinetics constraints that model the interactions between metabolites and the reactions whose fluxes they affect, including mass action kinetics and allosteric regulatory interactions. The number of parameters in LK-DFBA that need to be estimated can be far fewer than in ODE models due to these linear kinetics constraints. This enables LK-DFBA to potentially be applied to metabolic systems of all sizes, with a smaller increase in computational burden compared to ODE models. Furthermore, because LK-DFBA retains a linear structure, it can potentially be used with many existing metabolic modeling tools that require constraint-based models, such as OptKnock[126]. We have previously shown that LK-DFBA can outperform ODE-based modeling approaches when used in conditions most relevant to metabolomics data (low sampling frequency and high noise). A framework such as LK-DFBA that can model systems at the genome scale is essential to take full advantage of metabolomics data.

In our initial description of LK-DFBA, we explored two different approaches for model parameterization. The first approach, LK-DFBA (LR), parameterizes constraints solely via linear regression of interacting metabolite concentration and flux data. The second approach, LK-DFBA (LR+), uses the parameters from the linear regressions as initial seeding values for a secondary optimization to identify the optimal constraints for each interaction. While LK-DFBA (LR+) yields better fits to training data than LK-DFBA (LR), the latter approach estimates its parameters with trivial computational effort while still producing results that are similar in error to ODE models. As a result, LK-

DFBA (LR) may be the preferable approach for the efficient construction and parameterization of metabolic models at the genome scale.

However, the overall LK-DFBA framework still has some limitations in terms of how accurately it represents the underlying biology and biochemistry of the system. For example, the linear kinetics constraints used in LK-DFBA (LR) may be viewed as crude approximations of the interactions between metabolites and fluxes, which are typically non-linear in nature. While kinetic equations found in ODE models (such as Michaelis-Menten or biochemical system theory (BST) representations[97, 127]) can capture the non-linearity of these interactions, the current linear framework in LK-DFBA cannot. Additionally, when allosteric regulatory information is considered (which LK-DFBA includes in its framework), reaction fluxes are often controlled by multiple metabolites. Currently, LK-DFBA creates separate constraints for each metabolite that controls a flux, which precludes modeling how multiple metabolites simultaneously interact with a reaction flux.

Since the linear kinetics constraints are so critical to the function of LK-DFBA, it is likely that improving those constraints could have a substantial impact on LK-DFBA's ability to capture and predict biological phenomena. Accordingly, we devised three new types of kinetics constraints for LK-DFBA to account for biologically relevant features like non-linearity and simultaneous regulation by multiple metabolites. These new approaches were compared to the original LK-DFBA (LR) constraints by testing on synthetic model systems as well as models based on *Lactococcus lactis* and *Escherichia coli*[44, 50] metabolism. We also probed these constraint approaches for their robustness to model perturbation and their ability to predict phenomena not captured in training data.

We found that these new constraint approaches can improve model performance, and that the optimal constraint approach varied depending on the system being modeled but was consistent across perturbations for any given model. We also showed that the LK-DFBA approach chosen for the *L. lactis* and *E. coli* models can be used to predict changes in several critical metabolites and fluxes in agreement with literature experimental results. These improvements to LK-DFBA and demonstration of its effectiveness on new metabolic models support its attractiveness as a framework for modeling increasingly large metabolic systems in the future.

4.2 Linear Kinetics Dynamic Flux Balance Analysis (LK-DFBA)

Linear Kinetics-Dynamics Flux Balance Analysis (LK-DFBA) is a recently developed modeling strategy that is both scalable and capable of capturing metabolite dynamics. The full details of this approach have been described in detail previously[56], so we only outline the most important aspects of our framework here. In brief, LK-DFBA uses an LP-based structure with temporal dynamics modeled by discretizing time and unrolling the system into a larger matrix representing each timepoint separately, with an objective function that reflects the unrolling of the model. Linear inequality constraints that model mass action kinetics and metabolite-dependent regulation are included in the model; they are the driving force behind metabolite accumulation and depletion by limiting the maximum flux allowed based on the availability of metabolites over time. To parameterize these constraints, the LK-DFBA (LR) approach uses linear regression on assumed available metabolomics and fluxomics data, as described in the next section. If fluxomics data are unavailable, dynamic flux estimation (DFE) can be used to infer flux

values from concentration data[123]. In the LK-DFBA (LR+) method, the parameters from the LK-DFBA (LR) approach are used as initial conditions in a secondary optimization step that finds improved kinetics constraint parameters, though at the cost of computational time. Because LK-DFBA retains an LP structure, it is readily scalable and has the potential to be used with current constraint-based modeling tools.

4.3 Methods for Improving LK-DFBA

4.3.1 Constraint approaches

Throughout this work, we examined the original LK-DFBA (LR) approach and three new constraint approaches described in detail below.

4.3.1.1 LK-DFBA (LR)

The original LK-DFBA approach uses linear kinetics constraints to model the interaction between a metabolite and a flux, parameterized using available metabolomics and fluxomics data. These constraints take the form of $v < ax + b$, where v is the flux being constrained, x is the concentration of a metabolite that interacts with the flux, and a and b are the linear constraint parameters. These interactions may be due to mass action kinetics, where the interactions are known based on the stoichiometric topology of the system, or they may stem from allosteric regulation. While we have previously shown that these linear approximations of metabolic interactions can be effective for modeling metabolism, they are still approximations of the true non-linear and interconnected biochemical relationships in metabolism. Below, we discuss three new constraint approaches to address these potential limitations.

4.3.1.2 LK-DFBA (NLR)

While the key advantage of using constraint-based models is their LP structure that enables efficient identification of the optimal solution of the problem, most metabolite-flux interactions exhibit non-linear behavior that may not be captured well by linear equations. Recently, computational solvers have improved such that quadratically constrained programs (QCPs) are not much more computationally expensive than LPs. Accordingly, we implemented quadratic constraints into the LK-DFBA framework to explore their potential for improving model accuracy with only a modest increase in computational time. One important aspect of LPs and QCPs is that all of the constraints must create a convex feasible solution space in order to guarantee that a global optimum can be found[52]. If $v < ax^2 + bx + c$ represents a quadratic constraint, where *v* is the flux being constrained, *x* is the concentration of a metabolite that interacts with *v*, and *a*, *b*, and *c* are the parameters of the quadratic constraint, *a* must be a negative value to retain a convex solution space. If *a* is found to be a positive value during parameterization, we convert the quadratic constraint into its original linear form as found in LK-DFBA (LR). We refer to this overall approach as LK-DFBA (NLR).

4.3.1.3 LK-DFBA (DR)

Enzymatic reactions are often controlled by more than a single metabolite that can either induce or inhibit enzyme activity, which should ideally be captured in the model constraints. To model such regulation of a reaction by multiple metabolites, LK-DFBA (LR) creates individual linear constraints for each controller metabolite that are independent of each other and are thus unable to capture the synergistic or antagonistic

effects of multiple metabolites working in conjunction to regulate a flux. We implemented a new strategy that uses dimensionality reduction to consolidate information content from all controller metabolites for a flux into a single constraint. Dimensionality reduction is often used in data analysis, including analysis of metabolomics data, to more easily represent and digest datasets with many measured variables. Principal component analysis (PCA) is one of the most commonly used dimensionality reduction approaches; it linearly transforms the original variables into new, orthogonal composite variables called principal components that capture as much variance in the original variable data in as few principal components as possible[128]. Ideally, the first or first few principal components capture the majority of the variance in the original dataset, which allows one to focus only on those composite variables rather than all of the original variables at once. Here, we use PCA to capture the maximal variance of the controller metabolite data in a single principal component and use that composite variable as the regressor during linear regression with the target flux data. These new constraints are represented as $v < aPC_1 + b$, where *v* is the flux being constrained, PC_1 is the metabolite concentration data projected into the first principal component, and *a* and *b* are the constraint parameters. We refer to this dimensionality reduction approach as LK-DFBA (DR).

4.3.1.4 LK-DFBA (HP)

Another approach for modeling interactions with multiple metabolites is to use hyperplane constraints. Unlike LK-DFBA (DR), which always builds constraints in two dimensions (i.e. the target flux vs. the first principal component), the hyperplane constraint exists in (n + 1) dimensions, where n is the number of metabolites that control

a target flux. This approach may avoid loss of information content from metabolite data as is possible during dimensionality reduction: as the number of metabolites in an interaction increases, the likelihood of the first principal component not capturing the majority of variance in the data increases. The hyperplane constraint equation can be represented as $v < \sum_{i=1}^{n} a_i x_i + b$, where v is the flux being constrained, n is the number of metabolites that interact with v, x_i is the concentration of metabolite i, a_i is the constraint parameter for metabolite x_i, and b is another constraint parameter. We refer to the hyperplane approach as LK-DFBA (HP).

4.3.2 Translating constraints to contain training data

We found that translating the constraints such that all training data fall in the region under the inequality constraint decreased the possibility of the computational solver encountering infeasible solutions when simulating metabolite dynamics. Thus, for all LK-DFBA approaches, each constraint was translated to contain the training data by increasing the intercept of the constraint (i.e. the *b* parameter in LK-DFBA (LR), LK-DFBA (DR), and LK-DFBA (HP), and the *c* parameter in LK-DFBA (NLR)) until no training data were above each constraint.

4.3.3 Test models

In this work, we assessed several synthetic models and two biological systems described below.

4.3.3.1 Synthetic model

The first system we examine is a simple synthetic model with five metabolites and five fluxes that was derived from a branched pathway model used previously[56]. This system is based on an ODE-based modeling framework that uses power-law kinetics to represent reaction fluxes[97]. The kinetic equations for each pathway are shown in Figure 30. To create a variety of synthetic models with the same stoichiometric topology, we randomly generated *a* and *b* parameters in each kinetic equation. The parameters for each model can be found in Table 6. Time course metabolite and flux data were generated by solving the ODE system in MATLAB (2018b).

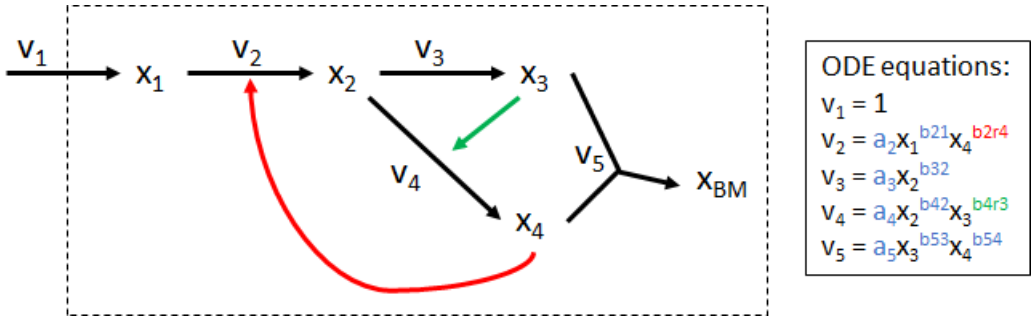

Figure 30: Synthetic model.
Adapted from another branched pathway model used in previous work[56]. v_1, v_2, v_3, v_4, and v_5 are system fluxes (black arrows) and x_1, x_2, x_3, x_4, and x_{BM} are metabolites, where x_{BM} is a metabolite representing biomass. Green and red arrows represent positive and negative regulatory interactions, respectively. ODE equations for the model are shown in the inset, where blue *a* and *b* parameters are mass action kinetic parameters and green and red *b* parameters are positive and negative regulatory parameters, respectively.

Table 6: Kinetic parameters in synthetic models.
Parameters were randomly generated using a uniform distribution between 0.1 and 1.0 (or -1.0 to -0.1 for b_{2r4}, which is a parameter that describes inhibition in the system). *k* represents the model number.

	Kinetics											Initial Conditions				
k	a_2	b_{21}	b_{2r4}	a_3	b_{32}	a_4	b_{42}	b_{4r3}	a_5	b_{53}	b_{54}	X_1	X_2	X_3	X_4	X_{BM}
1	0.8	0.5	-0.2	1.0	0.75	0.5	0.4	0.8	0.5	0.5	0.8	0.1	0.2	0.3	0.4	0.5
2	0.7	0.5	-0.2	0.8	0.2	0.2	0.9	0.7	0.9	0.4	0.9	0.3	0.3	0.6	0.5	0.2
3	0.1	0.5	-0.7	0.6	0.9	0.6	0.3	0.9	0.2	0.4	0.3	0.4	0.9	0.6	0.1	0.9
4	0.9	0.4	-0.2	0.6	0.8	0.4	0.9	0.6	0.5	0.2	0.7	0.8	0.1	0.5	0.7	0.8
5	0.2	0.4	-0.2	0.9	0.6	0.6	0.3	0.6	0.5	0.7	0.8	0.5	0.5	0.8	0.5	0.2
6	0.1	0.6	-0.3	0.3	0.9	0.5	0.1	0.7	0.6	0.8	0.6	0.2	0.9	0.2	0.1	0.3
7	0.5	0.2	-0.2	0.1	0.9	0.2	0.9	0.9	0.6	0.5	0.8	0.2	0.3	0.1	0.7	0.6
8	0.9	0.4	-0.3	0.8	0.1	0.2	0.9	0.7	0.2	0.6	0.8	0.6	0.4	0.1	0.4	0.4
9	0.6	0.5	-0.9	0.4	0.6	0.8	0.8	0.4	0.7	0.8	0.9	0.3	0.4	0.2	0.7	0.8
10	0.6	0.3	-0.9	0.3	0.2	0.6	0.6	0.7	0.6	0.8	0.9	0.4	0.2	0.8	0.1	0.7
11	0.8	0.2	-0.6	0.7	0.4	0.3	0.7	0.8	0.4	0.1	0.2	0.5	0.1	0.4	0.8	0.4
12	0.9	0.9	-0.1	0.1	0.4	0.7	0.3	0.3	0.8	0.1	0.8	0.1	0.9	0.9	0.6	0.1
13	0.2	0.9	-0.8	0.3	0.7	0.6	0.5	0.8	0.6	0.5	0.3	0.3	0.1	0.2	0.4	0.9
14	0.9	0.5	-0.9	0.1	0.8	0.2	0.7	0.3	0.5	0.8	0.5	0.9	0.7	0.3	0.5	0.9
15	0.6	0.8	-0.7	0.1	0.2	0.2	0.9	0.9	0.9	0.9	0.2	0.2	0.8	0.2	0.4	0.5
16	0.1	0.2	-0.7	0.8	0.5	0.5	0.9	0.4	0.3	0.2	0.6	0.8	0.8	0.2	0.1	0.5
17	0.3	0.4	-0.3	0.7	0.5	0.9	0.5	0.2	0.7	0.6	0.3	0.5	0.1	0.8	0.3	0.2
18	0.5	0.9	-0.4	0.3	0.6	0.4	0.2	0.3	0.7	0.5	0.6	0.9	0.4	0.6	0.2	0.9
19	0.9	0.8	-0.6	0.9	0.7	0.6	0.2	0.6	0.4	0.1	0.7	0.1	0.3	0.5	0.2	0.2
20	0.9	0.9	-0.2	0.1	0.7	0.3	0.3	0.5	0.6	0.4	0.7	0.4	0.8	0.2	0.3	0.4

4.3.3.2 Lactococcus lactis model

This model was created by Costa *et al*. and comprises central metabolism and production pathways for important metabolites such as mannitol and 2,3-butanediol [50]. The *L. lactis* model consists of 26 metabolites and 21 fluxes and is publicly available on KiMoSys[129]. Noiseless data were generated in COPASI 4.24 (Build 197) using the default initial conditions and parameters over a simulation time of two hours.

4.3.3.3 Escherichia coli model

The *E. coli* model developed by Chassagnole *et al*. encompasses glycolysis and the pentose phosphate pathway[44]. This model is publicly available on KiMoSys but was rebuilt within MATLAB to allow easy creation of new models that use the original *E.*

coli model's topology and stoichiometry. Noiseless data for the original *E. coli* model were generated in MATLAB (2018b) using the default initial conditions and parameters, while random initial conditions and parameters were used for the new models with the *E. coli* topology. To be consistent with our previous work, we used a simulation time of ten seconds[56].

4.3.4 Kinetic parameters in E. coli models

Parameters were randomly generated by drawing from the random normal distribution $N_i \sim (p_i,p_i)$ and taking the absolute value, where p_i is the original value of the *i*th parameter. Some parameters, such as the feed rate, dilution rate, and rates of synthesis reactions were kept at their original parameter values to ensure the models were viable.

4.3.5 LK-DFBA objective functions

Like other constraint-based methods, LK-DFBA requires an objective function, which is usually tied to some presumed goal of the system (such as maximizing biomass or ATP production). FBA models for specific organisms commonly have a separate flux reaction dedicated to biomass, made up of precise ratios of different metabolites. While LK-DFBA models with tuned objective functions can be created, the biological models we sought to use here do not have pre-existing tuned objective functions, so we instead focused on LK-DFBA's performance using generic objective functions.

Here, we have chosen flux v_5 as the objective function for the synthetic model, as it is the only efflux out of the system. For the *L. lactis* model, we use the LDH pathway as the objective function to maximize production of lactate because it is a key metabolite

in the organism (which is commonly used for dairy products) and was the metabolite produced at the highest levels in the original *L. lactis* model[50]. The objective function used for the *E. coli* model was to maximize all effluxes from the system, which included murein synthesis, glycerol-3-phosphate dehydrogenase, serine synthesis, PEP carboxylase, DAHP synthesis, pyruvate dehydrogenase, ribose phosphate pyrophosphokinase, glucose-1-phosphate adenyltransferase, the synthesis of murein and chorismate from PEP, and the synthesis of isoleucine, alanine, α-ketoisovalerate, and diaminopimelate from pyruvate. While we have observed that these objective functions can be further improved, and approaches have been developed for finding an optimal objective function for a model by creating a bilevel optimization problem and then leveraging the duality theorem[130, 131], our chosen objective functions were sufficient to at least qualitatively model the synthetic, *L. lactis*, and *E. coli* systems.

4.3.6 Pathway perturbations

To test the ability of LK-DFBA to predict metabolic behaviors not represented in the training data, we introduced perturbations into each system either through down-regulation (indicated with a prefix 'd' in all figures) or up-regulation (indicated with a prefix 'u') of reaction fluxes. For the synthetic models, we down-regulated v_2, v_3, and v_4 by multiplying their constraint equation parameters (which restricts their maximum allowable flux value) by 0.5x and up-regulated these pathways by doubling the constraint equation parameters. The pathways and reactions to be perturbed in the *L. lactis* [50, 132-135] and *E. coli*[45, 136-139] models were chosen based on previous literature. Reactions in the *L. lactis* model (lactate dehydrogenase, phosphofructokinase, acetate kinase, mannitol 1-

phosphatase) were down-regulated to 0.1x their original parameter values (since completely knocking out reactions would often produce infeasible solutions for the linear program) and up-regulated to 2x their original parameter values, magnitudes that were necessary to effect significant perturbations to the system's behavior. Reactions in the *E. coli* model (pyruvate kinase, phosphoglucose isomerase, glyceraldehyde-3-phosphate dehydrogenase, phosphofructokinase, triose-phosphate isomerase, ribulose-phosphate epimerase, phosphoglucomutase) were down-regulated to 0.1x and up-regulated to 2x their original parameter values.

4.3.7 Generating noisy data

Noise was introduced into the system by down-sampling the original noiseless data (originally 50 timepoints) into nT timepoints that are evenly spaced over the time interval of interest. Both metabolite and flux values were then replaced with a random value drawn from $N_{i,k} \sim (y_i(t_k), CoV \cdot y_i(t_k))$, where $y_i(t_k)$ is the value of species (metabolite or flux) i at timepoint k, and CoV is a coefficient of variance. For each sampling frequency and CoV condition, ten noisy datasets were generated.

4.3.8 Error calculation

The error of the predictions made by LK-DFBA was calculated using a normalized root mean squared error (NRMSE) between the LK-DFBA predicted metabolite concentrations and the noiseless ODE concentration or experimental data. P_{ik} and R_{ik} are the predicted (e.g. results from LK-DFBA) and reference (e.g. ODE model or experimental) data from a system with m metabolites and nT timepoints. $\bar{R}_i$ is the mean

of the concentrations of reference metabolite *i* across all timepoints to normalize the data and *N* is the total number of data points used in the NRMSE calculation.

$$\boldsymbol{NRMSE} = \sqrt{\frac{\sum_{i=1}^{m}\sum_{k=1}^{nT}\left(\frac{P_{ik}-R_{ik}}{\bar{R}_i}\right)^2}{N}} \qquad \text{(Equation 7)}$$

4.3.9 Pearson correlation calculation

The available *E. coli* knockout experimental data consisted of steady-state flux data, so to compare these to the knockout predictions made by LK-DFBA (which did not yield a steady state over the ten second time interval of the model) we used the average flux of our time course predictions. Because the average flux of our predictions and the steady-state fluxes of the experimental data are different measurements and therefore not directly comparable using NRMSE, we chose to use a Pearson correlation coefficient to evaluate our framework, which was recently used in a similar comparative analysis of metabolic models[140]. High correlations between steady-state flux experimental data and the average flux predictions would indicate that LK-DFBA can effectively predict if gene knockouts lead to an increase or decrease in flux for modeled reactions. The calculation for the Pearson correlation coefficient is shown in Equation 8, where A_i is the average of the predicted flux profile for the *i*th flux, v_i is the flux value of the *i*th flux from the experimental data, $\bar{A}$ is the mean across all fluxes for the average of computationally predicted fluxes, $\bar{v}$ is the mean flux value across all fluxes for the experimental data, and *n* is double the number of fluxes that are shared between both the *E. coli* model and experimental data because it includes flux values before and after the gene knockout (*n* =

28).

$$\boldsymbol{Pearson\ Correlation\ Coefficient} = \frac{\sum_{i=1}^{n}(A_i-\bar{A})(v_i-\bar{v})}{\sqrt{\sum_{i=1}^{n}(A_i-\bar{A})^2\sum_{i=1}^{n}(v_i-\bar{v})^2}} \qquad \text{(Equation 8)}$$

4.4 Results

4.4.1 Fitting and predicting phenotypes in synthetic models

We generated twenty random sets of parameters and initial conditions for the kinetic equations in the synthetic model to examine if different constraint approaches were more suitable for different models. We produced *in silico* metabolite concentration and flux data over a time interval of ten seconds by solving the ODEs in each synthetic system. The four constraint approaches were used for parameterization of LK-DFBA models to the twenty datasets. The fitted LK-DFBA models were then simulated over the same ten second interval using the initial conditions for each respective synthetic system to compare against the original ODE data. This process was performed on both noiseless (nT = 50, CoV = 0) and noise-added data with different sampling frequencies (nT = 50 or 15) and levels of noise (CoV = 0.05 or 0.15). To test the ability of each LK-DFBA approach to predict the effects of defined genetic perturbations, we down- and up-regulated the v_2, v_3, and v_4 pathways in the original kinetic equations by multiplying the kinetic coefficient parameters (*a* parameters in the inset of Figure 30) by 0.5x or 2x, respectively, and generating new ODE data. We then simulated the LK-DFBA model after adjusting the fitted LK-DFBA constraints to reflect the down- or up-regulation by multiplying the kinetics constraint parameters by 0.5x and 2x, respectively. The NRMSE between the predicted LK-DFBA metabolite concentrations and the ODE concentration

data from the perturbed synthetic models was then calculated.

For the noiseless cases with no genetic perturbation (WT) as shown in the first row below each bold line in Figure 31, the best-fitting constraint approach (dark green) varied across the different models. All four approaches performed best for at least one of the models. When fluxes were either down- or up-regulated via *in silico* genetic perturbations and LK-DFBA models fitted to the WT ODE data were used to predict these changes, the best constraint approach across all perturbations (dV2 through uV4) was generally consistent with the best approach in the absence of perturbations.

When using noisy data, similar trends were observed (representative example in Figure 32). While the best constraint approach for WT noisy data was not always the same as the best approach for noiseless data, the best constraint approach for a given noisy WT dataset was still generally the best for predicting the impacts of *in silico* genetic perturbations in the same model (dV2 through uV4). Interestingly, noisy data negatively affected the performance of LK-DFBA (HP) to a much greater extent than the other approaches, which caused LK-DFBA (HP) to never be identified as the best approach in the condition with the lowest sampling frequency and highest noise (nT = 15, CoV = 0.15). Similar results were found under other noisy conditions (Figure 50 through Figure 52).

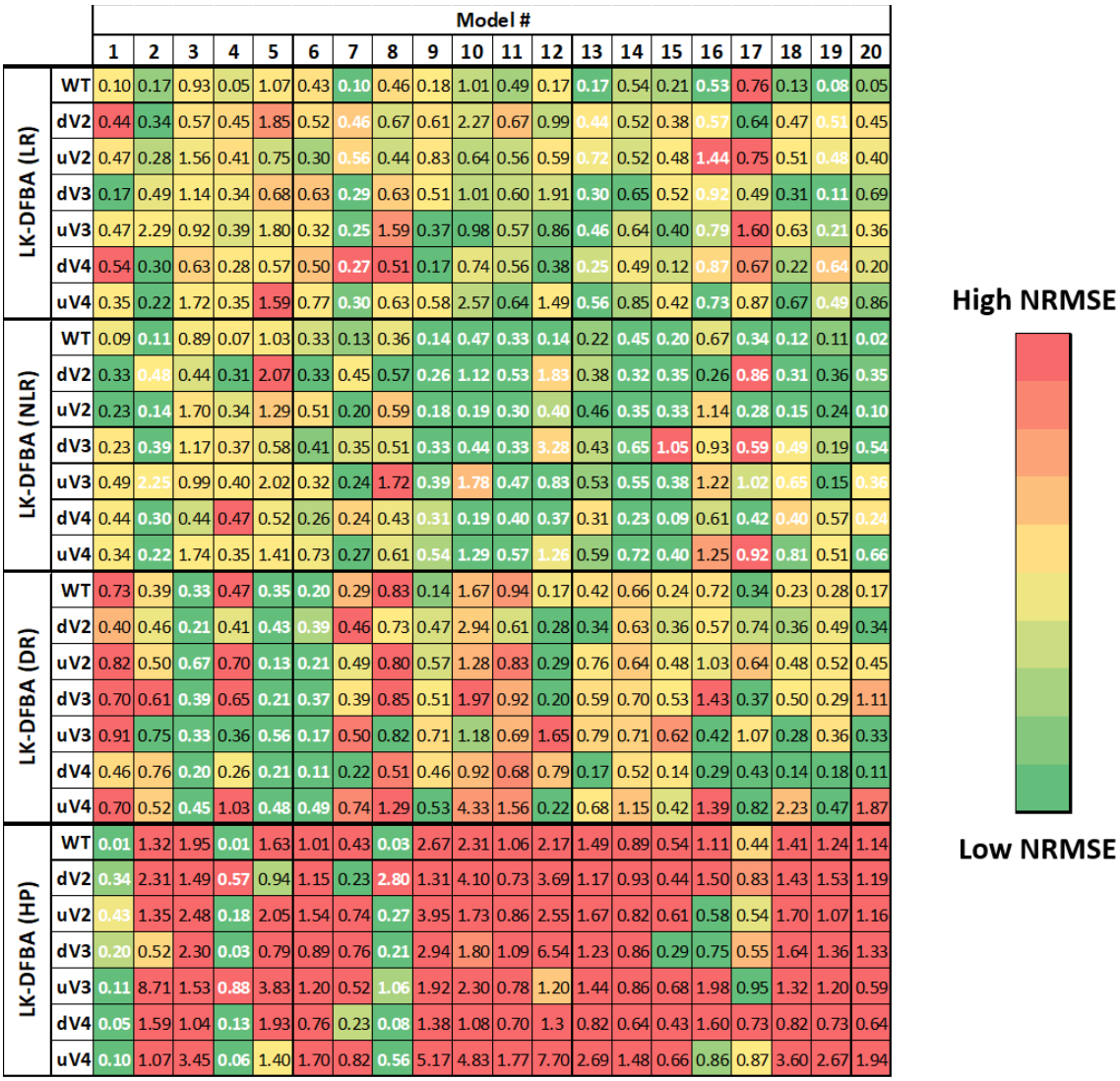

Figure 31: NRMSE heatmap of LK-DFBA approaches on different synthetic models using noiseless data.

Each constraint approach was used to fit parameters to wild-type (WT) noiseless data (50 timepoints) and then used to simulate the WT system and *in silico* genetic perturbations with fluxes v_2, v_3, or v_4 down- or up-regulated. Dark green boxes represent the lowest NRMSE within each perturbation for each synthetic model, while dark red boxes represent the highest NRMSE (meaning that the dynamic range of the color scale varies for each perturbation for each synthetic model to better convey the relative performance of different methods). The cells with bolded white numbers indicate the LK-DFBA approach that best fits the WT data. Cells with white numbers are generally consistently green, indicating that fitting to WT data is a good indicator of which approach will be optimal across all perturbations.

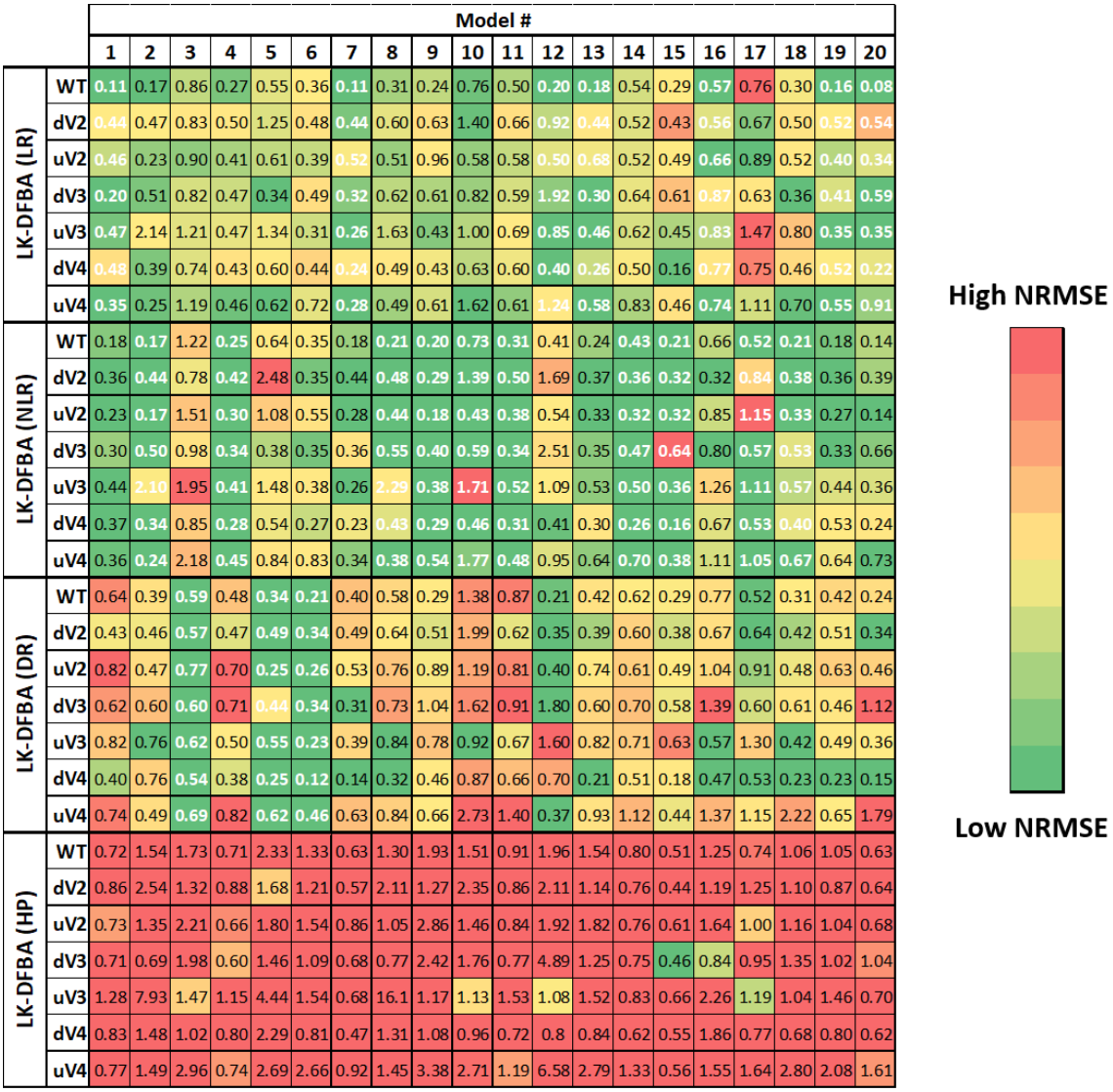

Figure 32: NRMSE heatmap of LK-DFBA approaches on different synthetic models using noise-added data (nT = 15, CoV = 0.15).
Each constraint approach was used to fit parameters to noisy (nT = 15, CoV = 0.15) wild-type (WT) data and then used to simulate the WT system and the system with *in silico* genetic perturbations with fluxes v_2, v_3, or v_4 down- or up-regulated. Dark green boxes represent the lowest average NRMSE (N = 10) within each phenotype for each synthetic model, while dark red boxes represent the highest average NRMSE. The cells with bolded white numbers indicate the LK-DFBA approach that best fits the WT data.

We also tested the effect of smoothing the noisy (nT = 15, CoV = 0.15) metabolite concentration and flux time course profiles by fitting to a previously described[100] impulse function (Figure 33). Smoothing the noisy data can often lead to

lower error of the final model but requires increased computation time for estimating the parameters of the impulse function and in certain cases can actually increase error if a specific dataset deviates significantly from all of the profiles that an impulse function can capture. The best constraint approach for WT smoothed data was the same as for unsmoothed data in 19 of the 20 models. As with the unsmoothed cases, the best constraint approach for smoothed data was typically consistent between WT and *in silico* genetic perturbations, and there were no cases where LK-DFBA (HP) performed the best (and it was generally the worst out of the four approaches) for smoothed data.

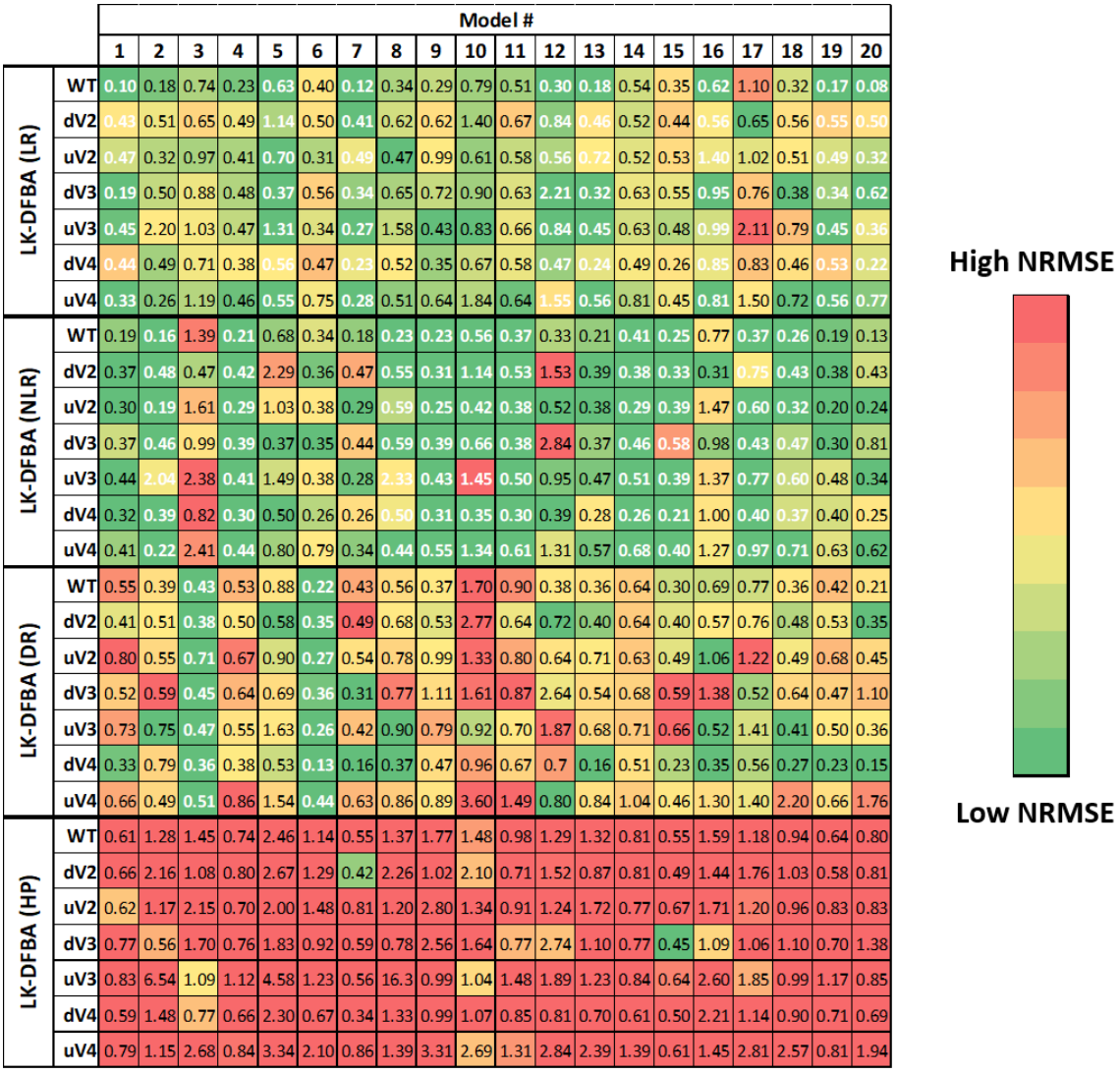

Figure 33: LK-DFBA performance on noisy synthetic model data after smoothing. Each constraint approach was used to fit parameters to noisy (nT = 15, CoV = 0.15) wild-type (WT) data that was smoothed with a previously described impulse function[100], and then used to simulate the WT system and the system with *in silico* genetic perturbations with fluxes v_2, v_3, or v_4 down- or up-regulated. Dark green boxes represent the lowest average NRMSE (N = 10) within each phenotype for each synthetic model, while dark red boxes represent the highest average NRMSE. The cells with bolded white numbers indicate the LK-DFBA approach that best fits the WT data.

4.4.2 Fitting and predicting phenotypes in L. lactis and E. coli models

For the *L. lactis* model, we tested the four constraint approaches on noiseless data and noisy data under different conditions (nT = 50 or 15, CoV = 0.05 or 0.15). On the noiseless data, the best constraint approach for the WT system was LK-DFBA (HP),

which also had the lowest NRMSE when predicting the results of perturbations to five different pathways (Figure 34). At high sampling frequencies and low noise (nT = 50, CoV = 0.05), LK-DFBA (HP) still performed the best, but as more noise was added or lower sampling frequencies were used, LK-DFBA (NLR) was optimal. This is consistent with the findings described above for the small synthetic systems where LK-DFBA (HP) can produce low NRMSE with noiseless data but has difficulties under more realistic conditions.

As with the *L. lactis* model, we tested all constraint approaches on both noiseless and noisy data from the *E. coli* model under different conditions (nT = 50 or 15, CoV = 0.05 or 0.15). For this model, LK-DFBA (NLR) was the best constraint approach for noiseless data (Figure 35). Noisy *E. coli* data produced the same results: for all noisy conditions, LK-DFBA (NLR) was optimal for the WT system. It was also optimal for almost all of the *in silico* genetic perturbations, showing once again that the same constraint approach that was optimal for the WT system at a given sampling condition was generally also optimal for the perturbed systems.

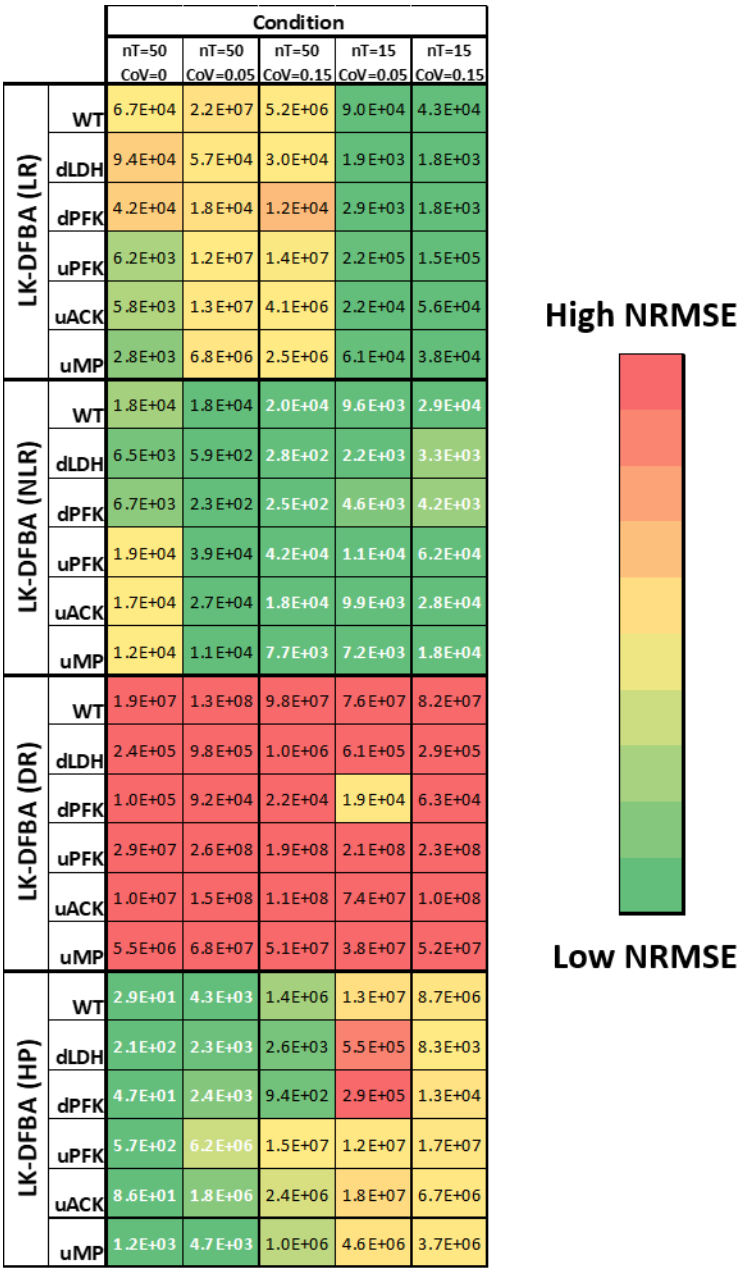

Figure 34: NRMSE heatmaps of constraint approaches on model of *L. lactis* metabolism.

Each constraint approach was used to fit parameters to wild-type (WT) *L. lactis* data and then used to simulate the WT system and the system with *in silico* genetic perturbations with literature-reported important pathways down- or up-regulated. Dark green boxes represent the lowest NRMSE within each phenotype for each model, while dark red boxes represent the highest NRMSE. Heatmaps show the mean of 10 noisy datasets, except for the noiseless condition (leftmost column).

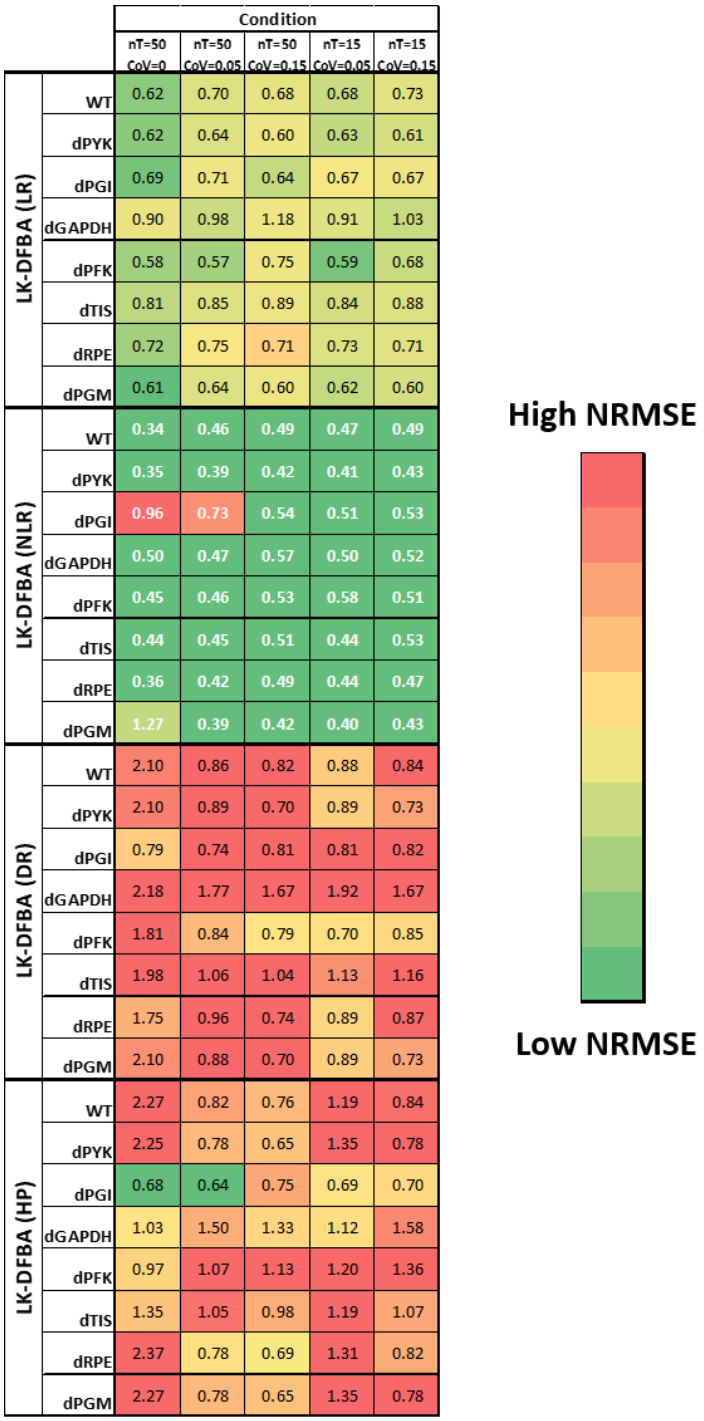

Figure 35: NRMSE heatmaps of constraint approaches on model of *E. coli* metabolism.

Each constraint approach was used to fit parameters to wild-type (WT) *E. coli* data and then used to simulate the WT system and the system with *in silico* genetic perturbations with literature-reported important pathways down- or up-regulated. Dark green boxes represent the lowest NRMSE within each phenotype for each model, while dark red boxes represent the highest NRMSE. Heatmaps show the mean of 10 noisy datasets, except for the noiseless condition (leftmost column).

We also perturbed the original parameters and initial conditions (drawing from the random normal distribution $N_i \sim (p_i, p_i)$ and taking the absolute value, where p_i is the original value of the ith parameter) of the *E. coli* model to create five new models with the same topology. As with the twenty different versions of the small synthetic system, we found that the best constraint approach was not conserved across models with the same topology as the original *E. coli* model when tested on noiseless data (Figure 36). Instead, the rates of individual reactions and how they affect overall model dynamics appear to be important factors in determining the optimal constraint approach.

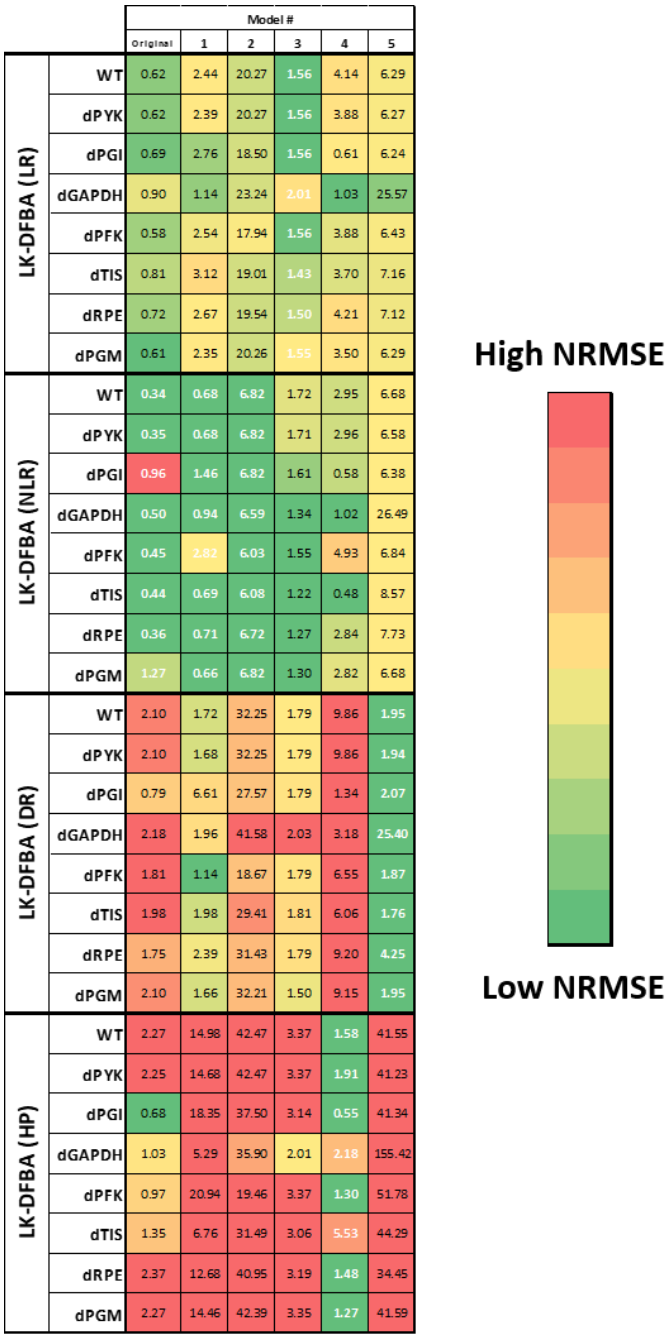

Figure 36: LK-DFBA performance on different models with the same stoichiometric topology as the *E. coli* model.

Five models with the same topology as the original *E. coli* model were created by randomizing the original kinetic parameters. The four LK-DFBA approaches were evaluated on noiseless data generated by these new models. Dark green boxes represent the lowest NRMSE within each phenotype for each model, while dark red boxes represent the highest NRMSE. The cells with bolded white numbers indicate the LK-DFBA approach that best fits the WT data. Cells with white numbers are generally consistently green, indicating that fitting to WT data is a good indicator of which approach will be optimal across all perturbations.

4.4.3 Improved LK-DFBA predictions yield qualitative consistency with experimental L. lactis metabolite concentration data

To further assess how well LK-DFBA performs when predicting different phenotypes, we compared the predictions of LK-DFBA to available experimental data for the first time. The previously described ODE-based *L. lactis* model was originally parameterized using experimental metabolite time course data from *L. lactis* cultures grown with an initial glucose concentration of 40 mM[50] and validated by comparison to experimental data from cultures grown at initial concentrations of 20mM and 80 mM glucose. Here, we similarly fitted all LK-DFBA approaches to data generated by the ODE model at 40 mM glucose and then simulated the LK-DFBA model using the best constraint approach at 20 mM and 80 mM initial concentrations of glucose for validation.

Figure 37 depicts the metabolite concentrations predicted by LK-DFBA (HP) (the best approach for noiseless data in the *L. lactis* model) for the three initial concentrations of glucose when trained on noiseless data. For multiple initial glucose concentrations, LK-DFBA (HP) captured the general qualitative trends of glucose (depletion) and lactate (accumulation), two key metabolites in *L. lactis* that are often studied[141, 142]. For cofactor metabolites that participate in many different reactions, such as ATP, NAD(H), and inorganic phosphate, it was more challenging for LK-DFBA (HP) to predict their concentration profiles over the simulation interval, which is a problem found in other modeling frameworks[52]. Although LK-DFBA's predictions were overall not as smooth or quantitatively accurate as the ODE model, this is to be expected due to the lack of a validated objective function for this constraint-based model; the objective function we used was a gross approximation that likely does not reflect the cell's true "goal", and it is

known that the objective function can significantly affect the predictions of FBA approaches. Nevertheless, as presented here, LK-DFBA can still qualitatively track important metabolite dynamics even when using a crude objective function. This is important to note, as many organisms that are not well-studied have no readily available objective function to use.

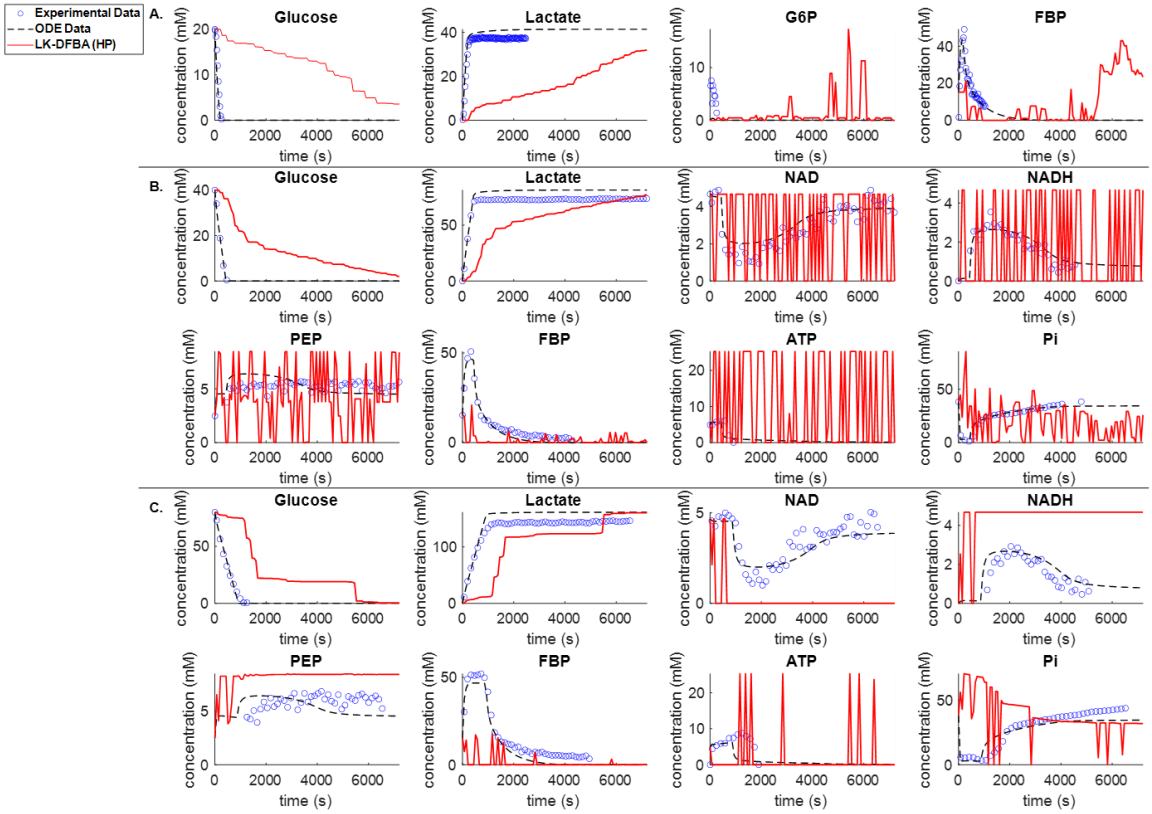

Figure 37: Comparison of LK-DFBA metabolite concentration predictions when fitted to noiseless ODE data against noiseless ODE data and all available *L. lactis* experimental data.

Panels A, B, and C depict concentration profiles for LK-DFBA (HP) and the ODE model compared to experimental data for initial glucose concentrations of 20 mM, 40 mM, and 80 mM, respectively. Cofactor concentrations were more challenging to predict, exhibiting many spikes upwards or downwards in concentration. Cofactors were involved in several kinetics constraints and the spikes in concentrations were likely due to changes in which constraints were currently active in the linear program. Nonetheless, the concentrations generally remained within an order of magnitude of the experimental data.

Figure 38 depicts the concentration profiles predicted by LK-DFBA (NLR) (the best approach for noisy data in the *L. lactis* model) after being fitted to 10 noisy datasets generated by the ODE model and simulated at 20 mM, 40 mM, and 80 mM initial glucose, respectively. Again, the LK-DFBA framework generally captured the qualitative trends of major metabolites such as glucose and lactate, though unsurprisingly not as accurately as when noiseless data are used and with difficulties predicting cofactor concentrations. Because LK-DFBA (NLR) contains quadratic constraints, its results are generally smoother compared to the other LK-DFBA approaches, which helped it predict some metabolites, such as PEP, arguably better than in the noiseless case. Furthermore, LK-DFBA (NLR) is less susceptible to noise for some metabolites, such as glucose and lactate, as observed in predicting similar time courses across the 10 noisy datasets. This could be advantageous if one is modeling a system with multiple noisy data sets and requires consistent predictions for certain metabolites. Likewise, if only using a single dataset, LK-DFBA (NLR) can ensure that these metabolic profiles would not dramatically change if a different dataset had been used. Other methods, such as the original LK-DFBA (LR) approach, can result in more varied predictions (Figure 39) depending on the noisy dataset used; some appear to produce better predictions than LK-DFBA (NLR), while others are worse (though all predictions follow the same trends). These observations reiterate that the best approach is dependent on the systems and datasets being studied, so having multiple LK-DFBA approaches available is an improvement over only using the LK-DFBA (LR) framework.

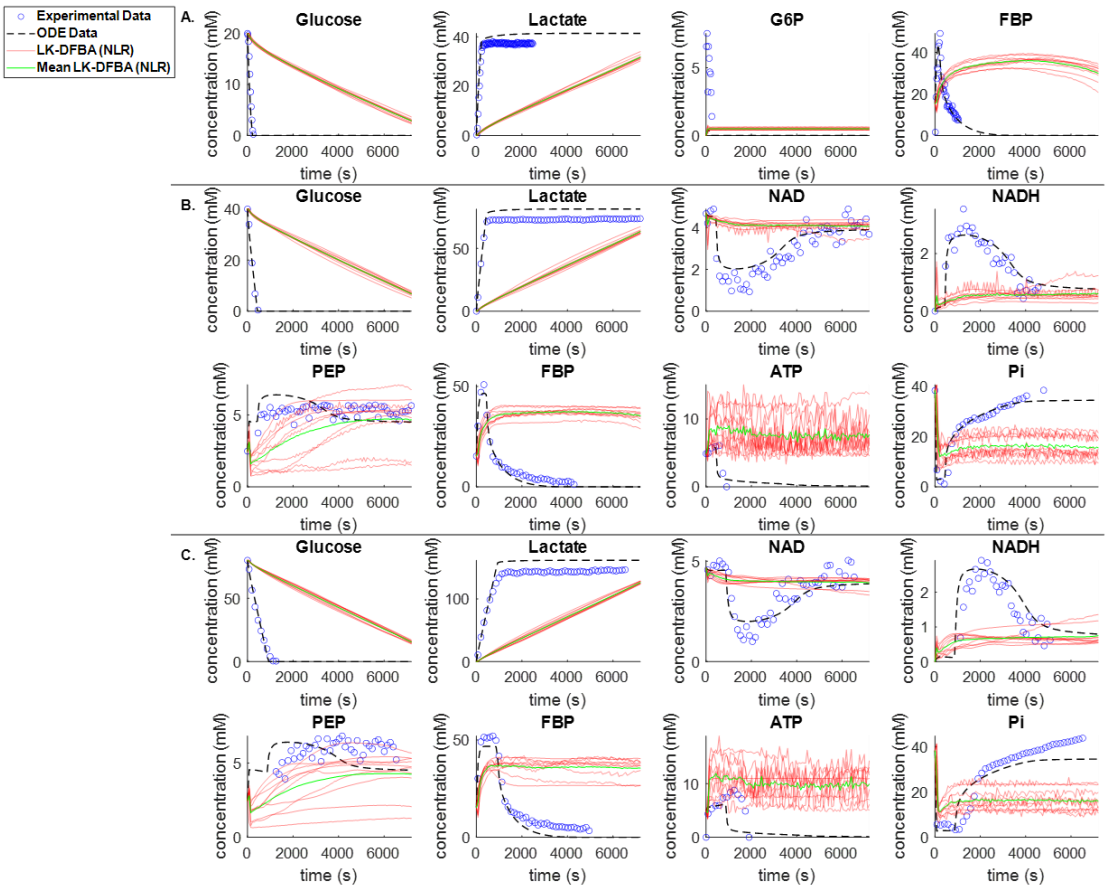

Figure 38: Comparison of LK-DFBA metabolite concentration predictions when fitted to noisy data against ODE and all available *L. lactis* experimental data.
A., B., and C. present concentration profiles for LK-DFBA (NLR) on 10 noisy datasets (nT = 15, CoV = 0.15) and the ODE model compared to experimental data for initial glucose concentrations of 20 mM, 40 mM, and 80 mM, respectively. The mean concentration profile (solid green line) is shown with each of the concentration profiles (solid red lines) from the 10 noisy datasets.

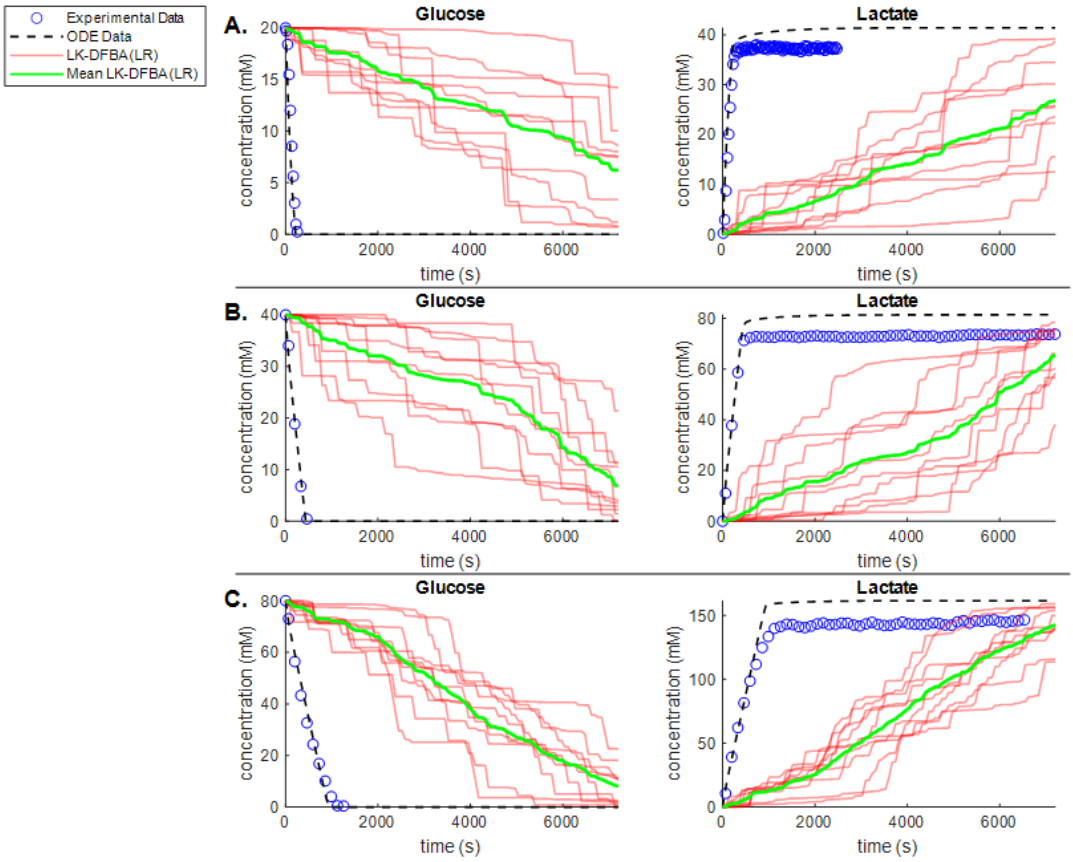

Figure 39: Comparison of LK-DFBA (LR) metabolite concentration predictions against ODE data and *L. lactis* experimental data when fitted to noisy ODE data. Panels A, B, and C depict concentration profiles for LK-DFBA (LR) on 10 noisy datasets (nT = 15, CoV = 0.15) and the ODE model compared to experimental data. The mean concentration profile (solid green line) is shown with each of the concentration profiles (solid red lines) from the 10 noisy datasets. While LK-DFBA (LR) is able to capture the general trends of glucose depletion and lactate accumulation, the best LK-DFBA approach on noisy data, LK-DFBA (NLR), showed much more consistency in its predictions across noisy datasets for glucose and lactate (Figure 38).

4.4.4 Changes in LK-DFBA flux profiles due to gene knockouts are correlated with experimental E. coli steady-state flux data

We also compared the predictions of the best LK-DFBA approach on the *E. coli* model to experimental steady-state flux data obtained through gene knockout experiments by Ishii *et al.*[137]. Because the Chassagnole model, which LK-DFBA is fitted to, only encompasses central carbon metabolism, we focused on 13 gene knockouts and

14 fluxes that are included in both the Chassagnole model and the Ishii steady-state flux results. We used the dilution rate of 0.2 h^{-1} for all experimental data. To emulate a gene knockout in the LK-DFBA model, we down-regulated the pathway(s) that correspond with the gene by multiplying the parameters of the relevant constraints by 0.1x instead of completely removing the reaction, as we found that this sufficiently reduced the possible flux reaction rate without causing infeasible solutions from the solver. Additionally, it is not uncommon for enzymatic activity to remain in a pathway after single gene knockouts due to paralogous enzymes and enzyme promiscuity. Because the LK-DFBA predictions do not reach steady-state for the simulation time examined in this work and our previous work (ten seconds), we instead used the average flux of the predicted time course to describe how LK-DFBA's predictions change from the wildtype to gene knockout phenotype. The average flux before and after a gene knockout should reflect whether the reaction rate generally increases or decreases across the studied time interval after a system perturbation. We used a Pearson correlation coefficient to determine if the average flux profiles predicted by LK-DFBA changed similarly to the experimental data after a gene knockout. This assessment method has been used previously by Lima et al. to compare multiple *E. coli* models, including the Chassagnole model, to the Ishii dataset[140].

To evaluate how our framework compares to *E. coli* experimental data, we examined LK-DFBA (NLR), as it was the best approach in the case of low sampling frequency and high noise (Figure 35). Figure 40 shows the average Pearson correlation of the LK-DFBA (NLR) flux predictions (after being fitted to ten noisy datasets with nT = 15 and CoV = 0.15) and the average correlation of the ODE model flux predictions with the experimental steady-state flux data. Of the gene knockouts and fluxes examined, LK-

DFBA (NLR) generally gave reliable predictions for whether fluxes increased or decreased due to gene knockouts, with correlation values greater than 0.6 in all but two cases and correlations greater than 0.7 in 6 out of 13 cases. These correlations were very similar to the correlations yielded by the ODE-based model. In 10 out of 13 knockouts, the correlations calculated for LK-DFBA outperformed or were within 5% of the correlations calculated with the ODE-based model. These results support the significant promise of LK-DFBA approaches for predictivity comparable to that of standard models but with the additional benefits (including relative model simplicity and potential scalability) that accrue from using a LP-based formulation.

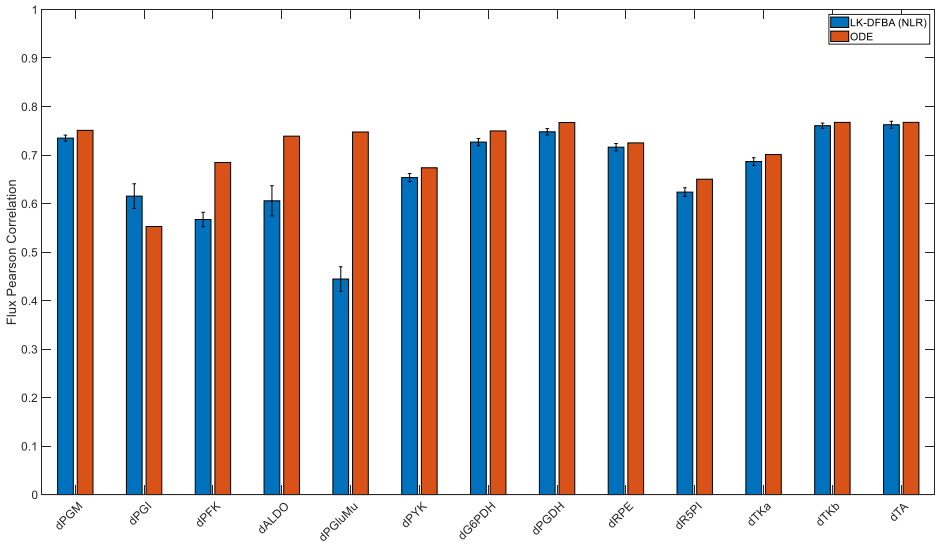

Figure 40: Pearson correlation coefficients of LK-DFBA and ODE model flux predictions with *E. coli* experimental data.
LK-DFBA (NLR) was the best approach when fitting on low sampling frequency (nT = 15) and high noise (CoV = 0.15) data. Blue and red bars represent LK-DFBA (NLR) and ODE model mean correlations, respectively, between the average predicted flux profiles and experimental steady-state flux data for various gene knockout conditions. Gene knockouts in the LK-DFBA and ODE-based models were simulated by down-regulating relevant pathways. Error bars for LK-DFBA represent one standard deviation (N = 10 runs).

4.5 Discussion

At the outset of this work, we sought to find a single LK-DFBA constraint approach that would improve upon the originally published framework. Instead, we have shown that the best constraint approach is highly dependent on the system being modeled. Despite each of the 20 small synthetic models having the exact same stoichiometry and allosteric regulatory interactions, the optimal LK-DFBA approach varied for both noiseless and noisy training datasets, with one of the new constraint approaches performing the best in the majority of cases. This finding suggests that the topology of the system is less important than the emergent dynamics from the collective metabolic reactions. It also supports the importance of having multiple types of constraints to choose from, as presented in this work, to allow more accurate modeling of any given system.

These conclusions are reinforced by analysis of biological systems, where LK-DFBA (HP) performs the best on *L. lactis* noiseless data and LK-DFBA (NLR) performs the best on *E. coli* noiseless data (though LK-DFBA (NLR) is superior for both systems when using low sampling frequency and high noise data). We further confirmed that topology is not the determining factor by randomizing parameters in the *E. coli* model (Figure 36): again, the best constraint approach varied across these topologically identical new models. Many metabolic pathways are conserved topologically across many species (e.g. glycolysis), though the kinetic parameters within these pathways can be vastly different. This suggests that having multiple LK-DFBA constraint approaches to choose from will improve our ability to model different systems.

While the best constraint approach varied across different model

parameterizations and topologies, the best approach (in terms of predicting metabolic phenotypes) for a given model was generally consistent across a wide range of pathway perturbations. This trend remained true whether using noiseless data, data with low sampling frequency and high noise, or noisy data that had been smoothed. These results instill confidence that the best constraint approach found when fitting to a wildtype metabolic system will also be the best approach when predicting changes to that system, meaning that an approach that can select the best-fitting of multiple constraint frameworks is viable and likely to be successful. One possible reason for the success of this approach is that when pathways are down- or up-regulated, it is common for only the nearest neighboring pathways to be significantly affected if the change to the system is not drastic or the perturbed pathway is not essential for cell survival, meaning that the emergent behavior from the system would not change too greatly and thus the same constraint approach would be optimal. To easily construct the optimal LK-DFBA model for a given biological system, we envision the workflow presented in Figure 41. After compiling the relevant system stoichiometry, regulatory information, and metabolomics and fluxomics data, one can fit each of the four LK-DFBA approaches to the data and determine which constraint approach most likely works best for predicting the results of different perturbations.

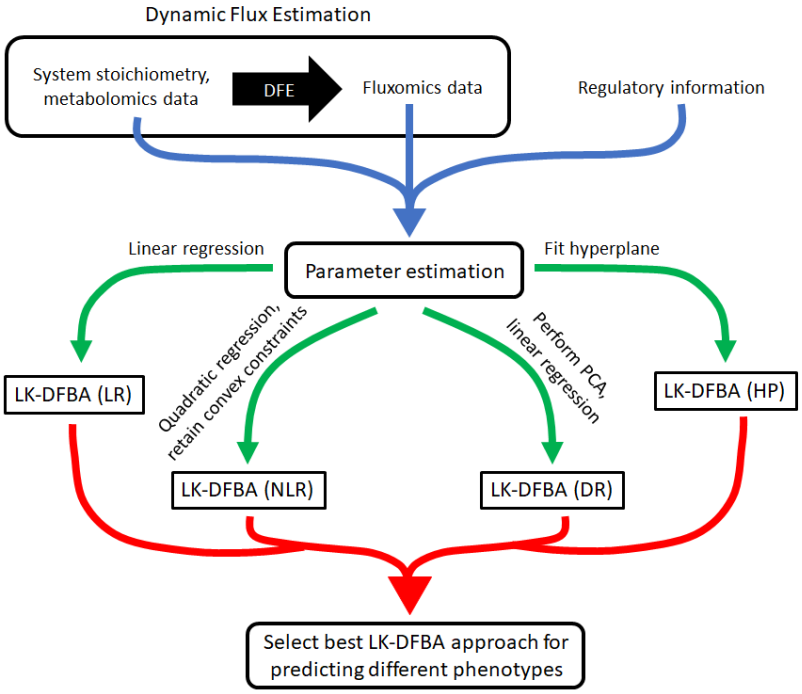

Figure 41: Workflow for selecting the best constraint approach for LK-DFBA when modeling metabolic systems.
Dynamic Flux Estimation (DFE) is applied to the system stoichiometry and available metabolomics data to infer instantaneous fluxes. The system stoichiometry, metabolomics data, inferred flux data, and system regulatory information are then used to estimate parameters for each LK-DFBA approach (blue arrow). Using multiple constraint approaches (green arrows), four different LK-DFBA models are created and tested for their respective abilities to recapitulate training data. The model with the lowest error is selected and can be used for future *in silico* predictions (red arrow).

Using ODE models and experimental data from *L. lactis* and *E. coli*, we found that LK-DFBA can effectively predict qualitative trends in concentration profiles of some important metabolites. While we have previously shown that LK-DFBA captures metabolite dynamics in synthetic data generated by ODE models, this is the first time LK-DFBA predictions have been validated with experimental data. For key metabolites that are important inputs or outputs of the system (e.g. carbon sources or end products), LK-DFBA can qualitatively predict if their concentration profiles are expected to decrease or increase, which is an important capability if one is using LK-DFBA to

engineer organisms to efficiently produce certain metabolites. Cofactors, on the other hand, are more difficult to model using LK-DFBA but are still typically predicted to be within an order of magnitude of the experimental data in most cases. This capability could be useful when assessing levels of accumulating toxic metabolites or cofactor imbalances if exact concentrations are not necessary. For LK-DFBA to accurately predict the metabolic profiles of cofactors, additional modifications to the framework may be necessary to account for metabolites that are involved in multiple reactions and are constantly consumed and produced.

We also found that LK-DFBA flux profile predictions were highly correlated with experimental flux data from genetic knockout experiments. Furthermore, these correlations were comparable to those found when using the ODE-based model. We note, though, that this comparable predictivity is limited by the fact that LK-DFBA was trained using ODE-generated data; if it had instead been fitted to actual metabolomics and fluxomics time course data used in the Ishii experiments (which is not available), these correlation values could possibly be even higher. Similarly, an improved objective function over the reasonable but arbitrary and unoptimized one used here could also lead to significant improvements in the performance of LK-DFBA.

By showing for the first time that LK-DFBA can predict changes in metabolite concentrations and flux profiles qualitatively, we have demonstrated LK-DFBA's potential as a widely-applicable metabolic modeling tool. Unlike many ODE-based modeling approaches that require specific kinetic equations for each flux reaction, LK-DFBA is generalizable. With four types of kinetics constraints that account for different biological interaction phenomena between metabolites and fluxes, we have improved

LK-DFBA to be amenable to many different systems. Additionally, applying the four LK-DFBA approaches to these models of *L. lactis* and *E. coli* has established that our framework can handle various biological systems of substantial size without the need for computationally taxing parameter estimation steps. Because each of the four LK-DFBA approaches maintains an easily solvable LP or QCP structure, LK-DFBA is a prime candidate for being one of the first frameworks able to model a variety of genome-scale systems while also capturing their metabolite dynamics.

While the addition of new constraint approaches has significantly improved the original LK-DFBA (LR) framework, there are still several areas where LK-DFBA can be improved. If computational resources when building the model are not a concern, a secondary optimization step can be used, as in the LK-DFBA (LR+) approach, to improve the parameters in each of the new constraint approaches. In addition, as previously noted the objective function used in LK-DFBA is also a ripe target for future efforts to improve this modeling framework. Here we have chosen objective functions that lead to the maximization of putatively important fluxes, but unlike many other constraint-based models, there was no specific biomass or other objective flux to use. Optimizing the weight of each flux or metabolite in the objective function could lead to even lower observed errors compared to experimental data and may also provide insight into what real biological systems tend to maximize.

4.6 Conclusions

In this work, we have shown that the LK-DFBA modeling framework can be improved by implementing more complex constraints with increased biological

relevance. We showed that there is no single best LK-DFBA constraint approach for all models, and the optimal approach depends not just on the topology of the biochemical system but also its kinetics and parameters. The constraint approach that performs the best in recapitulating training data consistently outperforms other constraint approaches at predicting the results of metabolic perturbations on the same system. With these new constraint approaches, we are able to model a variety of metabolic systems more accurately than if we were to just use the original LK-DFBA (LR) method. Moreover, based on comparisons to experimental data we showed that the improved LK-DFBA approaches can reasonably capture the qualitative dynamics of important metabolites and fluxes of interest to researchers. While these predictions may not be smooth or quantitative, the qualitative prediction of trends of metabolite dynamics in response to major perturbations is arguably the most critical aspect needed for creating metabolic models that give insight on how pathways can be further optimized or how metabolic resources can be rerouted to produce valuable chemicals: knowing that a specific knockout will increase or decrease flux is often sufficient to justify the expense of experimental implementation of such genetic perturbations. Moreover, we expect this computational framework to (with future effort) provide opportunities for computationally reasonable scale-up to the genome scale. While the acquisition of quality metabolomics and fluxomics data to build the constraints in LK-DFBA is still a challenge, the work we have presented here lays the groundwork needed to take full advantage of these types of datasets as they become increasingly more readily available.

CHAPTER 5: A Workflow Toward Modeling Dynamics in Metabolic Systems

5.1 Introduction

The three computational frameworks presented in the previous chapters were all built toward the goal of creating a comprehensive workflow for modeling metabolic systems, starting from the preprocessing of metabolomics data and ending with the construction of the model itself. While each framework was independently developed and may not be fully optimized to operate in unison yet, it is important to examine how the three platforms perform together in their current states. Not only will this further test the robustness of each framework, but it will also provide insight about possible areas of improvement that could lead to a more streamlined metabolic modeling workflow. In this section, I use a determined system with unknown regulation as a test case for combining SCOUR, MetaboPAC, and LK-DFBA to construct a metabolic model. This metabolic modeling workflow begins with relative abundance data that is used with SCOUR to predict the regulatory structure within the system. Next, this regulatory information is provided to MetaboPAC and the absolute concentrations are inferred from the relative abundances. Finally, LK-DFBA can construct a dynamic metabolic model with allosteric regulation using the regulatory information and absolute concentrations determined in the previous steps of the process. The workflow is depicted in Figure 42.

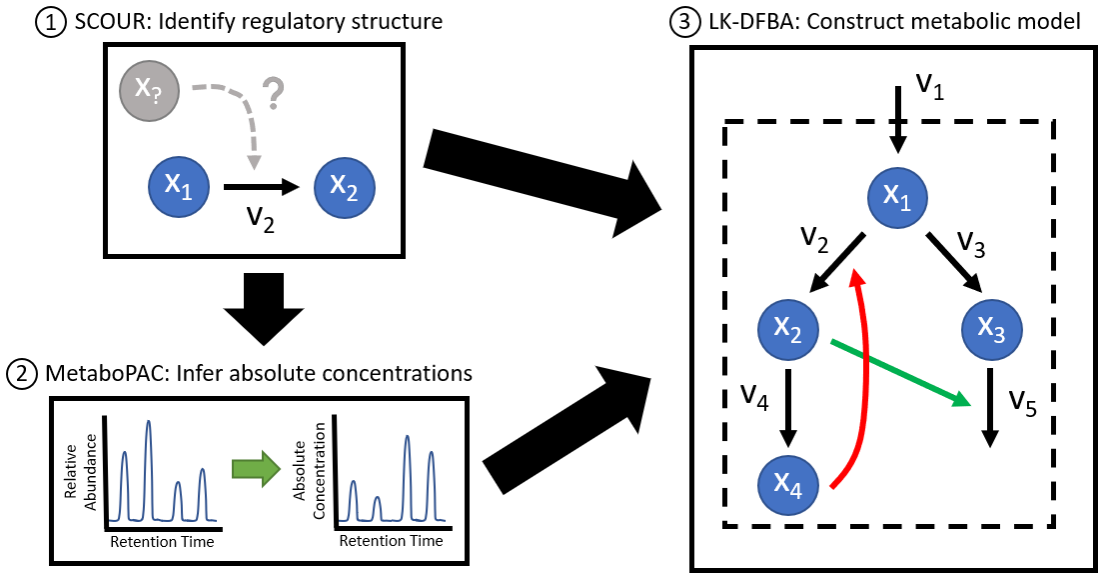

Figure 42: Workflow toward modeling dynamics in metabolic systems.
In the first step of the workflow, SCOUR identifies the regulatory structure of the system using relative abundance metabolomics data. Next, MetabocPAC uses the identified regulatory information to infer absolute concentrations. Finally, the identified regulatory structure and inferred absolute concentrations are used with LK-DFBA to construct a metabolic model.

5.2 Determined system with regulation

To assess LK-DFBA, SCOUR, and MetaboPAC together, we created a determined system containing four metabolites, fives fluxes, and two regulatory interactions that were initially unknown (Figure 43). Flux v_1 was assumed to be a constant known value. The reactions in the system were constructed using Michaelis-Menten kinetics and different initial conditions were used with the model to generate 15 datasets of metabolite concentration and flux time courses. Relative abundance data were simulated by randomly sampling 20 sets of response factors that were applied to each of the 15 ODE datasets for a total of 300 datasets that were used in both the SCOUR and MetaboPAC steps. For the LK-DFBA step of the process, only 20 datasets (simulated from different response factors but the same initial metabolite concentrations) were used to construct the LK-DFBA models and assess their ability to recapitulate the original

ODE data (i.e. the ODE dataset containing the true absolute concentration values and same initial metabolite concentrations).

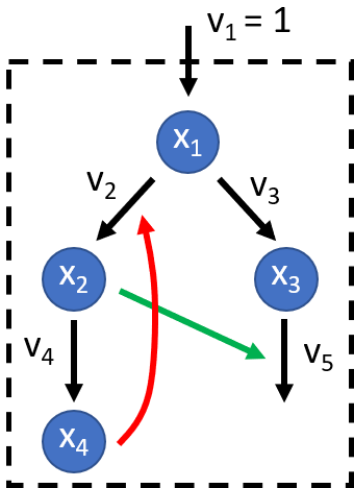

Figure 43: Determined system with regulation used to evaluate the metabolic modeling workflow.

5.3 Metabolic modeling workflow

5.3.1 Method for imputing missing metabolomics data

Before discussing the performance of all three frameworks together, I would be remiss if I did not mention a method for imputing missing values in raw metabolomics data that I developed during my thesis. In metabolomics data, it is common for some values to be missing due to random analytical instrument error (these values are described as missing completely at random (MCAR)) or because the abundance of a metabolite is below the limit of detection (these values are described as missing not at random (MNAR)). Because missing values in metabolomics data can bias downstream analyses or prevent the use of many analysis methods[143], researchers typically impute these missing values using several common methods, which include replacing missing values with zeros, the mean or median of all abundances, or using algorithms such as random forest or k-nearest neighbors (kNN)[75]. However, many of these imputation

approaches do not consider the difference between values that are MCAR and values that are MNAR, which can cause some missing values to be imputed inaccurately. To address this issue, we developed No Skip-kNN (NS-kNN)[144], a modified version of the original kNN formulation that determines which missing values are likely MCAR or MNAR and imputes them appropriately. Because MNAR values stem from sensitivity limitations from analytical instruments, NS-kNN imputes values identified as MNAR with lower abundances than the original kNN methodology. While NS-kNN was not used in the workflow presented in this chapter, it is an important tool in metabolomics data pre-processing and has already been shown in detail to effectively predict relative abundances of missing values better than several other imputation approaches.

5.3.2 Identifying metabolic regulation using relative abundances

I found that the regulation within the synthetic system could still be identified when using relative abundance data with SCOUR. When originally assessing SCOUR's performance in Chapter 3, the metabolomics data were assumed to be in terms of absolute concentrations and not relative abundances. This is a significant finding, as it may expand SCOUR's usability to the majority of current metabolomics datasets that have relative abundance values. SCOUR's performance is likely unhindered because relative abundances are essentially scaled values of the true absolute concentrations. Using relative abundances in lieu of absolute concentration would change how quickly the reaction rate of a flux would appear to increase or decrease based on the abundance of controller metabolite(s) (versus concentration), but it would not change the underlying kinetics or functionality of the reaction. For example, using relative abundances would

not cause a reaction to appear inhibited by a metabolite if the true absolute concentration actually induces the reaction. One limitation of using SCOUR in its current form is that the true flux reaction rates must be known to accurately predict regulation. When fluxes in the synthetic model were estimated directly from the determined system of mass balances (i.e. $\frac{dx}{dt} = S \cdot v$) using relative abundances, SCOUR could not identify regulatory interaction effectively.

When using SCOUR on noiseless relative abundance data (20 repetitions using different sets of response factors with each repetition containing relative abundance data simulated from 15 sets of initial conditions), we observed that SCOUR could accurately predict the regulatory structure of the determined system (Figure 44A). All performance metrics for noiseless data were above 97% when identifying interactions with one or two controller metabolites. For noisy data (nT = 50, CoV = 0.05), SCOUR was still able to identify regulatory interactions with PPVs above 65% (Figure 44B), which is similar to the results found with the synthetic systems in Chapter 3 under the same conditions (the system used in this chapter performed slightly worse for one-controller metabolite interactions, but slightly better for two-controller metabolite interactions). To keep this proof-of-principle study as simple as possible, we decided to move forward with the results from the noiseless data.

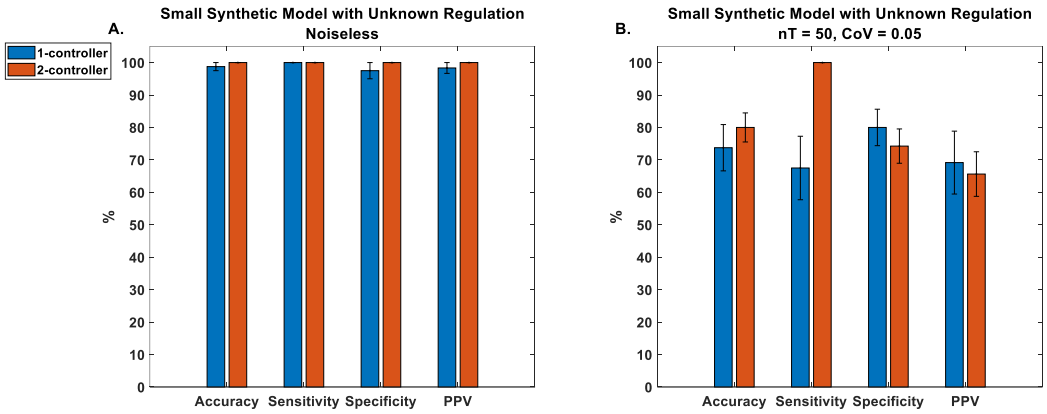

Figure 44: SCOUR performance using relative abundances from the small synthetic model with unknown regulation.
Accuracy, sensitivity, specificity, and positive predictive value metrics for SCOUR on A. noiseless and B. noisy (nT = 50, CoV = 0.05) data when predicting one- and two-controller metabolite interactions. Bars represent the mean performance (n = 20 for each set of different response factors) and error bars represent the standard error of the mean.

5.3.3 Inferring absolute concentrations from relative abundances and identified regulation

During my efforts to combine all three frameworks together, I found that SCOUR must be executed before MetaboPAC. When first attempting to combine all three frameworks together, I initially attempted to infer absolute concentrations from relative abundances using MetaboPAC before using SCOUR and presumed I would then identify the regulatory interactions in the next step. However, it soon became apparent that some of the penalties within the optimization approach in MetaboPAC require the regulatory structure of the system to be known *a priori* or else the penalties would be ineffective. For example, the correlation penalties in MetaboPAC require the number of controller metabolites of a reaction to be known, but if the regulation of the system is uncertain, these penalties cannot be accurately applied. For this reason, SCOUR must be the first step in the workflow so that the predicted regulatory structure can then be applied to MetaboPAC.

After identifying the regulatory topology of the determined system, we found that MetaboPAC was able to predict response factors with greater accuracy than randomly predicting response factors or using response factors of 500 when the kinetic structure of the system was unknown (i.e. only the optimization approach was used) (Figure 45A). If the kinetic equations of all reactions were assumed to be known, MetaboPAC predicted response factors with great accuracy (Figure 45B). 100% of the response factors were predicted within $\log_2(1.3)$ error if the kinetic structure of the system was fully known. We used the identified response factors from both the optimization approach and kinetic equations approach to calculate sets of inferred absolute concentrations that were used with LK-DFBA in the next section.

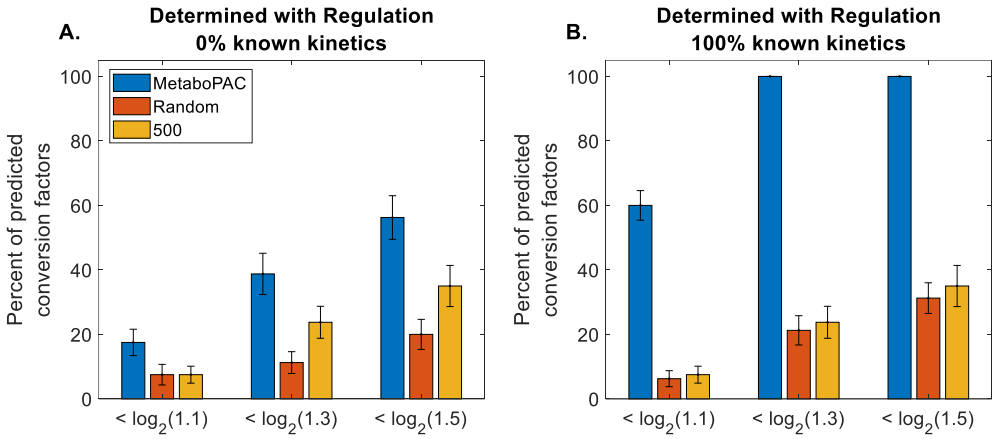

Figure 45: MetaboPAC performance on determined system with regulation. MetaboPAC compared to random response factors and response factors of 500 for the determined system with regulation using error ranges of $\log_2(1.1)$, $\log_2(1.3)$, and $\log_2(1.5)$ when A. 0% and B. 100% of the kinetic equations are known. Bars represent the mean percent of predicted response factors within the error ranges for each method. Error bars represent the standard error of the mean (n = 20 for different sets of true response factors).

When applying MetaboPAC in this workflow, we assumed that once the regulatory topology has been identified by SCOUR, the exact kinetics of the reactions

would also be known (if using the kinetic equations approach). In reality, there is a substantial amount of experimental work that needs to be performed to determine the kinetic formulation and kinetic parameters of a reaction, even if the substrates and allosteric regulators have been discovered. However, this is outside the scope of this thesis and the assumptions used here are reasonable for this proof-of-principle. Admittedly, one could argue that knowing the kinetic equations of all reactions in a system would defeat the need for LK-DFBA, which is why we apply both the results from the optimization approach and kinetic equations approach in the next section. However, it is important to reiterate that the results in Chapter 3 demonstrate that knowing only a percentage of kinetic equations can be beneficial and would further legitimize the use of LK-DFBA to model the remaining reactions in a system.

5.3.4 Creating a metabolic model using inferred absolute concentrations and identified regulation

Once both metabolite-flux interactions have been identified and absolute concentrations have been inferred, LK-DFBA can be applied to the determined system with regulation and a metabolic model can be constructed. Of the 15 ODE datasets used in the SCOUR and MetaboPAC steps, we chose one set (which includes 20 repetitions from different sets of response factors) to model with LK-DFBA. Along with the inferred absolute concentrations, the true flux profiles of each reaction were assumed to be known. We constructed four LK-DFBA models using the linear regression, non-linear regression, dimension reduction, and hyperplane kinetics constraint approaches and simulated each model using an objective function that maximized the fluxes through v_4

and v_5 (the two effluxes out of the system) to evaluate which approach could recapitulate the ODE data the best. Both the MetaboPAC absolute concentrations inferred using 0% and 100% known kinetics were applied to LK-DFBA.

The best LK-DFBA constraint approach was able to capture the general metabolic trends whether 0% or 100% of the kinetic equations were known. For the 0% known kinetic equations results, we observed that all LK-DFBA approaches were able to capture the general concentration profile trends of metabolites x_1 and x_3, but only LK-DFBA (LR) and LK-DFBA (NLR) could capture the trends of metabolites x_2 and x_4 (Figure 46). Similar trends were obtained when using the inferred absolute concentrations from the 100% known kinetics results (Figure 47). In both of these cases, LK-DFBA (LR) and LK-DFBA (NLR) could accurately capture the dynamics of each metabolite, though these two kinetics constraint approaches slightly underestimated the metabolite profile for x_1 when using only the optimization approach. Overall, LK-DFBA (LR) performed the best out of all the approaches for this system when using either 0% or 100% known kinetics results (Table 7), closely followed by the LK-DFBA (NLR) approach. LK-DFBA (LR) could track the metabolite dynamics of the system when 0% of the kinetics were known almost as closely as when 100% of the kinetics were known, highlighting the usefulness of the optimization approach in MetaboPAC. In Table 7, the standard error of the median when using the kinetic equations approach was substantially lower for each LK-DFBA model compared to the standard error of the median when using the optimization approach due to the former approach identifying similar sets of response factors more consistently across the repetitions.

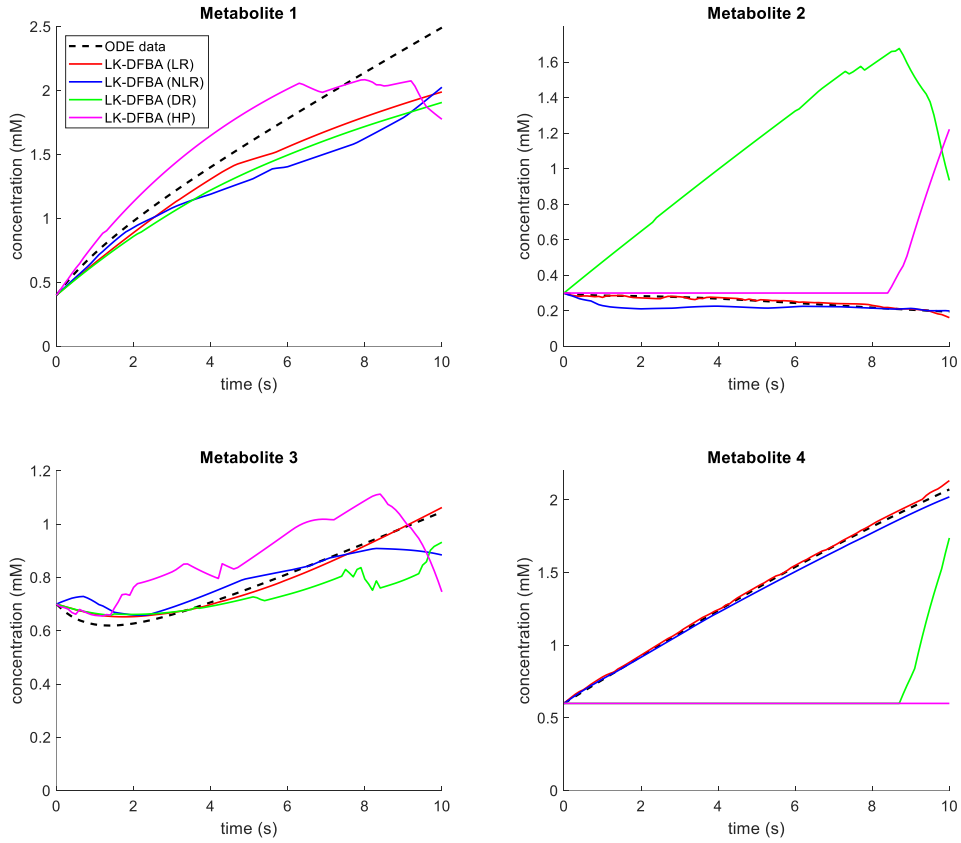

Figure 46: Comparison of LK-DFBA kinetic constraints on the determined system with regulation using inferred absolute concentrations from the MetaboPAC optimization approach.

Each LK-DFBA kinetics constraint approach was fitted to absolute concentration data (n = 20 for each set of response factors) inferred with MetaboPAC when 0% of the kinetic equations were known. The models were simulated to recapitulate the original ODE data and the median of these predicted metabolite concentration time course profiles was calculated and presented in this figure. The dashed line represents the original ODE data.

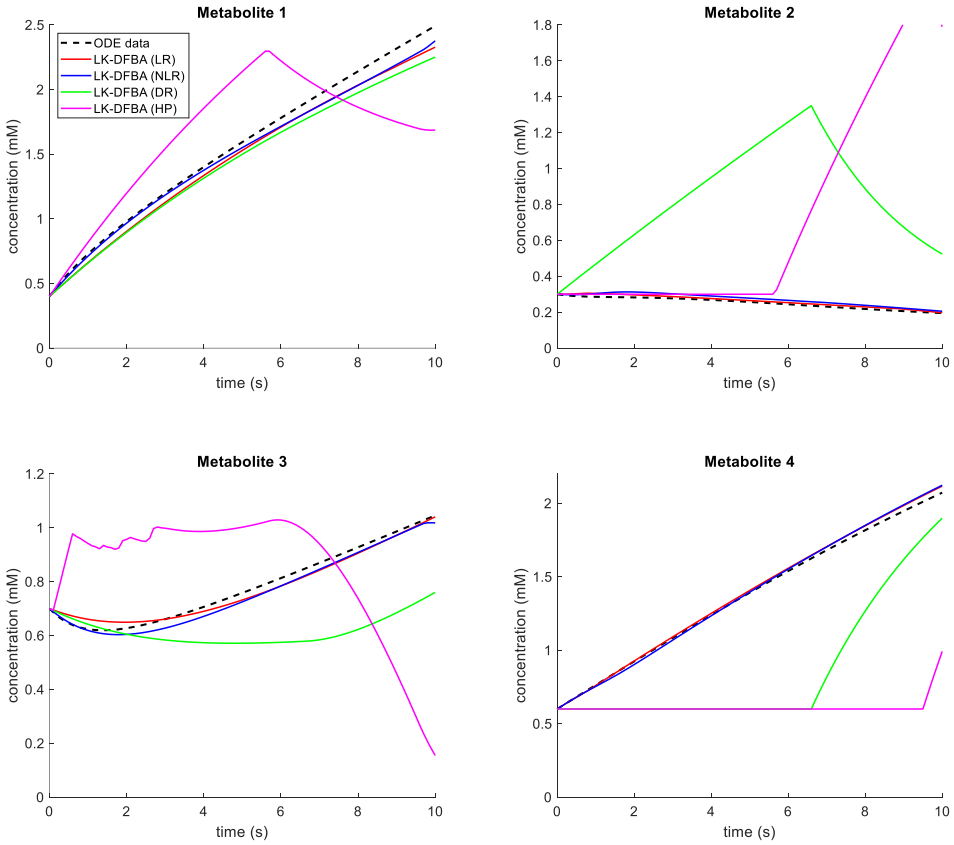

Figure 47: Comparison of LK-DFBA kinetic constraints on the determined system with regulation using inferred absolute concentrations from the MetaboPAC kinetic equations approach.
Each LK-DFBA kinetics constraint approach was fitted to absolute concentration data (n = 20 for each set of response factors) inferred with MetaboPAC when 100% of the kinetic equations were known. The models were simulated to recapitulate the original ODE data and the median of these predicted metabolite concentration time course profiles was calculated and presented in this figure. The dashed line represents the original ODE data.

We also fit each of the LK-DFBA approaches to the original ODE data and assumed all regulatory interactions were known (Figure 48). Interesting, we found that the LK-DFBA (LR) and LK-DFBA (NLR) models trained on inferred absolute concentration data (using the kinetic equations approach) and predicted regulatory interactions were more accurate than the LK-DFBA (LR) and LK-DFBA (NLR) models

trained on the ODE data with the correct regulation known *a priori* (Table 7). However, the differences in accuracy are negligible and the predicted profiles are very similar, which is not surprising because both SCOUR and MetaboPAC had predicted regulatory interactions (Figure 44A) and inferred absolute concentrations (Figure 45B) with very high accuracy.

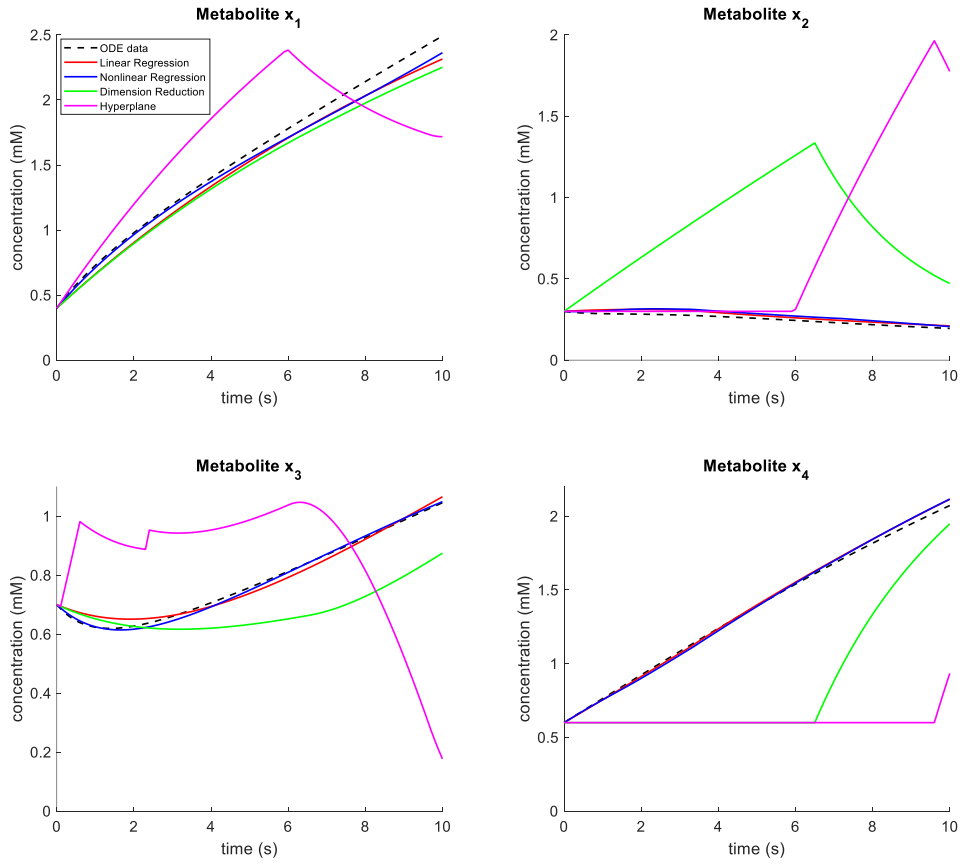

Figure 48: Comparison of LK-DFBA kinetic constraints on the determined system with regulation using the original ODE data and assuming all regulatory interactions were known.
Each LK-DFBA kinetics constraint approach was fitted to the original ODE data (n = 1). The models were simulated to recapitulate the original ODE data and are presented in this figure. The dashed line represents the original ODE data.

As a negative control, we fit each of the LK-DFBA approaches to the relative abundance data and assumed that the regulatory topology of the system was unknown (i.e. SCOUR and MetaboPAC were not used) (Figure 49). We found that LK-DFBA (LR) and LK-DFBA (NLR) performed substantially worse on relative abundance data with unknown regulation compared to our previous findings. These results demonstrate that SCOUR and MetaboPAC are significant steps in this metabolic modeling workflow. Interestingly, LK-DFBA (DR) and LK-DFBA (HP) improved in this case. This can be explained by the negative control assumption that there is no regulation in the system, meaning there are no reactions with multiple controller metabolites. Thus, the kinetics constraints in LK-DFBA (DR) and LK-DFBA (HP) are reduced to the constraints found in LK-DFBA (LR), which is why the NRMSE of these three approaches is equal in the negative control.

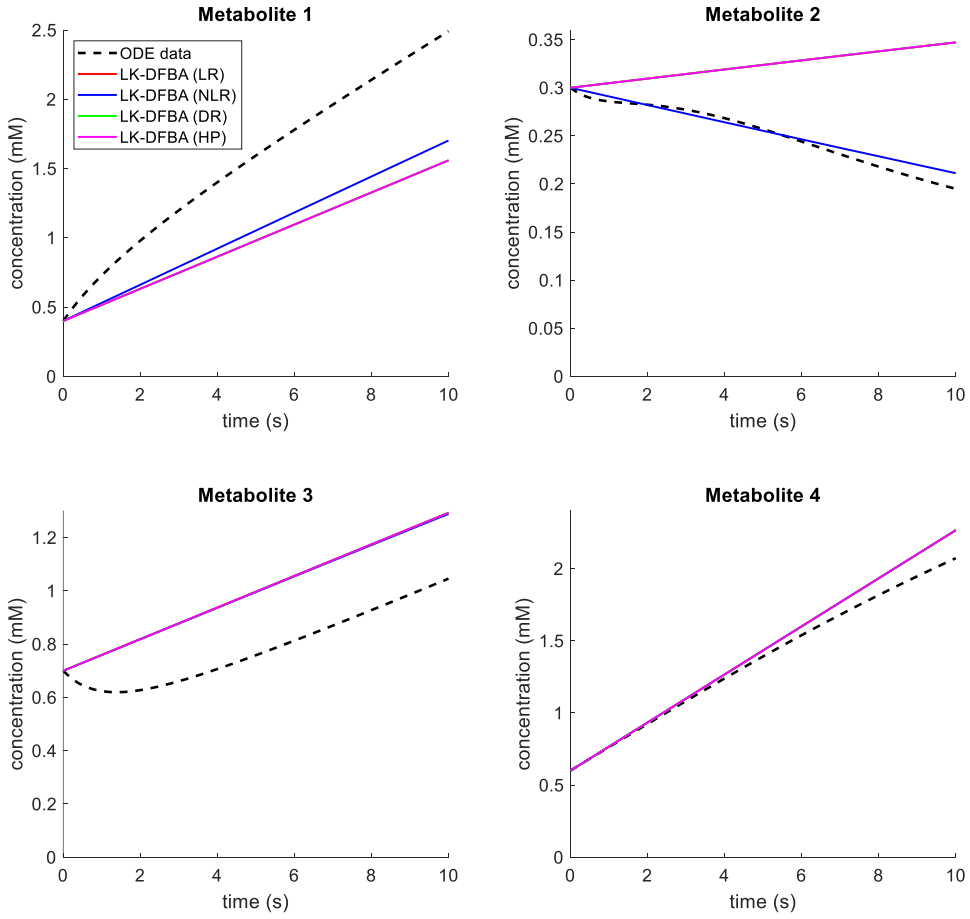

Figure 49: Comparison of LK-DFBA kinetic constraints on the determined system with regulation using the relative abundance data and assuming all regulatory interactions were unknown.
Each LK-DFBA kinetics constraint approach was fitted to relative abundance data (n = 20 for each set of response factors). The models were simulated to recapitulate the original ODE data and the median of these predicted metabolite concentration time course profiles was calculated and presented in this figure. The dashed line represents the original ODE data.

Table 7: Normalized root mean square error of the median concentration profile predictions by each LK-DFBA kinetics constraint approach compared to the ODE data.

The standard error of the median (n = 20) is provided to the right of the NRMSE for the conditions that used inferred absolute concentrations or relative abundances.

	LK-DFBA (LR)	LK-DFBA (NLR)	LK-DFBA (DR)	LK-DFBA (HP)
NRMSE 0% known kinetics	0.0820 ± 0.0409	0.1387 ± 0.0303	1.8832 ± 0.1979	0.6070 ± 0.1118
NRMSE 100% known kinetics	0.0397 ± 8.5766e-4	0.0513 ± 7.4302e-4	1.3288 ± 0.0104	1.6074 ± 0.0127
NRMSE ODE data and known regulation	0.0527	0.0582	1.2832	1.5070
NRMSE Relative abundances and no regulation	0.2991 ± 0.0018	0.2259 ± 0.0041	0.2991 ± 0.0018	0.2991 ± 0.0018

5.4 Conclusions

In this chapter, I have presented a proof-of-principle workflow that combines our three frameworks into a streamlined process for developing metabolic models. Despite the fact that the frameworks were developed individually and are not optimized to work together in their current states, the work in this chapter has illustrated that they can all be used collectively. Furthermore, we demonstrated that without SCOUR and MetaboPAC, the modeling accuracy of LK-DFBA is substantially diminished when using relative abundance data and no regulatory information. With some improvements, our modeling approach will be suitable for a variety of biological systems with little necessary *a priori* information about the system of interest. In the next chapter, we discuss in detail some of the key areas in each framework that should be explored to further improve their performances individually.

CHAPTER 6: Future Directions

Throughout this thesis I have presented three novel frameworks that solve some of the most significant challenges in modeling metabolic systems. I have developed an approach for identifying allosteric regulatory interactions, a method for inferring absolute concentrations, and improvements to a previous modeling framework. Together, these platforms create a cohesive workflow for modeling metabolite dynamics in systems of all sizes. In this chapter, I discuss the contributions of this research to the scientific community and several areas of improvement that should be explored in the future for each framework.

6.1 Thesis contributions

As stated at the outset of this thesis, the overall goal of this work was to create a streamlined process for modeling metabolic systems given only raw metabolomics data and the stoichiometry of the system. Currently, there are many different modeling frameworks available, but most assume that the available metabolic data used for modeling has already been pre-processed into a useable form and the regulation of reactions is known *a priori*. A comprehensive modeling framework that begins with processing of raw metabolomics data and ends with the modeling of an allosterically regulated system would be incredibly valuable to the modeling community and would be a significant step toward building a dynamic genome-scale model with metabolomics data. In this thesis, I have addressed three of the most important challenges facing metabolic modeling.

One of the most difficult aspects of modeling metabolic systems is that their regulatory topologies are often unknown, especially for systems that are not well-studied. To overcome this obstacle, I have developed SCOUR, a stepwise machine learning platform that can predict the regulatory structure of interactions. I demonstrated that SCOUR is particularly useful at classifying reactions that are only controlled by a single metabolite and can also identify two-controller metabolite interactions with accuracies that allow for experimental validation. SCOUR is the first machine learning approach to determine the allosteric regulation of metabolic systems and it will be a significant tool for modeling metabolic reactions accurately.

Another challenge in metabolic modeling frameworks is that although metabolomics data is primed to be a significant source of information for metabolic models, metabolomics is often omitted from these modeling platforms because the raw data are presented as relative abundances. To date, there have only been a few efforts that attempt to infer absolute concentrations from relative abundances without the need of chemical standards. In this thesis, I have presented a novel method, MetaboPAC, that infers absolute concentrations by leveraging the mass balances within a metabolic system. We determined that MetaboPAC can identify response factors (used to infer absolute concentrations) significantly more accurately than other methods when the kinetic equations of reactions are known. MetaboPAC is a powerful approach that will allow metabolomics data to be more easily integrated into metabolic modeling frameworks and other metabolic tools.

Finally, while there have been many modeling frameworks that can either model large-scale systems at steady-state or the metabolite dynamics of smaller-sized systems,

there have been very few platforms that can efficiently model metabolite dynamics at the genome-scale. Our group recently developed LK-DFBA, a linear programming framework that addresses this issue and can capture metabolite dynamics at all scales. However, the initial iteration of LK-DFBA uses crude approximations in its kinetics constraints to model the interactions between metabolites and fluxes. Here, I have present three new methods for constructing kinetics constraints that are more biologically relevant. We discovered that the optimal kinetics constraint approach was dependent on the system being modeled and that the best kinetics constraint approach for fitting to the wildtype data was typically also the best approach for predicting other metabolic phenotypes. Additionally, we determined for the first time that LK-DFBA could be used to predict general metabolic trends found in experimental data of two biological systems. The addition of new kinetics constraints will allow LK-DFBA to be used on a wider variety of metabolic systems in the future.

We have demonstrated that by combing the methods I have designed together, we can develop a dynamic metabolic model starting with only relative abundance data and the stoichiometry of the system. While the accomplishments in these aims have taken major strides toward achieving a cohesive metabolic modeling workflow, there are several areas of improvement discussed below that should be explored to make this an even more viable process.

6.2 Improvements to SCOUR

To the best of our knowledge, SCOUR is the first machine learning framework to use metabolomics data to identify allosteric regulation in metabolic systems. In Chapter

2, I demonstrated that SCOUR is particularly useful at predicting reaction fluxes controlled by one or two metabolites and is significantly better at predicting three-controller metabolite interactions than random classification. While these results are very promising, there are several avenues for improving the accuracy of SCOUR.

6.2.1 Different machine learning algorithms

In SCOUR, we use four different machine learning algorithms to construct the stacking classification model. In the first level of the model, random forest, k-nearest neighbors, shallow neural networks, and discriminant analysis were used to predict the set of controller metabolites that interact with a target flux. These outputs were fed to another discriminant analysis classifier in the second level of the stacking model that used the results from the four original algorithms to produce a final prediction. While these machine learning algorithms were found to be sufficient for the systems tested in this work and are some of the most common methods used in machine learning[145] due to their robustness, there are many more algorithms that should be tested and could lead to improvements in prediction accuracy. More specifically, these improvements would allow more regulatory interactions to be identified and lead to fewer false positives, which is particularly important when identifying three-controller metabolite interactions.

One potential algorithm is the support vector machine (SVM) approach. SVMs classify data using hyperplanes, known as support vectors[146], that separate classes by maximizing the distance between the data and the hyperplane. The most basic SVM method separates classes linearly, which can limit its effectiveness in many cases where non-linear classification is required. We initially tested SVM in an early version of

SCOUR but found it to classify metabolite-flux interactions poorly. However, there are many kernel functions that can be used with SVM to perform non-linear classification. Kernel functions transform and map the data to a different dimensional space so that SVM can more easily separate non-linear classes[147], but it is important to note that using kernel functions can lead to overfitting if they are not used correctly[148].

Another classifier that could be tested in the SCOUR framework is Naïve Bayes classification[149]. Naïve Bayes is a probabilistic algorithm that uses Bayes theorem to find the probability that a datapoint belongs to one class or another based on the given predictors (i.e. features). The biggest downside of using Naïve Bayes is that it assumes that the predictors are independent of each other, which is often not the case (and is not the case with the current features used in SCOUR). Nevertheless, Naïve Bayes has been used in metabolomics contexts before[150, 151]. Depending on the data being classified, there are several versions of Naïve Bayes, such as Bernoulli Naïve Bayes if features are Boolean or Gaussian Naïve Bayes if the features are continuous and exhibit a Gaussian distribution.

The second level classifier of the stacking model can accept many more than four inputs, so the addition of SVM, Naïve Bayes, or other machine learning algorithms could improve the prediction of the model. However, it is important to note that the inclusion of additional algorithms can also lead to bias toward specific classes of algorithms if there are multiple classifiers in the stacking model that are similar.

Another possibility for improvement of the overall approach could be the inclusion of unsupervised algorithms. All machine learning algorithms currently used in SCOUR and suggested in this section are supervised methods, meaning the user provides

class “labels” for the samples in the training data (i.e. the interactions in the training data are labeled as true positives or true negatives). Unsupervised learning does not require labeled classes and categorizes the data in groups or clusters of datapoints. Because the autogenerated training data are already labeled, it may not be appropriate to use unsupervised methods in the stacking model. However, unsupervised methods may be useful in developing new features that could improve the classification of different types of metabolic interactions. This process is called feature learning[152] and it can use unsupervised learning algorithms to identify information in the data that best separates true positive and true negative interactions. One common unsupervised machine learning method is principal components analysis[128], which linearly transforms the original variables in the data (e.g. concentrations, reaction rates, or features) into new orthogonal variables called principal components that capture as much of the variance in the data in as few variables as possible. Another unsupervised method is *k*-means clustering[153], which separates the data into *k* groups by identifying centroids in each group that optimally cluster the data. Besides feature learning, unsupervised learning methods are also often used in feature selection, whereby the features that contribute the most to the accuracy of the classification model are identified and the remaining features that do not contribute substantially, or even undermine classification, are removed[154].

6.2.2 Including more biologically relevant interactions in the training data

Along with creating a framework to classify metabolic interactions, we also developed a method for autogenerating biologically relevant training data. Metabolomics and fluxomics data can be difficult to obtain, which makes it a challenge to use machine

learning with these data, as most algorithms require an abundance of training information. The autogeneration method addresses this problem by emulating a multitude of different biological-like metabolite concentration and flux profiles. Our results demonstrate that these autogenerated datasets were able to successfully train the classification models used in SCOUR to predict regulatory interactions. To further improve the autogeneration method, some of the formulations for generating data could be modified.

To generate flux data, we use BST kinetic equations as the basis for all reactions in the autogeneration process. We chose to use BST because it is regarded as an all-purpose metabolic modeling approach that is flexible and can capture the kinetics of many different types of reactions[97]. However, many metabolic models use other types of kinetics, such as Michaelis-Menten or Hill kinetics, because they more closely represent the behavior of enzymes[155]. The two biological systems used to assess SCOUR use a mixture of Michaelis-Menten, Hill, and mass action kinetics, so it is possible that using other kinetic formulations in place of or in addition to BST would improve the potential for the autogenerated training data to lead to predictions of different types of metabolic interactions. In the future, when metabolomics and fluxomics data become more readily available, real data could supplement the autogenerated training data and provide additional insight about how biological interactions function and should be modeled.

6.2.3 Modification of autogeneration methods to include topological information

With all machine learning frameworks, there is always potential for improvement by adding or adjusting the features used to characterize the samples. Across the different

steps of SCOUR, we created different features that were specific to the number of putative controller metabolites that were being examined. One group of features that are not currently used in SCOUR but could be potentially applied across all steps are features that represent topological information about the system. The topology of the system could pertain to only the metabolites, the fluxes, or a combination of both.

Many possible topological features stem from graph theory, including the number of edges, centrality score, network diameter, network density, and vertex betweenness centrality[156]. Topological features have been previously used in other biological contexts with machine learning to predict novel metabolic pathways[99]. Because the topology of metabolites and fluxes is one of the defining features of a metabolic system, including this information could significantly improve the ability of SCOUR to identify regulatory interactions. The basis of including topological features is that reactions with similar topologies could also have comparable regulatory structures.

In its current state, SCOUR is unable to use topological information as features because the autogenerated training data has no inherent topology. As mentioned in the previous section, as both metabolomics and fluxomics data become more available, SCOUR is primed to use biological data as training data and topological features can then be implemented. Alternatively, the autogenerated training data method could be modified to include some pseudo-topological information for each artificial interaction. If topological information can be incorporated within SCOUR, several new types of features could be introduced to boost the accuracy of the framework

6.2.4 Using SCOUR on experimental data

To continue to prove SCOUR is a viable framework for identifying regulatory interactions, the next step is to test SCOUR on experimental data. Up to this point, we have assessed SCOUR on data simulated from two biological models and have added noise to emulate realistic data. When moving to experimental data, SCOUR should first be used on a well-studied system, such as *E. coli*, where the regulatory topology is well known. This will allow us to determine if SCOUR can truly identify regulatory interactions that are already known. Once this has been achieved, SCOUR can be used on other metabolic systems to discover new regulatory interactions that have not been established in the literature. Each of the new regulatory interactions predicted by SCOUR will need to be experimentally validated and through this process we can determine how many of these predicted interactions truly exist in the system.

6.3 Improvements to MetaboPAC

As the novelty of MetaboPAC is significant and the work presented in this thesis demonstrates its feasibility in a proof-of-principle context, the simplifying assumptions made in its implementation are well within reason. However, these assumptions entail some limitations that should be addressed in future work.

6.3.1 Non-linear relationships

Perhaps the biggest assumption of MetaboPAC is that the relationship between relative abundances and their absolute concentrations is linear. However, in reality some of these relationships may be nonlinear. Unlike LK-DFBA, which requires equality and

inequality constraints to be strictly linear (or quadratic if using the NLR method) to maintain an LP (or QP) structure, MetaboPAC does not use linear programming and already contains non-linearities within both the optimization and kinetic equations approach. This makes it easy to integrate non-linear absolute concentration relationships in MetaboPAC, whether they are simple polynomials or more complex. While the implementation of non-linear relationships is straightforward, determining which response factors should have linear or non-linear relationships is more difficult if not known *a priori*.

In the simplest non-linear scenario, we can assume that the equation used to infer absolute concentrations takes the form:

$$Relative\ Abundance = (Absolute\ Concentration)^{p} \times RF_T \qquad \text{(Equation 9)}$$

where p is some parameter that defines the non-linear relationship and RF_T is the true response factor. For linear relationships, p is set to one and this equation reduces to Equation 5. With the inclusion of p, the total number of parameters that need to be inferred (including the response factors) doubles. This will likely slow down MetaboPAC and make it more difficult to identify the response factors because not only are there more parameters, but it is also possible that identifiability of p could be more difficult (i.e. there are multiple values of p that are optimal). Nevertheless, the original underlying framework is still sound. The roles of the mass balances in the kinetic equations approach and the penalties in the optimization approach should not change whether the relationship between absolute concentrations and their relative abundances is linear or non-linear,

meaning that the biggest necessary adjustment is the implementation of Equation 9 in the mass balances and optimization penalties. One important matter to consider in the kinetic equations approach is that because there are double the number of parameters (i.e. unknowns), it is more likely for the system of non-linear equations to be underdetermined, and thus there may not be a unique solution for the response factors. A higher percentage of known kinetic equations or inclusion of more timepoints may be necessary to avoid this pitfall when considering non-linear relationships.

6.3.2 General improvements to the optimization approach

Both the kinetic equations approach and optimization approach are currently essential components of MetaboPAC. However, it is unsurprising that the kinetic equations approach can predict more accurate response factors, as it uses more mathematical and biological information. Unfortunately, in many biological systems, the kinetics of reactions are not generally known, making it critical to improve the optimization approach for effective predictions when there is minimal *a priori* biological knowledge. In some cases where 0% of the kinetics are known, we have observed that MetaboPAC still outperforms random response factors and response factors of 500, which illustrates that the optimization approach can be effective alone. There are a few opportunities to explore to improve the effectiveness of the optimization approach.

Because the principle behind the optimization approach is to eliminate response factors that infer absolute concentrations (and indirectly infer fluxes) that are not biologically feasible, there are several options for identifying infeasible response factors. First, any additional constraints to the minimum and maximum possible metabolite

concentrations and fluxes can greatly constrain the amount of allowable response factors when used in conjunction with the current penalties. However, this entails biological insight or information that may often not be available. Second, there have been other efforts in the FBA space that have attempted to remove flux distributions that are biologically infeasible whose methods could be translated to MetaboPAC. For example, several approaches have been developed to remove fluxes that are thermodynamically unlikely[37, 38] or have improbable metabolic energy totals (calculated as the norm of the flux distribution)[157]. Employing these methods in MetaboPAC could eliminate some unlikely response factors, but it is important to note that some of these methods assume that absolute concentrations are already known and thus would need to be adjusted accordingly to use relative abundances. Finally, the penalties used in the optimization approach could be further improved, whether it is the addition of new penalties that can identify poor response factors or adjusting the weights of each penalty.

6.3.3 Developing a platform for predicting confidence in inferred response factors

One significant extension to MetaboPAC that would improve its usability is a method for predicting the confidence in its inferred response factors. From our results, we found that the response factors identified when using the kinetic equations approach were relatively stable for all metabolites across the 48 repetitions of the non-linear least-squares solver with different initial seeds. In contrast, we observed that the optimization approach was often able to predict response factors for some metabolites more easily than others. The effectiveness of the penalties in the optimization problem are likely better for certain metabolite-flux interactions, which could narrow the range of biologically feasible

response factors for those particular metabolites. This would cause the width of the distribution of possible response factors in the 48 repetitions to be thinner for these metabolites and possibly increase the prediction accuracy when calculating the median of the response factor distribution.

To improve the usefulness of MetaboPAC, an auxiliary framework should be developed that assesses the confidence in the predictions generated by the optimization approach based on the distributions for each response factor. For example, thinner distributions would suggest higher confidence, whereas a wide distribution would signal low confidence. Other characteristics of the distribution of response factors that could be useful for predicting confidence include the standard deviation of the distribution, the number of peaks in the distribution, and the normality of the distribution. We currently only have an overall sense of the accuracy of the response factors inferred by MetaboPAC, but do not know which individual response factors are closest to their true values. Creating a platform for determining the confidence in the results produced by MetaboPAC would allow one to infer absolute concentrations of only metabolites for which the response factors were predicted with high confidence.

6.4 Further improvements to LK-DFBA

In this thesis, we have significantly improved on the original LK-DFBA framework, introducing three new approaches for constructing kinetics constraints. We have also demonstrated for the first time that LK-DFBA is a feasible platform for predicting different metabolic phenotypes in two biological systems. Because it is the kinetics constraints that allow LK-DFBA to capture metabolite dynamics, up to this point

we have focused on how these kinetics constraints can be improved. However, along with the constraints that define the feasible search space of the linear program, the objective function is the other key component in a CBM. Here, we discuss how optimizing the objective function in LK-DFBA could significantly improve modeling performance.

In CBMs, most stoichiometric matrices of biological systems are underdetermined (i.e. there are a greater number of reactions than metabolites), meaning there are an infinite number of possible flux profile solutions to the problem. To identify the most plausible flux profile, an objective function is used that is often based on some cellular goal, such as maximizing biomass. One could argue that because the objective function identifies a single best solution, it is as important as the equality and inequality constraints that define the feasible search space in the LP.

In our work to improve LK-DFBA with new kinetics constraints, we tested LK-DFBA on one synthetic system and two biological systems. For the synthetic system, we chose the only efflux out of the system as the objective function, as it seemed the most logical choice. For the *L. lactis* system we assumed the objective function was to maximize lactate production and for the *E. coli* system we assumed the objective function was to maximize all of the effluxes out of the cell. CBMs of biological systems typically use an objective function that maximizes a dedicated biomass flux reaction[158], but because we constructed the biological LK-DFBA models using information from two kinetic models rather than constraint-based models, there was no predetermined objective function. Nevertheless, the objective function we chose led to results sufficient to demonstrate that LK-DFBA could model metabolite dynamics in both organisms. However, the accuracy of the models could be improved if each objective function were

optimized instead of using our best guesses. In addition to examining how the objective function in LK-DFBA can be optimized, in this section we also discuss how modeling of cofactor metabolites can be improved and how LK-DFBA can be applied to metabolic engineering.

6.4.1 Existing methods for optimizing objective functions

Optimizing objective functions is not a novel idea in the realm of metabolic modeling. There have been several works that address the issue of objective function optimization, including ObjFind[130] and BOSS[131]. ObjFind creates a bilevel optimization problem that simultaneously attempts to maximize an objective function and minimize the flux distribution error between predicted and experimental data. This optimization approach identifies the optimal weights of the objective function that best fit the data (the *c* vector in Equation 1). BOSS is another bilevel optimization approach similar to ObjFind. Instead of determining the weights of available fluxes in the system, BOSS creates an additional flux that has the sole purpose of acting as the objective function. This extra flux term allows for more flexibility than ObjFind because it allows the stoichiometric amounts of individual metabolites to be more readily configured within the objective function. More recently, proteomics data have been used to determine the most essential fluxes in an objective function, arguing that the maximization of biomass is not suitable in certain conditions, such as when a cell is under stress[158]. Another approach was developed to complement DFBA by testing various objective functions simultaneously and pinpointing the best without the use of a bilevel optimization problem[159]. Applying these optimization methods to the objective function in LK-DFBA

could significantly improve modeling accuracy.

6.4.2 Optimizing the objective function in LK-DFBA

There are two obstacles that must be addressed in order to use ObjFind or BOSS with LK-DFBA. First, LK-DFBA contains novel linear kinetics constraints not found in the prototypical FBA framework, which can be challenging to implement within the bilevel optimization. Second, the solution vector in ObjFind and BOSS is a single set of flux values, while the solution vector in LK-DFBA is a time course of metabolite concentrations and flux values. If one were to use ObjFind or BOSS on LK-DFBA as is, the two optimization frameworks would identify an objective function that tries to simultaneously maximize metabolite concentrations and fluxes (whereas the current LK-DFBA only maximizes one or the other) and the objective function would simultaneously consider metabolites and fluxes across all timepoints as independent measures. As a result, different metabolites and fluxes could be maximized at different timepoints. While biologically it is possible and even likely that the cellular objective may not always be constant, it is unlikely to deviate in a short time span; this limitation would not be accounted for if one were to directly apply ObjFind and BOSS to LK-DFBA.

In an attempt to create a framework similar to ObjFind and BOSS that could work with LK-DFBA, I created two optimization approaches to search for the ideal objective function. To emulate ObjFind, the first approach attempted to optimize the weights of the *c* vector (Equation 1) for the available fluxes by minimizing the error between ODE data and the predictions generated by an LK-DFBA model fitted to the ODE data. In the second approach, instead of adjusting the weights for the *c* vector for the available fluxes,

a new flux was created and the stoichiometric contributions of each metabolite to the biomass was optimized, similar to BOSS. Like the first approach, the second approach minimized the prediction errors of the LK-DFBA model.

When I tested these two approaches on a small synthetic model as well as a model of *E. coli*, it quickly became apparent that the optimizer could easily become trapped at local minima. Unlike ObjFind and BOSS, which uses the duality principle in their bilevel optimizations to simultaneously identify the best objective function and the optimal solution to the FBA linear program, the approaches I have created are two nested optimization problems that are solved in serial. The objective functions determined by my approaches were not consistent across repetitions and did not seem to have any biological relevance for the *E. coli* model. For example, when using the first approach, there did not seem to be a preference for maximining any of the effluxes, which one might otherwise expect.

One important point to note is that when optimizing these objective functions, the LK-DFBA model used parameters estimated using the LR approach, which does not require an objective function to be known for parameter estimation. If using the LR+ approach, which has been shown to recapitulate training data better than the LR method, it is important to note that LR+ estimates parameters by minimizing the error between LK-DFBA and the training data, meaning LR+ uses an LK-DFBA objective function to predict metabolite and flux time course data during parameterization. If the LR+ approach is used during objective function optimization, whether the kinetics constraint parameters are identified before optimization of the objective function begins or at the same time will need to be considered.

While my initial attempts to optimize the objective function of LK-DFBA were unsuccessful, this undertaking should be reexamined in the future as it is a key area where LK-DFBA could be improved. With some modifications to ObjFind or BOSS, the kinetics constraints of LK-DFBA could be incorporated into the bilevel optimization problem, similar to the typical FBA flux constraints already included in both frameworks. Additionally, instead of creating a single linear program within LK-DFBA that includes all timepoints together, LK-DFBA could be reformulated so that each individual timepoint would be part of a separate linear program. This would possibly allow ObjFind or BOSS to more easily determine an optimal objective function at a single point in the time course that could then be applied to all timepoints. However, dividing LK-DFBA into separate timepoints would fundamentally change how the framework determines the optimal concentration and flux distributions and would need to be explored in more depth.

6.4.3 Improving predictions of cofactor metabolite concentrations

In Chapter 4, we determined that LK-DFBA could capture the metabolite dynamics of a few key metabolites in the biological systems examined. However, LK-DFBA struggled to accurately predict changes in cofactor concentrations. Cofactors, such as NADH and ATP, are involved in many different reactions, leading LK-DFBA to often predict that these metabolites are rapidly accumulating and depleting. Before it can become a widely used metabolic modeling framework, LK-DFBA will need to be able to model these cofactors and there are a few possible changes to LK-DFBA that could improve its accuracy when predicting the concentrations of these metabolites.

Because cofactors often participate in multiple reactions either as substrates or products, calculating the change in concentration of these metabolites involves many different kinetics constraints within LK-DFBA. On the other hand, other metabolites, such as lactate in the *L. lactis* model, only require a single constraint to calculate metabolite accumulation or depletion, which makes tracking these metabolites easier (Figure 37). A method for combining the constraints used to calculate the accumulation or depletion of cofactors into a single constraint could improve the predictions of these metabolite concentrations. Unlike the dimension reduction and hyperplane approaches in Chapter 4, which create a single constraint for reactions with multiple controller metabolites, this new approach would create a single constraint for multiple reactions. While this method could make it easier to track the dynamics of cofactors, it will be important to ensure that a single kinetics constraint does not oversimplify the framework.

Another possible approach for improving the prediction performance of cofactor concentrations is to add new constraints to the linear program that limit the change in concentration of these metabolites across each timepoint. These cofactors currently accumulate and deplete at rates that are not biologically likely. By restricting how quickly these concentrations can change based on the training data, LK-DFBA could more accurately model the behavior of these cofactors.

6.4.4 Using LK-DFBA in metabolic engineering

In this thesis, we have demonstrated that LK-DFBA can capture the general trends of different metabolic phenotypes in biological systems. As we continue to improve LK-DFBA, we will soon be able to use it as a tool for metabolic engineering.

We have already determined that LK-DFBA can predict some changes in metabolism when genetic perturbations are introduced to a system, but we have not tried to use LK-DFBA models to engineer an organism to efficiently produce a specific metabolite. Because LK-DFBA retains a linear programming structure, we envision that it could be integrated with current FBA strain design tools, including OptKnock[126], which identifies genetic knockouts for the overproduction of chemicals of interest. Like ObjFind for optimizing objective functions, OptKnock also uses a bilevel optimization structure and the same concerns discussed in the previous section will need to be investigated. Because OptKnock only focuses on gene knockouts and no other types of genetic perturbations that could add complexity, it is an ideal platform when first exploring how LK-DFBA can be used for strain design. Incorporating LK-DFBA in OptKnock is a critical first step toward demonstrating LK-DFBA can be a useful modeling framework for metabolic engineering and other applications in the future.

6.5 Closing remarks

The work presented in this thesis addresses three of the greatest challenges when attempting to model metabolic systems using metabolomics data. First, I introduced a stepwise machine learning platform for identifying the allosteric regulatory structure of metabolic systems, which is often unknown but critical to building accurate models. Second, I developed a novel method for inferring absolute concentrations from raw metabolomics data by leveraging the mass balances in a metabolic system. Finally, I presented improvements to the kinetics constraints in LK-DFBA and demonstrated for

the first time that the modeling framework can predict different phenotypes in two biological systems.

SCOUR, MetaboPAC, and LK-DFBA are each of significant importance in the process of developing metabolic models. The regulatory interactions predicted by SCOUR can lead to more accurate modeling of metabolic systems and the absolute concentrations inferred by MetaboPAC will improve the ability to integrate metabolomics data with computational tools. The new kinetics constraints implemented in LK-DFBA allow the framework to model a wider variety of metabolic systems. When combined together, these three platforms create a cohesive workflow that starts with the pre-processing of metabolomics data and ends with a fully constructed dynamic model with integrated allosteric regulatory information.

In this chapter, we have discussed several suggested areas to explore that could further improve SCOUR, MetaboPAC, and LK-DFBA. Even with each framework in its current state, we have already determined that they can be used together to model a simple, yet meaningful, metabolic system. As the amount of available metabolomics data continues to rapidly expand, this cohesive workflow will be ready to take advantage of these data toward building predictive dynamic metabolic models.

APPENDIX A

Supplementary Figures

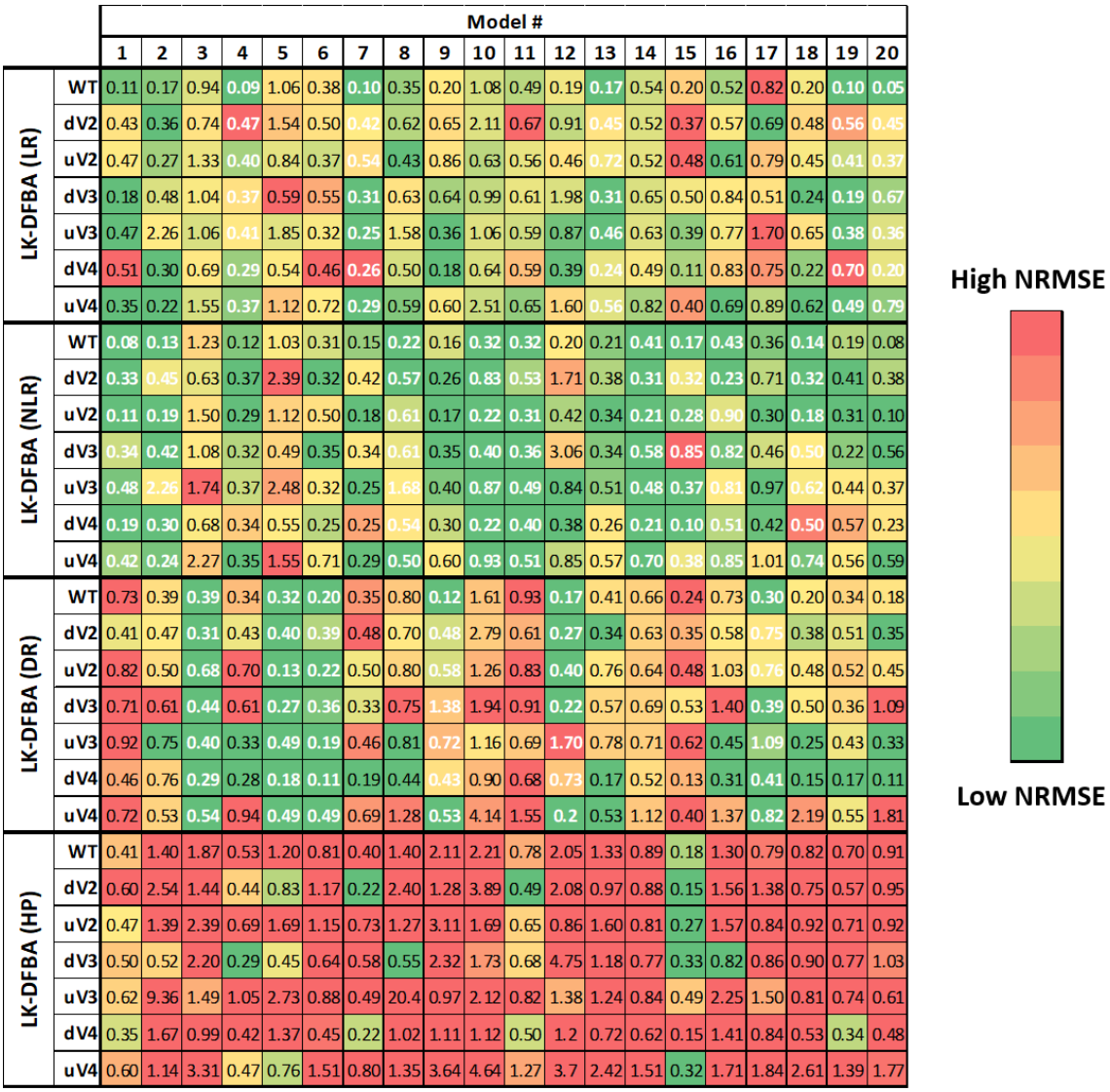

Figure 50: LK-DFBA performance on noisy synthetic model data, nT = 50, CoV = 0.05.

Each constraint approach was used to fit parameters to noisy (nT = 50, CoV = 0.05) wild-type (WT) data and then used to simulate the WT system and the system with *in silico* genetic perturbations with fluxes v_2, v_3, or v_4 down- or up-regulated. Dark green boxes represent the lowest average NRMSE (N = 10) within each phenotype for each synthetic model, while dark red boxes represent the highest average NRMSE. The cells with bolded white numbers indicate the LK-DFBA approach that best fits the WT data. Cells with white numbers are generally consistently green, indicating that fitting to WT data is a good indicator of which approach will be optimal across all perturbations.

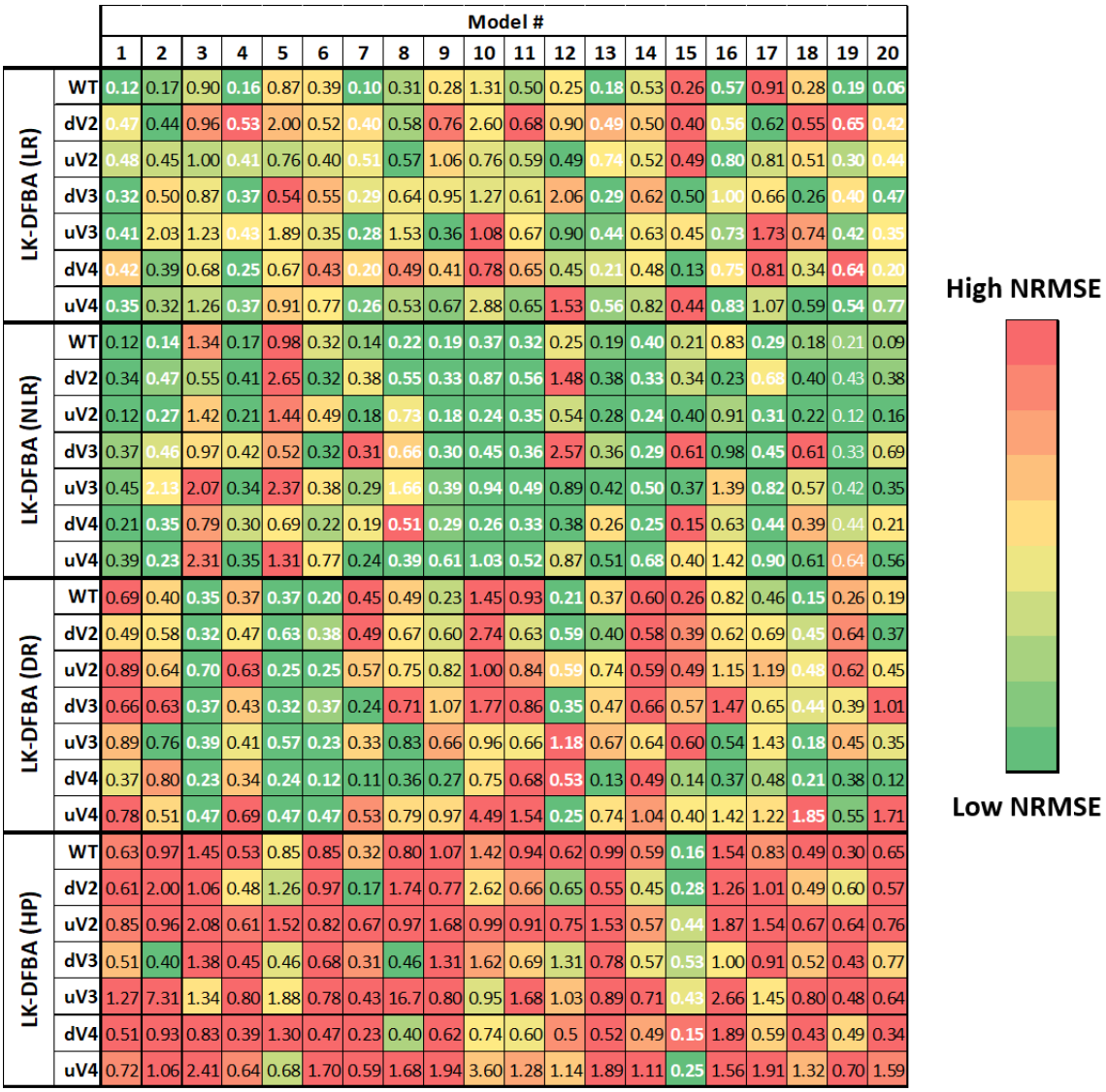

Figure 51: LK-DFBA performance on noisy synthetic model data, nT = 50, CoV = 0.15.

Each constraint approach was used to fit parameters to noisy (nT = 50, CoV = 0.15) wild-type (WT) data and then used to simulate the WT system and the system with *in silico* genetic perturbations with fluxes v_2, v_3, or v_4 down- or up-regulated. Dark green boxes represent the lowest average NRMSE (N = 10) within each phenotype for each synthetic model, while dark red boxes represent the highest average NRMSE. The cells with bolded white numbers indicate the LK-DFBA approach that best fits the WT data. Cells with white numbers are generally consistently green, indicating that fitting to WT data is a good indicator of which approach will be optimal across all perturbations.

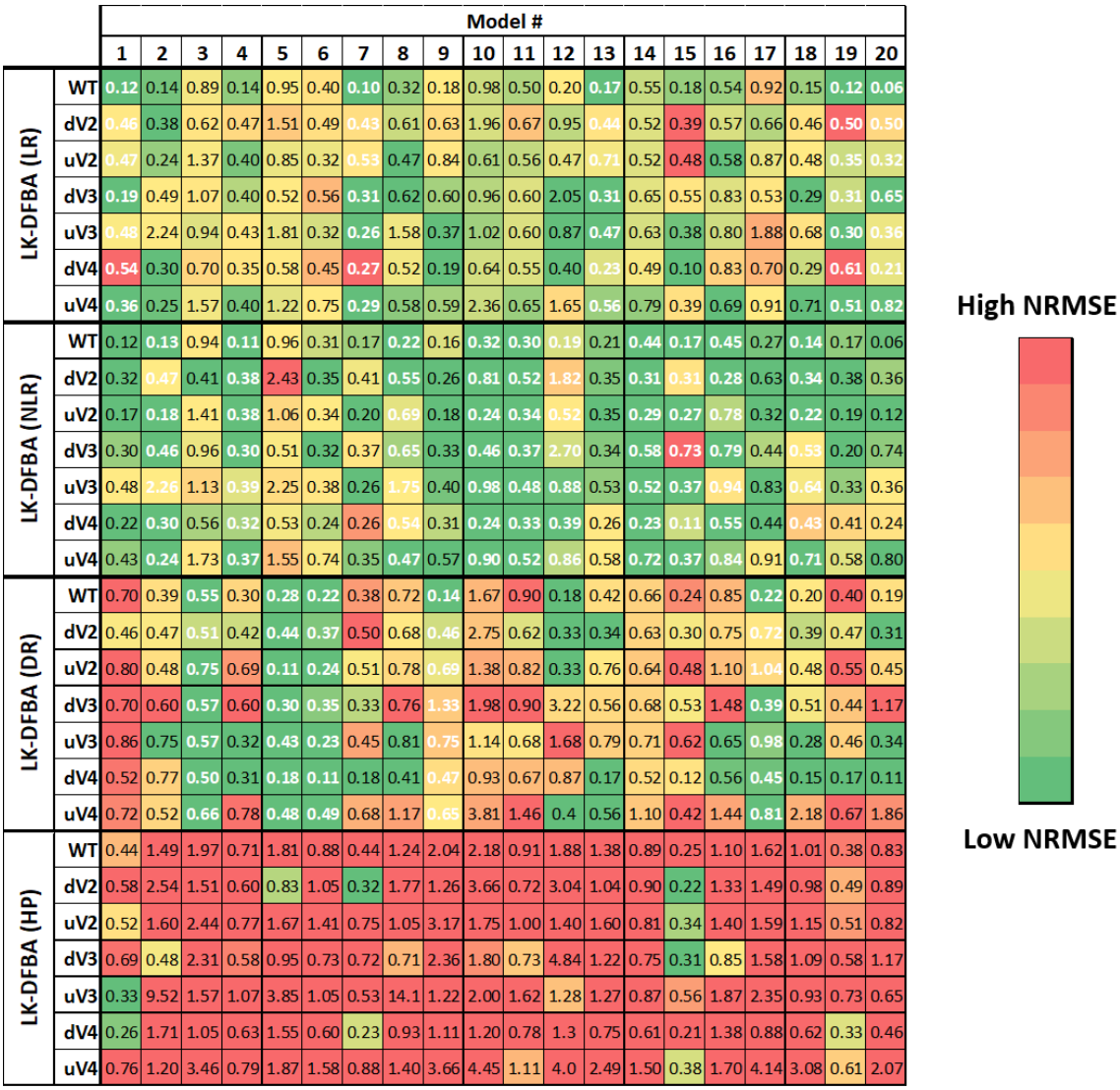

Figure 52: LK-DFBA performance on noisy synthetic model data, nT = 15, CoV = 0.05.

Each constraint approach was used to fit parameters to noisy (nT = 15, CoV = 0.05) wild-type (WT) data and then used to simulate the WT system and the system with *in silico* genetic perturbations with fluxes v_2, v_3, or v_4 down- or up-regulated. Dark green boxes represent the lowest average NRMSE (N = 10) within each phenotype for each synthetic model, while dark red boxes represent the highest average NRMSE. The cells with bolded white numbers indicate the LK-DFBA approach that best fits the WT data. Cells with white numbers are generally consistently green, indicating that fitting to WT data is a good indicator of which approach will be optimal across all perturbations.

REFERENCES

1. Mamas M, Dunn WB, Neyses L, Goodacre R. The role of metabolites and metabolomics in clinically applicable biomarkers of disease. Arch Toxicol. 2011;85(1):5-17.
2. Mastrangelo A, Armitage EG, Garcia A, Barbas C. Metabolomics as a tool for drug discovery and personalised medicine. A review. Curr Top Med Chem. 2014;14(23).
3. Wishart DS. Emerging applications of metabolomics in drug discovery and precision medicine. Nat Rev Drug Discov. 2016;15(7):473-84.
4. Noor E, Eden E, Milo R, Alon U. Central Carbon Metabolism as a Minimal Biochemical Walk between Precursors for Biomass and Energy. Molecular Cell. 2010;39(5):809-20.
5. Maicas S. The Role of Yeasts in Fermentation Processes. Microorganisms. 2020;8(8).
6. Keasling JD. Manufacturing Molecules Through Metabolic Engineering. Science. 2010;330.
7. Nielsen J. Metabolic engineering. Appl Microbiol Biotechnol. 2001;55(3):263-83.
8. Nielsen J, Fussenegger M, Keasling J, Lee SY, Liao JC, Prather K, et al. Engineering synergy in biotechnology. Nat Chem Biol. 2014;10(5):319-22.
9. Yadav VG, Stephanopoulos G. Metabolic engineering: the ultimate paradigm for continuous pharmaceutical manufacturing. ChemSusChem. 2014;7(7):1847-53.
10. Lau W, Fischbach MA, Osbourn A, Sattely ES. Key applications of plant

metabolic engineering. PLoS Biol. 2014;12(6):e1001879.

11. Aristidou A, Penttila M. Metabolic engineering applications to renewable resource utilization. Curr Opin Biotechnol. 2000;11(2):187-98.

12. Wishart DS. Metabolomics for Investigating Physiological and Pathophysiological Processes. Physiol Rev. 2019;99(4):1819-75.

13. Armitage EG, Barbas C. Metabolomics in cancer biomarker discovery: current trends and future perspectives. J Pharm Biomed Anal. 2014;87:1-11.

14. Gomez-Casati DF, Zanor MI, Busi MV. Metabolomics in plants and humans: applications in the prevention and diagnosis of diseases. Biomed Res Int. 2013;2013:792527.

15. Wishart DS. Applications of metabolomics in drug discovery and development. Drugs R D. 2008;9(5):307-22.

16. Beyoglu D, Idle JR. Metabolomics and its potential in drug development. Biochemical Pharmacology. 2013;85.

17. Puchades-Carrasco L, Pineda-Lucena A. Metabolomics in pharmaceutical research and development. Curr Opin Biotechnol. 2015;35:73-7.

18. Jorge TF, Mata AT, Antonio C. Mass spectrometry as a quantitative tool in plant metabolomics. Phil Trans R Soc A. 2016;374(2079).

19. do Prado RM, Porto C, Nunes E, de Aguiar CL, Pilau EJ. Metabolomics and Agriculture: What Can Be Done? mSystems. 2018;3(2).

20. Roessner U, Wagner C, Kopka J, Trethewey RN, Willmitzer L. Simultaneous analysis of metabolites in potato tuber by gas chromatography–mass spectrometry. The Plant Journal. 2001;23(1):131-42.

21. Schauer N, Fernie AR. Plant metabolomics: towards biological function and mechanism. Trends Plant Sci. 2006;11(10):508-16.

22. Diaz-Ruiz R, Rigoulet M, Devin A. The Warburg and Crabtree effects: On the origin of cancer cell energy metabolism and of yeast glucose repression. Biochim Biophys Acta. 2011;1807(6):568-76.

23. Mapelli V, Olsson L, Nielsen J. Metabolic footprinting in microbiology: methods and applications in functional genomics and biotechnology. Trends Biotechnol. 2008;26(9):490-7.

24. Tang J. Microbial metabolomics. Curr Genomics. 2011;12(6):391-403.

25. Volkova S, Matos MRA, Mattanovich M, de Mas IM. Metabolic Modelling as a Framework for Metabolomics Data Integration and Analysis. Metabolites. 2020;10(8).

26. Dromms RA, Styczynski MP. Systematic applications of metabolomics in metabolic engineering. Metabolites. 2012;2(4):1090-122.

27. Emwas AH. The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. Methods Mol Biol. 2015;1277:161-93.

28. Qiu Y, Reed D. Gas Chromatography in Metabolomics Study. In: Guo X, editor. Advances in Gas Chromatography: Books on Demand; 2014.

29. Emwas AH, Roy R, Mckay RT, Tenori L, Saccenti E, Gowda GA, et al. NMR Spectroscopy for Metabolomics Research. Metabolites. 2019;9(7).

30. Fernie AR, Aharoni A, Willmitzer L, Stitt M, Tohge T, Kopka J, et al. Recommendations for reporting metabolite data. Plant Cell. 2011;23(7):2477-82.

31. Khodadadi M, Pourfarzam M. A review of strategies for untargeted urinary metabolomic analysis using gas chromatography–mass spectrometry. Metabolomics. 2020;16.

32. Lu W, Su X, Klein MS, Lewis IA, Fiehn O, Rabinowitz JD. Metabolite Measurement: Pitfalls to Avoid and Practices to Follow. Annu Rev Biochem. 2017;86:277-304.

33. Machado D, Costa RS, Ferreira EC, Rocha I, Tidor B. Exploring the gap between dynamic and constraint-based models of metabolism. Metabolic Engineering. 2012;14.

34. Orth JD, Thiele I, Palsson BO. What is flux balance analysis? Nature Biotechnology. 2010;28:245-8.

35. Feist AM, Palsson BO. The biomass objective function. Curr Opin Microbiol. 2010;13(3):344-9.

36. Beard DA, Liang SD, Qian H. Energy balance for analysis of complex metabolic networks. Biophys J. 2002;83(1):79-86.

37. Henry CS, Broadbelt LJ, Hatzimanikatis V. Thermodynamics-based metabolic flux analysis. Biophys J. 2007;92(5):1792-805.

38. Kummel A, Panke S, Heinemann M. Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. Mol Syst Biol. 2006;2:2006 0034.

39. Ataman M, Hatzimanikatis V. Heading in the right direction: thermodynamics-based network analysis and pathway engineering. Curr Opin Biotechnol. 2015;36:176-82.

40. Segre D, Vitkup D, Church GM. Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci U S A. 2002;99(23):15112-7.

41. Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, et al. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Mol Syst Biol. 2010;6:390.

42. Bordbar A, Yurkovich JT, Paglia G, Rolfsson O, Sigurjonsson O, Palsson BO. Elucidating dynamic metabolic physiology through network integration of quantitative time-course metabolomics. Scientific Reports. 2017;7.

43. Mahadevan R, Edwards JS, Doyle FJ, 3rd. Dynamic flux balance analysis of diauxic growth in Escherichia coli. Biophys J. 2002;83(3):1331-40.

44. Chassagnole C, Noisommit-Rizzi N, Schmid JW, Mauch K, Reuss M. Dynamic modeling of the central carbon metabolism of Escherichia coli. Biotechnol Bioeng. 2002;79(1):53-73.

45. Kurata H, Sugimoto Y. Improved kinetic model of Escherichia coli central carbon metabolism in batch and continuous cultures. J Biosci Bioeng. 2018;125(2):251-7.

46. Millard P, Smallbone K, Mendes P. Metabolic regulation is sufficient for global and robust coordination of glucose uptake, catabolism, energy production and growth in Escherichia coli. PLoS Comput Biol. 2017;13(2).

47. Hynne F, Dano S, Sorensen PG. Full-scale model of glycolysis in Saccharomyces cerevisiae. Biophys Chem. 2001;94(1-2):121-63.

48. Rizzi M, Baltes M, Theobald U, Reuss M. In vivo analysis of metabolic dynamics in Saccharomyces cerevisiae: II. Mathematical model. Biotechnol Bioeng. 1997;55(4):592-608.

49. Sulieman AK, Putra MD, Abasaeed AE, Gaily MH, Al-Zahrani SM, Zeinelabdeen MA. Kinetic modeling of the simultaneous production of ethanol and fructose by Saccharomyces cerevisiae. Electronic Journal of Biotechnology. 2018;34:1-8.

50. Costa RS, Hartmann A, Gaspar P, Neves AR, Vinga S. An extended dynamic model of Lactococcus lactis metabolism for mannitol and 2,3-butanediol production. Mol Biosyst. 2014;10(3):628-39.

51. Tokic M, Hatzimanikatis V, Miskovic L. Large-scale kinetic metabolic models of Pseudomonas putida KT2440 for consistent design of metabolic engineering strategies. Biotechnol Biofuels. 2020;13:33.

52. Khodayari A, Maranas CD. A genome-scale Escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. Nat Commun. 2016;7:13806.

53. Smallbone K, Simeonidis E, Swainston N, Mendes P. Towards a genome-scale kinetic model of cellular metabolism. BMC Syst Biol. 2010;4:6.

54. Stanford NJ, Lubitz T, Smallbone K, Klipp E, Mendes P, Liebermeister W. Systematic construction of kinetic models from genome-scale metabolic networks. PLoS One. 2013;8(11):e79195.

55. Srinivasan S, Cluett WR, Mahadevan R. Constructing kinetic models of metabolism at genome-scales: A review. Biotechnol J. 2015;10(9):1345-59.

56. Dromms RA, Lee JY, Styczynski MP. LK-DFBA: a linear programming-based modeling strategy for capturing dynamics and metabolite-dependent regulation in metabolism. BMC Bioinformatics. 2020;21(1):93.

57. Covert MW, Palsson BO. Transcriptional regulation in constraints-based metabolic models of Escherichia coli. J Biol Chem. 2002;277(31):28058-64.

58. Link H, Kochanowski K, Sauer U. Systematic identification of allosteric protein-metabolite interactions that control enzyme activity in vivo. Nat Biotechnol. 2013;31(4).

59. Kern D, Zuiderweg ER. The role of dynamics in allosteric regulation. Curr Opin Struct Biol. 2003;13(6):748-57.

60. Covert MW, Schilling CH, Palsson B. Regulation of gene expression in flux balance models of metabolism. J Theor Biol. 2001;213(1):73-88.

61. Covert MW, Xiao N, Chen TJ, Karr JR. Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli. Bioinformatics. 2008;24(18):2044-50.

62. Shlomi T, Eisenberg Y, Sharan R, Ruppin E. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. Mol Syst Biol. 2007;3(101).

63. Machado D, Herrgard MJ, Rocha I. Modeling the Contribution of Allosteric Regulation for Flux Control in the Central Carbon Metabolism of E. coli. Front Bioeng Biotechnol. 2015;3:154.

64. Peregrin-Alvarez JM, Sanford C, Parkinson J. The conservation and evolutionary modularity of metabolism. Genome Biol. 2009;10(6):R63.

65. de Luis Balaguer MA, Fisher AP, Clark NM, Fernandez-Espinosa MG, Moller BK, Weijers D, et al. Predicting gene regulatory networks by combining spatial and temporal gene expression data in Arabidopsis root stem cells. Proc Natl Acad

Sci U S A. 2017;114(36):E7632-E40.

66. Hackett SR, Baltz EA, Coram M, Wranik BJ, Kim G, Baker A, et al. Learning causal networks using inducible transcription factors and transcriptome-wide time series. Mol Syst Biol. 2020;16.

67. Haque S, Ahmad JS, Clark NM, Williams CM, Sozzani R. Computational prediction of gene regulatory networks in plant growth and development. Curr Opin Plant Biol. 2019;47:96-105.

68. Lempp M, Farke N, Kuntz M, Freibert SA, Lill R, Link H. Systematic identification of metabolites controlling gene expression in E. coli. Nat Commun. 2019;10(1):4463.

69. Mochida K, Koda S, Inoue K, Nishii R. Statistical and Machine Learning Approaches to Predict Gene Regulatory Networks From Transcriptome Datasets. Front Plant Sci. 2018;9:1770.

70. Wang Y, Yang S, Zhao J, Du W, Liang Y, Wang C, et al. Using Machine Learning to Measure Relatedness Between Genes: A Multi-Features Model. Scientific Reports. 2019;9.

71. Yang Y, Fang Q, Shen HB. Predicting gene regulatory interactions based on spatial gene expression data and deep learning. PLoS Comput Biol. 2019;15(9):e1007324.

72. Oliveira AP, Dimopoulos S, Busetto AG, Christen S, Dechant R, Falter L, et al. Inferring causal metabolic signals that regulate the dynamic TORC1-dependent transcriptome. Mol Syst Biol. 2015;11(4).

73. Hackett SR, Zanotelli VR, Xu W, Goya J, Park JO, Perlman DH, et al. Systems-

level analysis of mechanisms regulating yeast metabolic flux. Science. 2016;354(6311).

74. Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, et al. Machine learning in bioinformatics. Briefings in Bioinformatics. 2006;7(1):86-112.

75. Gromski PS, Xu Y, Kotze HL, Correa E, Ellis DI, Armitage EG, et al. Influence of missing values substitutes on multivariate analysis of metabolomics data. Metabolites. 2014;4(2):433-52.

76. Dias-Audibert FL, Navarro LC, de Oliveira DN, Delafiori J, Melo C, Guerreiro TM, et al. Combining Machine Learning and Metabolomics to Identify Weight Gain Biomarkers. Front Bioeng Biotechnol. 2020;8:6.

77. Guarnera E, Berezovsky IN. Allosteric sites: remote control in regulation of protein activity. Curr Opin Struct Biol. 2016;37:1-8.

78. Rinschen M, Ivanisevic J, Giera M, Siuzdak G. Identification of bioactive metabolites using activity metabolomics. Nat Rev Mol Cell Biol. 2019;20(6):353-67.

79. Berggard T, Linse S, James P. Methods for the detection and analysis of protein-protein interactions. Proteomics. 2007;7(16):2833-42.

80. Macalino SJY, Basith S, Clavio NAB, Chang H, Kang S, Choi S. Evolution of In Silico Strategies for Protein-Protein Interaction Drug Discovery. Molecules. 2018;23(8).

81. Diether M, Sauer U. Towards detecting regulatory protein-metabolite interactions. Curr Opin Microbiol. 2017;39:16-23.

82. Huang M, Song K, Liu X, Lu S, Shen Q, Wang R, et al. AlloFinder: a strategy for

allosteric modulator discovery and allosterome analyses. Nucleic Acids Res. 2018;46(W1):W451-W8.

83. Savageau MA, Voit EO, Irvine DH. Biochemical systems theory and metabolic control theory: 1. fundamental similarities and differences. Mathematical Biosciences. 1987;86(2):127-45.

84. Ulusu NN. Evolution of Enzyme Kinetic Mechanisms. J Mol Evol. 2015;80(5-6):251-7.

85. Hoffmann J, Bar-Sinai Y, Lee LM, Andrejevic J, Mishra S, Rubinstein SM, et al. Machine learning in a data-limited regime: Augmenting experiments with synthetic data uncovers order in crumpled sheets. Sci Adv. 2019;5(4):eaau6792.

86. Le TA, Baydin AG, Zinkov R, Wood F. Using Synthetic Data to Train Neural Networks is Model-Based Reasoning. Ieee Ijcnn. 2017:3514-21.

87. Radivojevic T, Costello Z, Workman K, Garcia Martin H. A machine learning Automated Recommendation Tool for synthetic biology. Nat Commun. 2020;11(1):4879.

88. Schon M, Simeth J, Heinrich P, Gortler F, Solbrig S, Wettig T, et al. DTD: An R Package for Digital Tissue Deconvolution. J Comput Biol. 2020;27(3):386-9.

89. NS. Curve intersections MATLAB Central File Exchange2010 [Available from: https://www.mathworks.com/matlabcentral/fileexchange/22441-curve-intersections.

90. Galli S. Python Feature Engineering Cookbook: Over 70 Recipes for Creating, Engineering, and Transforming Features to Build Machine Learning Models: Packt Publishing Ltd; 2020.

91. Wolpert H. Stacked generalization. Neural Networks. 1992;5(2):241-59.

92. Breiman L. Random Forests. Machine Learning. 2001;45:5-32.

93. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 1967;13(1):21-7.

94. Fix E, Hodges JL. Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties. Randolph Field, Texas; 1951.

95. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. 1943. Bull Math Biol. 1990;52(1-2):99-115; discussion 73-97.

96. Fisher RA. The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics. 1936;7(2):179-88.

97. Voit EO. Biochemical Systems Theory: A Review. ISRN Biomathematics. 2013;2013:1-53.

98. Batushansky A, Toubiana D, Fait A. Using graph theory to analyze biological networks. BioMed Research International. 2016;2016:1-9.

99. Toubiana D, Puzis R, Wen L, Sikron N, Kurmanbayeva A, Soltabayeva A, et al. Combined network analysis and machine learning allows the prediction of metabolic pathways from tomato metabolomics data. Commun Biol. 2019;2:214.

100. Dromms RA, Styczynski MP. Improved metabolite profile smoothing for flux estimation. Mol Biosyst. 2015;11(9):2394-405.

101. Thonusin C, IglayReger HB, Soni T, Rothberg AE, Burant CF, Evans CR. Evaluation of intensity drift correction strategies using MetaboDrift, a normalization tool for multi-batch metabolomics data. J Chromatogr A. 2017;1523:265-74.

102. Wei X, Shi X, Kim S, Zhang L, Patrick JS, Binkley J, et al. Data preprocessing method for liquid chromatography-mass spectrometry based metabolomics. Anal Chem. 2012;84(18):7963-71.

103. Yang J, Zhao X, Lu X, Lin X, Xu G. A data preprocessing strategy for metabolomics to reduce the mask effect in data analysis. Front Mol Biosci. 2015;2:4.

104. Chen L, Zhong F, Zhu J. Bridging Targeted and Untargeted Mass Spectrometry-Based Metabolomics via Hybrid Approaches. Metabolites. 2020;10(9).

105. Reiekeberg E, Powers R. New frontiers in metabolomics: from measurement to insight. F1000Res. 2017;6(1148).

106. Fiehn O. Metabolomics by Gas Chromatography-Mass Spectrometry: the combination of targeted and untargeted profiling. Curr Protoc Mol Biol. 2016;114.

107. Veenstra TD. Metabolomics: the final frontier? Genome Med. 2012;4(4):40.

108. Worley B, Powers R. Multivariate Analysis in Metabolomics. Curr Metabolomics. 2013;1(1):92-107.

109. Kapoore RV, Vaidyanathan S. Towards quantitative mass spectrometry-based metabolomics in microbial and mammalian systems. Philos Trans A Math Phys Eng Sci. 2016;374(2079).

110. Moldoveanu SC, David V. Derivatization Methods in GC and GC/MS. Gas Chromatography - Derivatization, Sample Preparation, Application: Books on Demand; 2018.

111. Bennett BD, Kimball EH, Gao M, Osterhout R, Van Dien SJ, Rabinowitz JD.

Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. Nat Chem Biol. 2009;5(8):593-9.

112. Willemsen AM, Hendrickx DM, Hoefsloot HCJ, Hendriks MMWB, Wahl SA, Teusink B, et al. MetDFBA: incorporating time-resolved metabolomics measurements into dynamic flux balance analysis. Molecular BioSystems. 2015;11(137).

113. Schrimpe-Rutledge AC, Codreanu SG, Sherrod SD, McLean JA. Untargeted metabolomics strategies – Challenges and Emerging Directions. J Am Soc Mass Spectrom. 2016;27(12):1897-905.

114. Chalcraft KR, Lee R, Mills C, Britz-McKibbin P. Virtual quantification of metabolites by capillary electrophoresis-electrospray ionization-mass spectrometry: predicting ionization efficiency without chemical standards. Analytical Chemistry. 2009;81(7).

115. Wu L, Wu Y, Shen H, Gong P, Cao L, Wang G, et al. Quantitative structure–ion intensity relationship strategy to the prediction of absolute levels without authentic standards. Analytica Chimica Acta. 2013;794:67-75.

116. Liigand J, Wang T, Kellogg J, Smedsgaard J, Cech N, Kruve A. Quantification for non-targeted LC/MS screening without standard substances. Scientific Reports. 2020;10.

117. Tumanov S, Zubenko Y, Obolonkin V, Greenwood DR, Shmanai V, Villas-Boas SG. Calibration curve-free GC–MS method for quantitation of amino and non-amino organic acids in biological samples. Metabolomics. 2016;16(64).

118. Response Factors, Determination, Accuracy and Precision. In: Guiochon G,

Guillemin CL, editors. Quantitative Analysis By Gas Chromatography Journal of CHromatography Library. 421988. p. 587-627.

119. Koek MM, Jellema RH, van der Greef J, Tas AC, Hankemeier T. Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives. Metabolomics. 2011;7(3):307-28.

120. Barata JCA, Hussein MS. The Moore–Penrose Pseudoinverse: A Tutorial Review of the Theory. Braz J Phys. 2012;42:146-65.

121. Canelas AB, Ras C, ten Pierick A, van Dam JC, Heijnen JJ, van Gulik WM. Leakage-free rapid quenching technique for yeast metabolomics. Metabolomics. 2008;4.

122. Chou IC, Voit EO. Estimation of dynamic flux profiles from metabolic time series data. BMC Syst Biol. 2012;6:84.

123. Goel G, Chou IC, Voit EO. System estimation from metabolic time-series data. Bioinformatics. 2008;24(21):2505-11.

124. Stringer KA, Younger JG, McHugh C, Yeomans L, Finkel MA, Puskarich MA, et al. Whole Blood Reveals More Metabolic Detail of the Human Metabolome than Serum as Measured by 1H-NMR Spectroscopy: Implications for Sepsis Metabolomics. Shock. 2015;44(3).

125. Thonusin C, IglayReger HB, Soni T, Rothberg AE, Burant CF, Evans CR. Evaluation of intensity drift correction strategies using MetaboDrift, a normalization tool for multi-batch metabolomics data. J Chromatogr A. 2017;1523:265-74.

126. Burgard AP, Pharkya P, Maranas CD. Optknock: a bilevel programming

framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol Bioeng. 2003;84(6):647-57.

127. Chen WW, Niepel M, Sorger PK. Classic and contemporary approaches to modeling biochemical reactions. Genes Dev. 2010;24(17):1861-75.

128. Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometrics and Intelligent Laboratory Systems. 1987;2(1-3):37-52.

129. Costa RS, Verissimo A, Vinga S. KiMoSys: a web-based repository of experimental data for KInetic MOdels of biological SYStems. BMC Syst Biol. 2014;8:85.

130. Burgard AP, Maranas CD. Optimization-based framework for inferring and testing hypothesized metabolic objective functions. Biotechnol Bioeng. 2003;82(6):670-7.

131. Gianchandani EP, Oberhardt MA, Burgard AP, Maranas CD, Papin JA. Predicting biological system objectives de novo from internal state measurements. BMC Bioinformatics. 2008;9:43.

132. Gaspar P, Neves AR, Gasson MJ, Shearman CA, Santos H. High Yields of 2,3-Butanediol and Mannitol in Lactococcus lactis through Engineering of NAD(+) Cofactor Recycling. Appl Environ Microb. 2011;77(19):6826-35.

133. Neves AR, Ramos A, Shearman C, Gasson MJ, Santos H. Catabolism of mannitol in Lactococcus lactis MG1363 and a mutant defective in lactate dehydrogenase. Microbiology (Reading). 2002;148(Pt 11):3467-76.

134. Oh E, Lu M, Park C, Park C, Oh HB, Lee SY, et al. Dynamic Modeling of Lactic Acid Fermentation Metabolism with Lactococcus lactis. J Microbiol Biotechnol.

2011;21(2):162-9.

135. Wisselink HW, Moers AP, Mars AE, Hoefnagel MH, de Vos WM, Hugenholtz J. Overproduction of heterologous mannitol 1-phosphatase: a key factor for engineering mannitol production by Lactococcus lactis. Appl Environ Microbiol. 2005;71(3):1507-14.

136. Asmawaty T, Kadir A, Mannan AA, Kierzak AM, McFadden J, Shimizu K. Modeling and simulation of the main metabolism in Escherichia coli and its several single-gene knockout mutants with experimental verification. Microbiol Cell Factories. 2010;9(88).

137. Ishii N, Nakahigashi K, Baba T, Robert M, Soga T, Kanai A, et al. Multiple high-throughput analyses monitor the response of E. coli to perturbations. Science. 2007;316(5824):593-7.

138. Long CP, Gonzalez JE, Sandoval NR, Antoniewicz MR. Characterization of physiological responses to 22 gene knockouts in Escherichia coli central carbon metabolism. Metab Eng. 2016;37:102-13.

139. Usui Y, Hirasawa T, Furusawa C, Shirai T, Yamamoto N, Mori H, et al. Investigating the effects of perturbations to pgi and eno gene expression on central carbon metabolism in Escherichia coli using 13C metabolic flux analysis. Microbiol Cell Factories. 2012;11(87).

140. Lima AP, Baixinho V, Machado D, Rocha I. A Comparative Analysis of Dynamic Models of the Central Carbon Metabolism of Escherichia coli. Ifac Papersonline. 2016;49(26):270-6.

141. Kleerebezem M, Boels IC, Groot MN, Mierau I, Sybesma W, Hugenholtz J.

Metabolic engineering of Lactococcus lactis: the impact of genomics and metabolic modelling. J Biotechnol. 2002;98(2-3):199-213.

142. Papagianni M. Metabolic engineering of lactic acid bacteria for the production of industrially important compounds. Comput Struct Biotechnol J. 2012;3:e201210003.

143. Barnard J, Meng XL. Applications of multiple imputation in medical studies: from AIDS to NHANES. Stat Methods Med Res. 1999;8(1):17-36.

144. Lee JY, Styczynski MP. NS-kNN: a modified k-nearest neighbors approach for imputing metabolomics data. Metabolomics. 2018;14(12):153.

145. Bonaccorso G. Machine Learning Algorithms: Popular algorithms for data science and machine learning. Second ed: Packt Publishing; 2018. 522 p.

146. Cortes C, Vapnik V. Support-Vector Networks. Machine Learning. 1995;20(3):273-97.

147. Chauhan VK, Dahiya K, Sharma A. Problem formulations and solvers in linear SVM: a review. Artif Intell Rev. 2018;52:803-55.

148. Han H, Jiang X. Overcome support vector machine diagnosis overfitting. Cancer Inform. 2014;13(Suppl 1):145-58.

149. Lewis DD. Naive (Bayes) at forty: The independence assumption in information retrieval. European Conference on Machine Learning. 1998:4-15.

150. Giskeodegard GF, Grinde MT, Sitter B, Axelson DE, Lundgren S, Fjosne HE, et al. Multivariate Modeling and Prediction of Breast Cancer Prognostic Factors Using MR Metabolomics. J Proteome Res. 2010;9(2):972-9.

151. Trainor PJ, DeFilippis AP, Rai SN. Evaluation of Classifier Performance for

Multiclass Phenotype Discrimination in Untargeted Metabolomics. Metabolites. 2017;7(2).

152. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013;35(8):1798-828.

153. Morissette L, Chartier S. The k-means clustering technique: General considerations and implementation in Mathematica. Tutorials in Quantitative Methods for Psychology. 2013;9(1):15-24.

154. Chandrashekar G, Sahin F. A survey on feature selection methods. Computers and Electrical Engineering. 2014;40:16-28.

155. Hill CM, Waight RD, Bardsley WG. Does any enzyme follow the Michaelis-Menten equation? . Molecular and Cellular Biochemistry. 1977;15(3):173-8.

156. Toubiana D, Fernie AR, Nikoloski Z, Fait A. Network analysis: tackling complex data to study plant metabolism. Trends Biotechnol. 2013;31(1):29-36.

157. Faraji M, Voit EO. Stepwise inference of likely dynamic flux distributions from metabolic time series data. Bioinformatics. 2017;33(14):2165-72.

158. Montezano D, Meek L, Gupta R, Bermudez LE, Bermudez JCM. Flux Balance Analysis with Objective Function Defined by Proteomics Data—Metabolism of Mycobacterium tuberculosis Exposed to Mefloquine. PLoS ONE. 2015;10(7).

159. Nikdel A, Braatz RD, Budman HM. A systematic approach for finding the objective function and active constraints for dynamic flux balance analysis. Bioprocess Biosyst Eng. 2018;41(5):641-55.