UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Using metabolic network reconstructions to analyze complex data sets

Permalink

https://escholarship.org/uc/item/8sm6x1kp

Author

Zielinski, Daniel Craig

Publication Date

2015

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Using metabolic network reconstructions to analyze complex data sets

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioengineering

by

Daniel Craig Zielinski

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Thomas R. Bewley
Professor Terence T. Hwa
Professor Andrew D. McCulloch
Professor Kun Zhang

2015

Copyright
Daniel Craig Zielinski, 2015
All rights reserved.

The dissertation of Daniel Craig Zielinski is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Chair

University of California, San Diego

2015

DEDICATION

To my family.

EPIGRAPH

If people do not believe that mathematics is simple,
it is only because they do not realize how complicated life is.
—John von Neumann

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I have become indebted to a number of individuals throughout my time in graduate school and the Systems Biology Research Group (SBRG). First and foremost, I thank my advisor, Bernhard Palsson. There is so much to be said about Dr. Palsson, and unfortunately I cannot fit it all into this space. He is brilliant, intimidating, and kind. Inspirational in his success, infective with his passion for your work, he has fought for and supported so many, including myself, as they start their careers. His research in systems biology is why I joined UCSD, and through my time in his lab I have found my lifes work. I could not be more grateful. I would also like to thank those who have mentored me over the years, most notably Adam Feist, who brought me into the lab on my first project in metabolic engineering, and Neema Jamshidi, who introduced me to kinetics has since offered guidance on my cancer metabolism project. A large part of my practical understanding of navigating a graduate career comes from their guidance.

A number of my colleagues in the Palsson lab have influenced me through the years. Monica Mo introduced me to the drug response project and through our various collaborations I learned about academic collaborations, the biotech start-up industry, and some of the toughest issues related to choosing a career path. Aarash Bordbar, a classmate and co-author on several papers, has continually inspired me through the success he has brought about through work ethic and pursuit of excellence. Nikolaus Sonnenschein has been both a friend and my most important collaborator working on metabolic kinetics. Joshua Lerman and Miguel Campodonico have been both friends and constant sources of intellectual energy through our discussions of various ideas, both realistic and of the more ambitious variety. Among students I have mentored, I would like to especially mention James de Bree, Bin Du, and James Yurkovich. These students have helped me immeasurably, both through their efforts in our collaborations, and also their support, simply for believing that the science we do is important. In addition to these, others in the Palsson lab have also been friends and collaborators through the years, including: Nathan Lewis, Jan Schellenberger, Steve Federowicz, Haythem Latif, Mallory Embree, Jonathan Monk, Ali Ebrahim, Andreas Drager, Zachary

King, Teddy OBrien, Alex Thomas, Marta Matos, Austin Corbett, and Douglas McCloskey.

Finally, I thank my family, my parents Peter and Melissa, and my sister and brother-in-law Megan and Tavis. They have believed in me from day one, quite literally. Throughout my time in graduate school, they have patiently supported me, putting up with what must be some of the most boring conversations known to man as I obsess over my work, and provided a home I always look forward to returning to every chance I get. They've never asked a thing from me, but I dedicate this thesis to them.

Chapter 1 in part is a reprint of the material *Zielinski DC, Filipp F, Bordbar A, Jensen K, Smith J, Herrgard M, Mo ML, Palsson BO. Pharmacogenomic and clinical data link non-pharmacokinetic metabolic dysregulation to drug side effect pathogenesis. Nat Commun. (2015) 6:7101.* The dissertation author was the primary author.

Chapter 2 in part is a reprint of the material *Zielinski DC, Jamshidi N, Corbett AJ, Bordbar A, Thomas A, Palsson BO. Stoichiometric biomass synthetic requirements and metabolic stress resistance underlie hallmarks of cancer cell metabolism. In preparation.* The dissertation author was the primary author.

Chapter 3 in part is a reprint of the material *Zielinski DC, Matos M, De Bree J, Sonnenschein N, Palsson BO A data driven workflow for the construction of bottom-up kinetic models of metabolism. In preparation.* The dissertation author was the primary author.

VITA

2008	B.S. in Biomedical Engineering, University of Virginia
2015	Ph.D. in Bioengineering, University of California, San Diego

PUBLICATIONS

Du B, **Zielinski DC**, Drager A, Tan J, Zhang Z, Ruggiero K, Arzumanyan GA, Palsson BO. Constructing hybrid kinetic models of metabolism: Choice of approximate rate laws and effects of iterative refinement. *In Preparation*

Zielinski DC, De Bree J, Matos M, Sonnenschein N, Palsson BO. Bottom up reconstruction of enzyme kinetics: applications in E. coli core metabolism. *In Preparation*

Sonnenschein N, **Zielinski DC**, De Bree J, Palsson S, Thomas A, Bordbar A, Jamshidi N, Palsson BO. The MASS Toolbox: Accessible dynamic modeling. *In Preparation*

Zielinski DC, Jamshidi N, Corbett AJ, Bordbar A, Thomas A, Palsson BO. Stoichiometric biomass synthetic requirements and metabolic stress resistance underlie hallmarks of cancer cell metabolism. *In preparation.*

Dräger A, **Zielinski DC**, Keller R, Rall M, Eichner J, Palsson BO, Zell A. Context-sensitive creation of kinetic equations in biochemical networks. *Under Review*

Bordbar A, McCloskey D, **Zielinski DC**, Sonnenschein N, Jamshidi N, Palsson BO. Multi-omic data-driven construction of personalized kinetic models of metabolism. *Under Review*

Zielinski DC, Filipp FV, Bordbar A, Jensen K, Smith JW, Herrgard MJ, Mo ML, Palsson BO. Pharmacogenomic and clinical data link non-pharmacokinetic metabolic dysregulation to drug side effect pathogenesis. Nat Commun. (2015) 6:7101

Nam H, Campodonico M, Bordbar A, Hyduke DR, Kim S, **Zielinski DC**, Palsson BO. A systems approach to predict oncometabolites via context-specific genome-scale metabolic networks. PLoS Comput. Biol. (2014) 10(9) e 1003837

Olavarria K, De Ingeniis J, **Zielinski DC**, Fuentealba M, Muoz R, McCloskey D, Feist AM, Cabrera R. Metabolic impact of an NADH-producing glucose-6-phosphate dehydrogenase in Escherichia coli. Microbiology. (2014) 160:2780-93.

Schellenberger J, **Zielinski DC**, Choi W, Madireddi S, Portnoy V, Scott DA, Reed JL, Osterman AL, Palsson BO. Predicting outcomes of steady-state 13C isotope tracing experiments using Monte Carlo sampling. BMC Systems Biology (2012) 6:9

Zielinski DC, Palsson BO. Kinetic Modeling of Metabolic Networks. Systems Metabolic Engineering (2012) 25-55

Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, **Zielinski DC**, Bordbar A, Lewis NE, Rahmanian S, Kang J, Hyduke DR, Palsson BO. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v 2.0. Nat Protocols. (2011) 6:1290-1307.

Feist AM, **Zielinski DC**, Orth JD, Schellenberger J, Herrgard MJ, and Palsson BO. Model-driven evaluation of the production potential for growth-coupled products of Escherichia coli. Metabolic Engineering (2010) 12(3), 173-186

ABSTRACT OF THE DISSERTATION

Using metabolic network reconstructions to analyze complex data sets

by

Daniel Craig Zielinski

Doctor of Philosophy in Bioengineering

University of California, San Diego, 2015

Professor Bernhard Ø. Palsson, Chair

Understanding the behavior of complex biochemical networks is the primary goal of systems biology. This task is often addressed through the generation of large data sets consisting of measurements of biological components like mRNA transcripts, proteins, and metabolites. Although these methods have become increasingly accurate and comprehensive at measuring the state of the system, uncovering the function of the system then becomes a problem of analysis to extract understanding from the data. A key challenge in analyzing biological data sets is that determining the function of the system depends on a knowledge of the relationship between the components of the system. These relationships can be captured by grouping variables by known associations, such as pathways, or by explicitly modeling their relationships mathematically. Metabolic networks are particularly

primed for both of these approaches, because metabolic pathways are well-defined by network topology and the equation governing their function, the mass balance equation, is well understood. In this thesis, the capabilities of metabolic networks to interpret biological data are advanced through the development and application of models of increasing levels of detail. First, pathways systematically derived from a global human metabolic network reconstruction are used to identify metabolic perturbations tied to drug side effects from *in vitro* drug-treated gene expression data. Second, steady-state flux modeling of a core human metabolic network is used to identify factors underlying two hallmarks of cancer metabolism: the Warburg effect and glutamine addiction. Finally, the concept of a metabolic network reconstruction is extended by the definition of detailed enzyme kinetic mechanisms within E. coli central metabolism, integrating multiple data sets mechanistically to calculate dynamic functional states of enzymes. This work furthers the use of metabolic networks in analyzing complex biological data sets, showcasing the utility of these networks in addressing practical questions in systems biology using methods of increasing mechanistic resolution.

Chapter 1

Pharmacogenomic and clinical data link non-pharmacokinetic metabolic dysregulation to drug side effect pathogenesis

1.1 Abstract

Drug side effects cause a significant clinical and economic burden. However, mechanisms of drug action underlying side effect pathogenesis remain largely unknown. Here, we integrate pharmacogenomic and clinical data with a human metabolic network and find that non-pharmacokinetic metabolic pathways dysregulated by drugs are linked to the development of side effects. We show such dysregulated metabolic pathways contain genes with sequence variants affecting side effect incidence, play established roles in pathophysiology, have significantly altered activity in corresponding diseases, are susceptible to metabolic inhibitors, and are effective targets for therapeutic nutrient supplementation. Our results indicate that metabolic dysregulation represents a common mechanism underlying side effect pathogenesis that is distinct from the role of metabolism in drug clearance. We suggest that elucidating the relationships between the cellular response

to drugs, genetic variation of patients, and cell metabolism may help managing side effects by personalizing drug prescriptions and nutritional intervention strategies.

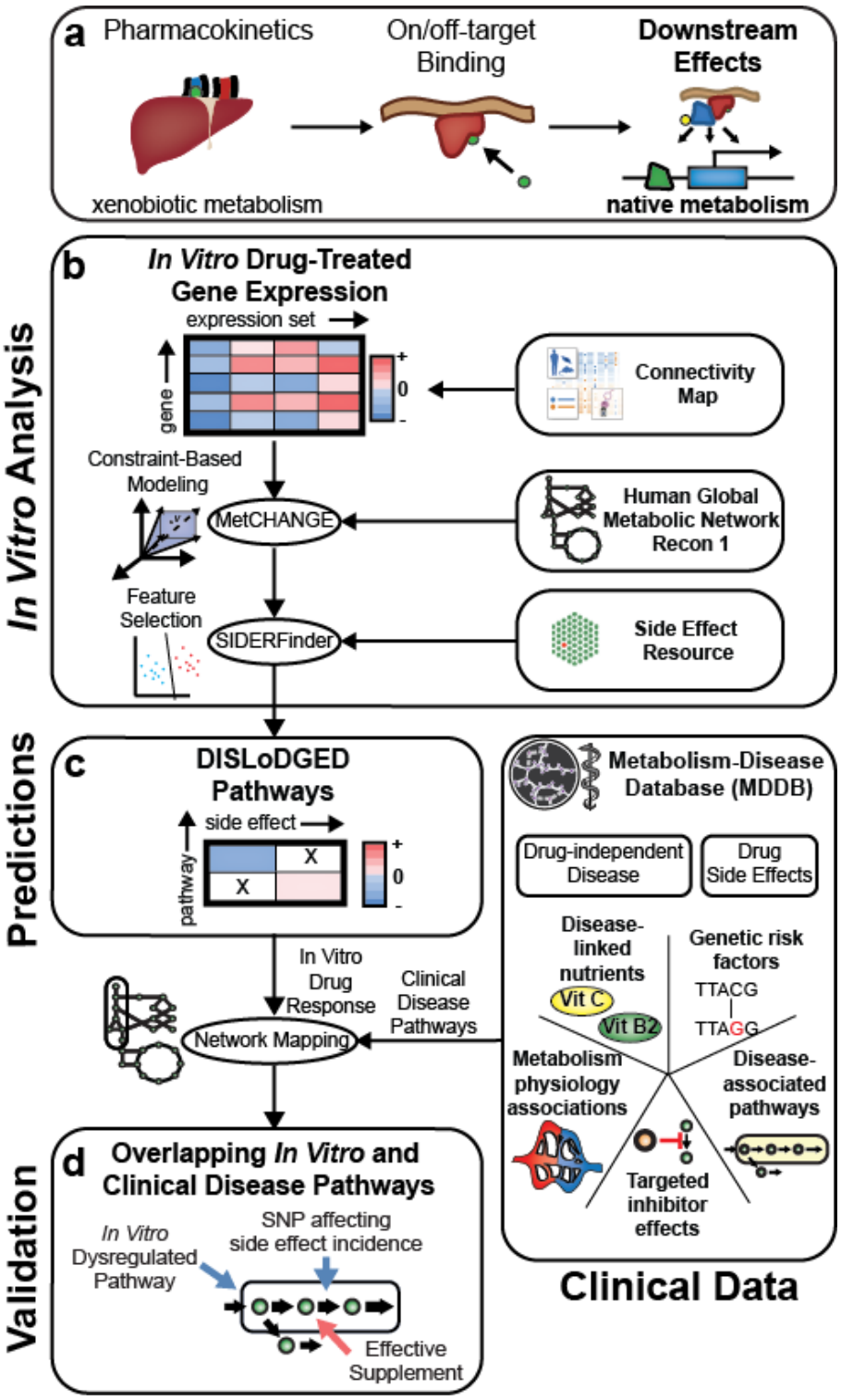
1.2 Introduction

Adverse drug reactions, commonly known as side effects, are thought to be responsible for as much as 11% of hospital admissions[**?**, 1], a fifth of both Phase II[2] and III[3] clinical trial failures, high-profile market withdrawals (e.g. Vioxx, Lipobay), and a large fraction of patient therapeutic non-compliance incidentsLoffler2003P. Risk factors associated with side effects have been identified, including number of drugs prescribed[4], patient age[5], and genetic variants[6]. Side effect-linked genetic variants identified so far are predominantly associated with drug pharmacokinetics, thereby affecting exposure of the body to a particular drug, but these variants do not give any indication of the mechanism by which pathogenesis is initiated. A recent study suggests that as many as half of drug side effects are related to known drug-protein binding events[7], and progress has been made toward systematically identifying drug binding events. However, only modest progress has been made toward elucidating specific drug-induced changes downstream of binding events for the majority of drugs (Figure 1.1a)[8]. These downstream effects in many cases may be most directly tied to side effect pathogenesis as well as patient genetic and environmental background.

Recent literature suggests that altered gene expression induced by drug treatment may be one mechanism by which drugs induce systemic off-target effects[9]. Unfortunately, the lack of clinical data has impeded the determination of causality of particular gene expression changes in side effect pathogenesis[10]. Recent studies have successfully utilized in vitro drug-treated gene expression profiles to predict clinical drug effectiveness[11], suggesting that *in vitro* data may contain features that are clinically conserved. However, demonstrating the relevance of *in vitro* drug response features to clinical side effect pathogenesis presents a significant challenge, due largely to the lack of suitable validating data sets and difficulty of clinical experimentation.

To address this challenge, we develop a network-based data analysis workflow built upon the use of *in vitro* drug treatment data to identify candidate side effect-linked features and a large collection of historical clinical and disease model data as a source of validation (Figure 1.1). First, we identify *in vitro* gene expression changes preferentially induced by drugs with clinically-defined side effects to identify candidate side effect-linked expression features. Then, we cross-reference these side effect-linked features with independent legacy clinical data found in the literature to corroborate their relevance in terms of five causal relationships. We implement this strategy within the context of the reconstructed global human metabolic network[12], which provides a biologically coherent structure for data integration due to the high-degree of network annotation and clear functional connectivity between genes via metabolic pathways[13].

Figure 1.1: **Overview and workflow used in this study.** a) Studies examining side effect pathogenesis focus primarily on drug pharmacokinetics, involving drug transport and clearance, and drug binding in terms of on and off target binding events. This study examines potential pathogenic mechanisms related to transcriptional changes downstream of clearance and binding events. b) Drug-treated gene expression profiles from the Connectivity Map database are analyzed in the context of the metabolic network reconstruction Recon 1 using constraint-based modeling to identify drug-induced pathway expression changes. Drug-induced metabolic pathway expression changes are analyzed in terms of drug side effects from the Side Effect Resource (SIDER) using a feature selection genetic algorithm to determine metabolic pathway perturbations conserved in particular side effects, termed DISLoDGED pathways. c) A new database, the Metabolism-Disease Database (MDDB), was generated by manual curation of literature to establish links between altered metabolic pathway function and pathologies, and this database was used to analyze DISLoDGED metabolic pathways. d) Five candidate causal mechanisms for metabolic changes in side effect pathogenesis (listed in the MDDB panel) are assessed in a large-scale fashion by comparing these *in vitro* perturbations to clinical data linking particular metabolic pathways to disease.


a
Pharmacokinetics
On/off-target Binding
Downstream Effects
xenobiotic metabolism
native metabolism
In Vitro Analysis
b
In Vitro Drug-Treated Gene Expression
expression set
gene
Connectivity Map
Constraint-Based Modeling
MetCHANGE
Human Global Metabolic Network Recon 1
Feature Selection
SIDERFinder
Side Effect Resource
Predictions
c
DISLoDGED Pathways
side effect
pathway
In Vitro Drug Response
Clinical Disease Pathways
Network Mapping
Metabolism-Disease Database (MDDB)
Drug-independent Disease
Drug Side Effects
Disease-linked nutrients
Vit C
Vit B2
Genetic risk factors
TTACG
TTAGG
Metabolism physiology associations
Disease-associated pathways
Targeted inhibitor effects
Clinical Data
Validation
d
Overlapping In Vitro and Clinical Disease Pathways
In Vitro Dysregulated Pathway
SNP affecting side effect incidence
Effective Supplement

1.3 Results

1.3.1 Calculation of drug-induced metabolic perturbations from gene expression data

We first identified drug-induced metabolic gene expression changes within 6,040 gene expression profiles in the Connectivity Map (CMap) dataset, representing three human cell lines exposed to 1,221 drug compounds[14] (Figure 1.1a). We analyzed the expression profiles using the reconstructed global human metabolic network Recon 1[12] with a novel metabolic pathway analysis algorithm, termed MetChange. MetChange is a constraint-based modeling[15] algorithm that computes a score for each metabolite summarizing the drug-induced gene expression changes along calculated production pathways for the metabolite (Figure 1.2). A MetChange score for a metabolite defines how expression has changed in a pathway containing these metabolite production reactions. Production in this case does not necessarily indicate secretion, as the majority of metabolites produced by one metabolic pathway are consumed in other metabolic pathways. We also note that gene expression is not the sole determinant of pathway activity, as gene and protein expression are imperfectly correlated and enzyme functional state may change due to perturbation as well. However, change in metabolic gene expression may still indicate a pathogenic metabolic functional change.

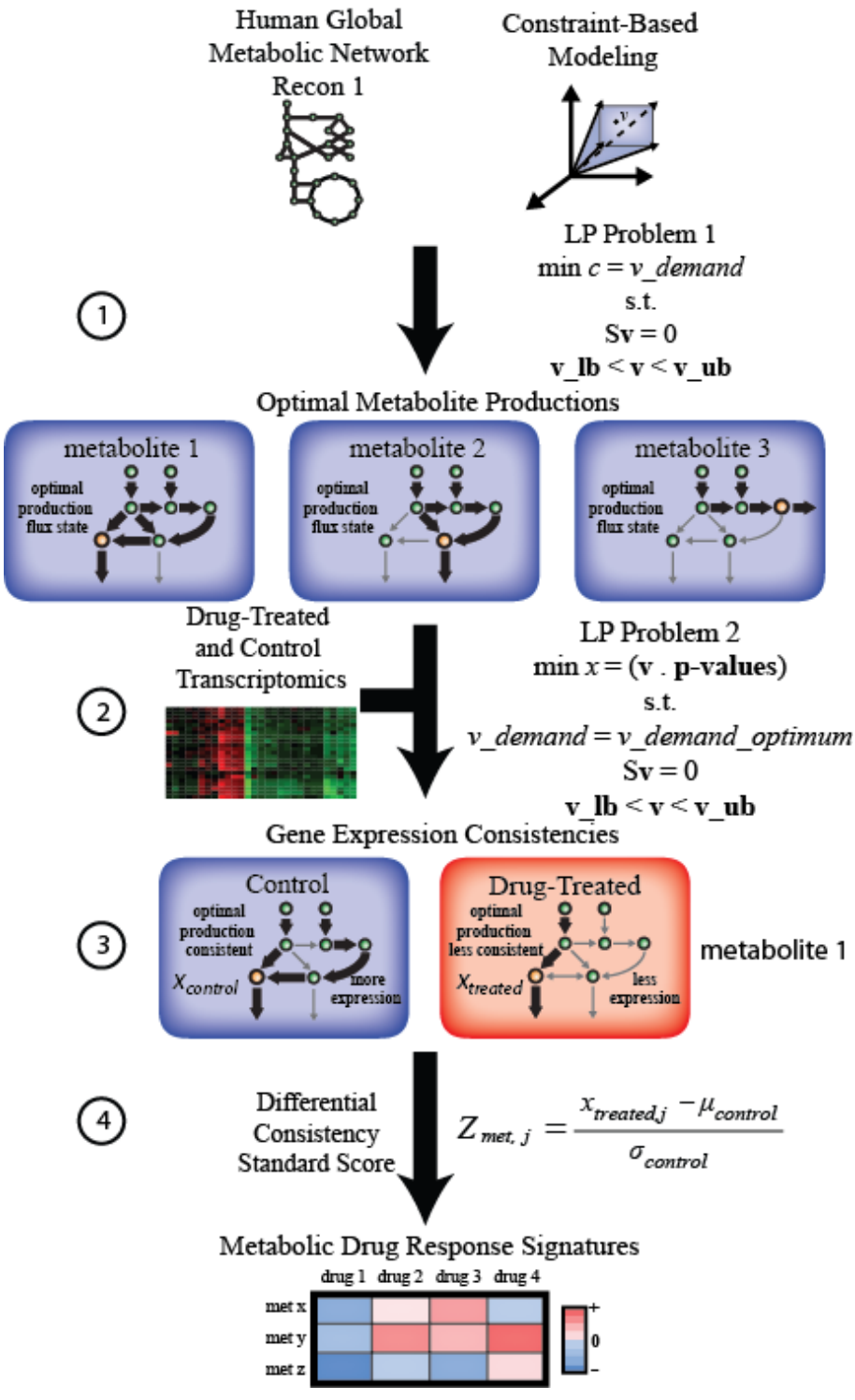
1.3.2 Validation of metabolic perturbations calculated by the MetChange algorithm

To compare the MetChange method against existing approaches that predict a metabolic outcome based on gene expression data[16], a published gene expression data from carbon and nitrogen starvation in S. cerevisiae was analyzed[17, 18]. A previously generated metabolic reconstruction of S. cerevisiae, iMM904[19], was used to compute MetChange scores for each condition. Scores were compared to metabolomics data generated for the same conditions using the relative changes from the initial time point. A total of 60 metabolites across 5 time

points for each carbon and nitrogen starvation were compared for both absolute (i.e. magnitude) correlation and directional (i.e. increased or decreased) correlation with MetChange scores. The MetChange algorithm compared favorably both with other network-based expression analysis methods and with the use of gene expression alone in predicting metabolic perturbations (Figure 1.3). Reassuringly, k-means clustering and principle component analysis (PCA) of MetChange scores, mapped metabolomics data, and mapped expression data suggest that MetChange scores maintain functional relationships between time points that are present in expression and metabolomics data (Figure 1.3).

To further validate the metabolic perturbations predicted by the MetChange method using the CMap data set, we performed a number of high throughput computational analyses comparing MetChange perturbations with drug response properties. First, metabolite scores were compared to co-occurrences of drug-metabolite text terms in the Pubmed database. A bootstrap analysis of Z-score permutations showed that PubMed drug-metabolite associations are recalled in a statistically significant (non-parametric perturbation p-value $p<10^{-3}$ for 1000 perturbations of MetChange scores). Second, it was found that known metabolic drug targets as found in the DrugBank database[20] are significantly closer in reaction proximity to known highly perturbed metabolites than less perturbed metabolites for these drugs (median Wilcoxon rank sum $p < 1.65 \times 10^{-10}$ for the highest scoring bin). Third, MetChange scores were found to be able to predict drug-protein literature co-associations within the Pubmed database in a statistically significant manner (non-parametric perturbation $p < 0.01$).

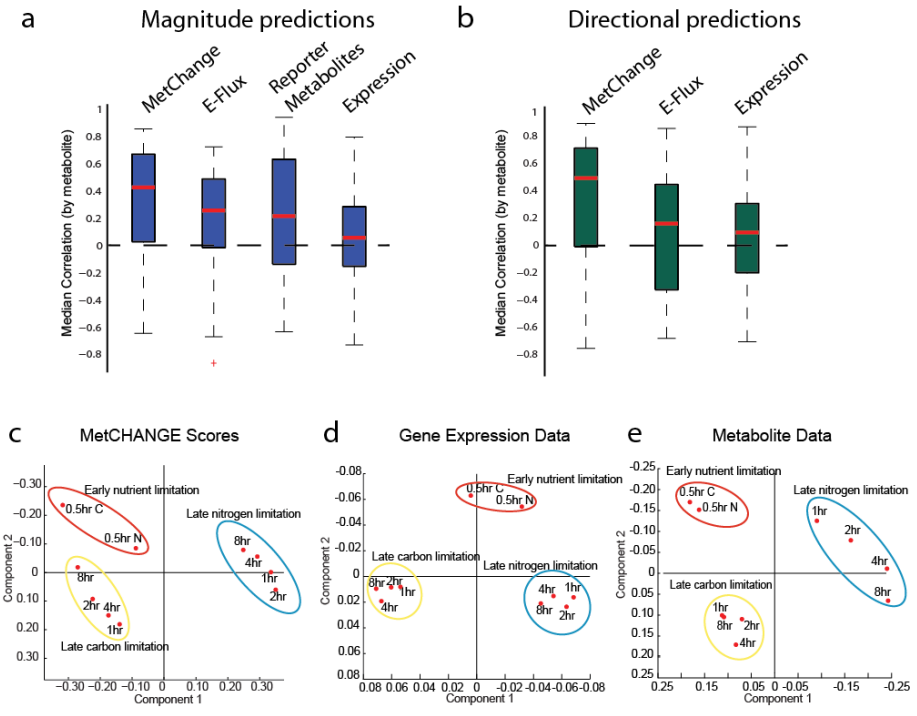
Figure 1.2: **Description of the MetChange algorithm.** 1) Using a metabolic network reconstruction, sink (demand) reactions are added for each metabolite. Demand reactions are irreversible with the stoichiometry. Each demand reaction is maximized in turn to obtain maximal production values for each metabolite using a linear programming problem (LP Problem 1). 2) Reaction presence/absence p-values are generated from gene expression data and mapped onto the metabolic network. A second linear programming problem is then solved (LP Problem 2) for each metabolite. LP Problem 2 identifies the flux solution that minimizes the inconsistency of the gene expression data with the optimal production of a metabolite by restricting the demand reaction for the metabolite to be at maximal flux, and subsequently minimizing an inconsistency score of (v x p-values). 3) An example case for metabolite 1. It is observed that the control data has greater expression (lower presence/absence p-value) for certain production reactions. Greater expression of production reactions results in a lower production inconsistency score for the control gene expression sample, compared with the drug-treated case, in which certain production reactions are less expressed (higher presence/absence p-value). 4) As different metabolites have different combinations of production reactions, they cannot be compared directly within samples. Instead, scores are compared for the same metabolite between control and treated samples to generate differential consistency scores using a simple standard score. Once standardized, metabolites can be compared within drugs to identify regions were perturbation in production potential has occurred due to gene expression changes.


Human Global Metabolic Network Recon 1
Constraint-Based Modeling
LP Problem 1
min c = v_demand
s.t.
Sv = 0
v_lb < v < v_ub
1
Optimal Metabolite Productions
metabolite 1
optimal production flux state
metabolite 2
optimal production flux state
metabolite 3
optimal production flux state
Drug-Treated and Control Transcriptomics
2
LP Problem 2
min x = (v . p-values)
s.t.
v_demand = v_demand_optimum
Sv = 0
v_lb < v < v_ub
Gene Expression Consistencies
3
Control
optimal production consistent
X control
more expression
Drug-Treated
optimal production less consistent
X treated
less expression
metabolite 1
4
Differential Consistency Standard Score
Z met, j = (x treated,j − μ control) / σ control
Metabolic Drug Response Signatures
drug 1 drug 2 drug 3 drug 4
met x
met y
met z
+
0
-

Finally, we validated perturbation predictions in targeted experimental measurements in MCF-7 cell culture under treatment by the drugs metformin, haloperidol, and genistein. These drugs were chosen due to previous evidence suggesting significant metabolic perturbation by these drugs. Measured metabolites were chosen to have broad coverage of pathways and target highly perturbed pathways as predicted by MetChange. Drug concentrations were chosen based on previous *in vitro* studies using these drugs.

First, treatment with the antipsychotic haloperidol revealed decreased uptake of Vitamin B6, consistent with the calculated decrease in the B6 processing pathway (Figure 1.4a). Second, treatment with metformin, an AMPK activator, showed significant perturbation of tricarboxylic acid cycle and fatty acid oxidation metabolites. The observed change was consistent with the large calculated perturbation but was in the opposite direction of the transcriptional change, indicating substantial non-transcriptional control of the metabolite levels (Figure 1.4b). Additional calculated metformin-induced changes supported by previous results include: 1) a down-regulation of folate metabolism consistent with reported folate deficiency in metformin-treated patients[21], 2) up-regulated oxidative stress response consistent with reported lower reactive oxygen species levels[22], and 3) increased polyamine synthesis and recycling pathways that may result from shared use of OCT transporters between metformin and polyamines. Third, treatment with genistein, an isoflavone with hypolipidemic effects[23], experimentally showed a preferential reduction of the fraction of mono-unsaturated fatty acids synthesized from glucose that was consistent with predictions (Figure 1.4c).

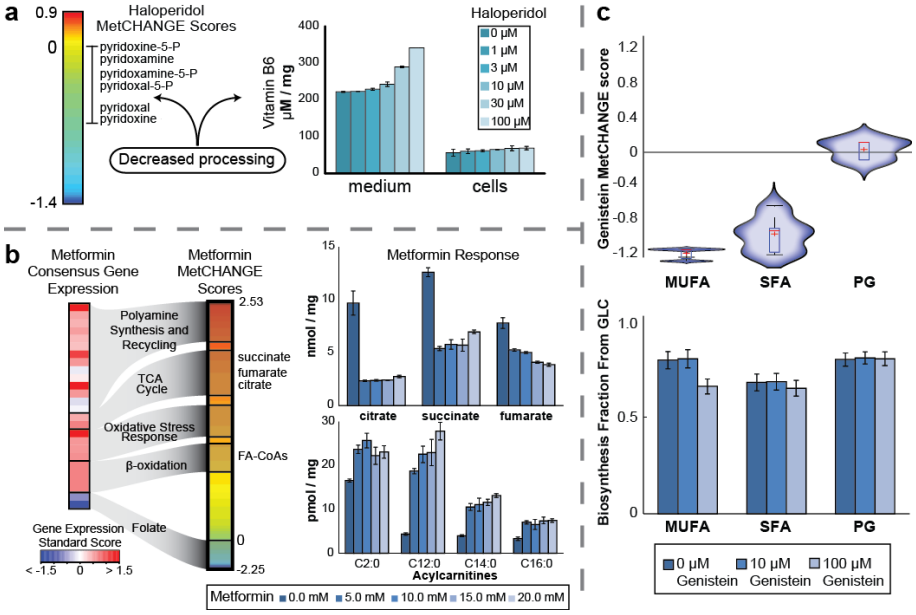
Figure 1.3: **Analysis of MetChange analysis of matched transcriptomic/metabolomics data in S. cerevisiae under carbon and nitrogen starvation conditions using the S. cerevisiae metabolic network reconstruction iMM904.** a) Comparison of the median correlations of computational metabolite absolute magnitude perturbation predictions with experimental data for several existing methods of integrating gene expression data with a metabolic reconstruction. b) The same comparison as in part a), but taking into account the direction of perturbation (the reporter metabolites method is not directional in its predictions, so both comparisons were made). Error bars are standard deviations. The MetChange algorithm performs favorably on this dataset in both absolute magnitude and directional predictions. c)-e) Principle Component Analysis (PCA) of the MetChange scores, gene expression data, and metabolite data for the 60 metabolites that mapped to iMM904. It is seen that the functional association of data is conserved after transformation to MetChange scores, and the MetChange principle component clustering has topological similarity to both gene expression and metabolite data clustering.


a
Magnitude predictions
MetChange
E-Flux
Reporter Metabolites
Expression
Median Correlation (by metabolite)
b
Directional predictions
MetChange
E-Flux
Expression
Median Correlation (by metabolite)
c
MetCHANGE Scores
Early nutrient limitation
0.5hr C
0.5hr N
Late nitrogen limitation
Late carbon limitation
Component 1
Component 2
d
Gene Expression Data
Early nutrient limitation
Late carbon limitation
Late nitrogen limitation
Component 1
Component 2
e
Metabolite Data
Early nutrient limitation
Late nitrogen limitation
Late carbon limitation
Component 1
Component 2

1.3.3 Identification of *in vitro* disease-linked metabolic pathways

Using MetChange scores generated from the Connectivity Map data set, we then identified the drug-induced metabolic gene expression changes that are most conserved among drugs with particular clinically-described side effects. We used a machine learning approach to select discriminating metabolite production pathway perturbations from the set of MetChange-calculated pathway changes based upon side effect frequency data reported in the Side Effect Resource (SIDER) database[24] (Figure 1.1b, Figure 1.5). A total of 357 side effects across 1417 gene expression profiles from CMap were analyzed, based on the criterion that at least 30 expression sets per side effect were available. The cutoff in number of expressions sets was chosen rationally as a balance between the scale of the study and the robustness of the side effect signature obtained. Overall, 2422 disease-linked drug-changed (termed DISLoDGED) metabolic pathways were identified with this analysis. These pathways are MetChange calculated metabolite production pathways that are over-represented in perturbations by drugs with particular side effects. DISLoDGED pathways are potential side effect to pathway relationships, hypothesizing that drug-induced perturbations away from metabolic homeostasis are involved in side effect pathogenesis. The calculated DISLoDGED pathways thus comprise a large set of omics-driven hypotheses of side effect pathogenic mechanisms (Supplementary Data 2). These associations were initially supported by automated co-association searches of PubMed abstracts to examine disease-linked nutrient deficiencies across all 357 side effects. Down-regulated DISLoDGED pathways were found to be marginally significantly predictive of deficiencies associated with corresponding specific diseases (permutation p= 0.055), while up-regulated pathways were not predictive (permutation p = 0.39). We then sought to determine whether the metabolic expression changes induced by drug treatment were involved in the pathogenesis underlying drug side effects.

Figure 1.4: **Validation in cell culture of drug-specific metabolic perturbations calculated from drug-induced transcription changes.** a) In the left panel, gene expression related to vitamin B6 metabolism is down-regulated by haloperidol, suggesting decreased utilization of vitamin B6. In the right panel, vitamin B6 levels were measured by an enzymatic assay in media supernatant and lysates of MCF-7 cells, showing decreased utilization of the essential nutrient. b) In the left panel, metformin gene expression perturbations are shown on the left, and resulting MetChange scores for related metabolites are shown on the right. The right panel shows response of TCA metabolites and acylcarnitines to metformin treatment. Metabolites show a large perturbation in the opposite direction of the observed transcriptional changes, validating the presence of a perturbation but indicating non-transcriptional control of metabolite levels in central metabolism as well. c) In the upper panel, genistein is predicted to preferentially decrease production of mono-unsaturated fatty acids. The lower panel shows measurement of biosynthesis fraction of different fatty acid groups and precursors from glucose following genistein treatment measured by NMR spectroscopy, which validate modeling predictions. Error bars represent standard deviations in all cases. Each experiment was performed in biological triplicate.


a
0.9
0
-1.4
Haloperidol
MetCHANGE Scores
pyridoxine-5-P
pyridoxamine
pyridoxamine-5-P
pyridoxal-5-P
pyridoxal
pyridoxine
Decreased processing
Vitamin B6
µM / mg
400
300
200
100
0
medium
cells
Haloperidol
0 µM
1 µM
3 µM
10 µM
30 µM
100 µM
b
Metformin
Consensus Gene
Expression
Metformin
MetCHANGE
Scores
2.53
0
-2.25
Polyamine
Synthesis and
Recycling
TCA
Cycle
Oxidative Stress
Response
β-oxidation
Folate
succinate
fumarate
citrate
FA-CoAs
Gene Expression
Standard Score
< -1.5
0
> 1.5
Metformin Response
nmol / mg
15
10
5
0
citrate
succinate
fumarate
pmol / mg
30
20
10
0
C2:0
C12:0
C14:0
C16:0
Acylcarnitines
Metformin
0.0 mM
5.0 mM
10.0 mM
15.0 mM
20.0 mM
c
Genistein MetCHANGE score
1.2
0.8
0.4
0
-0.4
-0.8
-1.2
MUFA
SFA
PG
Biosynthesis Fraction From GLC
1.0
0.5
0
MUFA
SFA
PG
0 µM
Genistein
10 µM
Genistein
100 µM
Genistein

1.3.4 Construction of a database of *in vivo* links between metabolism and disease

We assessed the relevance of DISLoDGED metabolic pathways to side effect pathogenesis through the use of a large body of clinical, biochemical and genetic literature on metabolism-disease relationships. We constructed a database of *in vivo* links found within the literature between metabolic function and disease, called the Metabolism Disease Database (MDDB), that consists of curated primary literature and existing databases (hosted at sbrg.ucsd.edu/mddb). Data collected through manual curation of the literature included disease-linked: 1) metabolic gene variants, 2) physiological system-specific metabolism, 3) metabolic pathways, 4) chemical inhibitors of metabolism, and 5) nutrient deficiencies and supplements. Data aggregated from existing databases included metabolic gene variants affecting disease incidences from a large GWAS database[25] and drug-metabolic enzyme target pairs from DrugBank. Studies on disease models were treated as acceptable sources where clinical data was not available. The resulting database encompasses 357 side effects, over 280 non-drug inhibitors, 600 drug molecules, 37 nutrients, and over 5000 investigated metabolic pathway-disease-link associations. The database includes information related to both drug side effects and non-drug-induced pathologies, and we examined predictions in terms of each of these separately.

1.3.5 Consistency of metabolism-disease relationships in the context of drug side effect pathogenesis

We first examined whether DISLoDGED metabolic pathways contain genes with variants that alter clinical side effect susceptibility, according to data in MDDB. Causal gene variants are typically considered to affect either drug pharmacokinetics, which involves drug exposure, or drug pharmacodynamics, which involves the interaction of the drug with the body. The majority of identified gene variants affecting drug side effect incidence are involved in drug pharmacokinetics, because these genes historically have been simpler to identify as drug

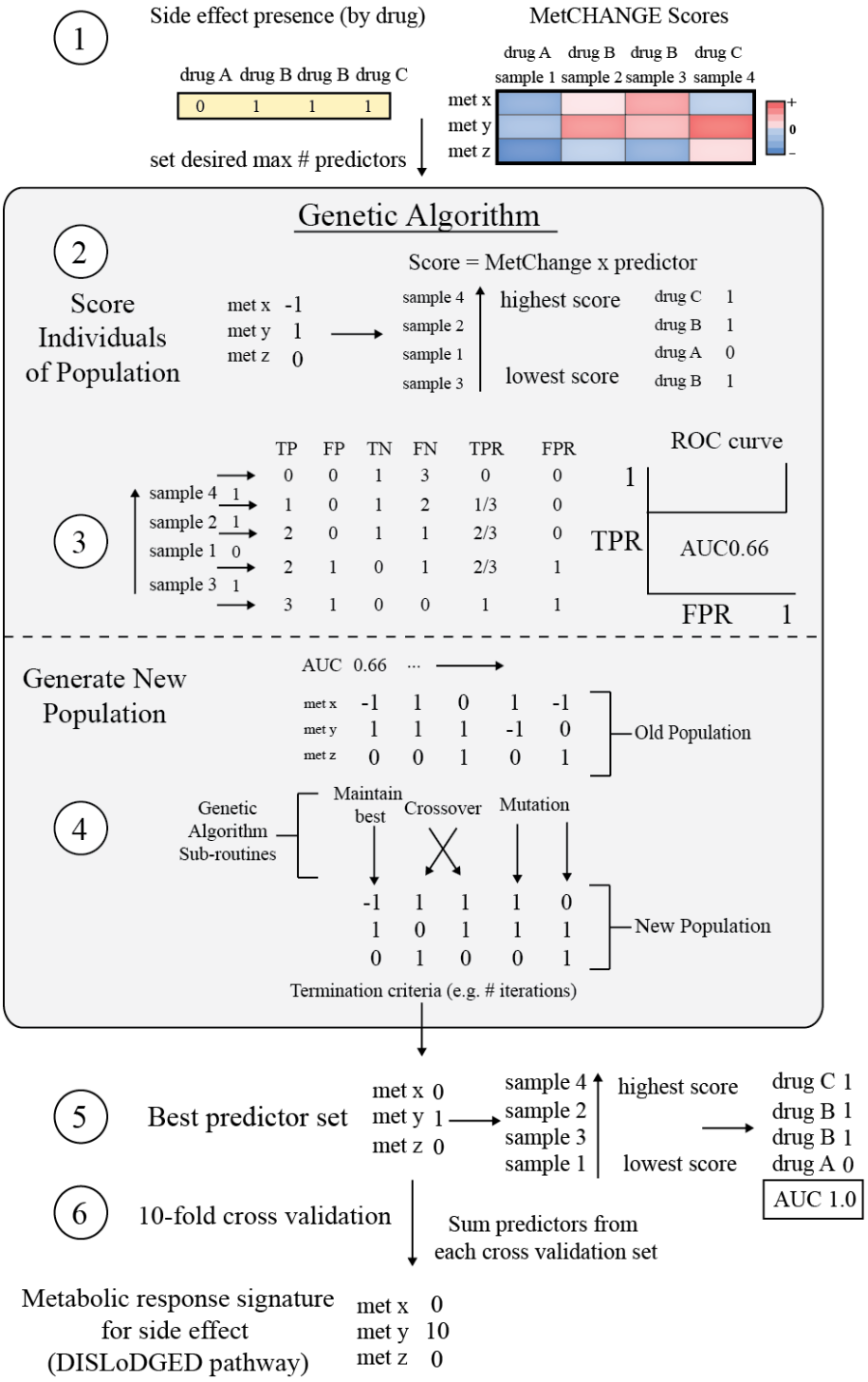
metabolism genes are largely known. However, we focused upon genes affecting pharmacodynamics, as these genes are more directly indicative of the mechanisms underlying pathogenesis. In MDDB, we identified nine side effect susceptibility gene variants that overlap directly with metabolism but are not involved in drug pharmacokinetics (Supplementary Note 1). We found that for each of the nine side effect susceptibility genes, at least one overlapping DISLoDGED pathway for this side effect had been identified through our analysis. This overlap between DISLoDGED pathways and non-pharmacokinetic side effect susceptibility altering genes was highly significant (joint hypergeometric tests, p-value 2.2 x 10^{-11}, see Supplementary Note 2 for calculation).

To support the relevance of this overlap, we assessed the nine overlapping DISLoDGED pathways in the context of additional known factors related to side effect pathogenesis. We found that, in each case, the overlapping DISLoDGED pathway had established ties to the clinical pathogenesis of the side effect. In the seven cases where the side effect pathology had been reported independently of the drug, alterations in the DISLoDGED pathway have been associated with the disease. Critically, in each of the nine cases, we found that drugs causing the side effect had been reported to induce a perturbation in the DISLoDGED pathway *in vivo*, demonstrating that the *in vitro*-derived DISLoDGED pathways are similarly perturbed *in vivo*. Furthermore, in seven of the nine cases, nutrient supplementation targeted at the DISLoDGED pathways have been shown to be effective in treating the drug side effect, while the remaining two cases are inconclusive due to reports of both positive and negative results.

For example, among the 38 drugs within the CMap database reported in SIDER as causing increased risk of cardiac arrhythmias, we identified five DISLoDGED pathways (Supplementary Note 1). Three of these DISLoDGED pathways, which were a down-regulation of oxidative pentose phosphate pathway and two related up-regulations in nitrogen metabolism, overlap with genetic polymorphisms known to cause increased susceptibility to arrhythmias[26]. These pathways are known to have physiological ties to the pathogenesis of arrhythmias[27] and have been shown to be perturbed *in vivo* by drugs causing arrhythmias[28]. Fur-

thermore, nutrient supplementation targeted at these pathways has been found therapeutic in drug-induced arrhythmias[29]. The remaining examples are presented in Supplementary Note 1.

Figure 1.5: **Description of the genetic algorithm used to identify metabolic signatures of side effects (DISLoDGED pathways).** 1) Inputs to the algorithm are a set of response variables for each gene expression set (either MetChange scores or gene expression changes), a binary presence/absence vector for whether each sample was treated with a drug that has the side effect, and the desired maximum number of predictor variables desired. 2) At initiation, the genetic algorithm generates a population of random guesses at the predictor variables, termed individuals, and assigns them either a value of -1, 0, or 1. For each individual, all gene expression samples are scored as the response variables (MetChange or gene expression changes) multiplied by the candidate signature. 3) Each gene expression sample is then ranked and a receiver operator characteristic (ROC) curve is generated and area under the curve (AUC) is calculated using the input presence/absence vector for the side effect or indication. The sample AUCs are the maximization objective of the genetic algorithm. 4) The genetic algorithm sub-routines are then used to generate a new population, biasing towards higher AUCs. Best solutions are maintained without modification, and lower scoring individuals are combined (crossed over) and modified (mutated) to search the solution space in a heuristic fashion. The termination criteria is typically a number of generations without improvement; however, we applied a simple maximum time termination criteria, as obtaining a global optimum was not deemed essential to gain biological insight. 5) The signature yielding the highest prediction AUC is considered the best predictor set. In the example case, the resultant AUC is 1.0, a perfect predictor for the sample set. 6) To assess over-fitting and hence the predictive potential of the metabolic signature, 10-fold cross validation is performed by generating 10 partitions of 90% of the data to train signatures and predict the remaining 10 partitions of 10% of the data. To find signatures that have constant predictive power, the cross validation signatures were summed, and high scoring metabolites were considered the conserved metabolic response signature (DISLoDGED pathway) for the side effect or indication.


1
Side effect presence (by drug)
drug A drug B drug B drug C
0 1 1 1
set desired max # predictors
MetCHANGE Scores
drug A drug B drug B drug C
sample 1 sample 2 sample 3 sample 4
met x
met y
met z
+
0
–
Genetic Algorithm
2
Score Individuals of Population
Score = MetChange x predictor
met x -1
met y 1
met z 0
sample 4
sample 2
sample 1
sample 3
highest score
lowest score
drug C 1
drug B 1
drug A 0
drug B 1
3
TP FP TN FN TPR FPR
sample 4 1 0 0 1 3 0 0
sample 2 1 1 0 1 2 1/3 0
sample 1 0 2 0 1 1 2/3 0
sample 3 1 2 1 0 1 2/3 1
3 1 0 0 1 1
ROC curve
1
TPR
AUC0.66
FPR 1
Generate New Population
AUC 0.66 ···
met x -1 1 0 1 -1
met y 1 1 1 -1 0
met z 0 0 1 0 1
Old Population
4
Genetic Algorithm Sub-routines
Maintain best
Crossover
Mutation
-1 1 1 1 0
1 0 1 1 1
0 1 0 0 1
New Population
Termination criteria (e.g. # iterations)
5
Best predictor set
met x 0
met y 1
met z 0
sample 4
sample 2
sample 3
sample 1
highest score
lowest score
drug C 1
drug B 1
drug B 1
drug A 0
AUC 1.0
6
10-fold cross validation
Sum predictors from each cross validation set
Metabolic response signature for side effect (DISLoDGED pathway)
met x 0
met y 10
met z 0

1.3.6 Consistency of DISLoDGED metabolic pathways with non-drug-induced pathogenesis

We further expanded the scope of validation beyond the nine available cases of non-pharmacokinetic genetic variants directly affecting side effect incidence. The MDDB database was used to determine whether the identified DISLoDGED pathways are conserved within *in vivo* data related to non-drug-induced pathologies as well, where a significantly larger body of literature exists than for side effect pathogenesis. This analysis hypothesizes that non-drug-induced pathogenesis and corresponding side effect pathogenesis share a common basis. Using the data collected, we assessed the calculated DISLoDGED pathways in terms of the five causal relationships in MDDB. Down-regulated and up-regulated DISLoDGED pathways were assessed independently to examine directional causal relationships. Approximately one sixth of calculated DISLoDGED pathways were investigated for validation, and compared to an equal number of randomized predictions as a control. The next sections describe each causal link examined.

1.3.7 Gene variants affecting disease incidence

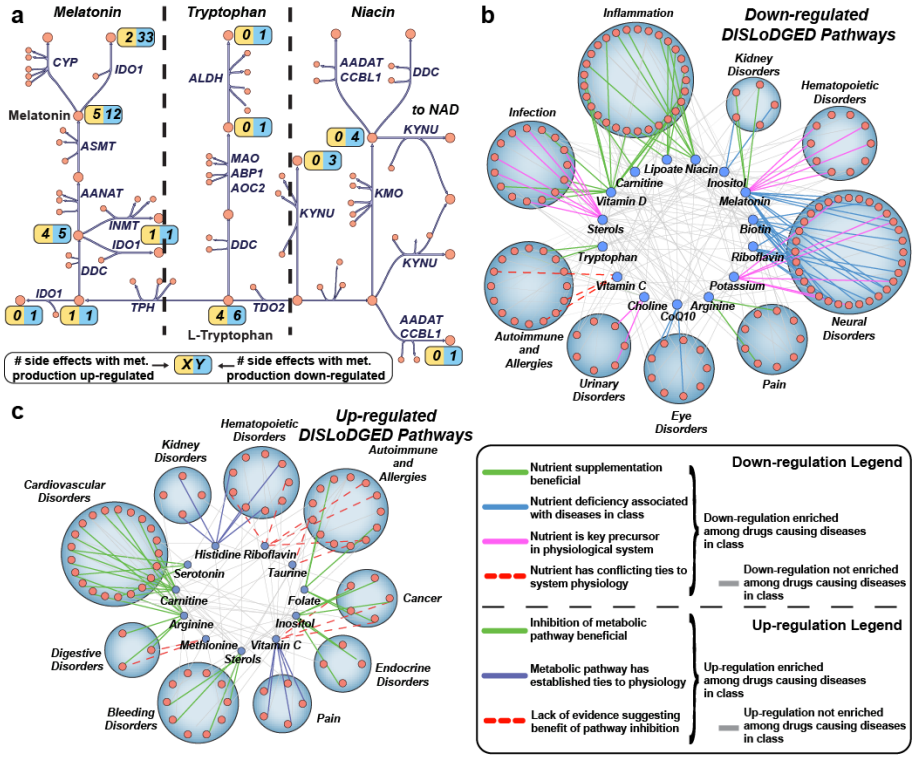
We first sought to determine whether DISLoDGED pathways contain known disease-linked gene variants. To do this, we compared the metabolic subsystems into which disease-linked gene variants and DISLoDGED pathways occur. We analyzed 239 metabolic disease-linked genes from MDDB and found an enrichment of disease-linked genetic variants among transporters (one-tailed hypergeometric $p = 0.0048$) and inositol metabolism (one-tailed hypergeometric $p = 0.02$), as well as a depletion of variants in central carbon metabolism (one-tailed hypergeometric $p = 0.035$). We found that DISLoDGED pathways showed similar results, including an enrichment of down-regulated DISLoDGED pathways in inositol metabolism ($p = 0.018$) mirroring the enrichment found in disease-linked genetic variants, as well as enrichment of DISLoDGED pathways in non-central metabolism (Supplementary Table 1). These results indicate certain metabolic pathways may be inherently less robust to pathological perturbation.

1.3.8 DISLoDGED pathway associations with disease physiology

Next, we examined whether DISLoDGED pathways are known be essential to system physiology in a manner that could result in disease when perturbed. To compare DISLoDGED pathways to disease pathophysiology, side effects were grouped based on affected physiological systems, such as renal diseases or autoimmune complications. We first grouped DISLoDED pathways by nearest nutrient for better coverage in the literature. DISLoDGED nutrient pathways preferentially altered by drugs causing side effects in specific physiological systems were then calculated. In 17 of the 18 cases of enrichments of down-regulated DISLoDGED nutrient pathways within particular physiological systems, the down-regulated pathways had directionally consistent links to the disease pathophysiology (Figure 1.6b, Supplementary Data 2). For example, inositol metabolism down-regulation was enriched among drugs with side effects affecting the kidney, including kidney failure. Supporting this relationship, the kidney is a primary site of inositol synthesis, and inhibition of inositol transport has been reported to cause renal failure.

Similarly, pathways that were up-regulated by drugs affecting particular physiological systems also showed directionally consistent links to pathophysiology. In nine out of 17 DISLoDGED nutrient pathways that were up-regulated in particular physiological systems, inhibitors targeted at the up-regulated pathway were established therapeutics within the disease class (Figure 1.6c). For example, drugs causing cancer- or autoimmune-related side effects were enriched in up-regulation of folate metabolism, and anti-folates are commonly used in treatment of diseases in both classes. Supporting the implications of this up-regulation in side effect pathogenesis, studies have shown that folate supplementation is tied to increased incidence of both cancer and childhood asthma.

Figure 1.6: **Maps of interactions between DISLoDGED pathways grouped by nutrient domain and corresponding side effects grouped by disease class.** Only nutrients and disease classes with at least one marginally enriched nutrient-class interaction (hypergeometric p <0.1) are shown. a) Using the metabolic network reconstruction Recon 1, side effect-specific metabolic perturbations (DISLoDGED pathways) are grouped into nutrient domains to enable comparison with existing disease-related genetic, clinical, and pre-clinical data to assess the potential causality of observed perturbations. In this figure, the number of side effects with an up-regulation in the production pathway for a metabolite is shown in yellow boxes, while blue boxes show the number of side effects with a down-regulation in the production pathway for a metabolite. b) Down-regulated DISLoDGED pathways. Nutrient-disease class interactions indicating an enrichment of down-regulations in drugs causing side effects within the class are colored according to the legend. c) Up-regulated DISLoDGED pathways. Nutrient-disease class interactions indicating an enrichment of up-regulations in drugs causing side effects within the class are colored according to the legend. Many of the enriched interactions are consistent with known effects of nutrient/pathway perturbation on the corresponding disease classes and physiological systems.


a
Melatonin
Tryptophan
Niacin
CYP
IDO1
Melatonin
ASMT
AANAT
INMT
IDO1
DDC
IDO1
TPH
ALDH
MAO
ABP1
AOC2
DDC
KYNU
TDO2
L-Tryptophan
AADAT
CCBL1
DDC
to NAD
KYNU
KMO
KYNU
AADAT
CCBL1
side effects with met. production up-regulated
XY
side effects with met. production down-regulated
b
Down-regulated DISLoDGED Pathways
Inflammation
Kidney Disorders
Hematopoietic Disorders
Infection
Lipoate
Niacin
Carnitine
Inositol
Vitamin D
Melatonin
Sterols
Biotin
Tryptophan
Riboflavin
Vitamin C
Potassium
Choline
CoQ10
Arginine
Autoimmune and Allergies
Neural Disorders
Urinary Disorders
Pain
Eye Disorders
c
Up-regulated DISLoDGED Pathways
Hematopoietic Disorders
Kidney Disorders
Autoimmune and Allergies
Cardiovascular Disorders
Histidine
Riboflavin
Serotonin
Taurine
Carnitine
Folate
Cancer
Arginine
Inositol
Digestive Disorders
Methionine
Vitamin C
Sterols
Endocrine Disorders
Bleeding Disorders
Pain
Down-regulation Legend
Nutrient supplementation beneficial
Nutrient deficiency associated with diseases in class
Nutrient is key precursor in physiological system
Nutrient has conflicting ties to system physiology
Down-regulation enriched among drugs causing diseases in class
Down-regulation not enriched among drugs causing diseases in class
Up-regulation Legend
Inhibition of metabolic pathway beneficial
Metabolic pathway has established ties to physiology
Lack of evidence suggesting benefit of pathway inhibition
Up-regulation enriched among drugs causing diseases in class
Up-regulation not enriched among drugs causing diseases in class

1.3.9 Altered activity of DISLoDGED pathways in disease

We then sought to determine whether DISLoDGED pathways are significantly over- or under-active in corresponding clinical disease and disease models, based on several metrics in the MDDB. To perform this analysis, DISLoDGED metabolic pathways were associated with dietary nutrients nearest in the metabolic network. In an initial analysis of 323 DISLoDGED nutrient pathways, down-regulated DISLoDGED pathways were found to be significantly enriched in disease-associated nutrient deficiencies (59% enrichment, binomial p-value 0.0017), while up-regulated DISLoDGED pathways were depleted in disease-associated nutrients deficiencies (47% depletion, binomial p-value 0.036) (Figure 1.7a).

These results were confirmed in an independent set of 453 investigated DISLoDGED pathways added to MDDB. Down-regulated DISLoDGED pathways were more likely to have a causal down-regulation associated with the corresponding pathology (17% enrichment of down-regulation, binomial one-tailed $p = 0.045$) (Figure 1.7b), while up-regulated DISLoDGED pathways were significantly predictive of consistently over-active pathways tied to corresponding pathologies (67% enrichment of over-activity, binomial one-tailed $p = 0.003$) (Figure 1.7c).

1.3.10 Effect of targeted inhibition of DISLoDGED metabolic pathways

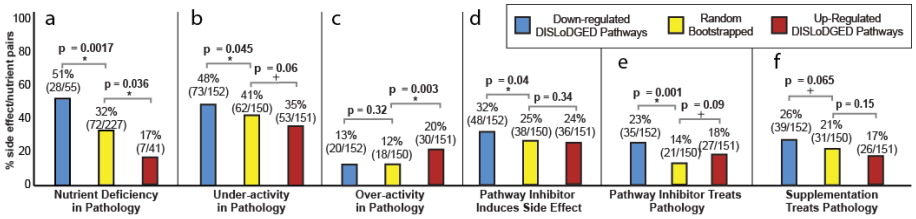
Next, we analyzed *in vivo* data on the effect of non-drug chemical inhibitors targeted at calculated DISLoDGED pathways. A causal relationship would be indicated by 1) inhibitors targeted at down-regulated DISLoDGED pathways reproducing the clinical side effect, and 2) inhibitors targeted at up-regulated DISLoDGED pathways treating the clinical disease. We found that metabolic inhibitors targeting down-regulated DISLoDGED pathways specifically were significantly more likely to cause corresponding side effects (28% enrichment, binomial one-tailed $p = 0.04$) (Figure 1.7d). Also, we found that both down-regulated and up-regulated DISLoDGED pathways were more likely to be known targets for effective metabolic inhibition to treat corresponding diseases, indicative of imperfect

directionality of predictions (64% enrichment, binomial one-tailed p = 0.0015 and 29% enrichment, binomial one-tailed = 0.09, respectively) (Figure 1.7e).

1.3.11 Effect of supplementation targeted at DISLoDGED pathways

Finally, we compared calculated DISLoDGED pathways grouped by nearest nutrient with clinical therapeutic nutrient supplementation data (Figure 1.6a). We sought to determine whether down-regulated DISLoDGED pathways might be targets for nutrient supplementation as a disease therapy. We observed that predicted down-regulated DISLoDGED pathways were preferentially targets of nutrients supplementation to alleviate corresponding pathologies (24% enrichment, binomial one-tailed p = 0.065) (Figure 1.7e), while over-expressed pathways were preferentially depleted as effective nutrient supplements (19% depletion, binomial p = 0.14), although the relationships did not reach statistical significance given sample sizes available.

Figure 1.7: **Multiple lines of clinical evidence in MDDB implicate a causal role for drug-induced metabolic transcription changes in side effect pathogenesis.** a-b) Down-regulated DISLoDGED pathways are found to be enriched in nutrient or pathway deficiencies associated with the corresponding pathology. c) Up-regulated DISLoDGED pathways are found to be enriched in consistently over-active pathways associated with corresponding pathologies. d) Inhibitors targeted at DISLoDGED metabolic pathways are more likely to effective treatments in corresponding pathologies. e) Metabolic inhibitors targeted at down-regulated DISLoDGED pathways are significantly more likely to cause corresponding side effects. f) Nutrient supplementation targeted at down-regulated DISLoDGED pathways was found to preferentially alleviate the corresponding pathology. p-values are defined for one-tailed binomial tests based on the level of the treated sample relative to the control (random). Sample sizes were chosen a priori based on feasibility of data collection, without an expectation for a particular effect size. Symbols: * indicates significant or p ¡ 0.05, + indicates marginally significant or p <0.1


a
b
c
d
e
f
Down-regulated DISLoDGED Pathways
Random Bootstrapped
Up-Regulated DISLoDGED Pathways
% side effect/nutrient pairs
100
80
60
40
20
0
p = 0.0017
*
51% (28/55)
p = 0.036
*
32% (72/227)
17% (7/41)
Nutrient Deficiency in Pathology
p = 0.045
*
48% (73/152)
p = 0.06
+
41% (62/150)
35% (53/151)
Under-activity in Pathology
p = 0.32
p = 0.003
*
13% (20/152)
12% (18/150)
20% (30/151)
Over-activity in Pathology
p = 0.04
*
32% (48/152)
p = 0.34
25% (38/150)
24% (36/151)
Pathway Inhibitor Induces Side Effect
p = 0.001
*
23% (35/152)
p = 0.09
+
14% (21/150)
18% (27/151)
Pathway Inhibitor Treats Pathology
p = 0.065
+
26% (39/152)
p = 0.15
21% (31/150)
17% (26/151)
Supplementation Treats Pathology

1.4 Discussion

The systems biology-based workflow developed in this study predicts side effect-linked dysregulated metabolic pathways (termed DISLoDGED pathways) from drug treatment of human cells in culture. We contextualize numerous independent data types that together support a key role of drug-induced non-pharmacokinetic metabolic dysregulation in side effect pathogenesis. Through the construction of a comprehensive resource on metabolic involvement in disease, we corroborate the predictions made using five independent clinical, genetic and biochemical bodies of literature. These results provide understanding of the mechanisms underlying side effect pathogenesis, which have remained largely opaque despite recent progress identifying causal protein binding events and genetic susceptibility factors.

The work presented relied upon the development of a systems biology workflow to identify side effect-linked features and validate the relevance of these features using historical *in vivo* data. The use of systems biology approaches to study drug side effects has become an active field in recent years. Previous studies have predicted drug-protein binding events responsible for side effects, effective combinations of drugs to minimize side effects, and drug mechanisms of action underlying side effect pathogenesis. Still, uncovering how causal binding events result in disease is an essential and largely unanswered question, as identified pathogenic mechanisms can potentially be used to design therapies to circumvent drug side effects. The present work we believe is the first to validate omics-driven predictions of post-binding mechanisms underlying side pathogenesis using clinical data at a large-scale, which was empowered by the generation of a new database through manual curation of the literature.

The workflow presented is highly dependent upon large amounts of gene expression data obtained under treatment by diverse drugs to filter out perturbations due to factors other than common side effects. Normalization of data across multiple studies and platforms is a substantial issue. Thus, data generated within a single study is ideal, but such large studies are rare. For this reason, the potential to extend the workflow to new data types, for example *in vivo* gene expression

data, may be limited. Due to the use of *in vitro* data, corroboration against *in vivo* data appears critical. Further deployment of the presented workflow may hinge upon expansion of curated disease data necessary to corroborate the clinical relevance of *in vitro* drug treatment features. Furthermore, the definition of pathways used to integrate disparate data types has yet to be concretely established and may be an area for further workflow optimization.

Due to the ubiquity of gene set enrichment analysis (GSEA) in pathway-based analysis of gene expression, a discussion on the differences between the MetChange algorithm and GSEA is warranted. Both methods attempt to aggregate signal in gene expression data along pathway definitions to increase the interpretability of the data and decrease the effect of noise. GSEA uses a variety of pathways, including manually curated metabolic pathways, and results are typically p-values of a Kolmogorov-Smirnov test for the likelihood that the cumulative distribution of expressions of genes in each pathway have not changed between conditions. MetChange defines a different production pathway for each metabolite in the network, based on calculation of functional states of the metabolic network, and scores for each metabolite how gene expression along this production pathway has changed between conditions. As a result, MetChange has the potentially to give finer resolution results, since its results are defined at the level of individual metabolite scores rather than pathway scores. However, the overall performance of the methods are difficult compare due to the fundamental difference in resolution of outputs.

In the comparison between MetChange and comparable methods (Figure 1.3), it is apparent that, on the data set used, output correlations of none of the methods with measured metabolite perturbations in yeast were particularly high. However, given that this analysis is possibly using these methods out of their intended use cases, no presumptions should be made regarding the general usefulness of any of these methods. Rather, this may highlight the difficulty in establishing standards by which to compare pathway analysis algorithms with statistical rigor. In this study, we performed the analysis primarily to provide some context for the relative performance of the MetChange algorithm at one specific task relevant

to predicting metabolic changes occurring within the cell. We note that a recent method to predict metabolomics changes from gene expression has reported statistically significant correlations on three other yeast data sets, including values that exceed those reported in Figure 1.3.

There are at least two obvious potential improvements that could be made to the workflow used in this work. First, as the cell lines within the Connectivity Map are within the NCI60 cell line panel, and expression, growth and most recently exometabolomic data have been measured for these lines, cell-specific metabolic models could be used in place of a global model. Second, the human metabolic network reconstruction has been updated with the publication of Recon 2, and thus improvements might be made from the increased scope of the model. We evaluated the potential for improvements using these changes by constructing cell-specific models for the MCF-7, HL-60, and PC-3 cell lines based on Recon 2 and running the MetChange algorithm on randomized simulated expression data drawn from an empirical distribution of MAS5.0 normalized p-values from the Connectivity Map data used in this study. Briefly, between cell-specific Recon 2 models, error was between 10-25% across replicates when given the same metabolite uptake constraints. The largest difference observed was by enforcing measured metabolite uptake constraints. Differences in MetChange scores between models constrained and unconstrained by metabolite uptakes were around 70%, indicating that significantly constraining the flux state of the model alters MetChange scores more so than topological differences. Thus, further improvements in prediction accuracy might be seen by accounting for the baseline metabolic differences between cell lines.

One interesting outcome of the work was that, while previously identified genetic susceptibility factors in the literature are dominated by genes involved in pharmacokinetics, we observe overlap of drug-induced metabolic changes with nine genes that affect pharmacodynamics. This may suggest that such non-pharmacokinetic genes may play a larger role in side effect pathogenesis than currently appreciated. We did however see some alteration of pharmacokinetic genes in a few cases as well (Supplementary Note 3). Additionally, DISLoDGED pathways

overlapped with clinical disease-linked pathways for both drug-treated and drug-independent studies, suggesting a common basis in pathogenesis. Furthermore, our results suggest that targeted nutrient supplementation may be a relatively simple and inexpensive path to broadly reduce side effect incidence. The impact of drugs on patient metabolic status is thought to be an underappreciated but important aspect of drug response, and this work further suggests this interaction is worthy of significant investigation. Patient attrition may be reduced through effective nutrient supplement to drug pairing during the development process.

A natural question that may arise is whether certain pathways or side effect disease classes are more successfully predicted by the method used in this work than others. A sensitivity analysis shows that the method appears to be fairly robust in being able to predict DISLoDGED pathways in various areas of the metabolic network and across disease classes. Restricting to at least 5 pairs investigated for both the side effect-linked deficiencies and random pairs, 9 nutrients deficiencies (polyunsaturated fatty acids, coenzyme Q10, melatonin, niacin, riboflavin, steroids, thiamine, vitamin A, and vitamin D) had disease associations better predicted by side effect-linked deficiencies compared with random, while one (choline) did not. Similarly, in side effects with at least two pairs investigated in both random and real, 7 side effect/nutrient deficiency relationships were better predicted by side effect-linked deficiencies (dyspepsia, epilepsy, hyperlipidemia, interstitial nephritis, tardive dyskinesia, testicular atrophy, and thrombocytopenia), while only 2 were better predicted by random nutrients (ecchymosis and tendonitis). These types of sensitivity analyses on predictive capability for particular nutrients and side effects are vulnerable to error due to small sample size but corroborate overall results that dysregulated DISLoDGED pathways are predictive of metabolic pathways associated with corresponding pathologies.

Although challenges remain, the ability to observe perturbations important to *in vivo* side effect susceptibility within *in vitro* data suggests that early-stage drug screening to identify and manage side effect risk may become possible, analogous to other disease in a dish efforts. Notably, high use drugs such as statins and antipsychotics, where patient populations are large enough for statistical analysis

of rare side effect events, dominate cases of side effects where genetic components have been identified. However, if vulnerable pathways can be identified through analysis of *in vitro* data, it may become easier to identify susceptibility factors for more rare disease classes as well. This workflow is currently limited to cases in which gene expression alteration underlies side effect pathogenesis, which is an undefined subset of all side effects. The results presented show the utility of integrating large, standardized datasets, such as the Connectivity Map, with clinical data types such as side effect incidence, genetic studies, and disease-nutrient associations, in the context of a highly-annotated network with clear functional connections. Such integration of disparate data sources is a key challenge in many areas of the life sciences today.

1.5 Methods

1.5.1 Overview of computational approach

As described schematically in Figure 1.1, with more detailed methods diagrams in Figure 1.2 and Figure 1.5, we employed a combination of constraint-based modeling and machine learning to look for metabolic gene expression perturbations that are conserved in adverse drug reactions. The first step is the analysis of drug-specific metabolic perturbations using the constraint-based MetChange algorithm described below and comparison of these perturbations to drug-specific response properties. We then combined MetChange scores based on side effects and used an AUC-maximizing classification genetic algorithm described below to determine small subsets of metabolic changes highly conserved in certain side effects. We note in general that the fields of constraint-based modeling and machine learning have developed a wide variety of methods that perform similar tasks. We compare our method with several other constraint-based methods and with gene expression data alone, as described in the Results. We also qualitatively contrast our method with gene set enrichment analysis in the Discussion. In general, performance of machine learning methods largely depends on the particular problem. We note that in cases of rare events, such that only a small fraction of samples

have a property, AUC (or rank) maximizing algorithms have been shown to perform particularly well. Additionally, we choose to place a hard constraint on the number of variables rather than using a traditional SVM with an L2-norm approach, for example. This was done to maximize interpretability of the resulting signatures, which was critical for later comparison of DISLoDGED pathways with pathology deficiency relationships and disease-linked and side effect-linked genetic susceptibilities. Thus, while we do not discount that other methods may exhibit better performance by certain standards, we chose our approach to meet the specific requirements of our problem, as is typical in the field. Full description of methods is below and a MATLAB implementation is provided (see Supplementary Software).

1.5.2 Connectivity Map data processing and integration

Gene expression data for the AffyMetrix HT Human Genome U133A platform was obtained from the Connectivity Map (CMap) database for MCF-7, PC-3, and HL-60 cell lines. Data was MAS5.0-normalized with the BioConductor package. Human Entrez Gene identifiers associated with probes were used to map detection p-values to their corresponding reactions based on the Boolean gene-protein-reaction (GPR) associations as was previously described. Reactions that were not associated with a gene were assigned a p-value of 0. Metabolic exchange was set to an exchange value of -1 for DMEM (MCF-7) or RPMI (PC-3 and HL-60) media metabolites. In cases where no metabolite production was possible with open constraints, the metabolite was removed from further analysis. For the cases of *in vitro* experimental validation under sodium phenylbutyrate and genistein treatment, for which data was generated for the fraction of metabolites generated from glucose, the MetChange algorithm was run using glucose as the sole carbon source to enable direct comparison with the data.

1.5.3 The MetChange algorithm

The Metabolite-Centered Hotspots of Altered Network Gene Expression (MetChange) algorithm was used to analyze differential quantitative gene expression profiles in the context of a genome-scale metabolic network. This algorithm is built upon the Gene Inactivity Moderated by Metabolism and Expression (GIMME) algorithm previously developed to build context-specific metabolic networks based on gene expression data.

The MetChange algorithm defines a consistency metric of an expression profile with optimally producing each metabolite in the metabolic network. For each metabolite, a sink reaction is created and flux through the sink reaction is maximized using flux balance analysis (FBA). This results in a set of optimal production fluxes for each network metabolite.

Each reaction is then weighted by a detection p-value from mapped expression data to solve a second optimization problem. To obtain metabolite production consistency scores, xi, for each ith metabolite, scores were generated by setting the lower bound of the metabolite sink reaction to its maximal production flux, vmax,i, and minimizing the inner product of the reaction detection p-values. This optimization is the GIMME-like step of the MetChange algorithm. Each ith component in cp serve as a cue for whether a reaction is detected or absent (i.e. a lower p-value indicates a reaction is more likely to be present). Hence, this second optimization determines: (1) a flux distribution which maximizes the production of a metabolite, and (2) generates a weighted flux distribution score defining consistency of the production of that metabolite with expression data. These scores cannot be directly compared across metabolites due to the fact that each metabolite has a different network flux state at maximal production. Thus, we compared metabolite production consistency scores between treated and control samples with a standardized score. A MATLAB implementation of the MetChange algorithm is provided in the Supplementary Software file.

1.5.4 Analysis of expression data for yeast under carbon and nitrogen starvation

Metabolite concentration and gene expression data for S. cerevisiae were mapped to a genome-scale yeast metabolic reconstruction, iMM904. For all methods, scores were compared to log-2 metabolite concentration changes relative to the initial time point for carbon or nitrogen data. Spearman correlations were calculated to avoid biases due to variable score distributions among the different methods. K-means clustering analysis was performed using 100 replicates. In all cases, the same clusters were obtained in repeated runs.

For the MetChange algorithm, reported expression values were used to generate reaction presence/absence p-values. Expression values were ranked, and based upon the apparent distribution; a value corresponding to the bottom 2nd percentile of expression was used to define the noise level. Expression values below this threshold were assigned a p-value of 1 as not present. The p-values below the threshold were then inverted across the threshold to generate a symmetric distribution. The mean and standard deviation of this approximated noise distribution were used to generate significance scores for the remaining expression values using a Z-test. These p-values were then mapped to reactions based on model-defined gene-reaction relationships. When multiple probes were assigned the same reaction, the minimum value was used. The MetChange algorithm was then applied using these mapped reaction presence/absence p-values. As the data was longitudinal in time and multiple controls were not available, log-2 scores were calculated with respect to the initial time point, rather than taking standard scores. Reporter metabolite analysis was then implemented. p-values for the significance in expression change were used as inputs to the reporter metabolite analysis. Each expression value was first standardized across conditions and then p-values were calculated using a standard Z test for each gene across condition. 10000 permutations of the data were used to generate the background levels of perturbation. In addition, the E-Flux method was implemented, with metabolite sinks set as separate objective reactions similar to the MetChange algorithm for comparison. Optimal flux through each metabolite sink reaction was then calculated using FBA

for each gene expression data set, and log-2 scores compared with the initial time point for each carbon and nitrogen data set were calculated.

For a direct comparison with gene expression level cues, we implemented the following approach. First, standardized scores of expression level changes were calculated with respect to all gene expression datasets for the particular perturbation (carbon or nitrogen starvation). The scores were mapped to their respective reactions according to the gene-reaction association in the metabolic network. We then assessed whether absolute changes in expression levels are predictive of the magnitude of metabolite change by adding the standardized reaction scores, weighted by the absolute stoichiometric coefficients. We then assessed whether higher gene expression for a reaction was indicative of a higher product concentration and lower reactant concentration. The mapped reaction score was added to the score for all metabolites in the reaction, multiplied by the signed stoichiometric coefficient for each metabolite.

1.5.5 Analysis of metabolic response phenotypes across drugs

The Recon 1 metabolic network was first converted into an irreversible network, such that each reaction proceeds only forward, and reactions that can proceed in multiple directions are split into two forward-proceeding reactions. The MetChange algorithm was run using gene expression presence/absence MAS5.0 p-values from the Connectivity Map (CMap) database build 02. When multiple controls were present, a standard score was generated. When a treated sample was from a batch with a single control, the mean and standard deviation of all control samples was used instead.

Cell line standard scores were then generated in the following manner. First, for each cell line, the median scores of all samples for each drug were found and used as the cell line-specific response. Then, to simplify compartment-specific scores to a general metabolite response, cytosolic metabolite scores were taken when available. If no cytosolic metabolite existed, the median of scores across all compartments was taken as the metabolite score. Finally, a standard score

across all drugs was calculated for each cell line. Consensus drug perturbations across cell lines were calculated by averaging cell-specific MetChange scores and standardizing across all drugs.

1.5.6 PubMed querying of drug associations

To identify drug-metabolite associations in the literature in an automated fashion, PubMed/MEDLINE records were downloaded from the National Library of Medicine and abstracts were parsed to raw text. Chemical entities were tagged using Reflect. A training set of 44 abstracts describing true drug-metabolite correlations (positive set) and 150 abstracts with other chemical entities (negative set) were used to train a Bayesian network to recognize abstracts that mention the causative relationship between an administrated drug and the presence of a certain metabolite . A second Bayesian network was trained to recognize sentences within the abstracts that refer to metabolites from those that refer to other chemical entities. For each drug-metabolite co-occurrence the two Bayesian networks were used to assign a posterior for abstract occurrence and a posterior for sentence occurrence and the joint probability of the two posteriors was used as the drug metabolite score. The evaluation of the output is provided in Supplementary Data 1. The text-mining output was evaluated using a random sample of 100 pairs. The abstracts that describe each sampled pair were manually checked and identified as true or false positives. The score cutoff was set to 0.33, which provides an 82% true positive rate and 70% accuracy. Text querying on side effect-metabolite, protein-metabolite, and drug-protein co-occurrences were performed using the Entrez Programming Utilities (NCBI) using a simple AND specification. Side effects were obtained from the Side effect Resource (SIDER) database. Signaling protein lists were obtained from various publically-hosted resources including the International Union of Basic and Clinical Pharmacology database (http://www.iuphar.org/). Metabolite common names were obtained from the Recon 1 network reconstruction. All associations are included in Supplementary Data 1.

1.5.7 Determining metabolite network distance from drug perturbation

To calculate network distance from metabolites, the DrugBank database was downloaded and cross referenced with drugs from cmap. A total of 134 drugs present in cmap have targets in Recon 1 reported in DrugBank. Across the three cell lines (HL60, MCF7, PC3) and multiple drug concentration ranges, there were a total of 611 expression profiles of drug-perturbed states. The median metabolite network distance of each metabolite was calculated for each of the 134 drugs, ignoring metabolites with reaction connectivities greater than 30, such as cofactors, protons, and water.

MetChange scores were compared against expression data, randomized data, and reporter metabolite analysis results. Data and scores were pooled into bins based on their standard scores. For each bin, the average of the median network distance of metabolites to metabolites of known drug-targeted enzymes was calculated for comparison. Metabolites predicted from the expression data set were determined by taking the highest expression change among reactions involving the metabolite and binning accordingly. The random dataset was generated by permuting the expression data set 1000 times and calculating the average bin value across all 1000 sets. The reporter metabolite data set was generated as described in the original publication using 10000 permutations to generate background perturbation levels.

1.5.8 Analysis of drug-signal protein signatures

Drug-protein and protein-metabolite literature co-associations were found as described. Associations were binary based on the presence or absence of known literature association. A receiver operating characteristic (ROC) curve was generated for the ability of MetChange scores to indirectly predict drug protein association. MetChange scores for all three cell lines (not averaged) were used together. At increasing thresholds from 0 to effectively infinity, metabolites with absolute MetChange scores greater than the threshold were scored as significantly changed

for each drug. Proteins associated with perturbed metabolites were then determined using the protein-metabolite literature co-associations. These proteins were used as guesses for true drug-protein literature associations for the drug corresponding to the sample, and the true positive rate (TPR) and false positive rate (FPR) were calculated. Varying the threshold from 0 to effectively infinity then generates the ROC curve. To assess statistical significance of the resulting area under the curve (AUC), 1000 permutations of MetChange scores were analyzed in the same way and a non-parametric rank test was conducted on the resulting AUCs.

1.5.9 Determination and analysis of drug side effect signatures

Drug side effects were taken from Side Effect Resource (SIDER) database for available drugs overlapping with the CMap database. The SIDER database contains minimum and maximum occurrence frequencies for a number of both treated and control studies for each drug-side effect pair. Side effect frequencies were processed in the following manner. The mean of the minimum and maximum frequency was calculated for each study, and then the median of frequencies from all studies was found for both treated and placebo studies. The difference between treated and placebo occurrence frequency was then calculated. If placebos were not available, the treated frequency was used. These frequencies were then mapped to all expression samples from CMap corresponding to the appropriate drugs. A minimum of 30 expression sets for a side effect were required for inclusion in the analysis. A total of 357 side effects were analyzed for 850 expression sets corresponding to 334 drugs.

A genetic algorithm was then implemented, termed SiderFinder (Figure 1.5). The matrix of MetChange scores for the 850 expression sets was input as well as the corresponding side effect frequencies for a particular side effect. A maximum number of predictor metabolites was set to 20 metabolites. A set of 125 candidate solutions was generated, assigning values of -1, 0, or 1 to each metabolite to indicate negative, no, or positive prediction of a high MetChange metabolite

score on occurrence of side effect. Each expression set was then scored for the side effect for each individual by multiplying the predictor set for the individual by the MetChange scores for the expression set. These side effect scores were ranked and a pseudo-ROC curve was generated by comparison of scores with the side effect frequencies for the current side effect. At each threshold, expression sets with a side effect score greater than the threshold were called as having the side effect. To weight more heavily samples with higher side effect frequency, the base 10 logarithm of each side effect frequency was taken and adjusted such that the lowest non-zero frequency had a value of 1, each order of magnitude greater in frequency is a unit greater, and all zero frequency side effects remain zero. An ROC curve was then calculated with true positive hits being assigned the value of the adjusted side effect frequency, with no effect to false positive, true negative and false negative values. The number of true values was taken to be the sum of the adjusted frequency vector so the AUC of the pseudo-ROC still spans 0 to 1. The AUC of this curve was then used as the objective function to maximize in the genetic algorithm. 10-fold cross validation was performed, and perturbations that appeared cumulatively in at least 4 of the 10 sets we selected. Using the genetic algorithm directly on gene expression data achieved similar classification performance (results not shown), but metabolite scores were chosen as the variables due to their previously established performance in predicting drug-specific metabolic perturbations.

Genetic algorithm creation, mutation, and crossover parameters were used as implemented in the OptGene function of the COBRA Toolbox 2.0. The genetic algorithm was solved using the Global Optimization Toolbox in MATLAB (MathWorks). A MATLAB implementation is provided.

1.5.10 Co-association analysis of drug side effect signatures

Statistical analysis of literature co-associations of side effects with side effect metabolite signatures was performed with 1) a non-parametric permutation test on the drug side effect metabolite signatures against 1000 permutations of the signatures for the AUC of predicting presence/absence of literature side ef-

fect/metabolite co-association in Pubmed abstracts, and; 2) a hypergeometric test for the enrichment of literature association among side effect/metabolite pairs in predicted signatures.

1.5.11 Construction of Metabolism-Disease Database (MDDB)

To enable statistical analysis of predicted DISLoDGED pathways, disease-nutrient pairs were randomly selected from observed down-regulations, observed up-regulations, and random associations chosen by resampling the former two lists through bootstrapping. These disease-nutrient pairs were then evaluated for existence of established relationships using manual literature searches while blind to the origin of the pair to prevent investigator bias. The list of collected metabolism-disease relationships is not yet comprehensive due to the scope of the effort, but instead relationships were investigated in a targeted manner.

Data collection for the database covered literature up to and including May 2013.

1.5.12 Pathway analysis

We curated a database of 1394 distinct GWAS publications and extracted 239 distinct disease-associated metabolic genes that overlapped with the 1496 genes in Recon 1. We then assigned pathways to each disease-associated metabolic gene as well as to each DISLoDGED pathway calculated in our analysis (Supplementary Data 1). Disease-associated genes were assigned pathways based on the previously assigned pathway of corresponding reaction assigned in Recon 1. DISLoDGED pathways (which are metabolite-centered) were associated with disease-linked pathways by determining the most frequent pathway among all reactions in which a metabolite takes part. To assess enrichment, hypergeometric tests were performed to determine whether the observed coverage of metabolites or genes was enriched or depleted in particular pathways, controlling for multiple hypothesis testing.

1.5.13 Side effect pathophysiology classifications

To assess whether common metabolic perturbations were observed among related diseases, we manually grouped side effects by pathophysiological disease class. Enrichment was assessed with the hypergeometric test, at a significance level of $= 0.1$. Enrichment where there was only a single disease within the class was discarded, as were conflicting cases (in which a nutrient was down-regulated in certain diseases in the class and up-regulated in others).

1.5.14 Side effect-linked gene polymorphism search

We attempted to identify all cases of genetic basis for side effect incidence reported in the literature, including GWAS and targeted genetic studies. To be eligible for comparison, we required that the side effect be a near or exact match and the metabolic pathway of the susceptibility gene be within the scope of our model. For example, immune-related genes were excluded due to non-specific metabolic association, while G protein-coupled receptors known be regulated by particular metabolic pathways were included. We generally excluded genes related to drug pharmacokinetics, including drug metabolism and transport, as the effects of polymorphisms on susceptibility are generally non-specific to the pathology. We also required that the pathology be manifested in nucleated cells (i.e. excluding red blood cell pathologies), as gene expression changes are assumed to be irrelevant to pathologies of enucleated cells. Based on these criteria, well over 20 studies were excluded, while 9 genetic susceptibilities spanning 6 side effects were valid for comparison with predictions. We also mention two cases in which pharmacokinetic genes do overlap with conserved side effect-linked gene perturbations, suggesting possible interactions between pharmacokinetic and gene expression effects of drug perturbation. Lists of included and excluded studies are found in Supplementary Note 1 and Supplementary Note 4, respectively.

1.5.15 Nutrient deficiency literature search

To populate MDDB, we searched the literature for associations between the pathology of the side effect and deficiency relationships between the closest nutrient to the metabolic perturbation and the occurrence of the pathology of the side effect. In an automated search, PubMed abstracts were searched for a number of side effect and nutrient synonyms along with a list of deficiency synonyms (Supplementary Data 2). Statistical significance of enrichment of co-association PubMed abstract hits (presence/absence) among down-regulated and up-regulated nutrient pathways was assessed through a dual permutation analysis. For each perturbed side effect-nutrient pair, 1000 permuted pairs were generated by first randomly selecting the presence/absence result for a random nutrient with the same side effect, and then randomly selecting the presence/absence result for a random side effect with the same nutrient, then averaging the result. Then the number of permutations with total presence calls greater than the true observation was counted and divided by the total permutations, as is typical in permutation tests.

In the manual alteration search, a number of possible nutrient-disease relationships were identified, such as an inhibitor of the metabolic pathway causing the side effect, an inhibitor of the metabolic pathway curing the side effect, etc. (Supplementary Data 2). Then, search terms were generated using synonyms, and PubMed was searched. "Results were then filtered such that each nutrient-side effect pair was assigned "up-regulated activity associated with the disease", "down-regulated activity associated with the disease", "no associated with the disease", or "conflicting associated with the disease". Inconsistent results were assigned as conflicting and were excluded from further analysis." Significance of enrichments of particular relationships among up-regulated and down-regulated pathways were then evaluated with the hypergeometric test.

In the manual deficiency search, deficiency relationships were defined such that the DISLoDGED nutrient pathway could meet any of three possible criteria to be considered a hit. First, the deficiency of the nutrient could be known to be associated with the occurrence of the side effect pathology. Second, supplementation with the nutrient is known to alleviate the side effect pathology. Third, a physio-

logical dysregulation of the pathway is known to be associated with the side effect pathology. PubMed and Google Scholar were both used for this study. Patents were not accepted as valid references, unless associated with a peer-reviewed publication.

To determine whether DISLoDGED pathways are significantly predictive of side effect/nutrient deficiency relationships, we generated a list of random side effect/nutrient pairs for comparison with DISLoDGED pathway-disease pairs through resampling of the pairs. Kolmogorov-Smirnov tests were used to ensure the distributions were not significantly different between DISLoDGED pathways and random side effect-nutrient pairs in terms of the frequency of occurrence of nutrients, as resampling should guarantee.

We then performed the nutrient deficiency literature search in two phases. The first was blinded and the second was an expansion and additional curation of the blinded study. The initial study was blinded to ensure that there was no selection bias in the search process. Random pairs were mixed with DISLoDGED pathways, both up-regulations and down-regulations, and information on their origin was removed. The literature searches for deficiency relationships were then conducted by investigators not involved in the generation of the distribution and thus were unaware of the treatments or treatment distributions. Statistical tests were only performed on the blinded literature results. The initial blinded list was then expanded to examine additional DISLoDGED pathways and curated to ensure consistent criteria between investigators. Although the curated list is unblinded, we have a greater confidence in this list for research purposes due to the consistent criteria and expanded list of investigated relationships.

1.5.16 Overview of experimental validation of MetChange results

Drug-induced metabolic perturbations calculated using the MetChange algorithm were validated using *in vitro* experiments in the MCF-7 cell line for three drugs: metformin, genistein, and haloperidol.

1.5.17 sample preparation

Human MCF-7 breast carcinoma cells (ATCC, HTB-22) were maintained in supplemented DMEM media (CellGro Mediatech, 10013CV) with 10% v/v fetal bovine serum (Hyclone, SH3039603), 1% v/v antibiotic/antimycotic solution (Omega Scientific, PS-20), 1% v/v non-essential amino acids (Hyclone, SH3023801), 1% v/v MEM vitamins (CellGro Mediatech, 25020CI), 1 mM l-glutamine (CellGro Mediatech, 25005CI). 4 106 cells were seeded in supplemented MEM medium (CellGro Mediatech, 15010CV) into 150 mm dishes. For labeling with [U-13C] glucose (Sigma-Aldrich, 389374) the medium was replaced with supplemented MEM with 2 g/l glucose total of which 50% was [U-13C] glucose. Approximately 2.0 109 cells were harvested by incubation with trypsin for 5 min at 37 C (Gibco, 25200-056) and adjusted for total protein amount (Thermo Scientific Pierce, 23227). Intracellular polar metabolites were extracted by rapid quenching with 50% methanol at -40C; total lipids were extracted using chloroform (Sigma-Aldrich, 366919). The cell extracts were dried by vacuum evaporation. Organic acids were derivatized to form the corresponding oximes and trimethylsilyl derivatives. Acyl-carnitines were derivatized to their corresponding methyl esters. Polar metabolites were dissolved in 99.9% 2H2O with 0.75% 3-(trimethylsilyl)propionic-2,2,3,3-d4 acid (Sigma-Aldrich, 293040). Lipophilic metabolites were dissolved in 2H-chloroform with 1% tetramethylsilane (Sigma-Aldrich, 151831).

1.5.18 Drug treatment

Cells were exposed for a 24h period to varying drug concentrations (as indicated in their respective citations below) while control cells were exposed to the corresponding DMSO (CAS 67-68-5, Sigma-Aldrich, D8418) dilution: metformin (glucophage) (CAS 1115-70-4, 15169101, MP Biomedicals, stock 3M in PBS), genistein (CAS 446-72-0, 10005167100 Cayman Chemicals, stock 100mM in DMSO), haloperidol (haldol) (CAS 52-86-8, 15369690 MP Biomedicals, stock 50mM in DMSO). For acyl-carnitine measurements, the medium was supplemented with 1mM L-carnitine (CAS 6645-46-1, Sigma-Aldrich, C0283). Each experiment was performed in three biological repeats.

1.5.19 NMR spectroscopy

NMR experiments were performed on a 500 MHz Bruker Avance spectrometer with 5 mm TXI z-gradient probe (Bruker-BioSpin, Karlsruhe, Germany) at 298 K. 13C enrichments were determined by 1D 1H NMR difference spectroscopy (13C-coupled spectrum minus the 13C-decoupled spectrum). 2D indirect-detected 13C, 1H J-resolved HSQC spectra were recorded over an experimental time of 1.5 d per spectrum with 5120 indirect 13C data points, at 80 ppm 13C sweep width, 40 ppm carrier position, 4096 direct 1H increments, 3s recycling delay, and 8 scans. Before Fourier transformation, the data were multiplied with a squared sine-bell window function, phase corrected, and zero-filled to 8192 data points, for indirect 13C sampling.

1.5.20 Mass spectrometry

Quantitation of organic acids was accomplished by GC/MS (ThermoFisher Trace GC Ultra/DSQ II single quadrupole mass spectrometer). Calibration standards were prepared in water and spiked with stable isotope-labeled internal standards. Detecting the derivatized organic acids was achieved by single ion monitoring of each derivative after GC separation. Acyl-carnitines were quantitated by LC/MS/MS (Waters Ultra Performance LC/triple quadrupole mass spectrometer) using flow injection analysis with electrospray ionization. Calibration standards of the acyl-carnitines were prepared in bovine serum spiked with stable isotope-labeled internal standards. Parent ion scanning was used to detect parent, acyl-carnitine molecular ions that produced a characteristic acyl-carnitine fragment ion, m/z 99, formed by collision-induced dissociation.

1.5.21 Enzymatic assays

Pyridoxal 5-phosphate or Vitamin B6 was determined in an enzymatic plasma assay (A/C Diagnostics, ACB6001, San Diego, CA), which we adapted to quantify tissue culture supernatant or intracellular Vitamin B6 content by using PBST buffer with 0.1% Tween-20 (CAS 9005-64-5, Sigma-Aldrich P9416).

1.5.22 Additional methods for analysis of metformin response

Acyl-carnitine species from two to sixteen carbons and organic acids of the TCA cycle were analyzed by liquid chromatography mass spectrometry and show dose-dependent associations upon metformin treatment in MCF-7 cells. Acetyl-carnitine (C2) values were determined in nmol / mg protein (scaled by factor 1/1000) and were compared with acyl-carnitine (C12-C16) levels in pmol / mg protein. Organic acids were determined in nmol / mg protein.

1.6 Acknowledgements

We thank Irene Kouskoumvekaki and Gianni Panagiotou from the Center for Biological Sequence Analysis at the Technical University of Denmark in Lyngby for help in determining drug-metabolite associations in the literature. We also thank Bin Du, Edward OBrien, and Hooman Hefzi for discussions and comments. This work was supported by NIH grants R01-GM071808, P01-R00-CA128814, K99-CA154887, and R01-GM068837.

Chapter 1 in part is a reprint of the material *Zielinski DC, Filipp F, Bordbar A, Jensen K, Smith J, Herrgard M, Mo ML, Palsson BO. Pharmacogenomic and clinical data link non-pharmacokinetic metabolic dysregulation to drug side effect pathogenesis. Nat Commun. (2015) 6:7101.* The dissertation author was the primary author.

Chapter 2

Stoichiometric biomass synthetic requirements and metabolic stress resistance underlie hallmarks of cancer cell metabolism

2.1 Abstract

Malignant transformation is often accompanied by significant metabolic changes. To identify drivers underlying these changes, we calculated metabolic flux states for the NCI60 cell line collection and correlated the variance between metabolic states of these lines with their other properties. The analysis revealed a remarkably consistent structure underlying high flux metabolism. The three primary uptake pathways, glucose, glutamine and serine, are each characterized by three features: 1) metabolite uptake sufficient for the stoichiometric requirement to sustain observed growth, 2) overflow metabolism, which scales with excess nutrient uptake over the basal growth requirement, and 3) redox production, which also scales with nutrient uptake but greatly exceeds the requirement for growth. We discovered that resistance to chemotherapeutic drugs in these lines broadly correlates with the amount of glucose uptake. These results support an interpre-

tation of the Warburg effect and glutamine addiction as features of a growth state that provides resistance to metabolic stress through excess redox and energy production. Several other proposed roles of these hallmarks, including production of growth precursors and minimization of cellular protein, were examined and found to be not supported by data. These results provide a greater context within which the metabolic alterations in cancer can be understood and demonstrate the utility of integrative data analysis using constraint-based methods of metabolic systems biology

2.2 Introduction

Over the past decade there has been a revival of metabolic research in oncology[30]. In particular, two defining characteristics of cancer metabolism have received much attention: 1) an increased glucose uptake rate accompanied by secretion of lactate even in the presence of oxygen, known as the Warburg effect[31], and 2) a high glutamine uptake rate essential for growth, known as glutamine addiction. Despite the central role these traits play in the discussion of cancer metabolism, the drivers underlying these traits are still debated[31, 32, 33]. Understanding these drivers will become important as cancer metabolism becomes increasingly a target for chemotherapeutics[30].

The NCI60 cell line collection consists of 60 cancer cell lines that have been extensively used as a model to study characteristics of cancer cells over the past quarter of a century[34]. Notably, the metabolite uptake and secretion profiles for these lines were recently published[35]. When coupled to growth[36] and cell size data[37], these data provide the opportunity to study cancer metabolic functional states at an unprecedented scale by utilizing flux balance analysis[15] (FBA). Fundamentally structured in the context of metabolic mass balance, FBA has been utilized successfully over the past decade as a method of data integration[38] as well as a number of other applications, including cancer metabolism[39]. Using FBA, we integrated available metabolic data to calculate metabolic flux states for the NCI60 panel. We then leveraged the differences in metabolic flux states across

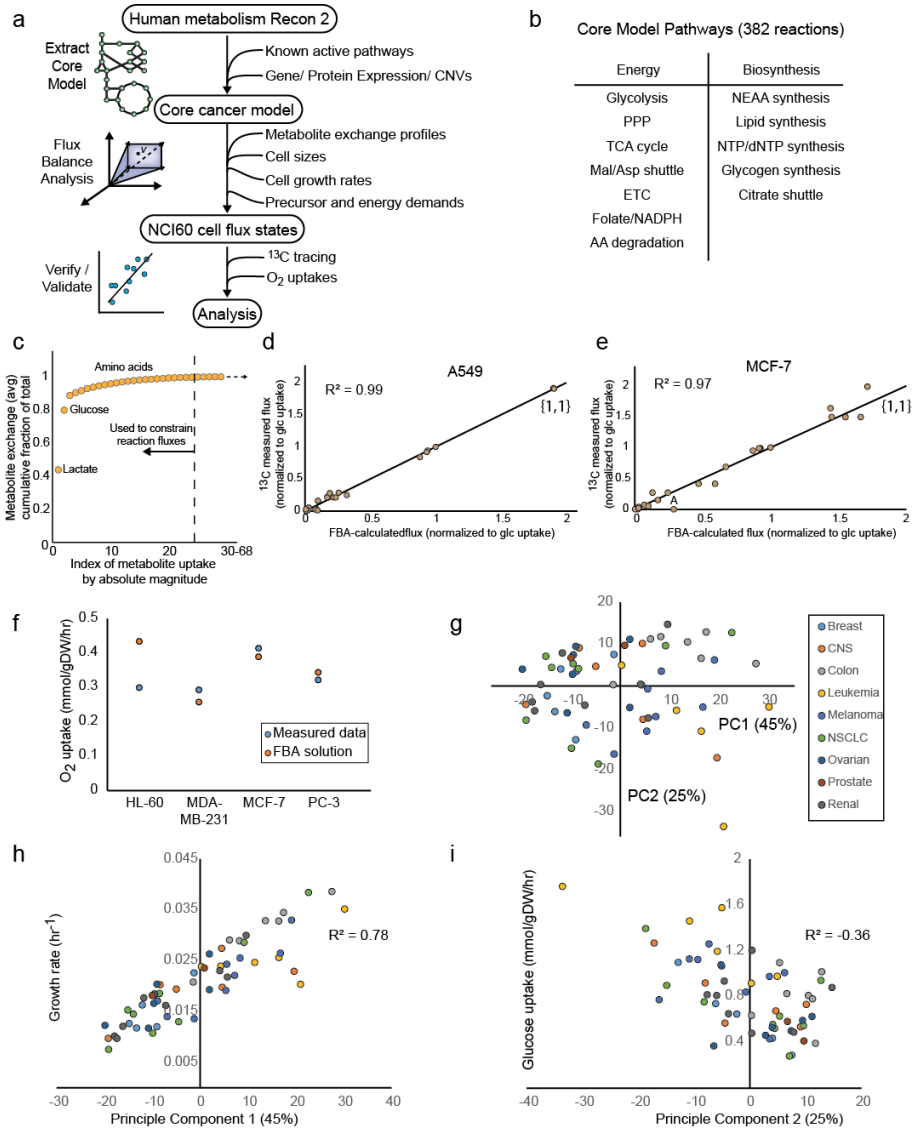
the NCI60 panel to identify drivers underlying two dominant features of cancer metabolism: the Warburg effect and glutamine addiction.

2.3 Results

First, we calculated metabolic reaction flux states for the NCI60 collection using flux balance analysis (FBA) constrained by measured uptake/secretion rates, growth rates, and cell sizes (Figure 2.1a-c). The computed flux states showed an excellent correlation with ^{13}C data available for the A549 and MCF-7 lines (Pearson $R^2 > 0.95$), after correcting for differences forced by disparities between measured uptakes in different data sets (Figure 2.1d, e). Computed flux states also showed good agreement with measured oxygen uptake rates (average error of 17%) for a subset of lines with available data[40] (Figure 2.1f). Principle component analysis (PCA) revealed that the metabolic states of the NCI60 lines are structurally very simple (Figure 2.1g), with large part of the variance explained by growth rate (Figure 2.1h) and glucose uptake differences (Figure 2.1i). Although there is no clear separation between distinct metabolic states apparent, examination of the states by tissue type shows some trends. For example, leukemia-derived lines appear characteristically fast growing with high glucose uptakes, while colon-derived lines appear characteristically fast growing with low glucose uptakes (Figure 2.1g).

We then analyzed flux states within amino acid metabolism in the context of biosynthetic demands on the NCI60 lines. The amino acids histidine and cysteine were not analyzed due to lack of uptake measurement, while methionine was excluded due to its close relationship with cysteine. Grouping amino acids by shared biosynthesis pathway, we first examined the essential amino acids. We found that essential amino acids, as well as the essential amino acid-derived tyrosine, are taken up at levels only slightly greater than is sufficient for protein synthesis. Each essential amino acid was found to be correlated with their biosynthetic demand, as has been reported[37].

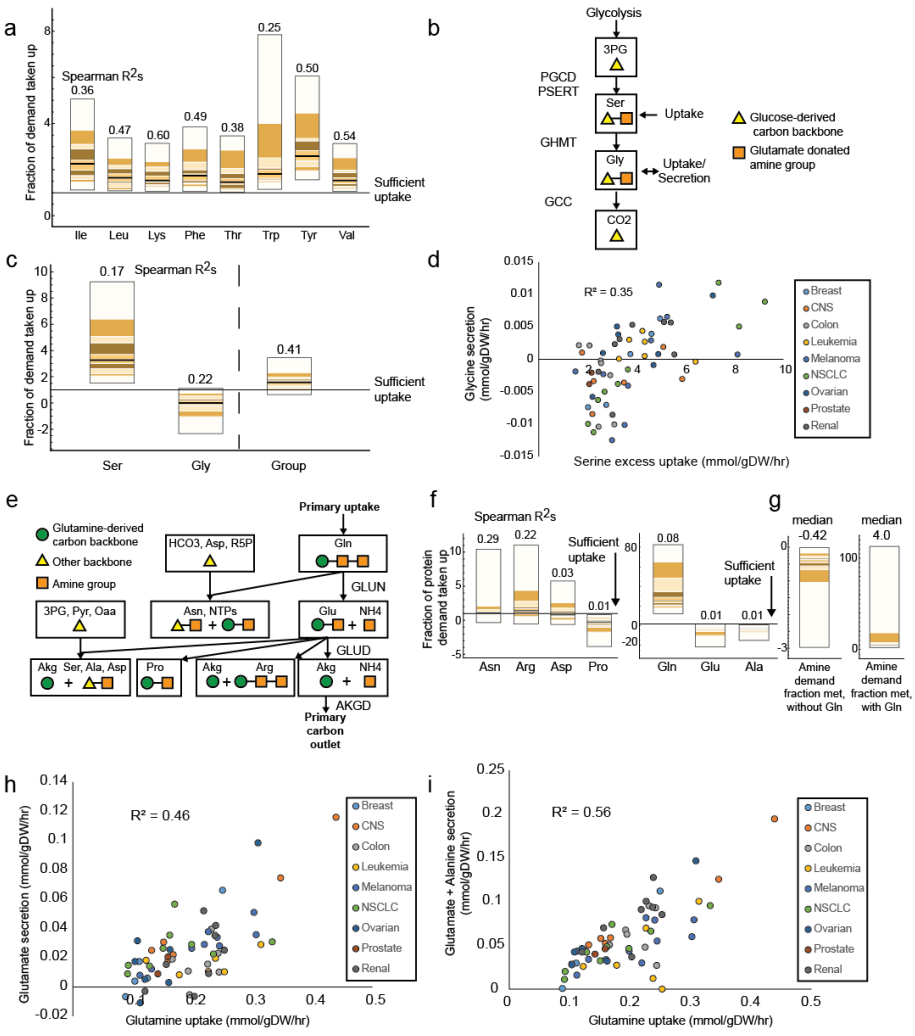
Figure 2.1: **Data-driven constraint-based modeling of a high confidence core cancer metabolic network results in accurate metabolic flux state calculations.** a) The workflow utilized in this study for the constraint-based calculation of metabolic flux states for the NCI60 panel using available data and a core metabolic model extracted from the global human metabolic network reconstruction Recon 2[41]. b) Summary of the metabolic pathways and functions represented by the core cancer metabolic network, consisting of 382 reactions. c) Cumulative distribution plot of absolute metabolite uptakes from a published data set[35]. Twenty-three of the highest flux metabolites were used as constraints on the core model (Supplementary Data 1), representing over 99% of the absolute metabolite exchange flux by mass. d) Comparison of flux balance analysis results to a previously published 13C-labeled glucose tracing experiment on the A549 line. The computed flux solutions were corrected for a substantial difference in measured lactate secretion prior to comparison. e) Comparison of flux balance analysis results to a previously published ^{13}C-labeled glucose tracing experiment on the MCF-7 line. The computed flux distributions were corrected for a difference in the active form of malic enzyme, which is forced by glutamine uptake to be the mitochondrial isozyme, consistent with other studies. Point labeled A: Citrate shuttling to lipid synthesis is not observed in the data but is forced to be active in the model by lipid synthesis requirements. f) Comparison of FBA-calculated and measured oxygen uptake rate data for a subset of the NCI60 lines. The average error was 17%, indicative of accurate prediction of electron transport chain activity. g) Principle component analysis (PCA) of the flux balance analysis (FBA) calculated flux distribution of the NCI60 cell lines. 60% of the variance is explained by the first two principle components, indicating that the flux solutions are relatively low dimensional. h) Correlation between the first principle component and the growth rate shows that the growth rate is a dominant determinant of the flux states. i) Correlation between the second principle component and glucose uptake, which was the highest correlated variable found, indicating that the second primary determinant of flux state is variance in glucose uptake at a particular growth rate.


a
Human metabolism Recon 2
Extract Core Model
Known active pathways
Gene/ Protein Expression/ CNVs
Core cancer model
Flux Balance Analysis
Metabolite exchange profiles
Cell sizes
Cell growth rates
Precursor and energy demands
NCI60 cell flux states
Verify / Validate
13C tracing
O2 uptakes
Analysis
b
Core Model Pathways (382 reactions)
Energy
Glycolysis
PPP
TCA cycle
Mal/Asp shuttle
ETC
Folate/NADPH
AA degradation
Biosynthesis
NEAA synthesis
Lipid synthesis
NTP/dNTP synthesis
Glycogen synthesis
Citrate shuttle
c
Amino acids
Glucose
Used to constrain reaction fluxes
Lactate
Metabolite exchange (avg) cumulative fraction of total
Index of metabolite uptake by absolute magnitude
d
A549
R² = 0.99
{1,1}
13C measured flux (normalized to glc uptake)
FBA-calculatedflux (normalized to glc uptake)
e
MCF-7
R² = 0.97
{1,1}
13C measured flux (normalized to glc uptake)
FBA-calculated flux (normalized to glc uptake)
f
O2 uptake (mmol/gDW/hr)
Measured data
FBA solution
HL-60
MDA-MB-231
MCF-7
PC-3
g
PC1 (45%)
PC2 (25%)
Breast
CNS
Colon
Leukemia
Melanoma
NSCLC
Ovarian
Prostate
Renal
h
Growth rate (hr-1)
R² = 0.78
Principle Component 1 (45%)
i
Glucose uptake (mmol/gDW/hr)
R² = -0.36
Principle Component 2 (25%)

Next we looked at the non-essential amino acids serine and glycine (Figure 2.22b). Serine uptake is consistently greater than its biosynthetic requirements and glycine secretion rate is highly variable. However, overall mass balance calculations show that if the glycine secretion is subtracted from the serine uptake, the difference matches the biomass synthetic requirements for these two amino acids combined, a feature previously observed[37] (Figure 2.2c). The correlation of the group uptake with biosynthetic demand is higher than either amino acid alone. Furthermore, glycine secretion increases as serine uptake increases above the protein requirement (Figure 2.2d), indicating glycine behaves as an overflow metabolite that regulates overall availability of the amino acid group when the glycine cleavage chain cannot sufficiently metabolize excess glycine.

The second non-essential amino acid group we examined consists of a set of amino acids linked to glutamine, including glutamine, glutamate, alanine, aspartate, asparagine, proline, and arginine (Figure 2.2e). We found that, on average, asparagine, arginine, and aspartate are taken up about the levels of their biosynthetic demand, but with a lower correlation with that demand than in the case of the essential amino acids, indicating greater variability (Figure 2.2f). Glutamate, alanine, and proline were all consistently secreted, in decreasing order magnitude.

Figure 2.2: **Amino acid metabolism is determined by protein synthesis demands and coupled to overflow metabolism.** a) Essential amino acid uptake compared with protein synthesis demand. Tyrosine is not an essential amino acid but was included in this set as it is taken up stoichiometrically with its protein synthesis requirement as are the other essential amino acids. Listed above each bar is the Spearman correlation R^2 of the amino acid uptake with biosynthetic requirement across lines b) Overview of serine and glycine synthesis pathway. c) Quantile plots of the uptakes of serine and glycine relative to their demand from protein synthesis. The sum of the uptake of the metabolites is approximately equal to the biosynthetic demand on the group as a whole, with any glycine deficiency provided by excess serine uptake. Listed above each bar is the Spearman correlation R^2 of the amino acid uptake with biosynthetic requirement across lines. d) Glycine secretion vs excess serine uptake, calculated as serine uptake minus serine and glycine protein synthesis demand. A significant correlation is observed, suggesting a role of glycine as an overflow metabolite when serine uptake exceeds the requirement for protein synthesis and ability to metabolize glycine through the glycine cleavage chain. e) Overview of the metabolism of glutamine and related biosynthetic precursors. f) Uptakes of glutamine-related amino acids compared to their protein synthesis requirement. It is observed that glutamine is taken up on average 32 times more than its requirement, while glutamate, alanine, and proline are secreted, in order of decreasing magnitude. g) Overall amine balance due to protein synthesis requirement. Without considering glutamine, cells do not take up enough amine to synthesis required amino acids. However, taking glutamine into consideration, the cell takes up 9.7 times more amine than necessary for growth, even after considering glutamate, alanine, and ammonia secretion. h) Secretion of glutamate versus glutamine uptake, showing the significant role of glutamate as an overflow metabolite coupled to glutamine. i) Secretion of glutamate and alanine vs the glutamine uptake. Although alanine secretion alone is not significantly correlated with glutamine uptake, considering glutamate and alanine secretion simultaneously improves the correlation with glutamine uptake over glutamate alone.

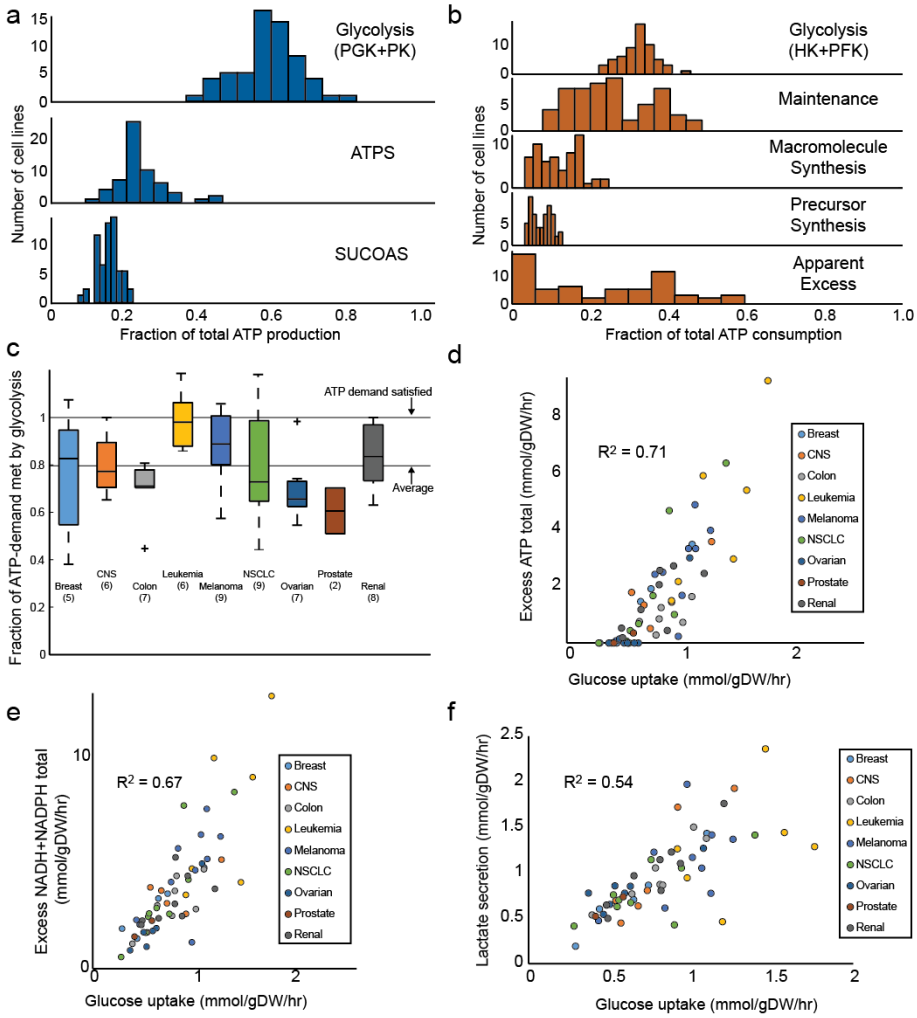

a
Fraction of demand taken up
Spearman R2s
0.36
0.47
0.60
0.49
0.38
0.25
0.50
0.54
Sufficient uptake
Ile Leu Lys Phe Thr Trp Tyr Val
b
Glycolysis
3PG
PGCD
PSERT
Ser
Uptake
GHMT
Gly
Uptake/ Secretion
GCC
CO2
Glucose-derived carbon backbone
Glutamate donated amine group
c
Fraction of demand taken up
Spearman R2s
0.17
0.22
0.41
Sufficient uptake
Ser Gly Group
d
Glycine secretion (mmol/gDW/hr)
R² = 0.35
Serine excess uptake (mmol/gDW/hr)
Breast
CNS
Colon
Leukemia
Melanoma
NSCLC
Ovarian
Prostate
Renal
e
Glutamine-derived carbon backbone
Other backbone
Amine group
Primary uptake
HCO3, Asp, R5P
Gln
GLUN
3PG, Pyr, Oaa
Asn, NTPs
Glu NH4
GLUD
Akg Ser, Ala, Asp
Pro
Akg Arg
Akg NH4
AKGD
Primary carbon outlet
f
Fraction of protein demand taken up
Spearman R2s
0.29
0.22
0.03
0.01
Sufficient uptake
0.08
0.01
0.01
Sufficient uptake
Asn Arg Asp Pro
Gln Glu Ala
g
median -0.42
median 4.0
Amine demand fraction met, without Gln
Amine demand fraction met, with Gln
h
Glutamate secretion (mmol/gDW/hr)
R² = 0.46
Glutamine uptake (mmol/gDW/hr)
Breast
CNS
Colon
Leukemia
Melanoma
NSCLC
Ovarian
Prostate
Renal
i
Glutamate + Alanine secretion (mmol/gDW/hr)
R² = 0.56
Glutamine uptake (mmol/gDW/hr)
Breast
CNS
Colon
Leukemia
Melanoma
NSCLC
Ovarian
Prostate
Renal

Glutamine however was taken up on average 32 times more than its biosynthetic requirement, consistent with the glutamine addiction phenotype (Figure 2.2f). To determine whether this uptake is tied to the biosynthetic requirement for nitrogen groups, we looked at the amine balance with and without glutamine uptake considered (Figure 2.2g). It was found that the cell is in a nitrogen deficient state without considering glutamine, with a median deficient of 42% of the demand. However, after considering glutamine uptake, nitrogen is in excess by a median of 400% of the biosynthetic requirement for amino acids. This suggests that glutamine uptake is a necessary feature of glutamine addiction, but is not sufficient alone to explain the very high level of glutamine uptake in these cell lines. We also observed overflow metabolism coupled to glutamine uptake, where glutamate secretion is significantly correlated with glutamine uptake (Figure 2.2h). Consideration of alanine secretion as well improves this correlation (Figure 2.2i), although it is not individually correlated with glutamine uptake. We then focused on energy metabolism in the form of ATP production and consumption (Figure 2.3a,b). We calculated that ATP is produced in similar parts between net glycolysis (mean 44%), ATP synthase (mean 35%), and succinyl-CoA synthetase (mean 21%), the latter of which produces GTP but is considered equivalent for this analysis (Figure 2.3a). Net glycolysis is defined as glycolytic production (phosphoglycerate kinase and pyruvate kinase) minus glycolytic consumption (hexokinase and phosphofructokinase). This split of ATP production is consistent with previous estimates[83], which place glycolysis at a slightly lower fraction of ATP production but agree that oxidative metabolism provides the majority of ATP. Looking at ATP demands, it is apparent that macromolecule and precursor synthesis only make up a minor part of the cellular ATP demand. A larger fraction is consumed by maintenance functions, such as the Na-K pump, or by undetermined functions that may consist of variable additional maintenance costs or energy drains such as futile cycles.

We find that net glycolytic ATP production is set at a level that mostly (80%) satisfies the known growth and maintenance ATP requirements of the cell, with little apparent tissue-specific bias (Figure 2.3c). However, due to the significant contribution of oxidative metabolism to ATP production, cell lines produce

an excess of ATP compared with known ATP demands[42] that scales with glucose uptake (Figure 2.3d). Additionally, there is an excess of redox production, both NADH and NADPH, over growth and energy demands that scales with glucose uptake as well (Figure 2.3e). Thus, although glycolysis alone balances the growth requirements for ATP, the overall energy balance is rapidly exceeded by ATP and NADH due to the contributions from oxidative phosphorylation in a manner scaling with glucose uptake per gram mass of cell. We hypothesize that a partial inability to metabolically process this excess of NADH production may be partly responsible for the overflow metabolism observed in the Warburg effect, where metabolized glucose is secreted as lactate facilitating the regeneration of NAD (Figure 2.3f).

Figure 2.3: **ATP production and utilization as determined by the Warburg effect.** a) Fraction of ATP production provided by glycolysis, the mitochondrial ATP synthase, and succinyl-CoA synthetase. It is observed that glycolysis provides on average 60% of total ATP production, in line with previous measurements. b) Fraction of ATP consumption by various cellular processes. c) Fraction of cellular ATP demand satisfied by glycolysis. On average, 80% of total ATP demand is provided by glycolysis. d) Comparison of glucose uptake with excess total ATP production. Although glycolysis satisfies 80% of cellular ATP demand, it produces only 60% of total ATP production, resulting in an excess of cellular ATP above growth and maintenance costs. e) Comparison of glucose uptake with excess total NADH+NADPH production. As with ATP, the high rate of glycolysis results in excess NADH+NADPH production above what is required for biosynthesis. f) Comparison of lactate secretion with the glucose uptake rate.


a
Glycolysis (PGK+PK)
ATPS
SUCOAS
Number of cell lines
Fraction of total ATP production
b
Glycolysis (HK+PFK)
Maintenance
Macromolecule Synthesis
Precursor Synthesis
Apparent Excess
Number of cell lines
Fraction of total ATP consumption
c
ATP demand satisfied
Average
Fraction of ATP-demand met by glycolysis
Breast (5)
CNS (6)
Colon (7)
Leukemia (6)
Melanoma (9)
NSCLC (9)
Ovarian (7)
Prostate (2)
Renal (8)
d
R² = 0.71
Excess ATP total (mmol/gDW/hr)
Glucose uptake (mmol/gDW/hr)
Breast
CNS
Colon
Leukemia
Melanoma
NSCLC
Ovarian
Prostate
Renal
e
R² = 0.67
Excess NADH+NADPH total (mmol/gDW/hr)
Glucose uptake (mmol/gDW/hr)
f
R² = 0.54
Lactate secretion (mmol/gDW/hr)
Glucose uptake (mmol/gDW/hr)

Next, we examined the role of the redox produced by the pathways associated with the high uptake metabolites glucose, glutamine, and serine. We found that growth-associated NADPH demands are dominated by fatty acid and steroid metabolism (Figure 2.4a), and are met by significant contributions to the cytosolic NADPH pool by the pentose phosphate pathway (66%) and folate metabolism (34%), which is in the range of previous estimates[43] (Figure 2.4b). Notably, a large majority of this NADPH production (mean 86%) originates from glucose. Thus, cytosolic NADPH production pathways meet the growth requirement of NADPH for the cell and is powered by glucose uptake.

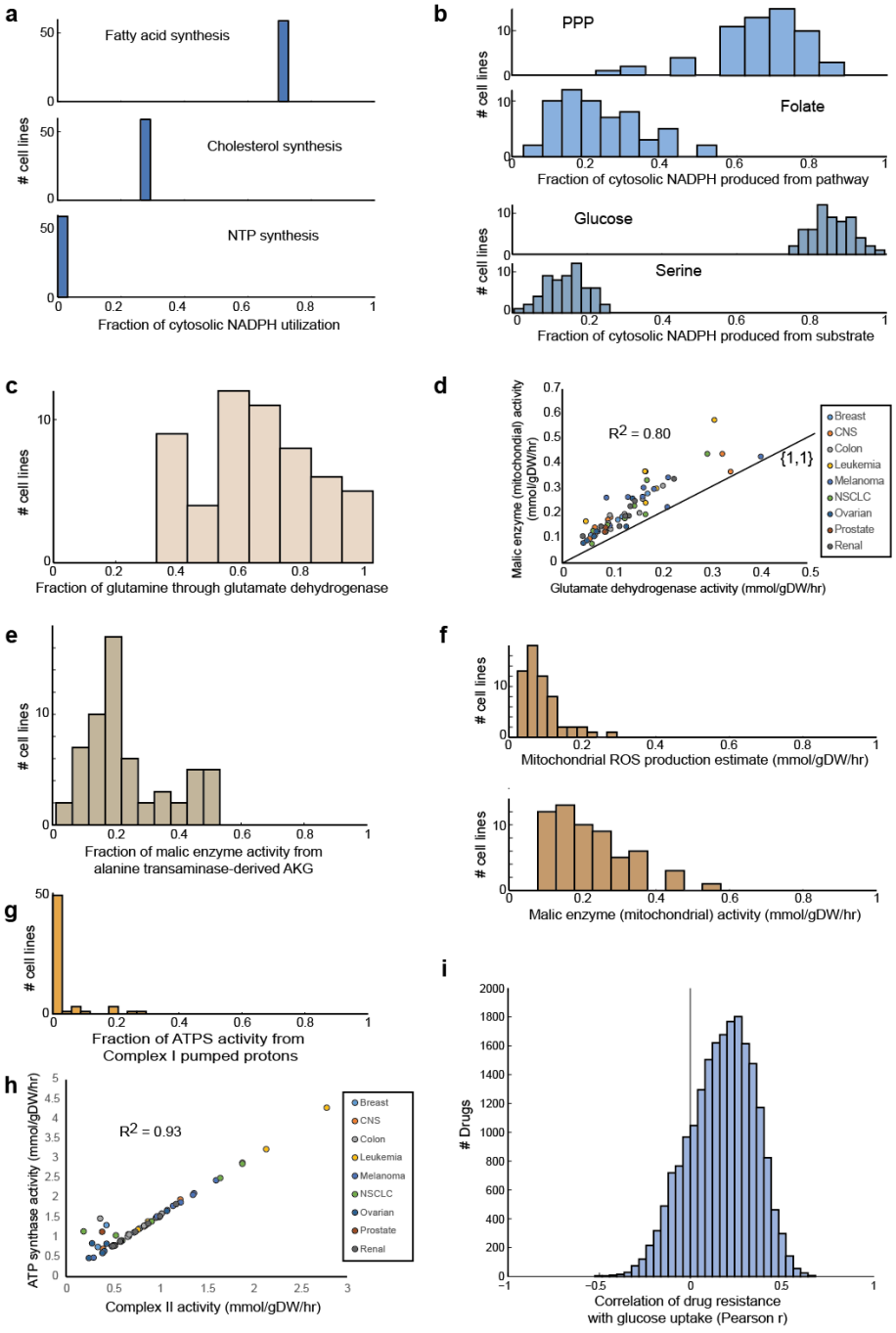
We then investigated the mitochondrial production of NADPH. We found that glutamate dehydrogenase (GLUD) metabolizes an average 70% of glutamine taken up (Figure 2.4c). However, alpha-ketoglutarate (AKG) produced from this reaction greatly exceeds the requirement for TCA metabolites and cannot be fully oxidized by the TCA cycle. To be fully catabolized, AKG produced from GLUD must exit the TCA cycle via malic enzyme (ME) to be converted into pyruvate for later oxidation (Figure 2.4d). This pathway produces NADPH through GLUD and ME in a manner that scales with glutamine uptake.

Simultaneously, an alternate pathway to glutamine catabolism through alanine transaminase exists that results in secretion of alanine, as observed (Figure 2.4e). These pathways can alternately be powered by glucose by producing oxaloacetate via pyruvate carboxylation, but this pathway is consistently measured to be minor. Because these pathways are quantitatively significant and tied specifically to glutamine uptake, we investigated whether the operation of these mitochondrial NADPH pathways might be a driving factor underlying glutamine addiction. We hypothesized that the function of mitochondrial NADPH production may be the mitigation of reactive oxygen species (ROS) stress. To estimate this potential demand, we identified mitochondrial enzymes previously shown to produce ROS through side reactions, and estimated the production of ROS as 2% of the flux through the enzyme. This analysis showed that the estimate of mitochondrial ROS production is on the same order of magnitude as glutamine fueled NADPH production through malic enzyme (Figure 2.4f). Although these estimations are

rough, this analysis provides some quantitative support that mitigation of mitochondrial ROS may be a driving factor underlying glutamine addiction, as has been suggested.

Finally, we sought to further understand the excess redox production tied to glucose uptake we identified previously. Surprisingly, we find that ATP synthase requires almost no NADH-powered proton pumping through Complex I (Figure 2.4g) to satisfy remaining energy demands not met by glycolysis. Instead, ATP synthase appears almost entirely powered by Complex II, due to the direct coupling of succinate dehydrogenase to Q10 in the membrane. We showed previously that O2 consumption rates are correctly predicted with primarily Complex II drive ATP synthase. However, if all NADH produced were fully oxidized through the electron transport chain, oxygen uptakes would greatly exceed measured values. Thus, we hypothesize that there is another NADH sink yet identified, which does not consume oxygen. Seeking to identify some relationship through which to understand this excess redox production, we correlated the glucose uptake with a large set of available data on the NCI60 lines, including gene expression, drug response, metabolomics, and proteomics. Remarkably, we found that higher glucose uptake broadly correlated with resistance to chemotherapeutic compounds in the NCI Developmental Therapeutics Program database (Figure 2.4i). This correlation was independent of the previously identified association between drug sensitivity and cell growth rate. Investigating the top correlated drugs, we identified a number of drugs that had previously been shown to induce apoptosis via an oxidative stress-dependent mechanism (Supplementary Data 1). Notably, imatinib resistance is among the highest correlated to glucose uptake, and imatinib resistance has previously been shown to be associated with higher glucose uptake[44]. Thus, it appears plausible that the excess redox production by high glucose uptake lines provides a measure of resistance to metabolic stress, such as those imposed by chemotherapeutics.

Figure 2.4: **NADPH balance and electron transport chain activity suggest role of excess glucose and glutamine uptake in metabolic stress resistance.** a) Cytosolic NADPH utilization, showing dominance of fatty acid and steroid synthesis. b) Cytosolic production of NADPH. The pentose phosphate pathway and folate pathway are both significant contributors, with the pentose phosphate pathway predicted to produce an average of 66% of cytosolic NADPH. However, an even higher fraction of cytosolic NADPH (86%) is glucose-derived rather than serine derived, due to additional NADPH from the glucose-driven de novo serine synthesis pathway. c) Fraction of glutamine uptake metabolized through glutamate dehydrogenase, with an average of 70%. d) Correlation between glutamate dehydrogenase and the mitochondrial malic enzyme demonstrates that glutamine is primarily converted to pyruvate through malic enzyme, producing mitochondrial NADPH. e) Fraction of malic enzyme flux due to alpha-ketoglutarate produced from alanine transaminase tied to alanine overflow, highlighting an alternate glutamine catabolic route. f) Comparison of mitochondrial malic enzyme flux to an estimation of reactive oxygen species production using a relationship of 2% of the flux of known ROS-producing reactions going to ROS-production[86]. g) Activity of mitochondrial complex I in the NCI60 cell lines. Although an excess of NADH is produced by glycolysis, ATP production and O2 consumption rates are not consistent with NADH oxidation by the ETC. h) Complex II activity vs ATP synthase activity shows that ATP synthase is predicted to be primarily driven by Q10H2 produced by complex II, rather than NADH-driven complex I. i) Correlation between resistance to chemotherapeutics and glucose uptake. Glucose uptake correlates highly with excess energy and redox production, and simultaneously correlates broadly with resistance to chemotherapeutic drugs in a manner independent of cellular growth rate.


a
Fatty acid synthesis
Cholesterol synthesis
NTP synthesis
cell lines
Fraction of cytosolic NADPH utilization
b
PPP
Folate
Fraction of cytosolic NADPH produced from pathway
Glucose
Serine
Fraction of cytosolic NADPH produced from substrate
c
Fraction of glutamine through glutamate dehydrogenase
d
R2 = 0.80
{1,1}
Malic enzyme (mitochondrial) activity (mmol/gDW/hr)
Glutamate dehydrogenase activity (mmol/gDW/hr)
Breast
CNS
Colon
Leukemia
Melanoma
NSCLC
Ovarian
Prostate
Renal
e
Fraction of malic enzyme activity from alanine transaminase-derived AKG
f
Mitochondrial ROS production estimate (mmol/gDW/hr)
Malic enzyme (mitochondrial) activity (mmol/gDW/hr)
g
Fraction of ATPS activity from Complex I pumped protons
h
R2 = 0.93
ATP synthase activity (mmol/gDW/hr)
Complex II activity (mmol/gDW/hr)
i
Drugs
Correlation of drug resistance with glucose uptake (Pearson r)

2.4 Discussion

The results of this work suggest a fundamental structure to cancer cell metabolism where biosynthetic demands determine much of the metabolic phenotype, but glucose and glutamine uptake scale above this requirement to result in an excess of redox production over biosynthetic and energy requirements. These results suggest an interpretation of the Warburg effect and glutamine addiction as motivated by metabolic stress resistance. Using the data available on the NCI60, we examined a number of other hypotheses for driving factors underlying the Warburg effect and glutamine addiction, notably the production of biomass precursors, the minimization of metabolic enzymes, and pH regulation (Supplementary Data 1). However, available data suggested these concepts are not quantitatively sufficient to explain the Warburg effect or glutamine addiction. Moving forward, further elucidation of the fate of the excess redox and energy produced in cancer cell lines should provide a better fundamental understanding of cancer metabolic hallmarks and fuel efforts at new therapeutic strategies.

2.5 Methods

2.5.1 Calculation of metabolic flux states

To calculate metabolic flux states, a stoichiometric model of central and growth metabolism was extracted from the latest global human metabolic network reconstruction, Recon 2[41]. This model was constrained by growth dataOconnor1997CR, metabolic uptake and secretion profiles[35], and an estimate of cell mass based on sustainable biomass that was validated against measured cell sizes[37]. ATP costs due to cellular maintenance functions was set to be 1.07 mmol/gDW/hr based on measurements[84]. Twenty-three metabolites had uptakes constrained by measured data, consisting of glucose, lactate, and amino acids, which accounted for greater than 99% of observed metabolic exchange fluxes in the NCI60 cell line data set. Furthermore, two flux splits, that between glycolysis and the pentose phosphate pathway, and between pyruvate dehydrogenase and pyru-

vate carboxylase, were constrained to be 90/10 each, based on previous ^{13}C tracing studies performed on the NCI60 lines. Metabolic flux states were then calculated by solving a flux balance analysis problem. A number of objectives were evaluated, including maximization and minimization of ATP production, maximization and minimization of cellular redox production, and minimization of overall cellular flux, a proxy for minimizing the proteomic cost of enzyme synthesis. We found that maximization of mitochondrial NAD(P)H gave the best agreement with ^{13}C tracing data thereby verifying the solutions on a subset of the NCI60 lines.

2.5.2 Constructing a core cancer model

To construct a core model, we included reactions necessary for biomass formation and primary metabolite catabolic pathways, while excluding anabolic pathways not associated with core biomass precursor production as well as secondary catabolic pathways. The primary distinction to be made was thus determining which reactions belong to primary and secondary catabolic pathways. This classification was performed algorithmically with manual justification based on 1) literature evidence and 2) feasibility based on modeling results. The following pathways were included: 1) pathways required for synthesis of biomass precursors, 2) core energy metabolism (glycolysis, the pentose phosphate pathway, the TCA cycle, the electron transport chain, and the malate aspartate shuttle), 3) pathways previously shown to be active in cancer cell lines under normal conditions 4) cofactor transfer reactions for NTPs and redox groups, 5) pathways involved with essential and/or high uptake/secretion metabolites, 6) pathways involved in catabolism of essential and/or high uptake/secretion metabolites, 7) pathways involved in cofactor regeneration and small metabolite processing/transport for cofactors/small metabolites produced in pathways gathered from the previous steps. Pathways that were excluded as a result of these criteria tend to have the following properties: 1) pathways that have unknown activity but due to small flux can be thought of as noise in approximations (e.g. glycosylation, EAA anabolic pathways, beta-alanine), 2) pathways removed because not measured to be major contributor to catabolism of high uptake metabolites (GLUDC, SERHL and methylglyoxyl), 3)

pathways that use high uptake metabolites only as cofactors (e.g. cysteine production from serine, where cysteine was not measured). Based on this approach, a core metabolic network of 382 reactions (Supplementary Data 1) was extract from the latest human metabolic network reconstruction[41]. Gene[45] and protein expression analysis[46] showed that this core set of reactions is highly conserved between cell lines, with the majority of reactions being highly expressed in all NCI60 lines. The core metabolic network was used to account for observed physiological functions, i.e. growth and measured uptakes, and minimize the influence of pathways with an unknown functional status on the networks flux state.

2.5.3 Cell-specific biomass determination

Cell biomass is composed of protein, lipids, DNA, RNA and small molecules, in weight fractions determined by cell composition studies. Average protein amino acid composition was taken from literature[47, 48]. Approximate DNA deoxyribonucleotide composition was set based on genomic base frequency taking into account the karyotype of the NCI60 lines[49]. RNA ribonucleotide composition was determined based on measured mass fractions. Lipid composition was set based on measured lipid composition for high concentration lipids. Small molecule weight fractions were determined for several high concentration non-essential metabolites using literature concentrations, using a typical cell dry weight of 0.2 ng/cell and cell volume of 2 pL/cell when unit conversions were necessary. We chose to set the macromolecule weight fractions to be constant between lines. Previous studies show minimal variance between macromolecule weight fractions for particular types of cells, such as hybridoma cells and Chinese hamster ovary cells. Other cell types, such as liver cells, may have significantly different macromolecule weight fractions, but cell lines derived from such tissues are not present in the NCI60 panel. Also, although cell composition has also been reported to change across growth conditions[50], the NCI60 panel was subject to uniform growth conditions in the studies generating the data used in this study. Furthermore, there is the question of whether cell composition changes with cell size. One study showed that doubling of cell size resulted in approximate doubling of respiration, suggesting the

protein content scales proportionally to size[51]. Also, as volume changes, the cell surface area (SA) to volume (V) ratio changes, and thus it is possible that the lipid weight fraction of the cell changes as well. However, compartment size has been shown to be approximately linearly correlated with total volume[52], and ER membrane alone is reported to be over 10 times the fraction of the total membrane as the cytoplasmic membrane[53], suggesting SA/V differences mean little in terms of lipid requirements. Thus, we assumed that the macromolecule composition was invariant across cell lines, although cell sizes differ. Protein content and cell volume data for the NCI60 was recently published. However, this data was insufficient to set cell-specific biomass macromolecule weight fractions, as the cell dry weight was not measured. To determine the cell-specific dry weights, we integrated cell volume data with the uptake rates as follows. First, the amount of biomass sustainable by each cell was determined by maximizing the growth through each line using FBA while constrained by measured uptakes in per cell units. Then, this sustainable biomass was corrected using measured protein content data as follows. If the sustainable protein, taking protein as 0.70 of total cell dry mass, is less than the measured protein, a value of 95% of the sustainable protein measurement was used as the estimate of cellular protein. This was done because the measured protein could not be sustained by the measured uptake rates, which we assumed was due to error in the measured protein. Measured protein was assumed to be the greater source of error because the measured uptake rates are highly correlated and there was no general bias of sustainable protein being greater or less than measured protein. Also the measured protein showed a relatively low agreement with cell volumes (Pearson $R^2 = 0.23$) and we observed certain spurious data points causing concern. For example, the SR line was reported to have a protein content of 0.021 ng/cell, which given the reported cell volume and average protein density would result in a dry weight fraction of protein of approximately 0.08, which is substantially lower than measured values around 0.7. Volume measurements were based in microscopy, and thus were seen as less error prone than protein content measurements which require cell count estimation, which can be a significant source of error. When sustainable protein was greater than measured protein, the measured

value was used to correct the sustainable protein, using the formula mestimate = mmeasured + 0.25*(msustainable-mmeasured). This formula was chosen based on resulting agreement with cell volume data. The correlation of estimated protein content with cell volume (Pearson $R^2 = 0.60$) was higher than either measured protein (Pearson $R^2 = 0.23$) or sustainable protein (Pearson $R^2 = 0.52$).

2.5.4 Curation of exometabolomic data

Published metabolite uptake and release (exometabolomic) data on the NCI60 lines was re-processed in a semi-automated process. The original dataset was processed by correcting for drift in the peak area standardization across runs by a linear L1 regression of blank media standards. However, upon detailed inspection of the drift for different metabolites, it was apparent that the drift was highly non-linear for some metabolites. The effect of applying a linear approximation in these non-linear cases was that metabolite uptakes were significantly mis-represented, and in some cases, metabolites were actually not exchanged substantially at all, once a non-linear drift correction was applied. We manually created non-linear approximations of drift for each metabolite in Mathematica based on the media standards for each metabolite. We applied these non-linear corrections to the raw data to recalculate the metabolite uptake and release profiles. The non-linear corrections used are available in the Supplementary Data 1.

2.6 Acknowledgements

We thank Joshua Lerman, Edward OBrien, Nikolaus Sonnenschein, and Nathan Lewis for helpful discussions. This work was supported by NIH grant GM068837.

Chapter 2 in part is a reprint of the material *Zielinski DC, Jamshidi N, Corbett AJ, Bordbar A, Thomas A, Palsson BO. Stoichiometric biomass synthetic requirements and metabolic stress resistance underlie hallmarks of cancer cell metabolism. In preparation* The dissertation author was the primary author.

Chapter 3

A data driven workflow for the construction of bottom-up kinetic models of metabolism

3.1 Abstract

Genome-scale metabolic network reconstructions provide a context within which omics data can be analyzed to understand phenotypic functions. Here, we develop a workflow to mechanistically integrate three disparate data types (metabolomic, fluxomic, and thermodynamic) within the context of a metabolic reconstruction. First we determine that the data sets are thermodynamically consistent. Then, we use the mass action stoichiometric simulation (MASS) model-building framework to develop a kinetic model *E. coli* core metabolism. We expanded this framework to include a database-driven semi-automated software package that enables the parameterization of mechanistic mass action modules from available kinetic data. The results from this study demonstrate that the MASS approach can generate network-scale dynamic models in a data-driven manner. The impending onslaught of quantitative *in vivo* metabolomics data can thus be converted into useful dynamic models in a data-driven fashion to generate descriptions of integrated network functions.

3.2 Introduction

In the past several years has witnessed a rapid development of new methods for metabolomics that are providing here-to-fore unmatched resolution of data describing the composition and dynamics of the cellular metabolome under many physiological conditions[46, 48]. As metabolomics data-generating methods continue to progress, so too must the development of in silico methods capable of analyzing this data and placing it in the context of what is known about cellular metabolic functions to interpret and understand the biological significance of the observations[49]. However, understanding the link between the concentration state of specific metabolites and cellular function is not a simple task. Metabolites typically participate in multiple reactions that may not be classifiable into canonical functional pathways due to the complexity of network structure. Metabolites also serve as signaling molecules, allosteric inhibitors, osmotic regulators, and other functions that further complicate the interpretation of metabolomic data. The role that a particular metabolite serves may also be context-dependent and change with environmental shifts, regulatory changes, and disease state.

It is desirable to obtain a quantitative understanding of the changes that occur in metabolites under various conditions and perturbations. The prediction of the dynamics of metabolite concentrations has historically been approached using computational modeling. Kinetic modeling of metabolism has classically consisted of defining reaction mechanisms, allosteric modifiers, and apparent constants for particular metabolic enzymes for pre-defined reaction conditions, measuring or approximating parameter values, and subsequently linking these modules to predict network function[54]. The applicability of these studies to *in vivo* processes is confounded by variation in numerical values for kinetic constants between *in vitro* and *in vivo* conditions, due to pH, concentration, molecular crowding, and viscosity differences, among other factors[55, 56, 57, 58]. Other efforts at constructing dynamic models focus on the reduction of networks based on quasi-equilibrium and quasi-steady-state assumptions in order to deal with the large number of parameters that can appear in kinetic models, but the application of these approximations for large systems becomes problematic. Genome-scale kinetic models have largely

been thought infeasible with current technology due to the large number of parameters to be defined and the difficulties in measuring such parameters.

As an alternative approach to fully defined kinetic models of metabolism, constraint-based modeling is a paradigm that has met with considerable success with practical applications at the genome scale. At the core of this methodology is the definition of the biochemically feasible system state space using parameter constraints based on mass balance and other principles[59, 60, 61, 62]. The space of network states can then be characterized in an unbiased fashion or searched for an optimal network state through the use of an assumed cellular objective in an optimization problem. These methods have been attractive due to that fact that little data beyond the network structure is required to make accurate predictions of phenotype.

It has been previously suggested that three generations of large-scale models will be developed with the third being the use of omics data to describe dynamic network states[63]. Advances in the field of metabolomics are beginning to provide the high-throughput *in vivo* data needed to develop network-scale dynamic models in a context-specific manner. An approach has been developed, termed the mass action stoichiometric simulation (MASS) approach, to construct kinetic models upon the stoichiometric models that classically have been used for constraint-based modeling. MASS modeling is a data-mapping approach that maps high-throughput in vivo concentration data, thermodynamic estimates, and calculated and in vivo flux data onto systems of mass action equations that are based off of validated metabolic network reconstructions. As the data mapped onto the networks is both high-throughput and specific to the measured state, MASS models allow a condition-specific analysis of in vivo dynamics at the genome scale.

The MASS modeling procedure has been described and there are metabolomics data sets now available that enable its meaningful deployment. Here we present a workflow for the development of dynamic MASS models for Escherichia coli[64]. In addition, we develop a powerful database-driven software package for the parameterization of mechanistic enzyme modules that can correct for a number of *in vitro*/*in vivo* differences. These goals represent an important

step towards developing a functional understanding metabolic system dynamics at a large scale.

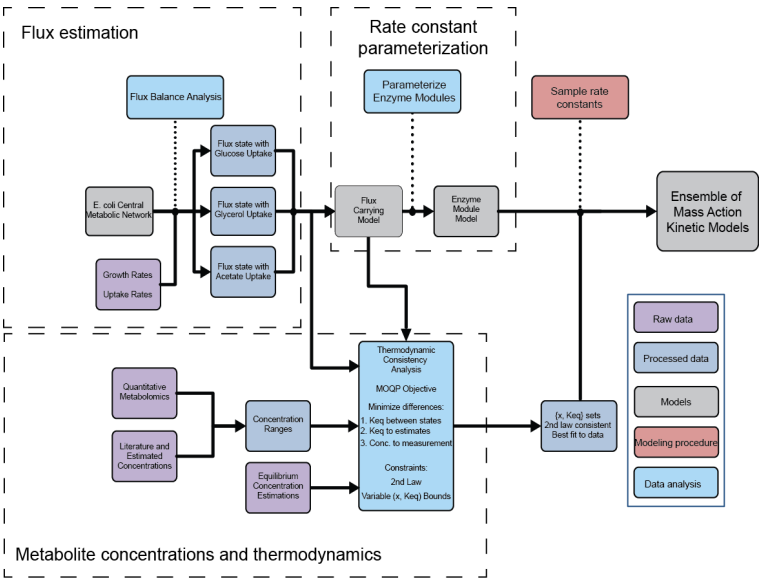


Figure 3.1: **Workflow for the bottom-up construction of data-driven kinetic models of metabolism.** Input data required are growth and uptake rates, metabolite concentrations, and equilibrium constant estimates, as well as enzyme kinetic data used in the parameterization of enzyme modules. Fluxes are calculated from flux balance analysis (FBA)[15] for all metabolic states for which concentration data was measured. Concentrations and equilibrium constants are checked for consistency with the 2^{nd} law of thermodynamics. Rate constants for mass action enzyme mechanisms are calculated from a specialized software package for parameterization. Concentrations, equilibrium constants, fluxes, and rate constants are then mapped onto the mass action enzyme module network to finish the parameterization of the kinetic model. Finally, alternate rate constant sets are sampled to account for uncertainty in rate constants, resulting in an ensemble of candidate kinetic models for subsequent simulation and analysis.

3.3 Results

3.3.1 Construction of a stoichiometric model of *E. coli* core metabolism

A core model of *E. coli* metabolism103, comprised of principally of glycolysis, the pentose phosphate pathway, the TCA cycle, the fermentation pathways, and the electron transport chain was used to analyze data generated from E. coli K-12 strain NCM3722 grown under three different carbon sources, glucose, glycerol, and acetate. The initial stoichiometric model contained 90 reaction and 71 metabolites. Flux balance analysis (FBA) was used to predict the steady-state flux distribution for the three growth conditions mentioned. The choice of constraints and objectives for the FBA optimization here is important for defining an accurate in vivo flux state, and these parameters are chosen to match the culture conditions that were used in acquiring the corresponding metabolomic data set[65, 66, 67, 68], (see Methods for details). This effort represents the first step of the workflow describing the process of creating data-driven MASS models of metabolism (Fig. 3.1).

Maximization of ATP was then used as the objective for FBA, subject to a growth constraint equal to the measured growth rate for each condition. The flux states calculated with this optimization were compared to available flux ratio, uptake rate, acetate secretion rate data from numerous studies and found to be in good agreement with measured data (Fig. 3.2)[65, 66, 67, 68]. Pearson correlation coefficients were calculated to provide an estimate for the difference between the flux states and were found to be 0.88, 0.62, and 0.56 for glucose/glycerol, glucose/acetate, and glycerol/acetate pairs, respectively. These differences are expected as acetate is the only gluconeogenic substrate of the three, eliciting the largest change for the latter two comparisons, while network states under all three carbon sources use the TCA cycle similarly, leading to an overall similarity in flux state. Reactions that were predicted to carry no flux in any of the conditions were removed from the model, and this reduced model, consisting of 75 reactions and 58 metabolites, was used for further analysis.

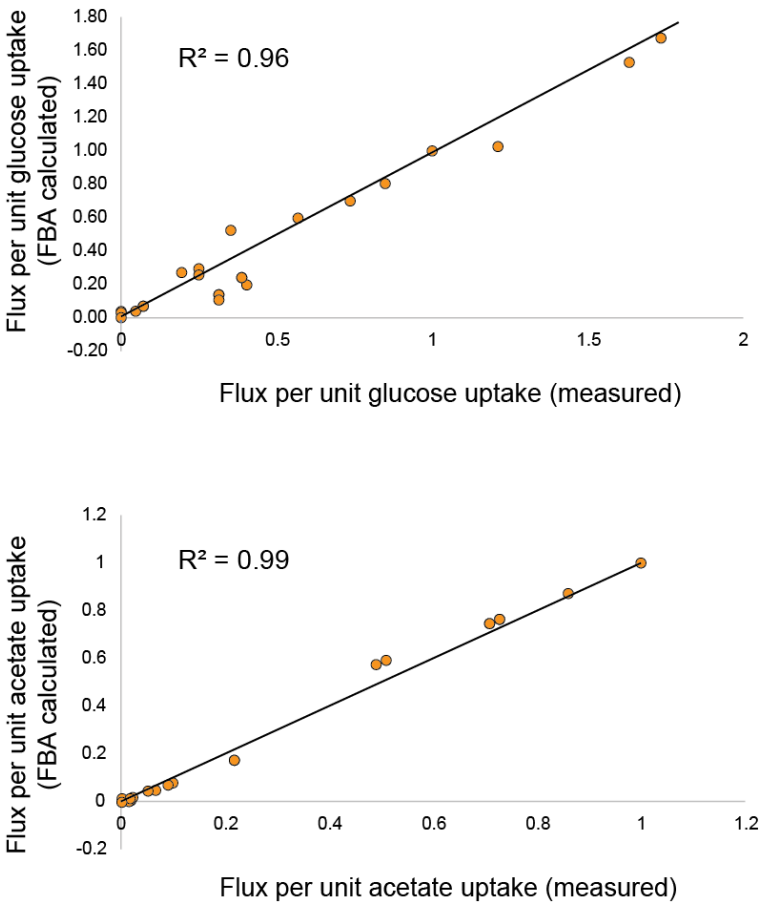


Figure 3.2: **Comparison of FBA calculated fluxes with ^{13}C labeled substrate data for glucose and acetate.** Data were taken from published studies [65, 66, 67, 68].

3.3.2 Multi-objective optimization approach to determining a thermodynamically consistent *in vivo* state

Thermodynamic constraints on metabolic network function have received much attention in the literature[69, 70, 71, 72]. We developed a constraint-based approach that integrated metabolomic measurements and calculated fluxes without violating thermodynamic principles. We sought to find a common set of equilibrium constants between the three experimental conditions, subject to the Second Law of Thermodynamics in each condition, while restricting metabolite concentrations to be within measured experimental error. This problem was addressed through formulation of a multi-objective optimization problem, simultaneously optimizing three objectives (Fig. 3.3a).

- The first objective is to minimize the distance between the equilibrium constants for each condition.
- The second objective is to minimize the distance between the equilibrium constants and the group contribution estimates.
- The third objective is to minimize the distance between concentration variables and the mean of their estimates from data (Panel B in Figure 3.3 illustrates this graphically).

A recent LC-MS/MS dataset provided quantitative metabolomic data for three nutrient growth conditions, covering 43% (25/58) of metabolites in the model[73]. The remaining metabolites largely consist of external and small metabolites, the concentrations of which can be reasonably estimated from literature and order of magnitude estimation. Equilibrium constants were back calculated from published group contribution estimates for Gibbs free energies assuming 1 mM concentration at pH 7.2113.

The results of the optimization (Fig. 3.3) show that the three disparate data types mapped onto the MASS model, fluxes, concentrations, and equilibrium constants, are largely thermodynamically consistent with each other across conditions, and the results remain close to the mean input values for the data. The

coefficient of variation of internal (non-exchange) reaction equilibrium constants between conditions was less than 10^{-4} for each reaction, showing that a common thermodynamic parameter set can be found that is consistent with the second law of thermodynamics for all conditions and for the variable bounds given. The agreement between the calculated consistent equilibrium concentration set and the group contribution estimates for internal reactions was good (median error $<$ 75%), but with a few significant outliers due to modeling assumptions (max error $\sim 10^6$%). The main source of outliers was transport reactions, which did not take into account electrochemical potential or pH gradient in the current mass action formulation and thus are subject to significant error. For example, the largest error occurred in the prediction for the ATP synthase reaction, which involves a proton pump that would not be correctly modeled, as a proton gradient was not taken into account.

When disregarding transport reactions, the error for the remaining reactions is decreased by an order of magnitude (median error $<$ 2.5%). A few notable internal, non-transport, equilibrium constants were predicted by flux and concentration data to be required to be significantly different than group contribution methods estimate. Reactions with calculated equilibrium constants at least an order of magnitude away from their group contribution estimates were enolase, triose phosphate isomerase, fructose bisphosphate aldolase, fumarase, GAP dehydrogenase, phosphoglucokinase, succinyl-CoA synthetase, and G3PD2, indicating that the *in vivo* thermodynamics for these reactions might not be accurately predicted computationally, although error in concentration estimates cannot be discounted. It is interesting to note that several of these enzymes occur in lower glycolysis, which is thought to be near equilibrium. The discrepancy between group contribution estimates and predictions from data for these reactions gives possible indication of a digression in the cellular microenvironment from the assumptions underlying solution thermodynamics, perhaps suggestive of substrate channeling61, which would have the effect of increasing the local concentration of metabolites and therefore alter the effective thermodynamics of the reaction from our calculations.

The error between calculated concentrations and reported data means necessary to create a common Keq set was also reasonable (median error $< 105\%$ and standard deviation of error $< 100\%$ in all conditions). Taken together, these results suggest that the condition-specific *in vivo* concentration data and validated flux estimations are consistent with group contribution estimates for *in vivo* thermodynamic data. Changes in the metabolome between conditions are consistent with the changes that flux calculations and Gibbs free energy estimates predict to be required to satisfy the 2^{nd} law of thermodynamics, showing how governing constraints determine the quantitative metabolome in a way that can be computationally rationalized. Additionally, a consistent set of equilibrium constants was found that satisfied the second law for all conditions studied.

Figure 3.3: **Non-linear optimization identifies reaction equilibrium constants consistent with multiple *in vivo* measured metabolomics data sets.** a) Definition of the optimization problem to identify reaction equilibrium constants that are consistent with the 2nd Law of Thermodynamics as well. The problem is a multi-objective non-linear optimization to find concentration and equilibrium constant sets that are as close to measured or estimated values as possible. b) A schematic of the optimization problem, in the case of the published E. coli metabolomics data set grown aerobically on glucose, acetate, and glycerol[73]. c) Consistency of equilibrium constants found to be consistent with metabolomics compared with the computational estimation of the equilibrium constants from group contribution[71]. A number of outliers are apparent. ATP synthase appears due to the proton differential not being considered in the mass action ratio for the reaction, and thus the Keq calculated can be thought of as an apparent Keq. Several reactions in lower glycolysis appear, which suggests that metabolic concentrations are not consistent with observed fluxes in lower glycolysis. Lower glycolysis has been thought to be subject to metabolite channeling[74], which if occurring would create local concentrations of the metabolites greater than those calculated assuming homogeneous distribution within the cell, consistent with the observed thermodynamic discrepancy.

a

Problem Formulation:

$$\min \quad \alpha_1\left(\left\|K_{eq,Glucose}-K_{eq,Glycerol}\right\|+\left\|K_{eq,Glucose}-K_{eq,Acetate}\right\|+\left\|K_{eq,Glycerol}-K_{eq,Acetate}\right\|\right)+$$
$$\alpha_2\left(\left\|K_{eq,mean}-K_{eq,GCEstimate}\right\|\right)+$$
$$\alpha_3\left(\left\|x_{Glucose,var}-x_{Glucose,data}\right\|+\left\|x_{Glycerol,var}-x_{Glycerol,data}\right\|+\left\|x_{Acetate,var}-x_{Acetate,data}\right\|\right)$$

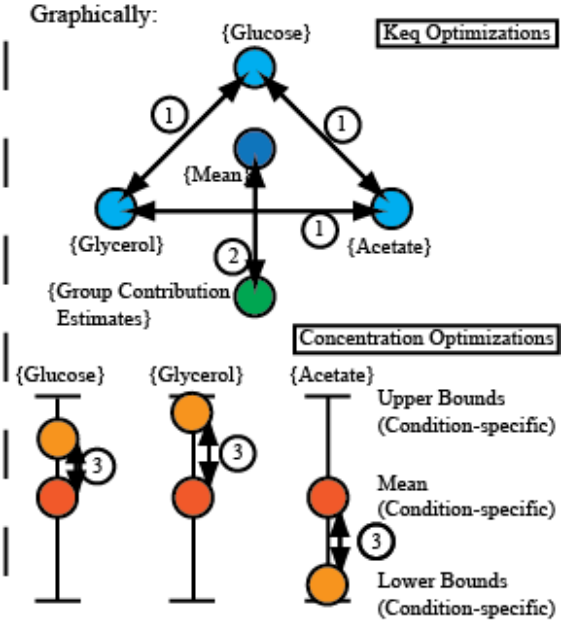
$$\text{s.t.} \quad \frac{\prod x_k^{s_{f,j}^k}}{\prod x_j^{s_{f,j}^*}} > K_{eq,f} \text{ for } v_f < 0 \text{ and } \frac{\prod x_k^{s_{f,j}^k}}{\prod x_j^{s_{f,j}^*}} < K_{eq,f} \text{ for } v_f > 0 \text{ and } \begin{matrix} K_{eq,lb,f} < K_{eq,f} < K_{eq,ub,f} \\ x_{lb,j} < x_j < x_{ub,j} \end{matrix}$$

b

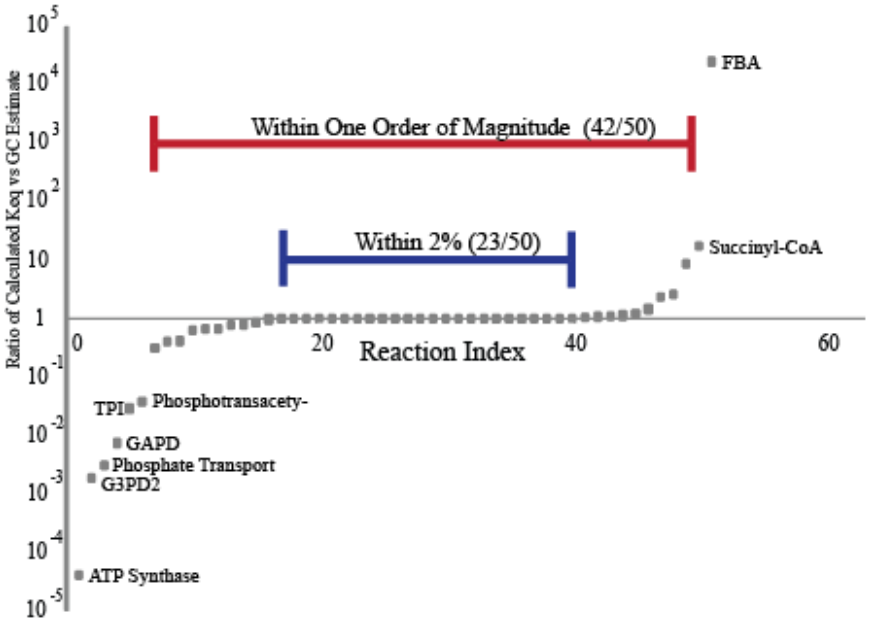
Goals:

1. Seek common set of Keq's
2. Mean of Keq's as close to estimates as possible
3. Concentration data as close to data means as possible

Graphically:



c



3.3.3 Construction of a software pipeline for the parameterization of mechanistic mass action kinetic modules

Thus far, we have determined that we can acquire consistent estimates of *in vivo* reaction fluxes, metabolite concentrations, and equilibrium constants. The last requirement for kinetic model specification is the rate constants of the enzyme mechanism. However, individual rate constants are typically not measured directly. Instead, the enzyme kinetic parameters measured are typically lump parameters from initial rate studies, such as the catalytic constant k_{cat}, the Michaelis constant K_M, and dissociation constants K_d. These lump parameters, other than the relatively simple K_d, are at best complicated non-linear expressions of the reaction elementary rate constants, and at worst do not have a comparable analytical expression that can be used to constrain selected rate constants. Furthermore, the experiments used to study enzyme kinetic parameters are often performed under non-physiological conditions, providing further cause for concern for their direct use in a kinetic model. Causing additional problems, these data have inherent error associated with measurement, and thus parameters may become inconsistent with each other, with little recourse. Finally, in practical situations the number of independent data points will be almost always be less than the number of parameters specified. Thus, some accounting for the uncertainty in free parameters should be done.

To address the challenges of parameterization of an elementary mass action enzyme module, we developed a novel pipeline to identify sets of elementary rate constants that reproduce lump enzyme kinetic parameters, accounting for uncertainty and many sources of *in vivo/ in vitro* differences. The overall strategy is to first calculate an algebraic expression that relates the elementary rate constants to the measured kinetic data and then to use a constrained, pseudo-randomized, non-linear least squares problem to find sets of rate constants that minimize the difference between the algebraic expression value and the measured data value. Importantly, these algebraic expressions can contain corrects for in vivo/in vitro differences by explicitly including models that represent these differences, such as ionic strength corrections to reaction equilibrium constants. Furthermore, by

pseudo-randomizing starting points in the non-linear squares problem, uncertainty and free parameters can be accounted for by repeated optimizations to generate equivalent parameter sets.

The resulting pipeline consists of 1) a database back end to store measured enzyme kinetic data, increasing automation and consistency, 2) powerful computer algebra systems to calculate expressions relating elementary rate constants to lump enzyme kinetic parameters for arbitrary enzyme reaction mechanisms, 3) a carefully optimized parameter fitting approach that solve a complicated constrained non-linear least squares problem to calculate the desired elementary rate constants.

3.3.4 Construction of a database to store and retrieve kinetic data

We first conducted a literature search for all available kinetic data on enzymes in the core *E. coli* model. Notably, much of this data was not found in the most popular enzyme kinetics database, the Braunschweig Enzyme Database (BRENDA)[75], indicating that prior collection efforts have not be fully comprehensive, and thus the availability of kinetic data may be higher than believed, for the model organism *E. coli* at least. Indeed, we found at least one kinetic data point for 54 of the 57 isozymes in the core kinetic model of *E. coli*. These data are provided in Supplementary Data 1.

3.3.5 Construction of a computer algebra pipeline to relate elementary mass action rate constants to measured kinetic parameters

Having collected the available data, the next challenge is to develop expressions that relate the rate constants to the measure kinetic data. Due to the fact that the vast majority of available data is derived from initial rate studies, we focus our discussion on handling the parameters k_{cat}, K_M, K_d, and K_{eq}. For arbitrary enzyme mechanisms, there is not guaranteed to be an expression directly relating k_{cat} or K_M to the elementary rate constants of the enzyme mechanism. However,

we can take advantage of the manner in which the k_{cat} and K_M are extracted from initial rate studies to provide these expressions. Both of these parameters first require solving for the overall steady-state rate of the enzyme from its elementary mass action mechanism. The overall steady-state is found by solving the mass balance equation $Sv = dx/dt = 0$ for the enzyme module along with the enzyme conservation equation E_{total} = sum of E_i for the individual enzyme forms E_i. This system of equations can be solved algorithmically using the King-Altman method, which is rooted in calculating determinants, or through the use of a computer algebra package such as provided by Mathematica. In practice, we have found the latter to be more efficient, but the King-Altman method provides a failsafe backup in cases where the computer algebra methods may fail. Then the expressions for E_i are substituted into the rate equation for the catalytic step of the reaction. The resulting expression is a function of the metabolite concentrations, the rate constants, and the total amount of enzyme E_{total}.

First, the K_M for a particular metabolite and enzyme is the concentration of the metabolite at which the enzyme operates at half of its maximal velocity. This mathematically means that the K_M value can be found by requiring that the quotient of the rate of the reaction to its maximal rate is 0.5 when the concentration of the metabolite equal to the K_M is substituted into the rate constant. In terms of the following non-linear least squares problem, this effectively requires a set of rate constants that causes the relative saturation of the enzyme to follow the typical Michaelis-Menten like saturation curve centered around the K_M value. To calculate the reaction quotient, the maximal velocity of the reaction must be found. To do this, we take advantage of powerful computer algebra tools offered by Mathematica, and algebraically computed the symbolic limit of the reaction rate as the metabolite concentration reaches infinity, subject to zero product concentration and a fixed co-substrate concentration. In practice, these limits have proven efficient to calculate. Thus, through algebraic manipulation of the mass action reaction rate laws, we can create a comparison equation relating elementary rate constants to measured K_M values.

In the case of the k_{cat}, it is universally defined as the constant relating the

v_{max} of the enzyme to the total enzyme concentration E_{total}. Thus, we can create a comparison equation of elementary rate constants by substituting an arbitrary amount of enzyme into the overall steady state rate equation, and set the concentrations equal to the measured saturating metabolite concentrations. The resulting equation will guarantee that the $v = v_{max}$ at the saturating metabolite concentrations, provided that it is coupled with K_M data that enforces that these metabolite concentrations are saturating. This equation thus relates the elementary rate constants to the measured k_{cat} values. Initially we examined the possibility of using the limit of the reaction rate as the metabolite concentrations become infinite, as was done in the creation of the K_M comparison equation. However, we discovered that this limit is poorly defined for multi-substrate reactions when the reaction exhibits a random order mechanism, because the final limit depends on the ratio of the substrates. For this reason, we decided to enforce the k_{cat} using the steady-state reaction rate coupled to assumed K_M parameters to enforce saturation, rather than using a limit-based approach. In practice, we did not find any cases of an enzyme having a measured k_{cat} without corresponding K_M data, so the need for K_M parameters with this approach did not cause any complications.

More simple to relate to elementary rate constants are dissociation constants and reaction equilibrium constants. Dissociation constants are simply defined as inverse equilibrium constants for metabolic binding steps. Thus, the corresponding comparison equations take the form of $K_d = k_r \,/\, k_f$ for a binding reaction E+I EI, where k_r and k_f are the reverse and forward elementary rate constants, respectively. Similarly, the overall reaction equilibrium constant K_{eq} is simply the multiplication of the equilibrium constants of sequential steps in the enzyme mechanism. As the individual K_{eq}s are defined in terms of elementary rate constants, the comparison equation is quite simple. Thus, using these methods we are able to construct algebraic expressions relating the individual rate constants to measured kinetic data. This will allow us to calculate sets of elementary rate constants that cause the mass action enzyme module to reproduce the measured kinetic behavior.

3.3.6 Sampling sets of equivalent rate constants satisfying measured kinetic data with a non-linear least squares approach

With the kinetic data and comparison equations now available, we then create a scoring function that consists of a least-squares problem with the elementary rate constants as variables, and the sum of squared errors between the kinetic data and comparison equations as the objective function to be minimized. This problem is constrained because the rate constants cannot be arbitrarily high. The diffusion limit on the second order rate constant k_{cat}/K_M is typically estimated around 10^9 s^{-1}, and this is the limit that we used as the upper bound on the rate constants. The lower bound was set at an arbitrarily low value of 10^{-6} s^{-1}, but it is unknown whether any rate constants could possibly reach slower rates. However, this wide range of rate constants creates a difficult problem for optimization algorithms. Also, the optimization problem is of course non-linear and thus will be subject to local minima, preventing a global optimum from being guaranteed to be found. The under-determined nature of the problem also suggests that many equivalently optimal solutions will exist. We addressed these difficulty by using a two stage optimization approach. The first stage uses a randomized non-derivative based particle swarm optimization with log-transformed rate constants as variables to efficiently scan over a large order of magnitude for local minima. This randomized first stage does not find perfect fits in practical situations, but instead finds decent fits that serve as starting points for the next stage optimization. The second stage optimization is a derivative-based Levenberg-Marquadt algorithm using linear variables that in practice quite efficiently perfects the solution based on starting points from the particle swarm algorithm. This two-stage optimization approach thus finds a set of equivalent rate constants with a high degree of fidelity, such that the majority of initial points optimized reach a perfect fit, where the sum of squared errors is very small. To minimize the number of repeated or highly similar rate constant sets, we then clustered these rate constants and extracted one rate constant set per cluster to serve as representative rate constant sets. The

appropriate number of clusters can be chosen by the user or suggested by a number of statistical methods. See Fig. 3.4 for an example fitting for adenylate kinase.

3.3.7 Estimation of enzyme concentrations and model simulation

Utilizing the collected *E. coli* data, we tested this workflow on a wide variety of mechanisms and data types and found it to enable robust and rapid enzyme module parameterization. With rate constants calculated, the only parameter is the total enzyme, which can thus be calculated from the enzyme conservation equation E_{total} = sum of E_i, substituting the previously found solutions of steady-state concentrations of enzyme forms as well as the calculated reaction fluxes, metabolite levels, reaction K_{eq}s, and rate constants. These enzyme levels can then be compared to quantitative data for validation. An alternate workflow when enzyme data is present and believed to be highly accurate would be to include this concentration in the least squares problem, along with the in vivo calculated fluxes and concentrations, to find a consistent system that satisfies all data.

With the total enzyme calculated, the kinetic model is now fully parameterized and can be simulated. Since multiple equivalent rate constant sets were found for each enzyme, we can now simulate an ensemble of models to account for uncertainty and undetermined parameters. This provides protection against the troubles of overfitting that have plagued the field of kinetic modeling.

The workflow developed is robust in application and allows the rapid development of thermodynamic and kinetically consistent mass action kinetic models. This approach should become very powerful in the development of practical kinetic models to address problems of importance in strain design and human health.

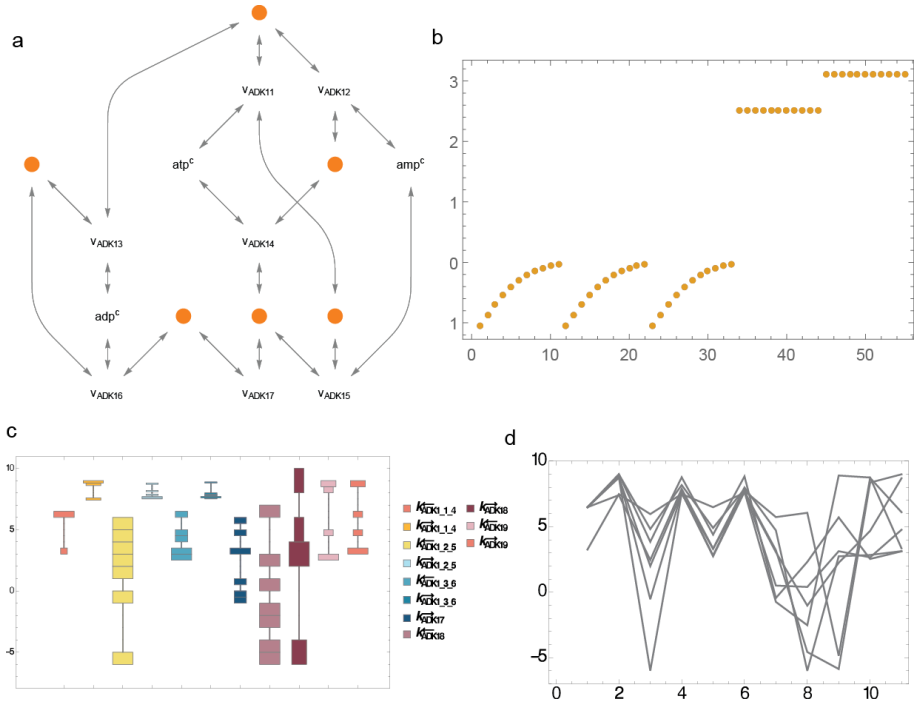


Figure 3.4: **Parameterization of enzyme module rate constants using a non-linear least squared approach.** a) The software fitting package automatically generates reaction maps based on the mechanism of the enzyme, in this case adenylate kinase. b) Best fit of the enzyme rate constants to enzyme kinetic parameters. In this case, three K_Ms, for atp, adp, and amp, and two k_{cat}s, for the forward and reverse reaction, are fitted. The k_{cat}s are printed multiple times for additional weighting in the optimization, an approach thats roughly equivalent to including a weighting parameter on the sum of squared error objective function. The y axis is in data-dependent units, while the x axis is the index of the data in the sum of squared errors term. c) Distribution plot of equivalent rate constant sets obtained by the two stage optimization algorithm. It is observed that wide ranges of rate constants are obtained, accounting for the fact that free variables exist while maintaining the non-linear constraints between rate constants imposed by data. The y axis is in units of s^{-1}, and it is observed that constraints on overall rate constants are obeyed to be between 10^9 and 10^{-6} seconds. d) Clustering of rate constants further shows the non-linear relationships between rate constants and demonstrates how characteristic rate constant sets can be chosen for later ensemble modeling.

3.4 Discussion

The work presented here seeks to functionally integrate quantitative metabolomic data with multiple data sources in the context of a metabolic network reconstruction. A workflow was generated that combines an optimization approach to obtain a thermodynamically-consistent data set, steady-state analysis to assess the value of the metabolome towards achieving a putative cellular objective, and the construction and analysis of condition-specific dynamic models using the MASS approach. The study shows that the MASS approach can generate network-scale dynamic models in a data-driven manner.

The MASS approach is a data mapping approach built upon the integration of multiple disparate data sets in the context of a network model. Through solving an optimization problem, we found that the condition-specific fluxomic and metabolomics data sets used were consistent with thermodynamic laws. The thermodynamic consistency of the data enables the successful integration of multiple experimental and experimentally validated data types within the framework of a stoichiometric model. It is notable that, despite the constraints given by the data error and flux directions, a common set of equilibrium constants was found for all conditions, and this common set was largely close to estimates from group contribution theory. It is noteworthy that in vivo equilibrium constants are estimated without a direct measurement but form high-throughput data mapping and consistency checking. As new metabolomic data sets for *E. coli* are generated for different conditions, this method can be repeatedly applied to further refine the values of equilibrium constants *in vivo*.

Omics datasets are typically analyzed by statistical methods based on qualitative and relative principles such as overrepresentation. This work takes an important step towards fully quantitative systems biology integrating quantitative data with functional models to make quantitative biological interpretation and prediction. By mapping quantitative omics data against the underlying systemic, chemical and thermodynamic structure, we are able to generate models that provide physiological insight and understanding. For example, by mapping the data onto a mass action kinetic scaffold, *in vivo* enzyme concentrations necessary to sus-

tain the observed flux can be calculated and compared to measured values. The enzyme saturation states calculated through this approach represent the functional state of the proteome, lending insight into systemic effects of enzyme saturation and allosteric regulation.

The scaling of kinetic modeling to the network level traditionally has been seen as troubled by the large number of parameters required. This study has shown that the integration of various data measurements and parameter estimation approaches in the context of a metabolic network reconstruction enables the construction of large-scale dynamic models. MASS models, due to their foundation in functional biochemistry and use of in vivo data, serve as an ideal starting point to begin to understand the complexity of in vivo biochemical system dynamics in a rigorous fashion. As the necessary quantitative data is generated, these models will continue to be developed for other systems, such as those relevant to disease, enabling a previously unattainable quantitative view into the physiological functions and role of metabolic dynamics in vivo.

3.5 Methods

3.5.1 Additional notes on calculating the condition-specific flux state

The oxygen and nutrient uptake conditions were set for each condition by first determining the minimum nutrient value predicted by the model to sustain the measured growth rate, then doing the same for literature data for growth on each substrate to estimate the excess nutrient uptake, then scaling the minimum value by this amount to estimate the actual uptake at the measured growth rate. This process was found to yield results largely consistent with available data[65, 66].

One discrepancy with in vivo data was that, for the acetate growth condition, the minimum oxygen uptake rate to support the measured growth rate and non-growth-associated ATP maintenance cost was calculated to be slightly higher than the previously established enzymatic capacity for oxygen incorporation into

the ETC. This is likely an artifact of the simplified ETC used in this stoichiometric model, which has a P/O ratio that is constrained to be an average of the maximum and minimum possible P/O ratio of the network as a simplifying measure and thus is lower than the capacity of the full *E. coli* metabolic network.

A recent study showed that the well-known phenomenon of acetate overflow is a result of catabolite repression, which results in a decrease of acetate intake[76]. Interestingly, the maximization of ATP for a constrained growth rate produces accurate estimates of acetate overflow[67]. This could indicate that the catabolite repression might act to maximize ATP production at a sub-optimal growth rate. However, this hypothesis is complicated by the fact that many flux states not involved in optimizing ATP could have similar acetate secretion.

Note that the inherent assumption in this model reduction is that the removed reactions remain inactive during perturbations, and also that reactions predicted to have zero flux in a particular condition but not in others remain inactive for that condition (i.e. the reaction rate, discussed later, is set to zero). This assumption is likely only valid near the steady-state studied, i.e. for small perturbations. This suggests a clear boundary on the types of questions that should be asked using such a model. For example, probing into the response of the system to a switch to different nutrient environments or after growth evolution would require the inclusion of additional reactions and likely regulatory mechanisms, as well as an approximation of the dynamics of these reactions. Accurately predicting dynamics of such complex processes is certainly an important goal of dynamic metabolic modeling and has been addressed previously elsewhere using even more course-grained modeling integrated with transcription regulation[77], but these types of predictions are out of the scope of the current study. The goal of the current study is to lay out a workflow for incorporating the necessary in vivo data into a large-scale kinetic model, and to show how these models can be used to compare the dynamic properties of the same network at different in vivo steady states.

Knowing a priori that the metabolomic data to be used was from non-growth-evolved cultures, the measured growth rates for these cultures was used as a minimal constraint, as opposed to an objective. To integrate flux data with

concentration data, fluxes were converted from mmol/gdW/hr to M/hr by assuming a cell volume 1.1 cubic microns and a cell weight of 150 pg, which were values obtained from the Bionumbers database.

3.5.2 Multi-objective optimization

To identify a set of parameters that is both consistent with the second law of thermodynamics and is as close as possible to the parameter estimates from various sources of data, a multi-objective non-linear programming problem was solved. The objective of the MONLP problem is a weighted sum of distances of the parameters from their estimated values. Equilibrium constants for exchange reactions were not included in the objective function as their values include the constant external metabolite concentrations, which are allowed to vary between conditions. The formulation of the problem is described in Figure 3.3, where 1, 2, and 3 are the weightings on the first, second, and third parts of the optimization. $K_{eq,GC\ Estimate}$ is the group contribution estimates for the equilibrium constants. xGlucose,data, etc, are the means of the concentration data used. S_p and S_n are matrices of the positive and negative stoichiometric coefficients. x_{pool} is the total concentration of concentrations that were indistinguishable by mass spectrometry.

The first set of constraints is the second law requirements that the PERCs be positive. This is equivalent to the statement that the flux must be going in the direction that is thermodynamically feasible as determined by the concentrations and equilibrium constant for that reaction. In other words, the difference between the mass action ratio and the equilibrium constant has to be the opposite sign as the flux for that reaction. As the equilibrium case leads to an undetermined PERC, the difference between the mass action ratio and the equilibrium constant was restricted to be greater or lesser than a specific tolerance, depending on the direction of flux. This tolerance defines the minimum distance of the reaction from equilibrium and thus effectively defines the upper limit on the rate of any reaction. The PERCs are potentially highly sensitive to the tolerance value, since, as mentioned above, states very close to the equilibrium state for a reaction result in very high PERCs, and thus slight changes in parameter values result in large

changes in PERCs. A sample of tolerances was tested, and a value was chosen that was large enough that the PERCs were no longer highly sensitive to the tolerance.

The second set of constraints is the requirements that the concentrations of metabolites with the same mass, and therefore could not be distinguished with the MS methods used, sum to the concentration of the measured corresponding pool. The first set of parameters is non-linear with linear variables, while the second set of parameters is linear in linear variables. As there are more constraints associated with the second law than with pools, the variables were transformed to log variables to make the second law constraints linear with log variables, while the pool constraints became non-linear with log variables. Note, if the concentrations of the metabolites within pools were measured separately, the entire problem could be transformed to linear constraints, a desirable trait for the scaling of this problem. For the purposes here, the presence of several non-linear constraints did not appear to affect the results. The remaining constraints are the variable bounds.

The choice of weightings on the three objectives is a practical consideration. The weightings were selected giving priority to the minimization of distances between equilibrium constants and the minimization of distances between the equilibrium constants and group contribution estimates. The requirement that the concentrations remain close to the mean is largely expected to be superficial, as any value within experimental error is equally valid. The underlying constraint given by the Second Law of Thermodynamics is manifested in the requirement that the chemical potential gradient and flux for each reaction are in the same direction116. While there may be many sets of parameters within constraints that satisfy the Second Law of Thermodynamics, we sought to find the set closest to the concentration data mean and group contribution estimates to try to define a most likely in vivo state for further analysis.

3.6 Supplementary Data

3.6.1 Enzyme module fits

Acetate kinase (ACKr)

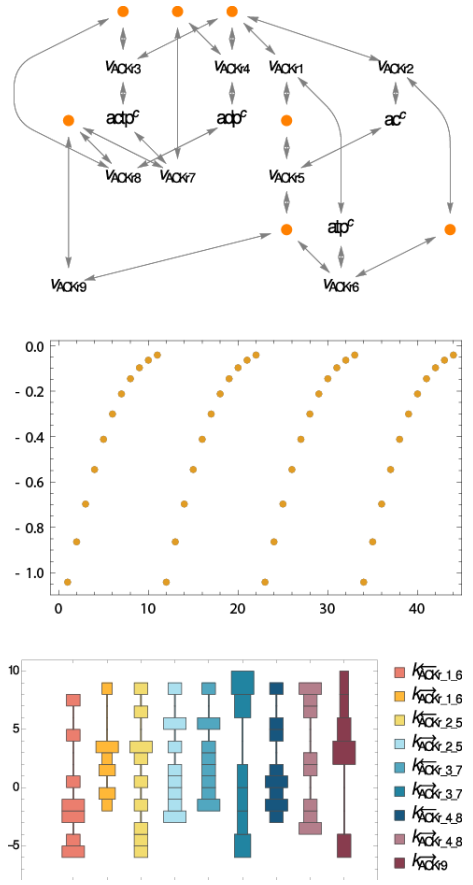


Figure 3.5: **Acetate kinase fitting results**

Aconitase A (ACONTa)

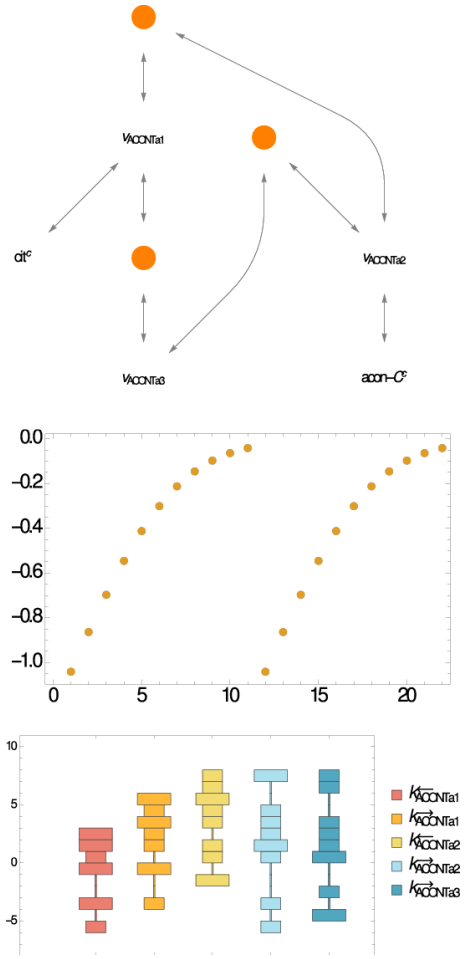


Figure 3.6: **Aconitase A fitting results**

Aconitase B (ACONTb)

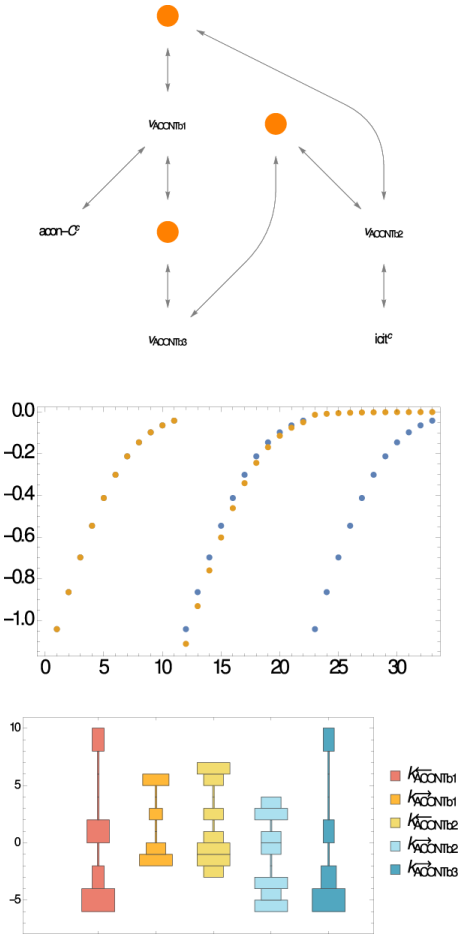


Figure 3.7: **Aconitase B fitting results**

ADK1 (ADK1)

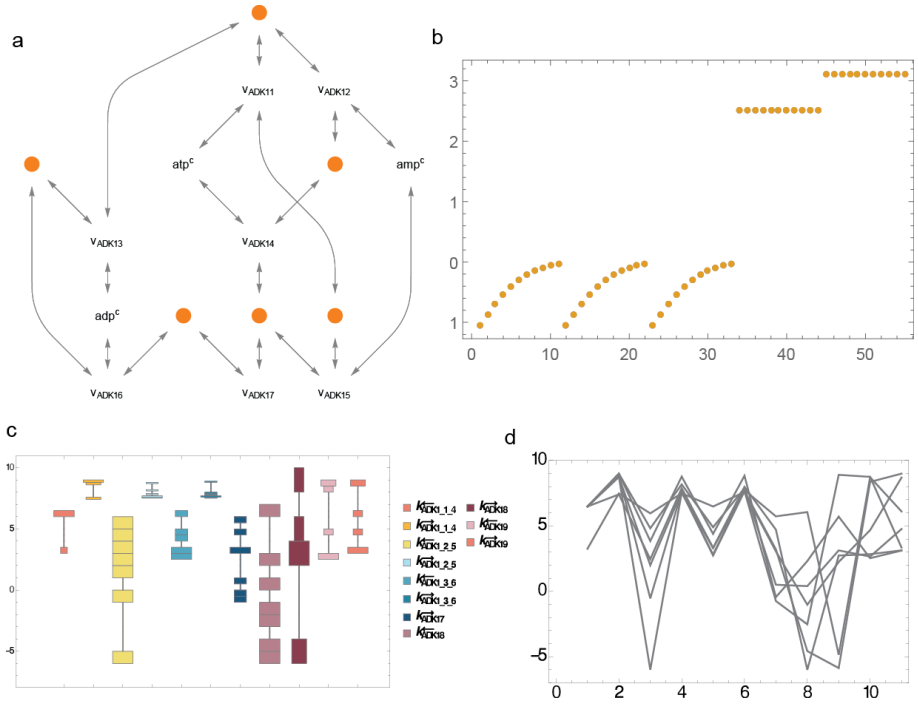


Figure 3.8: **Adenylate kinase fitting results**

Cytochrome *bd* Enzyme 1 (CYTBDpp)

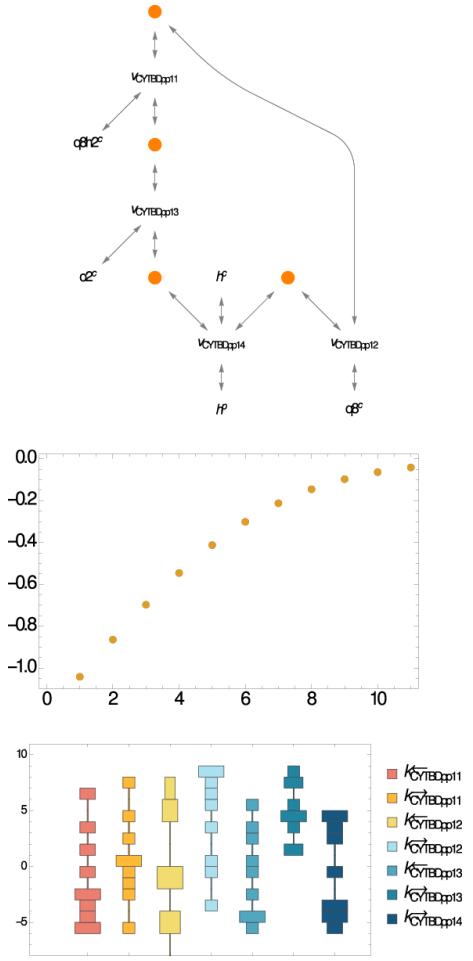


Figure 3.9: **Cytochrome bd 1 fitting results**

Enolase (ENO)

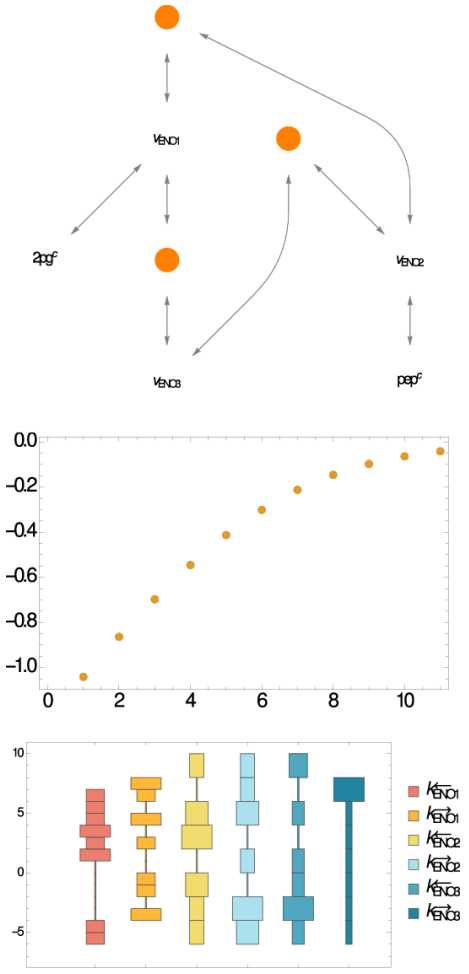


Figure 3.10: **Enolase fitting results**

Fumarase A (FUM)

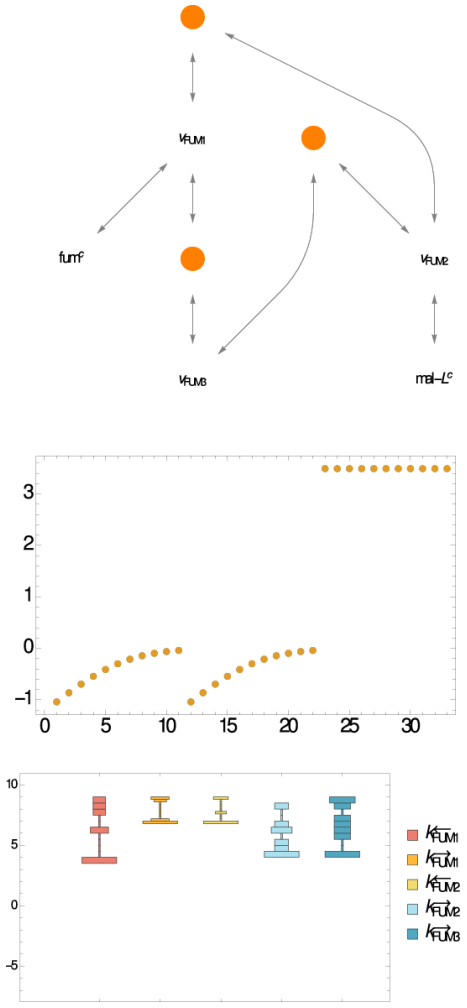


Figure 3.11: **Fumarase A fitting results**

Fumarase C (FUM)

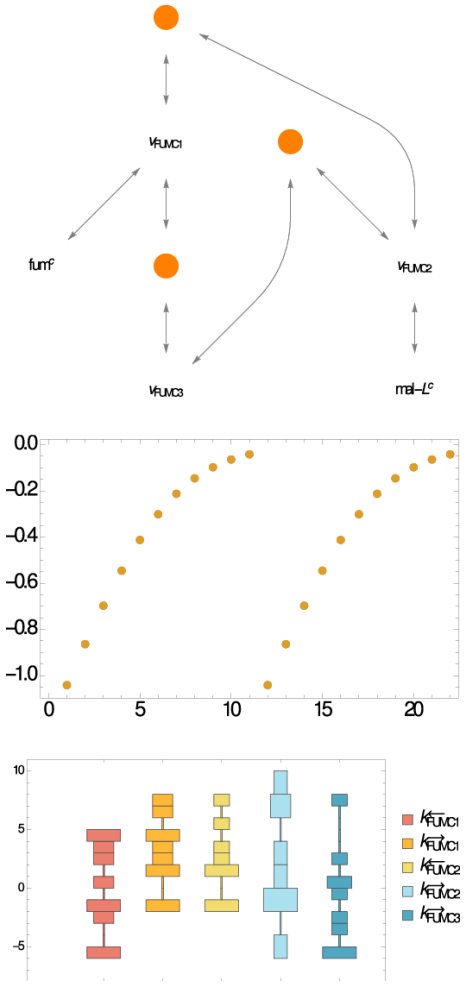


Figure 3.12: Fumarase C fitting results

Glycerol-3-phosphate dehydrogenase (G3PD5)

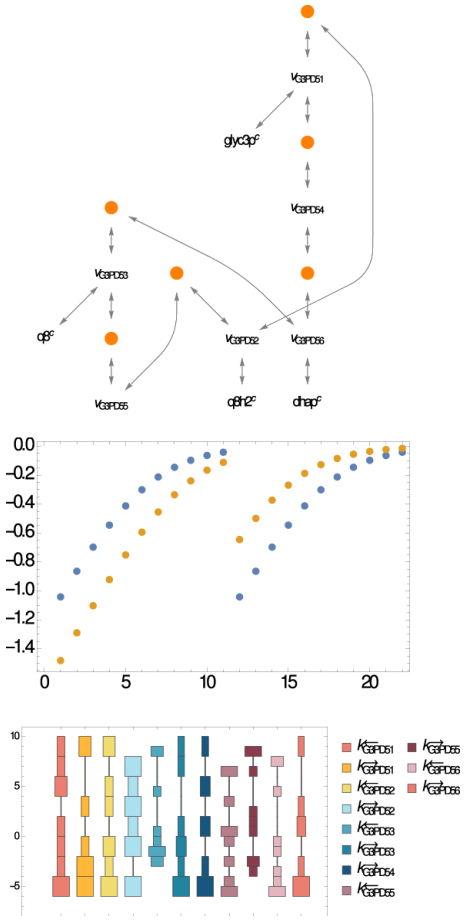


Figure 3.13: **Glycerol-3-phosphate dehydrogenase fitting results**

Glucose-6-phosphate dehydrogenase (G6PDH)

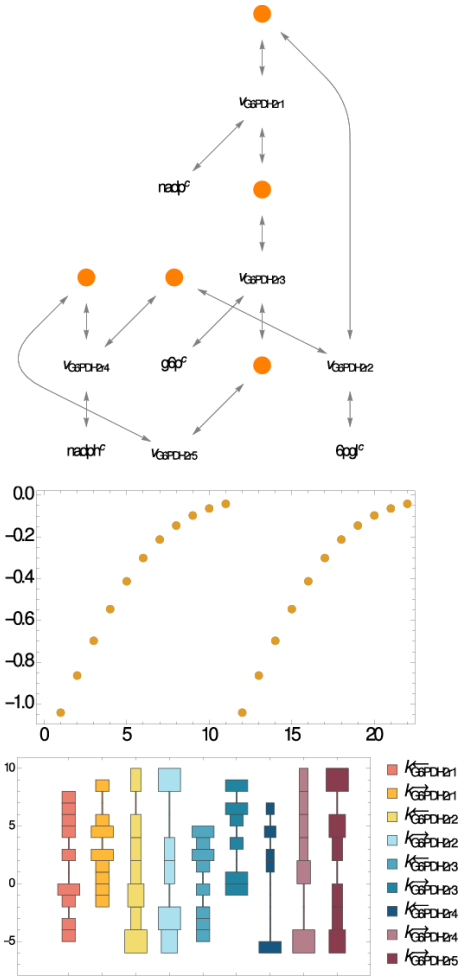


Figure 3.14: Glucose-6-phosphate dehydrogenase fitting results

Glyceraldehyde-3-phosphate dehydrogenase (GAPD)

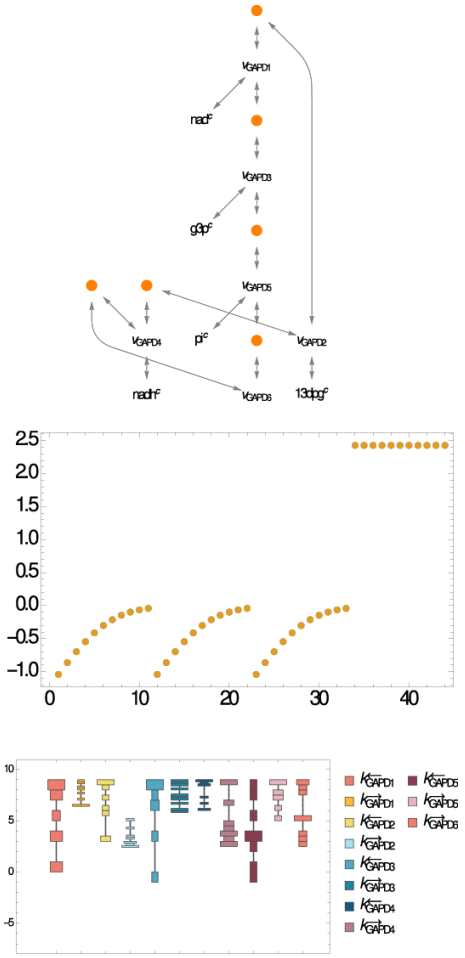


Figure 3.15: **Glyceraldehyde-3-phosphate dehydrogenase fitting results**

6-phosphogluconate dehydrogenase (GND)

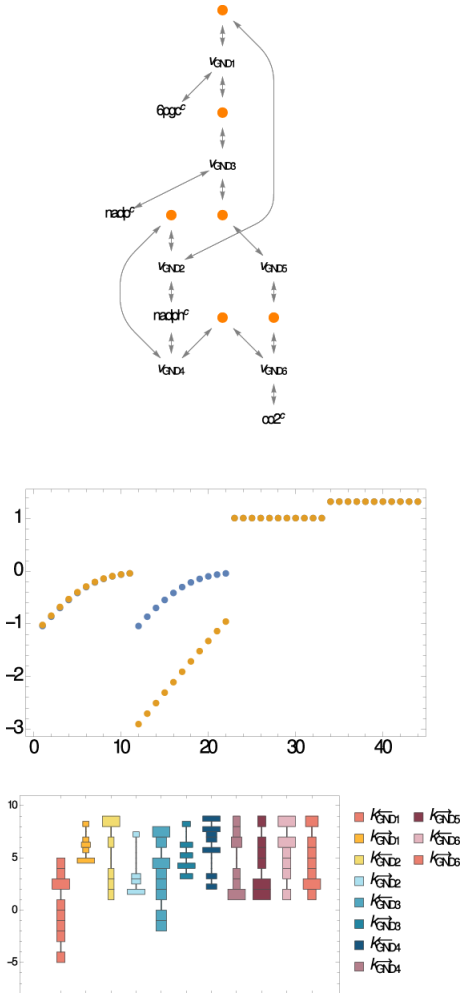


Figure 3.16: **6-phosphogluconate dehydrogenase fitting results**

Ribose-5-phosphate isomerase B (RPI)

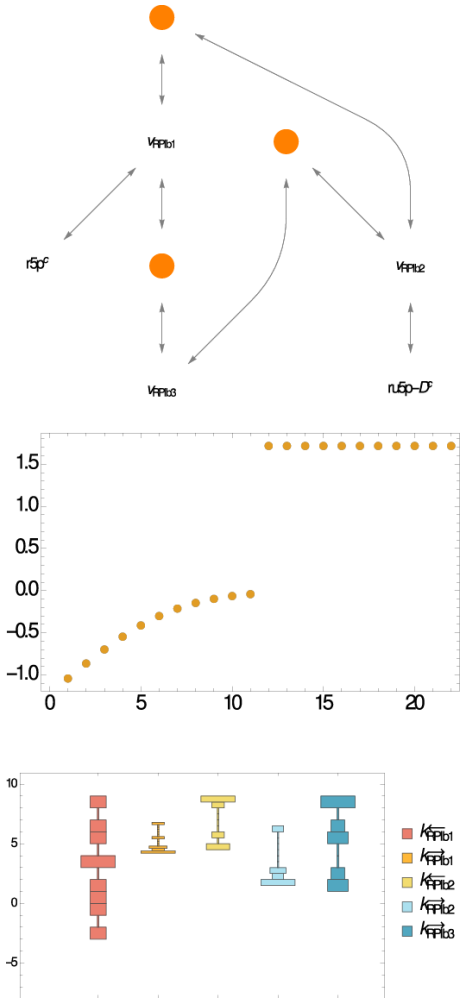


Figure 3.17: **Ribose-5-phosphate isomerase fitting results**

3.6.2 Model simulation

To show the viability of the approach, we integrated the glycolysis modules PGM, ENO, FBA, and GAPD into a kinetic model of glycolysis, where the rate law for non-module reactions is a mass action rate law based on stoichiometry. Steady-state concentrations and fluxes were set as described in the chapter, and the total enzyme for module reactions was back-calculated. The resulting model simulates stably.

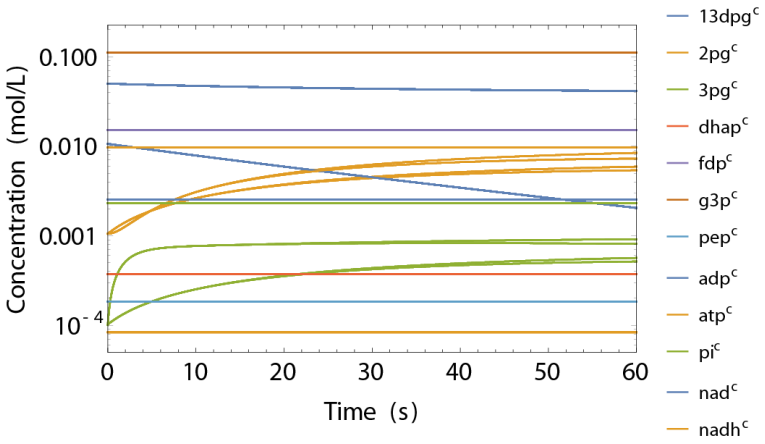


Figure 3.18: **Glycolysis simulation results**

3.7 Acknowledgements

Chapter 3 in part is a reprint of the material *Zielinski DC, Matos M, De Bree J, Sonnenschein N, Palsson BO A data driven workflow for the construction of bottom-up kinetic models of metabolism. In preparation.* The dissertation author was the primary author.

Chapter 4

Conclusions

4.1 Recapitulation

In this thesis, I have shown the utility of metabolic network reconstruction in analyzing complex biological data sets. Metabolic networks provide the functional connections between what would otherwise be treated a black box of statistical data points. It is the opinion of the author that these functional connections are absolutely necessary to uncover higher order associations that underlie system behavior. Although currently only primarily metabolic functions are analyzed with these methods, the scope of applications was nevertheless broad. In Chapter 1, I showed how metabolic pathways can be used based on topology alone to integrate drug-treated gene expression data with known disease signatures, such as altered pathway function, mutations, and nutrient associations, to identify candidate features underlying drug side effect pathogenesis. In Chapter 2, I showed how metabolite uptake and release profiles, cell growth rates, and cell compositions can be used in the context of the steady-state chemical mass balance equation to accurately calculate cancer cell flux states, providing context for observed metabolic function in terms of cellular growth requirements and potential for metabolic stress resistance. Finally in Chapter 3, I showed how data on in vitro enzyme kinetics can be integrated with in vivo flux and metabolic concentration data, correcting for many in vitro/in vivo differences, to fully parameterize a kinetic model of metabolism with a minimum of mechanistic assumptions.

4.2 Prospects of metabolic networks in analyzing complex data sets

The effect of drugs on native metabolism is an underappreciated aspect of drug response. The primary difficulty in further this type of effort is the availability of necessary data. The gene expression set used, the Connectivity Map data set, is near unique in its size and conformity to a single platform. Furthermore, the data used to validate candidate associations, in the form of curated disease-metabolism associations, was painstakingly extracted from the literature. In vitro data was also used to investigate individual drug-metabolism interactions, through targeted experiments testing three drugs on the MCF-7 cell line, but the in vivo applicability of these types of data is often a concern. The core utility of the metabolic network in this effort was to serve as a backbone onto which disparate data types can be mapped. In this sense, the work performed was not unique, as these types of pathway associations are commonly made in systems biology analysis efforts. The most unique aspect of the work was likely the combination of machine learning methods to extract candidate signatures tied to side effect pathogenesis with carefully curated in vivo disease data that is ideal to validate these types of predictions. The only better validation would be to conduct a clinical trial seeking to measure these metabolic variations directly in patient populations and associate them with side effect incidence. However, this scale of validation is simply not possible in an academic setting, and it is not clear whether the pharmaceutical industry has the incentive to conduct such studies were they available. The resurgence of interest in cancer metabolism has been accompanied by a number of efforts in systems biology to understand the principles underlying its hallmarks, most notably the Warburg effect. This work takes advantage of a recently published data set on cancer cell metabolic uptake and secretion profiles, the first of its size, to calculate accurate metabolic flux states for the NCI60 cell line panel and explore the Warburg effect from the perspective of the metabolic demands of growth and stress resistance. This effort most directly mirrors those in the constraint-based modeling field, where flux balance analysis is used in con-

junction with growth, uptake, and cell composition data to predict flux states. However, the work in this thesis presents by far the most accurate flux estimates yet obtained in a human system, as determined by agreement with ^{13}C-labeled glucose tracing and oxygen uptake data. A number of factors were necessary to obtain this agreement with experimental data. First, measurements on the exchange of metabolites were available on almost all highly exchanged metabolites, with notable exception to carbon dioxide, ammonia, and oxygen, where the latter was available for only 4 of the NCI60 lines. Second, growth data was available for the entire cell line panel, although cell content was not available and was forced to be estimated from uptake data, with good agreement with cell volume data. Third, the model used was a core model derived from the global human metabolic reconstruction Recon 2. This was beneficial because the global metabolic reconstruction has the potential to exhibit a number of unrealistic behaviors in practice. For example, NADH/NADPH transhydrogenase cycles, ATP-generating cycles, low flux alternate fermentative pathways like methylglyoxyl metabolism, are all commonly observed when attempting to calculate flux states with the global model. However, these types of activities are either infeasible thermodynamically or unrealistic kinetically, and should be excluded with little reservation when attempting to estimate real flux states under non-perturbed conditions. Looking forward, this work shows a clear utility of certain types of data, specifically growth and metabolic uptake data, that is relatively rarely generated in the modern era due in part to the excitement around sequencing and gene and protein expression data. It is hoped that more value will be given to obtaining this kind of data, especially in perturbed states such as hypoxia and clinically-relevant enzyme deficiencies. Developing kinetic models of metabolism was one of the first manifestations of systems biology. Originally developed as a method to enable the analysis of enzyme kinetic data, the Michaelis-Menten, or alternately the Briggs-Haldane, rate law has formed the core of kinetic modeling efforts as long as these efforts have existed, due to the correct recapitulation of saturation and regulation behavior and the reduction of parameters over a fully described mass action reaction system. It is particularly difficult to use parameters derived from initial rate experiments, most notably KM and

kcat, to parameter mass action systems because these parameters are extremely complicated non-linear combination of rate constants, leading to an underdetermined non-linear system. In this work, I showed how full mass action reaction systems can be parameterized using any kind of kinetic data, including steady-state rate laws, by setting up a non-linear least squares problem, dealing with the underdetermined system by sampling sets of equivalent rate constants that cause the enzyme reaction system to equally match measured data. The resulting kinetic modeling effort then becomes a matter of sampling these candidate rate constant sets and simulating or analyzing an ensemble of kinetic models. The benefits of this approach are numerous, but perhaps the most important is that the resulting rate constant sets can be thought of as uncertainties on the true underlying rate constant, and thus as more data is added, these uncertainties will inevitably become increasingly small. In my opinion, this concept has the potential to transforms the kinetic modeling field from a host of separate efforts with arbitrarily chosen modeling approaches and rate laws with little to learn from each other, to a concerted effort to iteratively improve our understanding of underlying kinetics and thermodynamics. I term this concept the kinetic reconstruction, analogous to the metabolic reconstructions discussed throughout the manuscript. These kinetic reconstructions would also be of interest to the protein modeling community, where the binding energies and rate constants that serve as the core of a kinetic reconstruction are estimated from protein structural simulations. It is hoped that this concept will prove formative in an era of kinetic modeling more focused on mechanistic integrity and practical applications of these models in health and metabolic engineering strain design applications.

4.3 Future

Looking forward, one immediate question is whether the utilization of biochemical network reconstructions in understanding complex data sets ends with metabolism, where the mass balance equation plays a central role. It has already been demonstrated that these approaches can be extended to transcription and

translation[76] as well as transcriptional regulation[77]. Pathway connectivity-based methods are certainly universal, although the definition of pathways in networks such as protein-protein interaction networks is a much more nebulous issue. Constraint-based steady-state flux calculation methods have already been applied at the genome scale to predict transcriptional and translational fluxes required to maintain a metabolic state given estimated flux per enzyme parameters[76]. Kinetic models of multi-scale systems exist already as well, with varying levels of detail, but it is not clear what final form these models will take. One key point in developing multi-scale models is that time and concentration scale differences exist in biochemical systems that can be exploited to model sub-systems individually without needing to consider all processes within the cell simultaneously. These modular approaches will likely simplify multi-scale model construction and analysis efforts. In summary, there is no indication that the utility of biochemical reconstructions in analyzing data sets will end with metabolism. It is likely that these reconstructions will only become more important in the future as data sets increasing in size and accuracy, and the expectations on the outcome of resulting analysis become greater.

Bibliography

[1] Pirmohamed M, James S, Meakin S, Green C, Scott AK, Walley TJ, Farrar K, Park BK, Breckenridge AM (2004) Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. BMJ 329: 15-9.

[2] Arrowsmith J (2011) Trial watch: Phase ii failures: 2008-2010. Nat Rev Drug Discov 10: 328-9.

[3] Arrowsmith J (2011) Trial watch: phase iii and submission failures: 2007-2010. Nat Rev Drug Discov 10: 87.

[4] Passarelli MC, Jacob-Filho W, Figueras A (2005) Adverse drug reactions in an elderly hospitalised population: inappropriate prescription is a leading cause. Drugs Aging 22: 767-77.

[5] Begaud B, Martin K, Fourrier A, Haramburu F (2002) Does age increase the risk of adverse drug reactions? Br J Clin Pharmacol 54: 550-2.

[6] Pirmohamed M, Park BK (2001) Genetic susceptibility to adverse drug reactions. Trends Pharmacol Sci 22: 298-305.

[7] Kuhn M, Al Banchaabouchi M, Campillos M, Jensen LJ, Gross C, Gavin AC, Bork P (2013) Systematic identification of proteins that elicit drug side effects. Mol Syst Biol 9: 663.

[8] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461: 747-53.

[9] Handschin C, Meyer UA (2003) Induction of drug metabolism: the role of nuclear receptors. Pharmacol Rev 55: 649-73.

[10] Wilke RA, Lin DW, Roden DM, Watkins PB, Flockhart D, Zineh I, Giacomini KM, Krauss RM (2007) Identifying genetic risk factors for serious adverse drug reactions: current progress and challenges. Nat Rev Drug Discov 6: 904-16.

[11] Sirota M, Dudley JT, Kim J, Chiang AP, Morgan AA, Sweet-Cordero A, Sage J, Butte AJ (2011) Discovery and preclinical validation of drug indications using compendia of public gene expression data. Sci Transl Med 3: 96ra77.

[12] Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BO (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci U S A 104: 1777-82.

[13] Bordbar A, Palsson BO (2012) Using the reconstructed genome-scale human metabolic network to study physiology and pathology. J Intern Med 271: 131-41.

[14] Lamb J, Crawford E, Peck D, Modell J, Blat I, Wrobel M, Lerner J, Brunet J, Subramanian A, Ross K, Reich M, Hieronymus H, Wei G, Armstrong S, Haggarty S, Clemons P, Wei R, Carr S, Lander E, Golub T (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. Science 313: 1929-35.

[15] Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? Nat Biotechnol 28: 245-8.

[16] Patil KR, Nielsen J (2005) Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proc Natl Acad Sci U S A 102: 2685-9.

[17] Brauer MJ, Yuan J, Bennett BD, Lu W, Kimball E, Botstein D, Rabinowitz JD (2006) Conservation of the metabolomic response to starvation across two divergent microbes. Proc Natl Acad Sci U S A 103: 19302-7.

[18] Bradley PH, Brauer MJ, Rabinowitz JD, Troyanskaya OG (2009) Coordinated concentration changes of transcripts and metabolites in saccharomyces cerevisiae. PLoS Comput Biol 5: e1000270.

[19] Mo ML, Palsson BO, Herrgard MJ (2009) Connecting extracellular metabolomic measurements to intracellular flux states in yeast. BMC Syst Biol 3: 37.

[20] Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y, Wishart DS (2014) Drugbank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 42: D1091-7.

[21] Wulffele MG, Kooy A, Lehert P, Bets D, Ogterop JC, Van der Burg BB, Donker AJM, Stehouwer CDA (2003) Effects of short-term treatment with metformin on serum concentrations of homocysteine, folate and vitamin b12 in type 2 diabetes mellitus: a randomized, placebo-controlled trial. Journal of Internal Medicine 254: 455-463.

[22] Chakraborty A, Chowdhury S, Bhattacharyya M (2011) Effect of metformin on oxidative stress, nitrosative stress and inflammatory biomarkers in type 2 diabetes patients. Diabetes Res Clin Pract 93: 56-62.

[23] Kim S, Shin H, Kim S, Kim J, Lee Y, Kim D, Lee M (2004) Genistein enhances expression of genes involved in fatty acid catabolism through activation of pparalpha. Mol Cell Endocrinol 220: 51-8.

[24] Kuhn M, Campillos M, Letunic I, Jensen L, Bork P (2010) A side effect resource to capture phenotypic effects of drugs. Molecular Systems Biology 6: 343.

[25] Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362-7.

[26] Ramirez AH, Shaffer CM, Delaney JT, Sexton DP, Levy SE, Rieder MJ, Nickerson DA, George J A L, Roden DM (2012) Novel rare variants in congenital cardiac arrhythmia genes are frequent in drug-induced torsades de pointes. Pharmacogenomics J .

[27] Valdivia CR, Ueda K, Ackerman MJ, Makielski JC (2009) Gpd1l links redox state to cardiac excitability by pkc-dependent phosphorylation of the sodium channel scn5a. Am J Physiol Heart Circ Physiol 297: H1446-52.

[28] El-Demerdash E, Mohamadin AM (2004) Does oxidative stress contribute in tricyclic antidepressants-induced cardiotoxicity? Toxicol Lett 152: 159-66.

[29] Altug S, Uzun O, Demiryurek AT, Cakici I, Abacioglu N, Kanzik I (1999) The role of nitric oxide in digoxin-induced arrhythmias in guinea-pigs. Pharmacol Toxicol 84: 3-8.

[30] Seyfried TN, Flores RE, Poff AM, D'Agostino DP (2014) Cancer as a metabolic disease: implications for novel therapeutics. Carcinogenesis 35: 515-27.

[31] Vander Heiden MG, Cantley LC, Thompson CB (2009) Understanding the warburg effect: the metabolic requirements of cell proliferation. Science 324: 1029-33.

[32] Shlomi T, Benyamini T, Gottlieb E, Sharan R, Ruppin E (2011) Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the warburg effect. PLoS Comput Biol 7: e1002018.

[33] Lunt SY, Vander Heiden MG (2011) Aerobic glycolysis: meeting the metabolic requirements of cell proliferation. Annu Rev Cell Dev Biol 27: 441-64.

[34] Shoemaker RH (2006) The nci60 human tumour cell line anticancer drug screen. Nat Rev Cancer 6: 813-23.

[35] Jain M, Nilsson R, Sharma S, Madhusudhan N, Kitami T, Souza AL, Kafri R, Kirschner MW, Clish CB, Mootha VK (2012) Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. Science 336: 1040-4.

[36] O'Connor PM, Jackman J, Bae I, Myers TG, Fan S, Mutoh M, Scudiero DA, Monks A, Sausville EA, Weinstein JN, Friend S, Fornace J A J, Kohn KW (1997) Characterization of the p53 tumor suppressor pathway in cell lines of the national cancer institute anticancer drug screen and correlations with the growth-inhibitory potency of 123 anticancer agents. Cancer Res 57: 4285-300.

[37] Dolfi SC, Chan LL, Qiu J, Tedeschi PM, Bertino JR, Hirshfield KM, Oltvai ZN, Vazquez A (2013) The metabolic demands of cancer cells are coupled to their size and protein synthesis rates. Cancer Metab 1: 20.

[38] Hyduke DR, Lewis NE, Palsson BO (2013) Analysis of omics data with genome-scale models of metabolism. Mol Biosyst 9: 167-74.

[39] Lewis NE, Abdel-Haleem AM (2013) The evolution of genome-scale models of cancer metabolism. Front Physiol 4: 237.

[40] Wagner BA, Venkataraman S, Buettner GR (2011) The rate of oxygen utilization by cells. Free Radic Biol Med 51: 700-12.

[41] Thiele I, Swainston N, Fleming RM, Hoppe A, Sahoo S, Aurich MK, Haraldsdottir H, Mo ML, Rolfsson O, Stobbe MD, Thorleifsson SG, Agren R, Bolling C, Bordel S, Chavali AK, Dobson P, Dunn WB, Endler L, Hala D, Hucka M, Hull D, Jameson D, Jamshidi N, Jonsson JJ, Juty N, Keating S, Nookaew I, Le Novere N, Malys N, Mazein A, Papin JA, Price ND, Selkov S E, Sigurdsson MI, Simeonidis E, Sonnenschein N, Smallbone K, Sorokin A, van Beek JH, Weichart D, Goryanin I, Nielsen J, Westerhoff HV, Kell DB, Mendes P, Palsson BO (2013) A community-driven global reconstruction of human metabolism. Nat Biotechnol 31: 419-25.

[42] Kilburn DG, Lilly MD, Webb FC (1969) The energetics of mammalian cell growth. J Cell Sci 4: 645-54.

[43] Fan J, Ye J, Kamphorst JJ, Shlomi T, Thompson CB, Rabinowitz JD (2014) Quantitative flux analysis reveals folate-dependent nadph production. Nature 510: 298-302.

[44] Kominsky DJ, Klawitter J, Brown JL, Boros LG, Melo JV, Eckhardt SG, Serkova NJ (2009) Abnormalities in glucose uptake and metabolism in imatinib-resistant human bcr-abl-positive cells. Clin Cancer Res 15: 3442-50.

[45] Zoppoli G, Solier S, Reinhold WC, Liu H, Connelly J J W, Monks A, Shoemaker RH, Abaan OD, Davis SR, Meltzer PS, Doroshow JH, Pommier Y (2012) Chek2 genomic and proteomic analyses reveal genetic inactivation or endogenous activation across the 60 cell lines of the us national cancer institute. Oncogene 31: 403-18.

[46] Moghaddas Gholami A, Hahne H, Wu Z, Auer FJ, Meng C, Wilhelm M, Kuster B (2013) Global proteome analysis of the nci-60 cell line panel. Cell Rep 4: 609-20.

[47] Sheikh K, Forster J, Nielsen LK (2005) Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of mus musculus. Biotechnol Prog 21: 112-21.

[48] Bonarius HP, Hatzimanikatis V, Meesters KP, de Gooijer CD, Schmid G, Tramper J (1996) Metabolic flux analysis of hybridoma cells in different culture media using mass balances. Biotechnol Bioeng 50: 299-318.

[49] Roschke AV, Tonon G, Gehlhaus KS, McTyre N, Bussey KJ, Lababidi S, Scudiero DA, Weinstein JN, Kirsch IR (2003) Karyotypic complexity of the nci-60 drug-screening panel. Cancer Res 63: 8634-47.

[50] Feijo Delgado F, Cermak N, Hecht VC, Son S, Li Y, Knudsen SM, Olcum S, Higgins JM, Chen J, Grover WH, Manalis SR (2013) Intracellular water exchange for measuring the dry mass, water mass and changes in chemical composition of living cells. PLoS ONE 8: e67590.

[51] Kit S, Fiscus J, Graham OL, Gross AL (1959) Metabolism and enzyme content of diploid and tetraploid lympnomas and carcinomas. Cancer Res 19: 201-6.

[52] De Menezes Y, De Faria FP, Sesso A (1996) In human hepatocellular carcinoma cells the total membrane surface area of each major organelle is a particular allometric function of the cytoplasmic volume. a morphometric study. J Submicrosc Cytol Pathol 28: 573-82.

[53] Frixione E, Porter RM (1986) Volume and surface changes of smooth endoplasmic reticulum in crayfish retinula cells upon light- and dark-adaptation. Journal of Comparative Physiology A 159: 667-674.

[54] Lowe RD, Szili EJ, Kirkbride P, Thissen H, Siuzdak G, Voelcker NH (2010) Combined immunocapture and laser desorption/ionization mass spectrometry on porous silicon. Anal Chem 82: 4201-8.

[55] Bennett BD, Yuan J, Kimball EH, Rabinowitz JD (2008) Absolute quantitation of intracellular metabolite concentrations by an isotope ratio-based approach. Nat Protoc 3: 1299-311.

[56] Palsson B, Zengler K (2010) The challenges of integrating multi-omic data sets. Nat Chem Biol 6: 787-9.

[57] Chassagnole C, Noisommit-Rizzi N, Schmid JW, Mauch K, Reuss M (2002) Dynamic modeling of the central carbon metabolism of escherichia coli. Biotechnol Bioeng 79: 53-73.

[58] Rizzi M, Baltes M, Theobald U, Reuss M (1997) In vivo analysis of metabolic dynamics in saccharomyces cerevisiae: Ii. mathematical model. Biotechnol Bioeng 55: 592-608.

[59] Vaseghi S, Baumeister A, Rizzi M, Reuss M (1999) In vivo dynamics of the pentose phosphate pathway in saccharomyces cerevisiae. Metab Eng 1: 128-40.

[60] Teusink B, Passarge J, Reijenga CA, Esgalhado E, van der Weijden CC, Schepper M, Walsh MC, Bakker BM, van Dam K, Westerhoff HV, Snoep JL (2000) Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? testing biochemistry. Eur J Biochem 267: 5313-29.

[61] Wright BE, Kelly PJ (1981) Kinetic models of metabolism in intact cells, tissues, and organisms. Curr Top Cell Regul 19: 103-58.

[62] Price ND, Papin JA, Schilling CH, Palsson BO (2003) Genome-scale microbial in silico models: the constraints-based approach. Trends Biotechnol 21: 162-9.

[63] Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ (2007) Quantitative prediction of cellular metabolism with constraint-based models: The cobra toolbox. Nat Protoc 2: 727-38.

[64] Feist AM, Zielinski DC, Orth JD, Schellenberger J, Herrgard MJ, Palsson BO (2010) Model-driven evaluation of the production potential for growth-coupled products of escherichia coli. Metab Eng 12: 173-86.

[65] Edwards JS, Ibarra RU, Palsson BO (2001) In silico predictions of escherichia coli metabolic capabilities are consistent with experimental data. Nat Biotechnol 19: 125-30.

[66] Fong SS, Palsson BO (2004) Metabolic gene-deletion strains of escherichia coli evolve to computationally predicted growth phenotypes. Nat Genet 36: 1056-8.

[67] Varma A, Palsson BO (1994) Predictions for oxygen supply control to enhance population stability of engineered production strains. Biotechnol Bioeng 43: 275-85.

[68] Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in escherichia coli. Mol Syst Biol 3: 119.

[69] Fleming RM, Thiele I, Nasheuer HP (2009) Quantitative assignment of reaction directionality in constraint-based models of metabolism: application to escherichia coli. Biophys Chem 145: 47-56.

[70] Qian H, Beard DA (2005) Thermodynamics of stoichiometric biochemical networks in living systems far from equilibrium. Biophys Chem 114: 213-20.

[71] Jankowski MD, Henry CS, Broadbelt LJ, Hatzimanikatis V (2008) Group contribution method for thermodynamic analysis of complex metabolic networks. Biophys J 95: 1487-99.

[72] Henry CS, Broadbelt LJ, Hatzimanikatis V (2007) Thermodynamics-based metabolic flux analysis. Biophys J 92: 1792-805.

[73] Bennett BD, Kimball EH, Gao M, Osterhout R, Van Dien SJ, Rabinowitz JD (2009) Absolute metabolite concentrations and implied enzyme active site occupancy in escherichia coli. Nat Chem Biol 5: 593-9.

[74] Shearer G, Lee JC, Koo JA, Kohl DH (2005) Quantitative estimation of channeling from early glycolytic intermediates to co in intact escherichia coli. FEBS J 272: 3260-9.

[75] Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D (2004) Brenda, the enzyme database: updates and major new developments. Nucleic Acids Res 32: D431-3.

[76] Valgepea K, Adamberg K, Nahku R, Lahtvee PJ, Arike L, Vilu R (2010) Systems biology approach reveals that overflow metabolism of acetate in escherichia coli is triggered by carbon catabolite repression of acetyl-coa synthetase. BMC Syst Biol 4: 166.

[77] Kotte O, Zaugg JB, Heinemann M (2010) Bacterial adaptation through distributed sensing of metabolic fluxes. Mol Syst Biol 6: 355.