UNIVERSITY OF CALIFORNIA, SAN DIEGO

Optimization of microbial cell factories with systems biology

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

in

Bioengineering

by

Zachary A. King

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Jeff Hasty
Professor Andrew D. McCulloch
Professor Christian M. Metallo
Professor Glenn Tesler

2016

The Dissertation of Zachary A. King is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2016

DEDICATION

For Elise,

what a delight

to see the world through your eyes

EPIGRAPH

When I'm working on a problem, I never think about beauty.
I think only how to solve the problem. But when I have finished,
if the solution is not beautiful, I know it is wrong.
—R. Buckminster Fuller

Hofstadter's Law: It always takes longer than you expect,
even when you take into account Hofstadter's Law.
—Douglas Hofstadter, Gödel, Escher, Bach: An Eternal Golden Braid

Coffee?
—@coffee_dad

TABLE OF CONTENTS

Signature Page iii

Dedication iv

Epigraph v

Table of Contents vi

List of Figures ix

List of Tables xi

Acknowledgements xii

Vita xvi

Abstract of the Dissertation xviii

Chapter 1 Introduction 1
1.1 Systems biology 2
1.2 Microbial cell factories 4
1.3 Introducing the Thesis 9

Chapter 2 BiGG Models: A platform for integrating, standardizing, and sharing genome-scale models 11
2.1 Introduction 11
2.2 Knowledge base content 15
2.3 Getting started with BiGG Models 17
2.3.1 BiGG website 17
2.3.2 Using BiGG Models for COBRA modeling 19
2.3.3 Using BiGG Models for building GEMs 20
2.3.4 Accessing the API 21
2.4 Implementation of standards 23
2.4.1 Loading genomes and GEMs 23
2.4.2 BiGG identifiers 24
2.4.3 ModelPolisher 26
2.4.4 Design and implementation 27
2.5 Conclusion 27
2.6 Availability and requirements 28

Chapter 3 Escher: A web application for building, sharing, and embedding data-rich visualizations of biological pathways 30
3.1 Introduction 30
3.2 Results 34
3.2.1 Building pathway maps 34
3.2.2 Visualizing data 37
3.3 Design and Implementation 40
3.4 Availability and Future Directions 46
3.5 Availability and requirements 47

Chapter 4 Optimizing Cofactor Specificity of Oxidoreductase Enzymes for the Generation of Microbial Production Strains—OptSwap . . . 49
4.1 Introduction 49
4.2 Methods 54
4.2.1 Modeling and computational tools 54
4.2.2 Model reduction and selection of reaction set for knockouts 55
4.2.3 Selection of reaction set for cofactor specificity swaps 56
4.2.4 MILP formulation 57
4.3 Results 63
4.4 Discussion 71

Chapter 5 Optimal cofactor swapping can increase the theoretical yield for chemical production in *Escherichia coli* and *Saccharomyces cerevisiae* 75
5.1 Introduction 75
5.2 Methods 79
5.2.1 Models and parameters 79
5.2.2 Non-native pathways 81
5.2.3 Selection of the reaction sets for cofactor-specificity swaps 81
5.2.4 MILP formulation 85
5.2.5 Non-unique solutions 87
5.2.6 Sensitivity analysis 90
5.2.7 Determining cofactor usage 91
5.3 Results 91
5.3.1 Native Pathways 91
5.3.2 Non-native pathways in *E. coli* 95
5.3.3 NADPH yield and parameter sensitivity 95
5.4 Discussion 99
5.4.1 Cofactor swaps for certain enzymes have a global impact on theoretical yields 99
5.4.2 Simulated theoretical yield improvement matches experimental observations for a GAPD swap 101

5.4.3 Optimal cofactor swaps increase ATP availability . 102
5.4.4 Cofactor swapping in yeast has a greater effect with D-xylose as a substrate . . . 103
5.4.5 Theoretical yield of non-native products increases with swaps . . . 105
5.4.6 Theoretical yields are sensitive to knowledge of cofactor preference and enzyme promiscuity . . . 106
5.5 Conclusion . . . 107

Chapter 6 Literature mining supports a next-generation modeling approach to predict cellular byproduct secretion . . . 108
6.1 Introduction . . . 108
6.2 Results . . . 111
6.2.1 Literature mining provides a diverse set of strains and phenotypes. . . . 111
6.2.2 Genome-scale models do not differentiate between isozymes. . . . 115
6.2.3 Larger models solve false predictions of cell death. . 118
6.2.4 Simulations suggest that some strains have room to evolve. . . . 119
6.2.5 Next-generation ME-models improve predictions but require parameterization. . . . 120
6.3 Discussion . . . 122
6.4 Methods . . . 124
6.4.1 Literature mining. . . . 124
6.4.2 Simulations. . . . 125
6.4.3 Parameter sampling. . . . 127
6.4.4 Failure model categorization. . . . 127

Chapter 7 Conclusions and Outlook . . . 128
7.1 Next-generation models and predictions . . . 128
7.1.1 New cellular networks . . . 130
7.1.2 Modularity . . . 132
7.2 Conclusion . . . 134

Bibliography . . . 135

LIST OF FIGURES

Figure 1.1: Three types of COBRA predictions have been successfully implemented for systems metabolic engineering. 9

Figure 2.1: BiGG Models content. 16
Figure 2.2: The BiGG Models homepage. 18
Figure 2.3: Accessing BiGG. 21
Figure 2.4: Standardizing GEMs. 25

Figure 3.1: The Escher interface. 35
Figure 3.2: Data visualization. 38
Figure 3.3: The organization of the Escher project. 41
Figure 3.4: The import and export types in Escher and the EscherConverter. 44

Figure 4.1: The OptSwap formulation for optimizing cofactor specificity of major metabolic enzymes (NAD(H) vs. NADP(H)) coupled to reaction knockouts. 59
Figure 4.2: Calculated production envelopes for OptSwap designs predicted to have significantly higher optimal production rate or substrate-specific productivity than designs with just reaction knockouts. . 66
Figure 4.3: Network diagrams showing the shift in reduced cofactor usage between wild-type and the L-alanine production design. 67
Figure 4.4: The predicted effect of combinatorial interventions on the design for anaerobic production of L-alanine on glucose minimal media. . 68

Figure 5.1: Cofactor specificity modifications of oxidoreductase reactions can improve the maximum theoretical yield of metabolic byproducts. 87
Figure 5.2: Results from optimizing the cofactor specificity of oxidoreductases in the *E. coli* *i*JO1366 model to increase the maximum theoretical yield of native metabolic compounds. 88
Figure 5.3: Results from optimizing the cofactor specificity of oxidoreductases in the *S. cerevisiae* *i*MM904 model to increase the maximum theoretical yield of native metabolic compounds. 89
Figure 5.4: The effect of swapping cofactor specificity of oxidoreductases for the production of non-native compounds in *E. coli*. 90
Figure 5.5: A sensitivity analysis on the impact of cofactor swapping when varying modeling parameters. 97

Figure 6.1: The bibliomic database. 112
Figure 6.2: The engineered fermentation pathways in *E. coli*. 113
Figure 6.3: Simulations of the bibliomic dataset in E. coli GEMs. 115
Figure 6.4: Comparing simulations with experiments. 117

Figure 7.1: COBRA tools have advanced through (a) increased scope of cellular systems which can be modeled, and (b) highly modular simulation strategies, built upon genome-scale models of metabolism. 130

LIST OF TABLES

Table 1.1: A comparison of modelling and analysis techniques for high-throughput data. 5

Table 2.1: Comparing BiGG (2010) to BiGG Models (2015). 14
Table 2.2: BiGG Identifiers. 26

Table 4.1: Oxidoreductase reactions targeted for analysis with OptSwap. . . 58
Table 4.2: Calculated maximum optimal production rate for designs selected by OptSwap and designs selected by RobustKnock. 69

Table 5.1: Non-native pathways reconstructed in *E. coli* for cofactor balance optimization. 81
Table 5.2: Oxidoreductase enzymes in the *E. coli* *i*JO1366 model chosen for cofactor balance optimizations. 83
Table 5.3: Oxidoreductase enzymes in the *i*MM904 model of *S. cerevisiae* chosen for cofactor balance optimizations. 84

Table 6.1: The increasing size and scope of genome-scale models of *E. coli*. . 114

Table 7.1: Current and next-generation COBRA models – types of predictions that are possible. 129

ACKNOWLEDGEMENTS

First, I must thank the graduate students, post-docs, and staff in the Palsson lab whose mentorship and ideas made this dissertation possible. Special thanks to Josh Lerman for being the world's most enthusiastic student of biology and for giving me an unparalleled opportunity to see my research applied in industry, to Teddy O'Brien for being a great sounding board and chill officemate, and to Gaby Guzman for answering all my questions about the mysterious wet lab. Thanks to Ali Ebrahim for being an equally obsessive coder and a great friend, Colton Lloyd for being an increasingly obsessive coder and for not letting his "buck-eyes" come up in conversation too often, and to Ke Chen for teaching me about structural modeling and for staying calm when her computer was infected by ghosts. Thanks to Aarash Bordbar, Dan Zielinski, Andreas Dräger, Justin Lu, Jon Monk, Jose Utrilla, Miguel Campodonico, Niko Sonnenschein, Steve Federowicz, and Richard Szubin. Thanks to Marc Abrams and Helder Balelo for always being available to help—we are lost when they go on vacation. Jan Lenington is the greatest administrator of all time; thanks Jan! And thank you Joanne Liu, Ryan LaCroix, James Yurkovich, Justin Tan, Troy Sandberg, Laurence Yang, Liz Brunk, Nathan Mih, Bin Du, Jenni Levering, Jarod Broddrick, and everyone who has come and gone from our lab over the last five years.

I owe Adam Feist a special thanks for giving me the opportunity to join the lab and for passing along this thing called "OptSwap" that became two chapters of this dissertation. He is also responsible for introducing me to at least 7 Mikkeller locations on 3 continents where we partook of many delicious beers. And thanks to Nate Lewis for providing an incredible amount of career advice and guidance on these projects.

Thanks to our many collaborators, including Markus Herrgård and João

Cardoso at the Center for Biosustainability and to Andreas Prlic and Ali Altunkaya at the Protein Data Bank. Thanks to Pierre Salvy, Amoolya Singh, Daniel Dougherty, Erin Wilson, Maxime Durot, and all the scientists at Amyris for their invaluable help and guidance during the development of Escher and for engendering a fantastic learning environment at Amyris.

To my advisor Bernhard Palsson, I am overwhelmingly grateful to have had the opportunity to work in your lab. I have not found a more exciting place to discover biology than this research group, and I owe much of that to you for having nurtured an environment of discovery, ambition, and intellectual rigor. Thank you for your excellent advice, your support of my work, and for bringing your diverse experiences into every meeting.

This work is made possible by funding from the National Science Foundation Graduate Research Fellowship Program (grant no. DGE-1144086). This grant gave me the flexibility to pursue some off-topic side projects, one of which became Chapter 3. Thanks to the Novo Nordisk Foundation and the Center for Biosustainability at the Technical University of Denmark for generously supporting so much of our work and for the unique (at least in academia) support of ongoing software development and maintenance. This work also received support from the Department of Energy (grant nos. DE-SC0008701 and DE-SC0004917), the European Commission as part of a Marie Curie International Outgoing Fellowship within the EU 7th Framework Program for Research and Technological Development (EU project AMBiCon, 332020), the National Institutes of Health (grant no. R21 HD080682), and the Novo Nordisk Foundation (grant no. NNF 132150-002).

Thanks to my wife Gaby; you make all of this brain-busting work worthwhile.

And thanks to my daughter Elise, who only came on the scene for the final year of my dissertation, but whose specific growth rate has been amazing to see. Thanks to my mom for instilling in me these academic aspirations and an empirical approach to everything. And thanks to my dad for a different kind of empiricism—how to work hard, be honest, support those around me, and strive to be great at the task at hand.

Chapter 1 is adapted from published manuscripts: King, Z. A., Lloyd, C. J., Feist, A. M., and Palsson, B. O. (2015b). "Next-generation genome-scale models for metabolic engineering". In: *Curr. Opin. Biotechnol.* 35, pp. 23–29. DOI: 10.1016/j.copbio.2014.12.016. The dissertation author was the primary author of the review. Bordbar, A., Monk, J. M., King, Z. A., and Palsson, B. O. (2014a). "Constraint-based models predict metabolic and associated cellular functions". In: *Nat. Rev. Genet.* 15.2, pp. 107–120. DOI: 10.1038/nrg3643. The dissertation author was one of the authors of the review.

Chapter 2 is a reprint of a published manuscript: King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O., and Lewis, N. E. (2016). "BiGG Models: A platform for integrating, standardizing and sharing genome-scale models". In: *Nucleic Acids Res.* 44.D1, pp. D515–22. DOI: 10.1093/nar/gkv1049. The dissertation author was the primary author of the paper and was responsible for the research.

Chapter 3 is a reprint of a published manuscript: King, Z. A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N. E., and Palsson, B. O. (2015a). "Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways". In: *PLoS Comput. Biol.* 11.8, e1004321. DOI: 10.1371/journal.pcbi.1004321. The dissertation author was the primary author of the paper and was

responsible for the research.

Chapter 4 is a reprint of a published manuscript: King, Z. A. and Feist, A. M. (2013). "Optimizing Cofactor Specificity of Oxidoreductase Enzymes for the Generation of Microbial Production Strains—OptSwap". In: *Ind. Biotechnol.* 9.4, pp. 236–246. DOI: 10.1089/ind.2013.0005. The dissertation author was the primary author of the paper and was responsible for the research.

Chapter 5 is a reprint of a published manuscript: King, Z. A. and Feist, A. M. (2014). "Optimal cofactor swapping can increase the theoretical yield for chemical production in *Escherichia coli* and Saccharomyces cerevisiae". In: *Metab. Eng.* 24, pp. 117–128. DOI: 10.1016/j.ymben.2014.05.009. The dissertation author was the primary author of the paper and was responsible for the research.

Chapter 6 is a reprint of a published manuscript: King, Z. A., O'Brien, E. J., Feist, A. M., and Palsson, B. O. "Literature mining supports a next-generation modeling approach to predict cellular byproduct secretion". In: *Metabolic Engineering.* under review. The dissertation author was the primary author of the paper and was responsible for the research.

Chapter 7 is adapted from a published manuscript: King, Z. A., Lloyd, C. J., Feist, A. M., and Palsson, B. O. (2015b). "Next-generation genome-scale models for metabolic engineering". In: *Curr. Opin. Biotechnol.* 35, pp. 23–29. DOI: 10.1016/j.copbio.2014.12.016. The dissertation author was the primary author of the review.

VITA

2011	B. S. E. in Biomedical Engineering, University of Michigan
2016	Ph. D. in Bioengineering, University of California, San Diego

PUBLICATIONS

King, Z. A., O'Brien, E. J., Feist, A. M., and Palsson, B. O. "Literature mining supports a next-generation modeling approach to predict cellular byproduct secretion". In: *Metabolic Engineering.* under review.

Hefzi, H., Ang, K. S., Hanscho, M., Bordbar, A., Ruckerbauer, D., Lakshmanan, M., Orellana, C. A., Baycin-Hizal, D., Huang, Y., Ley, D., Martinez, V. S., Kyriakopoulos, S., Jiménez, N. E., Zielinski, D. C., Quek, L.-E., Wulff, T., Arnsdorf, J., Li, S., Lee, J. S., Paglia, G., Loira, N., Spahn, P. N., Pedersen, L. E., Gutierrez, J. M., King, Z. A., Lund, A. M., Nagarajan, H., Thomas, A., Abdel-Haleem, A. M., Zanghellini, J., Kildegaard, H. F., Voldborg, B. G., Gerdtzen, Z. P., Betenbaugh, M. J., Palsson, B. O., Andersen, M. R., Nielsen, L. K., Borth, N., Lee, D.-Y., and Lewis, N. E. (2016). "A Consensus Genome-scale Reconstruction of Chinese Hamster Ovary Cell Metabolism". In: *Cell Syst* 3.5, 434–443.e8. DOI: 10.1016/j.cels.2016.10.020.

Gallina, A. A., Layer, M., King, Z. A., Levering, J., Palsson, B. Ø., Zengler, K., and Peers, G. (2016). "A Phaeodactylum tricornutum literature database for interactive annotation of content". In: *Algal Research* 18, pp. 241–243.

King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O., and Lewis, N. E. (2016). "BiGG Models: A platform for integrating, standardizing and sharing genome-scale models". In: *Nucleic Acids Res.* 44.D1, pp. D515–22. DOI: 10.1093/nar/gkv1049.

King, Z. A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N. E., and Palsson, B. O. (2015a). "Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways". In: *PLoS Comput. Biol.* 11.8, e1004321. DOI: 10.1371/journal.pcbi.1004321.

King, Z. A., Lloyd, C. J., Feist, A. M., and Palsson, B. O. (2015b). "Next-generation genome-scale models for metabolic engineering". In: *Curr. Opin. Biotechnol.* 35, pp. 23–29. DOI: 10.1016/j.copbio.2014.12.016.

King, Z. A. and Feist, A. M. (2014). "Optimal cofactor swapping can increase the theoretical yield for chemical production in *Escherichia coli* and Saccharomyces cerevisiae". In: *Metab. Eng.* 24, pp. 117–128. DOI: 10.1016/j.ymben.2014.05.009.

Bordbar, A., Monk, J. M., King, Z. A., and Palsson, B. O. (2014a). "Constraint-based models predict metabolic and associated cellular functions". In: *Nat. Rev. Genet.* 15.2, pp. 107–120. DOI: 10.1038/nrg3643.

King, Z. A. and Feist, A. M. (2013). "Optimizing Cofactor Specificity of Oxidoreductase Enzymes for the Generation of Microbial Production Strains—OptSwap". In: *Ind. Biotechnol.* 9.4, pp. 236–246. DOI: 10.1089/ind.2013.0005.

McCloskey, D., Gangoiti, J. A., King, Z. A., Naviaux, R. K., Barshop, B. A., Palsson, B., and Feist, A. M. (2013). "A model-driven quantitative metabolomics analysis of aerobic and anaerobic metabolism in *E. coli* K-12 MG1655 that is biochemically and thermodynamically consistent". In: *Biotechnol. Bioeng.* 111.4, pp. 803–815. DOI: 10.1002/bit.25133.

ABSTRACT OF THE DISSERTATION

Optimization of microbial cell factories with systems biology

by

Zachary A. King

Doctor of Philosophy in Bioengineering

University of California, San Diego, 2016

Professor Bernhard Ø. Palsson, Chair

Microbial cell factories are expected to have a transformative impact on the chemical industry, but, first, we must meet the challenges of designing and optimizing high-yield cell factory strains. The most popular conceptual model for cell factory optimization is the *design-build-test-learn* cycle. I present methods that use systems biology to improve the optimization process in each of these steps. First, the *build* step requires a parts list for a host organism and any heterologous pathways. I present BiGG Models, a database of more than 75 high-quality, manually-curated genome-scale metabolic models that comprise a standardized metabolic parts list. BiGG Models has

become the most popular resource in the community for gold-standard genome-scale metabolic models. For the *test* step, contextualization of omics data is an enormous challenge, and I present a visualization tool to address this challenge. Escher is a web application for visualizing data on biological pathways. With Escher, users can identify trends in common gene-oriented data types (e.g. RNA-Seq, proteomics) and metabolite- and reaction-oriented data types (e.g. metabolomics, fluxomics). For the *learn* step, genome-scale models can be used to identify general trends in cell factory performance. I introduce a computational method—OptSwap—to predict bioprocessing strain designs by identifying optimal modifications of the cofactor binding specificities of oxidoreductase enzymes and by identifying complementary reaction knockouts. I also present an optimization procedure that identifies optimal cofactor-specificity "swaps" for improving theoretical yield in genome-scale metabolic models. Swapping the cofactor specificity of central metabolic enzymes is shown to increase NADPH production and increase theoretical yields for many native and non-native products. Last, the *design* step requires models that can successfully predict phenotype from genotype. I assess the predictive capabilities of existing models of *Escherichia coli* through literature mining and simulate strains from the literature in six historical genome-scale models of *E. coli*. This study shows that the predictive power of models has increased as they have expanded in size and scope. Together, these studies provide a path toward successfully applying systems biology methods to optimizing microbial cell factories.

Chapter 1

Introduction

Biology is the only field of study that asks us to look deep inside ourselves, to the interior of our physical being. Somewhere in there is a mechanism that makes us tic. We are often tempted to view ourselves as mechanical beings. When a bone breaks, it cracks and fractures like inanimate rock. Our skin reacts with acids and bases. We physically respond to heat and cold, and even a lack of gravity. But with every discovery, biology seems to move further from these mechanistic and reductionist explanations, becoming richer in mathematics, physics, logic, and computation. If there is a more thrilling field of study, I have not found it.

Incredible revelations on the function of our internal "ticker" have come from many disciplines: microbiology and cell theory in the 19th century, developmental biology and genetics in the early 20th century, molecular biology beginning in the 1960s, and systems biology since the development of whole genome sequencing in the 1990s. With each wave of discovery, we develop a new vision of what our ticker looks like. Are we elaborate, skyskraping towers of protein and lipid? Or are we biochemical reactors where each stimulus sets off the an unthinkable network of toppling dominoes? Are

we information processing machines, stuffed full of tickertape written in a quaternary alphabet? Or maybe we are all of those at once—a network of networks—where physical structure, biochemical dynamics, and information processing are in constant communication.

In this dissertation, I employ the latter systems biology approach. I will defend the thesis that systems biology methods are essential to optimizing cell factory strains, especially for contextualizing and visualizing data and for generalizing observations about a model system to all cell factory strains. To begin, I will introduce the field of systems biology and the specific mathematical tools employed in this dissertation.

1.1 Systems biology

Whole genome sequencing, high-throughput data collection, and advances in computation have made it possible to construct a complete parts list for a model organism. Every gene in a complete genome sequence can be identified. Through a bottom-up reconstruction process, every biochemical reaction and metabolite in a metabolic network can be added to the parts list based on the annotated genome and on experimental literature (Feist et al. 2009; Thiele and Palsson 2010). A similar approach can be taken with other biological networks. The resulting *b*iochemical, *g*enetic, and *g*enomic (BiGG) knowledge base forms a scaffold upon which we can build predictive models, data analysis methods, and visualization tools.

A key goal of systems biology is to compute the relationship between genotype and phenotype. Diverse approaches that range from stochastic kinetic models to statistical Bayesian networks have been applied, and each of these approaches has differing rationales and advantages (Table 1.1). One of these approaches is constraint-

based reconstruction and analysis (COBRA), which can be implemented with a BiGG knowledge base. The BiGG knowledge base is first converted to a mathematically consistent format, primarily by generating a stoichiometric matrix. This matrix is the central component of an ever-growing set of COBRA modeling methods (Lewis, Nagarajan, and Palsson 2012). COBRA methods are primarily based on metabolic networks, including multicellular metabolic interactions (Zhuang et al. 2011; Klitgord and Segrè 2010; Bordbar et al. 2011; Bordbar et al. 2010; Lewis et al. 2010b). COBRA methods also exist for signaling (Papin and Palsson 2004; Li et al. 2009), transcriptional regulation (Gianchandani et al. 2009), and macromolecule synthesis (Thiele et al. 2009).

The first COBRA method for biological predictions was flux-balance analysis (FBA). Its formulation is rooted in the hypothesis that a cell is "striving" to achieve a metabolic objective. Studies have shown that, by optimizing the assumed cellular objectives of growth (Ibarra, Edwards, and Palsson 2002) and energy use (Carlson and Srienc 2004b; Carlson and Srienc 2004a) one can predict metabolic fluxes in microorganisms. The constraint-based modeling framework is also amenable to simultaneous integration of a range of omic data types (Hyduke, Lewis, and Palsson 2013). In particular, omic data have been used both to constrain calculated flux distributions and as a comparison and validation tool for model predictions. Such omic data integration has enabled context-specific studies of the metabolism of an organism, such as studies of enzyme promiscuity (Guzmán et al. 2015; Nam et al. 2012) and pathogenesis (Lobel et al. 2012). Recent work shows that COBRA methods can be used to place interaction networks of diverse biological components into context and to interpret these networks (Szappanos et al. 2011). Finally, the discrepancies

between COBRA predictions and experimental data have been used to design targeted experiments that correct inaccuracies in metabolic knowledge (Reed et al. 2006b; Orth and Palsson 2010).

Methods that model the genotype-phenotype relationship can also be applied to building and optimizing microbial cell factories. The promise of automating the design of microbial cell factories is still largely unrealized, but it has the potential to transform our built environment and our change our relationship with the natural environment.

1.2 Microbial cell factories

The demand for raw material inputs to agriculture, industry, and energy are growing steadily, and concerns about environmental sustainability are becoming more acute; thus, alternatives to traditional, fossil-fuel based chemical production are becoming economically viable (Johnson 2007). Cell factories, which use microorganisms to produce materials from renewable biomass, are an attractive alternative, and an increasing number of platform chemicals are being produced at industrial scale using engineered microorganisms (Manzer, Waal, and Imhof 2013). To meet the demand for robust, high-yield production strains, an initial metabolic engineering strategy was developed, which used random mutagenesis and screening to identify strains with improved production performance (i.e. titer, productivity, and yield). However, the permutations possible in genomic sequences are so numerous that a mutagenesis and screening approach can only explore a small subset of possible strains, so the highest performance strains might never be identified. On the other hand, if *targeted* strain improvements can be predicted, then strains can be engineered which would not be

Table 1.1: A comparison of modelling and analysis techniques for high-throughput data.

Method	Model systems	Parameterization	Typical prediction type	Advantages	Disadvantages
Stochastic kinetic modeling	Small-scale biological processes	Detailed kinetic parameters	Reaction fluxes, component concentrations and regulatory states	- Mechanistic - Dynamic - Captures biological stochasticity and biophysics	- Computationally intensive - Difficult to parameterize - Challenging to model multiple timescales
Deterministic kinetic modeling	Small-scale biological processes	Detailed kinetic parameters	Reaction fluxes, component concentrations and regulatory states	- Mechanistic - Dynamic	- Computationally intensive - Difficult to parameterize
Constraint-based modeling	Genome-scale metabolism	Network topology, and uptake and secretion rates	Metabolic flux states and gene essentiality	- Mechanistic - Large scale - No kinetic information is required	- No inherent dynamic or regulatory predictions - No explicit representation of metabolic concentrations
Logical, Boolean or rule-based formalisms	Signalling networks and transcriptional regulatory networks	Rule-based interaction network	Global activity states and on-off states of genes	Can model dynamics and regulation	Biological systems are rarely discrete
Bayesian approaches	Gene regulatory networks and signalling networks	High-throughput data sets	Probability distribution score	- Non-biased - Can include disparate and even non-biological data - Takes previous associations into account	- Statistical - Issues of over-fitting - Requires comprehensive training data
Graph and interaction networks	Protein-protein and genetic interaction networks	Interaction network that is based on biological data	Enriched clusters of genes and proteins	- Incorporates prior biological data - Encompasses most cellular processes	Dynamics are not explicitly represented
Pathway enrichment analysis	Metabolic and signaling networks	Pathway databases (e.g. KEGG, Gene Ontology, BioCyc)	Enriched pathways	- Simple and quick - Takes prior knowledge into account	- Biased to human-defined pathways - Non-modelling approach

found using untargeted mutagenesis and screening, and this can be accomplished with the tools of systems biology.

An important first step in designing a production strain for a non-native metabolite is to identify and build a synthetic pathway (Shin et al. 2013). COBRA methods have been employed successfully for pathway prediction and optimization (Shin et al. 2013; Campodonico et al. 2014). After a pathway has been designed, strain optimization is performed to increase the yield and productivity of the strain. The most common paradigm for managing strain optimization is the "design-build-test-learn" cycle, which will be used to help contextualize the methods in this dissertation (Smanski et al. 2014; Liu et al. 2015).

A great number of COBRA methods have been developed (Lewis, Nagarajan, and Palsson 2012), and the methods that have led to experimental improvements in production strains can be categorized according to the types of predictions made. The most common prediction of COBRA methods in systems metabolic engineering has been the calculation of maximum *theoretical yield*, the percentage of substrate carbon that can be converted to a target molecule, given the limitations of carbon and redox balance in the stoichiometric network (Fig. 1.1a). Yield is a critical consideration when considering the economic viability of a chemical production, so these analyses have direct consequences for cell factory design. Shen and Liao (2013) recently demonstrated the importance of theoretical yield calculations for designing production strains. In order to design a strain of *Escherichia coli* that produces 1-propanol, the authors used a simple mass- and redox-balanced stoichiometric model and FBA to compare the theoretical yields of three routes to 1-propanol: the native threonine pathway, the non-native citramalate pathway, and a synergistic employment of both

pathways (Fig. 1.1a). The calculations revealed that the two pathways together have a theoretical yield of 1.33 mol 1-propanol per mol glucose, 33% higher than either individual pathway. Indeed, the authors constructed production strains for all three 1-propanol routes, and the synergistic employment of both pathways had the highest observed yield, ~30% greater yield than the citramalate pathway and ~55% greater yield than the threonine pathway.

Another class of COBRA predictions uses the biomass objective function—a representation of all the metabolite demands required for cell growth (Feist et al. 2010)—to predict how gene deletions will affect cellular phenotypes. For instance, a number of COBRA algorithms have been developed to identify groups of gene knockouts that are predicted to change the fermentation profile of a cell when growing at a maximum growth rate, a characteristic known as *growth-coupling*; e.g. OptKnock (Burgard, Pharkya, and Maranas 2003), OptGene (Patil et al. 2005), RobustKnock (Tepper and Shlomi 2010), OptSwap (King and Feist 2013), for further discussion see (Lewis, Nagarajan, and Palsson 2012). However, only a few studies (Yim et al. 2011; Fong et al. 2005) have tested the validity of the predictions of these algorithms, and thus their significance to systems metabolic engineering is largely untested.

A similar COBRA method, called SIMUP, was recently shown to have direct usefulness for systems metabolic engineering (Gawand et al. 2013). The SIMUP method predicted groups of gene knockouts that forced co-utilization of two substrates. Lignocellosic biomass is typically hydrolyzed into a mixture of glucose and xylose, but industrial organisms preferentially consume glucose over xylose. Thus, the SIMUP algorithm was used to identify a group of three gene knockouts that disable upper glycolysis and part of the pentose phosphate pathway so that both glucose and xylose

consumption are necessary for rapid growth (Fig. 1.1b). The results of the simulations were tested *in vitro*, and it was found that SIMUP accurately predicted strain designs that co-utilized these substrates, albeit with substrate uptake rates lower than the wild type.

Some of the most successful COBRA methods for predicting modifications for systems metabolic engineering use empirical data (e.g. omics data) to generate a reference state for a host strain, then find the set of up-regulations, down-regulations, and/or knockouts necessary to increase the yield of a target molecule; e.g. MOMA (Segre, Vitkup, and Church 2002), OptForce (Ranganathan, Suthers, and Maranas 2010), FSEOF (Choi et al. 2010). In a recent example of this approach, overproduction of C14–C16 fatty acids was engineered in *E. coli* by implementing the predicted gene up-regulations and deletions (Ranganathan et al. 2012; Fig. 1.1c). Similar strategies have been employed to produce polylactatic acid (up to 11% yield by weight from glucose using metabolic flux analysis and MOMA; Jung et al. 2010), malonyl-CoA (4-fold increase from wild type using OptForce; Xu et al. 2011), lycopene (over 8-fold increase from control using FSEOF and MOMA; Choi et al. 2010), the antibiotic actinorhodin (52-fold increase from wild type using FSEOF; Kim et al. 2014), and the recombinant protein human Superoxide dismutase (up to 1.4-fold increase from control using FSEOF and MOMA; Nocon et al. 2014).

Thus, COBRA methods have been used to great effect in optimizing microbial cell factories, but the prediction types can be expanded. The tide is shifting toward a much greater range of predictive capabilities through use of more complex models that add additional constraints to the systematic analysis.

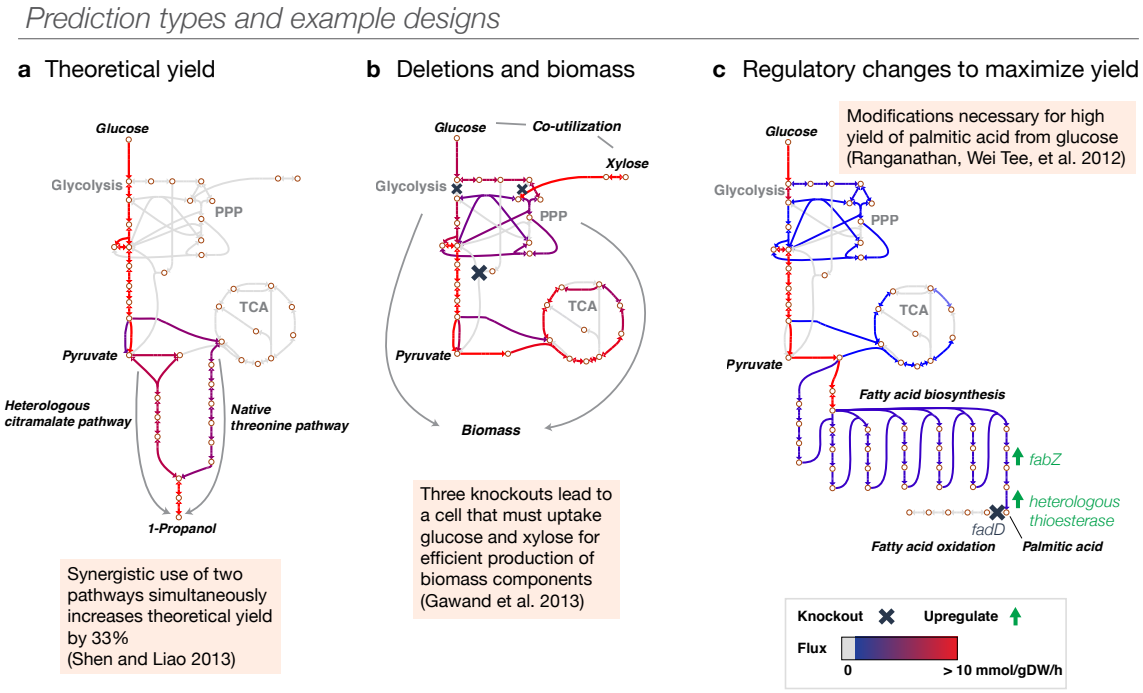

Figure 1.1: **Three types of COBRA predictions have been successfully implemented for systems metabolic engineering.** (a) Theoretical yield. As an example, theoretical yield was maximized through synergistic use of two production pathways (Shen and Liao 2013). (b) Gene deletions and biomass. As an example, the SIMUP algorithm identifies three gene knockouts that force co-utilization of glucose and xylose to achieve maximum growth (Gawand et al. 2013). (c) Regulatory changes to increase yield. As an example, gene up-regulations and deletions were used to increase the production of fatty acids (Ranganathan et al. 2012). Figures were generated using Escher (King et al. 2015a).

1.3 Introducing the Thesis

In this dissertation, I develop systems biology methods for the optimization of microbial cell factories. The *design-build-test-learn* cycle is widely used to conceptualize the optimization process, so the following chapters are organized around this cycle. Chapter 2 introduces BiGG Models, a public database of standardized genome-scale models of metabolism. Together, they represent a parts list that can be used during the *build* step in strain optimization when new reactions (via genes) are added to a strain. Chapter 3 introduces Escher, a web-based application for visualizing omics

data in the context of the metabolic network. During *test*, analytical data is collected and analyzed, and Escher provides a high-level view of many data types (including transcriptomics, fluxomics, and metabolomics) in the context of the metabolic network. Chapters 4 and 5 describe algorithms for optimizing cofactor usage in metabolic networks, specifically through modifying NADH/NADPH specificity of oxidoreductase enzymes. These methods contribute to a general picture of the importance of cofactor balance for product yields with NADPH-dependent pathways, and they demonstrate the value of COBRA methods in the *learn* step. Chapter 6 uses literature mining to assess the predictive strength of genome-scale metabolic models of *E. coli*. This study provides support for the use of COBRA methods and *de novo* predictions during *design*. Chapter 7 concludes the dissertation and offers an outlook on next-generation genome-scale modeling techniques for optimizing microbial cell factories.

Chapter 1 is adapted from published manuscripts: King, Z. A., Lloyd, C. J., Feist, A. M., and Palsson, B. O. (2015b). "Next-generation genome-scale models for metabolic engineering". In: *Curr. Opin. Biotechnol.* 35, pp. 23–29. DOI: 10.1016/j.copbio.2014.12.016. The dissertation author was the primary author of the review. Bordbar, A., Monk, J. M., King, Z. A., and Palsson, B. O. (2014a). "Constraint-based models predict metabolic and associated cellular functions". In: *Nat. Rev. Genet.* 15.2, pp. 107–120. DOI: 10.1038/nrg3643. The dissertation author was one of the authors of the review.

Chapter 2

BiGG Models: A platform for integrating, standardizing, and sharing genome-scale models

2.1 Introduction

Biological knowledge bases must evolve to keep pace with the incredible progress in experimental biology. Methods for collecting genome-scale 'omics' data have been widely adopted, and the resulting datasets can be difficult to understand, especially when multiple data types are collected in the same experiment (Dolinski and Troyanskaya 2015). These challenges are emblematic of the larger efforts to deal with and capitalize on Big Data (Margolis et al. 2014). A biological knowledge base can serve as a framework for interpreting omics data by providing biological context for each measurement. For this to work, the knowledge base must contain an accurate, genome-scale representation of the organism; it must use unique identifiers and links

to existing databases so that scientists can easily align data; and it must describe the relationships between between biological networks so that distinct omics data types can be connected during analysis.

Knowledge bases are widely available and commonly used by biologists. The most extensive pathway-oriented knowledge base is the Kyoto Encyclopedia of Genes and Genomes (KEGG) that contains 15 related databases with information on 3,982 organisms (Kanehisa et al. 2014). In contrast, BioCyc is best known for seven highly-curated, multi-scale knowledge bases for model organisms that include *Escherichia coli*, *Bacillus subtilis*, and *Homo sapiens* (Caspi et al. 2014). Similar databases are available for model organisms such as yeast (Costanzo et al. 2014) and mouse (Eppig et al. 2014). These knowledge bases are all generated through a combination of bioinformatics (e.g. identifying a gene function by sequence homology) and manual curation (e.g. assigning a pathway name to a set of gene products). A complimentary approach is to build a knowledge base around a mathematical model of an organism, and this approach has certain advantages.

Genome-scale metabolic models (GEMs) are mathematically-structured knowledge bases. They contain descriptions of all the biochemical reactions, metabolites, and genes in metabolism for a specific organism—a *Bi*ochemical, *G*enetic, and *G*enomic (BiGG) knowledge base (Feist et al. 2009). Additionally, GEMs contain descriptions of the biophysical constraints on metabolic systems (nutrient uptake, oxygen availability, reaction stoichiometry, and reversibility) (Feist et al. 2009). GEMs can be used to predict cellular phenotypes (Bordbar et al. 2014a), contextualize omics data (Lewis, Nagarajan, and Palsson 2012; Lewis and Abdel-Haleem 2013; Hyduke, Lewis, and Palsson 2013), design cell factories (King et al. 2015b; Machado and Herrgård 2015),

and understand evolutionary trajectories (McCloskey, Palsson, and Feist 2013). A further advantage of mathematical structure is that the accuracy of GEMs increases continuously through comparison with experimental data (Reed et al. 2006a).

GEMs have not generally been available through a centralized resource with reliable standards. A workflow for building high-quality GEMs has been described (Thiele and Palsson 2010), but this process is complex and the quality of published GEMs is highly variable (Monk, Nogales, and Palsson 2014). A number of challenges still exist in the reconstruction process. The workflow recommends that metabolites be linked against existing databases (Thiele and Palsson 2010), but this is not a formal requirement in the models. Visualization of GEMs has been an important feature since the first models were reconstructed, but accessible tools for visualizing GEMs have also been lacking. These challenges have been addressed in the past through unwritten "best practices" in individual labs, but they represent a general challenge when models from different labs are to be collected or compared.

The first BiGG knowledge base was published in 2010, and it addressed some of these challenges for a specific set of 10 GEMs generated at the Systems Biology Research Group at the University of California, San Diego (Schellenberger et al. 2010). With BiGG, reaction identifiers, metabolite identifiers, and pathway maps were formalized in a database, using the software package SimPheny (Genomatica, San Diego CA), and shared on a public website. In BiGG, users could export models in the SBML format (Hucka et al. 2003), visualize metabolic pathways, and search the database. BiGG was a widely-used community resource that was incorporated into other applications (Ganter et al. 2013; Wishart et al. 2013; Gavai et al. 2015; Kumar, Suthers, and Maranas 2012), but it was never extended to be a general resource for

storing large numbers of GEMs or for building new GEMs (Table 2.1. MetRxn is another curated and interactive database of GEMs (Kumar, Suthers, and Maranas 2012), but it focuses more on identifying metabolite structures and performing model comparisons.

Table 2.1: **Comparing BiGG (2010) to BiGG Models (2015).**

BiGG (2010)	BiGG Models (2015)
10 models	77 models
Pathway visualization with SVG	Pathway visualization with Escher
Export to SBML Level 2	Export to SBML Level 3, MAT, and JSON
	Standardized identifiers for metabolites, reactions, and genes
	Public, documented API
	Gene identifiers linked to NCBI RefSeq genome annotation

BiGG Models is a completely redesigned knowledge base that currently includes 77 GEMs linked to 71 genome annotations. It includes a workflow for integrating models built at different times so models can be improved and exported with the latest standards. Model, reaction, metabolite, compartment, and gene identifiers are standardized, and pathway maps are included using the Escher pathway visualization library (King et al. 2015a). A website allows users to search, browse, and visualize the networks. Models can be exported in various community standard formats (Dräger and Palsson 2014). BiGG Models has a comprehensive application programming interface (API) for accessing and building upon BiGG. With these features, BiGG Models is a platform for integrating, standardizing, and sharing knowledge of metabolism.

2.2 Knowledge base content

BiGG Models is built around a set of high-quality published GEMs. The original models were collected from the supplemental data provided with their publications. Only minimal changes to the models were made (changes are listed in Supplemental Data S6), and the updated models were validated by comparing content and predictions to the published models. These models were aligned in BiGG Models so that they share a common list of reactions and metabolites ("universal" reactions and metabolites). Thus, any curation of general attributes like metabolite formulae will apply to all models in the knowledge base, and therefore also provide a standard for future genome-scale metabolic network reconstructions. A total of 77 GEMs are included in BiGG Models as of publication, and more will be added over time.

Genome annotations for the models were downloaded from the NCBI RefSeq database (Pruitt et al. 2014). In total, 71 genome annotations were identified for the GEMs in BiGG Models (a full list of models and genome annotations can be found in Supplemental Data S1).

Pathway maps are included in BiGG Models using the Escher visualization library (King et al. 2015a). Maps are currently available for the most widely used models in the database, and more maps are under construction. External database links for metabolites and genes have been included in the database. The external databases include KEGG (Kanehisa et al. 2014), MetaCyc (Caspi et al. 2014), Reactome (Croft et al. 2014), HMDB (Wishart et al. 2013), RCSB PDB (Rose et al. 2015), Model SEED (Henry et al. 2010), and Entrez Gene (Brown et al. 2014). Finally, compartment names are often missing from publicly available GEMs, so a list of compartment names was collected in BiGG Models (Supplemental Data S3).

BiGG Models integrates these models, genome annotations, pathway maps, and additional data in order to provide a set of gold-standard models and a knowledge base of shared biological components (Fig. 2.1). This knowledge base can then be used for analyzing omics data related to reactions (fluxomics), genes (genomics, transcriptomics, proteomics), and metabolites (metabolomics). Recent work has extended GEMs to encompass gene expression (King et al. 2015b; O'Brien and Palsson 2015), and eventually these *ME-models* can be included in BiGG, where they can serve as a framework for analyzing protein-associated datasets (proteomics).

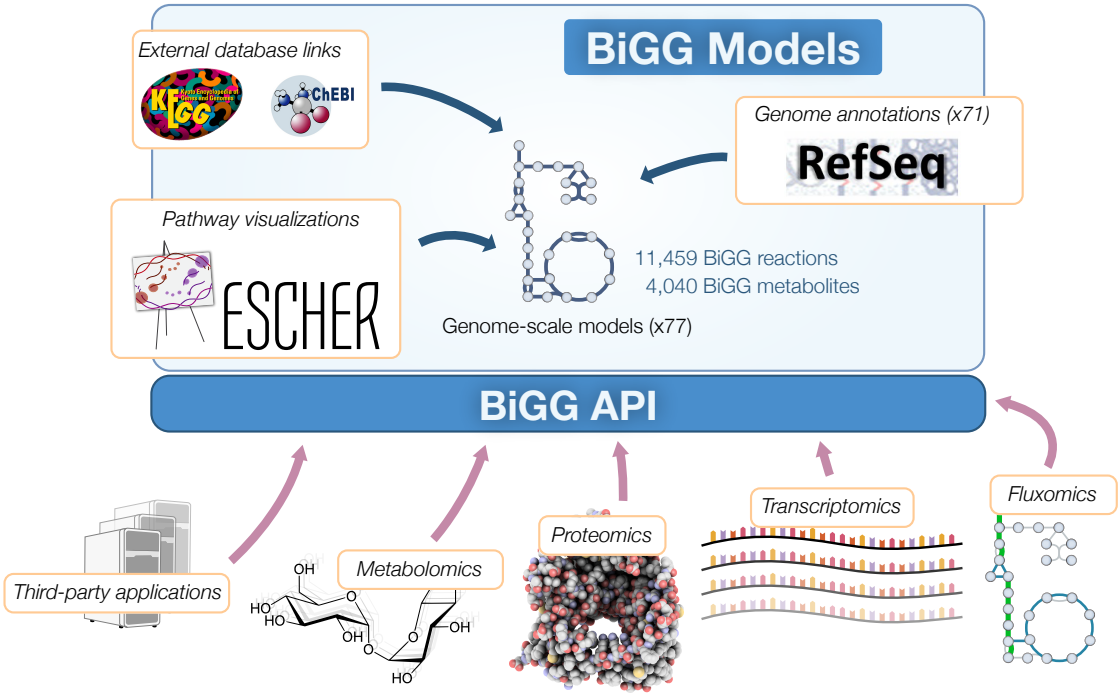

Figure 2.1: **BiGG Models content.** BiGG Models is built around a collection of 77 GEMs. The GEMs are integrated into a single database with shared reaction and metabolite identifiers. This core database is enriched with external database links, Escher pathway maps (King et al. 2015a), and genome annotations. As a result, BiGG Models is a resource that can be used to analyze and contextualize many omics data types.

2.3 Getting started with BiGG Models

2.3.1 BiGG website

BiGG Models has a user-friendly website (http://bigg.ucsd.edu) for browsing, searching, visualizing, and downloading content. The homepage for BiGG Models includes a search bar for finding models, reactions, metabolites, and genes for a search term (Fig. 2.2). It also includes links to lists of all the models, universal metabolites, and universal reactions in the knowledge base.

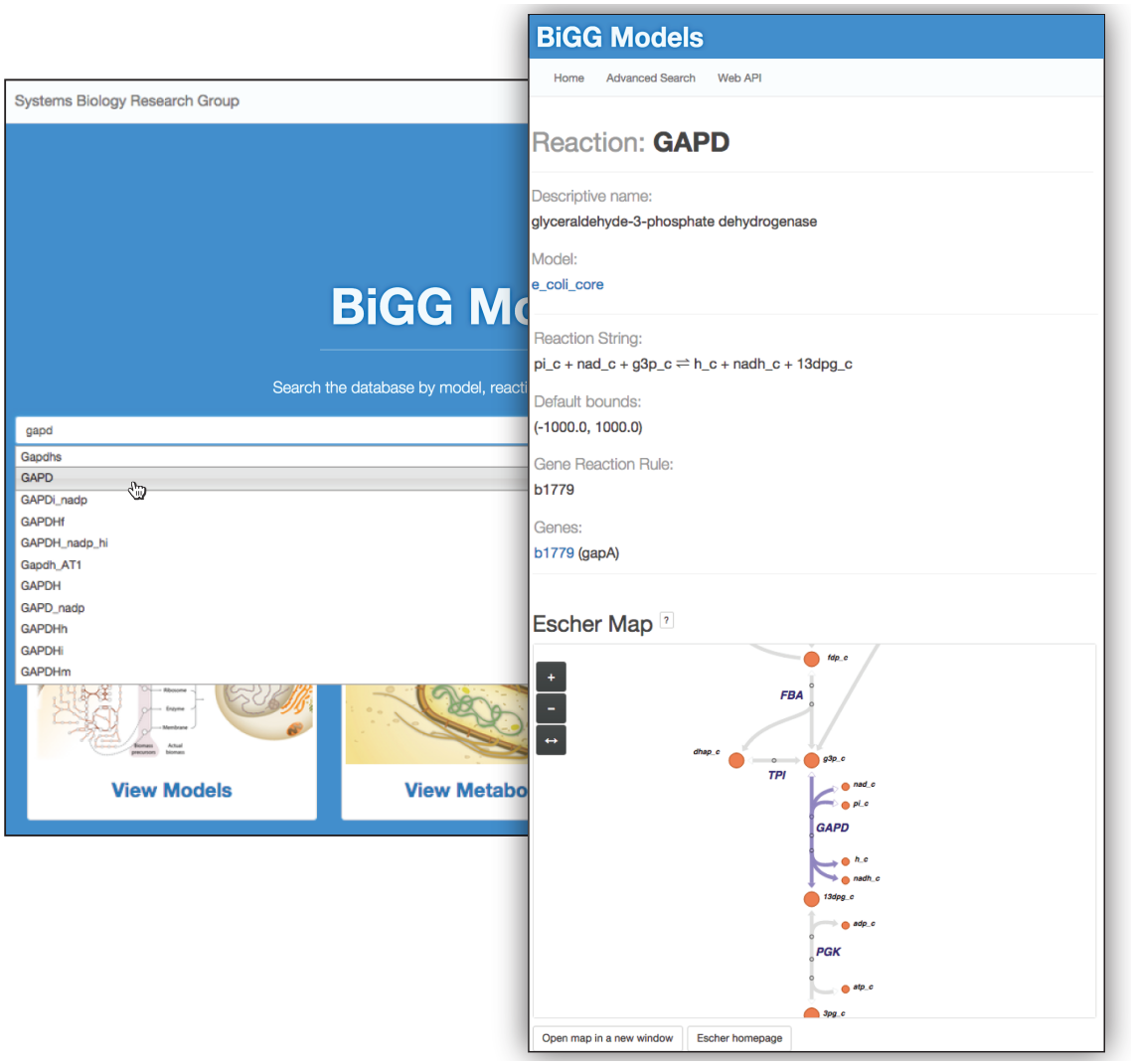

Figure 2.2: **The BiGG Models homepage.** The central text box allows users to search for pages in BiGG Models, including models and their reactions, metabolites, and genes. Convenient links to the most popular pages about models, metabolites, and reactions can be found below the search box. General information about BiGG Models can be found by clicking **About** at the top of the page.

The page for a BiGG model provides an overview of the model and options for downloading the model in community standard formats. The model page also provides a link to the corresponding genome annotation. Reactions and metabolites can be viewed on model-specific pages and universal pages, reflecting the organization of the knowledge base. Model-specific reaction pages include the stoichiometry of the

reaction, the reaction bounds within the GEM, and the gene-reaction rule for the reaction with links to the related genes. Metabolite pages show the molecular formula for the metabolite and provide external database links. Pages for each gene provide the position of the gene in the chromosome and a link to a page for the genome that contains the chromosome.

The website also includes pathway visualization, advanced search, and documentation of the web API. Model, reaction, and metabolite pages that have associated pathway visualizations include an embedded, interactive pathway map viewer powered by Escher (an example can be seen on the page http://bigg.ucsd.edu/models/e_coli_core/reactions/GAPD, Fig. 2.2). An **Advanced search** page gives users the option to search for metabolites by external identifier (e.g. KEGG ID) and to find BiGG pages for a specific model. And the **Web API** page has information and examples for using the web API.

2.3.2 Using BiGG Models for COBRA modeling

The GEMs in BiGG Models are can be used for modeling metabolism, interpreting omics data, visualizing metabolic phenotypes, and more (Lewis, Nagarajan, and Palsson 2012; Lewis and Abdel-Haleem 2013; King et al. 2015a). BiGG Models makes the models more accessible to users with a variety of options for browsing and downloading them. The GEMs in BiGG Models can be analyzed using the many available *Co*nstraint-*B*ased *R*econstruction and *A*nalysis (COBRA) methods (Bordbar et al. 2014a; Lewis, Nagarajan, and Palsson 2012; Ebrahim et al. 2013) or any software that reads SBML.

To use a model for COBRA analysis, first download the model in the appropriate

format. The most general and most highly annotated format is SBML (SBML Level 3 with FBC), which includes all the content of the model plus the external database links, compartment names, and license information. This is the preferred format for analysis in COBRApy (Ebrahim et al. 2013) and the 280+ existing tools can read SBML files (http://sbml.org/SBML_Software_Guide). Models are available in MATLAB MAT format for analysis with the MATLAB COBRA toolbox (Schellenberger et al. 2011) and the the JavaScript Object Notation (JSON) format for building visualizations with Escher (King et al. 2015a). With the BiGG Models API, software tools can also access the complete contents of these models programmatically.

2.3.3 Using BiGG Models for building GEMs

BiGG Models provides a set of identifiers and metabolic components that can be used for new models, as well as a set of standards for defining new IDs (Supplemental Data S4). The BiGG Models API can be used to directly access these identifiers using tools developed for building models.

Using BiGG for new reconstructions provides a number of benefits. Using BiGG IDs in a new model means that the model can easily be compared to the set of existing models that already in this knowledge base. BiGG Models, COBRApy (Ebrahim et al. 2013), and Escher (King et al. 2015a) can be deployed in other research labs, and using BiGG Models as a guide for new reconstructions will mean that the new reconstruction is compatible with these tools. Specifically, the Escher maps in BiGG Models can be adapted to new organisms if the new models utilize the same identifiers (see the Escher documentation for more details at https://escher.readthedocs.org). Finally, BiGG Models can be extended to include models built in other research groups, as

long as they conform to the standards set out with BiGG Models.

2.3.4 Accessing the API

BiGG Models includes a fully featured web API (Fig. 2.3). The API can be accessed from any programming language that supports Hypertext Transfer Protocol (HTTP) requests. Thus, BiGG can be used as a service from other applications; for example, a metabolic modeling toolkit could provide direct access to BiGG models *via* the BiGG API. The web API returns JSON formatted data. In the case of an error, an appropriate HTTP error code is returned. The full documentation of the API is provided on the **Web API** page of the BiGG website.

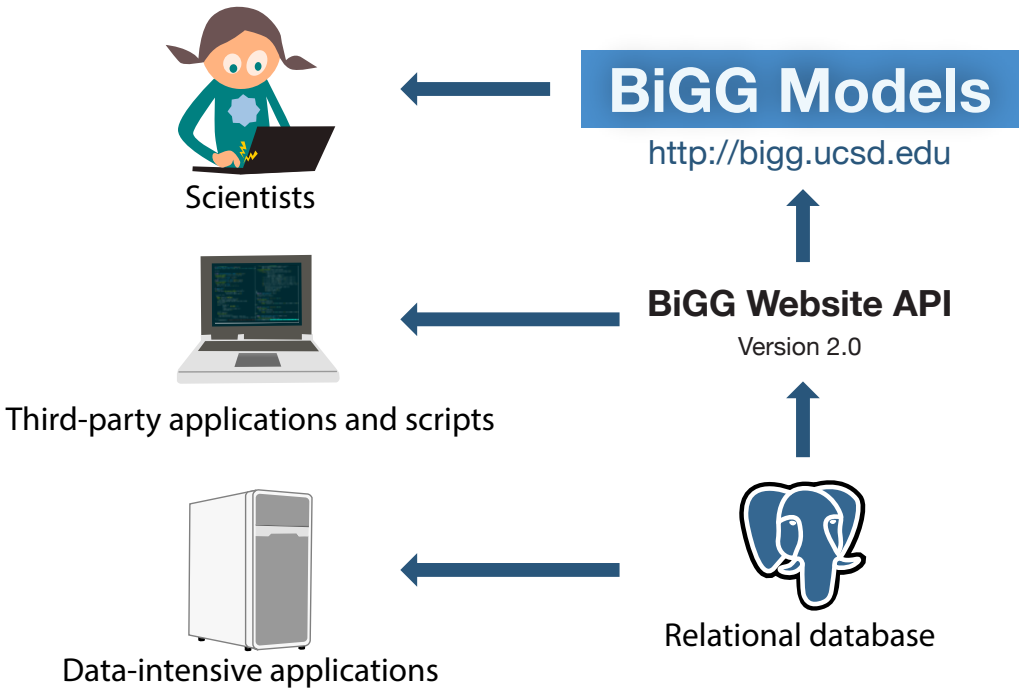

Figure 2.3: **Accessing BiGG.** BiGG Models has a user-friendly website for browsing and searching the knowledge base. The knowledge base can also be accessed programmatically using the web API. For more data-intensive applications, it is possible to run a local version of the BiGG database.

As an example, a list of the models in BiGG Models can be retrieved with the

following HTTP request:

```
GET http://bigg.ucsd.edu/api/v2/models HTTP/1.1
```

This can be accomplished by visiting http://bigg.ucsd.edu/api/v2/models in a web browser. Many programming languages provide functions for accessing resources on the web. For example, in Python 2.7, the following script will load the data and decode the JSON formatted results:

```
import urllib2
import json
# Run the HTTP request
response = urllib2.urlopen("http://bigg.ucsd.edu/api/v2/
   models")
# Should return the success code 200
assert response.code == 200
# decode the body of the response, and parse the resulting
    JSON
models = json.loads(response.read().decode("utf8"))
# print number of models
print models["results_count"]
```

The specific models can then be accessed with follow-up requests. For example, an overview of the first model with BiGG ID `e_coli_core` can be accessed with a request to the URL http://bigg.ucsd.edu/api/v2/models/e_coli_core, and the full model can be downloaded with a request to the URL http://bigg.ucsd.edu/api/v2/models/e_coli_core/download.

With these tools in hand, developers can use the BiGG API to access any content in the knowledge base from analysis scripts, modeling tools, and web applications.

2.4 Implementation of standards

2.4.1 Loading genomes and GEMs

A workflow was developed for integrating models and genome annotations into a single, coherent database. This workflow reconciles any conflicting information, links genes from GEMs to genes in the genome annotations wherever possible, and constructs a single database that serves as the basis for BiGG Models. The workflow proceeds as follows. First, a database is initialized in PostgreSQL (PostgreSQL Global Development Group), a high-performance, open-source, relational database. A total of 24 tables are necessary to store the content in BiGG Models (Fig. S2).

For each genome annotation, genes are loaded into the database with all available identifiers and external database references (Fig. S5). Genome annotations are used to fill the *genome*, *chromosome*, *genome region*, and *gene* tables (Fig. S2). A single genome can have multiple chromosomes, and genes in each chromosome are loaded from individual files in the Genbank file format (Benson et al. 2014). The positions of the genes are recorded, and the organism and taxon ID are stored for each genome annotation.

Next, GEMs are loaded into database by reaction, metabolite, and gene (Fig. S5). Efforts are made to separate general information about biological components from model-specific information. The information about reactions and metabolites that is not specific to an organism or a model is considered universal, and BiGG Models represents this information in database tables for universal reactions and universal metabolites. Model-specific information is stored in database tables for model-specific reactions and model-specific metabolites (Fig. S5). Analogously, information about

genes is separated into the annotation-specific gene table in the database and the model-specific gene table. Multiple GEMs may reference a genome annotation; thus, annotation-specific genes can be shared between models.

A further separation is made between metabolites (called *components* in the database tables), which can exist in any cellular compartment, and *compartmentalized metabolites*, which have a specific compartment and participate in reactions.

All the data in BiGG Models that are not found in the GEMs and the genome annotations are arranged in six preference files (Supplemental Data S3).

2.4.2 BiGG identifiers

BiGG Models uses a set of identifiers—BiGG IDs—that are unique, well-defined, human-readable, and memorable (Table 1). BiGG IDs have been used to build GEMs in many research groups, and they were available with BiGG 1, but problems have appeared with the quality and consistency of BiGG IDs. With BiGG Models, the goal is to provide a single source of correct BiGG IDs that are easy to discover and for other applications.

Now, BiGG IDs follow a simple, clear specification (Supplemental data S4). Reactions, metabolites, and genes are assigned unique alphanumeric identifiers, based on the IDs already found in most published GEMs (Fig. 2.4). Metabolites in compartments include a one or two letter compartment code (lowercase letters), and tissue-specific metabolites have a one or two letter tissue code (capital letters). BiGG IDs are now available in the MIRIAM registry with URIs from the identifiers.org service (Juty, Le Novère, and Laibe 2012) (Table 2.2).

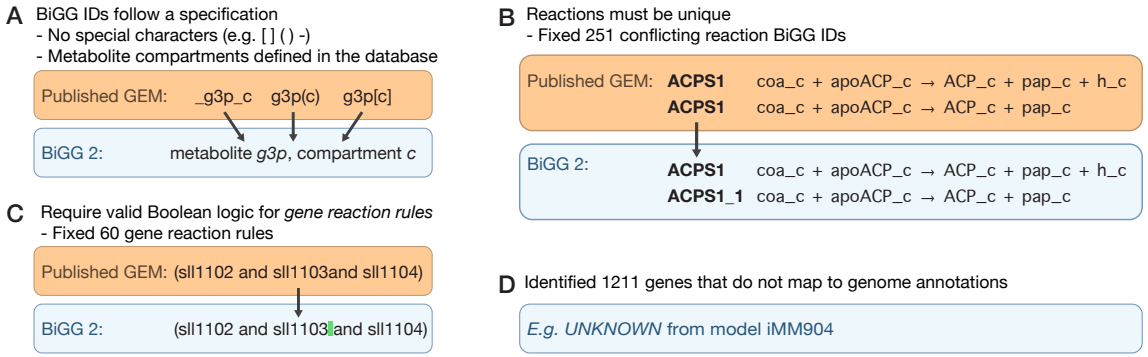

Figure 2.4: **Standardizing GEMs.** In order to standardize the GEMs in BiGG Models, a number of changes had to be made to the models. (A) First, metabolite and reaction IDs were standardized by removing extraneous characters and using a single format for referring to compartments. (B) In cases where the same reaction ID referred to different reactions, one of the reactions received a new identifier. (C) Invalid gene reaction rules were manually corrected. (D) All the genes that did not map to a genome annotation were recorded for future updates to both the GEMs and the genome annotations.

For compatibility with existing tools that do not allow numbers at the beginning of identifiers (e.g. SBML), a BiGG ID can be prefixed with `R_` for reactions and `M_` for metabolites. Unprefixed BiGG IDs are used on the BiGG website, in Escher, and in COBRApy, and prefixed BiGG IDs are automatically generated in exported SBML files.

Genes in BiGG Models have identifiers that are unique to a specific genome annotation. Thus, genes are referenced by their locus IDs in the genome annotation. Genes that do not map to a genome annotation retain the ID from the original model file. Gene BiGG IDs are prefixed with `G_` in exported SBML files, and unprefixed gene IDs are used in BiGG Models, Escher, and COBRApy.

Table 2.2: **BiGG Identifiers.** BiGG Models contains unique identifiers for models, reactions, metabolites, compartments, and genes. With the exception of genes, these elements are registered with MIRIAM.

Type	Example	BiGG ID	MIRIAM URI
Model	Latest *E. coli* model (Orth et al. 2011)	*i*JO1366	http://identifiers.org/bigg.model/iJO1366
Reaction	Glyceraldehyde-3-phosphate dehydrogenase	GAPD	http://identifiers.org/bigg.reaction/GAPD
Metabolite	Glyceraldehyde-3-phosphate	g3p	http://identifiers.org/bigg.metabolite/g3p
Compartment	Cytsosol	c	http://identifiers.org/bigg.compartment/c
Gene	*E. coli* gapA	b1779	

2.4.3 ModelPolisher

BiGG Models supports the latest SBML standard Level 3 Version 1 with FBC version 2 (Rodriguez et al. 2015). To generate compliant and highly-annotated files, the ModelPolisher application was developed (https://github.com/SBRG/ModelPolisher). SBML models are first generated using COBRApy (Ebrahim et al. 2013), then ModelPolisher inserts MIRIAM annotations and adds specific terms from the Systems Biology Ontology (SBO) (Courtot et al. 2011) to individual model components in order to better point out their role. The SBO is a collection of controlled vocabulary terms with clear definitions and references. For the annotation of BiGG models, the following new terms have been added to SBO: flux bound (SBO 625, SBO 626), exchange reaction (SBO 627), demand reaction (SBO 628), biomass reaction (SBO 629), and ATP maintenance (SBO 630). The resulting SBML files are available on the model pages of the BiGG Models website, and an overview of the model content can be seen by loading a downloaded SBML file in a web browser.

2.4.4 Design and implementation

BiGG Models is a modular application composed of a relational database, a web API, and a website (Fig. 2.3). It is primarily written in Python 2.7, SQL, JavaScript, HTML, and CSS.

BiGG Models is built with PostgreSQL 9.4.4 (PostgreSQL Global Development Group, http://www.postgresql.org/). The SQLAlchemy (http://www.sqlalchemy.org/) object relational mapper (ORM) is used to load and query from the database. The website and API servers are implemented with Tornado (http://www.tornadoweb.org/en/stable/). The website retrieves data through the same web API provided to users; thus, all the content in the website is available through the API. A number of other libraries were essential for building BiGG Models, including Jinja2 (http://jinja.pocoo.org/), JQuery (https://jquery.com), and TableSorter (https://mottie.github.io/tablesorter/docs/).

2.5 Conclusion

BiGG Models is a fully redesigned platform for integrating, standardizing, and sharing GEMs of metabolism. The knowledge base currently integrates the metabolic content from 77 GEMs and 71 genome annotations, and users can search and explore the knowledge base with the BiGG website. A web API is available for building new applications that extend the capabilities of BiGG. The result of these features is a knowledge base that can be used to understand a huge variety of experimental data.

BiGG is free for academic and non-profit use so that the community can easily use and extend the knowledge base. The BiGG Models source code is available on

GitHub (https://github.com/SBRG/bigg_models). If a user finds a model error or a website bug using the BiGG Models website, it is possible to submit a report to the maintainers so this issue can be resolved. Each page for a reaction, gene, metabolite, or model includes a form to submit such a report, along with instructions. Website bugs can be fixed with future software releases. Model issues, in contrast, cannot be immediately fixed because BiGG is meant to present GEMs that are mathematically equivalent to the published models (though identifiers have been modified). Therefore, model issues will be collected for future updates to the GEM for that organism.

BiGG Models will continue to be developed to meet the needs of experimental and computational biologists. New visualization tools and model analysis features are in the works. The next generation of models can eventually be included in BiGG; these models incorporate expression networks, increased spatial resolution, regulation, and protein structures into GEMs (Bordbar et al. 2014a; King et al. 2015b; O'Brien and Palsson 2015). Plans for future BiGG releases will be driven by ongoing feedback from the users of the BiGG Models knowledge base.

2.6 Availability and requirements

BiGG Models is freely available online for academic and non-profit use at http://bigg.ucsd.edu, and a JavaScript-enabled browser is required to access certain features. The requirements for viewing Escher maps can be found on the Escher website (https://escher.github.io). Installation of an independent system requires Python 2.7 and PostgreSQL 9.4.4 or later.

Chapter 2 is a reprint of a published manuscript: King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O., and

Lewis, N. E. (2016). “BiGG Models: A platform for integrating, standardizing and sharing genome-scale models”. In: *Nucleic Acids Res.* 44.D1, pp. D515–22. DOI: 10.1093/nar/gkv1049. The dissertation author was the primary author of the paper and was responsible for the research.

Chapter 3

Escher: A web application for building, sharing, and embedding data-rich visualizations of biological pathways

3.1 Introduction

The behavior of an organism emerges from the complex interactions between genes, proteins, reactions, and metabolites. With next-generation sequencing and various "omics" technologies, it is now possible to rapidly and comprehensively measure these components and interactions. These technologies have transformed the scientific process over the past decade. Data acquisition is substantially easier, but data analysis is increasingly becoming the primary bottleneck to discovery. To address the analysis bottleneck, there has been a demand for data visualization tools to complement

statistical and modeling methods.

Biological visualizations often fall into categories characterized by biological scale, and the style of a visualization reflects the type of information at that scale. Three-dimensional objects are often used for representing protein structures (Arnold et al. 2006; Herráez 2006), one-dimensional tracks for genome sequences (Skinner et al. 2009; Karolchik et al. 2014), force-directed graphs for interaction networks (Smoot et al. 2011), trees for phylogenetic relationships (Letunic and Bork 2007; Huson et al. 2007). And, finally, two-dimensional *pathway maps* have long been a popular visual representation of metabolic pathways and other biological pathways. For each type of visualization, data can be associated with the biological components in the visualization. Visualizing data in this way contextualizes and enriches the dataset for scientists. Data-rich visualizations have been extremely valuable for viewing, interpreting, and communicating data.

A tool for visualizing pathway maps must satisfy a set of core features. The tool must (1) visually represent reactions and pathways clearly and in a way that is biochemically correct, (2) allow users to navigate and search through the visualization, (3) allow users to design and customize pathway maps, (4) allow users to represent diverse data types within the map using visual cues like size and color, (5) provide import and export features so that maps can be stored, shared, and exported to other tools, and (6) provide an application program interface (API) so the tool can be used within data analysis pipelines.

The existing tools that satisfy these core features are all desktop applications. Briefly, these tools include Omix (Droste, Nöh, and Wiechert 2013), Cytoscape (Smoot et al. 2011), CellDesigner (Funahashi et al. 2008), Vanted (Rohn et al. 2012) with the

SBGN-ED add-on (Czauderna, Klukas, and Schreiber 2010), VisAnt (Hu et al. 2013) and PathVisio (Kutmon et al. 2015). Desktop applications have many advantages over web applications; including speed, stability, and integration with the operating system, and these merits have made desktop applications more popular.

The advantages of web applications include rapid deployment (no need to download an application or browser plug-in), greater cross-platform compatibility (e.g. mobile devices), flexible sharing, collaborating, and embedding features, as well as easy application development. Recently, a critical mass of performance enhancements and new libraries has made web tools comparable to desktop tools for many applications.

A number of web-based tools exist for visualizing pathway maps: ArrayXPath (Chung et al. 2004), Pathway Projector (Kono et al. 2009), iPath2.0 (Yamada et al. 2011), WikiPathways (Kelder et al. 2012), Biographer (Krause et al. 2013), and the BioCyc pathway viewer (Latendresse and Karp 2011). However, none of these satisfy all the core features for a pathway map visualization tool.

One of the key differentiating features of a web application is that modern web browsers come with a built-in software development platform (often called the Developer Tools). This development platform includes a JavaScript shell for directly interacting with the web page runtime and a tool for inspecting and modifying every element in the web page document object model (DOM). Thus, *any user can locally modify any element on the page at any time.* If a web application is built on the DOM, then users can rapidly prototype new features and build extensions to the application while it is running. (A comparable feature is the extensibility of the EMACS editor, which can be extended while the editor is running (Stallman 1981). On the strength of this feature, EMACS has remained popular for 30 years.) To utilize this powerful

feature, one must use a visualization library that is based on the DOM, the most popular of which is Data-Driven Documents (D3) (Bostock, Ogievetsky, and Heer 2011).

Escher is a web application for visualizing pathway maps, and it is designed to be a fully featured pathway visualization tool that also harnesses all the advantages of the web. Escher has three key features that distinguish it from all existing pathway visualization tools, including the popular desktop applications. First, Escher makes building pathway maps fast and easy, using the information in datasets and genome-scale models to suggest pathways to the user—with this, pathway map design can be semi-automated. Second, Escher connects genes and enzymes to the reactions they catalyze, so that genomic data can be visualized in the context of the reaction network. We show how Escher can be used to visualize reaction data (metabolic fluxes), metabolite data (metabolomics), and genomic data (transcriptomic data), bridging the gap between these data types. Third, Escher uses the advantages of web technologies so that pathway maps can be adapted, extended, shared, and embedded. We illustrate the export and development features of Escher, including native support for scalable vector graphics (SVG) export, a downloadable tool for converting Escher maps to common standards for representing layouts, and application program interfaces (APIs) for developing new applications that extend the functionality of Escher.

3.2 Results

3.2.1 Building pathway maps

To build a pathway map, one first needs a source for the names, stoichiometries, and associated genes for each biochemical reaction in an organism. This information is provided by a *c*onstraint-*b*ased *r*econstruction and *a*nalysis (COBRA) model, a collection of all the reactions, metabolites, and genes known to exist in an organism (also called a genome-scale model, GEM, or constraint-based model, CBM) (Bordbar et al. 2014a). While COBRA models have generally focused on metabolism, the COBRA modeling approach can be applied to any biochemical reaction network (Bordbar et al. 2014a), so Escher could be used to visualize pathways like gene expression and membrane translocation, which are now being incorporated into COBRA models (King et al. 2015b; Liu et al. 2014; O'Brien et al. 2013).

The Escher interface is centered around a canvas for the pathway map (Fig. 3.1A). In the Escher Builder, a number of editing modes are available in the **Edit** menu; these include tools for navigating the map (Pan mode), selecting and modifying elements (Select mode), adding reactions (Add reaction mode), rotating the current selection (Rotate mode), and adding and editing text annotations (Text mode).

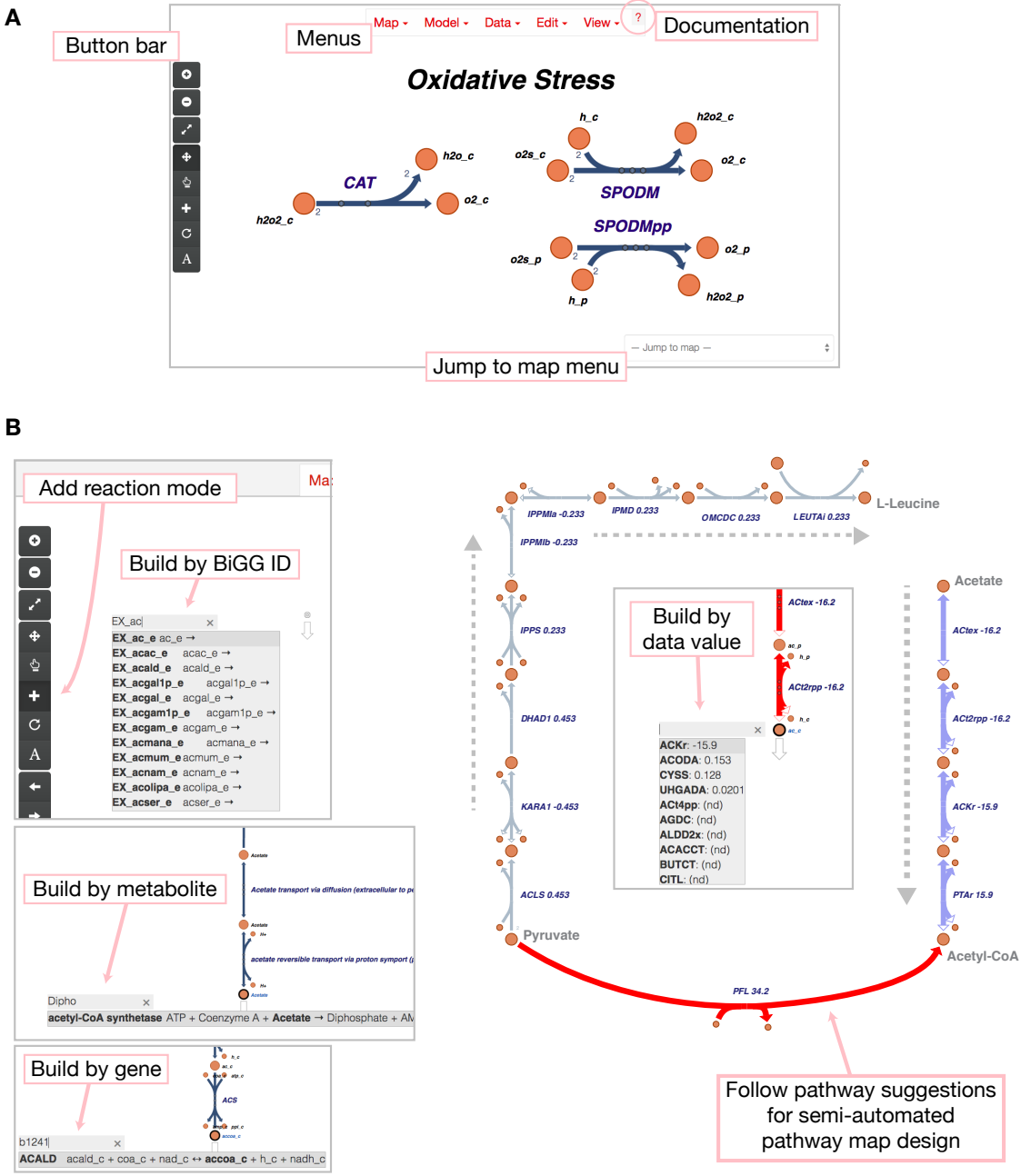

Figure 3.1: **The Escher interface.** A) The application includes a set of menus with a link to the documentation, a button bar for accessing common features, and a menu for jumping to maps that were built with the same model. B) To build pathway maps, enter the *Add reaction* mode using the **Edit** menu or the button bar. Click on the canvas or an existing metabolite to see a search menu. Reactions can be searched by reaction ID, by metabolite, and by gene. When a gene dataset or reaction dataset is loaded, suggestions appear for the reactions with the largest values in the dataset.

In *Add reaction mode*, a new pathway can be added to the canvas. Clicking on

the canvas or an existing metabolite opens the new reaction search box. The search box can find reactions with a number of queries: reaction identifiers (IDs) and display names, metabolite IDs and display names, and gene IDs and names (Fig. 3.1B). (IDs and names are based on those in the COBRA model.) If a reaction or gene dataset is loaded, then Escher provides suggestions of the next reaction to build, sorted by the data value for that reaction (Fig. 3.1B).

With this set of suggestions, a user can quickly build an Escher map based on previous knowledge of the organism or using the suggestion of a dataset. Data-driven map layout is also extremely useful for understanding an organism at the genome-scale—guided by the data, it is possible to find all the elements of a network that are, for example, highly upregulated without any bias toward well known pathways. To add the top suggested reaction, a user can simply press the Enter key. Thus, if a pathway is linear or has high values in a given dataset, then pressing Enter repeatedly will draw a linear pathway that is based entirely on the information in the data and the COBRA model. This process can be repeated to build perpendicular branches from metabolites in the pathway.

The Escher interface includes a general menu, a menu bar for accessing common functions, a tool for switching between maps, and a canvas containing the interactive pathway map (Fig. 3.1A). The **Map** and **Model** menus contain import and export functions for maps and COBRA models. The **Data** menu contains the data loading functions, and the **View** menu contains zoom options and access to the **Settings** page.

3.2.2 Visualizing data

Three types of data can be visualized on an Escher map: reaction data, metabolite data, and gene data. And Escher supports visualizing a single dataset, or visualizing the comparison of two datasets using a number of comparison functions (log, $\log_2$, and difference). The **Settings** page includes a detailed set of options for coloring and sizing elements based on statistical features of a dataset (min, max, quartiles, mean). Here, examples are provided for each data type, and the files required for recreating the visualizations are in the supplementary data.

Reaction data. To demonstrate the visualization of reaction fluxes, an *in silico* simulation of anaerobic growth was performed in the *Escherichia coli* COBRA model *i*JO1366 using parsimonious flux balance analysis (pFBA) (Lewis et al. 2010a; Orth et al. 2011). The Escher map of *i*JO1366 central metabolism was loaded (`iJO1366.Central Metabolism`) and the dataset (S1 Data) was loaded using the **Data>Load reaction data** function. (Datasets can be JavaScript Object Notation (JSON) or comma separated values (CSV) files, as described in the documentation.) Two settings were changed for this visualization: The absolute value of reaction data was visualized so that negative fluxes appear as large values, and the secondary nodes were hidden to simplify the visualization.

The resulting figure shows reaction fluxes for fermentation pathways (Fig. 3.2A). It was downloaded as a SVG image with the command **Map>Export as SVG**, and the text labels of the high flux reactions were made larger for the figure.

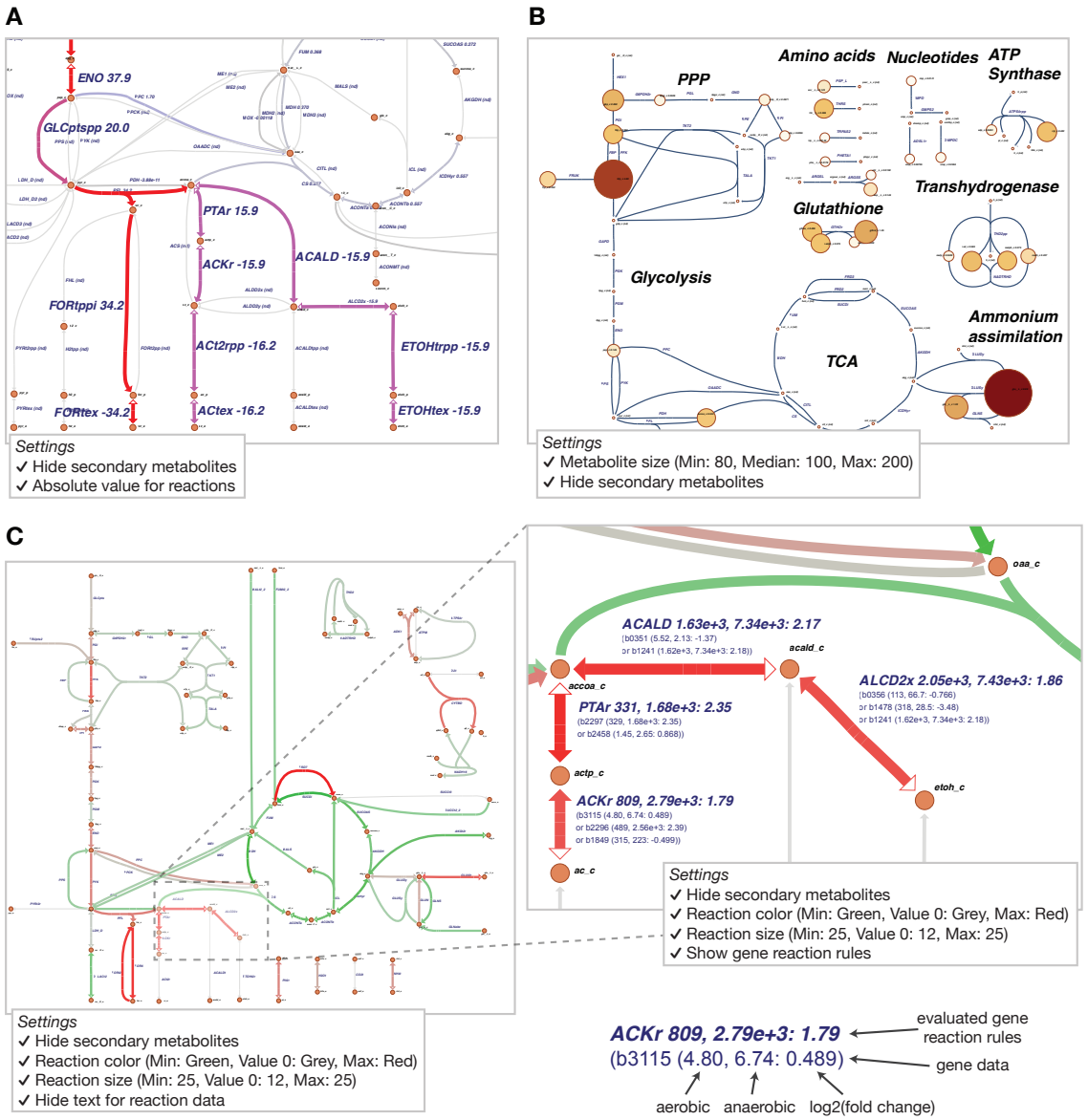

Figure 3.2: **Data visualization.** A) The results of an *in silico* flux simulation visualized on the reactions. B) Metabolomics data for *E. coli* aerobic growth visualized on the metabolites. C) RNA-Seq data showing the shift from aerobic to anaerobic conditions in *E. coli*. Green represents reactions downregulated in anaerobic growth and red represents gene upregulated in anaerobic growth, based on the $\log_2$ of the fold change.

Metabolite data. Metabolite concentrations are shown from a dataset recently reported by our research group (McCloskey et al. 2013), which were organized in a CSV file with metabolite BiGG IDs as keys (S2 Data). The example figure shows

aerobic metabolite concentrations on a modified map of *E. coli* central metabolism (S3 Data). To better identify metabolite concentration differences, the metabolite size was changed on the **Settings** page, and the secondary metabolites were hidden.

The resulting figure provides a high level view of the most abundant metabolites in the network aerobic growth of *E. coli* (Fig. 3.2B). It was downloaded as a SVG image with the command **Map>Export as SVG**, and the text annotations were made larger for the figure.

Gene data. To demonstrate the use of gene data on an Escher map, transcript abundances for aerobic and anaerobic growth of *E. coli* were calculated using RNA-Seq datasets from a recent publication (Bordbar et al. 2014b). The datasets were downloaded from the Gene Expression Omnibus (GEO) repository (Edgar, Domrachev, and Lash 2002) (accession number `GSE48324`), and fragments per kilobase of exon per million fragments mapped (FPKMs) were calculated using the Cufflinks functions cuffquant and cuffnorm (Trapnell et al. 2012), with appropriate parameters for the library type of the published data. These data were then collected, with locus tags as gene identifiers, in a single CSV file (D4 Data).

To connect genomic data with the reactions on an Escher map, Escher must consider which gene products are responsible for catalyzing each biochemical reaction. This association can be defined using Boolean *gene reaction rules* (also called gene-protein-reaction association (GPRs)) (Reed et al. 2003). When either of two enzymes can catalyze a reaction—as with isozymes—then these genes are connected with an OR rule. Escher *adds* the values of two genes connected with an OR rule. When two enzymes are required together for catalysis—as in an enzyme complex—these are connected with an AND rule. Escher can take the *mean* or the *minimum* of the two

values connected with an OR rule; this option is selected on the **Settings** page. For a comparison of two datasets, the gene reaction rules are evaluated for each dataset separately, then a comparison is made between the two resulting values (Fig. 3.2C).

The resulting figure shows the shift from aerobic to anaerobic conditions, where green reactions are downregulated anaerobically and red reactions are upregulated anaerobically (Fig. 3.2C). Escher shows the $\log_2$ of fold change between the conditions. However, Escher cannot yet display statistical significance for the datasets, so it should be paired with statistical tools (e.g. cuffdiff (Trapnell et al. 2012)).

3.3 Design and Implementation

JavaScript. Escher is a web application written primarily in JavaScript, using the libraries D3 (Bostock, Ogievetsky, and Heer 2011), and, optionally, JQuery (http://jquery.com), and Bootstrap (http://getbootstrap.com). The Escher JavaScript code can be compiled into a single JavaScript file, and a JavaScript API is available for interacting with and extending an Escher visualization (Fig. 3.3A). All layout, editing, import, and export features of Escher are included in the JavaScript library, and the default visual styles are defined in two cascading style sheets (CSS) files. The Escher website is built using the JavaScript API, and other web applications can be built on top of this library.

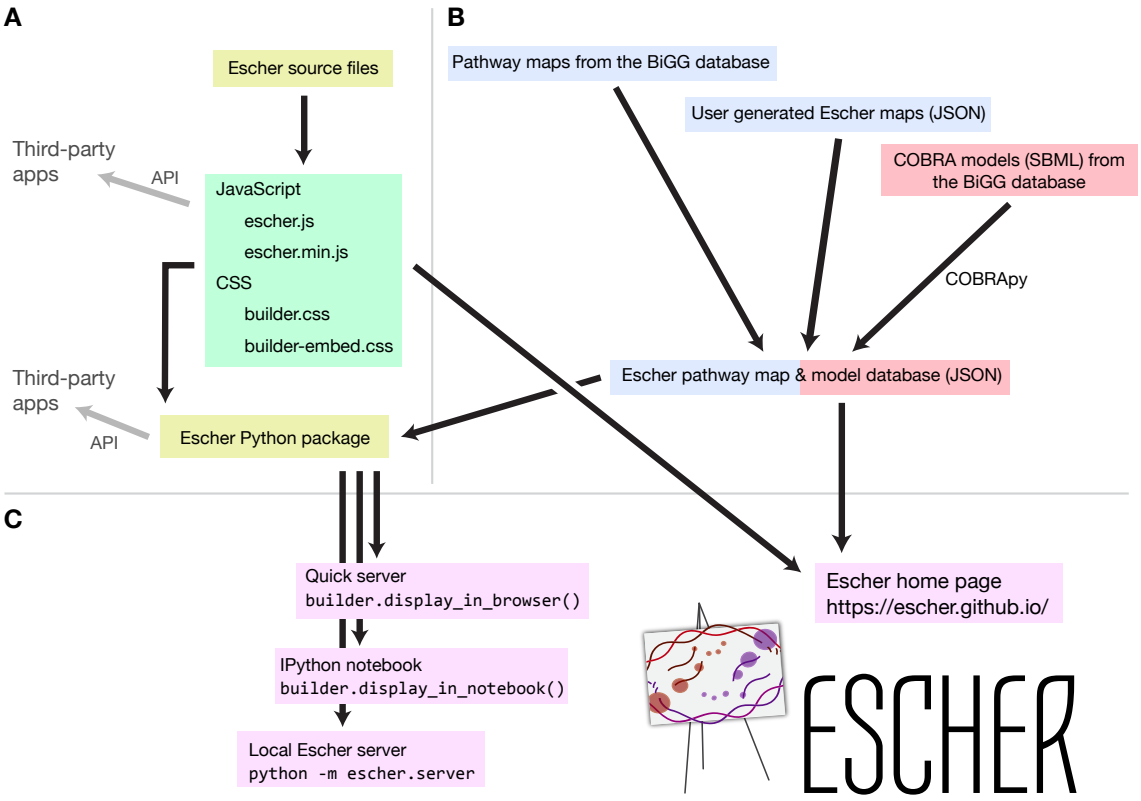

Figure 3.3: **The organization of the Escher project.** A) Escher source code can be compiled to a single JavaScript file (either minified or not minified) and two style sheets. The Python package is used to serve the Escher web application in various ways. APIs exist for both JavaScript and Python. B) Escher maps are generated from the BiGG database or built by users. COBRA models are generated using COBRApy. C) The Escher web application can be viewed on the Escher website, or, for local access, using various methods in the Python package.

Python. A Python package for Escher is also available (Fig. 3.3A), and this package includes a number of extra features: Access to Escher maps from Python terminals and IPython Notebooks, offline access to Escher, a local server with map and model caching, and a Python API for developing applications with these additional features. Accessing maps from Python and IPython Notebook allows Escher to be integrated directly with data analysis and modeling workflows. For example, within an IPython Notebook, the results of an *in silico* flux simulation can be applied to an Escher map, and the map will be embedded and shared with the notebook. Escher even supports NBViewer for

sharing static IPython Notebooks as websites (http://nbviewer.ipython.org).

Map and model database. Escher includes a database of pathway maps and genome-scale models. Pathway maps are currently available for a number of organisms, and new pathway maps will be continually added to the database from our group. The maps in the BiGG database are being converted to the new Escher format (Schellenberger et al. 2010). We also accept contributions from the community, and the method for submitting pathway maps is described in the documentation (S2 File).

JSON schema. Both Escher maps and COBRA models are stored as JavaScript Object Notation (JSON) files. JSON is a useful, plain-text format for storing nested data structures. For Escher maps, a JSON Schema has been defined (S1 File, see the schema file `escher/jsonschema/1-0-0`), and the schema can be enforced using the JSON Schema validators available in a number of languages (http://json-schema.org). Thus, Escher maps conform to a well-defined specification that can be generated by other tools and scripts.

Export. Escher represents biochemical reactions as transformations from a set of reactants to a set of products, and each reaction can be assigned enzymes using a Boolean *gene reaction rule*. Thus, Escher uses a well-defined representation of the biochemical network, but the scope of the Escher notation is much more specific than community standards such as Systems Biology Graphical Notation (SBGN) (Kitano et al. 2005) and Systems Biology Markup Language (SBML) with the layout extension (Hucka et al. 2003; Gauges et al. 2006; Dräger and Palsson 2014). Escher can be exported to both formats using the EscherConverter application (Fig. 3.4).

EscherConverter is written in Java™, and it is available as a standalone executable file (S3 File) that includes a graphical user interface with graph drawing capabilities and a command-line interface. Files can be opened through drag and drop or the file menu, and a history of up to 10 recent files is stored. Several user preferences allow flexible customization of the file conversion. The conversion to SBML and SBGN-ML (the XML implementation of SBGN) relies heavily on JSBML (Rodriguez et al. 2015) and libSBGN (Iersel et al. 2012).

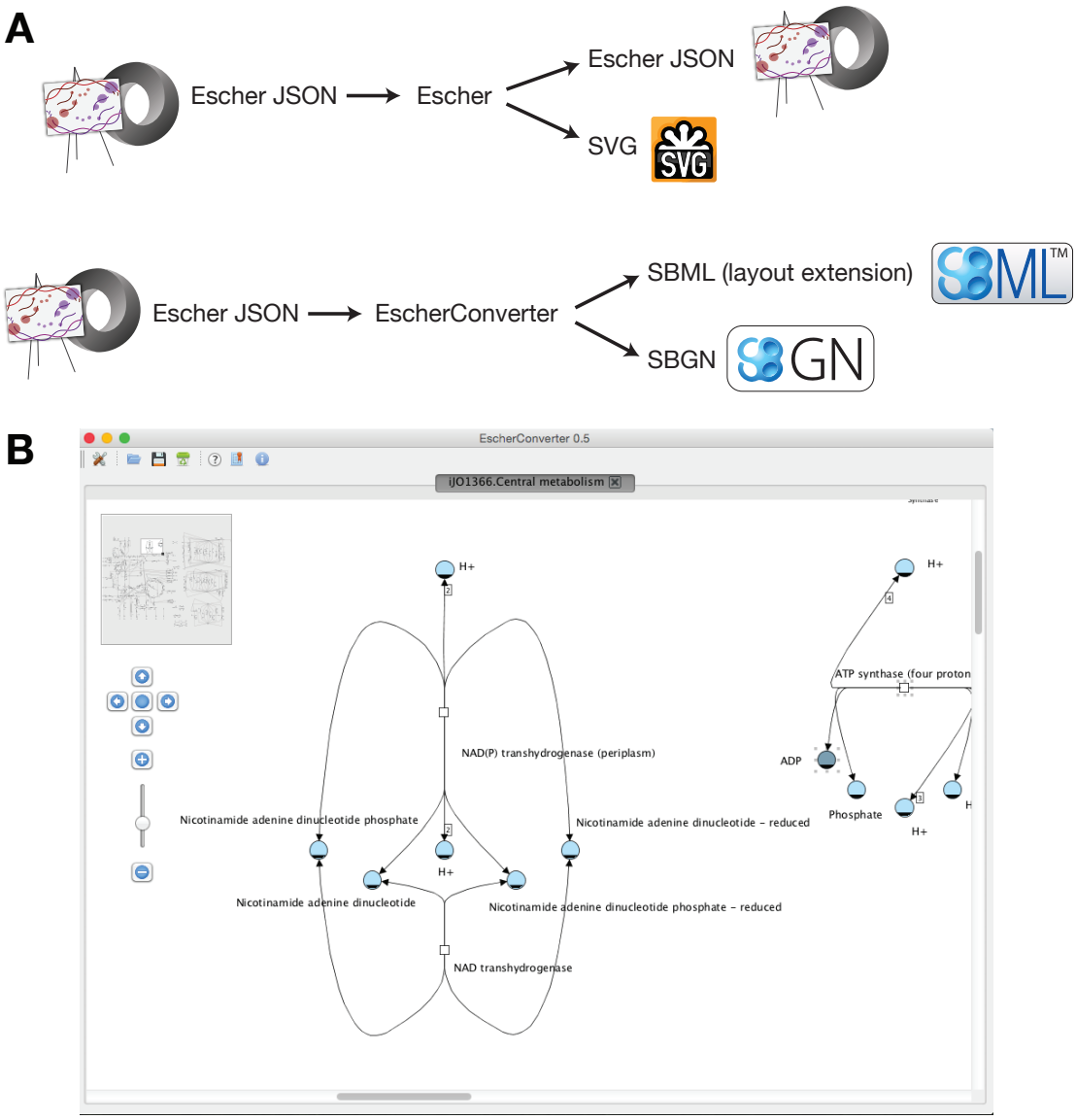

Figure 3.4: **The import and export types in Escher and the EscherConverter.** A) Escher can save to the Escher JSON file format or export to a SVG image. EscherConverter can be used to generate files in the SBML and SBGN-ML formats. B) The EscherConverter graphical user interface

Open-source development. Escher is hosted on GitHub, with a public bug tracker and tools for community contribution to the codebase (https://github.com/zakandrewking/escher). Documentation for Escher is available and was generated using Sphinx and ReadTheDocs (https://escher.readthedocs.org). This documentation

includes a description of the Escher features and detailed information on the JavaScript and Python APIs.

Integrating Escher with analysis workflows. The Escher Python package, which is available from the Python Package Index (PyPI, https://pypi.python.org), can be used to integrate Escher maps with data analysis and simulation workflows. Using the available functions, datasets can be applied to Escher maps, and the resulting maps can be saved as standalone web pages, saved as JSON or SVG, or exported using the EscherConverter as a command line utility. The Python package works directly with COBRA models using COBRApy (Ebrahim et al. 2013). It also includes functions for modifying all of the Escher map settings, including the color and size scales for all elements.

The Python package also includes a simple web server to run Escher locally. The web server caches maps and models for offline use, and users can also add maps to the cache directory so that they appear in the local web application. The following commands will install the package, print the location of the local cache directory, and run the Escher server:

```
# install escher
pip install escher
# print the cache directory
python -c "import␣escher;␣print␣escher.get_cache_dir()"
# run the local server (available at http://localhost
   :7778)
python -m escher.server
```

Developing with Escher. Application programming interfaces (APIs) are available for both JavaScript and Python to enable users to build, modify, and export maps programmatically. The specific functions in the APIs are defined in the Escher

Documentation. New web applications can be built on top of the basic Escher functions by developing with the Escher JavaScript API. The Documentation provides details on implementing a very simple web page with an embedded Escher map.

3.4 Availability and Future Directions

Escher version 1.1 is now available. Bug fixes and new pathway maps will be released regularly, and a number of Escher applications are currently in progress. Escher releases will follow the Semantic Versioning guidelines (http://semver.org) so that application developers can rely on new versions of Escher to be backwards compatible.

The Escher approach to web visualization. A major focus during development of future Escher versions will be to generalize and improve the approach to web visualization. As discussed in the Introduction, there are many types of biological visualizations that contribute to our interpretation of "omics" datasets. Successful user interface designs should be applicable to all of these visualization types, with modifications for the specific needs of a tool. As web platforms become ubiquitous for application development, it is important to consider what elements might be shared across a suite of visualization tools. This would make development of new tools easier, and improve interoperability between tools. For example, a genetic dataset in Escher could link directly to a visualization of the dataset on a genome browser.

The BiGG database. Escher will be included in the next release of the BiGG database (Schellenberger et al. 2010). The BiGG database is a repository for COBRA

models developed in the Systems Biology Research Group at the University of California, San Diego. BiGG already includes static pathway maps for many models in the database. Escher maps will be embedded in the web pages for models, reactions, and metabolites so that users can quickly see the network context of a biological component, and the maps will be available on both the BiGG and Escher websites.

A community effort. The Escher framework is highly amenable to improvements, such as new visual features. Example improvements include compartment membranes, representations of regulation and signaling such as those in the SBGN specification, better statistical tools for analyzing and comparing various data types, more import and export options, and direct integration of other visualizations (such as protein and metabolite structures). Because Escher is an open-source project, contributions from the community—bug fixes, use cases, code contributions, etc.—will be encouraged and will be an important factor in making Escher a sustainable, long-term solution to the challenges of visualizing biological pathways.

3.5 Availability and requirements

1. Project name: Escher
2. Project home page: https://escher.github.io
3. Project source: https://github.com/zakandrewking/escher
4. Open-source license: MIT license
5. Operating systems(s): Platform independent

6. Programming languages: JavaScript, Python, and Java

7. Other requirements: none

8. Any restrictions to use by non-academics: no limitations

Chapter 3 is a reprint of a published manuscript: King, Z. A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N. E., and Palsson, B. O. (2015a). "Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways". In: *PLoS Comput. Biol.* 11.8, e1004321. DOI: 10.1371/journal.pcbi.1004321. The dissertation author was the primary author of the paper and was responsible for the research.

Chapter 4

Optimizing Cofactor Specificity of Oxidoreductase Enzymes for the Generation of Microbial Production Strains—OptSwap

4.1 Introduction

In microorganisms, anabolism and catabolism are precisely controlled by regulation of flux through metabolic enzymes. This regulation of anabolic and catabolic activity is an important element of the adaptive system that allows a microorganism to find the optimal phenotype for growth and reproduction in its environment. The currency metabolites NAD(H) and NADP(H)—which carry and transfer reducing equivalents—play unique roles in metabolism. The primary role of the reduced respiratory cofactor NADH is to transfer electrons to oxygen via the electron transport chain,

generating the proton gradient that is used for oxidative phosphorylation of ADP to ATP (Russell and Cook 1995; Sauer et al. 2004; Gottschalk 1986). Concurrently, the reduced cofactor NADPH donates electrons to anabolic reactions and drives biosynthetic pathways in the cell (Sauer et al. 2004; Gottschalk 1986). Despite the chemical similarity between NAD(H) and NADP(H), many central dehydrogenase and reductase enzymes in the cell preferentially catalyze reduction or oxidation of a specific carrier (Kim et al. 2011; Bocanegra, Scrutton, and Perham 1993; Rodríguez-Arnedo et al. 2005; Lunzer et al. 2005). Thus, the functional separation of these electron carriers and the specificity of enzymes to one electron carrier allow the system to direct resources to energy production or anabolism on a whole-cell scale.

Cellular control over the direction of reducing equivalents to NADH or NADPH is facilitated both by the specificity and activity of dehydrogenase reactions and also by the activity of transhydrogenase enzymes, which transfer reducing equivalents between the two cofactors. Two transhydrogenases are encoded in the genome and expressed in *Escherichia coli*. The soluble transhydrogenase encoded by the gene *sthA* catalyzes the reaction

$$\mathrm{NAD^+ + NADPH \rightarrow NADP^+ + NADH}$$

The membrane-bound transhydrogenase encoded by *pntAB* couples reduction of $\mathrm{NADP^+}$ with inward proton translocation, catalyzing the reaction

$$\mathrm{NADP^+ + NADH + H^+_{\mathit{out}} \leftrightarrow NADPH + NAD^+ + H^+_{\mathit{in}}}$$

It has been reported that 35–45% of the NADPH required for biosynthesis is generated

by the transhydrogenase encoded by *pntAB* during aerobic batch growth on glucose (Sauer et al. 2004). The large contribution of this transhydrogenase enzyme to NADPH production reflects the importance of transhydrogenases for balancing cofactor generation to meet the needs of the cellular environment (Sauer et al. 2004; Fuhrer and Sauer 2009), so any strategy to modify cofactor production should consider the role that transhydrogenases play in attempting to maintain homeostasis.

Oxidoreductase specificity for NAD(H) or NADP(H) is a central parameter in determining the direction of cellular resources. Thus, strategies have been developed to modulate the cofactor specificity of dehydrogenase enzymes in order to increase production of desirable cellular products. Protein engineering has been used to switch the carrier specificity of dehydrogenase enzymes by modifying the amino acid residues of the nucleotide-phosphate binding site. In *E. coli*, the enzymes dihydrolipoadmide dehydrogenase (a component of the pyruvate dehydrogenase complex and the 2-oxoglutarate dehydrogenase complex) (Bocanegra, Scrutton, and Perham 1993; Guest et al. 2003) and isopropylmalate dehydrogenase (Lunzer et al. 2005) were engineered to prefer NADP(H) over NAD(H) as cofactor, and the enzyme isocitrate dehydrogenase (Hurley, Chen, and Dean 1996) was engineered to prefer NAD(H) over NADP(H). Furthermore, the cofactor specificity of isocitrate dehydrogenase was reversed by protein engineering, and a competition study was performed in which the modified strain outcompeted wild-type when grown on glucose (Zhu, Golding, and Dean 2005). Despite these efforts, dehydrogenase enzymes modified by protein engineering have not been exploited for metabolic engineering purposes.

An alternative strategy to engineer cofactor specificity is to replace native dehydrogenase reactions in central carbon metabolism with heterologous dehydro-

genases that have specificity for the opposite cofactor. These non-native enzymes may be more efficient than enzymes produced using protein engineering because the heterologous enzymes have benefited from evolutionary optimization. To this end, the NAD(H)-dependent glyceraldehyde-3-phosphate dehydrogenase in *E. coli* was replaced with the NADP(H)-dependent glyceraldehyde-3-phosphate from *Clostridium acetobutylicum* to increase lycopene yield (Martínez et al. 2008). In *Saccharomyces cerevisiae*, the native NAD(H)-dependent glyceraldehyde-3-phosphate dehydrogenase was replaced with the NADP(H)-dependent enzyme from *Kluyveromyces lactis* to increase fermentation of D-xylose to ethanol (Verho et al. 2003). Thus, researchers have exploited homologous swaps to boost production phenotypes.

The scientific interest shown in modifying electron carrier availability and the yield improvements seen in these modified strains suggest that a computational approach to model oxidoreductase specificity modifications could guide future experimental work and ultimately affect metabolic engineering. Which oxidoreductase specificity modifications are likely to have the greatest impact on the system? And how can oxidoreductase specificity changes be paired with reaction knockouts most effectively? These kinds of questions can be answered with an *in silico* modeling procedure.

As a tool for investigating metabolic networks *in silico*, constraint-based reconstruction and analysis (COBRA) methods have been shown to accurately predict bacterial behavior under many conditions (Feist and Palsson 2008; McCloskey, Palsson, and Feist 2013). COBRA methods utilize genome-scale models (GEMs), which are built by pairing the reconstructed metabolic network of an organism with governing constraints based on physico-chemical conservations (i.e. known reaction stoichiome-

tries), spatial limitations, and environmental parameters. Governing constraints are formulated as a set of linear inequalities that enclose the solution space available to metabolic fluxes (Price, Reed, and Palsson 2004). Solution spaces can be examined by optimizing for objectives using flux balance analysis (FBA) and other linear programming methods (Lewis, Nagarajan, and Palsson 2012). The optimal solutions predicted by FBA can match *in vivo* behavior in cases where the metabolic network of the cell is optimized for the same objective (e.g., growth). A powerful approach to achieve *in vivo* optimality is adaptive laboratory evolution (ALE). ALE optimizes the genotype of the organism, and the result is often a match between the observed growth phenotypes and the model predictions (Ibarra, Edwards, and Palsson 2002; Fong et al. 2006; Fong and Palsson 2004; Fong et al. 2005). Growth-coupled strains—strains where growth is directly coupled to the production of a given molecule—have drawn attention because these strains are predicted to produce high yields of the target molecule after ALE selecting for cell growth (Feist et al. 2010). Computational algorithms that identify growth-coupled designs have proliferated, including OptKnock (Burgard, Pharkya, and Maranas 2003), RobustKnock (Tepper and Shlomi 2010), and OptGene (Patil et al. 2005). (For a review, see Lewis, Nagarajan, and Palsson 2012.) Thus, *in silico* COBRA methods coupled with ALE constitute a strategy for rational engineering of high-yield production strains. A COBRA method for optimizing cofactor specificity has not been reported. Previous investigations have explored the importance of cofactor balancing for microbial production strains (Fuhrer and Sauer 2009; Verho et al. 2003; Ghosh, Zhao, and Price 2011), but *de novo* strain design has not been explored using non-biased cofactor optimization.

In this work, we present an *in silico*, constraint-based modeling technique—

OptSwap—for generating strategies to optimize the production of native cellular compounds by modifying the electron carrier specificity of oxidoreductase reactions in the metabolic network. A mixed-integer linear programming (MILP) method is presented that optimizes growth-coupled product yield by pairing oxidoreductase specificity swaps with reaction knockouts. Utilizing OptSwap, we predict novel, growth-coupled designs for the production of valuable compounds by *E. coli*, and we compare these designs to solutions that are predicted utilizing reaction knockouts alone.

4.2 Methods

4.2.1 Modeling and computational tools

The *i*JO1366 metabolic reconstruction of *E. coli* K-12 MG1655 was used for all simulations in this work (Orth et al. 2011). As described previously, FHL (formate-hydrogen lyase) and the oxidative stress reactions CAT (catalase), SPODM (cytosolic superoxide dismutase), and SPODMpp (periplasmic superoxide dismutase) were constrained to zero (Orth et al. 2011). The POR5 (pyruvate:ferredoxin oxidoreductase) reaction was made irreversible, as supported by biochemical data (Blaschkowski et al. 1982).

Flux balance analysis (FBA) (Kauffman, Prakash, and Edwards 2003), parsimonious flux balance analysis (pFBA) (Lewis et al. 2010a), flux variability analysis (FVA) (Mahadevan and Schilling 2003), and RobustKnock (Tepper and Shlomi 2010) were implemented in MATLAB as described in the literature. All simulations were performed using MATLAB (The MathWorks Inc., Natick, MA, USA) and the COBRA

Toolbox (Becker et al. 2007) software packages with TOMLAB/CPLEX (Tomlab Optimization Inc., San Diego, CA, USA) and Gurobi (Gurobi Optimization, Inc., Houston, TX, USA) LP/MILP solvers.

Substrate uptake rates for the solitary carbon substrates in each simulation were constrained to a maximum uptake rate of 20 mmol gDW^{-1} h^{-1}. For aerobic simulations, the oxygen uptake rate was set to a maximum of 20 mmol gDW^{-1} h^{-1}. These values were chosen based on experimental observations of aerobic and anaerobic growth of *E. coli* (Varma, Boesch, and Palsson 1993; Varma and Palsson 1994).

4.2.2 Model reduction and selection of reaction set for knockouts

The reaction set available for knockout was restricted based on a previously reported method for model reduction and target reaction selection (Feist et al. 2010). After setting the bounds for the primary carbon source and oxygen exchange, we performed the following procedure, as reported (Feist et al. 2010): (1) To generate a "reduced model," reactions that could not be utilized under the conditions of the simulation were removed, and upper and lower bounds of all reactions were set to maximum and minimum obtainable values, as determined by FVA. (2) Non-gene associated reactions, spontaneous reactions, transport reactions, reactions in peripheral metabolic pathways, and reactions acting on high-carbon containing molecules were removed from the knockout reaction set. (3) Sets of coupled reactions were identified by examining the null space of the stoichiometric matrix (Palsson 2006), and only one reaction in each correlated set was considered for knockout.

With glucose as the substrate, the reactions available for knockout during

optimization numbered 217 anaerobically and 243 aerobically. With D-xylose as the substrate, the reactions available for knockout numbered 216 anaerobically and 238 aerobically. (Supplementary Table S1; Supplementary Data are available online at http://dx.doi.org/10.1089/ind.2013.0005) Both the proton-translocating transhydrogenase reaction THD2pp (*pntAB*) and the energy-independent transhydrogenase reaction NADTRHD (*sthA*) were constrained to zero for the transhydrogenase-knockout design $\Delta pntAB\ \Delta sthA$.

4.2.3 Selection of reaction set for cofactor specificity swaps

The set of oxidoreductase reactions available for modification in the OptSwap procedure was determined by finding the oxidoreductase reactions in the high-flux backbone (Almaas et al. 2004). In the metabolic model *i*JO1366, all reactions that utilize NAD(H) or NADP(H) as a cofactor were located. The reactions were sorted by flux magnitude after pFBA optimization for flux through the biomass objective function under conditions of aerobic and anaerobic growth on glucose and D-xylose minimal media. The reactions with highest flux under each set of conditions were selected. D-Lactate dehydrogenase, malic enzymes, and L-1,2-propanediol oxidoreductase were added to the set based on interest in the literature (Zhang et al. 2007; Wang et al. 2011; Stols and Donnelly 1997). Malate dehydrogenase was removed from the pool of oxidoreductase enzymes that can be swapped with OptSwap because the NADPH-specific malate dehydrogenase allowed non-physiological loops to form in the flux solutions. Under anaerobic conditions, NADH:oxidoreductase I was removed from the pool of reactions during model reduction because the reaction cannot carry flux during simulations of anaerobic growth. Thus, 22 oxidoreductase enzymes were chosen

under aerobic conditions and 21 oxidoreductase enzymes under anaerobic conditions (Table 4.1).

4.2.4 MILP formulation

OptSwap is a bi-level MILP problem based on RobustKnock (Tepper and Shlomi 2010) with new constraints that enforce swaps of the cofactor specificity of oxidoreductase reactions (Fig. 4.1). The following procedure was used to incorporate these constraints into the RobustKnock problem. First, for each oxidoreductase enzyme in the pool of the reactions that can be "swapped," a reaction with opposite specificity (either NAD(H) or NADP(H)) was added to the model. New Boolean decision variables were utilized. The variables s_d represent the on/off state of the native oxidoreductase reactions, and t_d represent the on/off state of the "swapped" reactions, where a value of 0 means the reaction is off and a value of 1 means the reaction is on. D is the set of oxidoreductase reaction pairs (native and "swapped").

$$s_d \in \{0, 1\} \qquad \forall d \in D \tag{4.1}$$

$$t_d \in \{0, 1\} \qquad \forall d \in D \tag{4.2}$$

Table 4.1: **Oxidoreductase reactions targeted for analysis with OptSwap.**

Key oxidoreductase enzymes in *E. coli*	Gene symbol	Model reaction	Native electron carrier	Past studies exploring alternative cofactor uses
Glyceraldehyde-3-phosphate dehydrogenase	*gapA*	GAPD	NADH	non-native enzyme (Martínez et al. 2008; Verho et al. 2003)
Acetaldehyde dehydrogenase	*adhE* OR *mhpF*	ACALD	NADH	
Ethanol dehydrogenase	*adhP* OR *adhE*	ALCD2x	NADH	
Glutamate dehydrogenase	*gdhA*	GLUDy	NADPH	non-native enzyme (Yaoi et al. 1996)
Glucose-6-phosphate dehydrogenase	*zwf*	G6PDH2r	NADPH	
6-Phosphogluconate dehydrogenase	*gnd*	GND	NADPH	
FAD reductase	*fre*	FADRx	NADH	
Phosphoglycerate dehydrogenase	*serA*	PGCD	NADH	
Isocitrate dehydrogenase	*icd*	ICDHyr	NADPH	*in silico* structural analysis (Baba et al. 2006; Hurley, Chen, and Dean 1996) protein engineering (Rodríguez-Arnedo et al. 2005; Zhu, Golding, and Dean 2005; Wang et al. 2011; Baba et al. 2006)
Aspartate-semialdehyde dehydrogenase	*asd*	ASAD	NADPH	
Methylene tetrahydrofolate dehydrogenase	*folD*	MTHFD	NADPH	
Acetohydroxy acid isomeroreductase (2-Acetolactate)	*ilvC*	KARA1	NADPH	
Homoserine hydrogenase	*metL* OR *thrA*	HSDy	NADPH	
3-Isopropylmalate dehydrogenase	*leuB*	IPMD	NADH	protein engineering (Auriol et al. 2011; Lunzer et al. 2005)
Shikimate dehydrogenase	*aroE*	SHK3Dr	NADPH	
Dihydrodipicolinate reductase	*dapB*	DHDPRy	NADPH	
NADH:ubiquinone oxidoreductase I	*nuoF, nuoA–C, nuoE, nuoG–N*	NADH16pp	NADH	directed evolution (Heavner et al. 2012)
Pyruvate dehydrogenase	*lpd, aceE, aceF*	PDH	NADH	protein engineering (Bocanegra, Scrutton, and Perham 1993; Guest et al. 2003)
L-1,2-Propanediol oxidoreductase	*fucO*	LCARR	NADH	
D-Lactate dehydrogenase	*ldhA*	LDH_D	NADH	
Malic enzyme (NADH)	*maeA*	ME1	NADH	
Malic enzyme (NADPH)	*maeB*	ME2	NADPH	

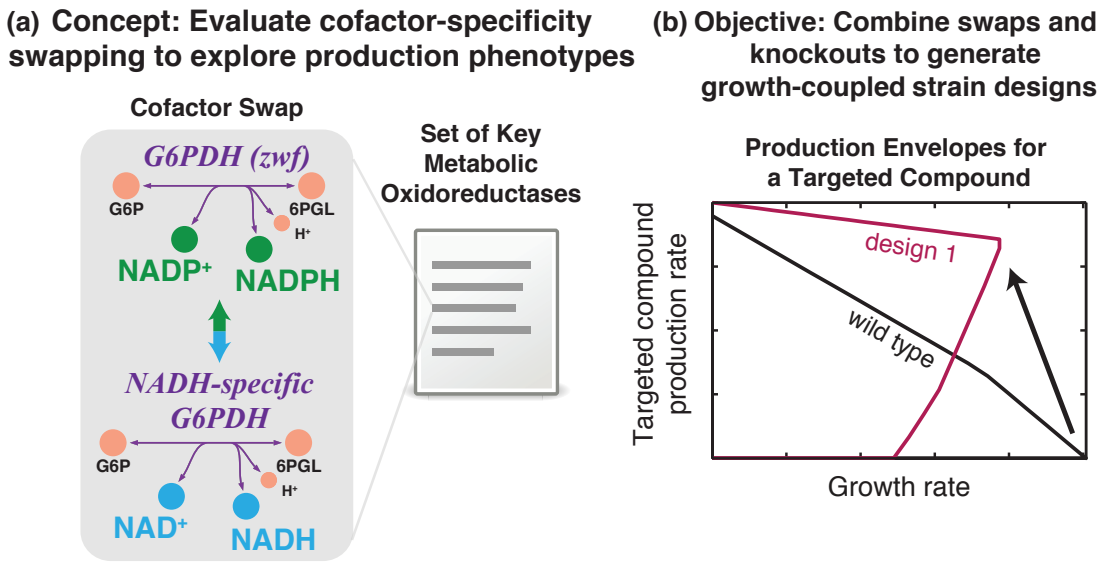

Figure 4.1: **The OptSwap formulation for optimizing cofactor specificity of major metabolic enzymes (NAD(H) vs. NADP(H)) coupled to reaction knockouts.** (a) Constraints are added to the MILP problem to enforce swapping the cofactor specificity of reactions catalyzed by oxidoreductase enzymes. (b) Production envelopes show the ability to growth couple a product of interest, which is not possible with native cofactor specificity. G6PDH: glucose-6-phosphate dehydrogenase; G6P: glucose-6-phosphate; 6PGL: 6-phophogluconolactone.

These variables are present in addition to the RobustKnock Boolean variables y_e which represent the on/off state of all reactions that can be knocked out in the model. E is the set of reactions that can be knocked out.

$$y_e \in \{0, 1\} \qquad \forall e \in E \tag{4.3}$$

Second, a constraint was added to the outer problem that requires either the native or the "swapped" reaction to be knocked out for each oxidoreductase reaction pair in D. This is the constraint that forces an oxidoreductase "swap."

$$s_e + t_d = 1 \qquad \forall e \in E \tag{4.4}$$

Third, a constraint was added to the outer problem to limit the number of swaps to be less than or equal to the parameter L.

$$\sum_{d \in D} (1 - s_d) \leq L \tag{4.5}$$

The knockout limitation constraint from RobustKnock ensures that the number of reaction knockouts is less than or equal to the parameter K.

$$\sum_{e \in E} (1 - y_e) \leq K \tag{4.6}$$

Fourth, a constraint was added to the outer problem to limit the number of interventions (oxidoreductase swaps and reaction knockouts) to be less than or equal to the parameter X.

$$\sum_{e \in E} (1 - y_e) + \sum_{d \in D} 1 - s_d \leq X \tag{4.7}$$

These three constraints on the number of knockouts and swaps can be included or excluded from the problem according to the desired simulation scenario. If the oxidoreductase set D is empty, then OptSwap reduces to the RobustKnock problem.

Fifth, a set of constraints was added to limit flux to zero for any native or swapped oxidoreductase reaction whose corresponding decision variable is equal to zero. The function x maps the set of oxidoreductase reaction pairs, D, to the corresponding fluxes and bounds for native oxidoreductase reactions in the model, and the function y maps the set of oxidoreductase reaction pairs to the corresponding fluxes and bounds for the non-native, "swapped" oxidoreductase reactions. Thus, when $s_d = 0$, the flux

$v_{x(d)}$ is constrained to zero.

$$s_d LB_{x(d)} \leq v_{x(d)} \leq s_d UB_{x(d)} \qquad \forall d \in D \tag{4.8}$$

$$t_d LB_{y(d)} \leq v_{y(d)} \leq t_d UB_{y(d)} \qquad \forall d \in D \tag{4.9}$$

These are present in addition to the knockout constraint from RobustKnock. The function z maps the set of reactions that can be knocked out, E, to the corresponding fluxes.

$$y_e LB_{z(e)} \leq v_{z(e)} \leq y_e UB_{z(e)} \qquad \forall e \in E \tag{4.10}$$

As an illustration of the variables s and t, consider the case where $s_d = 0$. Then, by Equation 4.4, $t_d = 1$. Equation 4.9 reduces to $LB_{y(d)} \leq v_{y(d)} \leq UB_{y(d)}$, so flux through the "swapped" oxidoreductase reaction is constrained only by the lower and upper bounds—the reaction is "on." For the same case, Equation 4.8 reduces to $0 \leq v_{x(d)} \leq 0$, so flux through the native oxidoreductase reaction is constrained to zero—the reaction is "off."

Thus, the final formulation of the OptSwap problem can be stated as follows.

J is the set of all reaction fluxes, and I is the set of all metabolites in the model.

$$
\begin{aligned}
&\max \min v_{chemical} \\
&\text{s.t.} \\
&\left[\begin{array}{ll}
\max v_{biomass} & \\
\text{s.t.} & \\
\sum_{j \in J} S_{ij} v_j = 0 & \forall i \in I \\
LB_j \leq v_j \leq UB_j & \forall j \in J \\
s_d LB_{x(d)} \leq v_{x(d)} \leq s_d UB_{x(d)} & \forall d \in D \\
t_d LB_{y(d)} \leq v_{y(d)} \leq t_d UB_{y(d)} & \forall d \in D \\
y_e LB_{z(e)} \leq v_{z(e)} \leq y_e UB_{z(e)} & \forall e \in E
\end{array}\right] \\
&s_d \in \{0,1\} \quad \forall d \in D \\
&t_d \in \{0,1\} \quad \forall d \in D \\
&y_e \in \{0,1\} \quad \forall e \in E \\
&s_d + t_d \leq 1 \quad \forall d \in D \\
&\sum_{d \in D} (1 - s_d) \leq L \\
&\sum_{e \in E} (1 - y_e) \leq K \\
&\sum_{e \in E} (1 - y_e) + \sum_{d \in D} 1 - s_d \leq X
\end{aligned}
\tag{4.11}
$$

To solve this bi-level max-min problem, we implemented the techniques that were used to simplify and solve RobustKnock (Tepper and Shlomi 2010). The function

dual_embed from the RobustKnock implementation was used to generate the dual of the inner problem and linearize the bilinear terms (Eq. 4.8–4.10). By the strong duality theory of linear programming, the outer problem and the dual of the inner problem could be integrated (Tepper and Shlomi 2010). Finally, the minimization problem in the resulting min-max problem was converted to a maximization problem using the *dual_embed* function. After this second conversion and integration, the OptSwap problem is a max-max optimization that can be solved with standard MILP solvers. (See Supplement S5 for the implementation of OptSwap in MATLAB).

A slight variation of the OptSwap MILP was also investigated. Equation 4.4 can be converted to an inequality so that oxidoreductase reactions can be swapped ($s_d = 0$, $t_d = 1$) or knocked out ($s_d = t_d = 0$).

$$s_d + t_d \leq 1 \qquad \forall d \in D \tag{4.12}$$

However, this more-general formulation proved to be more computationally intensive and presented numerical challenges that we could not resolve with the TOMLAB/CPLEX solver. Results with this formulation were limited to run for 12 hours, and many simulations did not solve to optimality (Supplementary Table S4).

4.3 Results

The *in silico* metabolic optimization procedure OptSwap was used to predict production strains that couple cellular growth to production of targeted products using a combination of oxidoreductase specificity swaps and reaction knockouts. In order to optimize for maximal growth-coupled yield, a MILP problem was developed

based on the RobustKnock problem (Tepper and Shlomi 2010). RobustKnock is a bi-level optimization problem similar to OptKnock, except that, where OptKnock maximizes production of a target molecule subject to maximizing biomass production, the RobustKnock problem maximizes the *minimum* production of a target product subject to maximizing growth rate. This procedure ensures that "non-unique" solutions are not produced—a concern that was first addressed by tilting the objective function prior to the appearance of RobustKnock (Feist et al. 2010). OptSwap contains additional constraints so that both reaction knockouts and oxidoreductase swaps can be utilized to identify a growth-coupled design.

OptSwap was demonstrated using the *i*JO1366 genome-scale metabolic model of *E. coli* (Orth et al. 2011). The pool of oxidoreductase enzymes that could be "swapped" in the analysis was determined by identifying central, high-flux-carrying reactions (as described in Methods). Where reaction knockouts were considered for use with OptSwap, a reaction selection workflow was utilized as previously reported (Feist et al. 2010). All swap designs were based on the transhydrogenase-deficient $\Delta pntAB$ $\Delta sthA$ genotype to ensure that unconstrained transhydrogenase activity in the model did not counteract the effect of swapping oxidoreductase specificity.

Growth-coupled designs predicted by OptSwap with both oxidoreductase specificity swaps and reaction knockouts were compared with solutions from RobustKnock where only reactions knockouts were considered. The designs were categorized by the number of interventions, where an intervention is a reaction knockout or a dehydrogenase swap. Thirteen high-value products were investigated based on a previous screening criteria for industrially significant compounds produced natively in *E. coli* (Feist et al. 2010). For each product, simulations were performed under conditions of

glucose and D-xylose minimal media for both aerobic and anaerobic growth (Supplementary Table S3).

For the conditions considered in these simulations, OptSwap produced ten designs where coupling oxidoreductase specificity swaps with reaction knockouts resulted in superior predicted production rate compared to just reaction knockouts (Table 4.2). The eight designs where OptSwap predicted significantly stronger growth coupling or higher substrate-specific productivity (yield ÃŮ growth rate) were investigated in more detail (Fig. 4.2). OptSwap predicted a strongly growth-coupled design for L-alanine production that secretes no coproducts and generates L-alanine with high yield under anaerobic conditions on glucose minimal media. The design requires only three knockouts and one swap of the native enzyme (Fig. 4.3); it utilizes an NAD(H)-specific glutamate dehydrogenase in place of the native NADP(H)-specific reaction (*ghdA*). NAD(H)-specific glutamate dehydrogenase transfers electrons from NADH to glutamate. Subsequently, L-alanine transaminase (*alaA* or *alaC*) produces L-alanine and 2-oxoglutarate from glutamate and pyruvate. Knockouts of D-lactate dehydrogenase (*ldhA*) and ethanol dehydrogenase (*adhP* or *adhE)* prevent D-lactate and ethanol production. With the new NADH-specific glutamate dehydrogenase pathway in place, L-alanine secretion is energetically favorable to succinate secretion. Knockout of acetyl-CoA acetyltransferase (*atoB*) is also predicted to be necessary for growth coupling (Fig. 4.4).

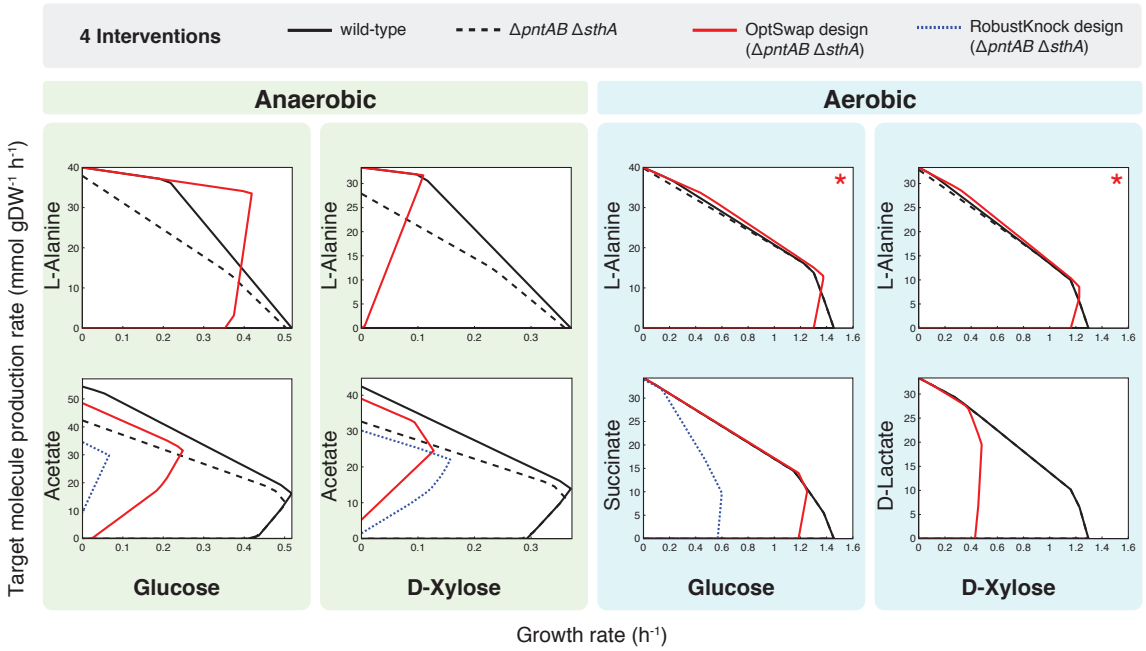

Figure 4.2: **Calculated production envelopes for OptSwap designs predicted to have significantly higher optimal production rate or substrate-specific productivity than designs with just reaction knockouts.** The wild-type (solid grey) and the *DeltapntAB DeltasthA* genotype (dashed black) are shown. All designs were found by first knocking out transhydrogenases (*DeltapntAB DeltasthA*). The optimized phenotypes for OptSwap (solid red) and RobustKnock (dashed orange) are compared. Each phenotype required four interventions (one intervention being either a reaction knockout or a oxidoreductase swap). The star (*) indicates that the same design was identified in both cases. The x indicates that no growth-coupled design was identified by RobustKnock under these conditions. OptSwap predicts growth-coupled designs for producing L-alanine under all four conditions and D-lactate aerobically on D-xylose substrate; growth coupling is not predicted for these products with just four or fewer reaction knockouts under identical conditions. The OptSwap designs for succinate and acetate are predicted to have higher yield than designs with just reaction knockouts. The designs for succinate production and acetate production on glucose substrate are also predicted to have higher substrate-specific productivities than the designs predicted by RobustKnock.

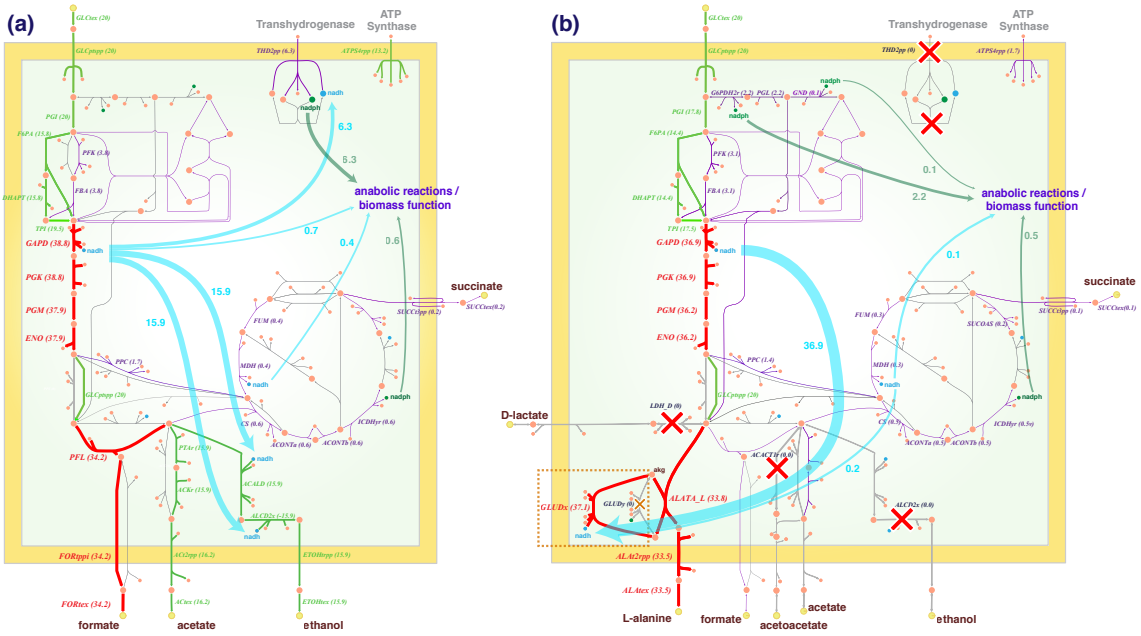

Figure 4.3: **Network diagrams showing the shift in reduced cofactor usage between wild-type and the L-alanine production design.** (a) Simulated wild-type flux distribution during anaerobic fermentation in *E. coli*. The fluxes shown are unique solutions calculated using pFBA, optimizing flux through the biomass objective function. (b) L-alanine production design. Red X's indicate reaction deletions, and the orange box indicates the oxidoreductase swap. Reactions are shown with arrows pointing in the direction of flux. Knockouts of ethanol dehydrogenase (*adhP* or *adhE*), acetyl-CoA acetyltransferase (*atoB*), and lactate dehydrogenase (*ldhA*) reactions prevent ethanol and D-lactate from being produced. With the NAD(H)-specific glutamate dehydrogenase (*gdhA_nadh*) in place, L-alanine is predicted to be the energetically-favorable final fermentation product.

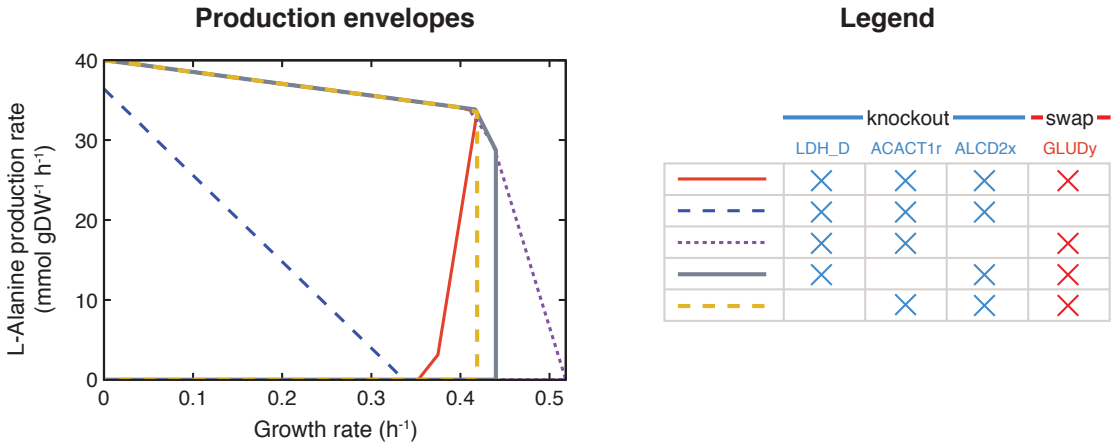

Figure 4.4: The predicted effect of combinatorial interventions on the design for anaerobic production of L-alanine on glucose minimal media. Glutamate dehydrogenase must be swapped for high yields of L-alanine to be produced at high growth rates. (Dashed blue envelope shows production with the native glutamate dehydrogenase.) With fewer than three reaction knockouts and the swapped glutamate dehydrogenase, high yield of L-alanine is possible, but L-alanine production is not growth-coupled. The final set of three reactions knockouts and one dehydrogenase swap causes a strong coupling between L-alanine production and cell growth. LDH_D: D-lactate dehydrogenase; ACACT1r: acetyl-CoA acetyltransferase; ALCD2x: ethanol dehydrogenase; GLUDy: glutamate dehydrogenase.

For anaerobic production of L-alanine from D-xylose and for aerobic production of L-alanine from glucose and D-xylose, slightly different designs were predicted. Phenotypes for aerobic production of L-alanine are predicted to have lower yield and higher maximum growth rate than the anaerobic designs (Fig. 4.2). In all four L-alanine production designs, swapping the cofactor specificity of glutamate dehydrogenase results in predicted phenotypes with strong growth coupling, while growth coupling is not predicted in any designs with four or fewer reaction knockouts. However, previously reported simulations suggest that growth-coupling of L-alanine production may be possible with five reaction knockouts in *E. coli* (Feist et al. 2010).

OptSwap predicted that anaerobic production of acetate can be increased with four interventions (Fig. 4.2). The acetate designs rely on replacing the native NAD(H)-dependent ethanol dehydrogenase (*adhP* or *adhE*) with a NADP(H)-dependent ethanol

Table 4.2: Calculated maximum optimal production rate for designs selected by OptSwap and designs selected by RobustKnock.

	Maximum optimal production rate, OptSwap/RobustKnock (mmol/gDW/h)															
	Glucose anaerobic				D-Xylose anaerobic				Glucose aerobic				D-Xylose aerobic			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Ethanol	22.3/22.3	34.7/34.7	36.4/36.4	37.6/37.6	23.4/23.4	31.9/31.9	32.3/32.3	32.4/32.4	2.6/2.6	18.0/18.0	24.8/24.8	24.9/25.0	0.0/0.0	13.8/13.8	22.4/22.4	**23.3/22.6**
Formate	36.5/36.5	36.9/36.9	37.2/40.3	34.6/40.7	30.6/30.6	31.5/31.5	32.5/33.5	29.5/33.8	29.8/29.8	36.3/38.5	39.1/39.1	39.1/39.1	25.2/25.2	29.1/29.1	32.5/32.5	34.8/34.8
Succinate	0.5/0.5	12.1/12.1	17.8/17.8	25.4/25.4	14.9/14.9	20.0/20.0	21.9/21.9	21.9/21.9	1.5/1.5	1.6/1.4	5.1/5.1	**10.1/9.8**	1.3/1.3	6.1/6.1	12.7/12.7	18.0/18.0
Acetate	17.7/17.7	16.1/24.8	25.1/26.8	**31.6/29.8**	19.7/19.7	19.7/20.3	19.7/21.4	**24.5/22.1**	25.1/25.1	32.2/36.6	33.3/37.3	**42.4/40.0**	20.8/20.8	32.6/32.6	38.4/38.4	39.0/39.0
D-Lactate	0.0/0.0	34.7/34.7	34.9/35.1	37.6/37.6	0.0/0.0	31.9/31.9	31.9/32.0	32.2/32.2	0.0/0.0	13.5/13.5	17.1/17.1	25.0/25.0	0.0/0.0	9.0/9.0	12.6/12.6	**19.2/0.0**
2-Oxoglutarate	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/7.0
L-Alanine	0.0/0.0	0.0/0.0	0.0/0.0	**33.5/0.0**	0.0/0.0	0.0/0.0	0.0/0.0	**31.7/0.0**	0.0/0.0	0.0/0.0	0.0/0.0	**13.0/0.0**	0.0/0.0	0.0/0.0	0.0/0.0	**8.5/0.0**
Glycerol	0.0/0.0	0.0/0.0	0.0/0.0	0.1/0.1	0.0/0.0	0.0/0.0	5.5/5.5	10.0/10.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	3.3/3.3
L-Serine	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Pyruvate	0.0/0.0	0.0/0.0	0.0/12.0	0.0/21.0	0.0/0.0	0.0/0.0	0.0/11.4	0.0/16.4	0.0/0.0	0.0/0.0	0.0/19.3	17.0/23.2	0.0/0.0	0.0/0.0	0.0/13.8	11.6/17.7
Fumarate	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
L-Malate	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
L-Glutamate	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0

dehydrogenase. On both glucose and D-xylose substrates, maximum optimal acetate production is predicted to be higher with the OptSwap design than with just four or fewer reaction knockouts. Furthermore, simulation predicted that acetate production on glucose substrate with the OptSwap design can be obtained at higher growth rates, and thus with higher substrate-specific productivity.

Simulations predicted that succinate can be produced aerobically with glucose as the substrate using three knockouts and one dehydrogenase swap. The OptSwap design for succinate production utilizes a NAD(H)-dependent glutamate dehydrogenase in place of the native enzyme, coupled to three reaction knockouts. The result is a small increase in the predicted succinate production rate at maximum growth rate but a large increase in the predicted substrate-specific productivity, so this design is interesting in the case where substrate-specific productivity is desirable.

The OptSwap design for aerobic production of D-lactate on D-xylose substrate is predicted to be strongly growth-coupled, whereas RobustKnock does not predict any growth-coupling for D-lactate with four or fewer interventions under identical conditions. The OptSwap design swaps the native NAD(H)-dependent ethanol dehydrogenase (*adhP* or *adhE*) with a NADP(H)-dependent ethanol dehydrogenase. Three knockouts were predicted: acetate kinase (*tdcD* or *ackA* or *purT*), ATP synthase (*atpA–H*), and pyruvate formate lyase (*pflA* and *pflB*). ATP synthase is not predicted to be essential by the metabolic model for growth on either of the substrates considered in these simulations (glucose amd D-xylose). Experimental evidence indicates that ATP synthase knockout strains grow slowly on glucose *in vivo* (Marx et al. 1999).

A number of cases exist where RobustKnock predicted a design with greater optimal yield than the OptSwap solution (Table 4.2). In these cases, the RobustKnock

solution contains a knocked out oxidoreductase reaction. In the formulation of OptSwap presented here, oxidoreductase reactions cannot be knocked out—Equation 4.4 enforces swaps and disallows knockouts. An alternative OptSwap formulation was considered in which oxidoreductase reaction knockouts are allowed. The equality constraint (Eq. 4.4) is replaced with an inequality (Eq. 4.12) that allows swaps or knockouts. However, the solution space of the more-general MILP problem caused challenges for the MILP solver used for this work. Numerical imprecision caused non-real solutions, and solution times increased to more than 12 hours. However, when the new problems solved to optimality (for a number of substrate/product combinations tested), they identified the same solutions found with RobustKnock (Supplementary Table S4).

4.4 Discussion

This study presents a computational method to predict optimal cofactor specificity of oxidoreductase reactions in the genome-scale metabolic model. The designs identified by OptSwap are non-intuitive solutions to produce desirable cellular products—solutions that are not possible with just reaction knockouts. Growth-coupled designs for L-alanine and D-lactate were identified by OptSwap using four interventions, where reaction knockouts could not be used to identify growth-coupled production phenotypes under identical conditions. The anaerobic production designs for L-alanine utilize a NAD(H)-specific glutamate dehydrogenase coupled to reaction knockouts that limit the pathways available for anaerobic fermentation in order to produce very high yields of L-alanine at maximum growth rate. For succinate and acetate, yield improvements were smaller, but large increases in substrate-specific productivity were predicted with OptSwap when compared to designs with just reaction

knockouts.

The unique anaerobic L-alanine production design reported here is simpler than a previously reported design requiring seven knockouts and one gene insertion (Zhang et al. 2007). In that study, the native D-lactate dehydrogenase of *E. coli* was replaced with alanine dehydrogenase from *Geobacillus stearothermophilus*. To ensure growth coupling of L-alanine, the authors also disabled fermentation pathways for ethanol, acetate, formate, D-lactate, L-lactate, and succinate. By utilizing the predictive power of constraint-based modeling, we predicted a design within high yield and fewer genetic manipulations.

The OptSwap *in silico* design strategies can be broadly applied to any product in the genome-scale model. A previous analysis demonstrated that growth coupling in genome scale models is most easily achieved for a class of products related to the native fermentation pathways in *E. coli* (Feist et al. 2010). However, techniques like OptSwap may expand the scope of products that can be growth coupled by exploring a broader set of solutions. The knockout and swap designs predicted by OptSwap are candidates for experimental validation, and can be built with current technologies. More complex designs than those presented here have been successfully implemented in *E. coli*, including the strain for production of L-alanine (Zhang et al. 2007).

The main limitation of OptSwap is that it is more computationally demanding than past methods, including RobustKnock and OptKnock. The greater complexity of the solution space means that finding an optimal solution is more difficult. MILP problems are NP complete, and solution time depends on the structure of the problem and the solving method. The solution times increased with the parameters K, L, and X (Supplementary Table S5), and adopting a more general problem increases solution

times further. While the simpler formulation of OptSwap reported here is computationally tractable, the more general problem allowing oxidoreductase knockouts could not be solved in many cases within 12 hours. CPLEX is a sophisticated MILP solver, and tweaking solver parameters did yield improvements in performance. OptSwap can be solved by choosing the less general formulation utilized for these results and restricting interventions (L or X) to be four or less. Improvements in MILP solvers and computational power will reduce these challenges in the future, and even more complex MILP problems will be solvable.

OptSwap adds a new level of complexity to the COBRA methods for identification of useful production phenotypes in microorganisms (Lewis, Nagarajan, and Palsson 2012). The constraints placed on the presence/absence of oxidoreductase reactions with specificity for different cofactors allow direct investigation and manipulation of cofactor pools that are central to directing cellular resources in microorganisms. Changing cofactor specificity alters the metabolic solution space. While reaction knockouts always decrease the size of the solution space, changing cofactor specificity can have more complex effects and may even increase the solution space in the direction of an objective function.

OptSwap can be readily implemented in metabolic models of other organisms. For example, the experimental findings with a dehydrogenase swap in *S. cerevisiae* (Verho et al. 2003) could be further investigated with the genome-scale metabolic model of that organism (Heavner et al. 2012). The constraints that enforce dehydrogenase specificity swaps can also be extended to target other metabolic specificity sets (e.g. nucleoside triphosphates) and can be incorporated into other *in silico* design strategies.

Chapter 4 is a reprint of a published manuscript: King, Z. A. and Feist,

A. M. (2013). “Optimizing Cofactor Specificity of Oxidoreductase Enzymes for the Generation of Microbial Production Strains—OptSwap”. In: *Ind. Biotechnol.* 9.4, pp. 236–246. DOI: 10.1089/ind.2013.0005. The dissertation author was the primary author of the paper and was responsible for the research.

Chapter 5

Optimal cofactor swapping can increase the theoretical yield for chemical production in *Escherichia coli* and *Saccharomyces cerevisiae*

5.1 Introduction

Division of the roles of currency metabolites is a well-conserved feature of metabolism in microorganisms. In *Escherichia coli* and *Saccharomyces cerevisiae*, the cofactors NAD(H) and NADP(H) are both responsible for transferring reducing equivalents between metabolic subsystems. NAD(H) is primarily generated by glycolytic enzymes and transfers reducing equivalents to the electron transport chain or to fermentation products (Russell and Cook 1995; Gottschalk 1986; Sauer et al. 2004). NADP(H) is produced primarily by the pentose phosphate pathway and tran-

shydrogenase enzymes, and it transfers reducing equivalents to provide energy for biosynthesis (Gottschalk 1986; Sauer et al. 2004). While this separation is not complete (for example, fungi utilize NADPH for pentose catabolism (Verho et al. 2003)), the functional separation of these electron carriers and the specificity of oxidoreductase enzymes to a specific electron carrier allow the cell to precisely partition resources between ATP production and anabolism.

When growing in a steady state, microorganisms coordinate the production of reduced cofactors to match cofactor consumption, and their metabolic network structures and regulatory systems are organized to carry out this balancing act in common environments (Sauer et al. 2004; Lunzer et al. 2005; Zhu, Golding, and Dean 2005). Consequently, the cofactor balance in microorganisms is poorly optimized for many synthetic cellular objectives (Ghosh, Zhao, and Price 2011; Lim et al. 2013; Jan et al. 2013). Thus, one should consider how cofactor balance can be optimized when formulating new cellular objectives for metabolic engineering and synthetic biology. Cofactor balance optimization is especially important for introducing non-native production pathways that are driven by cofactor concentration (Shen et al. 2011).

A number of experimental methods have been developed to increase the availability of the reduced cofactors NADH and NADPH to enzymes in production pathways that are cofactor-driven, and thereby increase yield of high-value byproducts (reviewed by Lee et al. 2013). One strategy to increase cofactor availability is overexpression of genes that generate cofactor-producing enzymes. Overexpression of NADH-producing formate dehydrogenase (*fdh1* from *Candida boidinii*) in *E. coli* was shown to increase production of ethanol during anaerobic fermentation and to cause production of fermentation byproducts during aerobic growth (Berríos-Rivera, San, and Bennett

2002; Berríos-Rivera et al. 2004; Berríos-Rivera, Bennett, and San 2002b; Berríos-Rivera, Bennett, and San 2002a). Similarly, the manipulation of transhydrogenase enzymes in *E. coli* can shift byproduct yield. The overexpression of *sthA*, which encodes the soluble transhydrogenase enzyme, was shown to increase yield of both (S)-2 chloropropionate and poly(3-hydroxybutyrate), two byproducts produced in *E. coli* by anabolic reactions that utilize NADPH (Sanchez et al. 2006; Jan et al. 2013). For (S)-2 chloropropionate production, the deletion of *pntAB*, which encodes the membrane-bound transhydrogenase enzyme, also increased product yield (Jan et al. 2013).

A second strategy to increase cofactor availability is to replace the native enzyme with a non-native oxidoreductase with specificity for the opposite cofactor. For example, the NAD(H)-dependent glyceraldehyde-3-phosphate dehydrogenase (GAPD) in *E. coli* (encoded by the gene *gapA*) was replaced with the NADP(H)-dependent GAPD from *Clostridium acetobutylicum* (encoded by the gene *gapC*) to increase the production of lycopene and to increase the NADPH yield to drive a bioprocessing reaction (cyclohexanone to ε-caprolactone) where *E. coli* acts only as a source of reducing equivalents (Martínez et al. 2008). In another study, the native NAD(H)-dependent GAPD of *S. cerevisiae* (encoded by the genes *TDH1–3*) was supplemented with the NADP(H)-dependent GAPD from *Kluyveromyces lactis* (encoded by the gene *GDP1*) to increase fermentation of D-xylose to ethanol (Verho et al. 2003). These studies show that experimental implementation of such cofactor "swaps" is feasible and can result in promising increases in product yield. However, the question that remains to be answered is, which enzymes should be modified for maximum yield?

Computational studies have also investigated cofactor balancing. Most studies

to date have utilized constraint-based modeling, which represents the metabolic network by formulating the stoichiometry of metabolic reactions as a linear system of equations. Thermodynamic constraints (e.g., reaction irreversibility) and environmental parameters (e.g., nutrient availability) can be included in the formulation, and, by assuming that the system is in a mass-balanced steady state, linear optimization techniques can be used to identify optimal metabolic flux states and modifications in well-understood organisms (Price, Reed, and Palsson 2004). In a previous study, the authors reported the development of a bilevel optimization method called OptSwap to identify growth-coupled designs using modifications of oxidoreductase specificity and knockouts (King and Feist 2013). Similarly, Chung et al. 2013 presented a method called cofactor modification analysis (CMA) which optimized modifications of oxidoreductase specificity to improve the yield of terpenoids in yeast, and Lakshmanan et al. 2013 used the method to identify growth-coupled bioprocessing designs. Ghosh, Zhao, and Price 2011 used constraint-based modeling to analyze cofactor balancing for the specific case of yeast producing ethanol from L-arabinose and D-xylose. Chin et al. 2009 utilized the constraint-based model of *E. coli* to calculate theoretical yields of xylitol with various knockouts that affected cofactor balance, but they did not report a strategy to improve the cofactor balance. Despite the success of these targeted studies, a comprehensive analysis of the effects of changing cofactor specificity on a system-wide scale does not exist.

In this work, constraint-based modeling is utilized to identify optimal cofactor-specificity swaps for increasing theoretical yield in the genome-scale metabolic models of *E. coli* and *S. cerevisiae*. This work presents a global analysis of cofactor swapping for a large number of products across two important production organisms, and the

optimizations identify the minimal cofactor swaps necessary to maximize theoretical yield in the metabolic network.

5.2 Methods

5.2.1 Models and parameters

The *i*JO1366 metabolic reconstruction of *E. coli* K-12 MG1655 (Orth et al. 2011) and the *i*MM904 metabolic reconstruction of *S. cerevisiae* (Mo, Palsson, and Herrgård 2009) were used for the simulations in this work. Flux balance analysis (FBA) (Kauffman, Prakash, and Edwards 2003) and parsimonious flux balance analysis (pFBA) (Lewis et al. 2010a) were implemented in MATLAB as reported. As described previously, the *i*JO1366 oxidative stress reactions CAT, SPODM, and SPODMpp and the FHL reaction were constrained to zero (Orth et al. 2011), and the *i*JO1366 POR5 (pyruvate:ferredoxin oxidoreductase) reaction was made irreversible, as supported by biochemical data (Blaschkowski et al. 1982). In *i*MM904 simulations, free exchange of six sterols and fatty acids—ergosterol, zymosterol, hexadecenoate, and octadecanoate (saturated, monounsaturated, and polyunsaturated)—was allowed under anaerobic conditions, as reported by Mo, Palsson, and Herrgård 2009.

For cofactor swap optimizations, the substrate uptake rates (SURs) for the solitary carbon substrates in each simulation were constrained to a maximum uptake rate of 10 mmol gDW^{-1} h^{-1}. For aerobic simulations, the oxygen uptake rate was set to a maximum of 10 mmol gDW^{-1} h^{-1}. For these simulations, the minimum flux through the biomass objective function was set to 0.1 h^{-1}. For simulations with *S. cerevisiae* under anaerobic conditions with D-xylose substrate and *E. coli* under

anaerobic conditions with glycerol substrate, the *in silico* growth rate was less than 0.1 h^{-1}. To explore the effect of cofactor swaps in these cases, the minimum biomass requirement was set to 10% of the maximum growth rate.

For substrate uptake, only native gene content was considered. For growth of *S. cerevisiae* on D-xylose, the preexisting *i*MM904 model includes a D-xylose catabolism pathway consisting of xylose reductase (XR, EC 1.1.1.307) and xylitol dehydrogenase (XDH, EC 1.1.1.10), which were included in the model based on annotation for the genes GRE3 and XYL2, respectively (Mo, Palsson, and Herrgård 2009). However, under experimental conditions, D-xylose catabolism in yeast cannot support growth, and recombinant XR and XDH are necessary for fermentation of D-xylose (Bettiga, Hahn-Hägerdal, and Gorwa-Grauslund 2008; Ghosh, Zhao, and Price 2011; Bengtsson, Hahn-Hägerdal, and Gorwa-Grauslund 2009). Therefore, these reactions were considered to be native in the simulations, but their usage requires heterologous gene expression. Xylose isomerase (XI, EC 5.3.1.5), an alternative enzyme for D-xylose uptake which is not in *i*MM904, was also considered in the analysis of *S. cerevisiae* xylose catabolism (Bettiga, Hahn-Hägerdal, and Gorwa-Grauslund 2008).

Theoretical maximum yield ($Y_{P/S}$) is reported as the percentage of carbon consumed during substrate uptake that is converted to production of the target byproduct, subject to a minimum growth rate requirement at steady state (Feist et al. 2010):

$$Y_{P/S} = \frac{\text{byproduct production rate (mmol carbon gDW}^{-1}\ \text{h}^{-1})}{\text{substrate uptake rate (mmol carbon gDW}^{-1}\ \text{h}^{-1})}$$

5.2.2 Non-native pathways

To simulate cofactor swapping in realistic production pathways for non-native compounds, a literature search was performed to identify the most recent and successful experimentally-validated strain designs for production of non-native compounds in *E. coli*. Pathways were reconstructed by creating *in silico* reactions corresponding to the genes used in these experiments (Table 5.1, Supplementary Tables). *In silico* cofactor specificities were determined based on reports in the literature. Transport was assumed to be non-energy-coupled unless otherwise specified in the *i*JO1366 reconstruction or in the literature.

Table 5.1: **Non-native pathways reconstructed in *E. coli* for cofactor balance optimization.** [a]Two alternate pathways were reconstructed for these cases, according to the literature.

Product	Number of reactions (+number of transporters) added to model	Reducing equivalents consumed in non-native pathway	Reference
1,3-Propanediol	2 (+2)	1 NADPH	(Tang et al. 2009)
1,4-Butanediol	7 (+2)	3 NADH	(Yim et al. 2011)
meso-2,3-Butanediol	3 (+1)	1 NADH	(Ui et al. 2004)
R,R-2,3-Butanediol	3 (+1)	1 NADH	(Yan, Lee, and Liao 2009)
R-3-Hydroxybutyrate	4 (+2)[a]	1 NADPH	(Tseng et al. 2009)
S-3-Hydroxybutyrate	4 (+2)[a]	1 NADH	(Tseng et al. 2009)
3-Hydroxypropanoate	2 (+2)	2 NADPH	(Rathnasingh et al. 2012)
R-3-Hydroxyvalerate	5 (+2)	1 NADPH	(Tseng et al. 2010)
S-3-Hydroxyvalerate	5 (+2)	1 NADH	(Tseng et al. 2010)
Styrene	2 (+1)	None	(McKenna and Nielsen 2011)
p-Hydroxystyrene	2 (+1)	None	(Qi et al. 2007)
Lycopene	3 (+1)	8 NADPH	(Martínez et al. 2008)

5.2.3 Selection of the reaction sets for cofactor-specificity swaps

The sets of oxidoreductase reactions available for modification during optimizations were determined by locating central, high-flux oxidoreductase reactions, as previously reported (King and Feist 2013). In the metabolic models *i*JO1366 and

*i*MM904, all reactions utilizing NAD(H) or NADP(H) as a substrate were located. The reactions were sorted by flux magnitude after pFBA optimization for flux through the biomass objective function under conditions of aerobic and anaerobic growth on glucose and D-xylose minimal media. The reactions with highest flux under all conditions were selected. Lactate dehydrogenase, malic enzymes, and lactaldehyde dehydrogenase were added to the set for *E. coli* based on interest in the literature (Stols and Donnelly 1997; Wang et al. 2011; Zhang et al. 2007). Under anaerobic conditions, the NADH:oxidoreductase I reaction in *i*JO1366 was removed from the pool of enzymes during model reduction. Thus, 21 oxidoreductase reactions were chosen for *E. coli* under anaerobic conditions and 22 oxidoreductase reactions under aerobic conditions (Table 5.2), and 22 reactions were chosen for *S. cerevisiae* under both aerobic and anaerobic conditions (Table 5.3).

Table 5.2: **Oxidoreductase enzymes in the *E. coli* *i*JO1366 model chosen for cofactor balance optimizations.** [a]The flux through each reaction for maximum growth simulation using pFBA, relative to carbon uptake rate.

Key oxidoreductase enzymes in *E. coli*	Gene name	Model reaction abbreviation	Native electron carrier in *E. coli*	pFBA, relative flux scaled to carbon uptake[a]			
				Glucose aerobic	Xylose aerobic	Glucose anaerobic	Xylose anaerobic
NADH:ubiquinone oxidoreductase I	*nuoF, nuoAâĂŞC, nuoE, nuoGâĂŞN*	NADH16pp	NADH	1.98	1.98	0.0	0.0
Glyceraldehyde-3-phosphate dehydrogenase	*gapA*	GAPD	NADH	1.68	1.38	1.89	1.59
Glucose-6-phosphate dehydrogenase	*zwf*	G6PDH2r	NADPH	0.45	0.40	0.15	0.11
6-Phosphogluconate dehydrogenase	*gnd*	GND	NADPH	0.45	0.40	0.15	0.11
Pyruvate dehydrogenase	*lpd, aceE, aceF*	PDH	NADH	0.37	0.66	0.0	0.0
FAD reductase	*fre*	FADRx	NADH	0.16	0.14	0.06	0.04
Phosphoglycerate dehydrogenase	*serA*	PGCD	NADH	0.12	0.11	0.04	0.03
Isocitrate dehydrogenase	*icd*	ICDHyr	NADPH	0.08	0.07	0.03	0.02
Aspartate-semialdehyde dehydrogenase	*asd*	ASAD	NADPH	0.08	0.07	0.03	0.02
Acetohydroxy acid isomeroreductase (2-Acetolactate)	*mdh*	MDH	NADH	0.08	0.07	0.02	0.01
Methylene tetrahydrofolate dehydrogenase	*folD*	MTHFD	NADPH	0.06	0.06	0.02	0.02
Acetohydroxy acid isomeroreductase	*ilvC*	KARA1	NADPH	0.06	0.06	0.02	0.02
Homoserine dehydrogenase	*metL* OR *thrA*	HSDy	NADPH	0.05	0.05	0.02	0.01
3-Isopropylmalate dehydrogenase	*leuB*	IPMD	NADH	0.03	0.03	0.01	0.01
Shikimate dehydrogenase	*aroE*	SHK3Dr	NADPH	0.03	0.02	0.01	0.01
Dihydrodipicolinate reductase	*dapB*	DHDPRy	NADPH	0.03	0.02	0.01	0.01
Acetaldehyde dehydrogenase	*adhE* OR *mhpF*	ACALD	NADH	0.0	0.0	0.93	0.78
Alcohol dehydrogenase (ethanol)	*adhP* OR *adhE*	ALCD2x	NADH	0.0	0.0	0.93	0.78
L-1,2-Propanediol oxidoreductase	*fucO*	LCARR	NADH	0.0	0.0	0.0	0.0
D-Lactate dehydrogenase	*ldhA*	LDH_D	NADH	0.0	0.0	0.0	0.0
Malic enzyme (NADH)	*maeA*	ME1	NADH	0.0	0.0	0.0	0.0
Malic enzyme (NADPH)	*maeB*	ME2	NADPH	0.0	0.0	0.0	0.0

Table 5.3: **Oxidoreductase enzymes in the *i*MM904 model of *S. cerevisiae* chosen for cofactor balance optimizations.** [a]The flux through each reaction for maximum growth simulation using pFBA, relative to carbon uptake rate. [b]The irreversible reaction ALCD2ir was removed from the model in favor of the equivalent but reversible ALCD2x.

Key oxidoreductase enzymes in *S. cerevisiae*	**Gene symbol**	**Model reaction abbreviation**	**Native electron carrier in *S. cerevisiae***	**pFBA, relative flux scaled to carbon uptake[a]**			
				Glucose aerobic	**Xylose aerobic**	**Glucose anaerobic**	**Xylose anaerobic**
Glyceraldehyde-3-phosphate dehydrogenase	TDH1 or TDH2 or TDH3	GAPD	NADH	1.68	1.20	1.88	1.63
Alcohol dehydrogenase (ethanol)	ADH1 or ADH4 or ADH5 or SFA1	ALCD2x[b]	NADH	1.01	0.78	1.67	1.57
NADH dehydrogenase	NDE1 or NDE2	NADH2-u6cm	NADH	0.69	1.44	0.03	0.01
Isocitrate dehydrogenase	IDP2	ICDHy	NADPH	0.30	0.11	0.02	0.01
Glutamate dehydrogenase	GDH1 or GDH3	GLUDyi	NADPH	0.26	0.22	0.00	0.00
Phosphoglycerate dehydrogenase	SER3 or SER33	PGCD	NADH	0.03	0.03	0.01	0.00
Acetaldehyde dehydrogenase	ALD6	ALDD2y	NADPH	0.02	0.02	0.01	0.00
6-Phosphogluconate dehydrogenase	GND1âĂŞ2	GND	NADPH	0.02	0.59	0.00	0.00
Glucose-6-phosphate dehydrogenase	ZWF1	G6PDH2	NADPH	0.02	0.59	0.00	0.00
Aspartate-semialdehyde dehydrogenase	HOM2	ASADi	NADPH	0.02	0.02	0.01	0.00
Homoserine dehydrogenase	HOM6	HSDxi	NADH	0.02	0.02	0.01	0.00
3-Isopropylmalate dehydrogenase	LEU2	IPMD	NADH	0.02	0.01	0.01	0.00
L-Aminoadipate-semialdehyde dehydrogenase	LYS2, LYS5	AASAD2	NADH	0.02	0.01	0.01	0.00
Saccharopine dehydrogenase (L-glutamate forming)	LYS9	SACCD1	NADPH	0.02	0.01	0.01	0.00
Saccharopine dehydrogenase (L-lysine forming)	LYS1	SACCD2	NADH	0.02	0.01	0.01	0.00
Shikimate dehydrogenase	ARO1	SHK3D	NADPH	0.01	0.01	0.01	0.00
Glycerol-3-phosphate dehydrogenase	GPD1	G3PD1ir	NADH	0.00	0.00	0.00	1.00
Xylose reductase	GRE3	XYLR	NADPH	0.00	1.00	0.00	1.00
Xylitol dehydrogenase	XYL2	XYLTD_D	NADH	0.00	1.00	0.00	1.00
Glutamate synthase	GLT1	GLUSx	NADH	0.00	0.00	0.10	0.03
Malate dehydrogenase	MDH2	MDH	NADH	0.00	0.00	0.08	0.03
Glycerol dehydrogenase	GCY1	GLYCDy	NADPH	0.00	0.00	0.00	1.00

5.2.4 MILP formulation

A MILP problem was formulated to find the set of cofactor-specificity swaps that maximize the theoretical yield of chemical production with a minimum biomass requirement (Fig. 5.1). The MILP formulation is functionally identical to the one used by Chung et al. 2013, but here it is implemented differently. As described in our previous report, the non-native oxidoreductase enzymes are added to the system and coupled so that either the native enzyme or the non-native enzyme is active (King and Feist 2013). The final formulation of the MILP problem can be stated follows, where S is the stoichiometric matrix, I is the set of metabolites, J is the set of reactions, D is the set of oxidoreductase reactions, s and t are decision variables defining the state of each swap, and L is the maximum number of swaps allowed:

$$
\begin{aligned}
&\text{maximize } v_{chemical} && \\
&\text{subject to } \sum_{j \in J} S_{ij} v_j = 0 && \forall i \in I \\
&\qquad LB_j \leq v_j \leq UB_j && \forall j \in J \\
&\qquad v_{biomass} \geq min\_biomass && \\
&\qquad s_d \in \{0,1\} && \forall d \in D \\
&\qquad t_d \in \{0,1\} && \forall d \in D \\
&\qquad s_d + t_d = 1 && \forall d \in D \\
&\qquad s_d LB_{x(d)} \leq v_{x(d)} \leq s_d UB_{x(d)} && \forall d \in D \\
&\qquad t_d LB_{y(d)} \leq v_{y(d)} \leq t_d UB_{y(d)} && \forall d \in D \\
&\qquad \sum_{d \in D} (1 - s_d) \leq L &&
\end{aligned}
$$

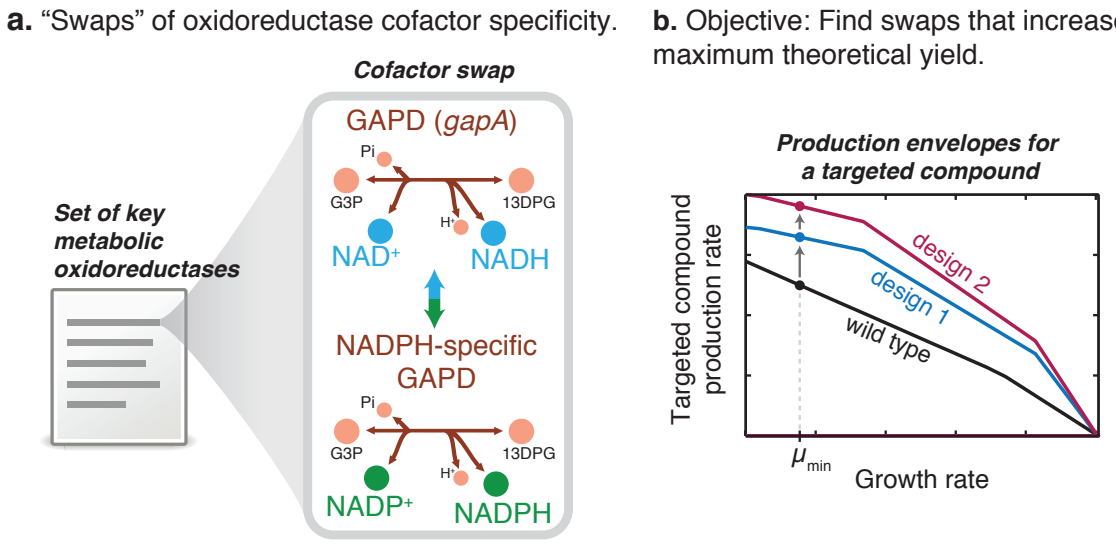

Figure 5.1: **Cofactor specificity modifications of oxidoreductase reactions can improve the maximum theoretical yield of metabolic byproducts.** In this work, (a) oxidoreductase reactions in the metabolic model are identified as candidates for *in silico* cofactor "swaps," and (b) an optimization procedure identifies swaps that maximize the theoretical yield of a targeted compound.

5.2.5 Non-unique solutions

In many cases, the solution to the MILP problem was non-unique. To investigate the diversity of cofactor swaps that could produce the same results, an exhaustive search method was utilized. For each solution, all swaps that improved maximum theoretical yield to within 99% of the optimal solution were found. These were identified as solution "groups" (Fig. 5.2–5.4). The same procedure was repeated to investigate secondary swaps, after first swapping a reaction from the first-swap group. Similar results could be achieved utilizing MILP solvers that enumerate solutions, such as the Solution Pool feature in CPLEX 11+ (IBM, Armonk, NY, USA).

a

		Anaerobic						Aerobic					
		Glucose			Xylose			Glucose			Xylose		
Product	No. of Carbons	wt	1 swap	2 swaps	wt	1 swap	2 swaps	wt	1 swap	2 swaps	wt	1 swap	2 swaps
L-Cysteine	3	0.10	0.18 GAPD		0.06	0.11 GAPD		0.39	0.43 GAPD	0.45 Group 8	0.41	0.44 GAPD	0.46 PGCD
L-Homoserine	4	0.26	0.45 GAPD	0.48 3+ rxns	0.15	0.31 GAPD		0.79	0.81 3+ rxns	0.83 GAPD	0.77	0.78 3+ rxns	0.80 Group 9
L-Threonine	4	0.23	0.40 GAPD		0.13	0.24 GAPD		0.77	0.79 3+ rxns	0.80 GAPD	0.75	0.76 3+ rxns	0.77 Group 9
1,5-Diaminopentane	5	0.21	0.37 GAPD		0.12	0.23 GAPD		0.59	0.60 GAPD	0.61 3+ rxns	0.57	0.58 GAPD	0.59 3+ rxns
L-Lysine	6	0.26	0.45 GAPD		0.15	0.27 GAPD		0.70	0.72 GAPD	0.73 3+ rxns	0.68	0.70 GAPD	0.70 3+ rxns
L-Isoleucine	6	0.28	0.47 GAPD	0.54 PDH	0.16	0.38 GAPD	0.39 3+ rxns	0.67	0.69 GAPD	0.70 3+ rxns	0.65	0.66 GAPD	0.67 3+ rxns
L-Leucine	6	0.40	0.56 GAPD	0.58 Group 1	0.25	0.40 GAPD		0.70	0.72 GAPD	0.72 3+ rxns	0.67	0.69 GAPD	0.70 3+ rxns
Glycine	2	0.17	0.24 GAPD	0.26 3+ rxns	0.10	0.15 GAPD		0.66	0.71 GAPD	0.74 3+ rxns	0.69	0.74 GAPD	0.78 PGCD
5-Methylthio-D-ribose	6	0.07	0.10 GAPD		0.04	0.06 GAPD		0.24	0.26 GAPD	0.26 3+ rxns	0.24	0.25 GAPD	0.25 3+ rxns
Spermidine	7	0.08	0.12 GAPD		0.05	0.07 GAPD		0.28	0.30 GAPD	0.30 3+ rxns	0.28	0.29 GAPD	0.29 3+ rxns
L-Proline	5	0.28	0.41 GAPD		0.16	0.25 GAPD		0.76	0.79 GAPD	0.79 3+ rxns	0.73	0.76 GAPD	0.76 3+ rxns
Cys-Gly	5	0.09	0.12 GAPD		0.05	0.08 GAPD		0.37	0.42 GAPD	0.43 Group 8	0.38	0.43 GAPD	0.44 3+ rxns
L-Aspartate	4	0.43	0.60 GAPD		0.25	0.36 GAPD		1.11	1.15 GAPD		1.08	1.11 GAPD	
Reduced glutathione	10	0.14	0.19 GAPD		0.08	0.11 GAPD		0.59	0.67 GAPD	0.68 3+ rxns	0.62	0.68 GAPD	0.68 3+ rxns
L-Serine	3	0.28	0.38 GAPD		0.16	0.23 GAPD		0.95	0.97 Group 6	0.97 3+ rxns	0.93	0.95 Group 6	0.95 3+ rxns
Thymine	5	0.13	0.17 GAPD		0.07	0.10 GAPD		0.60	0.67 GAPD	0.69 Group 8	0.62	0.70 GAPD	0.70 3+ rxns
Putrescine	4	0.17	0.22 GAPD		0.10	0.14 GAPD		0.58	0.60 GAPD		0.56	0.58 GAPD	
Glycerol	3	0.28	0.38 GAPD		0.16	0.23 GAPD		0.70	0.71 3+ rxns	0.71 3+ rxns	0.68	0.68 3+ rxns	0.69 3+ rxns

Percent increase in $Y_{P/S}$ from wildtype: 0% 10% 20% 30% 40% 50% 60% 70% >80%

b

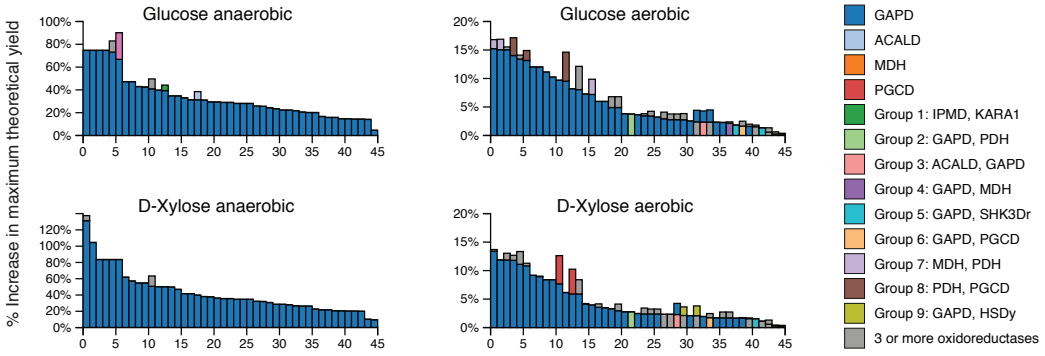

Figure 5.2: **Results from optimizing the cofactor specificity of oxidoreductases in the *E. coli* *i*JO1366 model to increase the maximum theoretical yield of native metabolic compounds.** (a) Maximum theoretical yield for wildtype (wt) metabolic content and after one oxidoreductase swap (1 swap) and two oxidoreductase swaps (2 swap). Colors indicate the percent increase in maximum theoretical yield compared to wildtype. (b) Reactions found to most influence production of products are ranked and plotted by increase in theoretical maximum yield under each condition, and each color indicates a reaction or group of reactions that can be swapped to reach the optimal theoretical maximum yield.

a

		Anaerobic				Aerobic			
		Glucose		Xylose*		Glucose		Xylose	
Product	No. of carbons	wt	swap	wt	swap	wt	swap	wt	swap
L-Aspartate	4	0.33	0.63 Group 2	0.20	0.70 ALCD2x	0.94		0.93	
L-Valine	5	0.41	0.78 Group 2	0.26	0.83 Group 1	0.75	0.78 Group 2	0.70	0.77 ALCD2x
L-Alanine	3	0.50	0.94 Group 2	0.31	1.00 Group 2	0.90	0.94 Group 2	0.84	0.93 ALCD2x
L-Serine	3	0.17	0.29 Group 1	0.10	0.50 ALCD2x	0.87	0.94 4+ rxns	0.76	0.93 4+ rxns
L-Cysteine	3	0.07	0.11 Group 1	0.04	0.18 ALCD2x	0.30	0.39 ALCD2x	0.29	0.42 ALCD2x
L-Lactate	3	0.17	0.26 Group 1	0.10	0.45 ALCD2x	0.77	0.80 ALCD2x	0.72	0.78 ALCD2x
L-Methionine	5	0.09	0.14 Group 1	0.06	0.23 ALCD2x	0.38	0.52 ALCD2x	0.37	0.50 ALCD2x
L-Phenylalanine	9	0.20	0.30 Group 1	0.13	0.51 ALCD2x	0.74	0.75 4+ rxns	0.69	0.73 ALCD2x
Pyruvate	3	0.40	0.53 Group 1	0.31	0.78 ALCD2x	0.94		0.93	
Thymine	5	0.08	0.10 Group 1	0.05	0.18 Group 1	0.36	0.39 Group 3	0.33	0.43 ALCD2x
sn-Glycero-3-phosphocholine	8	0.05	0.06 Group 1	0.03	0.11 Group 1	0.20	0.22 Group 3	0.20	0.24 ALCD2x
Glycine	2	0.11	0.15 Group 1	0.08	0.25 ALCD2x	0.53	0.76 ALCD2x	0.48	0.81 ALCD2x
L-Lysine	6	0.10	0.13 Group 2	0.06	0.23 Group 1	0.56	0.67 4+ rxns	0.50	0.65 ALCD2x
L-Proline	5	0.17	0.22 Group 2	0.10	0.38 Group 1	0.75	0.79 4+ rxns	0.70	0.77 ALCD2x
L-Asparagine	4	0.13	0.18 Group 2	0.08	0.30 Group 1	0.56	0.60 4+ rxns	0.55	0.65 Group 1
episterol	28	0.00		0.00		0.24	0.32 Group 1	0.21	0.36 ALCD2x
fecosterol	28	0.00		0.00		0.24	0.32 Group 1	0.21	0.36 ALCD2x
zymosterol	27	0.00		0.00		0.24	0.32 Group 1	0.21	0.36 ALCD2x
Ergosterol	28	0.00		0.00		0.23	0.30 Group 1	0.20	0.33 ALCD2x

Percent increase in $Y_{P/S}$ from wildtype: 0% 10% 20% 30% 40% 50% 60% 70% >80%

b

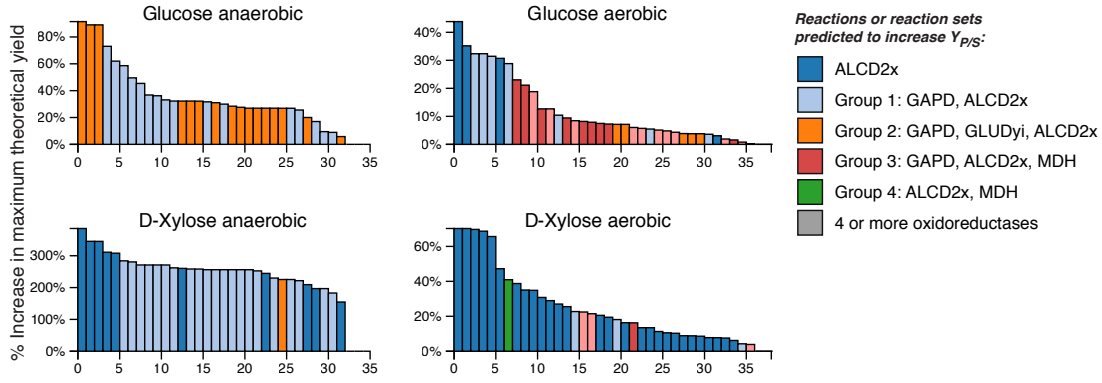

Figure 5.3: **Results from optimizing the cofactor specificity of oxidoreductases in the *S. cerevisiae* *i*MM904 model to increase the maximum theoretical yield of native metabolic compounds.** (a) Carbon yield for wildtype (wt) metabolic content and after one oxidoreductase swap (swap). Colors indicate the percent increase in maximum theoretical yield compared to wildtype. (b) Reactions found to most influence production of products are ranked and plotted by increase in theoretical maximum yield under each condition, and each color indicates a reaction or group of reactions that can be swapped to reach the optimal theoretical maximum yield. Note that the D-xylose uptake reactions (XR/XDH) exist in the *i*MM904 model and thus in these simulations, but this pathway is not active in wildtype *S. cerevisiae* (see Methods). *Minimum growth rate set to 10% of the maximum growth rate.

	Anaerobic									Aerobic								
	Glucose			Xylose			Glycerol*			Glucose			Xylose			Glycerol		
Product	wt	1 swap	2 swaps	wt	1 swap	2 swaps	wt	1 swap	2 swaps	wt	1 swap	2 swaps	wt	1 swap	2 swaps	wt	1 swap	2 swaps
1,3-Propanediol	0.29	0.39 GAPD	0.46 PDH	0.17	0.31 GAPD	0.32 PDH	0.69	0.75 Group 1	0.78 PDH	0.63	0.63 4+ rxns	0.63 4+ rxns	0.61	0.61 4+ rxns	0.61 4+ rxns	0.71	0.72 4+ rxns	0.72 4+ rxns
1,4-Butanediol	0.51			0.43			0.64			0.64			0.62			0.68		
R,R-2,3-Butanediol	0.63			0.51			0.33			0.66			0.64			0.65		
meso-2,3-Butanediol	0.63			0.51			0.33			0.66			0.64			0.65		
R-3-Hydroxybutyrate	0.53	0.72 GAPD	0.73 4+ rxns	0.40	0.60 GAPD		0.36	0.40 GAPD		0.78	0.79 Group 2	0.81 GAPD	0.75	0.77 Group 2	0.79 GAPD	0.80	0.82 GAPD	0.83 4+ rxns
S-3-Hydroxybutyrate	0.63	0.74 GAPD		0.49	0.60 GAPD		0.40			0.79	0.81 Group 2	0.81 4+ rxns	0.77	0.78 Group 2	0.79 4+ rxns	0.81	0.83 4+ rxns	
3-Hydroxypropanoate	0.32	0.55 GAPD	0.61 PDH	0.19	0.41 GAPD		0.39			0.84	0.87 Group 1	0.89 PDH	0.81	0.84 Group 1	0.86 PDH	0.91	0.92 4+ rxns	0.92 4+ rxns
R-3-Hydroxyvalerate	0.48	0.61 GAPD		0.28	0.40 GAPD		0.34			0.75	0.76 Group 1		0.72	0.73 Group 1		0.78		
S-3-Hydroxyvalerate	0.59	0.61 Group 1		0.36	0.40 GAPD		0.34			0.76	0.76 4+ rxns		0.73	0.73 4+ rxns		0.78		
Styrene	0.24	0.29 GAPD		0.14	0.18 GAPD		0.15			0.67	0.67 Group 3		0.65	0.66 Group 4		0.71		
p-Hydroxystyrene	0.23	0.27 GAPD		0.13	0.17 GAPD		0.14			0.69	0.70 Group 3		0.67	0.68 Group 3		0.72	0.72 4+ rxns	
Lycopene	0.36	0.41 GAPD		0.00			0.22			0.66	0.67 4+ rxns		0.00			0.70		

Percent increase in $Y_{P/S}$ from wildtype: 0% 10% 20% 30% 40% 50% 60% 70% >80%

Reaction sets predicted to increase $Y_{P/S}$:

Group 1: GAPD, PDH
Group 2: ACALD, GAPD, PDH
Group 3: GAPD, SHK3Dr
Group 4: GAPD, PDH, SHK3Dr

Figure 5.4: **The effect of swapping cofactor specificity of oxidoreductases for the production of non-native compounds in *E. coli*.** The table shows the maximum theoretical yield for wildtype (wt) and after one oxidoreductase swap (1 swap) and two oxidoreductase swaps (2 swaps), and the selected reaction or groups of reactions that can be swapped to reach the optimal theoretical maximum yield. Colors indicate the percent increase in maximum theoretical yield compared to wildtype. *Minimum growth rate set to 10% of the maximum growth rate.

5.2.6 Sensitivity analysis

A number of parameters can effect the simulation of theoretical yields; these include the oxygen uptake rate (OUR), the minimum growth rate (μ_{min}), and the SUR. A sensitivity analysis was performed for aerobic (OUR = 10 mmol gDW^{-1} h^{-1}) and anaerobic (OUR = 0) conditions. For a range of SUR values from 5 to 20 mmol gDW^{-1} h^{-1}, corresponding to a realistic range for glucose uptake in *E. coli* (Covert et al. 2004), and μ_{min} values from 0 to maximum growth, the optimal single swap was identified and the theoretical yield improvement was calculated.

5.2.7 Determining cofactor usage

To determine the NADPH usage of each production pathway, pFBA simulations were run optimizing for flux through each production pathway with zero growth ($\mu_{min} = 0$) and with the transhydrogenase enzymes constrained to zero flux. Then, the flux through all reactions consuming NADPH was summed. This procedure was repeated for all carbon-containing metabolites that could be exported by the metabolic model.

All simulations were performed using MATLAB (The MathWorks Inc., Natick, MA, USA) and the COBRA Toolbox (Becker et al. 2007) software packages with TOMLAB (Tomlab Optimization Inc., San Diego, CA, USA) and Gurobi (Gurobi Optimization, Inc., Houston, TX, USA) MILP solvers.

5.3 Results

5.3.1 Native Pathways

To determine the effect of cofactor swaps on product yield, simulations were performed to optimize production of 81 and 154 target compounds in *E. coli* and *S. cerevisiae*, respectively, while allowing one and two swaps of oxidoreductase specificity out of the pool of oxidoreductase reactions (Tables 5.2, 5.3). All carbon-containing molecules that can be exported by the metabolic models were considered. Increases in theoretical maximum yield were seen for many native products after one swap in both the *E. coli* *i*JO1366 and *S. cerevisiae* *i*MM904 models (Fig. 5.2–5.3). The theoretical yields for these pathways are dependent on the specific values selected for SUR, OUR, and μ_{min}. For these optimizations, the parameters were chosen to represent a possible

scenario for a bioprocessing strain (a maximum value of SUR = 10 mmol gDW^{-1} h^{-1} and OUR = 10 mmol gDW^{-1} h^{-1}, and a minimum value of μ_{min} = 0.1 h^{-1}). Studies have reported specific growth rates in the range of 0.03–0.43 h^{-1} for production strains of *E. coli* (Martínez et al. 2008; Murarka et al. 2008; Bettiga, Hahn-Hägerdal, and Gorwa-Grauslund 2008; Qian, Xia, and Lee 2009; Rathnasingh et al. 2009), and the SUR was chosen to match observed glucose uptake under aerobic conditions (Covert et al. 2004). To further justify the selection of these parameters for the global analysis, a sensitivity analysis was performed, discussed in Section 5.3.3.

E. coli

During anaerobic growth on glucose, cofactor optimizations of *i*JO1366 showed an increase in theoretical yield greater than 20% for 29 native products after one cofactor swap. The effect of a second swap was small in most simulations, and further swaps (> 2 swaps) had no effect on product yield (Fig. 5.2). For anaerobic growth with D-xylose as a substrate, the increases in maximum theoretical yield were slightly greater in magnitude but qualitatively similar to those with glucose as a substrate. Aerobically, smaller increases in maximum theoretical yield were observed for all native targets and cofactor swaps. A 10–15% increase in theoretical yields was seen for 15 products (e.g., thymine, agmatine, L-arginine, D-alanine) with glucose as a substrate (Fig. 5.2A), and similar increases were seen with D-xylose as a substrate.

In *i*JO1366, converting the cofactor specificity of GAPD (EC 1.2.1.13) from production of NADH to production of NADPH has a global effect on the theoretical yield of native products. When considering the 45 metabolites with the greatest change in yield after cofactor swapping, simulated under 4 media conditions, it was

observed that in all 180 cases swapping the native GAPD for the NADPH-dependent GAPD resulted in the greatest increase in yield, and for 155 of 180 cases no other cofactor swap can produce the theoretical maximum yield (the solutions are unique) (Fig. 5.2B). In order to investigate the relationship between cofactor swapping and the NADPH usage of production pathways, pFBA simulations (Lewis et al. 2010a) were run maximizing production for each target molecule with zero growth. Then, for each simulation, the sum of the fluxes through all reactions that consume NADPH was calculated. This sum is an indication of the NADPH usage for producing a particular metabolite at the theoretical limit of production at steady state. A correlation was observed between the NADPH usage of production pathways and the improvement in theoretical yield after swapping GAPD (Pearson's r = 0.76, Supplementary Fig. 1). This correlation points to the finding that producing more NADPH can can leads to a higher maximum theoretical yield for a number of desired production targets.

S. cerevisiae

Cofactor swapping optimizations for *S. cerevisiae* also showed increases in theoretical yield after one cofactor swap (Fig. 5.3). *S. cerevisiae* does not contain an active pathway for D-xylose catabolism. However, many studies have reported the introduction of a D-xylose uptake pathway containing the enzymes XR, which oxidizes NADPH, and XDH, which reduces NAD^+ to NADH, and high D-xylose consumption has been achieved with the pathway from *Pichia stipitis* (Ghosh, Zhao, and Price 2011; Bengtsson, Hahn-Hägerdal, and Gorwa-Grauslund 2009). These reactions were utilized in the analysis of native yeast metabolism because they are included in *i*MM904 (as discussed in the Methods). Maximum theoretical yields of many products increase

250–350% after one swap when this pathway is modeled and D-xylose is present as the substrate. XI is an alternative, cofactor-balanced pathway for xylose uptake (Bettiga, Hahn-Hägerdal, and Gorwa-Grauslund 2008). When this enzyme is simulated, the predicted flux distribution utilizes XI in preference to XR/XDH. Using XI increases yield to the maximum theoretical yield achieved by cofactor swaps. Thus, modifying the D-xylose uptake pathway is an alternative to implementing the optimal cofactor swap.

Under aerobic conditions, increases in maximum theoretical yield were smaller than under anaerobic conditions, as in the *E. coli* metabolic model. However, the theoretical yield increases were still significant. For example, maximum theoretical yields of episterol, fecosterol, zymosterol, and ergosterol increased 29–32% with glucose as a substrate and 66–70% with D-xylose as a substrate after swapping the cofactor specificity of the alcohol dehydrogenase reaction ALCD2x (EC 1.1.1.1).

In the yeast metabolic model, swapping the cofactor specificity of a second oxidoreductase had no impact on theoretical yield. Also, in comparison to the results for *E. coli*, more of the yeast yield optimizations resulted in non-unique solution groups (Fig. 5.3B). For instance, under anaerobic conditions with glucose substrate, the L-serine yield reached a maximum of 0.47 Cmol / Cmol substrate after swapping the cofactor specificity of either oxidoreductase enzyme in the solution group (Group 1: GAPD or ALCD2x). In contrast to *E. coli*, the reaction that appeared most often in solutions is ALCD2x, and GAPD appears secondarily.

5.3.2 Non-native pathways in *E. coli*

Major building block molecules that have been heavily studied as non-native products of *E. coli* were selected for the cofactor swapping analysis. These products include (1) the polyester building-blocks 1,3-propanediol, 1,4-butanediol, 2,3-butanediol, 3-hydroxybutyrate, and 3-hydroxyvalerate, and (2) the polyvinyl building blocks styrene and hydroxystyrene (Adkins et al. 2012). Production of lycopene, a red-colored carotenoid sold commercially as a colorant and a nutritional supplement, was also simulated. When applicable, production pathways for alternate sterioisomers were considered, so that, in total, twelve non-native pathways were reconstructed in the *E. coli* model (Table 5.1).

Optimization of cofactor swapping increased the theoretical maximum yield of 9 of 12 non-native pathways (Fig. 5.4). The effects were much greater anaerobically; under aerobic conditions, less than 10% increases were observed. For three products (1,4-butanediol, R,R-2,3-butanediol, and meso-2,3-butanediol), the optimization did not find any solution, indicating that modifying cofactor specificity is not predicted to increase the maximum yield of these products. For these products, a sensitivity analysis was performed (Section 5.3.3).

5.3.3 NADPH yield and parameter sensitivity

The effect of cofactor swaps was observed to depend on the values of SUR, OUR, μ_{min} selected for the simulation. To better understand the relationship between these parameters and the observed improvements in theoretical yield, a simple scenario where *E. coli* is used to regenerate reducing equivalents of NADPH for the conversion of cyclohexanone to ε-caprolactone was explored. Martínez et al. 2008 performed

a GAPD cofactor swap in *E. coli* and observed an increase in the yield of NADPH reducing equivalents for this chemical conversion. By examining this simple case, the effect of the cofactor swaps on NADPH yield can be isolated from other factors such as shifts in carbon metabolism.

Cofactor optimizations were performed for a range of SUR and μ_{min} values, under aerobic and anaerobic conditions. The NADPH yield in *E. coli* was observed to decrease as the biomass production approaches a maximum (Fig. 5.5A,D). However, swapping the cofactor specificity of the GAPD reaction resulted in increased maximum theoretical yield compared to the wildtype (Fig. 5.5C,F). It was also observed that the selection of GAPD as the optimal swap was consistent across parameters for aerobic and anaerobic conditions (see Discussion for further analysis).

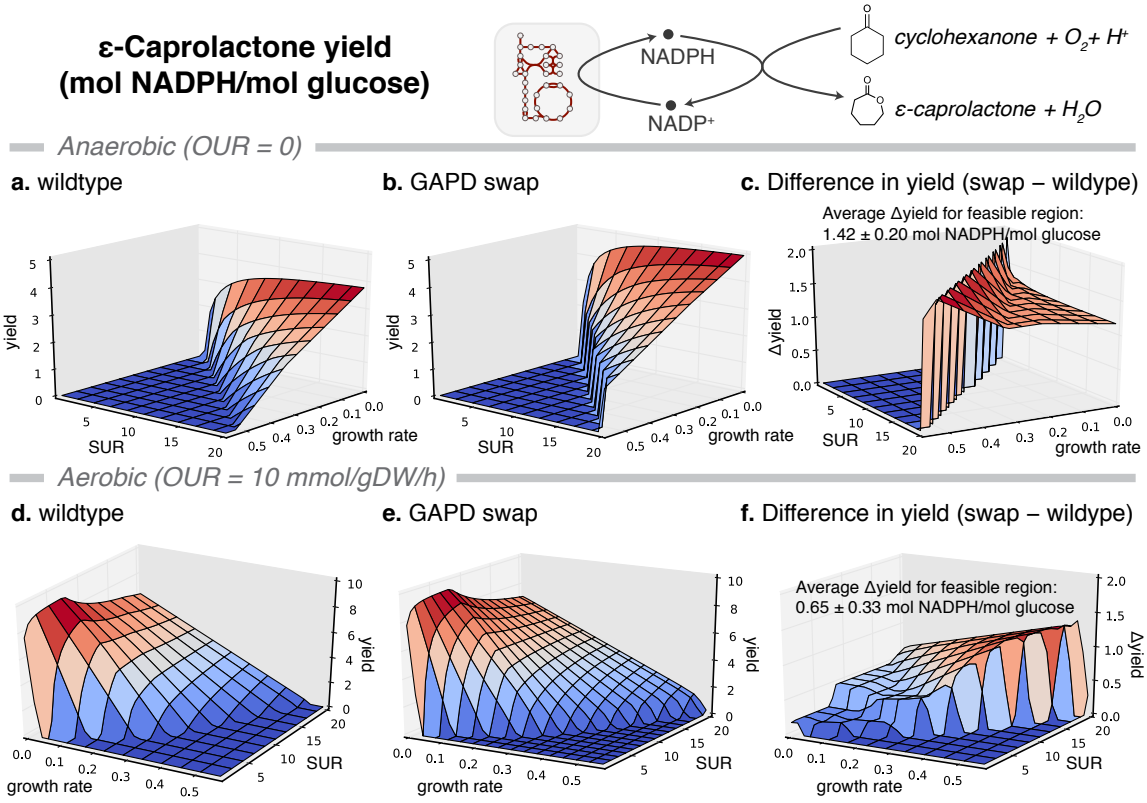

Figure 5.5: **A sensitivity analysis on the impact of cofactor swapping when varying modeling parameters.** The conversion of cyclohexanone to ε-caprolactone requires reducing equivalents which can be provided by *E. coli* growing in glucose minimal media. For this conversion, a sensitivity analysis was performed for (a–c) anaerobic and (d–f) aerobic conditions. Substrate uptake rate (SUR) and the minimum growth rate (μ_{min}) were varied across a wide range of values, and it was observed that swapping GAPD increased carbon yield across all parameters. For infeasible μ_{min} values, no growth is possible, so yield is zero (dark blue regions). Across all feasible datapoints, the average and standard deviation of the Δyield are given and can be compared to the increase in yield observed by Martínez et al. 2008 of 1.25 ± 0.19 mol NADPH/mol glucose.

For the non-native compounds 1,4-butanediol and 2,3-butanediol, running the cofactor swap optimization with the default parameters did not identify any opportunities to improve theoretical yield. The same sensitivity analysis was performed for these target molecules to determine whether, under any other conditions, a cofactor swap might be beneficial (Supplementary Fig. 5.2). It was observed that under a certain range of SUR and μ_{min} values, when growth is near the maximum growth rate, a swap does improve theoretical yield for 1,4-butanediol, under aerobic and anaerobic

conditions. Specifically, under anaerobic conditions, it was found that there were 790 feasible parameter sets based on the parameters of the sensitivity analysis (see Methods). Out of these, the cofactor swap was beneficial for 292 of the 790 feasible parameter sets (i.e. 37%, and the increase can be seen in Supplementary Fig. 5.2). Furthermore, when examining both 2,3-butanediol isomers during the sensitivity analysis, the cofactor swapping did not improve theoretical yield at any values of SUR and μ_{min}.

Finally, the sensitivity analysis was performed for L-cysteine production (on glucose minimal media in *E. coli*) to further examine whether the optimizations for a single parameter set (Fig. 5.2–5.4) are sensitive to the selection of SUR, OUR, and μ_{min}. It was observed that theoretical yields were improved by the GAPD cofactor swap across a broad range of SUR, OUR, μ_{min} values (Supplementary Fig. 5.3). Thus, the parameter set chosen for the global analysis, where SUR and μ_{min} are relatively small, leads to accurate estimations of the maximum effect of a cofactor swap (e.g. L-cysteine anaerobically), or even underestimations of the maximum effect of a cofactor swap (e.g. L-cysteine aerobically, 1,4-butanediol), and the general trends observed in the global analysis are a meaningful representation of the total impact of cofactor swapping on theoretical yields.

5.4 Discussion

5.4.1 Cofactor swaps for certain enzymes have a global impact on theoretical yields

This study presents a computational analysis to determine optimal cofactor-specificity swaps of oxidoreductase enzymes in genome-scale metabolic models, specifically for the production organisms *E. coli* and *S. cerevisiae*. Increases in theoretical yield were observed for many products of *E. coli* and *S. cerevisiae* metabolism after one cofactor swap, and swapping certain reactions (esp. GAPD, ALCD2x) was seen to have a global benefit for theoretical yields. The theoretical yield improvements were found for a number of native products that are produced on an industrial scale and have been considered for bioproduction using *E. coli* or *S. cerevisiae*. These include the native compounds L-lysine, L-isoleucine, L-proline, L-serine, L-threonine, L-aspartate, L-lactate, 1,5-diaminopentane (cadaverine), and putrescine, and the non-native compounds 3-propanediol, 3-hydroxybutyrate, 3-hydroxypropanoate, 3-hydroxyvalerate, styrene, and lycopene. (For review articles describing the bioproduction of these compounds in *E. coli*, see Becker and Wittmann 2012; Jang et al. 2012.)

Nearly all single swaps selected by the optimization for native and non-native products in *E. coli* were for the GAPD enzyme. This is a central enzyme in glycolysis, and the change in electron carrier specificity causes a shift in central metabolism toward NADPH production. Because the GAPD cofactor swap has such a global impact on biosynthesis in *E. coli* metabolism, a stable strain with NADPH-dependent GAPD could be a useful starting point for engineering *E. coli* for production of any of the products reported here (Fig. 5.2, 5.4). Published results have shown that

experimentally replacing the GAPD in *E. coli* with an NADPH-dependent GAPD increases yield and productivity of lycopene and e-caprolactone (Martínez et al. 2008). The optimizations presented here demonstrate that the GAPD cofactor swap is an ideal choice. However, for some products, swapping the cofactor specificity of other important oxidoreductase enzymes can have the same affects on yield (e.g., pyruvate dehydrogenase [PDH, EC 1.2.4.1], malate dehydrogenase [MDH, EC 1.1.1.37], and phosphoglycerate dehydrogenase [PGCD, EC 1.1.1.95]).

The environmental conditions (aerobicity, substrate) were a major determinant of the impact of cofactor swapping. Simulations with D-xylose as a substrate showed greater increases in yield after swapping than those with glucose substrate (Fig. 5.2–5.4), and, with glycerol as a substrate, very little improvement was possible with cofactor swapping (Fig. 5.4). Aerobicity also had a major effect. In anaerobic simulations, cofactor swapping had a much greater impact than in aerobic simulations. SUR and μ_{min} also affected the results of cofactor swapping (as described below). Some cases (e.g. aerobic production of glycine in yeast) are exceptions to these trends, so the relationship between environment conditions, cofactor swapping, and theoretical yield is complex, and the cofactor balance depends on the exact metabolic state of a cell.

Simulations utilizing the *S. cerevisiae* metabolic network differed from *E. coli* in that more cofactor swaps could theoretically produce the same improvement in yield, as seen in the many non-unique solutions (Fig. 5.3B). Thus, one has more flexibility to choose interventions as various swaps can produce the same effect. The exact enzyme to modify can be determined by other factors, such as the strength of regulation of the native enzyme or the availability of an appropriate enzyme with alternate cofactor specificity. As in *E. coli*, the native GAPD in *S. cerevisiae* has

been experimentally replaced with a NADPH-dependent GAPD, resulting in higher D-xylose fermentation (Verho et al. 2003) and the same design has been patented for the production of many products (ethanol, lactic acid, polyhydroxyalkanoates, amino acids, fats, vitamins, nucleotides) (Londesborough et al. 2003). However, the results of cofactor swapping optimizations demonstrate that modifying ALCD2x can have the same impact on theoretical yield, and, in many cases, any one of a group of enzymes can increase theoretical yield to optimal levels.

5.4.2 Simulated theoretical yield improvement matches experimental observations for a GAPD swap

To determine how cofactor swapping modifies the capabilities of the metabolic network, a simple scenario was explored where *E. coli* generates reducing equivalents of NADPH that drive a bioprocessing reaction. In such a scenario, the effect of cofactor balance can be isolated from biomass production and carbon metabolism. In an experimental study, Martínez et al. 2008 compared the ability of wildtype and GAPD-swap *E. coli* strains to produce reducing equivalents of NADPH for a reaction that converts cyclohexanone to ε-caprolactone (utilizing one mole of NADPH per mole of ε-caprolactone produced). The authors reported an increase in NADPH yield from 1.72 ± 0.19 to 2.97 ± 0.05 mol NADPH / mol glucose. For *in silico* optimizations of ε-caprolactone production, GAPD is the optimal cofactor swap across the range of SUR, OUR, and μ_{min} parameters, and swapping GAPD increases theoretical yield for all parameter values (Fig. 5.5). The increase in theoretical yield after swapping GAPD was somewhat consistent across parameters: 1.42 ± 0.20 mol NADPH / mol glucose under anaerobic conditions and 0.65 ± 0.33 mol NADPH / mol glucose under the

examined aerobic conditions (Fig. 5.5C,F). Thus, the simulated increase in theoretical yield is a plausible explanation for the 1.25 ± 0.19 mol NADPH / mol glucose increase in yield observed by Martínez et al. 2008.

5.4.3 Optimal cofactor swaps increase ATP availability

It has been reported that transhydrogenase enzymes in *E. coli* are responsible for producing 35–45% of the NADPH necessary for biosynthesis (Sauer et al. 2004). However, the transfer of reducing equivalents from NAD(H) to NADP(H) requires energy from proton translocation to proceed. Thus, *in vivo* and *in silico*, the amount of NADPH that can be produced by the transhydrogenase enzyme is limited by this energy requirement. Some yield increases for products of biosynthetic reactions can be achieved by overexpressing the membrane-bound transhydrogenase encoded by *sthA* or down-regulating the soluble transhydrogenase encoded by *pntAB* (Sanchez et al. 2006; Jan et al. 2013). However, the metabolic model predicts that modifying transhydrogenases will not lead to the maximum theoretical yield. Producing NADPH directly during a metabolic transformation is inherently more efficient than utilizing the transhydrogenase enzyme and the electrochemical gradient to generate NADPH from NADH. At the theoretical limit of NADPH production under anaerobic conditions (Supplementary Fig. 5.4), the GAPD cofactor swap allows for a decrease in NADPH production by the energy-coupled transhydrogenase enzyme. Thus, demand for transmembrane electrochemical potential (i.e. ATP equivalents) from the transhydrogenase enzyme decreases, and more flux can be directed away from glycolysis and ATP production and towards NADPH producing enzymes (e.g. in the pentose phosphate pathway). These trends were also investigated for native products of *E. coli*

*i*JO1366 for which the GAPD swap increased theoretical yield. For these products, it was observed that, at the theoretical maximum production state, the flux through glycolysis, the pentose phosphate pathway, and TCA cycle were not generally shifted after swapping GAPD (Supplementary Fig. 5.5). Thus, glycolytic flux remains high in the GAPD-swap simulation, and any ATP that was is no longer necessary for transhydrogenase activity can be directed to product and biomass generation.

In contrast to *E. coli*, no transhydrogenase enzymes have been identified in yeast (Nissen et al. 2001). Therefore, balancing cofactor production and consumption in *S. cerevisiae* is even more critical. In yeast, under anaerobic conditions, the theoretical yield of NADPH increases after the ALCD2x swap because NADH production decreases, so less flux is directed to ethanol fermentation (which would act to oxidize NADH), and more flux can be directed to acetate fermentation (which has a higher ATP yield than ethanol fermentation). Thus, the decrease in NADH production can theoretically increase the availability of ATP at steady state—even in the absence of transhydrogenase enzymes.

5.4.4 Cofactor swapping in yeast has a greater effect with D-xylose as a substrate

When D-xylose is used as a substrate, swapping the specificity of a single oxidoreductase can have an even greater effect on the production of native products in *S. cerevisiae*. There has been much interest in the use of 5-carbon sugars as a feedstock for *S. cerevisiae* production strains, which can be accomplished with heterologous expression of the *XYL1* and *XYL2* genes, encoding XR and XDH, respectively, from *P. stipitis* (Bengtsson, Hahn-Hägerdal, and Gorwa-Grauslund 2009). This XR is

NADPH-dependent, and modifications to the cofactor specificity of XR have been shown to increase xylose fermentation, experimentally (Bengtsson, Hahn-Hägerdal, and Gorwa-Grauslund 2009) and computationally (Ghosh, Zhao, and Price 2011). The simulations presented here show that swapping the cofactor specificity of high-flux, central metabolic enzymes (i.e., GAPD, ALCD2x) can increase theoretical yields by 2–3 fold across native metabolism in *S. cerevisiae* by generating more NADPH to drive D-xylose uptake by XR (Fig. 5.3, see Methods for how the native XR and XDH in the model are accounted for). Our results also show that swapping the cofactor specificity of these central metabolic enzymes has a greater impact on theoretical yield in yeast than cofactor swaps for XR or XDH. Utilizing the cofactor-balanced XI enzyme (Bettiga, Hahn-Hägerdal, and Gorwa-Grauslund 2008) is another approach to cofactor balancing D-xylose uptake, and the XI uptake pathway also maximizes the theoretical yield *in silico*.

While theoretical yield improves for many products with D-xylose as a substrate, the theoretical yield of ethanol fermentation was not much improved by cofactor swapping. Under aerobic conditions, simulations showed an increase in theoretical yield from 0.606 to 0.619 Cmol / Cmol (a 2% increase) after swapping the cofactor specificity of ALCD2x. Under anaerobic conditions, no increase was observed. In a previous study, Verho et al. 2003 compared ethanol fermentation from D-xylose with the native GAPD enzyme to ethanol fermentation with an NADPH-GAPD, and the authors reported a significant increase in ethanol yield (from 0.24 to 0.41 Cmol / Cmol). The results of this work suggest that the theoretical yield was not significantly increased in the GAPD-swap strain, and so other effects (e.g. regulation, kinetics) could be considered to explain the improved yield.

5.4.5 Theoretical yield of non-native products increases with swaps

Optimizations for the non-native pathways reconstructed in *E. coli* demonstrate that cofactor swapping can also improve maximum theoretical yields for non-native industrial products. In particular, the theoretical yields of 1,3-propanediol, 3-hydroxybutyrate, 3-hydroxypropanoate, 3-hydroxyvalerate, styrene, and lycopene were increased after one cofactor swap. For biomonomer production strains, yield is an extremely important consideration. Final titer, productivity, and yield are the three most important design parameters in bioprocessing, and yield plays a central role in determining the economic viability of a production process for high volume/low cost products like biomonomers (Villadsen, Nielsen, and Lidén 2011). The same type of modifications of cofactor specificity that have been used to increase yield of native products in *E. coli* (Martínez et al. 2008) and *S. cerevisiae* (Bengtsson, Hahn-Hägerdal, and Gorwa-Grauslund 2009; Verho et al. 2003) are predicted to increase the maximum yield of these non-native products.

In the global analysis, theoretical yields of 1,4-butanediol and 2,3-butanediol were not effected by cofactor swapping. For these products, a detailed analysis of swapping across the parameter space of SUR, OUR, and μ_{min} (Supplementary Fig. 2) showed that some improvement in theoretical yield *is* possible for 1,4-butanediol at certain parameter values. Specifically, improvements in theoretical yield are prevalent when μ_{min} is near its maximum value for a given SUR. Only at higher growth rates does cofactor balance become a limitation for production of 1,4-butanediol. However, across the parameter range, cofactor swaps did not increase the thoeretical yield for either of the 2,3-butanediol isomers. The case of 1,4-butanediol demonstrates that one

should pay close attention to SUR, OUR, and μ_{min} and their effects when simulating theoretical yield and optimizing cofactor balance.

5.4.6 Theoretical yields are sensitive to knowledge of cofactor preference and enzyme promiscuity

A limitation of this analysis is the possibility of alternative cofactor usage by promiscuous oxidoreductase enzymes (Olavarría, Valdés, and Cabrera 2012). This limitation can be addressed by improving the model to include additional reactions for those oxidoreductases that are known to catalyze flux with multiple cofactors, and, in fact, some reactions do have this kind of annotation (e.g. *frdABCD* in *i*JO1366 (Orth et al. 2011)). To engineer a strain that generates a product near the theoretical yield, it is still necessary to optimize the native regulation and the kinetics and thermodynamics of the optimal pathways. Also, the theoretical yields presented here are purely stoichiometric, so any theoretical kinetic or thermodynamic limitations beyond reaction reversibilities (Feist et al. 2007) are not included in the analysis. Furthermore, the kinetic and thermodynamic impacts of cofactor swapping a particular enzyme could lead to unintended consequences in the network that are not reflected in the thoeretical yield. Published cases of successfully swapping the cofactor specificity of oxidoreductase enzymes have not identify this as a major limitation (Verho et al. 2003; Martínez et al. 2008), but it is an important consideration, especially as one approaches the theoretical limits of bioproduction.

5.5 Conclusion

Constraint-based modeling is uniquely suited for modeling optimal metabolic states, because optimizations like cofactor swapping can be performed for large sets of products and environmental conditions. The optimizations reported here demonstrate the importance of cofactor swapping in *E. coli* and *S. cerevisiae* for native and non-native products, and show that microbial byproducts can be organized according to the need for a synthetic increase in NADPH production. These results also highlight the centrality of certain enzymes and that swapping the cofactor specificity of these enzymes (GAPD, ALCD2x) has a global effect on cofactor balance in the metabolic network. These methods are especially applicable for highly engineered strains can generate metabolic products with yields approaching the maximum theoretical yield (Murarka et al. 2008; Trinh 2012; Lin, Bennett, and San 2005). Cofactor swapping could be used to tune high-yield industrial bioprocessing strains, and experimental validation of the optimal swaps predicted for *E. coli* and *S. cerevisiae* could have an immediate application in industry.

Chapter 5 is a reprint of a published manuscript: King, Z. A. and Feist, A. M. (2014). "Optimal cofactor swapping can increase the theoretical yield for chemical production in *Escherichia coli* and Saccharomyces cerevisiae". In: *Metab. Eng.* 24, pp. 117–128. DOI: 10.1016/j.ymben.2014.05.009. The dissertation author was the primary author of the paper and was responsible for the research.

Chapter 6

Literature mining supports a next-generation modeling approach to predict cellular byproduct secretion

6.1 Introduction

All cells secrete metabolic byproducts in the course of growing and producing energy, and these byproducts play important roles in the study of biological systems. Byproducts are a readout of the cellular state; lactate excretion, for instance, is characteristic of tumor cell growth (Hanahan and Weinberg 2011; Basan et al. 2015). Byproducts can be engineered for bioproduction of commodity chemicals and biofuels (Lee and Kim 2015; Zhang, Rodriguez, and Keasling 2011; Chubukov et al. 2016). And byproducts of yeast fermentation – including ethanol – are responsible for the

most popular beverages in human history (Piškur et al. 2006). With the critical roles played by metabolic byproducts in disease and biotechnology, it is of great interest to be able to predict the byproducts that a cell will secrete under a specific condition. However, no published study has assessed whether existing computational methods are able to predict metabolic byproducts for a range of strains and conditions.

Computational models have been shown to correctly predict byproduct secretion under common laboratory conditions. During aerobic growth, the model bacterium *Escherichia coli* oxidizes substrate molecules to secrete CO_2 and water; during anaerobic fermentation, *E. coli* secretes mixed-acid fermentation products (ethanol, acetate, formate, D-lactate, and succinate) (Clark, David P 1989). Genome-scale models (GEMs) and constraint-based reconstruction and analysis (COBRA) methods rely on knowledge of the metabolic network and mass-balance during steady state growth to predict the optimal distribution of metabolic flux for growth (Bordbar et al. 2014a). GEMs have been shown to be able to predict *E. coli* byproduct secretions in certain cases (Varma, Boesch, and Palsson 1993; Fong et al. 2005). In the context of GEMs, the byproducts that must be secreted for optimal growth are called *growth-coupled*, and computational methods have been developed to predict and engineer growth-coupled chemical production (Burgard, Pharkya, and Maranas 2003; Feist et al. 2010; Lewis, Nagarajan, and Palsson 2012). However, few experimental studies have followed from the computational method development, among them (Fong et al. 2005; Yim et al. 2011), so it is unclear how these methods would scale up to a wide variety of strains and conditions.

Next-generation GEMs of metabolism and gene expression (called ME-models) are now available; ME-models predict the composition of the entire proteome of a

cell (O'Brien et al. 2013; O'Brien and Palsson 2015; Lerman et al. 2012). In contrast, GEMs of metabolism (M-models) predict only the reaction fluxes in a metabolic network (O'Brien and Palsson 2015). One new capability of ME-models is the ability to predict the bacterial Warburg effect, the tendency of bacteria to secrete acetate during aerobic growth in the presence of excess substrate (Basan et al. 2015; Molenaar et al. 2009). In ME-models, the limitations of ribosome efficiency lead to low-yield metabolic approaches like acetate secretion (O'Brien et al. 2013). The same effect can be seen in smaller-scale growth models and is supported by phenotypic data (Basan et al. 2015; Molenaar et al. 2009). Whether ME-models can correctly predict byproduct secretion for other conditions is not currently known.

High-quality genotypic and phenotypic data are required to test any model predictions, and such data have not been available for the study of byproduct secretion. The present study takes a novel approach by mining the research literature for examples of engineered strains of *E. coli* with diverse byproduct secretion mixtures. We collected 73 papers reporting a total of 89 strains of *E. coli* that have a wide range of gene knockouts, heterologous pathways, and growth conditions, and we simulated these paired genotype-phenotype data in 6 historical GEMs of *E. coli*, including the next-generation ME-model. We find that GEMs have been improving in their ability to recapitulate measured byproducts from experimental studies as the models have increased in size and scope. We explore the possible reasons for incorrect predictions and provide insights into the challenges of simulating byproduct secretion for any growing cell.

6.2 Results

6.2.1 Literature mining provides a diverse set of strains and phenotypes.

An impressive body of data on *E. coli* byproduct secretion can be found in the peer reviewed literature (Fig. 6.1). We generated a bibliomic database using a workflow for identifying relevant papers, extracting data, and performing quality assessment (Fig. S1; Supplementary Figures and Data available at http://dx.doi.org/10.1101/066944). Each paper in the database reported a strain design of *E. coli* in which the fermentation pathways were engineered to force the cell to secrete a target molecule (Fig. 6.2). The bibliomic database includes the gene knockouts, heterologous pathway descriptions, substrate conditions, oxygen availability, and the parent cell line for each strain (Supplementary Data 1). It is difficult to extract and normalize quantitative measures of byproduct secretion from the literature. Instead, we recorded the molecule that was targeted for overproduction in the study, and we confirmed that this byproduct was the major secretion product in each case (see Methods). The bibliomic database contains 73 papers and 89 strains of *E. coli*; this is approximately 20% of all papers on metabolic engineering of *E. coli* collected in the LASER database (Winkler, Halweg-Edwards, and Gill 2015).

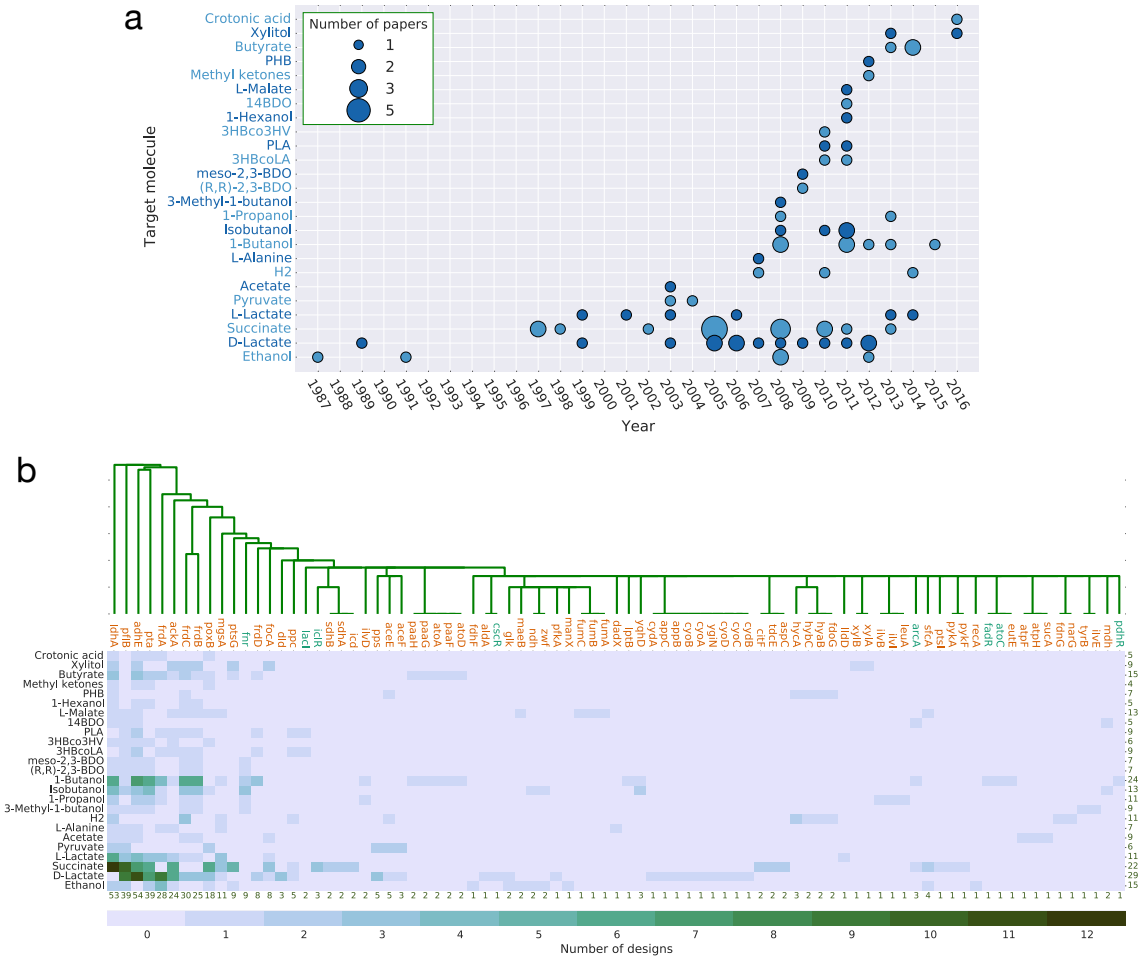

Figure 6.1: **The bibliomic database.** (a) The number of papers published for each target molecule over time. (b) The individual genes that have been knocked out for each target molecule. The sums across the bottom indicate the total number of designs that include a given gene, and the sums across the right indicate the total number of unique genes knocked out for a given target molecule. These common knockouts remove the routes to the native fermentation products acetate (*pta*, *ackA*, *pflB*), ethanol (*adhE*, *pflB*), formate (*pflB*), D-lactate (*ldhA*), and succinate (*frdABCD*). These knockouts represent a common strategy where the highest-yield fermentation pathways are knocked out, one by one, until the target pathway becomes the optimal route for balancing the redox state of the cell.

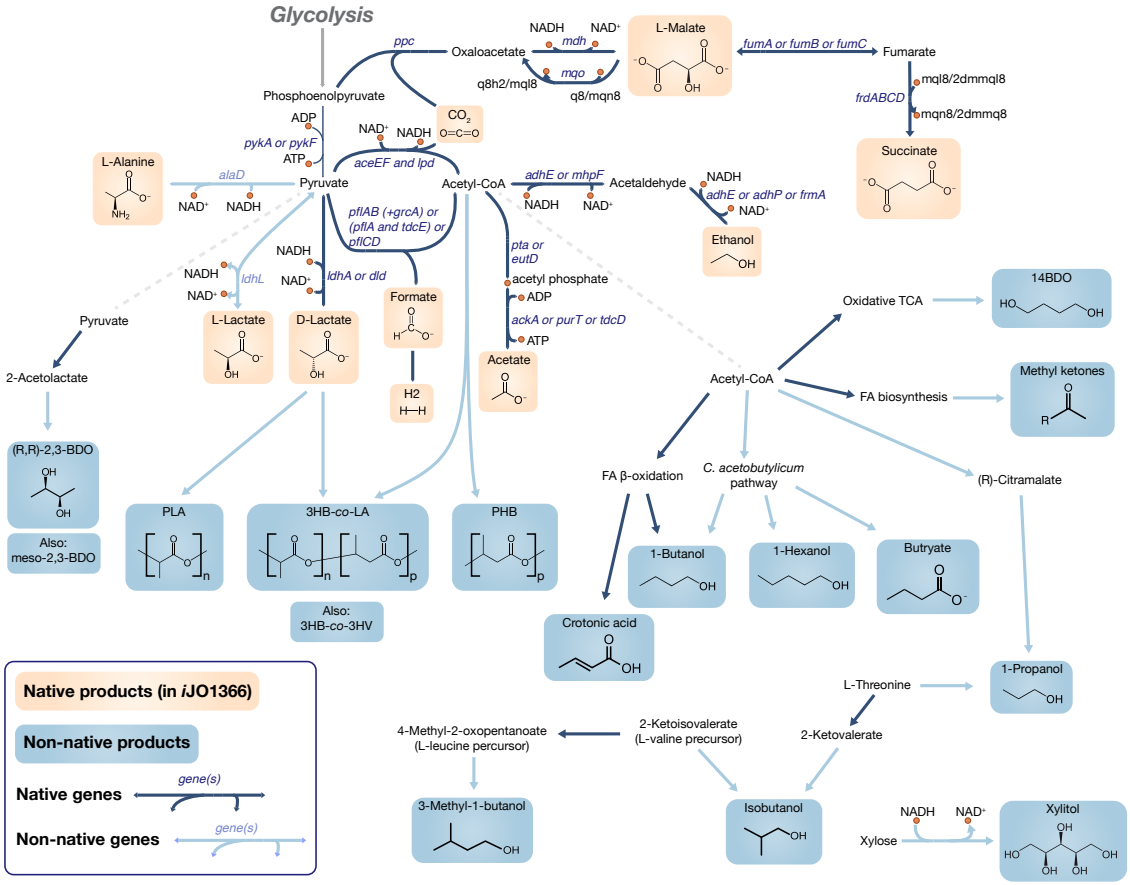

Figure 6.2: **The engineered fermentation pathways in *E. coli*.** All the engineering pathways in the bibliomic database are shown, along with their metabolic precursors. Native products (yellow) are those that appear in the genome-scale model *i*JO1366. Native pathways in *i*JO1366 (dark blue arrows) and non-native pathways (light blue arrows) are also differentiated.

The strains in the bibliomic database were simulated in six GEMs of *E. coli* (Table 6.1). The models have increased in size and complexity over the past decade; they include five M-models and one ME-model that includes 1,683 genes and accounts for 80% of the proteome by mass (O'Brien and Palsson 2015; O'Brien et al. 2013). Gene knockouts, heterologous pathways, and environmental conditions from the bibliomic database were recreated in each of the GEMs. For each strain, flux balance analysis (FBA) (Orth, Thiele, and Palsson 2010) was used to find the predicted growth rate and the growth-coupled yield, the carbon yield of a compound at the maximum growth

rate. The analysis began with two comparisons between the bibliomic database and the simulations: (1) whether the strain grew in a given environment and (2) whether the simulation predicted growth-coupled secretion of the target byproduct from the study.

Table 6.1: **The increasing size and scope of genome-scale models of *E. coli*.**

Model	Genes	Reactions	Metabolites / Components	Year (Reference)
Core model	137	95	72	2006 (Palsson 2006)
*i*JR904	904	1075	761	2003 (Reed et al. 2003)
*i*AF1260	1,260	2,382	1,668	2007 (Feist et al. 2007)
*i*AF1260b	1,260	2,388	1,668	2010 (Feist et al. 2010)
*i*JO1366	1,366	2,583	1,805	2011 (Orth et al. 2011)
*i*OL1650-ME	1,683	12,009	6,563	2013 (O'Brien et al. 2013)

The predictive power of GEMs has generally increased over time, with the increasing size and scope of the models. New GEMs provide better predictions of growth-coupled secretion compared to their predecessors ("Model accuracy" in Fig. 6.3). In order to understand the reasons for this trend, we designed a computational approach to categorize cases of incorrect prediction. Exhaustive search and parameter sampling were employed in the M- and ME-models, respectively, to determine what changes to the modeling approach might lead to *in silico* secretion of the target byproduct (see Methods). These categories provide insights into the general challenges of modeling byproduct secretion.

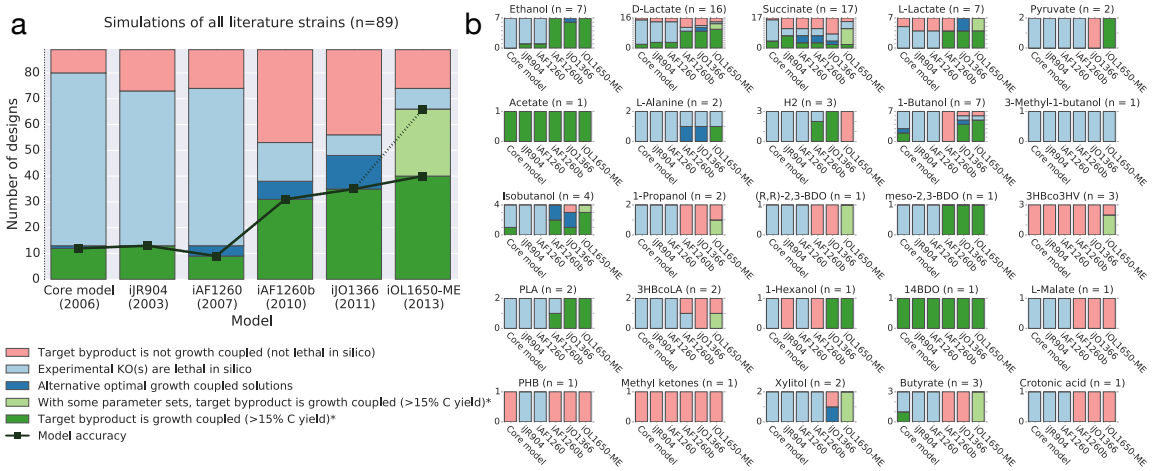

Figure 6.3: **Simulations of the bibliomic dataset in E. coli GEMs.** (a) The 89 strains in the bibliomic database were simulated in six GEMs of *E. coli*, and the incorrect predictions were categorized to suggest a reasons for the errors. The solid line signifies that the experimentally observed target byproduct is growth-coupled in the model. The dashed line represents the possibility of improving predictions in the ME-model by correctly determining the kinetic parameters (k_{eff}s). (b) The categories separated according to the target molecule.

6.2.2 Genome-scale models do not differentiate between isozymes.

Isozymes are common in metabolic networks, and they are represented in M-models, but their diverse regulatory and catalytic properties lead to a broad and complex set of challenges for metabolic modeling. Reactions are often catalyzed by a major isozyme that is responsible for most catalysis, while minor isozymes are also present in the cell but have a smaller role (they may not be expressed or have less-favorable kinetics) (Nakahigashi et al. 2009); recent progress in studying enzyme promiscuity and underground metabolism suggests that isozymes are even more widespread than previously thought (Guzmán et al. 2015). Many experimental studies report gene knockouts of major isozymes that decrease the activity of the associated reaction significantly, enough so that the minor isozymes can be ignored

(e.g. removing *ldhA* and ignoring *dld* (Trinh et al. 2011; Stols and Donnelly 1997; Zhou, Shanmugam, and Ingram 2003)). However, M-models do not distinguish between major and minor isozymes, so these cases are incorrectly predicted in the model; the minor isozyme catalyzes the reaction *in silico*, and the *in silico* gene knockout of the major isozyme has no effect. Therefore, to simulate byproduct secretion for real-world experiments, it was necessary to employ a "greedy knockout" strategy in which all reactions associated with a gene knockout are disabled, even if minor isozymes might be present (Fig. 6.4a).

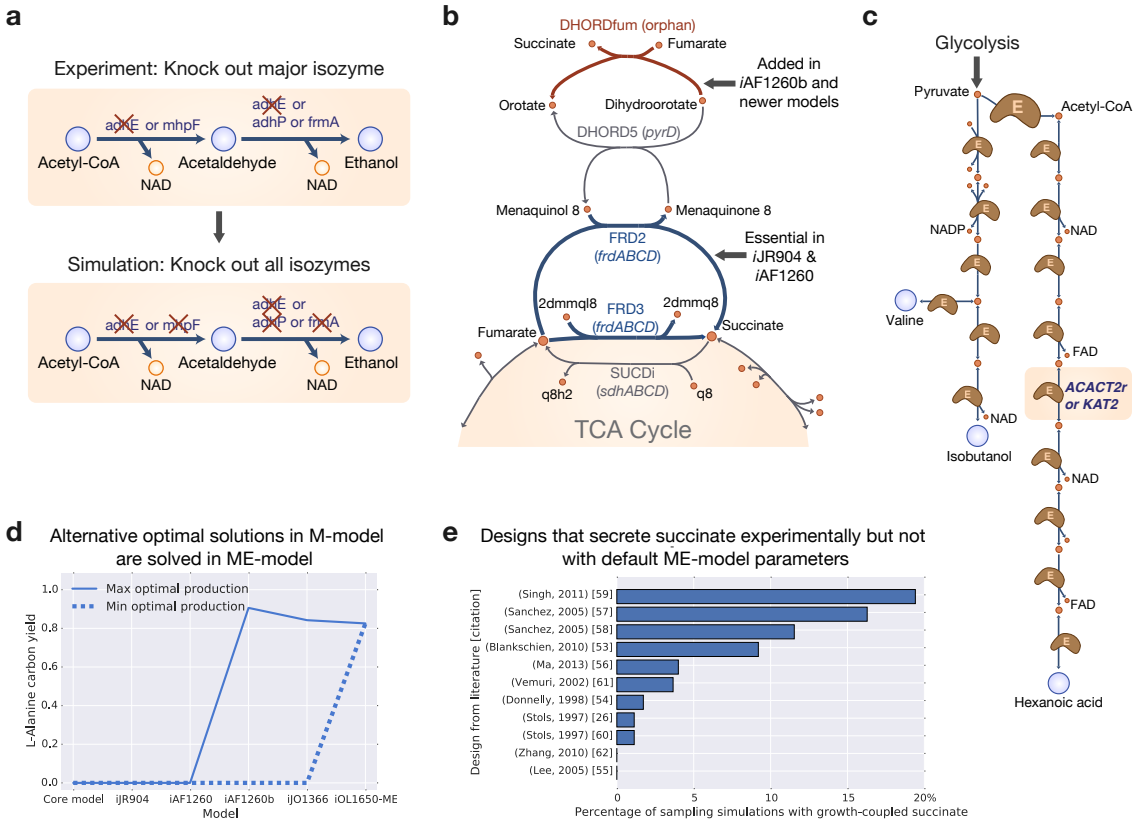

Figure 6.4: **Comparing simulations with experiments.** All modeling approaches have failure modes, and comparing model predictions to experimental results allows these failure modes to be analyzed. (a) A "greedy knockout" strategy is necessary to contend with major and minor isozymes that are difficult to simulate in GEMs. (b) The genes in the *frd* operon are responsible for most of the incorrect predictions of cell death in *i*JR904 and *i*AF1260. This error was fixed in *i*AF1260b and later models with the addition of the reaction DHORDfum. (c) For an isobutanol design, the ME model correctly predicts isobutanol secretion in preference to hexanoic acid secretion because the hexanoic acid pathway has greater protein cost (Atsumi, Hanai, and Liao 2008; Atsumi et al. 2010). (d) Alternative optimal phenotypes appear in M-models when two pathways have equivalent stoichiometries, as in this example for L-alanine secretion. ME-models explicitly account for the cost of producing pathway enzymes, so the shorter L-alanine production pathway is optimal in ME-models. (e) Succinate secretion is difficult to predict using existing GEMs, but an ensemble of ME-models with sampled kinetic parameters demonstrates that for certain parameter sets succinate secretion is correctly predicted.

There are exceptions where greedy knockouts are not appropriate. For example, the alanine racemase activity of isozymes *alrR* and *dadX* is necessary for *in silico* growth, so applying the greedy knockout strategy to the reported strain that has a knockout of *alrR* leads to a prediction of cell death (Zhang et al. 2007). In other words,

this strain can not be correctly simulated by M-models with or without the greedy knockout strategy. This issue can only be addressed through continued development of genome-scale modeling methods to address regulation, kinetics, allosteric inhibition, and the many biophysical properties that differentiate isozymes. Furthermore, ME-models can potentially select the appropriate enzyme based on protein cost, but ME-models do not include regulatory effects that often are responsible for the distinction between major and minor isozymes, so greedy knockouts are still generally required. In this study, the greedy knockout approach was sufficient to correctly simulate most of the gene knockouts in the bibliomic database.

6.2.3 Larger models solve false predictions of cell death.

Every strain in the bibliomic database was able to grow in the published experimental studies, but many simulations of these strains in early GEMs resulted in predictions of no growth (defined as *in silico* specific growth rate less than 0.005 hr^{-1}). These incorrect predictions have decreased as the GEMs have increased in size and scope ("Experimental KO(s) are lethal *in silico*" in Fig. 6.3). In most cases, the reason for the improved prediction is that the more comprehensive GEMs include a pathway that can rescue an essential cellular function when another important pathway is disabled by gene knockouts. In the five *E. coli* M-models, the lethal genotypes were analyzed by exhaustively searching for the minimal combinations of reactions that lead to *in silico* cell death (Fig. S2).

The biggest improvement in modeling the strains in the bibliomic database can be attributed to a single reaction. The models *i*JR904 and *i*AF1260 incorrectly predict that fumarate reductase (FRD, *frd*) is essential under anaerobic conditions, and 63%

of the designs in the bibliomic database include a knockout in the *frd* operon (see the large jump from *i*AF1260 to *i*AF1260b in Fig. 6.3a). These incorrect predictions were corrected in *i*AF1260b and later GEMs with the inclusion of a new reaction (DHORDfum) that rescues growth when FRD is removed (Fig. 6.4b). However, there is no experimental evidence to support the presence of the DHORDfum reaction. So why does this reaction exist in the models, and why does it improve predictions?

One explanation is that the DHORDfum reaction does not take place in the cell, and, instead, succinate dehydrogenase (SUCDi, *sdh*) acts in the reverse direction to rescue conversion of fumarate to succinate; this has actually been shown experimentally (Maklashina, Berthold, and Cecchini 1998). Thus, the evidence supports removing DHORDfum from the models and making SUCDi reversible. However, this change introduces the challenges associated with modeling isozymes for the activity catalyzed by *frd* and *sdh*, so the presence of DHORDfum has served as a convenient hack for modeling *E. coli*.

6.2.4 Simulations suggest that some strains have room to evolve.

When the experimental observations of byproduct secretion disagree with predictions, another possible explanation is that the experimental strain could evolve to grow faster by adopting the byproduct secretion strategy predicted by the model ("Target byproduct is not growth-coupled" in Fig. 6.33). FBA simulations predict the metabolic state of a cell that is operating close to optimal growth; GEMs are powerful for predicting cellular behavior precisely because fast growing cells often adopt a near-optimal strategy for growth (Ibarra, Edwards, and Palsson 2002; Edwards,

Ibarra, and Palsson 2001). Thus, some of the disagreement between observation and prediction might be caused, not by model errors, but rather by an assumption of the modeling approach (the optimality assumption). This hypothesis can be tested through laboratory evolution by passing the strain repeatedly (Fong et al. 2005). (The process is also called serial passage, metabolic evolution, growth rescue, or adaptive laboratory evolution (ALE).) Laboratory evolution was used in 14 studies (19 strains) in the bibliomic database to improve byproduct secretion, and the predictive power of the model is greater for these cases than for the bibliomic database in general (Fig. S3). This supports the hypothesis that FBA predicts byproduct secretions that are not correct for the reported strains but would be correct if the strains were evolved through growth selection.

6.2.5 Next-generation ME-models improve predictions but require parameterization.

ME-models expand upon M-models by explicitly accounting for all of the biochemical reactions in the gene expression machinery of the cell (including transcription and translation) (O'Brien et al. 2013; O'Brien and Palsson 2015). To include protein production in the ME-model, one must estimate the turnover rate of each enzyme (k_{eff}) that determines how many active proteins must be present to convert one set of reactants to products in a given time. ME-model simulations used a set of experimentally validated kinetic parameters from a recent study (Ebrahim et al. n.d.). For high-flux reactions, the k_{eff}s were shown to be consistent across four growth conditions. However, it is still possible for k_{eff}s to change between conditions, depending on metabolite concentrations and other variables (they range between 0 and k_{cat}). Therefore, we

sampled k_{eff}s in the ME-model to generate an ensemble of models for each strain that was not growth-coupled with default parameters (see Methods). We found that 26 / 41 strains in this set could be growth-coupled in the ME-model with at least one model in the ensemble, including 9 / 11 designs for succinate production (Fig. 6.4e). Addressing kinetic parameters will have to be a part of ME-model development going forward, and this should lead to better predictions of byproduct secretion.

The protein costs associated with metabolic pathways in the ME-model also solve another failure mode in M-models: alternative optimal solutions. Alternative optimal solutions occur in M-models when two metabolic states lead to the same growth rate, and this common failure mode has been solved with next-generation ME-models ("Alternative optimal growth-coupled solutions" in Fig. 6.3) (Lewis et al. 2010a). In ME-model simulations, each pathway has specific enzyme costs that must be precisely allocated using cellular resources. Therefore, pathways with the same metabolic contribution to cellular growth (e.g. same ATP production and redox balance) that are equivalent in the M-model have different proteomic costs in the ME-model. In all cases, this failure mode of M-models disappear in ME-model predictions (with one example provided in Fig. 6.4d).

In addition to removing alternative optimal solutions, the proteomic pathway costs in the ME-model can address challenges of encoding reversibility in the M-model. As an example, the production of isobutanol using a 2-keto acid based pathway was recently demonstrated (Atsumi et al. 2010; Atsumi, Hanai, and Liao 2008), and the optimal *in silico* phenotype of this production strain varies between models of *E. coli* (Fig. 6.3b). *i*AF1260b correctly predicts the production of isobutanol as the optimal fermentation product; in contrast, *i*JO1366 predicts that hexanoic acid, a 6-carbon

intermediate in the β-oxidation cycle, is the preferred product. This difference can be traced to the thermodynamic reversibility of the thiolase reaction in the second round of the reversed β-oxidation cycle – it is irreversible in iAF1260 (KAT2) and reversible in iJO1366 (ACACT2r) (Fig. 6.4c). The reversibility in iJO1366 is in line with experimental evidence (Dellomonaco et al. 2011), but it also leads to the seemingly incorrect prediction of hexanoic acid secretion. The ME-model suggests that the incorrect prediction of hexanoic acid secretion by iJO1366 is not so much a matter of thermodynamics as a matter of pathway length and thus proteomic cost. When the cost of producing enzymes for metabolic pathways is incorporated into genome-scale models, long pathways like the hexanoic acid production route through β-oxidation carry a greater cost than the shorter 2-keto acid route to isobutanol. This case shows the power of a constraint-based modeling approach: Properly encoding reversibility in M-models has been a long-standing challenge, so the ME-model applies a completely different constraint (pathway cost) that makes the reversibility of β-oxidation unimportant for correct predictions.

6.3 Discussion

As cellular models become larger and more complicated, the datasets used to validate them must also grow. This study presents a novel approach to model validation based on literature mining. In spite of the uneven quality of literature data, this approach was capable of generating important insights into the abilities of GEMs to predict byproduct secretion. Higher-quality data would enable an even more thorough model validation, and there is a great need in systems biology for standardizing genotype-phenotype datasets. Standards for storing phenotypic data

have been discussed (McMurry et al. 2016; Check Hayden 2015), and it is essential that progress be made.

There are a few challenges that will have to be addressed to scale these methods to larger and more complicated systems. First, many data points in the bibliomic database cannot be modeled in existing GEMs. For instance, regulatory knockouts are not in the scope of M- and ME-models, so they were ignored in this study. The correct predictions of strains in the bibliomic database draw largely from the concept of redox balance in the cell (NAD(P)H produced during glycolysis must be consumed by fermentation pathways), and extending prediction of byproduct secretion to other applications where redox balance is not the driving phenomenon may require further development of the modeling methods. However, constraint-based modeling methods are generally extensible, as we have seen with the development and implementation of ME-models. Exploration of constraint based approaches to other subsystems – including protein structures, membrane translocation, and regulation – are under way (King et al. 2015b).

Second, strains modeled using GEMs and FBA must be operating close to an optimal growth state. Understanding the byproduct secretion of strains that are not growing rapidly will require research into other objective functions that could make the models predictive for strains that are not optimizing for growth (Zhao et al. 2016; Schuetz, Kuepfer, and Sauer 2007). On the other hand, the optimality assumption of FBA offers an advantage: GEMs and laboratory evolution can be used together for systematic optimization of microorganisms (Fong et al. 2005; Yim et al. 2011).

Finally, the extension of these methods to larger and more complex organisms, such as tumor cells, will require rigorous development and assessment of GEMs. This

study provides an example of validating model predictions using genotype-phenotype data mined from the literature. The collection of these data will need to be scaled up to validate larger and more complex models. All cells have the same basic features that include gene expression, metabolism, and, by necessity, byproduct secretion; with targeted validation studies, we can feel increasingly confident in our ability to model and understand them.

6.4 Methods

6.4.1 Literature mining.

A literature mining search was performed to identify all papers reporting the construction of a cell factory strain of *E. coli* for the production of a fermentation product. A workflow was developed (Fig. S1), hundreds of papers were collected, and 73 were included in the bibliomic database based on their matching the following criteria:

- Utilized a strain of *E. coli*.
- Modified the strain for production of a native or heterologous metabolite.
- Removed alternative fermentation pathways using gene knockouts.

Metadata were collected from each paper, including the target production molecule, whether simulations were performed to identify knockouts, the parent *E. coli* strain, the genetic additions and deletions, the aerobicity and carbon sources during fermentation experiments, whether laboratory evolution was performed, and (when possible) the measured fermentation profile of the engineered strain.

A single target molecule was selected for each experiment, even though in some cases a mixture of products was reported. When papers reported mixtures of hydrogen or formate with a coproduct, the coproduct was considered the target molecule.

6.4.2 Simulations.

To simulate reported designs, the gene knockouts were implemented *in silico* using a "greedy knockout" strategy. For each gene that was knocked out experimentally, all reactions associated with that gene in the metabolic model are turned off. The alternative strategy is to evaluate the gene-protein-reaction (GPR) rules for each reaction in turn, to determine whether the reaction is turned off or remains unchanged; however, as discussed in the text, only the "greedy knockout" approach was able to correctly simulate strains in the bibliomic database.

For all non-native genes reported in the papers, pathways were reconstructed by creating *in silico* reactions corresponding to the genes used in these experiments. For transport reactions, transport was assumed to be non-energy-coupled unless otherwise specified in the *i*JO1366 reconstruction or in the literature.

Polymer production must be considered separately from ordinary metabolite secretions. To simulate these strains, the production of the monomer was optimized. It is unclear whether polymers such as polylactic acid (PLA) would be growth coupled. The PHA synthase is not energy coupled (Lee 1996), so an equilibrium between monomer and polymer would probably be achieved in the optimal state (this has been shown for soluble heteroglycans (Kartal et al. 2011)). However, by upregulating the PHA synthase in an strain optimized for monomer production, one can use the growth-coupling effect to perform much of the strain optimization. Thus, growth-coupling of

the monomer is of interest.

Five M-models and one ME-model of *E. coli* K-12 MG1655 were used for the simulations in this work. The M-models were collected from the BiGG Models database (King et al. 2016), and they were used as reported in their respective publications (Table 6.1). As described previously, the *i*JO1366 oxidative stress reactions CAT, SPODM, and SPODMpp and the FHL reaction were constrained to zero (Orth et al. 2011). A new software implementation of the ME model *i*OL1650-ME was used. Pathway diagrams were generated using Escher (King et al. 2015a), and COBRA simulations were performed with COBRApy (Ebrahim et al. 2013).

For M-model simulations, the substrate uptake rates (SURs) for the solitary carbon substrates in each simulation were constrained to a maximum uptake rate of 10 mmol gDW^{-1} hr^{-1}. The oxygen uptake rates were constrained to 0 for anaerobic conditions and 20 mmol gDW^{-1} hr^{-1} for aerobic conditions. For ME-model simulations, SURs were left unbounded and the ME-model optimization procedure chose optimal SURs. If LB or yeast extract was present in the medium, the simulations were still performed with an *in silico* minimal media based on the assumption that cells will preferentially consume glucose before more-complex carbon sources; however, if this approximation led to a lethal phenotype in *i*JO1366, then supplementations known to exist in rich media were added to alleviate the lethal phenotype. Microaerobic designs were assumed to be anaerobic because it has been observed that even under aerobic conditions the anaerobic physiology contributes to fermentation (Ingram et al. 1987).

FBA was used to find the maximum and minimum secretion of each metabolite in the network when the growth rate is near its maximum (within 0.01%) (Orth, Thiele, and Palsson 2010). The key outputs of these simulations are *predicted growth*

rate – the flux through the biomass objective function – and the *growth-coupled yield* – the minimum carbon flux through the target molecule exchange reaction at the maximum growth rate

6.4.3 Parameter sampling.

Parameter sampling in the ME-model was employed to determine the sensitivity of ME-model simulations to k_{eff} values. For each sampling simulation, an ensemble of 200 models was generated with k_{eff} values selected randomly from a lognormal distribution of possible k_{cat}s. The distribution was determined from a collection of all k_{cat}s in the BRENDA enzyme database ($\mu = 2.48$ and $\sigma = 3.29$) (Bar-Even, Noor, and Savir 2011).

6.4.4 Failure model categorization.

Growth-coupling was defined as secretion of the target molecule with greater than 15% carbon yield or, for hydrogen production, greater than 2 mmol gDW^{-1} hr^{-1}. Lethal phenotypes were defined as having an *in silico* growth rate below 0.005 hr^{-1}. Alternative optima were identified by finding designs whose maximum secretions were above the threshold for growth coupling but whose minimum secretions were below this threshold.

Chapter 6 is a reprint of a published manuscript: King, Z. A., O'Brien, E. J., Feist, A. M., and Palsson, B. O. "Literature mining supports a next-generation modeling approach to predict cellular byproduct secretion". In: *Metabolic Engineering*. under review. The dissertation author was the primary author of the paper and was responsible for the research.

Chapter 7

Conclusions and Outlook

7.1 Next-generation models and predictions

As the applications of COBRA methods have multiplied, there has also been a continuous expansion in the scope and complexity of new models (Feist et al. 2009) and methods (Lewis, Nagarajan, and Palsson 2012). Together, these new methods deliver a vision of a COBRA modeling framework encompassing many biological networks and simulation strategies (Fig. 7.1), so that many new types of predictions can be made (Table 7.1).

Table 7.1: **Current and next-generation COBRA models – types of predictions that are possible.**

Model Scope	Enabled Predictions	References
Metabolism (M-Model)	- Genetic manipulations for metabolite overproduction/consumption - Gene essentiality - Growth rate given constraining uptake rates - Metabolite secretion rates - Metabolic fluxes, considering stoichiometry of network, at given growth rate	(Orth et al. 2011)
Metabolism and gene expression (ME-Model)	- Maximum feasible growth rate without explicit constraining substrate uptake rates - Cellular macromolecule composition - Gene expression levels - Metabolic fluxes, considering enzyme cost, at maximum feasible growth rate - Nutrient limited phenotypes	(Lerman et al. 2012; O'Brien et al. 2013)
ME-Model with cell membrane protein translocation	- Membrane protein composition - Protein compartmentalization - Protein excretion rates	(Liu et al. 2014)
Protein structural properties and metabolism	- Effect of enzyme structural characteristics on cell metabolism - Enzymes with low thermal stability which have greatest effect on cellular thermotolerance - Drug binding sites of enzymes and resultant cellular phenotype	(Zhang et al. 2009; Chang et al. 2013a; Chang et al. 2013b)
Probabilistic transcriptional regulation and metabolism	- Quantitative impact on growth rate following genetic perturbation of transcriptional regulators - Downstream effects of gene up-/down-regulation on pathways effected by regulatory network	(Chandrasekaran and Price 2010)
"Whole-cell" model	- Role of metabolism in cell-cycle regulation - Interaction of many cellular processes	(Karr et al. 2012)

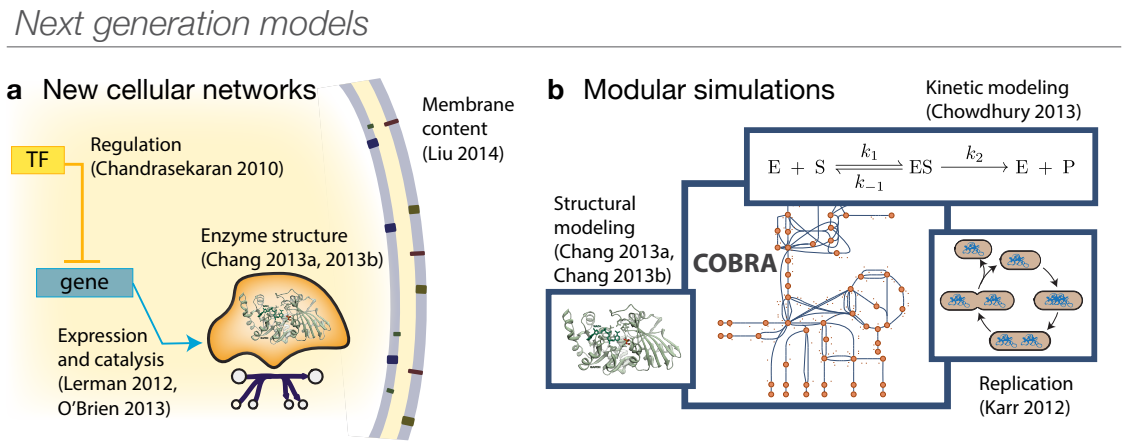

Figure 7.1: **COBRA tools have advanced through (a) increased scope of cellular systems which can be modeled, and (b) highly modular simulation strategies, built upon genome-scale models of metabolism.**

7.1.1 New cellular networks

Many of the latest improvements in COBRA modeling are based on collecting and reconstructing knowledge of cellular networks beyond metabolism (Fig. 7.1a). These next-generation networks include gene product expression coupled to metabolism, protein translocation in the cell membrane, protein structures of metabolic enzymes, and transcriptional regulation. Each of these network reconstructions has been integrated with the metabolic network, so that additional costs and constraints can be directly incorporated with existing COBRA methods.

The first model to integrate *m*etabolism with gene *e*xpression (ME-model) was developed for the minimal thermophilic bacterium *Thermotoga maritima* (Lerman et al. 2012), followed closely by the development of a ME-model for *Escherichia coli* (O'Brien et al. 2013; Thiele et al. 2012). With the inclusion of gene expression, ME-models have a host of new predictive capabilities. They directly account for the protein investment necessary for operating a metabolic pathway. In a metabolic model (M-

model), enzymes are "free;" a pathway of ten reactions has the same metabolic cost as a pathway of three reactions, as long as the overall stoichiometries of the pathways are equivalent. This often comes into play when simulating knockout mutant phenotypes that shift flux in exotic ways. In reality, longer pathways have a significantly greater enzyme production cost, and this cost is directly predicted in ME-model simulations. As an example of new predictions possible for systems engineering, the ME-model predicted acetate overflow metabolism as a result of rate-yield tradeoffs between metabolic pathways (O'Brien et al. 2013). Furthermore, the ME-model was shown to simulate batch growth conditions where the availability of enzyme protein limits the maximum substrate uptake rate and, therefore, the maximum growth rate. Both of these examples highlight important concepts in metabolic engineering that are captured by ME-models and not in models of metabolism.

The *E. coli* ME-model has been extended to include protein translocation in the cell membrane (Liu et al. 2014). In addition to the predictions described above, this extended ME-model can predict how spatial limitations in the inner and outer membranes (i.e. membrane crowding) lead to tradeoffs between energy-efficient pathways that require membrane space (e.g. electron transport chain) and less-efficient pathways that require less membrane space (e.g. fermentation). These tradeoffs are not present in other COBRA models, and they could have major effects on metabolic engineering strategies that employ the expression of a large numbers of membrane-bound enzymes for substrate and product transport.

Another recent model expansion involved the collection of all the available enzyme structures for enzymes in the metabolic network of *T. maritima* (Zhang et al. 2009), followed by the collection of enzyme structures for *E. coli* in a model called

GEM-PRO (Chang et al. 2013a). In addition to the application to thermotolerance discussed in the GEM-PRO publication, GEM-PRO can eventually be applied to exploring the multitude of effects that protein structures have on metabolic activity and regulation, including enzyme promiscuity, catalytic rates, complex formation, substrate channeling, and allosteric regulation (Zhang et al. 2009; Beltrao, Kiel, and Serrano 2007; Fisher et al. 2014).

A final area of active expansion in the scope of COBRA models is transcriptional regulation. This topic has received much attention (Gonçalves et al. 2013), and so, in this article, it will only be noted that many strain engineering strategies are based on regulatory effects. Thus, techniques like the *p*robabilistic *r*egulation *o*f *m*etabolism (PROM) approach (Chandrasekaran and Price 2010), which can predict the impact of transcriptional regulation on metabolic activity, will eventually allow for the direct prediction of regulatory modifications for strain engineering. A major limitation of PROM is that it requires large-scale empirical datasets to make accurate predictions. It is not yet possible to make forward predictions of transcriptional regulation using only the structure of the regulatory network. However, this kind of prediction will eventually be possible as the quality of binding information for transcription factors increases and the knowledgebase of regulatory interactions becomes more comprehensive (Cho et al. 2012; Lee et al. 2007; Federowicz et al. 2014; Cho et al. 2007; Carrera, Estrela, and Luo 2014; Cho et al. 2014; Seo et al. 2014)

7.1.2 Modularity

A second theme in the evolution of COBRA models is modularity of simulation strategies (Fig. 7.1b). There have been long-standing challenges associated with

genome-scale modeling of cellular networks using fine-grained approaches like stochastic and deterministic kinetic modeling, and so COBRA methods, which are tractable at the genome-scale, have risen in popularity (Bordbar et al. 2014a). By embracing modular simulations embedded in genome-scale COBRA models, it is possible to explore complex topics—including dynamics, concentrations, and physical structures—without losing genome-scale context.

While kinetic modeling is very difficult at the genome scale, the k-OptForce algorithm combines a genome-scale model of metabolism with a smaller-scale, deterministic kinetic model to identify strategies for metabolic engineering that incorporate knowledge about metabolite concentrations and kinetics (Chowdhury, Zomorrodi, and Maranas 2014). Additionally, the GEM-PRO model of metabolism and enzyme structure brings algorithms for protein structure analysis and simulation into the context of a COBRA model (Chang et al. 2013a; Chang et al. 2013b).

Finally, a 'whole-cell' model of *Mycoplasma genitalium* has been reported, and this model includes modular simulations of all annotated gene functions in that minimal organism (Karr et al. 2012). The authors employ a highly modular platform where many types of simulations, including a COBRA model of metabolism, are executed at discrete time points. This model represents an extreme approach to modular simulation where every system can exist somewhat separately but with a shared notion of time. To manage this modularity computationally, all the constituent cellular processes are decoupled and simulated over a single time step (1 second), and then the modules are synchronized before the next time step is taken. A challenge associated with of this strategy is that it is difficult to establish clear numerical convergence criteria.

7.2 Conclusion

With next-generation COBRA models, new predictions will be possible, ranging from enzyme promiscuity to allosteric regulation, from membrane crowding to overflow metabolism, and from metabolite concentration to cell cycle effects. To automate the design and creation of engineered cells, it will be necessary to incorporate detailed knowledge of interconnected cellular processes and to simulate these processes across time- and size-scales, using modular simulations. Furthermore, these new methods and models will need to be rigorously validated by repeatedly comparing predictions to experimental outcomes, in order to determine key modeling parameters and to close the gaps in our knowledge of biological systems. COBRA methods have proven their usefulness in a growing number of studies, and, as they expand to include many new cellular networks and types of simulation and prediction, the value and adoption of these methods are likely to grow.

Chapter 7 is adapted from a published manuscript: King, Z. A., Lloyd, C. J., Feist, A. M., and Palsson, B. O. (2015b). "Next-generation genome-scale models for metabolic engineering". In: *Curr. Opin. Biotechnol.* 35, pp. 23–29. DOI: 10.1016/j.copbio.2014.12.016. The dissertation author was the primary author of the review.

Bibliography

Adkins, J., Pugh, S., McKenna, R., and Nielsen, D. R. (2012). "Engineering microbial chemical factories to produce renewable "biomonomers"". In: *Front. Microbiol.* 3.August, p. 313. DOI: 10.3389/fmicb.2012.00313.

Almaas, E., Kovács, B., Vicsek, T., Oltvai, Z. N., and Barabási, A.-L. (2004). "Global organization of metabolic fluxes in the bacterium *Escherichia coli*". In: *Nature* 427.6977, pp. 839–843. DOI: 10.1038/nature02289.

Arnold, K., Bordoli, L., Kopp, J., and Schwede, T. (2006). "The SWISS-MODEL workspace: A web-based environment for protein structure homology modelling". In: *Bioinformatics* 22.2, pp. 195–201. DOI: 10.1093/bioinformatics/bti770.

Atsumi, S., Hanai, T., and Liao, J. C. (2008). "Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels". In: *Nature* 451.7174, pp. 86–89. DOI: 10.1038/nature06450.

Atsumi, S., Wu, T.-Y., Eckl, E.-M., Hawkins, S. D., Buelter, T., and Liao, J. C. (2010). "Engineering the isobutanol biosynthetic pathway in Escherichia coli by comparison of three aldehyde reductase/alcohol dehydrogenase genes". In: *Appl. Microbiol. Biotechnol.* 85.3, pp. 651–657. DOI: 10.1007/s00253-009-2085-6.

Auriol, C., Bestel-Corre, G., Claude, J.-B., Soucaille, P., and Meynial-Salles, I. (2011). "Stress-induced evolution of *Escherichia coli* points to original concepts in respiratory cofactor selectivity". In: *Proc. Natl. Acad. Sci. U. S. A.* 108.4, pp. 1278–1283. DOI: 10.1073/pnas.1010431108.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. a., Tomita, M., Wanner, B. L., and Mori, H. (2006). "Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection". In: *Mol. Syst. Biol.* 2, p. 2006.0008. DOI: 10.1038/msb4100050.

Bar-Even, A., Noor, E., and Savir, Y. (2011). "The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters". In: *Biochemistry*, pp. 4402–4410.

Basan, M., Hui, S., Zhang, Z., Shen, Y., Williamson, J. R., and Hwa, T. (2015). "Overflow metabolism in bacteria results from efficient proteome allocation for energy biogenesis". In: *Nature* 1. DOI: 10.1038/nature15765.

Becker, J. and Wittmann, C. (2012). "Systems and synthetic metabolic engineering for amino acid production – the heartbeat of industrial strain development". In: *Curr. Opin. Biotechnol.* 23.5, pp. 718–726. DOI: 10.1016/j.copbio.2011.12.025.

Becker, S. A., Feist, A. M., Mo, M. L., Hannum, G., Palsson, B. Ø., and Herrgård, M. J. (2007). "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox". In: *Nat. Protoc.* 2.3, pp. 727–738. DOI: 10.1038/nprot.2007.99.

Beltrao, P., Kiel, C., and Serrano, L. (2007). "Structures in systems biology". In: *Curr. Opin. Struct. Biol.* 17.3, pp. 378–384. DOI: 10.1016/j.sbi.2007.05.005.

Bengtsson, O., Hahn-Hägerdal, B., and Gorwa-Grauslund, M. F. (2009). "Xylose reductase from *Pichia stipitis* with altered coenzyme preference improves ethanolic xylose fermentation by recombinant *Saccharomyces cerevisiae*". In: *Biotechnol. Biofuels* 2, p. 9. DOI: 10.1186/1754-6834-2-9.

Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2014). "GenBank". In: *Nucleic Acids Res.* 43.D1, pp. D30–D35. DOI: 10.1093/nar/gku1216.

Berríos-Rivera, S. J., Sánchez, a. M., Bennett, G. N., and San, K.-Y. (2004). "Effect of different levels of NADH availability on metabolite distribution in *Escherichia coli* fermentation in minimal and complex media". In: *Appl. Microbiol. Biotechnol.* 65.4, pp. 426–432. DOI: 10.1007/s00253-004-1609-3.

Berríos-Rivera, S. J., Bennett, G. N., and San, K.-Y. (2002a). "Metabolic Engineering of *Escherichia coli*: Increase of NADH Availability by Overexpressing an NAD^+-Dependent Formate Dehydrogenase". In: *Metab. Eng.* 4.3, pp. 217–229. DOI: 10.1006/mben.2002.0227.

— (2002b). "The Effect of Increasing NADH Availability on the Redistribution of Metabolic Fluxes in *Escherichia coli* Chemostat Cultures". In: *Metab. Eng.* 4.3, pp. 230–237. DOI: 10.1006/mben.2002.0228.

Berríos-Rivera, S. J., San, K.-Y., and Bennett, G. N. (2002). "The Effect of NAPRTase Overexpression on the Total Levels of NAD, The NADH/NAD^+ Ratio, and the Distribution of Metabolites in *Escherichia coli*". In: *Metab. Eng.* 4.3, pp. 238–247. DOI: 10.1006/mben.2002.0229.

Bettiga, M., Hahn-Hägerdal, B., and Gorwa-Grauslund, M. F. (2008). "Comparing the xylose reductase/xylitol dehydrogenase and xylose isomerase pathways in arabinose and xylose fermenting *Saccharomyces cerevisiae* strains". In: *Biotechnol. Biofuels* 1.1, p. 16. DOI: 10.1186/1754-6834-1-16.

Blankschien, M. D., Clomburg, J. M., and Gonzalez, R. (2010). "Metabolic engineering of *Escherichia coli* for the production of succinate from glycerol". In: *Metab. Eng.* 12.5, pp. 409–419. DOI: 10.1016/j.ymben.2010.06.002.

Blaschkowski, H. P., Neuer, G., Ludwig-Festl, M., and Knappe, J. (1982). "Routes of flavodoxin and ferredoxin reduction in *Escherichia coli*". In: *Eur. J. Biochem.* 123.3, pp. 563–569. DOI: 10.1111/j.1432-1033.1982.tb06569.x.

Bocanegra, J. A., Scrutton, N. S., and Perham, R. N. (1993). "Creation of an NADP-dependent pyruvate dehydrogenase multienzyme complex by protein engineering". In: *Biochemistry* 32.11, pp. 2737–2740. DOI: 10.1021/bi00062a001.

Bordbar, A., Feist, A. M., Usaite-Black, R., Woodcock, J., Palsson, B. O., and Famili, I. (2011). "A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology". In: *BMC Syst. Biol.* 5, p. 180. DOI: 10.1186/1752-0509-5-180.

Bordbar, A., Lewis, N. E., Schellenberger, J., Palsson, B. Ø., and Jamshidi, N. (2010). "Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions". In: *Mol. Syst. Biol.* 6, p. 422. DOI: 10.1038/msb.2010.68.

Bordbar, A., Monk, J. M., King, Z. A., and Palsson, B. O. (2014a). "Constraint-based models predict metabolic and associated cellular functions". In: *Nat. Rev. Genet.* 15.2, pp. 107–120. DOI: 10.1038/nrg3643.

Bordbar, A., Nagarajan, H., Lewis, N. E., Latif, H., Ebrahim, A., Federowicz, S., Schellenberger, J., and Palsson, B. O. (2014b). "Minimal metabolic pathway structure is consistent with associated biomolecular interactions". In: *Mol. Syst. Biol.* 10, p. 737. DOI: 10.15252/msb.20145243.

Bostock, M., Ogievetsky, V., and Heer, J. (2011). "D3: Data-Driven Documents". In: *IEEE Trans. Vis. Comput. Graph.* 17.12, pp. 2301–2309. DOI: 10.1109/TVCG.2011.185.

Brown, G. R., Hem, V., Katz, K. S., Ovetsky, M., Wallin, C., Ermolaeva, O., Tolstoy, I., Tatusova, T., Pruitt, K. D., Maglott, D. R., and Murphy, T. D. (2014).

"Gene: a gene-centered information resource at NCBI". In: *Nucleic Acids Res.* 43.D1, pp. D36–D42. DOI: 10.1093/nar/gku1055.

Burgard, A. P., Pharkya, P., and Maranas, C. D. (2003). "Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization". In: *Biotechnol. Bioeng.* 84.6, pp. 647–657. DOI: 10.1002/bit.10803.

Campodonico, M. A., Andrews, B. A., Asenjo, J. A., Palsson, B. O., and Feist, A. M. (2014). "Generation of an atlas for commodity chemical production in *Escherichia coli* and a novel pathway prediction algorithm, GEM-Path". In: *Metab. Eng.* 25, pp. 140–158. DOI: 10.1016/j.ymben.2014.07.009.

Carlson, R. and Srienc, F. (2004a). "Fundamental *Escherichia coli* biochemical pathways for biomass and energy production: creation of overall flux states". In: *Biotechnol. Bioeng.* 86.2, pp. 149–162. DOI: 10.1002/bit.20044.

— (2004b). "Fundamental *Escherichia coli* biochemical pathways for biomass and energy production: identification of reactions". In: *Biotechnol. Bioeng.* 85.1, pp. 1–19. DOI: 10.1002/bit.10812.

Carrera, J., Estrela, R., and Luo, J. (2014). "An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of *Escherichia coli*". In: *Mol. Syst. Biol.* 10.735, pp. 1–13.

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., Holland, T. A., Keseler, I. M., Kothari, A., Kubo, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D. S., Weerasinghe, D., Zhang, P., and Karp, P. D. (2014). "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases". In: *Nucleic Acids Res.* 42, pp. D459–D471. DOI: 10.1093/nar/gkt1103.

Chandrasekaran, S. and Price, N. D. (2010). "Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and Mycobacterium tuberculosis". In: *Proc. Natl. Acad. Sci. U. S. A.* 107.41, pp. 17845–17850. DOI: 10.1073/pnas.1005139107.

Chang, R. L., Andrews, K., Kim, D., Li, Z., Godzik, A., and Palsson, B. O. (2013a). "Structural Systems Biology Evaluation of Metabolic Thermotolerance in *Escherichia coli*". In: *Science* 340.6137, pp. 1220–1223. DOI: 10.1126/science.1234012.

Chang, R. L., Xie, L., Bourne, P. E., and Palsson, B. O. (2013b). "Antibacterial mechanisms identified through structural systems pharmacology". In: *BMC Syst. Biol.* 7.1, p. 102. DOI: 10.1186/1752-0509-7-102.

Check Hayden, E. (2015). "Synthetic biologists seek standards for nascent field". In: *Nature News* 520.7546, p. 141. DOI: 10.1038/520141a.

Chin, J. W., Khankal, R., Monroe, C. a., Maranas, C. D., and Cirino, P. C. (2009). "Analysis of NADPH supply during xylitol production by engineered *Escherichia coli*". In: *Biotechnol. Bioeng.* 102.1, pp. 209–220. DOI: 10.1002/bit.22060.

Cho, B.-K., Charusanti, P., Herrgård, M. J., and Palsson, B. O. (2007). "Microbial regulatory and metabolic networks". In: *Curr. Opin. Biotechnol.* 18.4, pp. 360–364. DOI: 10.1016/j.copbio.2007.07.002.

Cho, B.-K., Federowicz, S., Park, Y.-S., Zengler, K., and Palsson, B. Ø. (2012). "Deciphering the transcriptional regulatory logic of amino acid metabolism". In: *Nat. Chem. Biol.* 8.1, pp. 65–71. DOI: 10.1038/nchembio.710.

Cho, B.-K., Kim, D., Knight, E. M., Zengler, K., and Palsson, B. O. (2014). "Genome-scale reconstruction of the sigma factor network in *Escherichia coli*: topology and functional states". In: *BMC Biol.* 12, p. 4. DOI: 10.1186/1741-7007-12-4.

Choi, H. S., Lee, S. Y., Kim, T. Y., and Woo, H. M. (2010). "In silico identification of gene amplification targets for improvement of lycopene production". In: *Appl. Environ. Microbiol.* 76.10, pp. 3097–3105. DOI: 10.1128/AEM.00115-10.

Chowdhury, A., Zomorrodi, A. R., and Maranas, C. D. (2014). "k-OptForce: integrating kinetics with flux balance analysis for strain design". In: *PLoS Comput. Biol.* 10.2, e1003487. DOI: 10.1371/journal.pcbi.1003487.

Chubukov, V., Mukhopadhyay, A., Petzold, C. J., Keasling, J. D., and Martín, H. G. (2016). "Synthetic and systems biology for microbial production of commodity chemicals". In: *npj Systems Biology and Applications* 2, p. 16009. DOI: 10.1038/npjsba.2016.9.

Chung, B. K.-S., Lakshmanan, M., Klement, M., Mohanty, B., and Lee, D.-Y. (2013). "Genome-scale *in silico* modeling and analysis for designing synthetic terpenoid-producing microbial cell factories". In: *Chem. Eng. Sci.* 103.15, pp. 100–108. DOI: 10.1016/j.ces.2012.09.006.

Chung, H. J., Kim, M., Park, C. H., Kim, J., and Kim, J. H. (2004). "ArrayXPath: Mapping and visualizing microarray gene-expression data with integrated

biological pathway resources using Scalable Vector Graphics". In: *Nucleic Acids Res.* 32, pp. 621–626. DOI: 10.1093/nar/gkh476.

Clark, David P (1989). "The fermentation pathways of *Escherichia coli*". In: *FEMS Microbiol. Rev.* 63, pp. 223–234.

Costanzo, M. C., Engel, S. R., Wong, E. D., Lloyd, P., Karra, K., Chan, E. T., Weng, S., Paskov, K. M., Roe, G. R., Binkley, G., Hitz, B. C., and Cherry, J. M. (2014). "Saccharomyces genome database provides new regulation data". In: *Nucleic Acids Res.* 42, pp. D717–D725. DOI: 10.1093/nar/gkt1158.

Courtot, M., Juty, N., Knüpfer, C., Waltemath, D., Zhukova, A., Dräger, A., Dumontier, M., Finney, A., Golebiewski, M., Hastings, J., Hoops, S., Keating, S., Kell, D. B., Kerrien, S., Lawson, J., Lister, A., Lu, J., Machne, R., Mendes, P., Pocock, M., Rodriguez, N., Villeger, A., Wilkinson, D. J., Wimalaratne, S., Laibe, C., Hucka, M., and Le Novère, N. (2011). "Controlled vocabularies and semantics in systems biology". In: *Mol. Syst. Biol.* 7, p. 543. DOI: 10.1038/msb.2011.77.

Covert, M. W., Knight, E. M., Reed, J. L., Herrgård, M. J., and Palsson, B. O. (2004). "Integrating high-throughput and computational data elucidates bacterial networks". In: *Nature* 429.6987, pp. 92–96. DOI: 10.1038/nature02456.

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shamovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L., and D'Eustachio, P. (2014). "The Reactome pathway knowledgebase". In: *Nucleic Acids Res.* 42, pp. D472–D477. DOI: 10.1093/nar/gkt1102.

Czauderna, T., Klukas, C., and Schreiber, F. (2010). "Editing, validating and translating of SBGN maps". In: *Bioinformatics* 26.18, pp. 2340–2341. DOI: 10.1093/bioinformatics/btq407.

Dellomonaco, C., Clomburg, J. M., Miller, E. N., and Gonzalez, R. (2011). "Engineered reversal of the β-oxidation cycle for the synthesis of fuels and chemicals". In: *Nature* 476.7360, pp. 355–359. DOI: 10.1038/nature10333.

Dolinski, K. and Troyanskaya, O. G. (2015). "Implications of Big Data for cell biology". In: *Mol. Biol. Cell* 26.14, pp. 2575–2578. DOI: 10.1091/mbc.E13-12-0756.

Donnelly, M. I., Millard, C. S., Clark, D. P., Chen, M. J., and Rathke, J. W. (1998). "A novel fermentation pathway in an *Escherichia coli* mutant producing succinic

acid, acetic acid, and ethanol". In: *Appl. Biochem. Biotechnol.* 70.1, pp. 187–198.

Dräger, A. and Palsson, B. Ø. (2014). "Improving collaboration by standardization efforts in systems biology". In: *Frontiers in Bioengineering and Biotechnology* 2, p. 61. DOI: 10.3389/fbioe.2014.00061.

Droste, P., Nöh, K., and Wiechert, W. (2013). "Omix - A visualization tool for metabolic networks with highest usability and customizability in focus". In: *Chemie-Ingenieur-Technik* 85.6, pp. 849–862. DOI: 10.1002/cite.201200234.

Ebrahim, A., Brunk, E., Tan, J., O'Brien, E. J., Kim, D., Szubin, R., Lerman, J. A., Lechner, A., Sastry, A., Bordbar, A., Feist, A. M., and Palsson, B. O. "Multi-omic data integration enables discovery of hidden biological regularities". In: *Nature Communications*. In press.

Ebrahim, A., Lerman, J. A., Palsson, B. O., and Hyduke, D. R. (2013). "COBRApy: COnstraints-Based Reconstruction and Analysis for Python". In: *BMC Syst. Biol.* 7, p. 74. DOI: 10.1186/1752-0509-7-74.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". In: *Nucleic Acids Res.* 30.1, pp. 207–210. DOI: 10.1093/nar/30.1.207.

Edwards, J. S., Ibarra, R. U., and Palsson, B. O. (2001). "In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data". In: *Nat. Biotechnol.* 19.2, pp. 125–130. DOI: 10.1038/84379.

Eppig, J. T., Blake, J. A., Bult, C. J., Kadin, J. A., Richardson, J. E., and The Mouse Genome Database Group (2014). "The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease". In: *Nucleic Acids Res.* 43, pp. D726–D736. DOI: 10.1093/nar/gku967.

Federowicz, S., Kim, D., Ebrahim, A., Lerman, J., Nagarajan, H., Cho, B.-K., Zengler, K., and Palsson, B. (2014). "Determining the control circuitry of redox metabolism at the genome-scale". In: *PLoS Genet.* 10.4, e1004264. DOI: 10.1371/journal.pgen.1004264.

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V., and Palsson, B. Ø. (2007). "A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information". In: *Mol. Syst. Biol.* 3.121, p. 121. DOI: 10.1038/msb4100155.

Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L., and Palsson, B. Ø. (2009). "Reconstruction of biochemical networks in microorganisms". In: *Nat. Rev. Microbiol.* 7.2, pp. 129–143. DOI: 10.1038/nrmicro1949.

Feist, A. M. and Palsson, B. Ø. (2008). "The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*". In: *Nat. Biotechnol.* 26.6, pp. 659–667. DOI: 10.1038/nbt1401.

Feist, A. M., Zielinski, D. C., Orth, J. D., Schellenberger, J., Herrgård, M. J., and Palsson, B. Ø. (2010). "Model-driven evaluation of the production potential for growth-coupled products of *Escherichia coli*". In: *Metab. Eng.* 12.3, pp. 173–186. DOI: 10.1016/j.ymben.2009.10.003.

Fisher, A. K., Freedman, B. G., Bevan, D. R., and Senger, R. S. (2014). "A review of metabolic and enzymatic engineering strategies for designing and optimizing performance of microbial cell factories". In: *Comput. Struct. Biotechnol. J.* 11.18, pp. 91–99. DOI: 10.1016/j.csbj.2014.08.010.

Fong, S. S., Burgard, A. P., Herring, C. D., Knight, E. M., Blattner, F. R., Maranas, C. D., and Palsson, B. O. (2005). "In silico design and adaptive evolution of *Escherichia coli* for production of lactic acid". In: *Biotechnol. Bioeng.* 91.5, pp. 643–648. DOI: 10.1002/bit.20542.

Fong, S. S., Nanchen, A., Palsson, B. Ø., and Sauer, U. (2006). "Latent pathway activation and increased pathway capacity enable *Escherichia coli* adaptation to loss of key metabolic enzyme". In: *J. Biol. Chem.* 281.12, pp. 8024–8033. DOI: 10.1074/jbc.M510016200.

Fong, S. S. and Palsson, B. Ø. (2004). "Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes". In: *Nat. Genet.* 36.10, pp. 1056–1058. DOI: 10.1038/ng1432.

Fuhrer, T. and Sauer, U. (2009). "Different biochemical mechanisms ensure network-wide balancing of reducing equivalents in microbial metabolism". In: *J. Bacteriol.* 191.7, pp. 2112–2121. DOI: 10.1128/JB.01523-08.

Funahashi, A., Matsuoka, Y., Jouraku, A., Morohashi, M., Kikuchi, N., and Kitano, H. (2008). "CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks". In: *Proc. IEEE* 96.8, pp. 1254–1265. DOI: 10.1109/JPROC.2008.925458.

Gallina, A. A., Layer, M., King, Z. A., Levering, J., Palsson, B. Ø., Zengler, K., and Peers, G. (2016). "A Phaeodactylum tricornutum literature database for interactive annotation of content". In: *Algal Research* 18, pp. 241–243.

Ganter, M., Bernard, T., Moretti, S., Stelling, J., and Pagni, M. (2013). "MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks". In: *Bioinformatics* 29.6, pp. 815–816. DOI: 10.1093/bioinformatics/btt036.

Gauges, R., Rost, U., Sahle, S., and Wegner, K. (2006). "A model diagram layout extension for SBML". In: *Bioinformatics* 22.15, pp. 1879–1885. DOI: 10.1093/bioinformatics/btl195.

Gavai, A. K., Supandi, F., Hettling, H., Murrell, P., Leunissen, J. A. M., and Beek, J. H. G. M. van (2015). "Using Bioconductor Package BiGGR for Metabolic Flux Estimation Based on Gene Expression Changes in Brain". In: *PLoS One* 10.3, e0119016. DOI: 10.1371/journal.pone.0119016.

Gawand, P., Hyland, P., Ekins, A., Martin, V. J. J., and Mahadevan, R. (2013). "Novel approach to engineer strains for simultaneous sugar utilization". In: *Metab. Eng.* 20, pp. 63–72. DOI: 10.1016/j.ymben.2013.08.003.

Ghosh, A., Zhao, H., and Price, N. D. (2011). "Genome-scale consequences of co-factor balancing in engineered pentose utilization pathways in *Saccharomyces cerevisiae*". In: *PLoS One* 6.11, e27316. DOI: 10.1371/journal.pone.0027316.

Gianchandani, E. P., Joyce, A. R., Palsson, B. Ø., and Papin, J. a. (2009). "Functional states of the genome-scale *Escherichia coli* transcriptional regulatory system". In: *PLoS Comput. Biol.* 5.6, e1000403. DOI: 10.1371/journal.pcbi.1000403.

Gonçalves, E., Bucher, J., Ryll, A., Niklas, J., Mauch, K., Klamt, S., Rocha, M., and Saez-Rodriguez, J. (2013). "Bridging the layers: towards integration of signal transduction, regulation and metabolism into mathematical models". In: *Mol. Biosyst.* 9.7, pp. 1576–1583. DOI: 10.1039/c3mb25489e.

Gottschalk, G. (1986). *Bacterial metabolism*. 2nd. New York, NY: Springer-Verlag.

Guest, J. R., Abdel-Hamid, A. M., Auger, G. A., Cunningham, L., Henderson, R. A., Machado, R. S., and Attwood, M. M. (2003). "Physiological Effects of Replacing the PDH Complex of *E. coli* by Genetically Engineered Variants or by Pyruvate Oxidase". In: *Thiamine: Catalytic Mechanisms in Normal and Disease States*. Ed. by F. Gordon and S. P. Mulchand. New York, NY: CRC Press. Chap. 22, pp. 387–407. DOI: 10.1201/9780203913420.ch22.

Guzmán, G. I., Utrilla, J., Nurk, S., Brunk, E., Monk, J. M., Ebrahim, A., Palsson, B. O., and Feist, A. M. (2015). "Model-driven discovery of underground

metabolic functions in *Escherichia coli*". In: *Proceedings of the National Academy of Sciences* 112.3, pp. 929–934. DOI: 10.1073/pnas.1414218112.

Hanahan, D. and Weinberg, R. A. (2011). "Hallmarks of cancer: the next generation". In: *Cell* 144.5, pp. 646–674. DOI: 10.1016/j.cell.2011.02.013.

Heavner, B. D., Smallbone, K., Barker, B., Mendes, P., and Walker, L. P. (2012). "Yeast 5 – an expanded reconstruction of the Saccharomyces cerevisiae metabolic network". In: *BMC Syst. Biol.* 6.1, p. 55. DOI: 10.1186/1752-0509-6-55.

Hefzi, H., Ang, K. S., Hanscho, M., Bordbar, A., Ruckerbauer, D., Lakshmanan, M., Orellana, C. A., Baycin-Hizal, D., Huang, Y., Ley, D., Martinez, V. S., Kyriakopoulos, S., Jiménez, N. E., Zielinski, D. C., Quek, L.-E., Wulff, T., Arnsdorf, J., Li, S., Lee, J. S., Paglia, G., Loira, N., Spahn, P. N., Pedersen, L. E., Gutierrez, J. M., King, Z. A., Lund, A. M., Nagarajan, H., Thomas, A., Abdel-Haleem, A. M., Zanghellini, J., Kildegaard, H. F., Voldborg, B. G., Gerdtzen, Z. P., Betenbaugh, M. J., Palsson, B. O., Andersen, M. R., Nielsen, L. K., Borth, N., Lee, D.-Y., and Lewis, N. E. (2016). "A Consensus Genome-scale Reconstruction of Chinese Hamster Ovary Cell Metabolism". In: *Cell Syst* 3.5, 434–443.e8. DOI: 10.1016/j.cels.2016.10.020.

Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., and Stevens, R. L. (2010). "High-throughput generation, optimization and analysis of genome-scale metabolic models". In: *Nat. Biotechnol.* 28.9, pp. 977–982. DOI: 10.1038/nbt.1672.

Herráez, A. (2006). "Biomolecules in the computer: Jmol to the rescue". In: *Biochem. Mol. Biol. Educ.* 34, pp. 255–261. DOI: 10.1002/bmb.2006.494034042644.

Hu, Z., Chang, Y. C., Wang, Y., Huang, C. L., Liu, Y., Tian, F., Granger, B., and Delisi, C. (2013). "VisANT 4.0: Integrative network platform to connect genes, drugs, diseases and therapies". In: *Nucleic Acids Res.* 41.May, pp. 225–231. DOI: 10.1093/nar/gkt401.

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I. I., Hedley, W. J., Hodgman, T. C., Hofmeyr, J. H., Hunter, P. J., Juty, N. S., Kasberger, J. L., Kremling, A., Kummer, U., Le Novère, N., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E. D., Nakayama, Y., Nelson, M. R., Nielsen, P. F., Sakurada, T., Schaff, J. C., Shapiro, B. E., Shimizu, T. S., Spence, H. D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., and Wang, J. (2003). "The systems biology markup language (SBML): A medium for representation and exchange

of biochemical network models". In: *Bioinformatics* 19.4, pp. 524–531. DOI: 10.1093/bioinformatics/btg015.

Hurley, J. H., Chen, R., and Dean, A. M. (1996). "Determinants of Cofactor Specificity in Isocitrate Dehydrogenase: Structure of an Engineered $NADP^+ \rightarrow NAD^+$ Specificity-Reversal Mutant". In: *Biochemistry* 35.18, pp. 5670–5678.

Huson, D. H., Richter, D. C., Rausch, C., Dezulian, T., Franz, M., and Rupp, R. (2007). "Dendroscope: An interactive viewer for large phylogenetic trees". In: *BMC Bioinformatics* 8, p. 460. DOI: 10.1186/1471-2105-8-460.

Hyduke, D. R., Lewis, N. E., and Palsson, B. Ø. (2013). "Analysis of omics data with genome-scale models of metabolism". In: *Mol. Biosyst.* 9.2, pp. 167–174. DOI: 10.1039/c2mb25453k.

Ibarra, R. U., Edwards, J. S., and Palsson, B. O. (2002). "*Escherichia coli* K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth". In: *Nature* 420.November, pp. 20–23. DOI: 10.1038/nature01195.1..

Iersel, M. P. van, Villéger, A. C., Czauderna, T., Boyd, S. E., Bergmann, F. T., Luna, A., Demir, E., Sorokin, A., Dogrusoz, U., Matsuoka, Y., Funahashi, A., Aladjem, M. I., Mi, H., Moodie, S. L., Kitano, H., Le novère, N., and Schreiber, F. (2012). "Software support for SBGN maps: SBGN-ML and LibSBGN". In: *Bioinformatics* 28.15, pp. 2016–2021. DOI: 10.1093/bioinformatics/bts270.

Ingram, L. O., Conway, T., Clark, D. P., Sewell, G. W., and Preston, J. F. (1987). "Genetic engineering of ethanol production in *Escherichia coli*". In: *Appl. Environ. Microbiol.* 53.10, pp. 2420–2425.

Jan, J., Martinez, I., Wang, Y., Bennett, G. N., and San, K.-Y. (2013). "Metabolic engineering and transhydrogenase effects on NADPH availability in *Escherichia coli*". In: *Biotechnol. Prog.* 29.5, pp. 1124–1130. DOI: 10.1002/btpr.1765.

Jang, Y.-S., Kim, B., Shin, J. H., Choi, Y. J., Choi, S., Song, C. W., Lee, J., Park, H. G., and Lee, S. Y. (2012). "Bio-based production of C2-C6 platform chemicals". In: *Biotechnol. Bioeng.* 109.10, pp. 2437–2459. DOI: 10.1002/bit.24599.

Johnson, F. X. (2007). *Industrial Biotechnology and Biomass Utilisation: Prospects and Challenges for the Developing World*. Tech. rep. Vienna, Austria: United Nations Industrial Development Organization.

Jung, Y. K., Kim, T. Y., Park, S. J., and Lee, S. Y. (2010). "Metabolic engineering of *Escherichia coli* for the production of polylactic acid and its copolymers". In: *Biotechnol. Bioeng.* 105.1, pp. 161–171. DOI: 10.1002/bit.22548.

Juty, N., Le Novere, N., and Laibe, C. (2012). "Identifiers.org and MIRIAM Registry: Community resources to provide persistent identification". In: *Nucleic Acids Res.* 40, pp. D580–D586. DOI: 10.1093/nar/gkr1097.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). "Data, information, knowledge and principle: back to metabolism in KEGG". In: *Nucleic Acids Res.* 42, pp. D199–D205. DOI: 10.1093/nar/gkt1076.

Karolchik, D., Barber, G. P., Casper, J., Clawson, H., Cline, M. S., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M., Harte, R. A., Heitner, S., Hinrichs, A. S., Learned, K., Lee, B. T., Li, C. H., Raney, B. J., Rhead, B., Rosenbloom, K. R., Sloan, C. A., Speir, M. L., Zweig, A. S., Haussler, D., Kuhn, R. M., and Kent, W. J. (2014). "The UCSC Genome Browser database: 2014 update". In: *Nucleic Acids Res.* 42.D1, pp. 764–770. DOI: 10.1093/nar/gkt1168.

Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B., Assad-Garcia, N., Glass, J. I., and Covert, M. W. (2012). "A whole-cell computational model predicts phenotype from genotype". In: *Cell* 150.2, pp. 389–401. DOI: 10.1016/j.cell.2012.05.044.

Kartal, O., Mahlow, S., Skupin, A., and Ebenhöh, O. (2011). "Carbohydrate-active enzymes exemplify entropic principles in metabolism". In: *Mol. Syst. Biol.* 7.542, p. 542. DOI: 10.1038/msb.2011.76.

Kauffman, K. J., Prakash, P., and Edwards, J. S. (2003). "Advances in flux balance analysis". In: *Curr. Opin. Biotechnol.* 14.5, pp. 491–496. DOI: 10.1016/j.copbio.2003.08.001.

Kelder, T., Iersel, M. P. van, Hanspers, K., Kutmon, M., Conklin, B. R., Evelo, C. T., and Pico, A. R. (2012). "WikiPathways: Building research communities on biological pathways". In: *Nucleic Acids Res.* 40.November 2011, pp. 1301–1307. DOI: 10.1093/nar/gkr1074.

Kim, M., Sang Yi, J., Kim, J., Kim, J.-N., Kim, M. W., and Kim, B.-G. (2014). "Reconstruction of a high-quality metabolic model enables the identification of gene overexpression targets for enhanced antibiotic production in Streptomyces coelicolor A3(2)". In: *Biotechnol. J.* 9.9, pp. 1185–1194. DOI: 10.1002/biot.201300539.

Kim, S., Lee, C. H., Nam, S. W., and Kim, P. (2011). "Alteration of reducing powers in an isogenic phosphoglucose isomerase (pgi)-disrupted *Escherichia coli* expressing NAD(P)-dependent malic enzymes and NADP-dependent glyceraldehyde 3-phosphate dehydrogenase". In: *Lett. Appl. Microbiol.* 52.5, pp. 433–440. DOI: 10.1111/j.1472-765X.2011.03013.x.

King, Z. A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N. E., and Palsson, B. O. (2015a). "Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways". In: *PLoS Comput. Biol.* 11.8, e1004321. DOI: 10.1371/journal.pcbi.1004321.

King, Z. A. and Feist, A. M. (2013). "Optimizing Cofactor Specificity of Oxidoreductase Enzymes for the Generation of Microbial Production Strains—OptSwap". In: *Ind. Biotechnol.* 9.4, pp. 236–246. DOI: 10.1089/ind.2013.0005.

— (2014). "Optimal cofactor swapping can increase the theoretical yield for chemical production in *Escherichia coli* and Saccharomyces cerevisiae". In: *Metab. Eng.* 24, pp. 117–128. DOI: 10.1016/j.ymben.2014.05.009.

King, Z. A., Lloyd, C. J., Feist, A. M., and Palsson, B. O. (2015b). "Next-generation genome-scale models for metabolic engineering". In: *Curr. Opin. Biotechnol.* 35, pp. 23–29. DOI: 10.1016/j.copbio.2014.12.016.

King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O., and Lewis, N. E. (2016). "BiGG Models: A platform for integrating, standardizing and sharing genome-scale models". In: *Nucleic Acids Res.* 44.D1, pp. D515–22. DOI: 10.1093/nar/gkv1049.

King, Z. A., O'Brien, E. J., Feist, A. M., and Palsson, B. O. "Literature mining supports a next-generation modeling approach to predict cellular byproduct secretion". In: *Metabolic Engineering.* under review.

Kitano, H., Funahashi, A., Matsuoka, Y., and Oda, K. (2005). "Using process diagrams for the graphical representation of biological networks". In: *Nat. Biotechnol.* 23.8, pp. 961–966. DOI: 10.1038/nbt1111.

Klitgord, N. and Segrè, D. (2010). "Environments that induce synthetic microbial ecosystems". In: *PLoS Comput. Biol.* 6.11, e1001002. DOI: 10.1371/journal.pcbi.1001002.

Kono, N., Arakawa, K., Ogawa, R., Kido, N., Oshita, K., Ikegami, K., Tamaki, S., and Tomita, M. (2009). "Pathway projector: Web-based zoomable pathway browser

using KEGG Atlas and Google Maps API". In: *PLoS One* 4.11, e7710. DOI: 10.1371/journal.pone.0007710.

Krause, F., Schulz, M., Ripkens, B., Flöttmann, M., Krantz, M., Klipp, E., and Handorf, T. (2013). "Biographer: Web-based editing and rendering of SBGN compliant biochemical networks". In: *Bioinformatics* 29, pp. 1467–1468. DOI: 10.1093/bioinformatics/btt159.

Kumar, A., Suthers, P. F., and Maranas, C. D. (2012). "MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases". In: *BMC Bioinformatics* 13, p. 6. DOI: 10.1186/1471-2105-13-6.

Kutmon, M., Iersel, M. P. van, Bohler, A., Kelder, T., Nunes, N., Pico, A. R., and Evelo, C. T. (2015). "PathVisio 3: An Extendable Pathway Analysis Toolbox". In: *PLoS Comput. Biol.* 11, e1004085. DOI: 10.1371/journal.pcbi.1004085.

Lakshmanan, M., Chung, B. K. S., Liu, C., Kim, S.-W., and Lee, D.-Y. (2013). "Cofactor Modificiation Analysis: A Computational Framework to Identify Cofactor Specificity Engineering Targets for Strain Improvement". In: *Journal of bioinformatics and biotechnology* 11.6, p. 1343006.

Latendresse, M. and Karp, P. D. (2011). "Web-based metabolic network visualization with a zooming user interface". In: *BMC Bioinformatics* 12.1, p. 176. DOI: 10.1186/1471-2105-12-176.

Lee, S. J., Lee, D. Y., Kim, T. Y., Kim, B. H., Lee, J., and Lee, S. Y. (2005). "Metabolic Engineering of *Escherichia coli* for Enhanced Production of Succinic Acid , Based on Genome Comparison and In Silico Gene Knockout Simulation". In: *Appl. Environ. Microbiol.* 71.12, p. 7880. DOI: 10.1128/AEM.71.12.7880.

Lee, S. Y. (1996). "Bacterial polyhydroxyalkanoates". In: *Biotechnol. Bioeng.* 49, pp. 1–14. DOI: 10.1002/(SICI)1097-0290(19960105)49:1<1::AID-BIT1>3.0.CO;2-P.

Lee, S. Y. and Kim, H. U. (2015). "Systems strategies for developing industrial microbial strains". In: *Nat. Biotechnol.* 33.10, pp. 1061–1072. DOI: 10.1038/nbt.3365.

Lee, S. G., Park, J. H., Hou, B. K., Kim, Y. H., Kim, C. M., and Hwang, K. S. (2007). "Effect of weight-added regulatory networks on constraint-based metabolic models of *Escherichia coli*". In: *Biosystems.* 90.3, pp. 843–855. DOI: 10.1016/j.biosystems.2007.05.003.

Lee, W.-H., Kim, M.-D., Jin, Y.-S., and Seo, J.-H. (2013). "Engineering of NADPH regenerators in *Escherichia coli* for enhanced biotransformation". In: *Appl. Microbiol. Biotechnol.* DOI: 10.1007/s00253-013-4750-z.

Lerman, J. A., Hyduke, D. R., Latif, H., Portnoy, V. A., Lewis, N. E., Orth, J. D., Schrimpe-Rutledge, A. C., Smith, R. D., Adkins, J. N., Zengler, K., and Palsson, B. O. (2012). "In silico method for modelling metabolism and gene product expression at genome scale". In: *Nat. Commun.* 3.may, p. 929. DOI: 10.1038/ncomms1928.

Letunic, I. and Bork, P. (2007). "Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation". In: *Bioinformatics* 23.1, pp. 127–128. DOI: 10.1093/bioinformatics/btl529.

Lewis, N. E. and Abdel-Haleem, A. M. (2013). "The evolution of genome-scale models of cancer metabolism". In: *Front. Physiol.* 4.237, pp. 1–7. DOI: 10.3389/fphys.2013.00237.

Lewis, N. E., Hixson, K. K., Conrad, T. M., Lerman, J. a., Charusanti, P., Polpitiya, A. D., Adkins, J. N., Schramm, G., Purvine, S. O., Lopez-Ferrer, D., Weitz, K. K., Eils, R., König, R., Smith, R. D., and Palsson, B. Ø. (2010a). "Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models". In: *Mol. Syst. Biol.* 6.390, p. 390. DOI: 10.1038/msb.2010.47.

Lewis, N. E., Nagarajan, H., and Palsson, B. Ø. (2012). "Constraining the metabolic genotype-phenotype relationship using a phylogeny of *in silico* methods". In: *Nat. Rev. Microbiol.* 10.4, pp. 291–305. DOI: 10.1038/nrmicro2737.

Lewis, N. E., Schramm, G., Bordbar, A., Schellenberger, J., Andersen, M. P., Cheng, J. K., Patel, N., Yee, A., Lewis, R. A., Eils, R., König, R., and Palsson, B. Ø. (2010b). "Large-scale *in silico* modeling of metabolic interactions between cell types in the human brain". In: *Nat. Biotechnol.* 28.12, pp. 1279–1285. DOI: 10.1038/nbt.1711.

Li, F., Thiele, I., Jamshidi, N., and Palsson, B. Ø. (2009). "Identification of potential pathway mediation targets in Toll-like receptor signaling". In: *PLoS Comput. Biol.* 5.2, e1000292. DOI: 10.1371/journal.pcbi.1000292.

Lim, J. H., Seo, S. W., Kim, S. Y., and Jung, G. Y. (2013). "Model-driven rebalancing of the intracellular redox state for optimization of a heterologous n-butanol pathway in Escherichia coli". In: *Metab. Eng.* 20, pp. 56–62. DOI: 10.1016/j.ymben.2013.09.003.

Lin, H., Bennett, G. N., and San, K.-Y. (2005). "Metabolic engineering of aerobic succinate production systems in *Escherichia coli* to improve process productivity and achieve the maximum theoretical succinate yield". In: *Metab. Eng.* 7.2, pp. 116–127. DOI: 10.1016/j.ymben.2004.10.003.

Liu, J. K., O'Brien, E. J., Lerman, J. A., Zengler, K., Palsson, B. Ø., and Feist, A. M. (2014). "Reconstruction and modeling protein translocation and compartmentalization in *Escherichia coli* at the genome-scale". In: *BMC Syst. Biol.* 8.110. DOI: 10.1186/s12918-014-0110-6.

Liu, R., Bassalo, M. C., Zeitouna, R. I., and Gill, R. T. (2015). "Genome scale engineering techniques for metabolic engineering". In: *Metab. Eng.* Pp. 1–12. DOI: 10.1016/j.ymben.2015.09.013.

Lobel, L., Sigal, N., Borovok, I., Ruppin, E., and Herskovits, A. a. (2012). "Integrative genomic analysis identifies isoleucine and CodY as regulators of Listeria monocytogenes virulence". In: *PLoS Genet.* 8.9, e1002887. DOI: 10.1371/journal.pgen.1002887.

Londesborough, J., Penttilae, M., Richard, P., and Verho, R. (2003). "Fungal microorganism having an increased ability to carry out biotechnological process(es)". 2003038067 A1.

Lunzer, M., Miller, S. P., Felsheim, R., and Dean, A. M. (2005). "The biochemical architecture of an ancient adaptive landscape". In: *Science* 310.5747, pp. 499–501. DOI: 10.1126/science.1115649.

Ma, J., Gou, D., Liang, L., Liu, R., Chen, X., Zhang, C., Zhang, J., Chen, K., and Jiang, M. (2013). "Enhancement of succinate production by metabolically engineered *Escherichia coli* with co-expression of nicotinic acid phosphoribosyltransferase and pyruvate carboxylase". In: *Appl. Microbiol. Biotechnol.* DOI: 10.1007/s00253-013-4910-1.

Machado, D. and Herrgård, M. (2015). "Co-evolution of strain design methods based on flux balance and elementary mode analysis". In: *Metabolic Engineering Communications* 2, pp. 85–92. DOI: 10.1016/j.meteno.2015.04.001.

Mahadevan, R. and Schilling, C. H. (2003). "The effects of alternate optimal solutions in constraint-based genome-scale metabolic models". In: *Metab. Eng.* 5.4, pp. 264–276. DOI: 10.1016/j.ymben.2003.09.002.

Maklashina, E., Berthold, D. a., and Cecchini, G. (1998). "Anaerobic expression of *Escherichia coli* succinate dehydrogenase: functional replacement of fumarate

reductase in the respiratory chain during anaerobic growth". In: *J. Bacteriol.* 180, pp. 5989–5996.

Manzer, L. E., Waal, J. C. v. d., and Imhof, P. (2013). "The Industrial Playing Field for the Conversion of Biomass to Renewable Fuels and Chemicals". In: *Catalytic Process Development for Renewable Materials*. Ed. by J. C. v. d. Waal and P. Imhof. 1st ed. Weinheim, Germany: Wiley-VCH, pp. 1–24.

Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J., Guyer, M., and Green, E. D. (2014). "The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data". In: *J. Am. Med. Inform. Assoc.* 21, pp. 957–958. DOI: 10.1136/amiajnl-2014-002974.

Martínez, I., Zhu, J., Lin, H., Bennett, G. N., and San, K.-Y. (2008). "Replacing *Escherichia coli* NAD-dependent glyceraldehyde 3-phosphate dehydrogenase (GAPDH) with a NADP-dependent enzyme from *Clostridium acetobutylicum* facilitates NADPH dependent pathways". In: *Metab. Eng.* 10.6, pp. 352–359. DOI: 10.1016/j.ymben.2008.09.001.

Marx, A., Eikmanns, B. J., Sahm, H., Graaf, A. A. de, and Eggeling, L. (1999). "Response of the central metabolism in *Corynebacterium glutamicum* to the use of an NADH-dependent glutamate dehydrogenase". In: *Metab. Eng.* 1.1, pp. 35–48. DOI: 10.1006/mben.1998.0106.

McCloskey, D., Gangoiti, J. A., King, Z. A., Naviaux, R. K., Barshop, B. A., Palsson, B., and Feist, A. M. (2013). "A model-driven quantitative metabolomics analysis of aerobic and anaerobic metabolism in *E. coli* K-12 MG1655 that is biochemically and thermodynamically consistent". In: *Biotechnol. Bioeng.* 111.4, pp. 803–815. DOI: 10.1002/bit.25133.

McCloskey, D., Palsson, B. Ø., and Feist, A. M. (2013). "Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*". In: *Mol. Syst. Biol.* 9.1, p. 661. DOI: 10.1038/msb.2013.18.

McKenna, R. and Nielsen, D. R. (2011). "Styrene biosynthesis from glucose by engineered *E. coli*". In: *Metab. Eng.* 13.5, pp. 544–554. DOI: 10.1016/j.ymben.2011.06.005.

McMurry, J., Kohler, S., Balhoff, J., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., Foster, E., Gourdine, J.-P., Jacobsen, J., Keith, D., Laraway, B., Lewis, S., Xuan, J. N., Shefchek, K., Vasilevsky, N., Yuan, Z., Washington, N., Hochheiser, H., Mungall, C., Groza, T., Smedley, D., Robinson,

P., and Haendel, M. (2016). “Navigating the phenotype frontier: The Monarch Initiative”.

Mo, M. L., Palsson, B. O., and Herrgård, M. J. (2009). “Connecting extracellular metabolomic measurements to intracellular flux states in yeast”. In: *BMC Syst. Biol.* 3, p. 37. DOI: 10.1186/1752-0509-3-37.

Molenaar, D., Berlo, R. van, Ridder, D. de, and Teusink, B. (2009). “Shifts in growth strategies reflect tradeoffs in cellular economics”. In: *Mol. Syst. Biol.* 5.323, p. 323. DOI: 10.1038/msb.2009.82.

Monk, J., Nogales, J., and Palsson, B. O. (2014). “Optimizing genome-scale network reconstructions”. In: *Nat. Biotechnol.* 32.5, pp. 447–452. DOI: 10.1038/nbt.2870.

Murarka, A., Dharmadi, Y., Yazdani, S. S., and Gonzalez, R. (2008). “Fermentative utilization of glycerol by *Escherichia coli* and its implications for the production of fuels and chemicals”. In: *Appl. Environ. Microbiol.* 74.4, pp. 1124–1135. DOI: 10.1128/AEM.02192-07.

Nakahigashi, K., Toya, Y., Ishii, N., Soga, T., Hasegawa, M., Watanabe, H., Takai, Y., Honma, M., Mori, H., and Tomita, M. (2009). “Systematic phenome analysis of *Escherichia coli* multiple-knockout mutants reveals hidden reactions in central carbon metabolism”. In: *Mol. Syst. Biol.* 5.306, p. 306. DOI: 10.1038/msb.2009.65.

Nam, H., Lewis, N. E., Lerman, J. a., Lee, D.-H., Chang, R. L., Kim, D., and Palsson, B. O. (2012). “Network context and selection in the evolution to enzyme specificity”. In: *Science* 337.6098, pp. 1101–1104. DOI: 10.1126/science.1216861.

Nissen, T. L., Anderlund, M., Nielsen, J., Villadsen, J., and Kielland-Brandt, M. C. (2001). “Expression of a cytoplasmic transhydrogenase in Saccharomyces cerevisiae results in formation of 2-oxoglutarate due to depletion of the NADPH pool”. In: *Yeast* 18.1, pp. 19–32. DOI: 10.1002/1097-0061(200101)18:1<19::AID-YEA650>3.0.CO;2-5.

Nocon, J., Steiger, M. G., Pfeffer, M., Sohn, S. B., Kim, T. Y., Maurer, M., Rußmayer, H., Pflügl, S., Ask, M., Haberhauer-Troyer, C., Ortmayr, K., Hann, S., Koellensperger, G., Gasser, B., Lee, S. Y., and Mattanovich, D. (2014). “Model based engineering of Pichia pastoris central metabolism enhances recombinant protein production”. In: *Metab. Eng.* 24, pp. 129–138. DOI: 10.1016/j.ymben.2014.05.011.

O'Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R., and Palsson, B. Ø. (2013). "Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction". In: *Mol. Syst. Biol.* 9.1, p. 693. DOI: 10.1038/msb.2013.52.

O'Brien, E. J. and Palsson, B. O. (2015). "Computing the functional proteome: recent progress and future prospects for genome-scale models". In: *Curr. Opin. Biotechnol.* 34, pp. 125–134. DOI: 10.1016/j.copbio.2014.12.017.

Olavarría, K., Valdés, D., and Cabrera, R. (2012). "The cofactor preference of glucose-6-phosphate dehydrogenase from *Escherichia coli*–modeling the physiological production of reduced cofactors". In: *FEBS J.* 279.13, pp. 2296–2309. DOI: 10.1111/j.1742-4658.2012.08610.x.

Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M., and Palsson, B. Ø. (2011). "A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism—2011". In: *Mol. Syst. Biol.* 7.535, p. 535. DOI: 10.1038/msb.2011.65.

Orth, J. D. and Palsson, B. (2010). "Systematizing the generation of missing metabolic knowledge". In: *Biotechnol. Bioeng.* 107.3, pp. 403–412. DOI: 10.1002/bit.22844.

Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). "What is flux balance analysis?" In: *Nat. Biotechnol.* 28.3, pp. 245–248. DOI: 10.1038/nbt.1614.

Palsson, B. Ø. (2006). *Systems Biology: Properties of Reconstructed Networks*. Cambridge, UK: Cambridge University Press.

Papin, J. A. and Palsson, B. O. (2004). "The JAK-STAT signaling network in the human B-cell: an extreme signaling pathway analysis". In: *Biophys. J.* 87.1, pp. 37–46. DOI: 10.1529/biophysj.103.029884.

Patil, K. R., Rocha, I., Förster, J., and Nielsen, J. (2005). "Evolutionary programming as a platform for *in silico* metabolic engineering". In: *BMC Bioinformatics* 6, p. 308. DOI: 10.1186/1471-2105-6-308.

Piškur, J., Rozpedowska, E., Polakova, S., Merico, A., and Compagno, C. (2006). "How did Saccharomyces evolve to become a good brewer?" In: *Trends Genet.* 22.4, pp. 183–186. DOI: 10.1016/j.tig.2006.02.002.

Price, N. D., Reed, J. L., and Palsson, B. Ø. (2004). "Genome-scale models of microbial cells: evaluating the consequences of constraints". In: *Nat. Rev. Microbiol.* 2.11, pp. 886–897. DOI: 10.1038/nrmicro1023.

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., Murphy, M. R., O'Leary, N. A., Pujar, S., Rajput, B., Rangwala, S. H., Riddick, L. D., Shkeda, A., Sun, H., Tamez, P., Tully, R. E., Wallin, C., Webb, D., Weber, J., Wu, W., Dicuccio, M., Kitts, P., Maglott, D. R., Murphy, T. D., and Ostell, J. M. (2014). "RefSeq: An update on mammalian reference sequences". In: *Nucleic Acids Res.* 42, pp. D756–D763. DOI: 10.1093/nar/gkt1114.

Qi, W. W., Vannelli, T., Breinig, S., Ben-Bassat, A., Gatenby, A. A., Haynie, S. L., and Sariaslani, F. S. (2007). "Functional expression of prokaryotic and eukaryotic genes in *Escherichia coli* for conversion of glucose to p-hydroxystyrene". In: *Metab. Eng.* 9.3, pp. 268–276. DOI: 10.1016/j.ymben.2007.01.002.

Qian, Z.-G., Xia, X.-X., and Lee, S. Y. (2009). "Metabolic engineering of *Escherichia coli* for the production of putrescine: a four carbon diamine". In: *Biotechnol. Bioeng.* 104.4, pp. 651–662. DOI: 10.1002/bit.22502.

Ranganathan, S., Suthers, P. F., and Maranas, C. D. (2010). "OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions". In: *PLoS Comput. Biol.* 6.4, e1000744. DOI: 10.1371/journal.pcbi.1000744.

Ranganathan, S., Wei Tee, T., Chowdhury, A., Zomorrodi, A. R., Moon Yoon, J., Fu, Y., Shanks, J. V., and Maranas, C. D. (2012). "An integrated computational and experimental study for overproducing fatty acids in *Escherichia coli*". In: *Metab. Eng.* 14.6, pp. 687–704. DOI: 10.1016/j.ymben.2012.08.008.

Rathnasingh, C., Raj, S. M., Jo, J.-E., and Park, S. (2009). "Development and evaluation of efficient recombinant Escherichia coli strains for the production of 3-hydroxypropionic acid from glycerol". In: *Biotechnol. Bioeng.* 104.4, pp. 729–739. DOI: 10.1002/bit.22429.

Rathnasingh, C., Raj, S. M., Lee, Y., Catherine, C., Ashok, S., and Park, S. (2012). "Production of 3-hydroxypropionic acid via malonyl-CoA pathway using recombinant *Escherichia coli* strains". In: *J. Biotechnol.* 157.4, pp. 633–640. DOI: 10.1016/j.jbiotec.2011.06.008.

Reed, J. L., Famili, I., Thiele, I., and Palsson, B. O. (2006a). "Towards multidimensional genome annotation". In: *Nat. Rev. Genet.* 7.2, pp. 130–141. DOI: 10.1038/nrg1769.

Reed, J. L., Patel, T. R., Chen, K. H., Joyce, A. R., Applebee, M. K., Herring, C. D., Bui, O. T., Knight, E. M., Fong, S. S., and Palsson, B. O. (2006b). "Systems

approach to refining genome annotation". In: *Proc. Natl. Acad. Sci. U. S. A.* 103.46, pp. 17480–17484. DOI: 10.1073/pnas.0603364103.

Reed, J. L., Vo, T. D., Schilling, C. H., and Palsson, B. Ø. (2003). "An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR)". In: *Genome Biol.* 4.9, R54. DOI: 10.1186/gb-2003-4-9-r54.

Rodriguez, N., Thomas, A., Watanabe, L., Vazirabad, I. Y., Kofia, V., Gómez, H. F., Mittag, F., Matthes, J., Rudolph, J., Wrzodek, F., Netz, E., Diamantikos, A., Eichner, J., Keller, R., Wrzodek, C., Fröhlich, S., Lewis, N. E., Myers, C. J., Le Novère, N., Palsson, B. Ø., Hucka, M., and Dräger, A. (2015). "JSBML 1.0: providing a smorgasbord of options to encode systems biology models". In: *Bioinformatics*, Advance access. DOI: 10.1093/bioinformatics/btv341.

Rodríguez-Arnedo, A., Camacho, M., Llorca, F., and Bonete, M.-J. (2005). "Complete reversal of coenzyme specificity of isocitrate dehydrogenase from Haloferax volcanii". In: *Protein J.* 24.5, pp. 259–266. DOI: 10.1007/s10930-005-6746-8.

Rohn, H., Junker, A., Hartmann, A., Grafahrend-Belau, E., Treutler, H., Klapperstück, M., Czauderna, T., Klukas, C., and Schreiber, F. (2012). "VANTED v2: a framework for systems biology applications". In: *BMC Syst. Biol.* 6.3, p. 139. DOI: 10.1186/1752-0509-6-139.

Rose, P. W., Prlić, A., Bi, C., Bluhm, W. F., Christie, C. H., Dutta, S., Green, R. K., Goodsell, D. S., Westbrook, J. D., Woo, J., Young, J., Zardecki, C., Berman, H. M., Bourne, P. E., and Burley, S. K. (2015). "The RCSB Protein Data Bank: views of structural biology for basic and applied research and education". In: *Nucleic Acids Res.* 43, pp. D345–D356. DOI: 10.1093/nar/gks1200.

Russell, J. B. and Cook, G. M. (1995). "Energetics of bacterial growth: balance of anabolic and catabolic reactions". In: *Microbiol. Mol. Biol. Rev.* 59.1, pp. 48–62.

Sanchez, A. M., Andrews, J., Hussein, I., Bennett, G. N., and San, K.-Y. (2006). "Effect of overexpression of a soluble pyridine nucleotide transhydrogenase (UdhA) on the production of poly(3-hydroxybutyrate) in *Escherichia coli*". In: *Biotechnol. Prog.* 22.2, pp. 420–425. DOI: 10.1021/bp050375u.

Sánchez, A. M., Bennett, G. N., and San, K.-Y. (2005a). "Efficient succinic acid production from glucose through overexpression of pyruvate carboxylase in an *Escherichia coli* alcohol dehydrogenase and lactate dehydrogenase mutant". In: *Biotechnol. Prog.* 21.2, pp. 358–365. DOI: 10.1021/bp049676e.

Sánchez, A. M., Bennett, G. N., and San, K.-Y. (2005b). "Novel pathway engineering design of the anaerobic central metabolic pathway in *Escherichia coli* to increase succinate yield and productivity". In: *Metab. Eng.* 7.3, pp. 229–239. DOI: 10.1016/j.ymben.2005.03.001.

Sauer, U., Canonaco, F., Heri, S., Perrenoud, A., and Fischer, E. (2004). "The soluble and membrane-bound transhydrogenases UdhA and PntAB have divergent functions in NADPH metabolism of *Escherichia coli*". In: *J. Biol. Chem.* 279.8, pp. 6613–6619. DOI: 10.1074/jbc.M311657200.

Schellenberger, J., Park, J. O., Conrad, T. M., and Palsson, B. Ø. (2010). "BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions". In: *BMC Bioinformatics*. DOI: 10.1186/1471-2105-11-213.

Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., Bordbar, A., Lewis, N. E., Rahmanian, S., Kang, J., Hyduke, D. R., and Palsson, B. Ø. (2011). "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0". In: *Nat. Protoc.* 6.9, pp. 1290–1307. DOI: 10.1038/nprot.2011.308.

Schuetz, R., Kuepfer, L., and Sauer, U. (2007). "Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*". In: *Mol. Syst. Biol.* 3.119, p. 119. DOI: 10.1038/msb4100162.

Segre, D., Vitkup, D., and Church, G. M. (2002). "Analysis of optimality in natural and perturbed metabolic networks". In: *Proceedings of the National Academy of Sciences* 99.23, pp. 15112–15117.

Seo, S. W., Kim, D., Latif, H., O'Brien, E. J., Szubin, R., and Palsson, B. O. (2014). "Deciphering Fur transcriptional regulatory network highlights its complex role beyond iron metabolism in *Escherichia coli*". In: *Nat. Commun.* 5, p. 4910. DOI: 10.1038/ncomms5910.

Shen, C. R., Lan, E. I., Dekishima, Y., Baez, A., Cho, K. M., and Liao, J. C. (2011). "Driving forces enable high-titer anaerobic 1-butanol synthesis in *Escherichia coli*". In: *Appl. Environ. Microbiol.* 77.9, pp. 2905–2915. DOI: 10.1128/AEM.03034-10.

Shen, C. R. and Liao, J. C. (2013). "Synergy as design principle for metabolic engineering of 1-propanol production in *Escherichia coli*". In: *Metab. Eng.* 17, pp. 12–22. DOI: 10.1016/j.ymben.2013.01.008.

Shin, J. H., Kim, H. U., Kim, D. I., and Lee, S. Y. (2013). "Production of bulk chemicals via novel metabolic pathways in microorganisms". In: *Biotechnol. Adv.* 31.6, pp. 925–935. DOI: 10.1016/j.biotechadv.2012.12.008.

Singh, A., Cher Soh, K., Hatzimanikatis, V., and Gill, R. T. (2011). "Manipulating redox and ATP balancing for improved production of succinate in *E. coli*". In: *Metab. Eng.* 13.1, pp. 76–81. DOI: 10.1016/j.ymben.2010.10.006.

Skinner, M. E., Uzilov, A. V., Stein, L. D., Mungall, C. J., and Holmes, I. H. (2009). "JBrowse : A next-generation genome browser". In: *Genome Res.* 19, pp. 1630–1638. DOI: 10.1101/gr.094607.109.

Smanski, M. J., Bhatia, S., Zhao, D., Park, Y., B A Woodruff, L., Giannoukos, G., Ciulla, D., Busby, M., Calderon, J., Nicol, R., Gordon, D. B., Densmore, D., and Voigt, C. A. (2014). "Functional optimization of gene clusters by combinatorial design and assembly". In: *Nat. Biotechnol.* 32.12, pp. 1241–1249. DOI: 10.1038/nbt.3063.

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., and Ideker, T. (2011). "Cytoscape 2.8: New features for data integration and network visualization". In: *Bioinformatics* 27.3, pp. 431–432. DOI: 10.1093/bioinformatics/btq675.

Stallman, R. M. (1981). *EMACS the extensible, customizable self-documenting display editor*. DOI: 10.1145/872730.806466.

Stols, L., Kulkarni, G., Harris, B. G., and Donnelly, M. I. (1997). "Expression of Ascaris suum malic enzyme in a mutant Escherichia coli allows production of succinic acid from glucose". In: *Appl. Biochem. Biotechnol.* 63-65.4, pp. 153–158.

Stols, L. and Donnelly, M. I. (1997). "Production of succinic acid through overexpression of NAD^+-dependent malic enzyme in an *Escherichia coli* mutant". In: *Appl. Environ. Microbiol.* 63.7, pp. 2695–2701.

Szappanos, B., Kovács, K., Szamecz, B., Honti, F., Costanzo, M., Baryshnikova, A., Gelius-Dietrich, G., Lercher, M. J., Jelasity, M., Myers, C. L., Andrews, B. J., Boone, C., Oliver, S. G., Pál, C., and Papp, B. (2011). "An integrated approach to characterize genetic interaction networks in yeast metabolism". In: *Nat. Genet.* 43.7, pp. 656–662. DOI: 10.1038/ng.846.

Tang, X., Tan, Y., Zhu, H., Zhao, K., and Shen, W. (2009). "Microbial conversion of glycerol to 1,3-propanediol by an engineered strain of *Escherichia coli*". In: *Appl. Environ. Microbiol.* 75.6, pp. 1628–1634. DOI: 10.1128/AEM.02376-08.

Tepper, N. and Shlomi, T. (2010). "Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways". In: *Bioinformatics* 26.4, pp. 536–543. DOI: 10.1093/bioinformatics/btp704.

Thiele, I., Fleming, R. M. T., Que, R., Bordbar, A., Diep, D., and Palsson, B. O. (2012). "Multiscale modeling of metabolism and macromolecular synthesis in *E. coli* and its application to the evolution of codon usage". In: *PLoS One* 7.9, e45635. DOI: 10.1371/journal.pone.0045635.

Thiele, I., Jamshidi, N., Fleming, R. M. T., and Palsson, B. Ø. (2009). "Genome-scale reconstruction of *Escherichia coli*'s transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization". In: *PLoS Comput. Biol.* 5.3, e1000312. DOI: 10.1371/journal.pcbi.1000312.

Thiele, I. and Palsson, B. Ø. (2010). "A protocol for generating a high-quality genome-scale metabolic reconstruction". In: *Nat. Protoc.* 5.1, pp. 93–121. DOI: 10.1038/nprot.2009.203.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks". In: *Nat. Protoc.* 7, pp. 562–578. DOI: 10.1038/nprot.2012.016.

Trinh, C. T. (2012). "Elucidating and reprogramming *Escherichia coli* metabolisms for obligate anaerobic n-butanol and isobutanol production". In: *Appl. Microbiol. Biotechnol.* 95.4, pp. 1083–1094. DOI: 10.1007/s00253-012-4197-7.

Trinh, C. T., Li, J., Blanch, H. W., and Clark, D. S. (2011). "Redesigning *Escherichia coli* metabolism for anaerobic production of isobutanol". In: *Appl. Environ. Microbiol.* 77.14, pp. 4894–4904. DOI: 10.1128/AEM.00382-11.

Tseng, H.-C., Harwell, C. L., Martin, C. H., and Prather, K. L. J. (2010). "Biosynthesis of chiral 3-hydroxyvalerate from single propionate-unrelated carbon sources in metabolically engineered *E. coli*". In: *Microb. Cell Fact.* 9, p. 96. DOI: 10.1186/1475-2859-9-96.

Tseng, H.-C., Martin, C. H., Nielsen, D. R., and Prather, K. L. J. (2009). "Metabolic engineering of *Escherichia coli* for enhanced production of (R)- and (S)-3-hydroxybutyrate". In: *Appl. Environ. Microbiol.* 75.10, pp. 3137–3145. DOI: 10.1128/AEM.02667-08.

Ui, S., Takusagawa, Y., Sato, T., Ohtsuki, T., Mimura, A., Ohkuma, M., and Kudo, T. (2004). "Production of L-2,3-butanediol by a new pathway constructed in *Escherichia coli*". In: *Lett. Appl. Microbiol.* 39.6, pp. 533–537. DOI: 10.1111/j.1472-765X.2004.01622.x.

Varma, A., Boesch, B. W., and Palsson, B. O. (1993). "Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates". In: *Appl. Environ. Microbiol.* 59.8, pp. 2465–2473.

Varma, A. and Palsson, B. Ø. (1994). "Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use". In: *Nat. Biotechnol.* 12.October, pp. 994–998.

Vemuri, G. N., Eiteman, M. A., and Altman, E. (2002). "Effects of Growth Mode and Pyruvate Carboxylase on Succinic Acid Production by Metabolically Engineered Strains of *Escherichia coli*". In: *Appl. Environ. Microbiol.* 68.4, pp. 1715–1727. DOI: 10.1128/AEM.68.4.1715.

Verho, R., Londesborough, J., Penttilä, M., and Richard, P. (2003). "Engineering Redox Cofactor Regeneration for Improved Pentose Fermentation in *Saccharomyces cerevisiae*". In: *Appl. Environ. Microbiol.* 69.10, p. 5892. DOI: 10.1128/AEM.69.10.5892.

Villadsen, J., Nielsen, J., and Lidén, G. (2011). "Chemicals from Metabolic Pathways". In: *Bioreaction Engineering Principles*. Boston, MA: Springer US. Chap. 2, pp. 7–62. DOI: 10.1007/978-1-4419-9688-6.

Wang, B., Wang, P., Zheng, E., Chen, X., Zhao, H., Song, P., Su, R., Li, X., and Zhu, G. (2011). "Biochemical properties and physiological roles of NADP-dependent malic enzyme in *Escherichia coli*". In: *J. Microbiol.* 49.5, pp. 797–802. DOI: 10.1007/s12275-011-0487-5.

Winkler, J. D., Halweg-Edwards, A. L., and Gill, R. T. (2015). "The LASER database: Formalizing design rules for metabolic engineering". In: *Metabolic Engineering Communications* 2, pp. 30–38. DOI: 10.1016/j.meteno.2015.06.003.

Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., Bouatra, S., Sinelnikov, I., Arndt, D., Xia, J., Liu, P., Yallou, F., Bjorndahl, T., Perez-Pineiro, R., Eisner, R., Allen, F., Neveu, V., Greiner, R., and Scalbert, A. (2013). "HMDB 3.0—The Human Metabolome Database in 2013". In: *Nucleic Acids Res.* 41, pp. D801–D807. DOI: 10.1093/nar/gks1065.

Xu, P., Ranganathan, S., Fowler, Z. L., Maranas, C. D., and Koffas, M. a. G. (2011). "Genome-scale metabolic network modeling results in minimal interventions that cooperatively force carbon flux towards malonyl-CoA". In: *Metab. Eng.* 13.5, pp. 578–587. DOI: 10.1016/j.ymben.2011.06.008.

Yamada, T., Letunic, I., Okuda, S., Kanehisa, M., and Bork, P. (2011). "IPath2.0: Interactive pathway explorer". In: *Nucleic Acids Res.* 39.May, pp. 412–415. DOI: 10.1093/nar/gkr313.

Yan, Y., Lee, C.-C., and Liao, J. C. (2009). "Enantioselective synthesis of pure (R,R)-2,3-butanediol in *Escherichia coli* with stereospecific secondary alcohol dehydrogenases". In: *Org. Biomol. Chem.* 7.19, pp. 3914–3917. DOI: 10.1039/b913501d.

Yaoi, T., Miyazaki, K., Oshima, T., Komukai, Y., and Go, M. (1996). "Conversion of the Coenzyme Specificity of Isocitrate Dehydrogenase by Module Replacement". In: *J. Biochem.* 1018, pp. 1014–1018.

Yim, H., Haselbeck, R., Niu, W., Pujol-Baxley, C., Burgard, A., Boldt, J., Khandurina, J., Trawick, J. D., Osterhout, R. E., Stephen, R., Estadilla, J., Teisan, S., Schreyer, H. B., Andrae, S., Yang, T. H., Lee, S. Y., Burk, M. J., and Van Dien, S. (2011). "Metabolic engineering of *Escherichia coli* for direct production of 1,4-butanediol". In: *Nat. Chem. Biol.* 7.7, pp. 445–452. DOI: 10.1038/nchembio.580.

Zhang, F., Rodriguez, S., and Keasling, J. D. (2011). "Metabolic engineering of microbial pathways for advanced biofuels production". In: *Curr. Opin. Biotechnol.* 22.6, pp. 775–783. DOI: 10.1016/j.copbio.2011.04.024.

Zhang, X., Jantama, K., Moore, J. C., Shanmugam, K. T., and Ingram, L. O. (2007). "Production of L-alanine by metabolically engineered Escherichia coli". In: *Appl. Microbiol. Biotechnol.* 77.2, pp. 355–366. DOI: 10.1007/s00253-007-1170-y.

Zhang, X., Shanmugam, K. T., and Ingram, L. O. (2010). "Fermentation of glycerol to succinate by metabolically engineered strains of *Escherichia coli*". In: *Appl. Environ. Microbiol.* 76.8, pp. 2397–2401. DOI: 10.1128/AEM.02902-09.

Zhang, Y., Thiele, I., Weekes, D., Li, Z., Jaroszewski, L., Ginalski, K., Deacon, A. M., Wooley, J., Lesley, S. A., Wilson, I. A., Palsson, B. Ø., Osterman, A. L., and Godzik, A. (2009). "Three-Dimensional Structural View of the Central Metabolic Network of Thermotoga maritima". In: *Science* September, pp. 1544–1549.

Zhao, Q., Stettner, A. I., Reznik, E., Paschalidis, I. C., and Segrè, D. (2016). "Mapping the landscape of metabolic goals of a cell". In: *Genome Biol.* 17.1, p. 109. DOI: 10.1186/s13059-016-0968-2.

Zhou, S., Shanmugam, K. T., and Ingram, L. O. (2003). "Functional Replacement of the *Escherichia coli* d-(–)-lactate dehydrogenase gene (ldhA) with the l-(+)-lactate dehydrogenase gene (ldhL) from Pediococcus acidilactici". In: *Appl. Environ. Microbiol.* 69.4, p. 2237. DOI: 10.1128/AEM.69.4.2237.

Zhu, G., Golding, G. B., and Dean, A. M. (2005). "The selective cause of an ancient adaptation". In: *Science* 307.5713, pp. 1279–1282. DOI: 10.1126/science.1106974.

Zhuang, K., Izallalen, M., Mouser, P., Richter, H., Risso, C., Mahadevan, R., and Lovley, D. R. (2011). "Genome-scale dynamic modeling of the competition between Rhodoferax and Geobacter in anoxic subsurface environments". In: *ISME J.* 5.2, pp. 305–316. DOI: 10.1038/ismej.2010.117.