UNIVERSITY OF CALIFORNIA, SAN DIEGO

Causation in Biology: A Biomolecular Systems View

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

in

Bioengineering

by

Nathan Enoch Lewis

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Jeff Hasty, Co-chair
Professor Steven P. Briggs
Professor Trey Ideker
Professor Milton H. Saier, Jr.

2012

The dissertation of Nathan Enoch Lewis is approved, and it is acceptable in quality for publication on microfilm:

____

____

____

____
Co-chair

____
Chair

University of California, San Diego

2012

Dedication

To Maria

For your sacrifices, support, and enduring love.

You truly are my best friend.

To Anabelle

For the many hand-squeezes, hugs, and verses of "Butterfly Kisses".

May your love and concern for others flourish throughout your life.

To Livya

For your inquisitive smile. May your curiosity and goodness fuel your life.

Epigraph

We make a living by what we get, we make a life by what we give.

Winston Churchill

Table of Contents

List of Figures

Acknowledgements

I owe countless people greatly for their capacitation and support, which has culminated in the drafting of this thesis. This work could not have been done without their sacrifices, both big and small, which have touched my life.

First, I am indebted to millions who, for centuries, have tirelessly laid the scientific groundwork upon which this thesis stands. In particular, I thank Professor Bernhard Palsson for fostering a collaborative and innovative team of researchers. However, I particularly thank him for his care in mentoring and guiding me through my scientific development. I will be forever grateful for the training I have received under his hospice, with respect to framing meaningful research questions, tailoring paper- and grant-writing efforts for the appropriate audiences, and for guiding me in my career development. His scientific vision has been particularly enlightening and will influence my lifelong scientific career.

I also want to thank the many researchers with whom I have worked over the past several years. At UCSD, I've enjoyed the company and collaboration of many people. In particular, Joshua Lerman, Harish Nagarajan, Jan Schellenberger, Hojung Nam, Aarash Bordbar, and Roger Chang have all provided long hours of deep discussion and fruitful collaboration over the last few years. I also thank Christian Barrett, Markus Herrgard, Iman Famili, Daniel R. Hyduke, Pep Charusanti, and Neema Jamshidi, who have all been senior mentors and role models to me. In addition, I thoroughly enjoyed the numerous researchers with whom I have extensively collaborated, including Vasily Portnoy, Tom Conrad, DaeHee Lee, Steve Federowicz, Teddy O'Brien, Hooman Hefzi, Donghyuk Kim, Tenai Eguen, Akshay Chaudhari, M. Paul Andersen, and the various undergraduates that have worked with me. I also appreciate the frequent deep conversations shared with Daniel Zielinski, and insightful conversations shared with Karsten Zengler, Adam Feist, Monica Mo, Ines Thiele, Ronan Fleming, Eric Knight, Young Seoub Park, Tzu-Wen Huang, Kenyon Applebee, Jay SJ Hong, Yanming Gong, Nikolaus Sonnenschein, Juan Nogales Enrique, Brian Schmidt, Addiel U. de Alba Solis, Ali Ebrahim, Mallory

Embree, Ryan LaCroix, Haythem Latif, Jonathan Monk, Jeff Orth, Alex Thomas, BK Cho, Sorena Rahmanian, Douglas Taylor, Osbaldo Resendis-Antonio, Natalie Hurlen, and Jennifer Reed. Lastly, without the support of Marc Abrams and Kathy Andrews, much of my work would have been, at best, delayed.

I also want to thank those with whom I collaborated in Israel during a year-long visit to Israel in the middle of my doctoral work. In particular, I thank Dr. Eran Segal and his lab members for their kindness and support, and Dr. Ron Milo and his students Arren Bar-Even and Elad Noor for including me in their work while in Israel. Lastly I thank Dr. Eytan Ruppin, Dr. Tomer Shlomi, Yedael Waldman, Tamir Tuller, and Shira Mintz for their support and collaboration.

I also thank the members of my thesis committee collectively for their candor and for requiring my best efforts, and for being willing to take their personal time to mentor me. In particular I want to thank Professor Steve Briggs for his continued mentorship and guidance with respect to scientific communication and thoughtful planning of research. I also want to thank Professor Jeff Hasty for the opportunities he provided me to teach with him.

I also thank the other advisors that prepared me for my doctoral research, including Qiang Tian, Li Ma, and Leroy Hood at the Institute for Systems Biology, Emanuel Petricoin III and Marc Lowenthal at the FDA and NIH, and most importantly Craig Thulin who, at Brigham Young University immediately asked me to join his lab when I told him I had switched from the nursing major to biochemistry. His personal mentorship opened my eyes to how much I love scientific research, and prepared me for a valuable scientific career.

Lastly, none of this would have been possible without the loving support of my family. I thank my parents, Mark and Fae Lewis, who instilled in me a love for learning and a desire to work hard and serve others. I also am grateful for the daily loving smiles from my two little sweethearts, Anabelle and Livya. Their zest for life has always picked me up, no matter how my research was or was not progressing. Most importantly, I am eternally grateful for by best friend, therapist, and wife, Maria P. Lewis since her daily love and support got me to this point. I could not have finished so well without her, as she truly completes me.

Chapter 1 is a modified version of material in Lewis, N.E., Nagarajan, H., and Palsson, B.Ø. Constraining the metabolic genotype-phenotype relationship using a phylogeny of *in silico* methods. *Nature Reviews Microbiology* (2012). I was the primary author, while the co-authors participated in and supervised the drafting of this review.

Chapter 2, in part, is a reprint of the material as it appears in 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., Palsson, B.Ø. Large-scale *in silico* modeling of metabolic interactions between cell types in the human brain. *Nature Biotechnology*, 28:1279–1285 (2010). I was the primary author, while the co-authors participated in the research that served as the basis for this study.

Chapter 3 contains some material from Lewis, N.E., Chang, R.L., Kim, D., Hefzi, H.H., Palsson, B.Ø. Prokaryotes use enzyme post-translational modification to globally regulate metabolism. *In preparation*. I was the primary author, while the co-authors provided support in the research that served as the basis for this study.

Chapter 4, in part, is a reprint of the material as it appears in Nam, H.J., Lewis, N.E., Lerman, J.A., Lee, D.H., Chang, R.L., Kim, D., Palsson, B.Ø. Network context and selection in the evolution to enzyme specificity. *Submitted*. I was a joint-primary author and the corresponding author, while the remaining co-authors provided support in the research that served as the basis for this study.

Chapter 5, in part, is a reprint of the material as it appears in 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., Palsson, B.Ø. Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.*, 6:390 (2010). I was the primary author, while the co-authors provided support in the research that served as the basis for this study.

Chapter 6, in part, is a reprint of the material as it appears in Lewis, N.E., Lee, D.H., Rutledge, A., Conrad, T.M., Kim, D., Adkins, J.A., Smith, R.D., Palsson, B.Ø. *E. coli* learns to grow optimally on a non-native carbon substrate through laboratory evolution. *Under revision*. I was a joint-primary author, while the co-authors provided support in the research that served as the basis for this study.

Vita

2006 B.S., Biochemistry, Brigham Young University

2008 M.S., Bioengineering, University of California, San Diego

2012 Ph.D., Bioengineering, University of California, San Diego

Publications

25. **Lewis, N.E.**, et al. The genomic basis for the phylogeny of Chinese hamster ovarian cell lines. *In preparation.*
24. Baycin, D., Zhang, H., **Lewis, N.E.**, Nagarajan, H., Palsson, B.Ø., Betenbaugh, M.J. Large-Scale Proteomic and Glycoproteomic Analysis of Chinese Hamster Ovary Cells. *In preparation.*
23. **Lewis, N.E.**, Chang, R.L., Kim, D., Hefzi, H.H., Palsson, B.Ø. Prokaryotes employ enzyme post-translational modification to globally regulate metabolism. *In preparation.*
22. Bordbar, A., Schellenberger, J., **Lewis, N.E.**, Nagarajan, H., Palsson, B.Ø. A biochemically meaningful coordinate system for biological networks. *In preparation.*
21. Noor, E., **Lewis, N.E.**, Milo, R. A proof for loop-law constraints in stoichiometric metabolic networks. *Under revision.*
20. Nam, H.J.*, **Lewis, N.E.*‡**, Lerman, J.A., Lee, D.H., Chang, R.L., Kim, D., Palsson, B.Ø. Network context provides a selective pressure in the evolution of enzyme promiscuity. *Under review at Science.*
19. Hefzi, H.H., Palsson, B.Ø., **Lewis, N.E.‡**. Reconstruction of genome-scale metabolic networks. In Handbook of Systems Biology. *Submitted.*
18. Hyduke, D.R., **Lewis, N.E.**, Palsson, B.Ø. Genome-scale network reconstructions as contextual frameworks for integrated analysis of omics data. *Under revision for Nature Reviews Genetics.*
17. 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.A., Palsson, B.Ø. Quantitative genome-scale simulation of molecular biology and metabolism. *Submitted.*
16. **Lewis, N.E.***, Lee, D.H.*, Rutledge, A., Conrad, T.M., Kim, D., Adkins, J.A., Smith, R.D., Palsson, B.Ø. E. coli learns to grow optimally on a non-native carbon substrate through laboratory evolution. *Under revision.*
15. **Lewis, N.E.**, Nagarajan, H., Palsson, B.Ø. Constraining the metabolic genotype-phenotype relationship using a phylogeny of *in silico* methods. *Nature Reviews Microbiology.* In press.
14. Xu, X.*, Nagarajan, H.*, **Lewis, N.E.***, et al. The Genomic Sequence of the Chinese Hamster Ovary (CHO) K1 cell line. *Nature Biotechnology*, 29:735-41 (2011).
13. Schellenberger, J., Que, R., Fleming, R.T., Thiele, I., Orth, J., Feist, A.M., Zielinski , D.C., Bordbar, A., **Lewis, N.E.**, Rahmanian, S., Kang, J., Hyduke, D., Palsson, B.Ø. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nature Protocols*, 6:1290-307 (2011).
12. Nam, H.J.*, Conrad, T.M., **Lewis, N.E.*‡**. The role of cellular objectives and selective pressures in metabolic pathway evolution. *Current Opinions in Biotechnology*, 22:595-600 (2011).
11. Conrad, T.M., **Lewis, N.E.**, Palsson, B.Ø. Microbial Laboratory Evolution in the Era of Genome-Scale Science. *Molecular Systems Biology*, 7:509 (2011).
10. Schellenberger, J., **Lewis, N.E.**, Palsson, B.Ø. Elimination of thermodynamically infeasible loops in steady state metabolic models. *Biophysical Journal*, 100:544-53 (2011).
9. **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., Palsson, B.Ø. Large-scale *in silico* modeling of metabolic interactions between cell types in the human brain. *Nature Biotechnology*, 28:1279–1285 (2010).

8. Conrad, T.M., Frazier, M., Joyce, A.R., Cho, B. K., Knight, E. M., **Lewis, N.E.**, Landick, R, Palsson, B.Ø. RNA polymerase mutants found through adaptive evolution re-program *Escherichia coli* K-12 MG1655 for optimal growth in minimal media. *Proc. Nat. Acad. Sci. USA*, 107:20500-5 (2010).
7. Bordbar, A., **Lewis, N.E.**, Schellenberger, J., Palsson, B.Ø., Jamshidi, N. Insight into human alveolar macrophage and *M. tuberculosis* interactions via metabolic reconstructions. *Mol. Syst. Biol.*, 6:422 (2010).
6. Portnoy, V.A., Scott, D.A., **Lewis, N.E.**, Tarasova, Y., Osterman, A.L., Palsson, B.Ø. Deletion of genes encoding cytochrome oxidases and quinol monooxygenase blocks the aerobic-anaerobic shift in *Escherichia coli* K-12 MG1655. *Appl. Environ. Microbiol.*, 76:6529-40 (2010).
5. **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., Palsson, B.Ø. Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.*, 6:390 (2010).
4. Bar-Even, A., Noor, E., **Lewis, N.E.**, Milo, R. Design and analysis of synthetic carbon fixation pathways. *Proc. Natl. Acad. Sci. USA.,* 107:8889-8894 (2010).
3. **Lewis, N.E.**, Cho, B.K., Knight, E.M. Palsson, B. Ø. Gene expression profiling and the use of genome-scale in silico models of *Escherichia coli* for analysis: providing context for content. *J. Bacteriol.*, 191:3437-44 (2009).
2. **Lewis, N.E.**, Jamshidi, N., Thiele, I. & Palsson, B.Ø. Metabolic systems biology: a constraint-based approach. In Encyclopedia of Complexity and Systems Science 5535 (Springer, New York, 2009).
1. Merrell, K., Southwick, K., Graves, S.W., Esplin, M.S., **Lewis, N.E.**, Thulin, C.D. Analysis of low-abundance, low-molecular-weight serum proteins using mass spectrometry. *J Biomol Tech.,* 15:238-48 (2004).

* Authors contributed equally, ‡ corresponding author

ABSTRACT OF THE DISSERTATION

Causation in Biology: A Biomolecular Systems View

by

Nathan Enoch Lewis

Doctor of Philosophy in Bioengineering

University of California, San Diego, 2012

Professor Bernhard Ø. Palsson, Chair
Professor Jeff Hasty, Co-chair

Fundamental physical phenomena are studied with a "cause and effect" approach. This enables understanding and prediction by employing mathematically formulated physical laws. Such approaches are less successful in biological systems, since they are subject to dual causation. That is, both physicochemical laws and evolving genetic constraints govern organisms. Biological systems respond immediately to stimuli (proximal causation) against a constant genetic background; however, these responses depend upon evolving genetic programs. Alterations in genetic programs are manifestations of distal causation, representing changes induced by genetic drift and natural selection. Constraint-based reconstruction and analysis is an emerging modeling approach that can account for both physicochemical constraints in biological systems and some evolutionary selective pressures. Here, constraint-based modeling is deployed to integrate disparate data types with genome-scale metabolic models to gain insight into mechanisms in proximal and distal causation, and conceptual advances are presented with respect to how these data are interpreted using constraint-based models. Specifically, these advances are used to suggest mechanisms determining proximal responses with respect to disease progression in human brain metabolism and the regulation of prokaryotic metabolism in dynamic environments. In addition, methods are presented that use genome-scale models of metabolism to analyze various data types to identify determinants of distal causation. Specifically, these methods are deployed to show that the evolution of enzyme specificity is guided by network context and the need to produce biomass. Moreover, these pressures further tune expression levels of metabolic pathways in laboratory evolved bacteria. Thus, through network reconstruction and data integration, vast amounts of data can be queried and provide detailed insight into proximal and distal causation in complex biological networks.

Chapter 1: Constraint-based modeling as a tool in predicting dual causality in biology

"Causality in biology is a far cry from causality in classical mechanics"

- Ernst Mayr[1]

Fundamental physical phenomena are studied with a "cause and effect" approach that enables understanding and prediction by employing mathematically formulated "physical laws" (Figure 1.1.a). Biological systems, however, are subject to dual causation, since both physicochemical laws and genetic constraints govern their functions. Biological systems respond immediately to stimuli (proximal causation) against a constant genetic background (Figure 1.1.b); however, these responses depend on evolving genetic programs (Figure 1.1.c). The alterations in genetic programs are manifestations of distal causation, representing changes induced by genetic drift and natural selection. Within the constraints of physical law, evolution alters the proximal responses in an unpredictable manner. Therefore, as Ernst Mayr stated 50 years ago, "[for] almost any biological phenomenon . . . there is always a proximate set of causes and [a distal] set of causes; both have to be explained and interpreted for a complete understanding of the given phenomenon."[1]

These are significant considerations since the genotype-phenotype relationship is the most fundamental relationship in biology. For decades this relationship has been subject to mostly argument, reasoning and qualitative analysis. However, our ability to fundamentally understand the genotype-phenotype relationship began changing in the mid-1990s, upon completion of first prokaryotic genome sequencing projects. Full genome sequences provide comprehensive, albeit not yet complete, information about the genetic elements that create the form and function of an organism. The comprehensive understanding for some cellular processes, such as metabolism, has resulted in structured knowledge-bases that can be mathematically represented[2-4]. This mathematical representation enables the computation of phenotypic states[5-8] based on genetic and environmental parameters. Remarkably, this provides a mechanistic representation of the microbial metabolic genotype-phenotype relationship.

Constraint-based models of genome-scale metabolic networks capture the inherent dual causation in the genotype-phenotype relationship by simultaneously accounting for constraints from physicochemical laws and genetics. The realization that these quantitative genotype-phenotype relationships could be constructed from a genome has driven the emergence of a whole new area of research that began after the first full genome sequences became available in the mid 1990s. More recently, the flood of increasingly rich high-throughput data has accelerated the evolution of constraint-based reconstruction and analysis (COBRA) methods from a set of basic tools for metabolic network analysis into a powerful analytical framework that is now widely used. Here we describe: 1) the basic features of the COBRA framework, 2) the 'phylogeny' of the evolving groups of COBRA methods, and 3) the COBRA 'ecology,' i.e., how various COBRA methods complement each other to answer larger questions in biology.

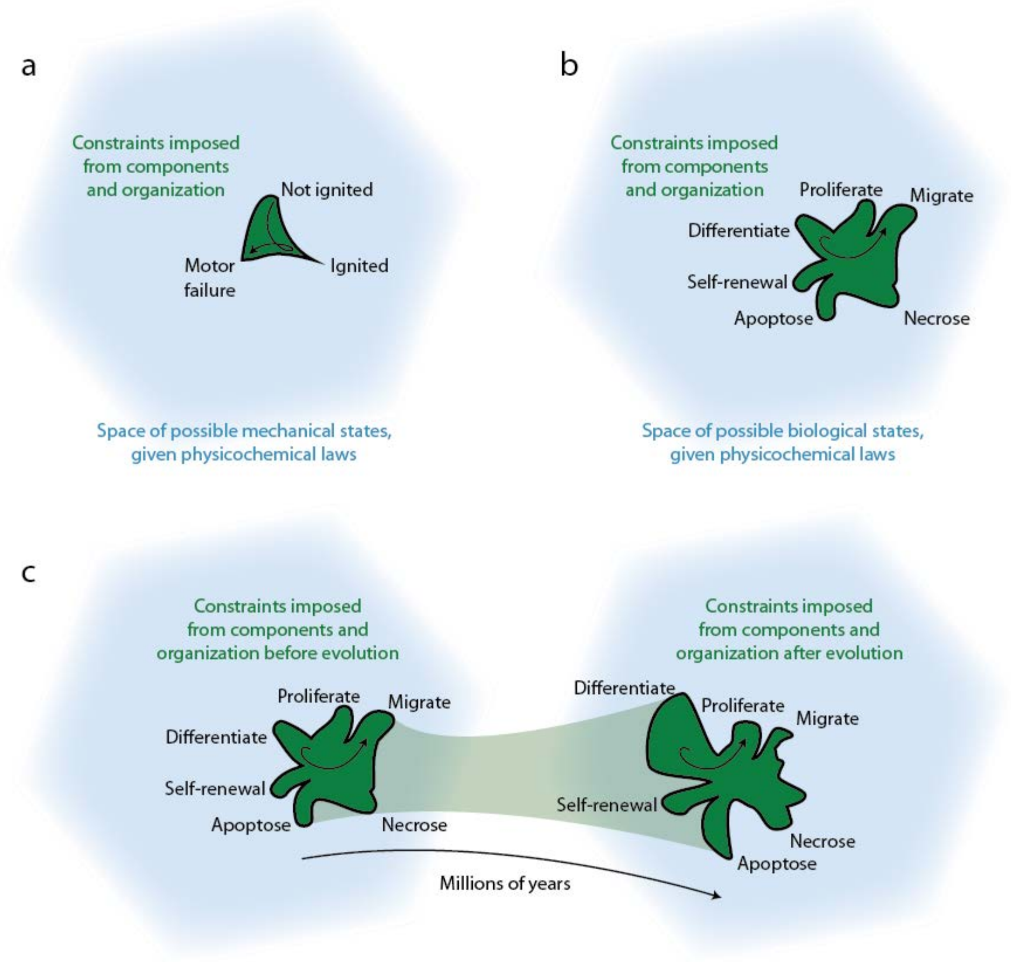

a
Constraints imposed from components and organization
Not ignited
Motor failure
Ignited
Space of possible mechanical states, given physicochemical laws
b
Constraints imposed from components and organization
Proliferate
Migrate
Differentiate
Self-renewal
Apoptose
Necrose
Space of possible biological states, given physicochemical laws
c
Constraints imposed from components and organization before evolution
Proliferate
Migrate
Differentiate
Self-renewal
Apoptose
Necrose
Constraints imposed from components and organization after evolution
Differentiate
Proliferate
Migrate
Self-renewal
Necrose
Apoptose
Millions of years

Figure 1.1. Proximal and distal causation. (a) A physical system can exist in any state, subject to physicochemical laws (blue). However, the space of allowed physical states is constrained by the selection of materials used to build the physical system, how these materials are connected, and how the system interacts with processes outside of the system boundaries (green). For example, a solid rocket motor can exist in a few states, and the motor composition and connections define these states. However, a failure in one component in the complex system can move the actual system from an ignited state to a failed state. Physical laws can be employed to predict how one might transition the rocket from a non-ignited state to the ignited and functioning state. These predictions are used to design the rocket. In like manner, an understanding of physical laws and the system can help identify what caused a rocket motor to fail. The processes that cause the motor to change states are the determinants of proximal causation. (b) A cell also has constraints on phenotypes it can display, imposed by its components and how they are organized. Through detailed study of these components and their interactions, we hope to understand proximal causation in cells, i.e., the physiological and environmental factors that cause a cell to change its phenotype. Despite the complexity, progress continues in our understanding of proximal causation in biology. (c) Understanding causation in biological systems is further complicated by evolution. Distal causation addresses the factors driving phenotype changes through evolution, and the mechanisms are often poorly understood.

Constraint-based modeling defined and its relation to causation

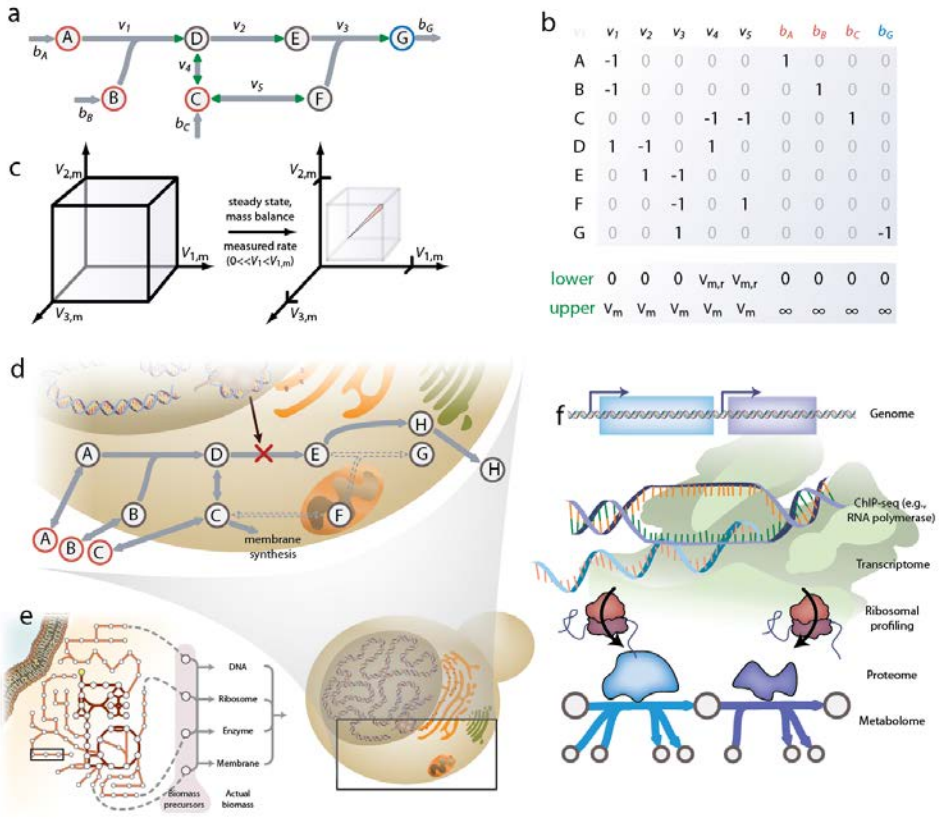
The COBRA approach is based on a few fundamental concepts. These concepts include 1) the imposition of physicochemical constraints that limit computable phenotypes (Figure 1.2.a-d), 2) the identification and mathematical description of evolutionary selective pressures (Figure 1.2.e), and 3) a genome-scale perspective of cell metabolism that accounts of all metabolic gene products in a cell (Figure 1.2.d,f). These fundamental concepts are briefly described here.

Constraints on reaction networks: determinants of proximal causation.

Metabolism is a complex network of biochemical reactions. The occurrence of any reaction is limited by three primary constraints: reaction substrate and enzyme availability, mass and charge conservation, and thermodynamics. For metabolism, reaction substrates must be present in a cell's microenvironment or produced from other reactions, and enzymes must be available. Mass conservation further limits the possible reaction products and their stoichiometry, while thermodynamics constrain reaction directionality. This information for all possible metabolic reactions in an organism can be detailed and catalogued in metabolic reconstruction knowledgebases [2, 3].

In the COBRA framework, the constraints detailed in a metabolic reconstruction are converted into an *in silico* model by mathematically describing the metabolic network and adding network inputs and outputs (e.g., uptake and secretion products). Much like a cell has one genome and many transcriptional states, an organism has one metabolic reconstruction from which context-specific models can be derived, each representing cellular functions under different conditions.

Model constraints are mathematically described by a matrix representing the stoichiometric coefficients of each reaction (Figure 1.2.a-b)[9]. Known upper and lower bounds on the flux through each reaction are imposed as additional constraints. Mathematically, these constraints define a multi-dimensional "solution space" of allowable reaction flux distributions. An actual expressed flux state resides within this solution space. Additional constraints can further shrink the solution space to focus in on the actual flux state of the network (Figure 1.2.c). These additional constraints may include enzyme capacity, spatial localization, metabolite sequestration, and multiple levels of gene, transcript, and protein regulation (Figure 1.2.d). Such constraints mechanistically describe proximal causation in a cell, since they account for the genetics and physical laws constraining metabolism and the cell phenotype.



	v_1	v_2	v_3	v_4	v_5	b_A	b_B	b_C	b_G
A	-1	0	0	0	0	1	0	0	0
B	-1	0	0	0	0	0	1	0	0
C	0	0	0	-1	-1	0	0	1	0
D	1	-1	0	1	0	0	0	0	0
E	0	1	-1	0	0	0	0	0	0
F	0	0	-1	0	1	0	0	0	0
G	0	0	1	0	0	0	0	0	-1
lower	0	0	0	$V_{m,r}$	$V_{m,r}$	0	0	0	0
upper	V_m	V_m	V_m	V_m	V_m	∞	∞	∞	∞

Figure 1.2. Fundamentals of the genome-scale metabolic genotype-phenotype relationship. COBRA is based on three primary fundamental concepts: network constraints (a-d), objective functions (e), and the association of reactions with the genome. (a) A mixture of molecules (red) can react to yield end products (blue). (b) The stoichiometry of this reaction network is described mathematically in a stoichiometric matrix, with each column representing the reaction stoichiometry. Negative and positive values represent reactants and products, respectively. Reaction flux is limited by thermodynamics and catalytic capacities ($V_m = V_{max}$), described by upper and lower bounds on each reaction flux (green). (c) Reaction constraints result in a "solution space" that contains all feasible flux distributions. Additional constraints (e.g., mass balance, the steady-state assumption, and measured metabolite consumption rates) reduce the space of feasible flux distributions, as shown by the pink line. (d) *In vivo* biochemical networks involve additional complexity. Gene regulation can change the abundance of catalysts (e.g., the transformation of D to E). Often components are localized in different organelles (e.g., E and F), thereby blocking reactions. (e) The biomass objective function describes an evolutionary pressure for microbial growth, and describes the metabolic demands to make basic metabolite building blocks for all of the cellular components (e.g., membranes, macromolecules, ATP, etc.). (f) The association of metabolism with the genome is done by mathematically linking the genome to transcripts, proteins, and chemical reactions. The gene-protein-reaction (GPR) schema is used to describe gene association in the models, and provide an interface for the integration of high-throughput data.

Mathematical statement of cell objectives: determinant of distal causation.

In non-biological chemical networks, the material flow through pathways can be predicted in a "cause and effect" manner, using mathematical models that describe the associated physical laws. This description can be achieved in a "time-invariant" manner, since reproducing the same physical conditions will drive flux through the same pathways.

In contrast, causation in biology is "time-variant". A vast array of chemical reactions may occur inside a cell, and many "pathways" can link a starting molecule to a given product. However, regulatory mechanisms have evolved to select when and where pathways will be used in an organism under a given condition. The selection of active pathways is a reflection of evolution and the result distal causation. Thus, if the cellular objectives that drive evolution are understood or can he hypothesized, optimal flux states of biochemical reaction networks can be predicted. In the COBRA framework these cellular objectives are described mathematically and used for computation of phenotypic states.

Many cellular objectives can be defined in the context of metabolism (Figure 1.2.e). For example, as a proxy for growth, a biomass reaction can be defined that contains all necessary precursors for synthesizing the cell components for growth (e.g., with amounts of amino acids for proteins and the nucleic acids for RNA)[10]. The biomass function and other objective functions can be used with optimization algorithms, such as linear programming[11, 12] to predict metabolic pathway usage and cellular phenotypes[12-14].

Since these objective functions mathematically state cellular aims and can predict phenotypes, they capture pressures guiding evolution, and therefore represent a determinant of distal causation. The objective function is thus an important part of the COBRA framework. It is not based on fundamental physical principles, but based on biological functions that are selected for over many generations.

A genome-wide basis for modeling metabolism.

Constraint-based modeling has rapidly developed since the advent of whole-genome sequencing in 1995[15, 16]. A genome provides the genetic basis for the metabolic network in a target organism, and genome annotation allows the delineation of gene-protein-reaction associations (GPRs), which describe the relationships between genes, enzymes, and the reactions they catalyze (Figure 1.2.f)[17]. Annotated genomes and associated biochemical and genetic data have facilitated the development of carefully curated and validated metabolic network reconstructions containing thousands of reactions. When a reconstruction knowledgebase for an organism is converted into a genome-scale model (GEM), the mathematical representation forms the constraints, and the objective function can be used to represent the optimal biological functions that it strives to achieve. Thus, simulation of phenotypic states can be performed using a GEM for a target organism.

There are two primary implications stemming from the genome-scale view of metabolism of GEMs. First, since, in principle, they account for all known metabolic genes in a cell, they can be used for analyzing genome-scale datasets (e.g., proteomic, transcriptomic, metabolomic, etc.)[18], while accounting for how the components are chemically connected (Figure 1.2.f). Second, since each metabolic gene is associated with the biochemical functions of its gene product, simulations of metabolite flow through the network can provide clear, mechanistic predictions of how each gene product affects the metabolic network function. Thus, cell phenotypes can be readily computed and data can be interpreted with GEMs, thereby providing mechanistic insight into how the cell genotype may contribute to the cell phenotype.

A "phylogeny" of constraint-based methods and applications

COBRA methods have 'evolved' and 'diversified' over the past decade, leading to the development of more than 100 different methods, many of which have been implemented in available software packages. These developments may be likened to an evolutionary process, in which specific scientific questions have selected for algorithmic innovations, yielding a phylogenetic tree of COBRA methods (Figure 1.3). Here we classify these methods into major groups and describe examples that address the broader scientific questions.

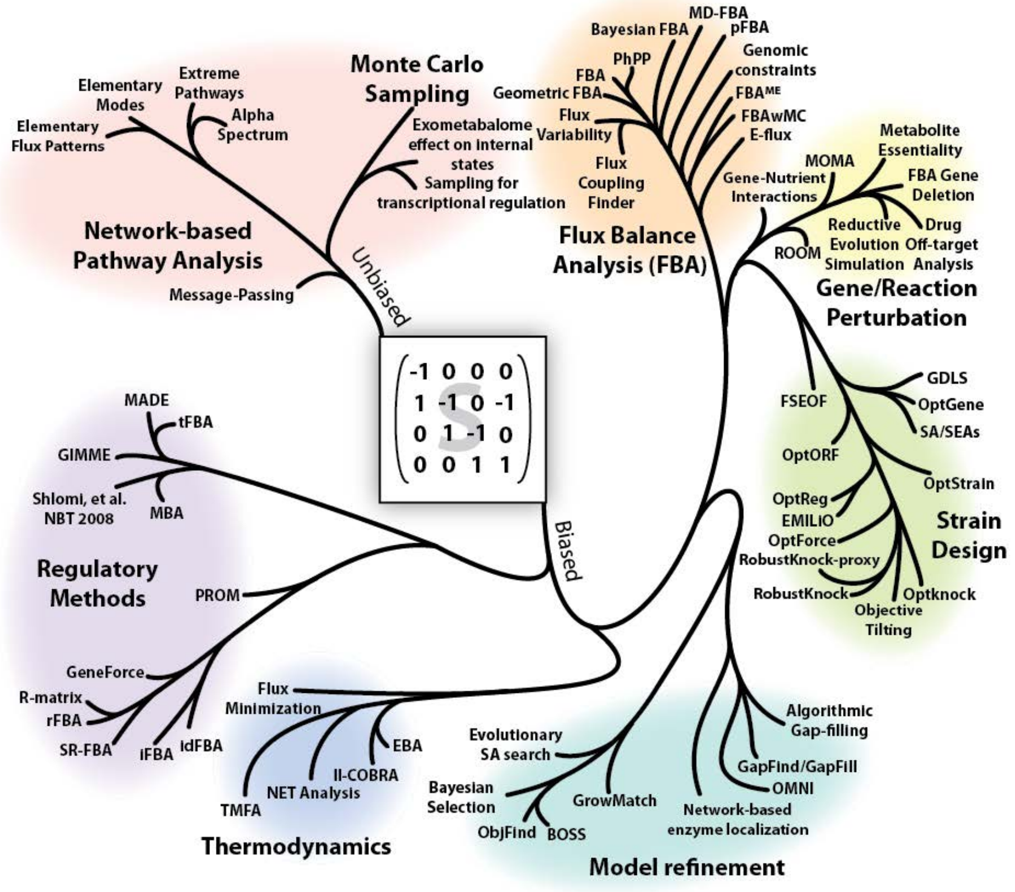


Figure 1.3. The "phylogeny" of constraint-based modeling methods. Over the past years, the constraint-based modeling community has rapidly expanded. Because of the versatility and scalability of these models, more than 100 methods have been developed for their modeling and analysis, all based on the analysis of the underlying metabolic network structure (i.e., the stoichiometric matrix). A phylogenetic tree is used to depict the similarities between applications of the methods and the underlying algorithms for many of the methods.

Global characterization of solution spaces

Metabolic pathways are conceptual abstractions that group reactions together. However, sometimes these "pathways" fail to reflect the actual usage of metabolic networks[19]. Fortunately, the "pathways" needed for specific metabolic functions can be identified without biases from traditional pathway concepts. In constraint-based modeling, this is approached through unbiased and biased methods, represented by the two primary branches of the phylogenetic tree (Figure 1.3). While biased methods will be described later, unbiased methods describe all steady-state flux distributions, including reaction sets that function together without belonging to the same traditional pathway concepts.

Elementary flux mode (EFM) analysis and extreme pathway (ExPa) analysis provide global and unbiased characterization of allowable phenotypes, and have been reviewed and compared previously[20-22]. These methods identify reaction sets (i.e., pathways) that carry flux through the network, and combinations of these reaction sets describe the entire solution space (i.e., all steady-state phenotypes). These methods have enjoyed many applications, including assessing pathway regulation[23], aiding in designing ethanol-secreting *E. coli*[24], identifying synthetic lethals[25], and demonstrating the trade-off between reducing translation costs and rapidly responding to environmental changes[26]. These methods are generally applied to small models or small portions of GEMs[27], since their computational complexity scales exponentially[28, 29]. However, simplifications are beginning to permit their use on larger models[30-33].

Alternative approaches have concurrently been developed to describe the entire "solution space" in an unbiased fashion[34, 35]. For example, Markov-chain Monte Carlo sampling (MCMC) methods[35] characterize all feasible steady-state reaction fluxes. This provides a probability distribution of feasible fluxes for each reaction under the user-provided growth conditions. These methods have successfully provided insight into several biological properties, such as the high flux backbone of central metabolism in *E. coli*[36], condition-specific regulation of yeast[37, 38] and *E. coli*[39] metabolism, and disease states in cardiac myocytes[40], erythrocytes[41], and the human brain[42].

Finding the "optimal" metabolic state with FBA methods

EFM, ExPa, and MCMC methods characterize all possible flux states a metabolic network can deploy. However, a cell does not use most possible flux states. Thus, biased COBRA methods include the optimization of an objective function to focus in on physiologically relevant flux distributions. Flux Balance Analysis (FBA) is the most basic and commonly used biased method for simulating genome-scale metabolism. In FBA, the cellular objective is defined, and metabolites in the media are supplied to the metabolic network. Linear programming is then used to optimize an objective function (e.g., the biomass objective function) subject to the constraints imposed by the metabolic network and metabolite uptake rates[11, 12, 43]. This calculation finds one solution in the solution space that is believed to best represent the true cellular phenotype. The solution includes a prediction of the optimal objective magnitude (e.g., biomass yield or growth rate) and potential flux values for each reaction (Figure 1.4.a).

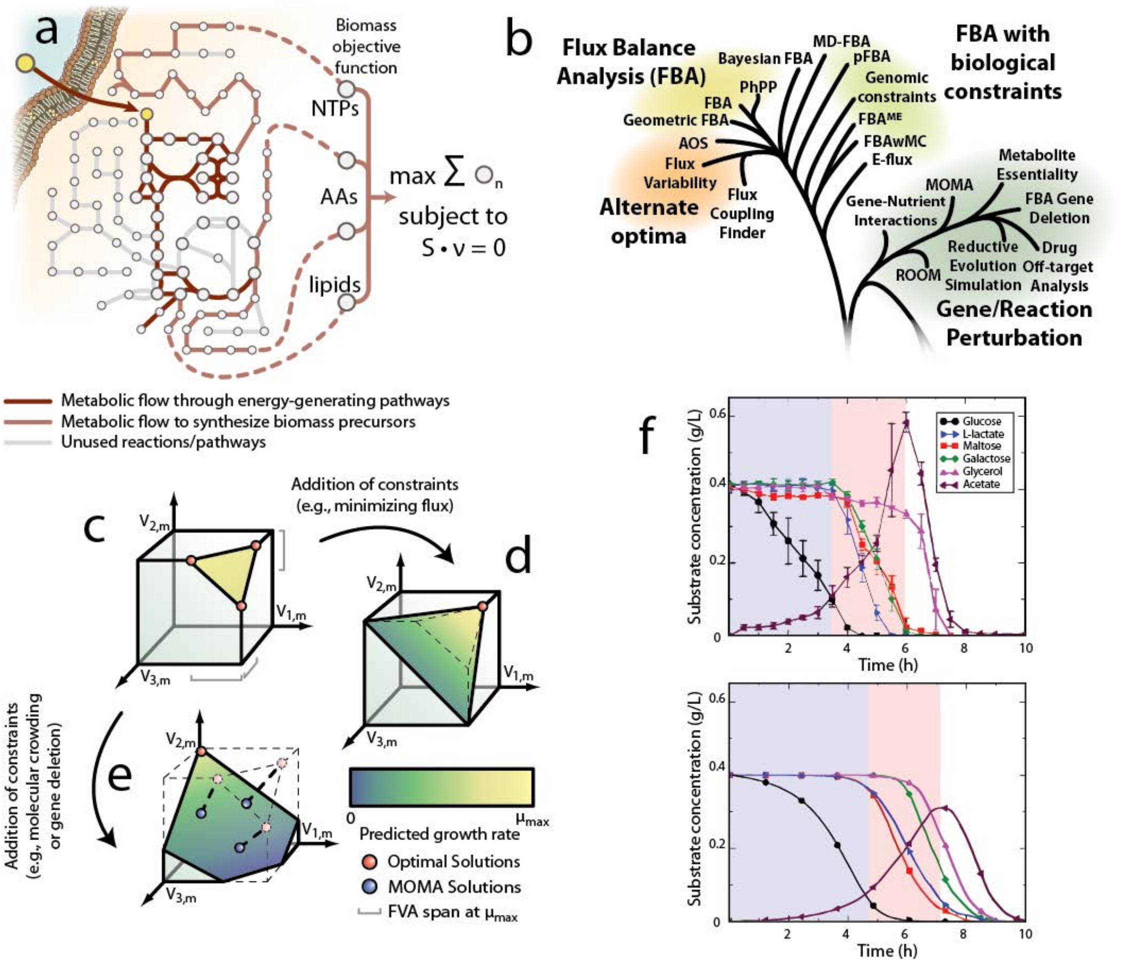

a
Biomass objective function
NTPs
AAs
lipids
max Σ ○n
subject to
S • v = 0
Metabolic flow through energy-generating pathways
Metabolic flow to synthesize biomass precursors
Unused reactions/pathways
b
Flux Balance Analysis (FBA)
Bayesian FBA
MD-FBA
pFBA
PhPP
FBA
Geometric FBA
Genomic constraints
FBA with biological constraints
FBA^ME
FBAwMC
E-flux
AOS
Flux Variability
Alternate optima
Flux Coupling Finder
MOMA
Metabolite Essentiality
Gene-Nutrient Interactions
FBA Gene Deletion
Reductive Evolution Simulation
Drug Off-target Analysis
ROOM
Gene/Reaction Perturbation
c
Addition of constraints (e.g., minimizing flux)
d
V2,m
V1,m
V3,m
e
Addition of constraints (e.g., molecular crowding or gene deletion)
0
μmax
Predicted growth rate
Optimal Solutions
MOMA Solutions
FVA span at μmax
f
Glucose
L-lactate
Maltose
Galactose
Glycerol
Acetate
Substrate concentration (g/L)
0
0.2
0.4
0.6
Time (h)
2
4
6
8
10

Figure 1.4. Flux balance analysis (FBA). (a) In FBA, a cellular objective (e.g., biomass production) is optimized. This provides the predicted flux for each reaction in the network. (b) FBA and related methods can be classified into groups that use FBA in its purest form, methods that assess alternate optimal solutions, and methods with additional biological constraints. Several variants on FBA also exist that apply perturbations to genes or reactions. (c) FBA solutions are typically not unique, i.e., there are alternate optimal solutions that use different pathways to achieve the same objective value (e.g., growth rate). (d) Additional constraints can be applied to reduce the solution space size, and may remove competing optimal solutions, or (e) change the optimal solution. If the optimal solution is moved, then the choice of the new optimal solution may depend on the solver and/or algorithm, as shown for the MOMA method. (f) The addition of constraints can enhance predictions. For example, when constraints on molecular crowding are added, the model-predicted order of substrate metabolism is consistent with experimental observation. Panel f reproduced from [44], Copyright 2007, National Academy of Sciences, USA. NTPs, nucleotide triphosphates; AAs, amino acids; FVA, flux variability analysis; v, reaction flux; μ_{max}, predicted maximum growth rate.

FBA successfully makes quantitative predictions using a few governing constraints on the model. For example, a pre-genome era application of FBA recapitulated the acetate overflow phenotype of *E. coli*[45]. Using GEMs, FBA has since predicted growth rates[46], pathway usage[47, 48], reaction stoichiometry[49], and the effect of gene expression noise on fitness[50]. It allowed the analysis complex phenotypes, such as metabolism in non-growing cells[51]. In addition, numerous variations on FBA have been developed to assess alternative optimal solutions or to account for additional constraints on metabolic flux in cells (Figure 1.4.b).

Predicted flux values from FBA can vary due to alternate optimal solutions (i.e., the same objective value using different reactions). Alternate optimal solutions are enumerated using mixed-integer linear programming (MILP)[52] and the ranges spanned by alternate optima are found for each reaction using flux variability analysis (Figure 1.4.c)[53, 54]. Some unlikely alternate optima can be removed by employing additional model constraints (Figure 1.4.d). The consideration of all alternate optima is critical when interpreting an FBA solution, since the flux through a single reaction can vary considerably depending on which solution is found. For example, the COBRA method Minimization of Metabolic Adjustment (MOMA)[55] predicts a new flux vector and objective value after a perturbation (e.g., gene deletion). To do this, MOMA calculates one “wild type” FBA solution, and finds the nearest solution after perturbing the network (i.e., the minimum change to reaction fluxes from the FBA solution). Since the new predicted flux vector and growth rate can differ considerably depending on which alternate optimal solution is used (Figure 1.4.e), all possible results from alternate optima must be assessed.

To focus in on realistic microbial phenotypes in FBA predictions, additional biologically-relevant constraints have been proposed. These include constraints imposed by genomic organization[56], economy in enzyme usage[47, 57-59], metabolite dilution[60], and changes in transcript level[61, 62]. These refinements of FBA further decrease the range of feasible reaction fluxes to obtain solutions closely resembling cellular physiology under certain growth conditions. For example, constraints from molecular crowding have been applied to FBA solutions (FBAwMC)[44]. In FBAwMC, reaction flux is constrained to reflect internal limitations on enzyme abundance in the crowded cytoplasm. This method predicted that molecular crowding contributes to substrate preferences in *E. coli*. In a medium with multiple carbon substrates, FBAwMC accurately predicted that glucose would be consumed first, followed by mixed-substrate consumption and a late utilization of glycerol and the excreted acetate (Figure 1.4.f), suggesting that molecular crowding may contribute to substrate preference. A similar variation on FBA accounts for crowding on the cytoplasmic membrane (FBA^{ME})[63] by limiting the flux through the glucose transporter and the three cytochromes in *E. coli*. This constraint recapitulated the simultaneous use of respiratory and fermentative pathways and predicted the effect of glucose and oxygen availability on cytochrome oxidase expression. Thus, by imposing molecular crowding constraints on metabolic flux, both FBAwMC and FBA^{ME} have provided additional insights into cell physiology.

Modeling gene, reaction, or metabolite perturbation

Since genome-scale metabolic networks capture the activities of hundreds of enzymes, mutant phenotypes can be assayed through *in silico* gene perturbation and simulation. Such approaches immediately demonstrated the predictive power of COBRA methods on the first GEMs[15, 16]. In these studies, metabolic genes were “knocked out” in the model by restricting the flux through their associated reactions to zero. When growth of mutant *E. coli* was simulated with FBA, 86% of the mutant phenotypes (i.e., growth or no growth) were accurately predicted[16]. This success rate was far greater than any other phenotype-predicting algorithm at the time. These initial analyses were followed by variations of the gene deletion concept, with methods such as MOMA[55], Regulatory On/Off Minimization[64], and Metabolite Essentiality Analysis (MEA)[65](Figure 1.4.b).

Gene and reaction perturbation studies have aided health-related applications, such as assessing unexpected metabolic effects of off-target protein-drug interactions[66] and predicting novel anti-microbial targets[67]. For example, MEA[65] was applied to the metabolic GEM of *Vibrio vulnificus*[68] in an effort to identify potential antibiotic targets for this pathogenic relative to the causative agent of cholera. MEA identifies metabolites that, if removed, inhibit biomass production. Such metabolites could be blocked *in vivo* with analogues that bind or modify active sites on enzymes that normally synthesize or catabolize the associated essential metabolite[69]. Here, this analysis yielded five candidate metabolites that could be targeted. Thus, only 352 analogues had to be screened for antimicrobial properties, which is much fewer than commonly used for drug discovery screens. One of several small molecules that showed antimicrobial properties was subjected to additional study, and this candidate molecule considerably out-performed sulfamethoxazole, an existing therapeutic for *V. vulnificus* infection. While additional drug safety assessment and optimization is required for this candidate drug, this study demonstrates how COBRA methods can guide antibiotic screens and provide immediate insight into their mode of action.

In silico **design of production strains**

Metabolic engineering approaches often perturb and screen cells for desired phenotypes. However, engineered strains can decrease product yield over time, since products drain cellular energy and resources. Several COBRA methods aim to address this by predicting perturbations (e.g., gene deletions or additions) that force the strain to couple product yield to a cellular objective, such as growth. Thus, as cells grow exponentially, they may also improve their productivity[70] (Figure 1.5.a).

Most COBRA strain-design methods systematically identify reactions that, when perturbed, may couple a product to a selective pressure (Figure 1.5.b). For example, OptKnock[71] employs MILP on a wild-type model (Figure 1.5.c.i) to find reaction deletion sets that force product secretion under optimal growth (Figure 1.5.c.ii). However, since OptKnock optimizes both the biomass objective function and product yield, strain designs occasionally have alternate optima with other secretion products (Figure 1.5.c.iii). To avoid this challenge, the product can be added to the biomass function (Objective Tilting[72]) or MILP can be used to find designs that provide the maximum lower bound on product yield while maximizing growth (RobustKnock[73]) (Figure 1.5.c.iv).

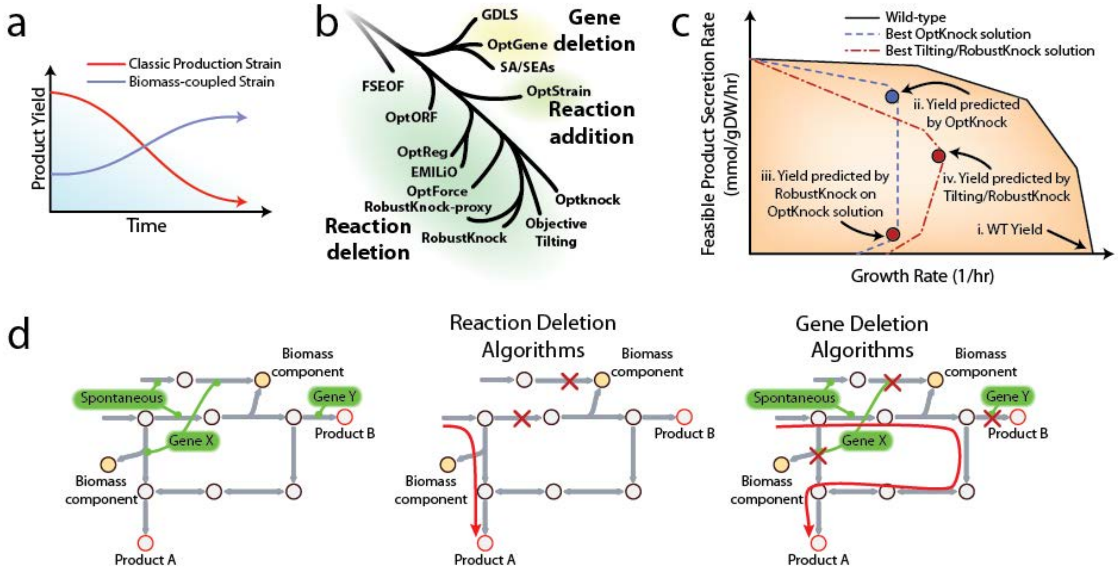


Figure 1.5. Principles of model-guided strain design. (a) Non-growth-coupled production strains witness a decrease in product yield over time, while growth-coupled strains can enhance product yield. (b) A number of methods have been developed to predict growth-coupled production strains by modeling reaction deletion, gene deletion, or reaction addition. (c) Growth-coupled strain designs are predicted to force product secretion while growing optimally. Different algorithms, such as OptKnock, Objective tilting, and RobustKnock can provide different optimal growth-coupled strain designs. (d) Many algorithms predict the set of reactions that must be blocked to obtain a desired product. However, a few methods provide a more realistic view by modeling genetic modifications, since some genes catalyze multiple reactions, and other reactions are spontaneous.

For algorithmic simplicity, most strain design methods focus on perturbing reactions. However, strain designs based on reactions can require additional gene deletions (isozymes). Moreover, predictions are occasionally not feasible when they require the removal of one reaction catalyzed by a multi-specific enzyme (Figure 1.5.d). To avoid such predictions, heuristic approaches, such as OptGene[74] and GDLS[75], identify growth-coupled production strain designs that directly involve gene deletions. Thus, strain designs from such methods are more realistic and easier to test *in vivo*.

Strain-design predictions are not limited to manipulations of the host cell's metabolic pathways. The repertoire of products may be expanded *in silico* by adding genes from other organisms to confer novel metabolic functions. *In silico* methods have used graph theoretical approaches[76-78] or kinetic parameters[78] to build novel biosynthetic pathways, which were subsequently tested or ranked using FBA or related methods. Unfortunately, without accounting for the host metabolic network, these approaches cannot guarantee growth-coupled strain designs. Thus, without further engineering (e.g., with scaffolds[79-81]) predicted biosynthetic pathways may not yield product when implemented *in vivo*. However, this concern has been addressed by 1) manually removing genes to growth-couple the new pathways [78], 2) systematically following pathway prediction with OptKnock[82], or 3) conducting the novel pathway search within the host-cell metabolic network to optimize the balance between added and deleted reactions, as done in OptStrain[83]. Thus, approaches have been developed to couple the synthesis of a non-native product to a cellular objective.

The concept of designing strains that couple a product to a defined selective pressure is not only intriguing, but COBRA-based *in silico* predictions have been successfully implemented *in vivo*[70, 82]. It is anticipated that these tools will continue to aid metabolic engineering projects.

Refining representations of dual causation

Simulating proximal and distal causality with COBRA requires accurate representations of metabolic network stoichiometry and objective functions. While metabolic reconstructions are usually carefully built and rigorously tested, they are often incomplete. In addition, metabolic networks may contain a few errors in stoichiometry, thermodynamics, gene associations, or biomass composition, resulting from ambiguities in associated biochemical studies [84] or genome annotation [85]. Moreover, biomass compositions can vary between environments[86], and biomass optimization does not always describe the cellular objective[87], especially under nutrient limitation, stationary phase, or stress[51, 88]. To address these concerns, phenotypic screens have been analyzed with gap-filling COBRA methods (Figure 1.6.a) to predict missing pathways and their associated genes[89, 90], to identify reactions with incorrect directionality or inclusion[84, 91-93], and to predict subcellular localization of reactions within microbes with multiple organelles[94]. Complementary COBRA methods also improve the definition of cellular objectives by integrating data to systematically assess[95-97], predict[98], or modify objective functions[84, 86, 92].

Recently, high-throughput genetic interaction screens have helped refine metabolic networks and the biomass objective function of yeast[84, 99]. For example, model-predicted epistasis in *S. cerevisiae* was compared with 176,821 experimentally measured genetic interaction pairs. Although the COBRA model predictions were enriched for high-confidence measured genetic interactions, it did not predict many epistatic interactions. The authors developed an algorithm that reconciled discrepancies between model-predicted and experimentally measured interactions. Several predicted model improvements were experimentally validated. For example, the authors found that quinolinate formation from aspartate was wrongly included in the yeast reconstruction. In addition, the algorithm predicted that glycogen should be removed as an essential component in the biomass objective function, since it is not essential for growth. Thus, this study demonstrated that COBRA methods could be deployed to improve the yeast metabolic network and provide condition-specific updates to the biomass objective function.

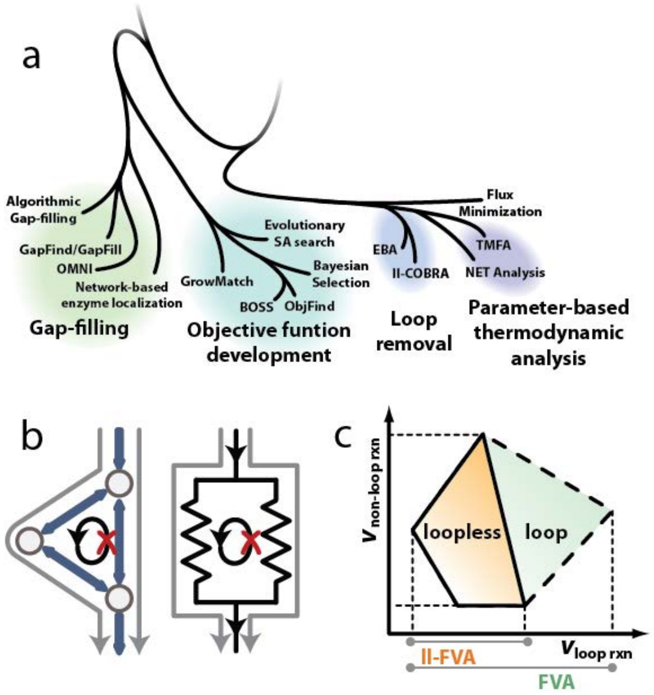


Figure 1.6. Model-guided model refinement. COBRA methods can systematize the refinement of biochemical knowledge. (a) Computational methods can be used improve network topology by filling network gaps and determining subcellular localization. In addition, several methods exist to reconcile model objectives with experimental data in an effort to understand cellular objectives under a given condition. Lastly, the refinement of model thermodynamics is an important part of constraint-based modeling. Thus, methods exist to remove thermodynamically infeasible solutions, to derive thermodynamic data from the model, or use thermodynamic parameters to improve reaction directionality constraints. (b) When a metabolic network is not adequately constrained, metabolites can cycle infinitely in loops. Akin to Kirchhoff's loop law for electrical circuits, this property is thermodynamically infeasible. Thus, methods like ll-FVA, which uses the loopless-COBRA constraints on flux variability analysis, are able to systematically remove these loops by adding a constraint that limits solution to the regions of the solution space that are not involved in these loops.

Thermodynamics

COBRA methods provide quantitative predictions without detailed parameterization of each reaction, beyond declaring directionality to reflect reaction thermodynamics. Directionality is often determined from biochemical assays, but such assays may not recapitulate the conditions and metabolite concentrations inside the cell. Therefore, reaction directionality *in vitro* may be inconsistent with *in vivo* flux. In addition, unrealistic fluxes can be predicted *in silico* if a reaction is reversible in a model, but irreversible *in vivo*. Thus, to improve model predictions, methods are now applying more rigorous thermodynamic constraints (Figure 1.6.a) by removing thermodynamically infeasible pathway usage[100-102] or constraining flux based on Gibbs free energy calculations[57, 103, 104]. Methods are also being used to infer thermodynamic parameters [105].

Most COBRA models contain sets of reactions that can cycle metabolites amongst themselves (Figure 1.6.b). In these cases, FBA cannot predict reliable flux values for these reactions, since their metabolites can be cycled infinitely. Such "loops" are biologically unrealistic since no net thermodynamic driving force exists, akin to Kirchhoff's second law for electric circuits. Thus the net flux around these loops should be zero[100]. While these loops often do not affect predicted non-loop reaction flux, their existence can severely upset predictions from other methods, such as MOMA (E. Ruppin, personal communication). Approaches to systematically remove loops have been proposed[100-102]. For example, loopless-COBRA[101] improves FBA solutions by employing MILP to cancel out loop flux (Figure 1.6.c).

While loop-removal methods can be easily deployed without extra parameterization, detailed thermodynamic approaches may provide more biologically meaningful reaction flux predictions. Thermodynamic parameters for many metabolites are not known. Fortunately, recent advances in group contribution theory provide Gibbs free energy of formation estimates for metabolites in COBRA models[106]. With these predicted values, standard Gibbs free energy change of every reaction can be predicted. These values can help determine reaction directionality[57, 107], predict reasonable concentration levels[103], and allow the use of metabolite concentrations[108] and ranges on kinetic parameters[104] as constraints. A recent study[109] used estimated metabolite free energy with experimentally measured equilibrium constants to quantitatively assign reaction directionality. This approach also incorporated *in vivo* pH, temperature, and ionic strength to quantitatively assign reaction directionality to the *E. coli* metabolic network. When the authors compared the model-predicted and experimentally measured growth rates, they found that the qualitative assignment of directionality to certain reaction classes (e.g., ABC and proton coupled transporters) was necessary, in addition to quantitative assignment, to match model predictions with experimental data. Since thermodynamics represents one primary model constraint necessary for accurate COBRA predictions, it is expected that further developments in this area will be of great importance to the field.

Incorporating regulatory constraints and signaling

Transcriptional regulation and signaling networks interface extensively with metabolism to produce cellular phenotypes (Figure 1.7.a). By incorporating regulatory and signaling constraints into the function of metabolic networks, interactions between the systems can be captured to enhance COBRA predictions. There are two primary paradigms that dictate how regulatory constraints are implemented in constraint-based models (Figure 1.7.b). Either experimental data is overlaid on the metabolic network[61, 62, 110-113] to constrain flux through specific reactions (Figure 1.7.c), or a mathematical representation of transcription regulation[114, 115] or signaling[116, 117] is interfaced directly with the metabolic network to aid in modeling (Figure 1.7.d).

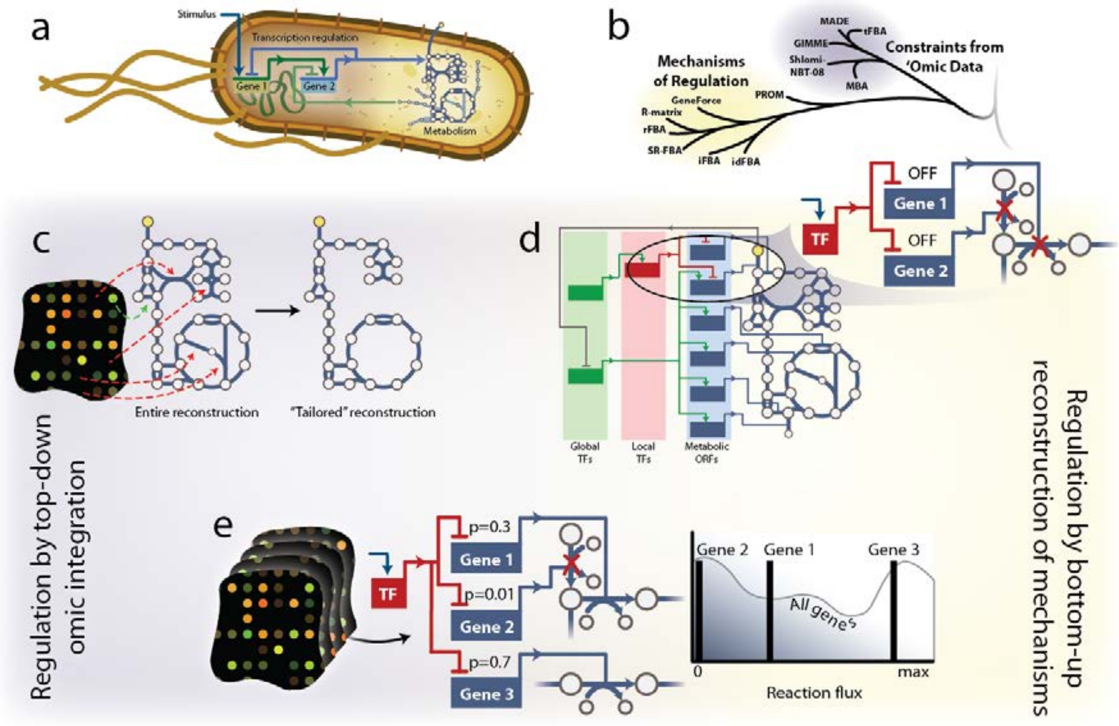


Figure 1.7. Incorporating and inferring regulation. (a) Signaling, transcription regulation, and metabolism are interlinked in the cell. Therefore it is desirable that the networks be integrated for more holistic modeling of organisms. (b) Two primary paradigms exist in COBRA modeling for integrating transcription regulation and metabolism. (c) Algorithms such as GIMME and MBA use high-throughput data and model simulations to identify which pathways are likely expressed and active in the cells when the data were sampled. This results in a tailored context-specific representation of the metabolic network. (d) Algorithms such as rFBA, iFBA, and SR-FBA incorporate detailed mathematical representations of the known molecular mechanisms of transcription regulation. These approaches contain binary regulatory logic that dictates, under a specific signal, which metabolic pathways are suppressed and cannot carry flux. (e) Hybrid methods, such as PROM are arising, in which transcriptomic data are used to infer the levels of constraints imposed by the regulatory network. PROM also uses probabilistic measures to allow for a more continuous regulation of reaction flux. For example, Gene 2 is tightly regulated by a transcription factor (TF). Thus, when the TF is activated by a signal, reaction flux is more tightly constrained than Gene 1, which is only loosely regulated.

Not all pathways are active under all growth conditions. Thus, 'omic data can be used to constrain models accordingly (Figure 1.7.c)[61, 62, 110-113]. Methods such as GIMME[110], Shlomi-NBT-08[111], and MBA[112] each remove pathways lacking expression in 'omic data to obtain functional models that are consistent with the data. In particular, these approaches have provided novel insights and discoveries in tissue-specific human metabolism[42, 66, 118, 119]. However, they were also recently used to model metabolic interactions between *M. tuberculosis* and a human alveolar macrophage[86], which will be discussed later.

To expand model predictions beyond metabolism, mathematical descriptions of regulatory mechanisms are being integrated with metabolic models (Figure 1.7.d). Such integrated metabolic and regulatory models can improve phenotype predictions and even find novel regulatory interactions. This was done for the nutrient-controlled transcriptional regulatory network for *S. cerevisiae*[120], which included Boolean regulatory interactions between 55 transcription factors and 750 metabolic genes. This integrated regulatory-metabolic network was used to simulate growth under different environmental and genetic perturbations using regulatory FBA (rFBA)[121]. The model predicted new transcriptional regulatory interactions, and elucidated regulatory cascades using chromatin immunoprecipitation data and transcription factor binding motifs. While integrated models of metabolism and transcription regulation can provide improved phenotype predictions, this study showed they can also expand regulatory knowledge. It is anticipated that these models may further demonstrate metabolic pathway usage in conditions for which 'omic data are not available.

A few variations on rFBA have been suggested[115, 116]. Despite their success, rFBA and related methods have two primary weaknesses. First, they assume binary responses for all transcriptional regulatory interactions, when real biological systems exhibit a range of behavior in transcriptional regulation, from binary to continuous. Second, few organisms have been studied enough to provide adequate regulatory information for rFBA. However, a method called probabilistic regulation of metabolism (PROM) addresses these concerns[122]. When ample transcriptomic data are available, PROM uses an organism's known transcriptional regulatory network to infer probabilities in how it integrates with the metabolic network, yielding an improved regulatory-metabolic network model. Moreover, PROM can apply intermediate responses (as opposed to binary), since it uses conditional probabilities for modeling transcription regulation instead of hard Boolean rules (Figure 1.7.e).

PROM was deployed to model the integrated regulatory-metabolic network of *M. tuberculosis* [122]. All TFs modulating metabolic gene expression were systematically deleted from the model and *in silico* growth phenotypes were compared with experimentally measured phenotypes. PROM correctly predicted 96% of the TF knockout phenotypes, including 5 of the 6 TFs that were essential for optimal growth. This suggests that this method may help predict antibiotic targets for both regulatory and metabolic genes.

An ecosystem of COBRA methods to address larger scientific questions

Individual COBRA methods have answered numerous scientific questions, and many with respect to biological causation. However, a great strength of the COBRA framework is that many methods can be deployed in parallel or in series to obtain additional insights into a question of interest. Moreover, different models can be easily swapped or combined to test hypotheses relevant to different species. Thus by using a community of methods and several data types, deeper insights into larger questions may be attained. Here I provide examples of how COBRA methods have complemented each other and provided insight into microbial community interactions.

The community structure in an organism's microenvironment helps to shape metabolic pathways usage. Thus the prediction of proximal and distal responses in biology can be affected by neighboring organisms. Organisms will compete for scarce resources and/or depend on the metabolic capabilities of their cohabitants. Evolution often selects for cells that leverage this community structure, as is regularly manifested in cellular metabolism[123, 124]. COBRA methods are now modeling and characterizing several aspects of metabolism's role in microbial community structure[125-127]. These studies are providing insight into mutualism[128], competition[129], parasitism[86, 130], and community evolution[124, 131].

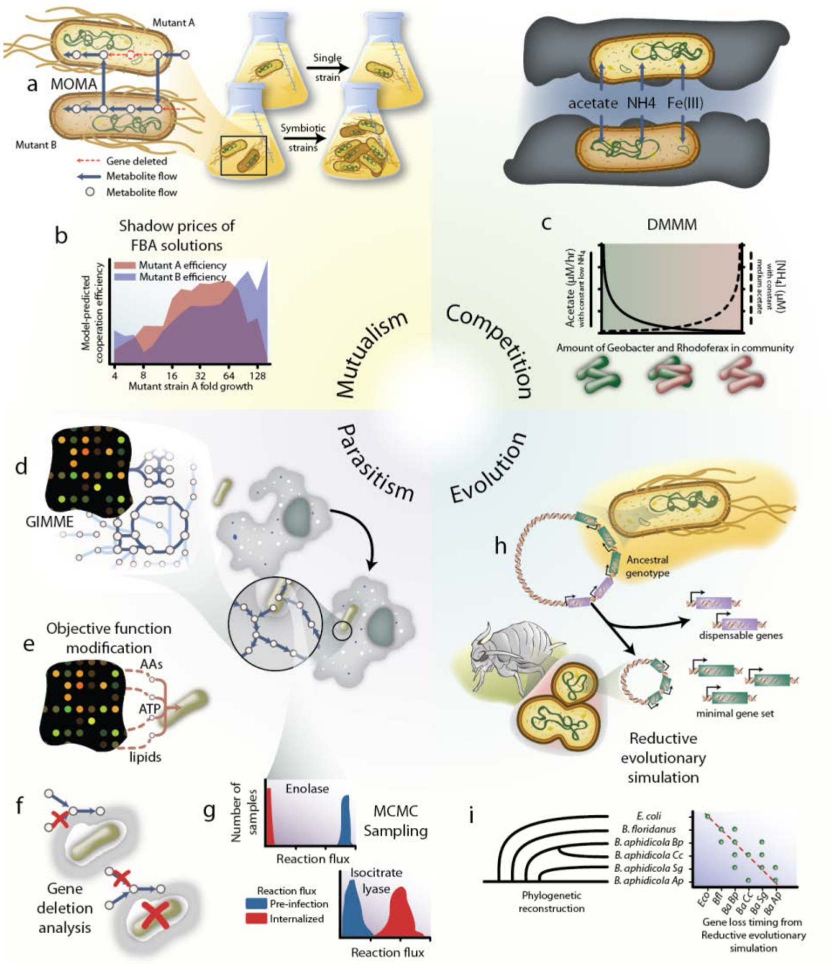


Figure 1.8. Integrating COBRA methods to study community interactions. COBRA methods are providing insight into the metabolic interactions in various types of microbial communities. (a) To study the mutualistic behavior of co-dependent mutant *E. coli*, researchers used MOMA to simulate synergistic growth of pairs of auxotrophic *E. coli*. (b) Shadow prices from FBA simulations of these pairs were used to compute cooperation efficiencies between strains, which were subsequently compared with measured fitness improvements. (c) Competition in communities was modeled using DMMM to understand how communities of *Geobacter* and *Rhodoferax* compete for resources, and how the demographics varies under different nutrient ratios, thereby affecting the efficiency of bioremediation efforts. Host-pathogen interactions between *M. tuberculosis* and a human macrophage were studied using COBRA. (d) GIMME leveraged transcriptomic data to build host-pathogen models at different stages of infection. (e) Since the cellular objective of internalized *M. tuberculosis* is not known, refinements to the objective function were predicted from transcriptomic data. (f) The reliability of these models was assessed by comparing gene deletion analysis simulations with experimental gene essentiality data. (g) Flux states of internalized *M. tuberculosis* were simulated using MCMC sampling and found a suppression of central metabolism and activation of the glyoxylate shunt, represented here by enolase and isocitrate lyase, respectively. The role of communities in evolution has been studied using Reductive evolutionary simulation. In particular, this method predicted the minimal set of genes needed to for *Buchnera* to grow in the rich innards of the aphid. The predicted minimal gene sets (h) and order of gene loss (i) were consistent with the gene content and phylogenetic structure of several *Buchnera* species.

Mutalism

Synthetic mutualism between auxotrophic mutants of *E. coli* was recently studied using COBRA methods[128]. The authors grew pairs of auxotrophic mutants and then modeled their coupled metabolism using MOMA to identify mutant pairs that complement each other's growth by exchanging essential metabolites (Figure 1.8.a). Shadow prices from FBA were used to assess balance between the cost (from metabolite loss) and the benefit (from receiving missing essential metabolites) to each rescued auxotroph. The cooperative efficiency (i.e., the ratio of uptake benefit to production cost) recapitulated the observed growth of the co-cultures. Significant increases in growth (Figure 1.8.b) were witnessed in co-cultures that exchanged beneficial, but less costly metabolites (i.e., higher cooperative efficiency). While it is difficult to directly measure the exchange of metabolites between the auxotrophs, the computed cooperation efficiency provides an indirect quantitative assessment of the metabolite cross-feeding in this mutualistic system.

Competition

Metabolic competition for scarce nutrients has also been assessed with COBRA methods. Dynamic multi-species metabolic modeling (DMMM) characterized the competition for acetate, Fe(III), and ammonia between *Geobacter sulfurreducens* and *Rhodoferax ferrireducens* in subsurface anoxic environments (Figure 1.8.c)[129]. DMMM simulates the growth rate of both organisms and the rates of change of external metabolites, to dynamically predict population changes in the community. Using DMMM, the community composition was predicted under geochemically distinct conditions of low, medium, and high acetate flux. Under low acetate flux, DMMM predicts *Rhodoferax* dominates the community when sufficient ammonia is available, whereas *Geobacter* dominates under low ammonia and high acetate flux. This difference was attributed to the nitrogen fixation abilities of *Geobacter*, as well as its higher acetate uptake rate compared to *Rhodoferax*. Moreover, it was also predicted that under nitrogen fixing conditions, *Geobacter* increases its respiration at the expense of biomass production, thus showing how balancing community structure can impact the efficacy of uranium bioremediation in low ammonium zones.

Parasitism

Host-pathogen interactions are now being studied with COBRA methods[130]. A recent study modeled the metabolic interactions between a human alveolar macrophage and *M. tuberculosis*[86]. Context-specific models of infection were built with GIMME[110] and Shlomi-NBT-08[111] using transcriptomic data from three types of *M. tuberculosis* infections (Figure 1.8.d). Next, the objective function was iteratively revised using infection-specific gene expression data in order to represent the metabolic activity of the pathogen *in vivo* (Figure 1.8.e). Gene deletion analysis was compared with *in vivo* gene essentiality data (Figure 1.8.f), and MCMC sampling was also used to demonstrate a significant alteration in metabolic pathway usage in *M. tuberculosis* during macrophage infection, including a suppression of glycolysis and an increased dependency on glyoxylate metabolism (Figure 1.8.g). This constraint of central metabolism during *M. tuberculosis* infection was also suggested by DCP, another method related to FBA[132]. This suppression of certain metabolic pathways with an increased dependency upon normally latent pathways may provide novel antibiotic targets.

Community evolution

A central tenet of evolutionary theory is that over time, genetic drift and selective pressure causes organisms to optimize their cellular machinery for a particular niche. This assumption of cellular optimization has made COBRA methods useful tools to investigate hypotheses concerning organismal evolution, as recently reviewed by Papp, et al.[7] In nature, the optimization and evolution of microbial metabolism is a multi-species affair, as demonstrated by the aphid endosymbiant *Buchnera aphidicola*. This descendant of the *Enterobacteriaceae* family has suffered drastic loss of genomic material as it evolved in its host's nutrient-rich innards. Since *B. aphidicola* is related to *E. coli*, reductive evolutionary simulation (a derivative of gene deletion analysis)[124] on the *E. coli* model was used to predict the minimal metabolic gene sets. It was found that these minimal sets are highly consistent with the metabolic gene content of *B. aphidicola* (Figure 1.8.h). In addition, the predicted order of gene loss could explain ~40% of the variation in the phylogenetically reconstructed gene loss time among the genomes of five *Buchnera* species (Figure 1.8.i)[131], thus suggesting that the bacterium optimized its pathway usage for its new rich habitat. However, metabolic pathways retained in the computed minimal gene sets highlight the bacterium's role in community evolution. Retained pathways contained reactions needed for producing riboflavin and essential amino acids lacking from the aphid diet, thereby highlighting their role in the symbiotic relationship[124]. Thus, COBRA methods are helping to describe the role of the community environment to distal causation, and how evolution shapes gene content in symbiotic communities[7].

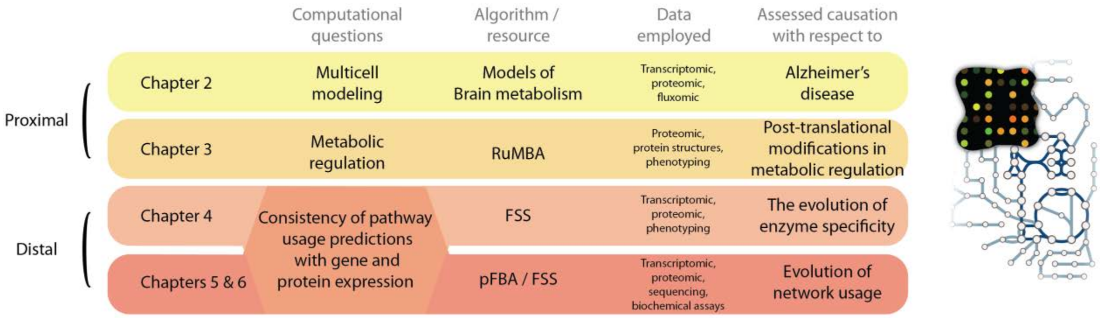
On the pathway to causation in biology

Dual causation and the sheer complexity of biological systems have presented considerable challenges in achieving the quantitative rigor in the life sciences that is enjoyed in other fields of science. This has been in part because the assessment of causation requires 1) a means to quantitatively model physical phenononae under the constraints of physical laws, 2) detailed knowledge of the properties of components in the system, and 3) an understanding of how these components interact.

Fortunately, in the past decade, cell and molecular biology has become increasingly quantitative as the molecular parts that define the form and function of organisms have been measured and their interactions have been characterized. COBRA has arisen as a quantitatively rigorous framework that integrates these data from high-throughput technologies with knowledge from decades of careful studies on the biochemical properties of the gene products. Lastly, COBRA provides a quantitative framework that accounts for the constraints imposed by physical laws and allows for the quantitative prediction of phenotypes and the underlying mechanisms.

The remaining chapters of this dissertation aim to demonstrate how COBRA methods contribute to the goal of assessing causation in complex biochemical networks by showing how these methods can be integrated with various data sources (Figure 1.9). For each case study, an open question with respect constraint-based modeling is addressed. To do this, an algorithmic or conceptual advance to COBRA is presented. Ultimately, each of these studies integrates COBRA methods with numerous data types to assess causation with respect to several biological questions.

Thus, by mathematically describing cell biochemistry and evolutionary selective pressures, reconstructed networks and COBRA methods are now presenting a basis for the quantitative analysis and prediction of complex metabolic genotype-phenotype relationships.



		Computational questions	Algorithm / resource	Data employed	Assessed causation with respect to
Proximal	Chapter 2	Multicell modeling	Models of Brain metabolism	Transcriptomic, proteomic, fluxomic	Alzheimer's disease
	Chapter 3	Metabolic regulation	RuMBA	Proteomic, protein structures, phenotyping	Post-translational modifications in metabolic regulation
Distal	Chapter 4	Consistency of pathway usage predictions with gene and protein expression	FSS	Transcriptomic, proteomic, phenotyping	The evolution of enzyme specificity
	Chapters 5 & 6		pFBA / FSS	Transcriptomic, proteomic, sequencing, biochemical assays	Evolution of network usage

Figure 1.9. A roadmap for this dissertation, in which high-throughput data can be integrated with genome-scale simulations of metabolism to study causation in biochemical systems. In each chapter, an open question in constraint-based reconstruction and analysis is addressed with the development on new algorithms and resources. Several data types are integrated to answer these questions, and then further used to address causation with respect to a biological question of interest.

Chapter 1 is a modified version of material in Lewis, N.E., Nagarajan, H., and Palsson, B.Ø. Constraining the metabolic genotype-phenotype relationship using a phylogeny of *in silico* methods. *Nature Reviews Microbiology* (2012). I was the primary author, while the co-authors participated in and supervised the drafting of this review.

Section I: Proximal causation in metabolism

Life continues because the chemical organization therein is structured such that it can respond and adapt to moderate fluctuations in environment, and utilize the resources in these habitats to propagate its genetic programs. These responses to variations in environment represent proximal responses.

For decades, scientists have been trying to predict these proximal responses and identify their causative factors through the use of various modeling techniques, such as statistical models, phenomenological models, and models that detail the biomolecular components in a cell. As demonstrated in the previous chapter, constraint-based modeling and analysis has been successful over the past decade because it allows one to model complex phenotypes, using detailed biochemical knowledge on hundreds or thousands of genes and proteins, simultaneously. In the following two chapters, conceptual developments are presented with respect to the usage of constraint-based modeling as a context in which large high-throughput transcriptomic and proteomic data sets are analyzed. These approaches subsequently provide insight into proximal causation in response to internal pathologies and environmental variation. Specifically, the chapters presented here demonstrate causation in human neurological disease phenotypes and microbial metabolic regulation in response to fluctuating environmental conditions. Through both of these chapters it is clear that across widely different organisms, proximal responses may be modeled and causation in physiology and pathophysiology can be predicted.

Chapter 2: Large-scale *in silico* modeling of metabolic interactions between cell types in the human brain.

Constraint-based reconstruction and analysis of genome-scale microbial metabolic networks has matured over the past decade, and has provided a wealth of insight into how organisms function and respond to environmental cues. For example, these predictions have aided in metabolic engineering, and provided a mechanistic bridge between genotypes and complex phenotypes[5, 6]. Computational methods[133] and a detailed SOP[3] have been outlined for the reconstruction of high-quality prokaryotic metabolic networks, and many methods can be deployed for their analysis[43, 134]. Constraint-based modeling of metabolism entered a new phase with the publication of the human metabolic network (Recon 1)[135], based on build-35 of the human genome. Methods allowing tissue-specific model construction have followed[110-112].

Tissue metabolic functions often rely on interactions between many cell types. Thus, methods are needed that integrate the metabolic activities of multiple cells. Here, using Recon 1, we analyze and integrate omics data with information from detailed biochemical studies to build multicellular constraint-based models of metabolism. We demonstrate this process by constructing and analyzing models of human brain energy metabolism, with an emphasis on central metabolism and mitochondrial function in astrocytes and neurons. Moreover, we provide three detailed examples, demonstrating the use of models to provide insight into the metabolic mechanisms underlying physiological and pathophysiological states in brain.

Building metabolic models of multiple cell types

Omics datasets can be difficult to analyze due to their size. However, such datasets can be used to construct large mechanistic models for specific tissues and cell types[110, 111] that serve as a context for further analysis. The workflow for generating multicellular models, as depicted in Figure 2.1, consists of the following four steps:

Step 1. An organism-specific metabolic network is reconstructed from genome annotation, lists of biomolecular components, and the literature[3]. Metabolic pathways and associated gene products are not completely known for any species. Thus, a reconstruction is refined through iterations of manual curation, hypothesis generation, experimental validation, and incorporation of new knowledge. Recon 1 has been through five iterations[135].

Step 2. Many gene products are not expressed in all cells at any given time[136]. Therefore, gene product presence from omic data is mapped to Recon 1 using the gene-protein-reaction associations to obtain a draft reconstruction for the tissue of interest. This process may be performed manually or algorithmically[110, 111].

Step 3. Initial context-specific reconstructions are incomplete and may contain false positives due to proximal tissue contamination. Moreover, few high-throughput datasets are cell-type specific. Thus, the initial reconstruction represents the union of metabolic networks from various cell types. To address this problem, the literature is searched to verify enzyme localization and partition the model into compartments representing different cell types and organelles. Upon completion, the reconstruction is converted into a model by specifying inputs, outputs, relevant parameters, and by representing the network mathematically[9]. See Thiele and Palsson[3] for details of proper manual curation.

Step 4. Once the network is accurately reconstructed and converted into an *in silico* model, it is used for simulation and analysis[5, 6, 134] for hypothesis generation and to obtain insight into systems-level biological functions.

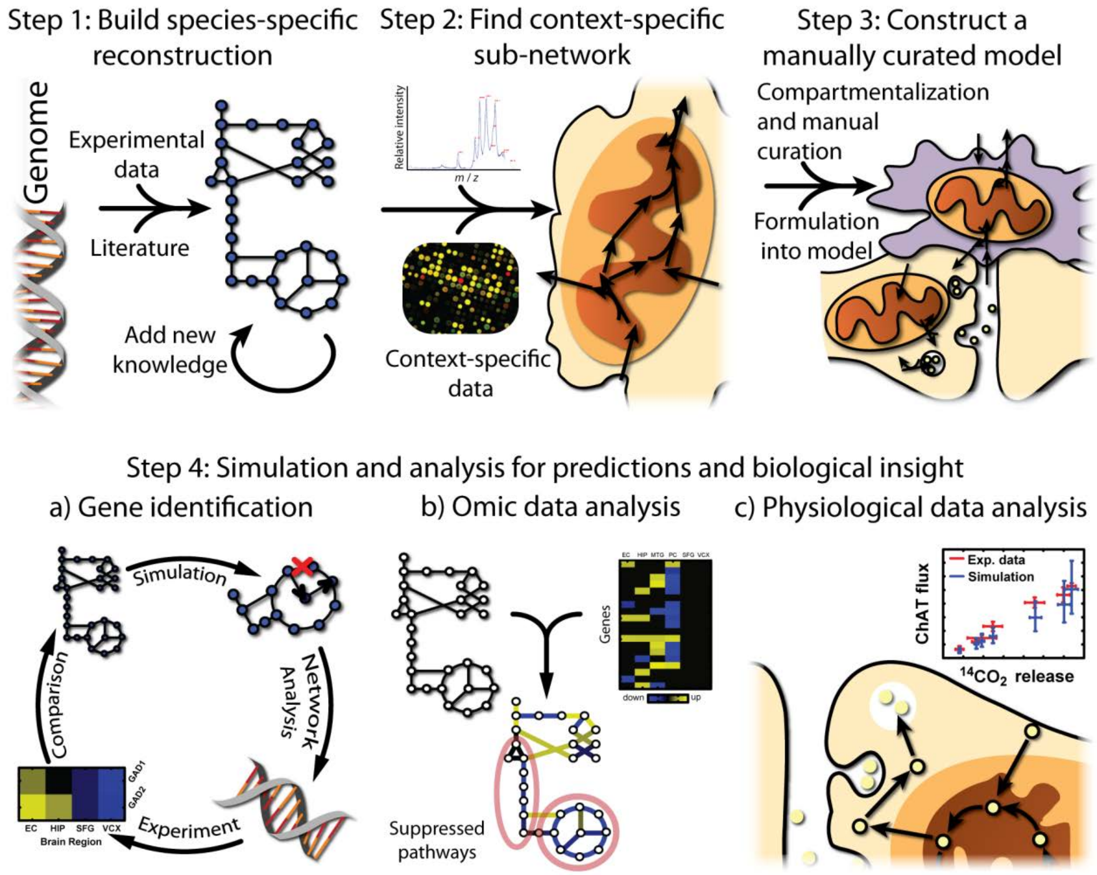


Figure 2.1. A workflow for bridging the genotype-phenotype gap with the use of high-throughput data and manual curation for the construction of multicellular models of metabolism. Metabolic models of multicellular tissues can be constructed to gain insight into biology and make testable hypotheses. First, a species-specific reconstruction is built based on genome annotation, experimental data, and knowledge obtained from the literature. Second, high-thoughput data can be mapped to the reconstruction in order to find a context-specific network (e.g., representing a tissue). Third, multicellular models are constructed as the context-specific network is organized into compartments representing different cell types, based on cell-specific knowledge and data. These networks are linked together with the transport of shared metabolites, and then formulated into a model. Fourth, the models are utilized for simulation and analysis to gain insight and generate testable hypotheses. For example, the models can be used to a) predict disease-associated genes, such as glutamate decarboxylase in this work. b) High-thoughput data can be analyzed in the network context to identify sets of genes that change together and affect specific pathways, such as the brain-region-specific suppression of central metabolism in Alzheimer’s disease patients. c) Physiological data can be analyzed in the context of the model, therefore allowing, for example, the calculation of the percentage of the brain that is cholinergic.

This workflow was used to build three different multicellular models of brain energy metabolism. Each model represents one canonical neuron type (i.e., glutamatergic, GABAergic, or cholinergic), its interactions with the surrounding astrocytes, and the transport of metabolites through the blood-brain barrier (Figure 2.2). This reconstruction focuses on the core of cerebral energy metabolism, including central metabolism, mitochondrial metabolic pathways, and pathways relevant to anabolism and catabolism of three neurotransmitters: glutamate, γ-aminobutarate (GABA), and acetylcholine. Thus, the three models contain the high flux pathways and important reactions in neuron and astrocyte metabolism. These models currently represent the largest and most detailed models of brain energy metabolism [137-139]. Our models contain 1066, 1067, and 1070 compartment-specific reactions, transformations, and exchanges, involving 983, 983, and 987 metabolite/compartment combinations, for the glutamatergic, GABAergic, and cholinergic models, respectively, and are associated with a total of 403 genes. The validity of these models is demonstrated through various tests and comparisons to physiological data. Specifically, our model predicts ATP production rates within 8% of the average published value, and internal flux measurements are consistent with experimentally measured values [42]. Moreover, three analyses using the models are detailed here. Since most of these analyses cannot be done on previous brain models or Recon 1 as published, this work provides novel insight into brain energy metabolism.

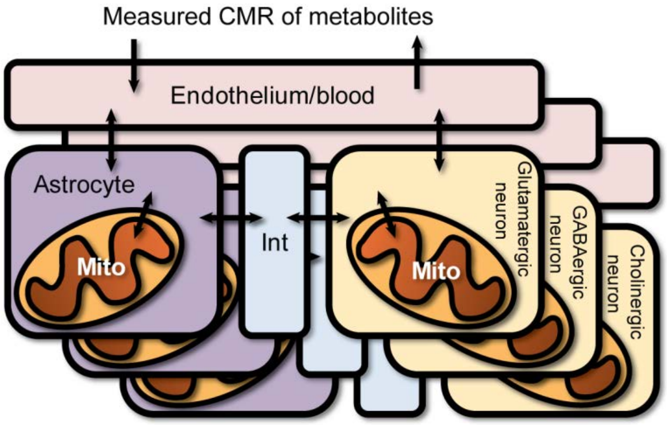


Figure 2.2. General structure of the models. Three models were built from the brain reconstruction. Each model consists of various compartments: 1) the endothelium/blood, 2) astrocytes, 3) astrocytic mitochondria, 4) neurons, 5) neuronal mitochondria, and 6) an interstitial space between the cell types. Each neuron metabolic network was tailored to represent a specific neuron type, containing genes and reactions generally accepted to be unique to the neuron type. Mito = mitochondrion, Int = interstitial space, CMR = cerebral metabolic rate.

Identifying a potentially neuroprotective gene in AD

Alzheimer's disease (AD) is characterized by histopathological features, including neurofibrillary tangles and β-amyloid plaques. Moreover, there is a strong metabolic component, in which metabolic rates of various brain regions decrease years before the onset of dementia[140].

Several central metabolic enzymes exhibit altered expression or activity in AD, such as pyruvate dehydrogenase (PDHm), α-ketoglutarate dehydrogenase (AKGDm), and cytochrome c oxidase (CYOO)[141-143]. The activities of these enzymes are affected by the AD-related proteins β-amyloid and Tau kinase[144, 145]. *In silico*, as the activity of these enzymes decreases, neurons demonstrate impaired metabolic capacity, and deficiencies in PDHm activity leads to a decreased cholinergic neurotransmission capacity (see [42] for more details).

AKGDm deficiency shows the greatest impairment in post-mortem AD brain (57% decrease in activity)[142]. An *in silico* analysis shows that this deficiency impairs the metabolic rate in glutamatergic and cholinergic neurons (Figure 2.3.a), since it limits oxidative phosphorylation (OxPhos) capacity in neurons (Figure 2.3.b-c). Such impairment of OxPhos leads to neuronal apoptosis[146]. Surprisingly, however, OxPhos is not impaired in the GABAergic neuron model (Figure 2.3.d). This model-derived result is consistent with the experimental observation that glutamatergic and cholinergic neurons are more affected in moderate stages of AD[147], while most GABAergic cells are relatively unaffected until late stages[148]. Therefore, the models were further interrogated to identify a mechanism that allows GABAergic neurons to absorb the perturbation, thereby leading to the cell-type-specific effects.

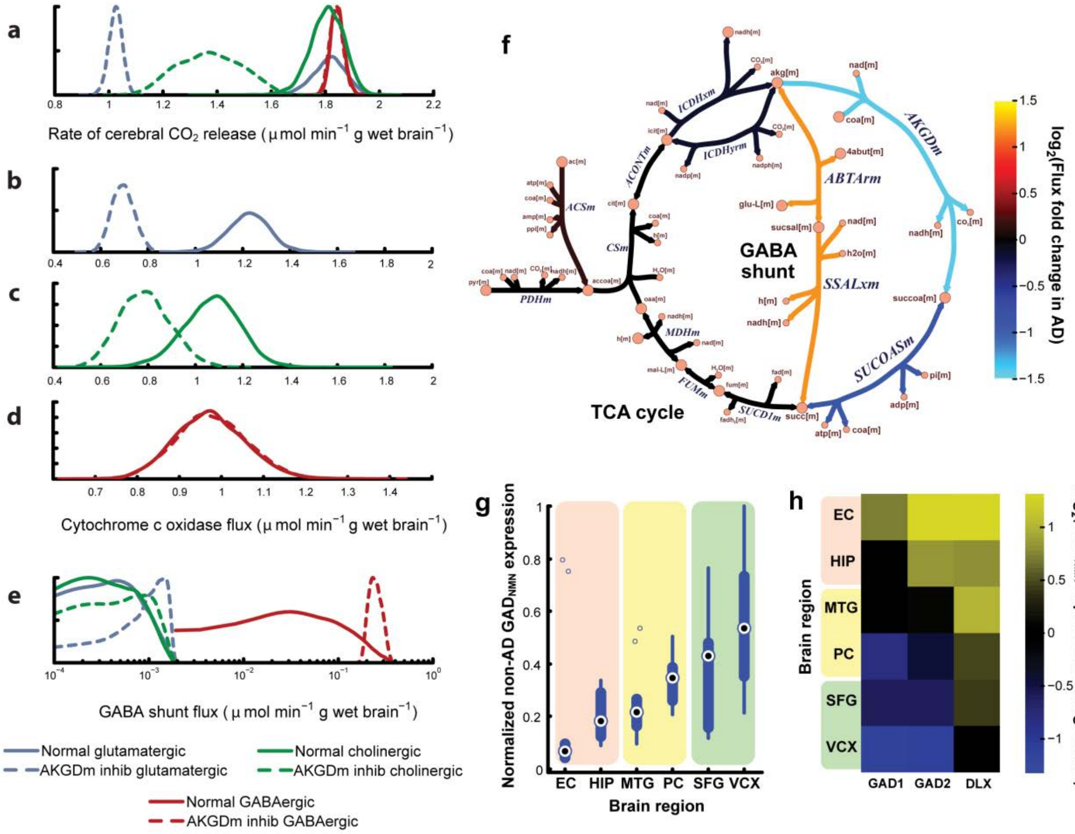


Figure 2.3. Decrease in α-ketoglutarate dehydrogenase (AKGDm) activity, associated with Alzheimer's disease (AD), shows cell-type and regional effects *in silico* consistent with experimental data. Kernel density plots show the distribution of feasible fluxes for various reactions (a-e). An *in silico* reduction of AKGDm flux from normal activity (a-e, solid lines) to AD brain activity (a-e, dashed) decreases (a) the oxidative metabolic rate for glutamateric and cholinergic neurons, but not GABAergic neurons. This results from a decrease in the feasible fluxes for oxidative phosphorylation (e.g., cytochrome c oxidase) for both (b) glutamatergic and (c) cholinergic neurons, but not (d) GABAergic cells. This cell-type-specific protection from the AKGDm deficiency results from (e) an increased flux through the GABA shunt in GABAergic cells, by bypassing the damaged AKGDm (f). GABAergic cells maintain a higher GABA shunt flux because of the expression of glutamate decarboxylase (GAD). Neuroprotective properties of GAD are supported by gene expression. (g) Severely damaged brain regions in AD patients have lower GAD$_{NMN}$ expression in control brain, while high GAD$_{NMN}$ regions (SFG and VCX) show little damage. In AD brain, (h) severely affected regions (HIP and EC) show an increase in GAD$_{NMN}$ and the GAD-inducing DLX family, suggesting that non-GAD expressing neurons may be lost in AD. EC = entorhinal cortex, HIP = hippocampus, MTG = middle temporal gyrus, PC = posterior cingulate cortex, SFG = superior frontal gyrus, VCX = visual cortex, NMN = neuron marker normalized, inhib = inhibited. All reaction and metabolite abbreviations are defined in [42].

Simulations show that GABAergic neurons absorb the AKGDm perturbation through the GABA shunt (Figure 2.3.f), a pathway that uses 4-aminobutyrate transaminase and succinate-semialdehyde dehydrogenase to bypass part of the TCA cycle. However, our models suggest that glutamatergic and cholinergic neurons cannot, despite carrying a small flux through the shunt enzymes (Figure 2.3.e). Support for these results includes recent evidence that suggests that cerebellar granule neurons, which have higher levels of GABA, can absorb perturbations to AKGDH through this shunt[149].

To identify the mechanism allowing only GABAergic neurons to use the GABA shunt to absorb the AKGDm perturbation, an *in silico* analysis was performed to identify genes that contribute to this (see [42] for details). This analysis suggests that the two isoforms of glutamate decarboxylase (GAD) could provide the cell-type specific neuroprotection. GAD allows the GABA shunt to carry a higher flux following the AKGDm perturbation in GABAergic neurons; however, the lack of GAD in other neuron types greatly limits the use of the GABA shunt *in silico* (Figure 2.3.e). Therefore, by fueling the GABA shunt, GAD may play a neuroprotective function, thus contributing to the sparing of GABAergic systems in earlier AD[148].

Certain populations of glutamatergic and cholinergic cells tend to be lost earlier in AD, but others survive. Interestingly, while GAD is canonically a GABAergic gene, it occasionally shows low expression in other neuron types, including glutamatergic and cholinergic cells[150]. Therefore, such populations of non-GABAergic, GAD-expressing neurons would also be protected. Thus, we follow with an analysis of the correlation of GAD expression and AD pathology for further validation.

If GAD has neuroprotective capacity *in vivo*, two properties of GAD expression are expected. First, brain regions with less GAD per neuron will be more affected in AD, while regions with abundant GAD will be spared. To test this hypothesis, we used a compendium of published microarrays of non-tangle-bearing neurons from six brain regions in Alzheimer patients and age-matched non-Alzheimer controls[151]. Among control patients, GAD expression levels among the brain regions is consistent with the extent of neuron loss found in AD patients; i.e., brain regions with more neuron loss in AD (e.g., the entorhinal cortex and hippocampus) have lower GAD expression in control patients, while relatively unaffected regions in AD (e.g., superior frontal gyrus and visual cortex) show much higher levels of GAD expression (Figure 2.3.g).

For the second property, neurons with low GAD expression should be lost in AD; therefore, GAD expression per neuron should increase in histopathologically affected regions. Consistent with the hypothesis, there is a significant increase in the expression of the brain-specific GAD2 in the entorhinal cortex and hippocampus in AD (p=0.0050 and 0.018, respectively) (Figure 2.3.h). As a control, all other neuron-specific genes tested failed to show a correlation with AD pathology, except the Dlx genes, which induce GAD expression in the brain (Figure 2.3.h)[152]. Therefore, these results lend additional support to the possibility that GAD may be providing a neuroprotective effect, and that this effect is correlated with the regional specificity of AD. Moreover, the model was able to guide the identification of a gene and the mechanism for its role in AD, while the subsequent microarray analysis provides experimental support for the model prediction.

Microarray analysis shows pathway suppression in AD

Atrophy alone cannot explain the extent of impaired metabolism in AD in many brain regions[153]. Therefore, in such regions, there must be metabolic pathway suppression within surviving cells. To test this, we used PathWave[154], a method that identifies differentially expressed pathways in omic data based on intra-pathway connectivity in a user-provided metabolic network (see Methods).

In this analysis, each brain region demonstrates different amounts of changed metabolic pathways. The visual cortex and superior frontal gyrus lack any differentially expressed pathways, consistent with previous work that shows little change in metabolic rate in AD in these regions[151]. However, the posterior cingulate cortex (PC) and middle temporal gyrus (MTG) have the largest numbers of significantly differentially expressed pathways (23 and 18 pathways, respectively). These two regions show significantly decreased metabolic rates in AD, but show fewer histopathological effects[151]. Both the entorhinal cortex (EC) and hippocampus (HIP) also show decreases in expression of nine metabolic pathways, though the number of suppressed pathways may be lower since these regions suffer a high amount of neuron loss, and only histopathologically healthy neurons were expression profiled. Therefore, more affected neurons may already have been lost or not profiled.

In the PathWave analysis, the four brain regions that show significantly low metabolic rates in AD (PC, MTG, HIP and EC)[153] all show a significant suppression of glycolysis and the TCA cycle (Figure 2.4). In addition, the HIP, MTG and PCC show a suppression of the malate-aspartate shuttle and OxPhos. Individual regions also show a suppression of other pathways, such as heme biosynthesis (MTG, PC), ethanol metabolism (EC, PC), and several amino acid metabolism pathways. Thus, using PathWave with our models, we find that the decreased metabolic rate in specific regions in AD is associated with the down-regulation of central metabolic gene expression in histopathologically normal neurons.

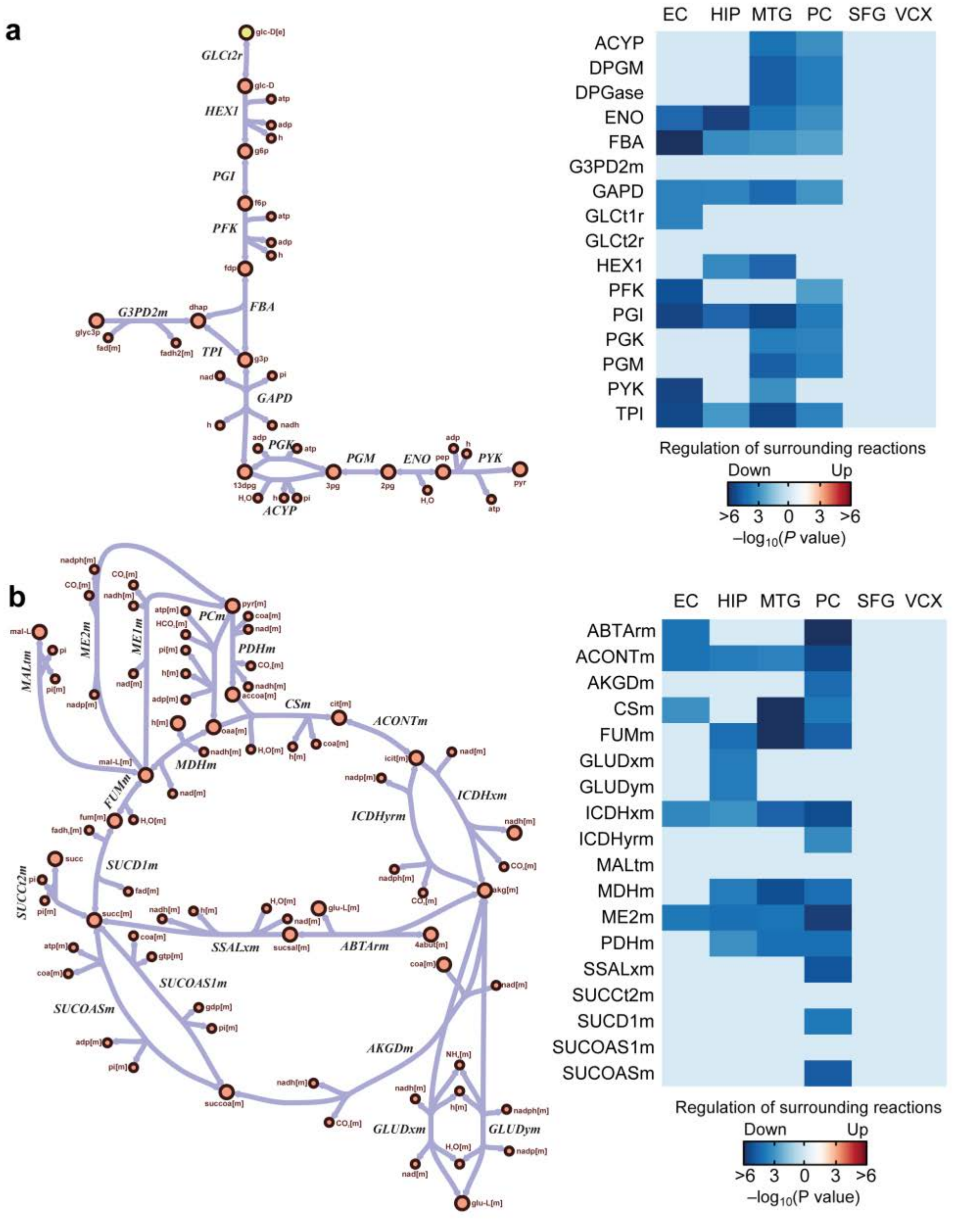


Figure 2.4. Metabolically affected brain regions in AD show significant suppression of central metabolic pathways. In certain AD brain regions, the metabolic rate of glucose decreases more than can be explained by brain atrophy. PathWave analysis demonstrates that histopathogically normal cells from the metabolically affected brain regions (EC, HIP, MTG, and PC) demonstrate a significant suppression of central metabolic pathways, such as (a) glycolysis and (b) the TCA cycle and surrounding reactions. Metabolically less affected regions (SFG and VCX) show no significant suppression. Reaction suppression shown here is a composite expression of the reaction associated genes and the genes of closely connected reactions. Only significantly changed reactions are shown (FDR = 0.05). EC = entorhinal cortex, HIP = hippocampus, MTG = middle temporal gyrus, PC = posterior cingulate cortex, SFG = superior frontal gyrus, VCX = visual cortex. All reaction and metabolite abbreviations are defined in [42].

Models detail properties of cholinergic metabolic coupling

Studies have demonstrated that the use of cytosolic acetyl-CoA for the synthesis of the neurotransmitter acetylcholine comes from the acetyl-CoA formed in the mitochondria. This tight coupling of acetylcholine to mitochondrial metabolism allows treatments that increase glucose uptake in the brain to improve cognitive functions in rats[155] and humans with severe cholinergic cognitive pathologies, such as Alzheimer's Disease and Trisomy-21[156]. Pathways that transport acetyl-CoA carbon to the cytosol have been suggested; however, the mechanism is still not clear[157].

Constraint-based modeling was used to aid in the identification of two pathways that could indirectly transport acetyl-CoA into the cytosol, and provide insight into needed complementary pathways. To identify possible pathways, reaction sets were identified by randomly removing reactions from Recon 1 until a minimum set was determined that couples the mitochondrial and cytosolic acetyl-CoA pools. This was repeated until more than 21,000 unique minimal reaction sets were identified. Singular value decomposition was then used to identify dominant pathway features that frequently co-occur.

The first singular vector is dominated by reactions that occur most frequently in the reaction sets (e.g., water transport across the cell membrane). However, the second and third singular vectors are dominated by reaction sets that usually co-occur or never co-occur (Figure 2.5). These reactions cluster into three distinct pathways, providing hypotheses to aid in the reconstruction process. The omic data used in the reconstruction process, and a thorough literature search eliminated the pathway using cytosolic acetyl-CoA sythetase (Figure 2.5.a), and validated the other two, involving the transport and metabolism of acetyl-CoA-derived citrate or acetoacetate, using ATP-citrate lyase (ACITL) or cytosolic acetyl-CoA C-acetyltransferase (ACACT1r), respectively (Figure 2.5.b-c).

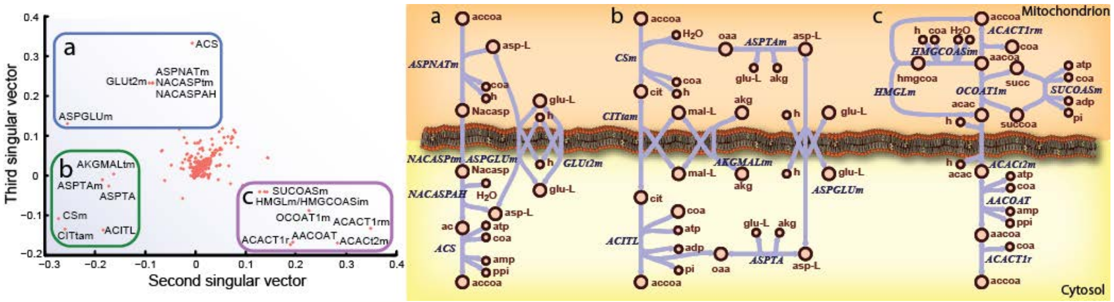


Figure 2.5. Singular Value Decomposition (SVD) of feasible pathways elucidates potential pathways that allow for coupling of mitochondria acetyl-CoA metabolism and cytosolic acetylcholine production. 21,000 unique feasible reaction sets were computed, each showing transport of mitochondrial acetyl-CoA carbon to the cytosol in human metabolism. SVD of a matrix of all 21,000 pathways yielded 3 primary pathways that allow this coupling of mitochondrial metabolism to acetylcholine production, by carrying the acetyl-CoA carbon on (a) N-acetyl-L-apartate, (b) citrate, or (c) acetoacetate. As shown by the second singular vector, reactions in the pathway with citrate tend to be missing from pathways when the reactions for the acetoacetate pathway are included. The third singular vector shows a similar relationship of the N-acetyl-L-aspartate pathway. The omic data and known enzyme localization only support the usage of citrate and acetoacetate as potential carriers in neurons.

Our model contains these two pathways and shows a correlation between the flux through mitochondrial pyruvate dehydrogenase and choline acetyltransferase ($r = 0.45$, $p = 3x10^{-247}$), consistent with the experimental observation of a tight coupling between mitochondrial metabolism and acetylcholine production. ACITL and ACACT1r also correlate with choline acetyltransferase flux ($p < 3 \times 10^{-90}$). Moreover, it has been reported that the inhibition of ACITL reduces the acetylcholine production rate by 30%[157]. The *in silico* inhibition of ACITL reduces acetylcholine production by 7.3%. The *in silico* decrease is smaller because the model can immediately adapt to the perturbation, while *in vivo* regulatory responses would take time to adapt. Therefore, it is expected that the model change is smaller. Interestingly, the inhibition of ACACT1r reduces acetylcholine production by 39%. Thus, it seems that cholinergic neurotransmission depends on redundant pathways, and that acetoacetate may play a more dominant role in transporting mitochondrial acetyl-CoA to the cytosol.

The coupling of mitochondrial metabolism to acetylcholine synthesis aids in the treatment of cholinergic disorders. However, knowledge of the abundance of cholinergic neurotransmission also aids in this purpose. It is difficult to identify cholinergic neurons, based solely on cell morphology, since cholinesterases and immunohistochemical markers for cholinergic neurons are also found in noncholinergic neurons and other tissues[158]. Therefore, it is unknown what percentage of all neurotransmission is cholinergic.

Using our cholinergic model, we compute the percent contribution of cholinergic neurotransmission based on published data[159]. The data used for this purpose were obtained from rat brain minces, incubated in solutions containing [1-^{14}C]pyruvate or [2-^{14}C]pyruvate. Both acetylcholine and radiolabeled CO_2 were measured at various titrations of several different pyruvate dehydrogenase inhibitors.

The cholinergic model was subjected to similar levels of pyruvate dehydrogenase inhibition. The simulations successfully reproduced the experimentally-witnessed linear relationship between acetylcholine production and metabolic rate, and acetylcholine production was correlated with CO_2 release ($r = 0.68$).

The fraction of cholinergic neurotransmission for the brain was computed by randomly choosing points from both the distributions of experimental data and distributions predicted by the simulations. A scaling factor was subsequently found that reconciles the two. This was repeated for 14 different combinations of pyruvate labeling and pyruvate dehydrogenase inhibitors[159], yielding a median predicted cholinergic portion of total brain neurotransmission of 3.3% (Figure 2.6.a). After adding this new parameter to the model, the predictions corresponded well with the experimental data sets (Figure 2.6.b), including six datasets representing three pyruvate dehydrogenase inhibitors withheld from the previous computations (Figure 2.6.c-d). Thus, the model was used in conjunction with experimental data to gain insight into physiological observations and derive important physiological parameters dependent on systems-level activity.

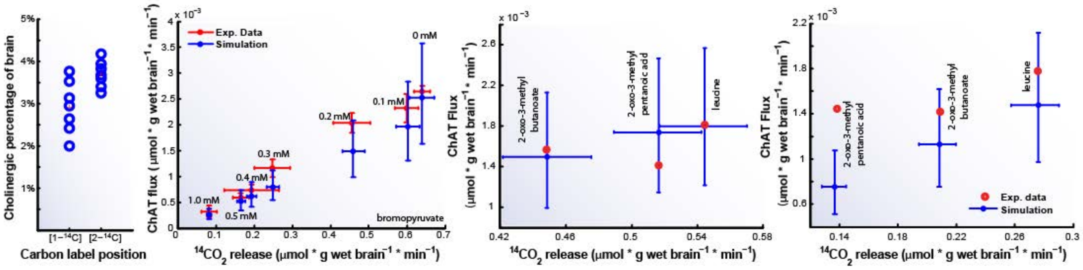


Figure 2.6. Model-aided prediction of cholinergic contribution is consistent with experimental acetylcholine production. Percent brain cholinergic neurotransmission was predicted based on 14 sets of experimental data in which brain minces were fed [1-^{14}C]-pyruvate or [2-^{14}C]-pyruvate, followed by measurement of ^{14}C-labeled CO_2 and acetylcholine. (a) For each experiment, the feasible amount of the brain that can generate the experimental response was computed, centering at 3.3%. (b) This parameter was employed in the analysis, and the updated model predictions were consistent with experimental data, such as seen in the case of treating the brain minces with [1-^{14}C]-pyruvate and increasing levels of the pyruvate-dehydrogenase inhibitor bromopyruvate. Moreover, the updated model predictions were consistent with measured ^{14}C-labeled CO_2 and acetylcholine production for brain minces that were treated with three PDHm inhibitors withheld from previous computations for both supplementation with (c) [1-^{14}C]-pyruvate and (d) [2-^{14}C]-pyruvate. Error bars on the simulation results represent 25th and 75th percentiles. ChAT = choline acetyltransferase.

Implications of these proximal responses

In this study a workflow was presented for generating tissue-specific, multicellular metabolic models. Through the analysis and integration of omic data, followed by manual curation, this workflow was used to build a first-draft manually-curated multicellular metabolic reconstruction of brain energy metabolism. Three models were generated from this reconstruction, representing different types of neurons coupled to astrocytes. We employed these models in three distinct analyses, each of which yielded predictions and insights into proximal responses and causation in AD and cholinergic neurotransmission, with respect to how the cells respond to metabolic perturbations. For example, we found that glycolysis seems suppressed in seemingly healthy neurons, and predicted a mechanism by which neurons selectively are lost in AD. Moreover, we were able to provide additional support for pathways that contribute to acetylcholine synthesis, a process that doctors have tried to enhance in AD patients in the past.

As experimental methods and data resolution improve, the accuracy of these models and their ability to predict causation in disease and responses to treatment may also improve. Improvements in neuroimaging and metabolomics will allow for more precise quantification of metabolite flow through the blood-brain-barrier, which is of interest since dysfunction of this system accompanies many neurological disorders and injuries [160]. In addition, improvements in transcriptomics and proteomics will provide higher-resolution quantification of cell- and organelle-specific genes and proteins. This data will allow models to account for neuron groups in specific brain regions, subcellular heterogeneity within cells, and the inclusion of less abundant glial cells. For example, higher-resolution models may provide insight into proximal causation during metabolic changes in specific cell populations, such as the structures closely related to the olfactory system, which are affected in the early stages of Alzheimer's disease [161].

As seen in this work, novel insight into mammalian tissue-specific metabolism may be gained as more multicellular models are constructed. Our models demonstrate metabolic coupling and synergistic activities that more coarse-grained models miss, since the three analyses presented here were not possible using Recon 1 or the previous models of brain metabolism. The compartmentalization of metabolic processes within cells[162], between cells[163], and in host-pathogen interactions has an important role in normal physiology. Therefore, such models may provide greater insight and more accurately predict how cells respond to the environment and carry out their true cellular functions.

In a broad sense, this study serves as an example of how mechanistic genotype-phenotype relationships can be built. From the genotype one can begin to reconstruct the network for an organism. The integration of high-throughput data and careful manual curation can add context-specific mechanistic network structure to genomic information. Thus, this network becomes a representation of the complex genetic interactions and biochemical mechanisms underlying observed phenotypes. This complex, but mechanistic, relationship between the genotype and phenotype can be used as a foundational structure upon which additional high-throughput data can be analyzed and predictive simulations can be conducted, thus leading to improved understanding, testable hypotheses, and increased knowledge[5, 6, 164]. Ultimately, this network encapsulates the functions of all known components, its use for elucidating mechanisms of proximal causation and using the network for predictive simulation.

Methods

Reconstruction of iNL403: This work focuses on the core of cerebral energy metabolism and the pathways that play a critical role in cell-type specific functions in the brain. The pathways included in this work include mitochondrial metabolic pathways, central metabolic pathways closely tied to mitochondrial function, and additional pathways that are needed for modeling neuron and astrocyte functions. To reconstruct these pathways, a list of known human mitochondrial, glycolytic, and transport reactions were extracted from the manually-curated human metabolic reconstruction (Recon 1)[135]. From this list, reactions were directly added to the brain reconstruction if brain-localized protein or gene expression was suggested by the Human Protein Reference Database (release 5) (HPRD)[165] or H-inv (version 4.3) (HINV)[166], both of which provide tissue expression presence calls for each gene. Proteomic data from live human brain, acquired for the HUPO brain proteome project (BPP) (www.ebi.ac.uk/pride)[167], were also used (See Supplementary Table 7 in [42] for accession numbers). Additional reactions were added as dictated by biochemical data from the literature. Reactions and pathways were manually curated to verify presence in the human brain and to determine cell-type localization, thus yielding a first-draft metabolic reconstruction of the brain metabolic network. Reactions unique to the different neuron types were determined from the literature (see notes in the Supplement of [42]), and consist largely of the reactions needed to make and metabolize their associated neurotransmitters. A list of all reactions, supporting data, citations, and a comparison with previous brain metabolism models can be found elsewhere [42]. Models in SBML format and model updates can be obtained from http://systemsbiology.ucsd.edu/In_Silico_Organisms/Brain.

Constraint-based modeling: Constraint-based modeling and analysis of metabolic networks has been previously described[9, 134]. Briefly, all of the reactions are described mathematically by a stoichiometric matrix, S, of size *m* x *n*, where *m* is the number of metabolites and *n* is the number of reactions, and each element is the stoichiometric coefficient of the metabolite in the corresponding reaction. The mass balance equations at steady state are represented as

$$S \bullet v = 0 ,$$

where v is the flux vector[9]. Maximum and minimum fluxes and reaction reversibility, when known, are placed on each reaction, further constraining the system as follows:

$$v_{\min} \leq v \leq v_{\max}\,.$$

At this point the model can then be used with many constraint-based methods[134] to study network characteristics.

The S matrix was constructed with the mass and charge balanced reactions from the reconstruction. Select metabolites, known to cross the blood-brain barrier, were added as exchange reactions, allowing those metabolites to leave or enter the extracellular space in the model. A few metabolites from network gaps were allowed to enter or leave the system from the cytosol or mitochondria. This was only used when transporter mechanisms or subsequent pathway steps where not known, and when their entrance or removal from the system was necessary for model function. When available, cerebral metabolic rates were used from published data to constrain the upper and lower bounds of the exchange reactions[168, 169]. All parameters are detailed in [42].

Monte Carlo sampling: Monte Carlo sampling was used to generate a set of feasible flux distributions (points). The method is based on the artificially centered hit and run algorithm with slight modifications. Initially, a set of non-uniform pseudo-random points, called warm-up points, is generated. In a series of iterations, each point is randomly moved, always remaining within the feasible flux space. This is done by 1) choosing a random direction, 2) computing the limits of how far one can travel in that direction, and 3) choosing a new random point along this line. After many iterations, the set of points is mixed and approach a uniform sample of the solution space, thus providing a distribution for each reaction that represents the range and probability of the flux for each reaction, given the network topology and model constraints. For more detail, see the Supplementary Notes of [42].

Simulating enzyme deficiencies: Enzyme deficiencies were obtained from the literature[142]. To simulate each deficiency, the distribution for all candidate flux states was determined using Monte Carlo sampling. From this distribution, the most probable flux was found and reduced by the fraction reported in the literature. All candidate states were then recomputed and compared with normal candidate flux states.

Alzheimer's disease microarray analysis: Microarrays were obtained from the Gene Expression Omnibus (GSE5281). Arrays consist of 161 Affymetrix Human Genome U133 Plus 2.0 Arrays that profile the gene expression from laser-capture microdissected histopathologically normal neurons from six different brain regions of Alzheimer's disease patients and age-matched controls. These arrays were not used in model construction.

Arrays were gcrma normalized using the bioconductor package for R. Pearson's correlation coefficients were computed for all array pairs, and arrays with $r < 0.8$ were discarded (i.e., GSM119643, GSM119661, GSM119666, and GSM119676).

Different arrays had different levels of glial contamination. Therefore, to assess the amount of GAD (neuron-specific), the GAD1 and GAD2 levels on each array were normalized as follows. For each array, the relative amount of neuron material was determined by computing a ratio for four neuron-specific genes to the median level across all arrays. Neuron-specific genes were chosen to represent different neuron parts, including the soma, axon, and synaptic bouton (TUBB3[170], NeuN[171], SYN1[172], and ACTL6B[173]). These were summed to compute a relative amount of neuron material (*NM*) for each array, *j*,

$$NM_j = \sum_i \frac{g_{i,j}}{\overline{g}_i},$$

for each neuron marker gene g_i. Since GAD genes are neuron-specific in the central nervous system, these were normalized for each array by the associated relative amount of neuron material, thus termed GAD_{NMN} for neuron-marker normalized GAD. It is assumed in this study that the neuron markers used here do not change their expression level per neuron between Alzheimer's patients and age-matched control, since no published studies have demonstrated that these genes change expression in healthy cells through the progression of Alzheimer's disease. It is possible that there is down-regulation of some neuron markers among neurons bearing neurofibrillary tangles, since synapse loss is a hallmark of Alzheimer's disease[147]. However, the arrays used in this study profile histopathologically normal neurons and the surrounding glial cells. Therefore, it is not expected that there will be significant changes in the expression of these key neuronal genes in the data used here. Lastly, the inclusion of multiple genes from different cell regions helps to minimize the effects from expression changes not attributable to glial cell contamination. The results presented in this work are robust to the removal of each neuron marker gene (See [42] for details).

PathWave analysis: PathWave allows for the elucidation of pathways that significantly change together. Its advantage over other methods, such as Gene Set Enrichment Analysis, is that it takes metabolic network connectivity into account in order to identify changes in pathways.

PathWave was used as published previously[154]. The reactions in each model were subdivided into biologically relevant functional pathways. Reactions that were involved in multiple pathways were added to each associated pathway.

Since PathWave analyzes microarray data based on closely connected reactions, pathways were simplified by removing all metabolites with connectivity greater than eight in the metabolic network. Exceptions are listed in [42]. For each of these simplified metabolic pathways, reactions were laid into a 2-dimensional, regular square lattice grid. To optimally preserve neighborhood relations of the reactions, adjacent nodes of the network were placed onto the grid as close to each other as possible. We mapped each expression data set, obtained from the Gene Expression Omnibus (GSE5281), onto the corresponding reactions of the transcribed enzymes. If a reaction was catalyzed by a complex of proteins, the average expression was taken. The resulting expression values of each reaction were z-transformed. Haar wavelet transforms on the optimized grid representation of each pathway were performed to explore every possible expression pattern of neighboring reactions and to define groups of reactions within a pathway that showed significant differences between samples of different conditions.

To obtain significance values, the sample labels were permutated (n = 10,000) and scores were calculated for each wavelet and permutation. The scores represent the absolute value of the logarithm of the p-value for each wavelet feature, calculated by t-tests. For the best hit (highest score of the non-permutated wavelet features) a p-value was obtained from the reference distributions and represented the significance for the corresponding pathway. The p-value for each pathway was corrected for multiple testing (FDR = 0.05)[174]. Only pathways with more than three significantly differentially regulated reactions were further considered (FDR = 0.05). To obtain local patterns in the pathways, all wavelet features were statistically tested applying t-tests and corrected for multiple testing. Statistically significant features contained those sub-graphs of the metabolic network that showed differentially regulated patterns. Reconstructing these sub-graphs allowed us to directly detect the regions of interest in the metabolic network (see [42] for details).

Identifying pathways for acetylcholine synthesis: An FBA-derived approach was employed to identify all possible pathways coupling the mitochondrial and cytosolic acetyl-CoA pools using known reactions in human metabolism. First, the potential pathways were identified using Recon 1[135]. A reaction that supplies mitochondrial acetyl-CoA was added to the model. A second reaction was added to remove cytosolic acetyl-CoA from the model. Lastly, all other metabolite uptake and secretion constraints were opened. Reactions were randomly removed until a minimum pathway was identified, capable of carrying flux between mitochondrial and cytosolic acetyl-CoA. This was repeated until more than 21,000 unique sets of reactions were identified. An r x p binary matrix was then built with the p unique reaction sets consisting of r reactions. Each element (i,j) of this matrix was 0 if reaction i was absent from pathway j or 1 if reaction i was in pathway j. Rows for all reactions that were never necessary were subsequently removed from the matrix. Singular value decomposition was then used, followed by varimax factor rotation of the first five singular vectors. Singular vector loadings demonstrated the dominant sets of reactions, and their major dependencies that could be used to couple mitochondrial acetyl-CoA metabolism and cytosolic acetylcholine metabolism.

Predicting cholinergic neurotransmission: The percentage cholinergic neurotransmission was computed based on published data[159]. The previously published data were obtained from rat brain minces that were incubated in solutions containing [1-^{14}C]pyruvate or [2-^{14}C]pyruvate. Both acetylcholine and radiolabeled CO_2 were measured at various titrations of several different inhibitors of pyruvate dehydrogenase (PDHm) (see [42] for all inhibitors).

Simulations were conducted using the cholinergic model. The models were allowed to take up the same substrates provided experimentally[159], at rates consistent with the data (see [42] for details).

Monte Carlo sampling was used to identify all feasible flux states. This was done for various levels of PDHm inhibition, ranging from 0 to 90% inhibition. The percentage cholinergic neurotransmission was computed by randomly selecting a feasible flux state from each level of PDHm inhibition and computing the slope of the sum of labeled CO_2-producing fluxes and choline acetyltransferase for the different simulations. A similar slope was computed from randomly sampled points from the reported experimental distributions. The ratio of these slopes represents a feasible percentage cholinergic neurotransmission. This was repeated 1000 times and the median value was reported. Comparisons with the experimental data were done by suppressing the *in silico* pryruvate dehydrogenase flux until the measured CO_2 release rate was obtained. At this level of suppression, the resulting predicted acetylcholine production rate was compared with the experimentally measured rates. See [42] for more details.

Chapter 2, in part, is a reprint of the material as it appears in 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., Palsson, B.Ø. Large-scale *in silico* modeling of metabolic interactions between cell types in the human brain. *Nature Biotechnology*, 28:1279–1285 (2010). I was the primary author, while the co-authors participated in the research that served as the basis for this study.

Chapter 3: Prokaryotes use enzyme post-translational modification to globally regulate metabolism

When cells meet a new environment, genetic programs can be wired to respond. For example, if a bacterium senses an increase in a desirable sugar, the protein sensing the sugar may send a signal through the cell to express the necessary transporter and enzymes. Thus, the cell has genetically-encoded regulatory programs that can behave in a cause and effect manner to defined stimuli.

Among their many regulatory programs in a cell, post-translational modifications are known to modulate the activity of many eukaryotic metabolic enzymes. However, for decades it has been commonly assumed that prokaryotes do not widely utilize PTMs such as phosphorylation and acetylation, except to regulate a handful of enzymes and two component systems. This assumption has been challenged when several recent studies found a large number of proteins with lysine acetylation [175, 176] or succinylation [177] and numerous serine, threonine, and tyrosine phosphorylation sites [178] in *E. coli*. Similar results have surfaced for other prokaryotes and archaea [179-182]. A common finding in these studies is that there is an abundance of protein phorphorylation or acetylation events, and that many of these PTMs are found on metabolic enzymes. Initial biochemical assays have demonstrated that a few of these PTMs may have roles in metabolic regulation [177, 180]. However, the question remains if remaining PTMs on metabolic enzymes also exhibit regulatory functions, or if the regulatory roles are isolated to a few enzymes in central metabolism.

Here we address this question from many angles and demonstrate that prokaryotes employ enzyme acetylation and phosphorylation to regulate metabolic flux in fluctuating nutritional environments. First, we show that PTMs occur on enzymes that require regulation, and that they are complementary to non-covalent metabolic regulation. Second, across most changes in nutrient availability, PTMs are enriched among enzymes at branch points in which flux must be partially diverted from one pathway to another. Third, the PTMs are usually found close to the enzyme catalytic site, suggesting that many PTMs induce changes in protein structure by modifying electrostatic interactions in the enzymes. Lastly, an assessment of enzymes that require regulation suggested different metabolic conditions that may lead to the activation of protein kinases and acetyltransferases, and these predictions were supported by growth phenotypes of mutant *E. coli* in which the associated genes were deleted. Together, all of these lines of evidence clearly show that many protein PTMs in prokaryotes regulate metabolic flux in response to dynamic variations in environmental conditions.

Metabolically-regulated enzymes can be predicted in silico

When cells are confronted with a change in nutritional environment, they often respond by changing their metabolic pathway usage accordingly. Metabolic pathway usage can be approximated using constraint-based modeling, which estimates the steady-state flux through all pathways and reactions in a genome-scale metabolic network [183].

A constraint-based modeling method, called Regulated Metabolic Branch Analysis (RuMBA), simulates the coordinated regulation required to immediately adapt pathway usage in response to fluctuations in nutrient availability. Specifically, RuMBA predicts which enzymes will require immediate regulation to guide the metabolic flux from one growth condition to the steady-state flux needed for the second condition (see Methods and Appendix 3.1 for details). Thus, this method is tailored to predict metabolic regulation, as opposed to other modes of regulation that act at longer time scales, such as transcriptional regulation and protein degradation.

To predict which enzymes will require regulation in a sudden nutritional shift, this method analyzes metabolic flux through each branch point in the network and determines where flux must be diverted from one pathway to another. This diversion of flux by metabolic regulation at a branch point allows the network to rapidly shift toward the desired steady-state flux distribution for the new nutritional environment. Such a mechanism would provide a clear fitness advantage for a few reasons. First, it would waste less energy on transcription and translation in response to transient fluctuations in metabolite concentrations in the microenvironment or within the cell. It would only require that each cell maintain a small number of modifying enzymes. Second, this mechanism would allow for immediate responses to nutritional changes, while transcription and translation can take as long as a generation or two in rapidly growing cells to respond to sudden nutrient changes [184], and likely longer to fine-tune expression [47, 185]. Third, the response could provide immediate fitness improvement for both external environmental changes and internal imbalances resulting from noisy gene and protein expression in each cell.

As an example of how RuMBA predicts sites of metabolic regulation, consider the scenario of metabolism shifting from glucose to acetate. As *E. coli* grows on glucose, fermentation products are usually secreted. As the primary substrate is exhausted, the cells often will change their metabolic network expression to maintain growth on a fermentation product, such as acetate [186]. To rapidly achieve this new steady-state of growth on acetate, the cells may employ metabolic regulatory mechanisms to force the flux at each branch-point toward the reaction that ideally would carry more flux. This can be done through metabolic regulation until transcription and translation can catch up and fine-tune enzyme levels.

One extensively-studied metabolic branch-point relevant to the glucose-acetate shift is the split between the TCA cycle and the glyoxylate shunt [186, 187]. At this flux split, isocitrate is consumed using either isocitrate dehydrogenase (ICDH) to synthesize alpha-ketoglutarate, or isocitrate lyase (ICL) to synthesize glyoxylate (Figure 3.1.a). ICDH is used almost exclusively during glucose metabolism, while it is used less during growth on acetate [186, 187], since ICL is used substantially for anaplerosis. Consistent with this, RuMBA predicts that this branch point must be regulated ($p << 1 \times 10^{-5}$) since there is a significant diversion of flux from ICDH to ICL during the shift from glucose to acetate metabolism (Figure 3.1.b).

How extensively do model-predicted regulation sites coincide with known sites of metabolic regulation? To assess this, 1219 non-covalent metabolic regulation events in *E. coli* were collected and compared with the model-predicted regulation sites. Many of the most significant predictions are known to be regulated (Figure 3.1.c), and far more are regulated than expected by chance (Figure 3.1.d). Allosteric regulation is particularly enriched, which further supports this concept since this class of regulation is more fit for guiding flux towards a new steady state, while competitive inhibition may be more fit for stabilizing a current steady state. Thus, these results suggest that RuMBA can be used to reliably predict which enzymes may require regulation in fluctuating nutritional environments.

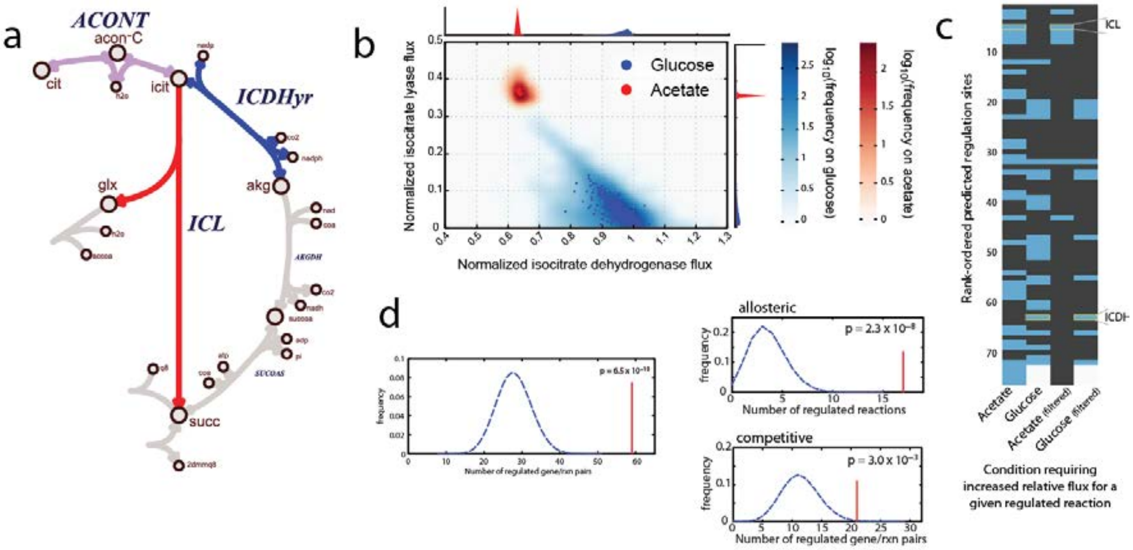


Figure 3.1. Metabolic regulation is important for fluctuating nutritional environments, and metabolically-regulated enzymes can be predicted *in silico*. As external nutrient availability varies in the cell microenvironment or variations in enzyme copy number in cells occur, metabolic regulation can help a cell rebalance its metabolism while the transcription and translation programs are deployed and fine-tuned. (a) This often occurs at branch-points in the metabolic network, such as the branch point at isocitrate, which divides flux between the TCA cycle and the glyoxylate shunt. (b) Randomly-sampled flux distributions for growth on acetate and glucose show that isocitrate dehydrogenase (ICDH) is used almost exclusively for growth on glucose minimal media, while a significant amount of flux is diverted to isocitrate lyase when the cell is metabolizing acetate. (c) RuMBA was used to identify the reactions and their associated enzymes that require significant regulation (i.e., enzymes for which their relative flux at a branch point must increase or decrease in the shift form glucose to acetate). Many of these predictions are enzymes that are known to undergo metabolic regulation (blue;left) and known to be regulated by metabolites that have recently been shown to significantly change in their intracellular concentration between glucose and acetate metabolism (right). (d) In fact, the RuMBA predictions are significantly enriched in known metabolically-regulated enzymes, particularly for allosteric regulation.

PTMs cluster on regulated enzymes in E. coli

In addition to small molecule-mediated metabolic regulation, post-translational modifications (PTMs) have been long known to regulate many eukaryotic metabolic enzymes. However, in prokaryotes, this role for PTMs has only been shown for a few enzymes, because of the dearth of known prokaryotic phosphorylation and acetylation sites. Through advances in proteomics, a few hundred PTMs have been recently identified in prokaryotes, so the question remains if PTM-mediated metabolic regulation is used widely in prokaryotes. A recent report provided evidence that acetylation is important in the regulation of a few enzymes in central metabolism in *Salmonella enterica* [180]; however, it is still commonly assumed that phosphorylation is relevant almost solely to two-component signaling in prokaryotes, and a recent study has suggested that some PTM sites on eukaryotic metabolic enzymes arose after eukaryotes diverged from prokaryotes [188].

If PTMs are more globally relevant to metabolic regulation in prokaryotes, they should be enriched in metabolism, especially among branch-points where flux is shifted between pathways during a substrate shift. Four recent proteomic studies identified a few hundred *E. coli* peptides with serine, threonine, or tyrosine phosphorylation [178], lysine acetylation [175, 176], and lysine succinylation[177]. Of these proteins, 56% are metabolic enzymes represented in the *E. coli* metabolic network [189] (Figure 3.2.a), which is far more than expected by chance ($p = 3x10^{-15}$ when accounting for all proteins and $p < 3x10^{-8}$ when only accounting for proteins with measured gene expression across 12 growth conditions). Furthermore, these metabolic PTMs are also more likely functional. When the conservation of modified residues on the metabolic enzymes is compared to the proteomes of 1057 other prokaryotic genomes, they are more highly conserved than similar non-modified residues on the same proteins (p = 0.0041;rank-sum test).

The immediate question is if these PTMs regulate metabolic flux under variations in environmental condition. When compared to the predicted enzymes requiring regulation for the glucose-acetate shift, enzymes with PTMs are complementary to enzymes with known regulation via non-covalent interactions (Figure 3.2.b). Furthermore, PTMs are enriched among these predicted sites of regulation ($p = 9.4\text{x}10^{-6}$; hypergeometric test), suggesting that PTM-mediated metabolic regulation may play an important complementary role to other modes of regulation.

PTMs are associated with enzymes requiring regulation between many nutritional changes

To test if PTMs are generally associated with enzymes requiring regulation to rapidly adapt to variations in the nutritional environment, we used MCMC sampling, followed by RuMBA to identify enzymes requiring metabolic regulation for 15,051 shifts between 174 different media compositions. Across these 15,051 shifts, enzymes with PTMs are regulated in more substrate shifts than enzymes without PTMs ($p = 6.0\ \text{x}10^{-6}$; Wilcoxon rank-sum test). In addition, in 92% of the 15,051 substrate shifts, PTMs were enriched (hypergeometric test; $FDR < 0.01$) in the RuMBA predictions. When a media shift failed to show significant enrichment of PTMs, this tended to stem from high structural similarity between the primary carbon sources in both media formulations (as measured by Tanimoto coefficients). In these non-significant shifts, the media are often metabolized in a similar manner, and so few branch points will require regulation.

It seems that PTMs are associated with enzymes requiring regulation in fluctuating nutritional environments. Are the PTMs more frequently associated with enzymes that require regulation in more growth condition changes? K-means clustering was conducted on a binary matrix detailing which enzymes need to be metabolically regulated for each substrate pair. In this, four clear clusters of enzymes appear. In the cluster with the highest frequency of predicted regulation, 43% of the enzymes have known PTM sites. In contrast, only 9% of the two clusters with occasional regulation have measured PTMs, and only 3% of the reactions with no predicted regulation are associated with PTMs. These PTMs may be either non-functional, functional outside of metabolism, intermediates in their catalytic function (e.g., metabolite phosphorylation), or examples of limitations in the model's predictive capacity. However, this significant overlap of known PTM targets and model-predicted regulation suggests that PTMs are indeed important for regulating the diversion of flux at important branch points to facilitate the fast transition to the new substrate's steady state flux distribution.

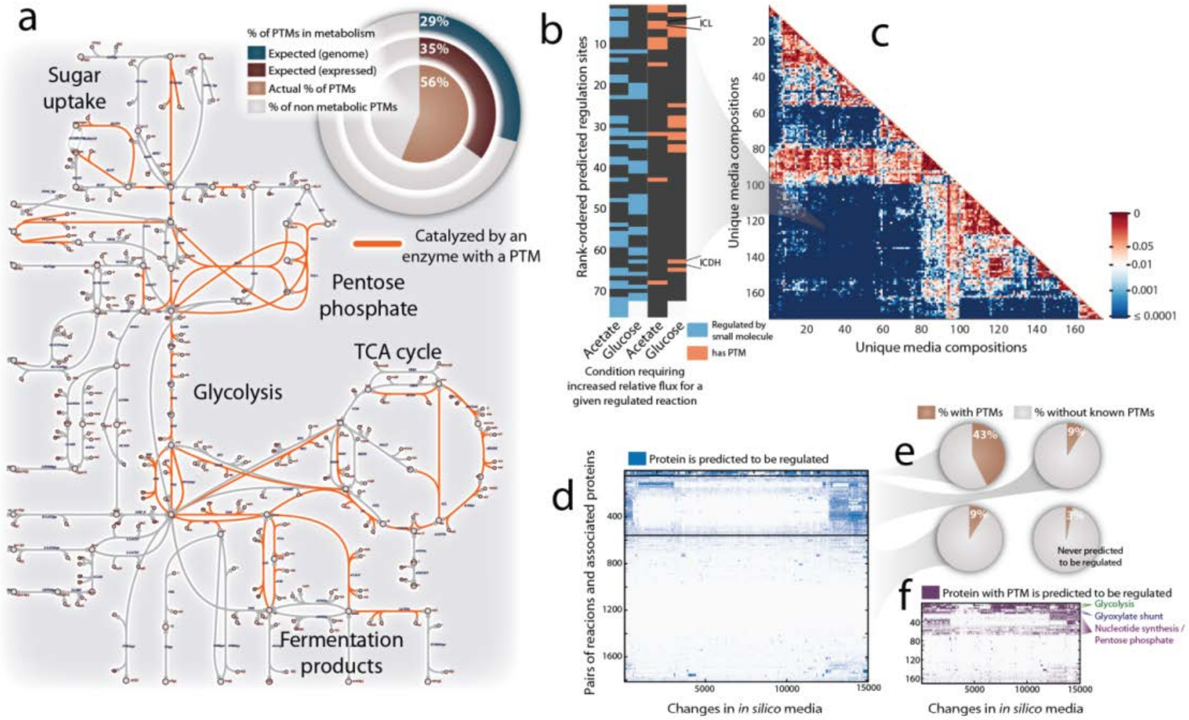


Figure 3.2. Post-translational modifications are associated with enzymes requiring metabolic regulation in the *E. coli* metabolic network. (a) PTMs are associated with many enzymes in central metabolism and other metabolic pathways. 56% of measured PTM proteins are associated with proteins in the *E. coli* metabolic reconstruction. The expected percent when accounting for protein-coding genes is 29%, and 35% when accounting for genes that are expressed across 12 different growth conditions. (b) These PTMs may aid in metabolic regulation by providing a complementary effect, since within the most significant RuMBA regulated enzyme predictions, PTMs (red) complement non-covalent small-molecule regulation (blue) for the glucose-acetate shift. (c) PTMs are enriched in the RuMBA regulated enzymes for most pairwise shifts between 174 different media conditions. (d) All enzymes that were predicted to require regulation for at least one of the 15,051 media shifts were clustered. (e) 43% of the enzymes in the highly regulated cluster have at least one known PTM, while only ~9% of the enzymes in the less regulated clusters are associated with PTMs. Only 3% of the enzymes that are never predicted to require regulation are associated with PTMs. (f) When only RuMBA regulated enzymes with PTMs are clustered, the clusters reveal key pathway features underlying different conditions that may stimulate regulation.

PTMs on in silico-regulated enzymes are more likely functional

The results presented here suggest that many newly-discovered PTMs are associated with enzymes that require metabolic regulation. This would help cells make a rapid transition to a new steady state when there is a sizable fluctuation in nutrient availability. However, it is also possible that some of these PTMs are non-functional since they may pose little or no burden on cell fitness[190]. Thus, the question remains as to if the majority of these PTMs are functional for metabolic regulation or not. This fraction of non-functional PTMs is expected to be low, since only 3% of enzymes that are never predicted to require regulation actually have detected PTMs on them (Figure 3.2.e). A thorough assessment of PTMs in the context of their protein structures provides further support that many detected PTMs are indeed functional.

PTMs occur near active-site residues and alter protein structure

More often, PTMs that regulate metabolism are located near the catalytic site of the enzyme, thereby making conformational changes to the active site or blocking substrate binding. Thus, if many of the PTMs in *E. coli* provide regulatory roles, they should be located near active sites. To test this, all available protein structures for modified proteins were acquired (n = 62), and distances were computed between residues with PTMs and all other residues. These were compared to distances between PTM residues and residues that are known to modulate enzyme activity (i.e, known catalytic sites, residues used for substrate binding, and residues that modify enzyme activity if mutated).

PTMs were significantly closer to residues relevant to enzyme function than amino acids without annotation on 48% of the 62 modified proteins (Wilcoxon rank-sum test; FDR < 0.07) (Figure 3.3.a). Moreover, 37% of the proteins have at least one residue of known catalytic importance within 10Å of the nearest PTM. It is likely that this value is an underestimate, since functional residues are not fully annotated on most of the proteins.

To further assess the regulatory potential of these PTMs at the active sites, we looked to see if they were enriched among the most frequently regulated enzymes, as predicted by RuMBA. Of the 30 enzymes with PTMs nearest to their active sites, 16 were within the top 5% most frequently regulated

RuMBA predictions (i.e., regulated in more than 2617 of the 15051 possible media shifts), which is more than expected by chance (hypergeometric test; p = 0.0062).

In enzymes, PTMs often regulate function by altering protein structure. These changes can alter the binding affinity of substrates or cofactors[191], or affect protein complex formation. In some cases such regulation occurs through steric blocking of binding sites. However, frequently PTMs function by changing electrostatic interactions between residues [188]. For example, some PTMs, such as acetylation can disrupt salt-bridges between lysines and acidic residues. These disrupted salt-bridges would otherwise aid the stabilization of one protein conformation, possibly changing enzyme activity by modifying the catalytic site. Similarly, other PTMs, such as phosphorylation might disrupt or create new salt bridges with basic residues.

Consistent with these properties, many of the metabolic PTMs may disrupt or form salt bridges. In fact, without accounting for modifications in protein structure following post-translational modification, at least 31 of the 62 modified metabolic proteins with known structures have a PTM which would disrupt or add a salt bridge. Of these, 58% of the enzymes are within the 5% most regulated enzymes according to RuMBA, which is more than expected (p = 0.001). Thus it seems that salt bridges may occur on these modified enzymes to modulate protein structure and regulate activity.

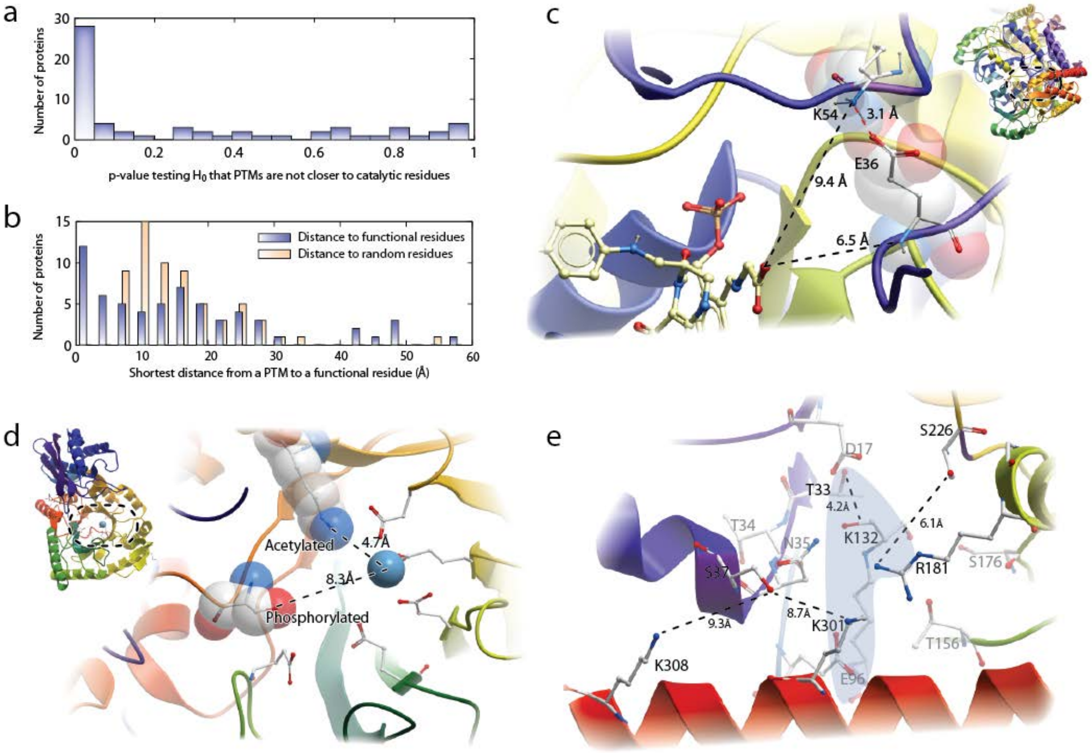


Figure 3.3. PTMs are preferentially located near enzyme active sites. (a) The distance between each PTM and all functional residues (i.e., active site residues or amino acids used for substrate binding) was significantly shorter than expected for at least 48% of the proteins. (b) In addition, 37% of the proteins had a modified residue closer than 10Å to a functional residue. (c) For example, lysine-54 is acetylated in serine hydroxymethyltransferase. This acetylation would likely disrupt a salt bridge with glutamate-36 and possibly affect enzyme activity since these residues are near the substrate binding site. Moreover this PTM may inhibit dimerization since the salt bridge residues are on different subunits of the active dimer. (d) In the enolase active site, serine-371 can be phosphorylated and lysine-341 can be acetylated, which may respectively stabilize or destabilize the Mg^{2+} required for enzyme activity. (e) Transaldolase B has a deep catalytic pocket with several residues contributing to catalysis. In particular, D17, E96, and K132 (highlighted in blue) are important to the reaction mechanism. However, two phosphorylation sites on S37 and S226 will form salt bridges that may occlude the active site, while a third phosphate on T33 will directly block K132.

PTMs and associated salt bridges may inhibit dimerization or alter active sites

The results presented above suggest that many of the measured PTMs may provide a regulatory role since they are often located near functional sites and may often modulate electrostatic interactions. Moreover, PTMs with these properties seem to correlate with RuMBA-regulated enzymes. These can be explained statistically, but can detailed mechanisms be identified for specific proteins? Here we describe details for a few representative enzymes.

Serine hydroxymethyltransferase converts serine to glycine to form 5,10-methylene-tetrahydrofolate, which is an important source of C1 units in the cell. This enzyme can also catalyze a D-alanine transaminase reaction, forming pyruvate from D-alanine. Both of these reactions are predicted to require regulation in ~12% of all possible substrate shifts. This enzyme has several PTMs near its active site. For example, K54 is acetylated in this enzyme (Figure 3.3.c), and this acetylation would likely disrupt a salt bridge with E36. This may subsequently disturb Y55, which is essential for correct positioning of the covalently-attached pyridoxal 5'-phosphate (PLP) cofactor [192], and therefore the acetylation of K54 may possibly disturb enzyme activity. It is also interesting to note that this enzyme functions as a homodimer, and since the aforementioned salt bridge is between residues on two different subunits, an acetylation may also decrease the efficacy of dimerization, which would inhibit catalysis by decreasing the affinity of PLP for the enzyme [193].

Another example of particular interest is enolase. Several modifications have been detected on this enzyme, most of which are near either its RNaseE interaction site, or its glycolytic active site. Two modified residues are relevant to the reversible conversion of 2-phospho-D-glycerate to phosphoenolpyruvate. The active site where this occurs contains a Mg^{2+} ion (Figure 3.3.d). The loss of this ion inhibits the enzyme as enolase rapidly denatures [194]. In the enolase active site, S371 can be phosphorylated and K341 can be acetylated, which may respectively stabilize or destabilize the Mg^{2+} required for enzyme activity. A few biochemical assays support this conclusion, since dephosphorylation [195] of enolase has been shown to inhibit the enzyme, and acetylation increases in *Salmonella* enolase under conditions with lower enolase flux [180]. However, neither of these previous studies identified the modified residue responsible for the inhibition.

One modified enzyme that is predicted to be regulated by RuMBA is transaldolase B, a member of the non-oxidative branch of the pentose phosphate pathway (PPP) that contributes to the reversible link between the PPP and glycolysis. The protein structure of transaldolase B in *E. coli* (b0008) contains a deep catalytic pocket in which the reversible transfer of a dihydroxyacetone moiety to erythrose 4-phosphate is catalyzed to form sedoheptulose 7-phosphate. This reaction employs several residues, but in *E. coli*, the primary residues involved are D17, E96, and K132 [196] (Figure 3.3.e).

Interestingly, four phosphorylated residues have been detected within a few angstroms of the active site. Two of these (S37 and S226) are positioned, such that the phosphorylation may form salt bridges with basic residues, and these salt bridges may occlude the active site of the enzyme. Another phosphorylation site, T33, resides within 4Å of K132, the central active site residue. Therefore, its phosphorylation should directly block enzyme activity.

PTMs may regulate flux responses under general nutritional changes

PTMs are enriched in most substrate shifts and particularly among enzymes in more sensitive branch-points. Moreover, these PTMs are often located near functional residues and may play important roles in blocking active sites, regulating protein stability, and altering complex formation. What causes these modifications? There are not many known protein kinases, phosphatases, and acetyltranferases in most prokaryotes. While some of these likely have activities specific to maybe one or a few proteins [197], it is possible that some modifying enzymes add and remove PTMs in response to changes in specific metabolites or redox states. Furthermore, these proteins may target metabolic enzymes needed to rectify these general metabolite imbalances. For such a situation, it would be expected that many of the reactions experiencing significant diversion of flux will be shared by many substrate shifts. Therefore, the activity of one kinase, for example, might phosphorylate several proteins relevant to the metabolism of one general class of enzymes.

To investigate the possible types of nutritional changes that might modulate the activity of different kinases, phosphatases, and acetyltranferases, the RuMBA predictions involving enzymes with known modifications were analyzed. Through k-means clustering (k=4) of RuMBA results from all shifts, metabolic modules were identified that involve metabolites relevant to specific pathways (Figure 3.2.f). The three regulated clusters are enriched in enzymes that interact with glycolytic intermediates, the glyoxylate shunt, and purine metabolism, respectively (Figure 3.4.a-c). These modules are associated with specific shifts in nutritional environment, and the third module interestingly is associated with shifts that predict high changes in growth rate (Figure 3.4.c).

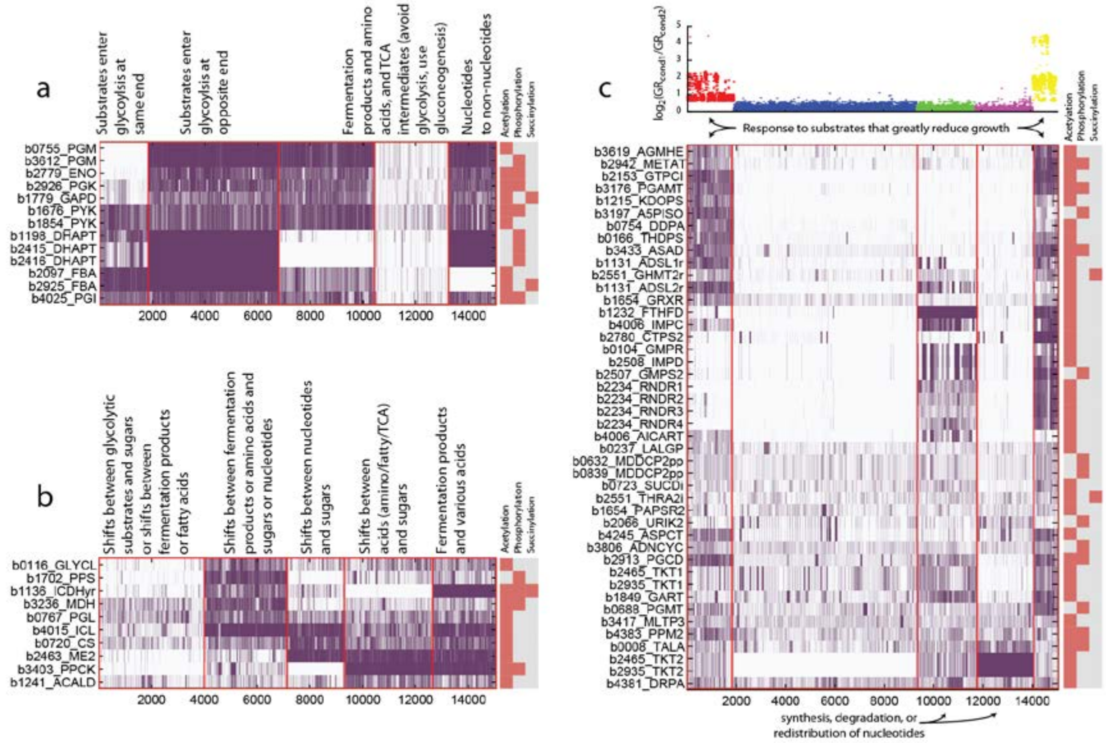


Figure 3.4. Clusters of PTM-associated RuMBA predictions provide insight into environmental shifts that are associated with the regulation of different pathways. Clustering of RuMBA predictions identifies modules with similar regulation patterns, including glycolysis, the glyoxylate shunt, and nucleotide metabolism and the pentose-phosphate pathway. Regulation of nucleotide metabolism is particularly high when model-predicted growth rates significantly change between two media conditions. Within these clusters, (a) shifts that significantly changed glycolysis included shifts between sugars and organic and amino acids. (b) The glyoxylate shunt was usually regulated in shifts between fermentation products and amino acids, sugars, or nucleotides. (c) Nucleotide metabolism and the pentose phosphate pathway were frequently required regulation when shifts involved nucleotides and various acids, and often involved significant changes in growth rates between the substrates.

To see if growth in these different conditions is affected to the removal of kinases, phosphatases, and acetyltransferases, mutant strains of *E. coli* were grown on a few different media that were enriched in the different clusters. Specifically, mutants were grown on M9 minimal media supplemented with glucose, L-lactate, or inosine to cover different classes of carbon sources (i.e., a glycolytic sugar, an organic acid byproduct, and a nucleoside). Mutants grown on these different substrates included Δ*aceK*, Δ*cobB*, Δ*pphA*, Δ*yeaG*, Δ*yfiQ*, Δ*yiaC*, Δ*yihE*, and Δ*ynbD*. Interestingly, several of these mutants showed faster growth than the wild-type strain on a given substrate, but slower growth on another (Figure 3.5). It should be noted, that while many of these showed significantly different growth rates compared to WT, the magnitude of the difference averaged around 7%-9% of the WT growth rate. If these enzymes are primarily for regulating flux under short fluctuations, it is anticipated that transcription regulation will usually balance the growth after several generations, so the effect on growth rate will be small at steady state. Consistent with this, we saw a small but significant difference in growth rate for different enzymes that control post-translational regulation. Thus, even though the targets for these modifying enzymes have yet to be elucidated, they clearly contribute to growth fitness and show preferential response on different media conditions. Furthermore, the removal of these enzymes more often inhibited growth on glucose, but frequently increased the growth rate on L-lactate and/or inosine. Thus, there may likely be a preference for the usage of these enzymes to enhance growth on substrates that confer a higher growth rate, since WT growth is much slower on L-lactate and inosine. Thus, it is possible that they are either modifying metabolic enzymes or modifying other proteins that affect the metabolic rate.

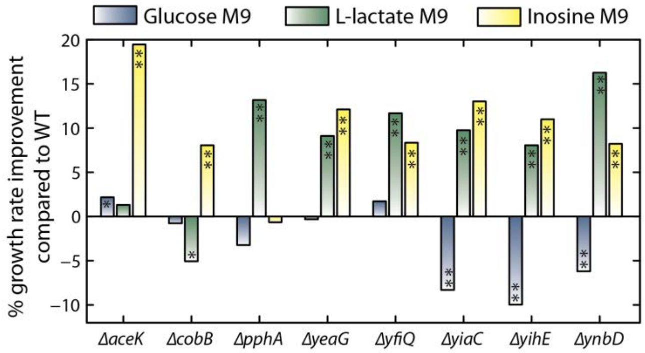


Figure 3.5. Protein kinase, phosphatase, acetyltranferase, and deacetylase mutants show variable fitness with compared to wild type on different media conditions. Mutants were grown on glucose, L-lactate, and inosine M9 minimal media, and many showed decreased or increased fitness on the substrates, suggesting their potential role in regulating metabolism. Difference in growth rate is shown with a significance lower that p = 0.05 (*) or lower than p = 0.01 (**).

PTM regulation, optimality, noise, and proximal causation

In a cause and effect-like manner, cells are programmed to respond to changes in their microenvironment. This has been studied extensively with respect to transcription regulation. In addition, non-covalent allosteric and competitive enzyme regulation has been studied for decades in prokaryotes. Except for a few cases, prokaryotic PTMs have been ignored with respect to metabolic regulation. For example, the general assumption has been that phosphorylation was isolated primarily to two-component signaling. This perception began to change when recent proteomic studies identified a large amount of lysine acetylation and serine, threonine, and tyrosine phosphorylation in multiple bacteria and archaea. Surprisingly, these were usually enriched in metabolic enzymes. However, since then, only a few hand-picked enzymes from the pentose-phosphate pathway and the TCA cycle have been tested for their dependence on PTMs to modulate enzyme activities [177, 180]. Here we have taken a multi-faceted approach to provide support that many of the known PTMs modulate enzyme activity and will be important for regulating flux throughout the *E. coli* metabolic network, especially in dynamic nutritional environments.

Many properties of microbial metabolism have arisen as organisms evolve toward optimality, given their environment and historical contingency. For example, through adaptive laboratory evolution (ALE), it was demonstrated that bacteria will evolve transcription regulatory programs to enhance growth [198-202], improve gene and protein expression efficiency [185], and optimize metabolism [47]. Since metabolism provides the energy and resources for all cell functions, microbes can evolve to enhance metabolic efficiency and reduce waste [47, 185]. However, many of these ALE studies were conducted in well-mixed and static nutritional environments. Therefore, the microbes had adequate time to adapt at the levels of transcription and translation. Such adaptations require many generations to fine-tune [47, 184, 185], and as such, response time is much longer than might be required for microbes in more dynamic natural environments. Different adaptive mechanisms are needed for optimizing growth in natural environments, since the periodicity of micro-environmental changes can be shorter than needed to optimize growth with transcriptional regulation.

At the cellular level, the external metabolic microenvironment and internal expression of metabolic enzymes are dynamic and fraught with noise. Several sources of noise affect the ability of cells to optimize metabolism and growth. Both extrinsic (e.g., fluctuations in metabolite concentration) and intrinsic (e.g., gene and protein levels) sources of noise [203] limit the efficacy of metabolism and decrease the rate of biomass formation [50]. While noise is more carefully controlled for genes catalyzing essential reactions [50], the repertoire of essential genes changes extensively for different growth conditions in fluctuating environments [204, 205]. Moreover, even when redundant pathways exist, some of these can require higher protein costs [47], suboptimal cofactor usage, or decreased catalytic efficiency [206]. Bursting in transcription [207, 208] further complicates this issue, since this process can periodically infuse a non-optimal surplus of an enzyme in a single cell.

Variations in expression level complicate the achievement of optimal expression for a given cell. Thus, it is anticipated that the sub-optimal enzyme expression on the single-cell level will lead high variability in metabolite concentrations internally and possibly yield a variety of byproducts in the external cellular microenvironment. These fluctuations may be stabilized through gene expression in a single cell and through the community, since each cell has its own unique perturbed metabolic network. However, fine-tuning of transcription and translation for a pathway can take more than a generation for exponentially-growing cells [184], and the ability for communities to balance levels of secreted byproducts is limited by diffusion and cell density. Thus, it would be beneficial for the cells to have mechanisms to partially rectify aberrant metabolism and metabolic fluctuations in the cell microenvironment, without having to rely on transcription and translation.

Metabolic regulation plays such a role in reducing metabolic noise to allow for more optimal metabolism. For that reason, numerous feed-back mechanisms have been identified in which products from metabolic pathways can inhibit up-stream enzymes if end-product begins to build up [26]. Such feedback provides a critical role in reducing transient noise in the metabolic network. Enzyme post-translational modification can play a similar role; however, variations in a metabolite of interest would instead stimulate a kinase or acetyltransferase, and this enzyme would subsequently modify its target protein(s). These covalent modifications would allow for more prolonged regulation of target enzymes. Moreover, due to the relatively small number of known protein kinases and acetyltransferases in prokaryotes, we anticipate that these mechanisms are primarily reserved for reacting to more prominent changes in the metabolic environment, such as when *E. coli* moves through the digestive tract from regions in which lactose and arabinose are preferentially metabolized to regions in the intestine where maltose is metabolized [209-211]). Thus, these rapid, more general modes of regulation may provide temporary enhancements of metabolic efficiency while transcription and translation catch up. These rapid responses will provide an immediate boost in fitness. While this only provides a short advantage in each fluctuation, more dynamic conditions and variations in internal enzyme expression [50], including bursting [207, 208], will increase the frequency of metabolic state shifts and thereby compound each fitness boost. Thus, the rapid-response advantage would quickly provide a substantial fitness gain.

One question that remains to be answered here is which kinases, phophatases, and acetyltransferases are modifying the various targets. Unfortunately, little research has been done in this realm, and methods are still being optimized. However, most common prokaryotes have at most a couple dozen enzymes that can add or remove the PTMs. Moreover, the field has not addressed the possibility of how many of these PTMs stem from self-catalysis (e.g., autophosphorylation) or non-enzymatic chemical addition of reactive acetyl- or phopho- moieties. However, irrespective of the mode of addition of these PTMs to the enzymes, our results suggest that many likely regulate metabolism since 1) they are often localized close to active sites, 2) the modified residues are more conserved, 3) the modifications tend to modulate salt-bridge formation, and 4) they occur on enzymes that are predicted here to require regulation.

In conclusion, prokaryotes undergo proximal responses to dynamic metabolic conditions through various regulatory mechanisms. Through a multi-faceted approach, we demonstrated that post-translational modifications play a previously underappreciated role in regulating prokaryotic metabolism. By analyzing proteomic data in the context of protein sequence, enzyme structures, and genome-scale metabolic modeling, it is clear that PTMs are employed to respond to familiar extrinsic fluctuations and intrinsic expression noise with their concomitant variability in metabolite concentrations. Through these mechanisms cells gain a fitness advantage while more costly and slower transcriptional processes catch up. As these mechanisms are further studied and quantitatively assessed, models will be able to more accurately predict the proximal responses of microbial metabolism and growth.

Methods

Acquisition of post-translational modifications and metabolic regulation: Lists of metabolic proteins with post-translational modifications (PTMs) were obtained from studies that identified sites of protein acetylation, phosphorylation, and succinylation in *E. coli* by mass spectrometry [175-178]. All reported occurrences of non-covalent metabolite-mediated metabolic regulation were obtained from Ecocyc [212]. Metabolic regulation events labeled in Ecocyc as allosteric, noncompetitive, uncompetitive, and competitive were used in this analysis to distinguish between different regulatory properties of these enzymes.

Model parameterization: The iAF1260 *E. coli* metabolic model was used with published uptake and secretion rates [189]. A few irreversible reactions were removed because they had reversible duplicates in the model. These include: GLCtexi, URIt2pp, URAt2pp, THMDt2pp, KAT1, INSt2pp, INDOLEt2pp, ICHORSi, CYTDt2pp, and ADNt2pp.

For the 174 simulated media formulations in *E. coli*, glucose uptake was set to zero in the iAF1260 model, and flux balance analysis was used to find which of all other carbon sources could support growth, as reported in the reconstruction of iAF1260. For each of the 174 growth-supporting carbon sources, an uptake rate was set, which was consistent with uptake rate of glucose in the published iAF1260 model (i.e., 8 mmol grDW^{-1} hr^{-1}), normalized by the number of carbons in the metabolite. For example, since glucose has 6 carbons, the uptake rate of glycerol, with 3 carbons, was set as 16 mmol grDW^{-1} hr^{-1} (which is similar to the actual reported glycerol uptake rate in M9 minimal media [213]). While this was used to standardize the media conditions, variations in carbon uptake rates did not significantly impact the results presented in this work.

Markov chain Monte Carlo sampling: The distribution of feasible fluxes for each reaction in the models used here were determined using Markov chain Monte Carlo (MCMC) sampling [35], as previously described [42, 86], and was implemented with the COBRA Toolbox v2.0 [214]. Uptake rates were used to constrain the models as detailed above. To model more realistic growth conditions [87], sub-optimal growth was modeled. Specifically, the biomass objective function (a proxy for growth rate) was provided a lower bound of 90% of the optimal growth rate as computed by flux balance analysis [43]. Thus, the sampled flux distributions represented sub-optimal flux-distributions, while still modeling fluxes relevant to cell growth and maintenance.

MCMC sampling was used to simulate thousands of feasible flux distributions (referred to here as “points”) using the artificially centered hit-and-run algorithm with slight modifications, as described previously [42, 86]. Briefly, a set of non-uniform points was generated. Each point was subsequently moved in random directions, while remaining within the feasible flux space. To do this, a random direction is first chosen. Next, the limit for how far the point can travel in the randomly-chosen direction is calculated. Lastly, a new random point on this line is selected. This process is repeated until the set of points approaches a uniform sample of the solution space, as measured using the mixed fraction metric, which measures uniformity by measuring how many of the sample points pass through the middle line of the solution space [101]. A mixed fraction of approximately 0.50 was obtained, suggesting that the space of all possible flux distributions is nearly uniformly sampled.

Regulated Metabolic Branch Analysis: Regulated Metabolic Branch Analysis (RuMBA) provides a list of enzymes and reactions that may need to be metabolically regulated to immediately adapt to a fluctuation in the nutritional environment. See Appendix 3.1 for a detailed discussion and a validation of the method.

Markov chain Monte Carlo sampling of the metabolic solution space is used to obtain a uniformly distributed assessment of feasible flux values each reaction can have at steady state. Subsequently, flux through each branch point metabolite in the network with a connectivity less than 30 is assessed. For each metabolite, all reactions that can produce or consume it are identified. For each MCMC sample point in the solution space, all incoming fluxes are summed up, as are all outgoing fluxes. Then, for each i^{th} reaction, the fraction of total flux through the metabolite, v_{met}, that is contributed by the reaction of interest, is computed as follows:

$$f_i = \frac{v_i}{v_{met}},$$

where v_i is the flux through reaction i and f_i is the fraction of all flux passing through the metabolite of interest, that is passing through reaction i. Since this is done for many random feasible sets of flux values through all of the reactions at the branch point, a distribution of f_i fractions is computed for each reaction for the two growth conditions of interest. A p-value is computed that measures the overlap of the f_i values for that reaction under the given growth condition, i.e., the probability of finding an f_i value in the first growth condition that is equal to or more extreme than an f_i value for the same reaction in the second growth condition. The p-values are subsequently corrected for multiple hypotheses (FDR < 0.01).

A small fraction of reactions can show miniscule, but significant changes due mostly to slight differences in predicted growth rates. Thus, the list of the regulated reactions and their associated enzymes is filtered to focus on the more significant results. Reactions that change their predicted flux level by less that 50% are filtered out from the list of reactions requiring regulation. This was done by simulating changes in reaction flux occurring in a shift between two conditions, as done previously [78, 86]. The distributions of sampled fluxes for each reaction were compared between two media conditions. First, flux magnitudes were normalized between each pair of media conditions (media *A* and *B*). To do this, a ratio of total flux through the metabolic network was computed and used to normalize each sample point. To compute this ratio, each sample point was taken and the magnitudes of all *n* non-loop-associated reaction fluxes were summed to acquire a value for the total network flux. For both media conditions, the median total network flux was taken and used to normalize each reaction flux for all sample points in media B, as follows:

$$v_{i,j,B}^{*} = v_{i,j,B}\frac{median(\{\sum_{r=1}^{n}|v_{r,1,A}|,\ldots,\sum_{r=1}^{n}|v_{r,j,A}|,\ldots,\sum_{r=1}^{n}|v_{r,p,A}|\})}{median(\{\sum_{r=1}^{n}|v_{r,1,B}|,\ldots,\sum_{r=1}^{n}|v_{r,j,B}|,\ldots,\sum_{r=1}^{n}|v_{r,p,B}|\})},$$

where $v*_{i,j,B}$, is the normalized flux through reaction *i* in sample point *j* under media condition *B*, obtained after multiplying the sampled flux $v_{i,j,B}$, by the ratio of the median total flux magnitude for the reaction for all *p* sample points under growth on medium *A* to the median total flux magnitude for the reaction for all *p* sample points under growth on medium *B*.

Once the flux values were normalized, the changes of fluxes between two conditions were determined as previously described [86]. Briefly, calls on differential reaction activity were made when the distributions of feasible flux states (obtained from MCMC sampling) under two different conditions did not significantly overlap. For each metabolic reaction, a p-value was obtained by computing the probability of finding a flux value for a reaction in one condition that is equal to or more extreme than a given flux value in the second condition. Significance of p-values was adjusted for multiple hypotheses (FDR ≤ 0.01). When the magnitude of flux changed less than 50% of the initial flux magnitude, these reactions were filtered out from the set of predicted regulation sites and excluded from further analysis. However, results were robust for a wide range of filter levels.

Clustering of reaction changes: An *m* x *n* matrix with *m* gene-reactions pairs (predicted to be regulated in at least one media shift; m = 1814) and *n* total media shifts (n = 15,051) was made, detailing in which shifts each gene-reaction pair is predicted to require regulation (FDR < 0.01). The gene-reaction pairs were subjected to k-means clustering (k = 3). Clustering was repeated 100 times with different seed values to find consensus clusters.

Determination of expressed genes: Expression profiles were obtained from previous studies [114, 215-217]. The Affymetrix CEL files were normalized using gcrma, implemented in R. Genes were considered not expressed if they did not have a mean expression level across biological replicates that were significantly higher than the five highest-expression non-*E. coli* negative control probe sets on the array (1-tail t-test; FDR = 0.05). The sets of expressed genes from each study were used to estimate the number of expressed proteins.

Residue conservation: All protein sequences of 1057 prokaryotic species were acquired from the KEGG database. Homologs to all *E. coli* proteins containing at least one known PTM were identified by using the Smith-Waterman algorithm. When more than two proteins in one species had the same percent identity, the protein with the lowest e-value was chosen. In the rare case in which multiple proteins from a species had identical % identity scores and e-values, all qualifying proteins were included.

Each metabolic *E. coli* protein with a PTM (n=109) was thus grouped with its homologs, and the pair-wise Smith Waterman alignment between the individual *E. coli* protein and each of the homologs was used to quantify the conservation of post-translationally modified residues, as calculated (i.e., the percent of pair-wise comparisons where the aligned residue was identical in the homolog). Conservation of non-modified residues for these amino acids was calculated in an identical fashion.

Salt bridge prediction and measurement of distance from PTMs to active site residues: Protein structures for modified enzymes were obtained from the Protein Data Bank. Potential salt bridges that could be disrupted by a PTM were determined by finding all residues within 4Å of a lysine or serine that could form a salt bridge. Potential new salt bridges were found by searching for basic residues within 8Å of a phosphorylated serine, threonine, or tyrosine.

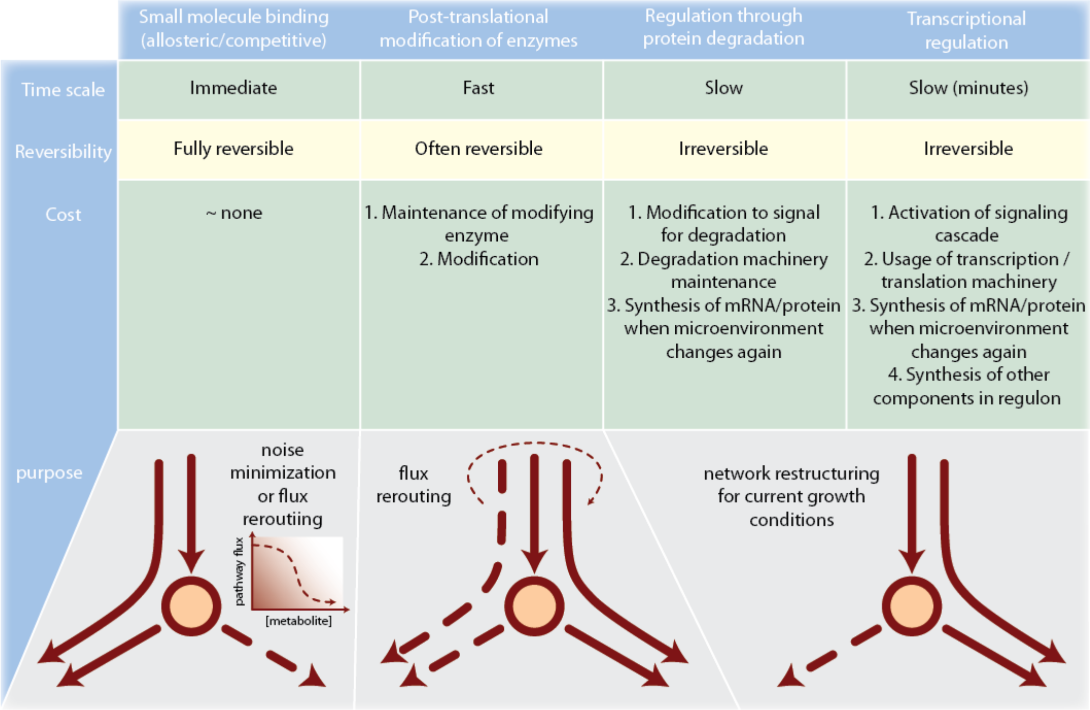
Distances between modified residues and all other amino acids were calculated between centroids of each amino acid. These were used to compare distance between random residues and modified residues with distances between modified residues and functional residues. Functional residues are defined as active sites on proteins, substrate binding sites, and residues which modulate enzyme activity if replaced, and were all acquired from Ecocyc, Uniprot, and the literature.

Mutant growth assays: Wild type *E. coli* and several mutants missing kinases, phosphatases, or acetyltransferases (Δ*aceK*, Δ*cobB*, Δ*pphA*, Δ*yeaG*, Δ*yfiQ*, Δ*yiaC*, Δ*yihE*, and Δ*ynbD*) were obtained from the Kieo collection [218]. Gene deletion was verified by PCR of the scar region, and strains were subsequently grown overnight M9 media, supplemented in 2g/L glucose, L-lactate, or inosine in a seeding culture. An aliquot of culture was returned to fresh media such that the OD600 was ~0.03. Cultures were subsequently grown at 37°C with constant stirring. Turbidity was periodically measured at OD600 as a proxy for cell count, and growth rates were computed from OD measurements at mid-exponential phase.

Appendix 3.1: A detailed assessment of RuMBA

Metabolic regulation is a rapid means to redirect flux in a metabolic network, while transcriptional regulation and regulation of enzyme abundance are processes that act on a longer time scale. Therefore, it is expected that following a shift to a new growth condition, allosteric regulation and post-translational enzyme modification will redirect flux at important branch points. The rational for this response is that, *in vivo*, there are regular fluctuations in the cellular microenvironment and frequent environmental changes [209-211]. Thus, it would be advantageous for the cell to have a means to rapidly regulate metabolic pathway usage using reversible mechanisms while slower and more permanent regulatory mechanisms are being activated. The relative costs and timescale of a few types of regulation are listed below in Table 3.1.

Table 3.1. Potential responses to changes in media composition in the microenvironment.



	Small molecule binding (allosteric/competitive)	Post-translational modification of enzymes	Regulation through protein degradation	Transcriptional regulation
Time scale	Immediate	Fast	Slow	Slow (minutes)
Reversibility	Fully reversible	Often reversible	Irreversible	Irreversible
Cost	~ none	1. Maintenance of modifying enzyme 2. Modification	1. Modification to signal for degradation 2. Degradation machinery maintenance 3. Synthesis of mRNA/protein when microenvironment changes again	1. Activation of signaling cascade 2. Usage of transcription / translation machinery 3. Synthesis of mRNA/protein when microenvironment changes again 4. Synthesis of other components in regulon
purpose	noise minimization or flux reroutiing	flux rerouting	network restructuring for current growth conditions	

Two methods have been developed to predict which enzymes will require significant changes in activity level following a change in carbon substrate for shorter and longer timescales. Tentatively, I call these RuMBA and FSS, respectively.

A variant on FSS has been used previously [78, 86]. Another method similar to FSS has also been recently published, showing its conceptual accuracy [37]. A brief discussion of this method provides a conceptual basis to understand RuMBA. Constraint-based modeling, the framework upon which both RuMBA and FSS are based, uses the metabolic network topology to define a space of possible phenotypes by adding a series of known biologically-relevant governing constraints (e.g., uptake rates for media components, byproduct secretion rates, growth rates, etc.). This space of possible phenotypes represents all possible combinations of metabolic steady-state pathway usage that a cell can use in the given growth conditions. Assuming the constraints are accurate, the actual steady state flux distribution (or pathway usage) should be within the *in silico* solution space (Figure 3.6.a). The range and distribution of flux through each reaction within these solution spaces are dependent on the constraints, such as reaction thermodynamics, metabolite uptake rates, etc. Therefore, the space is condition-specific, i.e., the various dimensions of the space might move when the model is simulated under two different growth conditions. For example, as shown in Figure 3.6.b-c, the flux may be significantly higher in the second growth condition (reaction 2), or show no significant change between the two growth conditions (reaction 1).

The predicted changes in pathway use from FSS represent the changes that lead to the optimal pathway usage in different growth conditions. However, to achieve this optimality, the activity of numerous enzymes must be fine tuned, and often, many proteins need to be up-regulated to meet this requirement. These adjustments require significant changes in transcription and translation, which can take a generation or two for entire pathways.

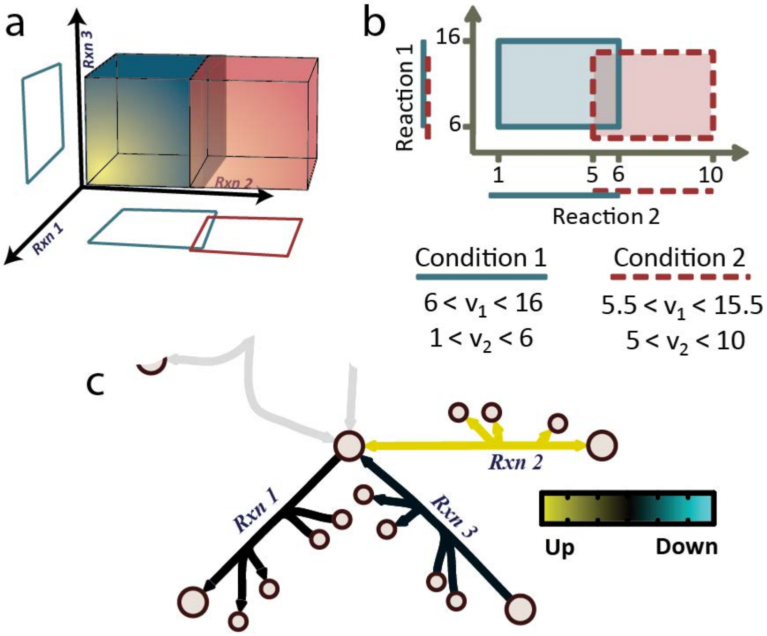


Figure 3.6. Condition-specific shifts in the flux solution space. (a) Constraint-based modeling employs governing constraints to define a space of feasible phenotypes, which are represented by allowable steady-state fluxes for each reaction. When growth conditions change (e.g., a change in carbon source, or aerobicity), the space of feasible fluxes can change. (b) For example, reaction 2 shows a change in the range of feasible flux levels under the new growth condition, which can be shown in the metabolic map (c). These changes can be mapped back to the genes and proteins associated with each reaction.

On a shorter time scale, when changes in enzyme level are either less efficient (e.g., protein degradation) and/or not feasible to obtain, a more reasonable adaptive response involves a temporary suppression of the activity of an enzyme to avoid sending metabolites down less efficient pathways, or to boost the activity of present enzymes that will be needed in the new growth conditions. Thus regulation at metabolic branch-points becomes of great importance, so that metabolites can be shuttled down the most efficient pathways.

RuMBA leverages this idea to compute the shift of the solution space for short-time scale changes in metabolic pathway activity at metabolic branch points. To do this, Markov chain Monte Carlo sampling of the metabolic solution space is used to obtain a uniformly distributed assessment of feasible flux values each reaction can have at steady state.

To assess each branch point metabolite in the network, all reactions that can produce or consume it are identified. For example, aconitase produces isocitrate, while isocitrate dehydrogenase and isocitrate lyase both consume it (Figure 3.7.a). For each sample point in the solution space (Figure 3.7.b-c), all incoming fluxes are summed up, as are all outgoing fluxes. Then, for each i^{th} reaction, the fraction of total flux through the metabolite, v_{met}, that is contributed by the reaction of interest, is computed as follows:

$$f_i = \frac{v_i}{v_{met}},$$

where v_i is the flux through reaction i and f_i is the fraction of all flux passing through the metabolite of interest, that is passing through reaction i. Since this is done for many random feasible sets of flux values through all of the reactions at the branch point, a distribution of f_i fractions is computed for each reaction for the two growth conditions of interest (Figure 3.7.d). Therefore a p-value can be computed that measures the overlap of the f_i values for that reaction under the given growth condition, thus quantifying how significantly the flux changes from one enzyme to another when environmental conditions change. The function of a phosphorylation event can subsequently be predicted if the change in phosphorylation is also known.

To test this method, three *E. coli* enzymes were identified (in the literature) that undergo differential protein phosphorylation between growth on glucose and acetate. RuMBA was employed to predict the effect of phosphorylation on these three enzymes (Figure 3.7.e). At late log phase, enolase has been shown to have seven times higher phosphorylation when *E. coli* was grown on glucose than when grown on acetate [195]. *In silico*, RuMBA predicts that enolase will have a reduced flux level on acetate. Therefore, one may predict that the phosphorylation event would activate its forward flux. It was determined that when treated with acid phosphatase, enolase was inhibited [195]. Similarly, RuMBA predicts that on acetate, the flux through isocitrate dehydrogenase (ICDHyr) decreases, while the flux through isocitrate lyase (ICL) should increase. Experimentally, the phosphorylation of ICDHyr increases and may increase for ICL (phosphorylation is high when grown on acetate, but has not been rigorously tested on glucose). Thus, it is predicted that phosphorylation of ICDHyr inhibits enzyme activity, while it activates ICL. Both of these predictions are consistent with published data [219, 220].

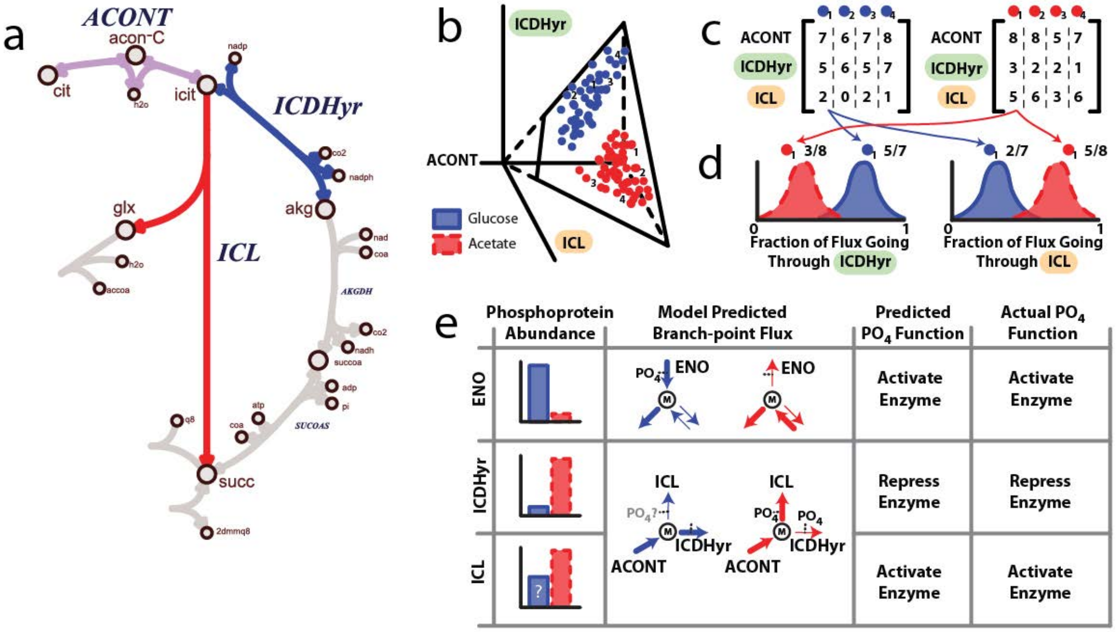


Figure 3.7. RuMBA accurately predicts the known metabolic regulatory function of protein phosphorylation. RuMBA computes the shift of flux from one pathway to another when growth conditions are changed. At metabolic branch points, such as the split between isocitrate dehydrogenase (ICDHyr) and isocitrate lyase (ICL), RuMBA predicts that different branches will be used under different growth conditions. (b) When *E. coli* is grown on glucose, RuMBA predicts that most of the flux through aconitase (ACONT) continues in the TCA cycle through (ICDHyr). However, when switched to acetate minimal media, a significant amount of flux is siphoned off into the glyoxylate shunt through ICL. (c) RuMBA predicts this shift by using MCMC sampling to compute a uniform sample of feasible steady-state flux values (points) for all reactions that produce or consume a metabolite of interest, such as isocitrate. (d) The fraction of flux that goes through each branch is computed for each point, yielding a distribution of fractional split values. (e) RuMBA results can then be compared to experimentally measured flux values for the given growth conditions, yielding predictions for the metabolic regulatory function of phosphorylation events. RuMBA accurately predicts experimentally measured effects.

Chapter 3, contains some material from Lewis, N.E., Chang, R.L., Kim, D., Hefzi, H.H., Palsson, B.Ø. Prokaryotes use enzyme post-translational modification to globally regulate metabolism. *In preparation*. I was the primary author, while the co-authors provided support in the research that served as the basis for this study.

Section II: Distal causation in the evolution of proximal capabilities

Proximal causation is limited by physicochemical constraints, such as the conservation of mass and thermodynamic laws. However, proximal causation in biological systems is also constrained by their genetic programs. These genetic programs often evolve to improve an organism's fitness and its likelihood to reproduce. Thus, evolution is considered to be a driven by factors of distal causation in biology. One does not have to travel far for examples of distal causation shaping an organism's genetic program, thereby further shaping limitations and capabilities of proximal responses.

The following chapter provides one such example of how selective pressures have shaped the catalytic promiscuity of microbial enzymes. Importantly, by using constraint-based modeling to integrate and analyze various high-throughput data types, it is clear that there is an interplay between metabolic network context, environmental conditions, and the need to generate biomass, which imposes distal effects on the evolution of enzyme function and ultimately the metabolic capabilities of microbes.

Chapter 4: Network context and selection in the evolution to enzyme specificity

It is a widely held view that ancestral enzymes had low substrate specificity and catalytic efficiency [221]. Through mutation, duplication, and horizontal gene transfer, gene families diversified and promiscuous enzyme functions were refined to exhibit specific and more efficient catalytic abilities [222]. While several models of this evolutionary process have been suggested [222-224], it is clear that these evolutionary mechanisms continue to refine catalytic activities of enzymes today as organisms continually adapt to environmental changes in nature [223, 225] and in the laboratory [225-228].

In contrast to this view in which proteins continuously evolve towards absolute-specificity (i.e., catalyzing one physiologically relevant reaction in an organism), increasing evidence shows that most enzymes actually display non-physiological promiscuous activities, and that many are multi-specific, where multiple physiologically important are catalyzed by one enzyme [222]. Thus, a fundamental question arises: why would some enzymes evolve to absolute-specificity, while others maintain the promiscuous and multi-specific, characteristics (Figure 4.1)?

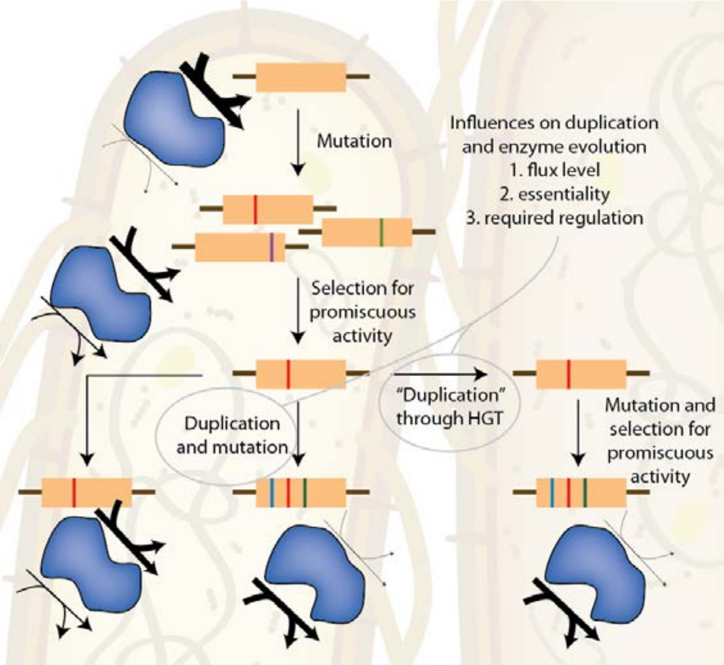


Figure 4.1. Evolution of enzyme specificity through the divergence prior to duplication model. When an organism is introduced into a new environment, an existing enzyme with promiscuous activity may be mutated or amplified if one promiscuous activity provides a beneficial function in the new environment. This results in a multi-specific enzyme, carrying multiple physiologically-relevant activities. It is believed that the difficulty of tuning multiple enzymatic activities on one enzyme drives the duplication (or horizontal gene transfer (HGT)) and further refinement of individual enzyme activities. Why, then, are there so many multi-specific enzymes? Here it was found that the network context, environment, and the need to synthesize biomass all impose stronger selection for absolute-specificity when enzymes maintain a higher flux, are more essential, and/or require more regulation of flux.

This question is addressed here through an *in silico* metabolic network analysis, fortified with new and published experimental data. We demonstrate striking differences in how hundreds of absolute-specific and multi-specific enzymes are used in the context of the entire metabolic network of *E. coli*. Moreover, these characteristics are found to be general properties of metabolic networks across archaea, bacteria, and eukaryotes suggesting that enzyme usage in the network context may influence the evolution of enzyme specificity.

Multi-specificity is abundant in microbial metabolism

How extensive is multi-specificity amongst known metabolic enzymes? A carefully curated comprehensive reconstruction of the *E. coli* metabolic network has been developed, validated and extensively used over the past 20 years [5, 189]. We use it here to study enzyme specificity in a network context. The reconstructed network includes the biochemical functions of 1,260 genes (36% of the functionally annotated ORFs), 1,147 proteins and complexes, and 2,382 reactions [189]. Moreover, >92% of the biochemical functions of the gene products in the reconstruction have been experimentally determined [189] and studied in more than 61,727 published studies.

In this reconstruction, 671 enzymes are absolute-specific and catalyze 450 metabolic reactions, while 410 multi-specific enzymes catalyze 866 metabolic reactions. Enzyme classification is not significantly influenced by depth of study on specific enzymes, since the degree of multi-specificity does not correlate with knowledge depth of the enzymes, and several deeply studied pathways actually contain fewer multi-specific enzymes than expected by chance. Interestingly, multi-specific enzymes catalyze 66% of the non-spontaneous metabolic reactions in the network (termed here “multi-specific reactions”), and more than 80% of these can be active in common growth conditions. Thus, contrary to the general view of enzymes, multi-specificity plays a prominent role in *E. coli* metabolism.

Absolute-specific enzymes carry a higher flux load

It is possible that multi-specific enzymes are retained where there is weaker selection based on the enzyme usage in the metabolic network. Higher demands on enzyme usage may provide an evolutionary selective pressure to enhance catalysis and reduce the required quantity of enzyme. Absolute-specificity may also be selected since catalytic improvements for one substrate on a multi-specific enzyme can decrease the efficiency of its other catalytic activities [228]. Thus, is absolute-specificity selected based on enzyme usage? While technical challenges limit the resolution and scope of experimental flux determination, genome-scale metabolic networks can estimate reaction flux after being converted into computational models [183]. Importantly, these estimates have shown consistency with smaller-scale ^{13}C-based studies [97] and various -omic datatypes [47]. Thus, we computed steady-state flux loads for E. coli, using a Markov-chain Monte Carlo (MCMC) sampling method [36]. This approach allows the simulation of flux for suboptimal growth, thereby better reflecting growth in nature. Flux loads were computed under 174 different media compositions in which E. coli can grow. For each growth condition, the median flux load for each reaction was rank-ordered to determine the relative flux loads between reactions.

Across all simulated growth media, absolute-specific reactions tend to maintain a higher flux than multi-specific reactions (Figure 4.2.B; $p = 1.73 \times 10^{-43}$). Thus, enzyme specialization may allow for an increased specific activity of high-flux enzymes, thereby partially offsetting the cost of gene duplication [229] by requiring fewer enzyme molecules to maintain the required flux load. While in general, absolute-specific reactions carry a higher flux load the existence of low-flux absolute-specificity suggests that additional drivers for protein evolution exist. An assessment of these low-flux absolute-specific reactions shows an abundance of enzymes that synthesize essential cell components (e.g., cofactors and prosthetic groups).

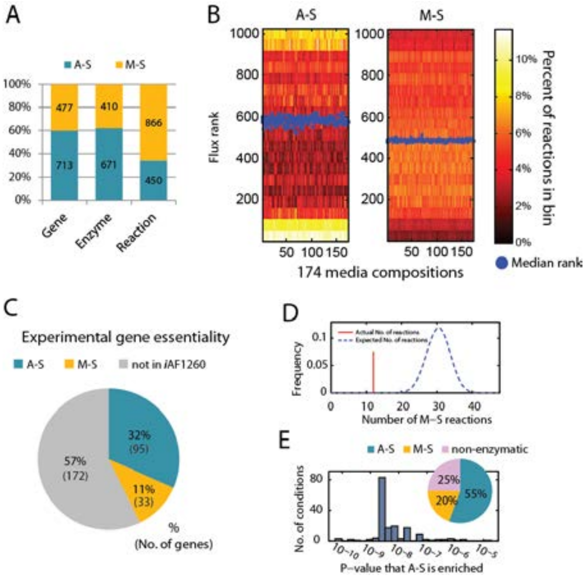


Figure 4.2. Flux level and essentiality correlate with enzyme specificity. (A) The number of absolute- and multi-specific genes, proteins, and reactions in E. coli metabolism. (B) Reaction flux magnitudes are rank-ordered and binned for 174 different media conditions. Color intensity shows the percentage of reactions within the given flux rank range. (C) A higher percentage of essential genes *in vivo* are absolute-specific. (D) *In silico* growth (i.e., biomass function flux) on glucose minimal medium is predicted to be dependent on few multi-specific reactions. (E) For all 174 simulated growth conditions, absolute-specificity is significantly enriched among in silico-predicted reactions essential for growth, representing 55% of the essential reactions (inset). A-S: absolute-specific, M-S: multi-specific.

Absolute-specific enzymes are more essential

Thus, to carefully regulate the synthesis of cofactors and other essential cell components, are the absolute-specific enzymes more essential for cell growth than multi-specific enzymes? Absolute-specific enzymes are significantly enriched among experimentally-determined essential genes [218] ($p = 8.65 \times 10^{-5}$, Figure 4.2.C). In addition, *in silico* simulation demonstrates that cell growth is rarely directly dependent on multi-specific reaction flux (Figure 4.2.D), while absolute-specific reactions are more frequently essential for growth across all 174 tested media conditions (Figure 4.2.E). Possibly, the abundance of absolute-specificity among essential enzymes stems from selection to limit substrate competition in the synthesis of necessary biomass components.

Absolute-specific enzyme flux varies more in fluctuating environments

Natural environments are dynamic and metabolite concentrations fluctuate considerably at the microbial scale [230]. As nutrient concentrations change, gene essentiality [204] and reaction flux also vary [37, 97]. Thus, in view of the flux load and essentiality results, does the need to control variations in flux in dynamic environments induce selective pressures against multi-specificity? Changes in pathway usage in dynamic environments were predicted by simulating several shifts in carbon, oxygen, and nitrogen sources for *E. coli*, using experimentally measured phenotypic data to parameterize the model. For each substrate shift, the model predicted whether metabolic reaction flux should increase or decrease. To provide some support for these predictions, we compared them with microarray data, obtained for each condition (both new data and from previous studies [215]), and found that these predictions are consistent with measured differential gene expression.

Across all shifts in media, there is a considerable difference in the percentages of active absolute-specific and multi-specific reactions that significantly change their flux. In most substrate shifts, absolute-specific reaction flux is more than twice as likely to change than multi-specific reaction flux. This result is robust for more stringent classifications of multi-specificity (Figure 4.3.A). Thus, flux through absolute-specific reactions is considerably more sensitive to environmental change, while multi-specific reaction flux varies less.

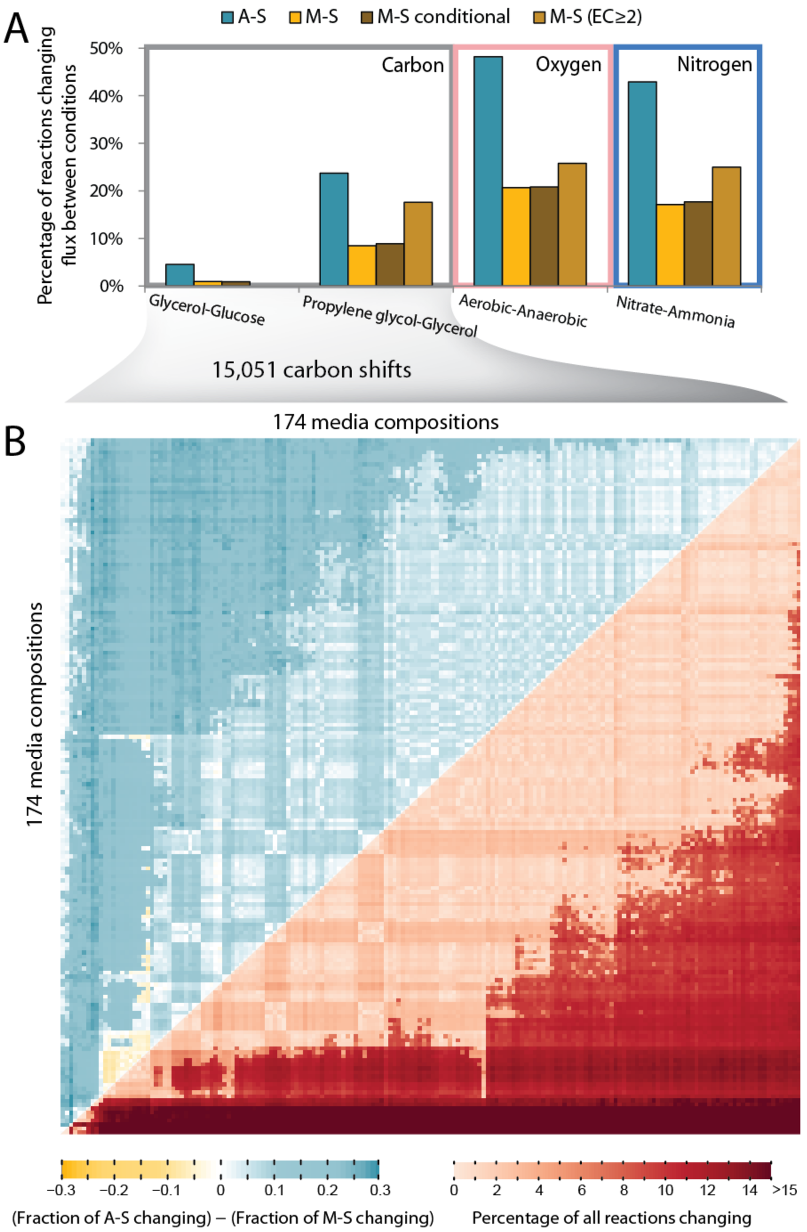


Figure 4.3. Specificity correlates with differential flux in dynamic environments. (A) Experimental-data were acquired and used with the model to predict the percentage of reactions that change in four nutritional shifts. The percentage of reactions changing is reported here for the following reaction classes: absolute-specific (green), multi-specific (yellow), conditional multi-specific reactions (brown, see SOM for details), and multi-specific with more than two E.C. numbers (tan). (B) A systematic computational screen of all 15,051 possible shifts between 174 carbon substrates shows that absolute-specific reactions tend to change more frequently (upper), and that this difference is particularly clear for shifts that cause more reactions to change (lower).

To show that this is a general property of E. coli metabolism, 15,051 pairwise environmental shifts were simulated. In 96% of these shifts, absolute-specific reactions change more frequently than multi-specific reactions (Figure 4.3.B, upper). This property is especially apparent for environmental shifts that induce more than ~5% of the model reactions to change flux (Figure 4.3.B, lower). Since absolute-specific reactions are subject to greater flux changes in nutritionally dynamic environments, it is anticipated that these environmental fluctuations may guide evolution towards absolute-specificity to allow for focused enzyme regulation associated with sensitive flux levels.

Absolute-specific enzymes are subject to more metabolic regulation

Does this more selective variation in absolute-specific enzyme flux result in more regulation of their activity than multi-specific enzymes? Metabolic regulation of enzyme activity is commonly mediated through metabolite-protein interactions or post-translational modifications (PTMs) [177, 231]. To quantify the prevalence of metabolic regulation, a few hundred known metabolite-mediated regulatory interactions and enzyme PTMs were identified for *E. coli* and assessed with respect to absolute- and multi-specific enzymes. Allosteric, uncompetitive, and noncompetitive regulatory interactions are enriched in absolute-specific enzymes (p-value = $9x10^{-4}$), as are PTMs (p-value = $5x10^{-3}$). Metabolic regulation is depleted among multi-specific enzymes, presumably reflecting the decreased need to change flux through these reactions in varying environments. Moreover, when their flux does change, it tends to do so in the same direction, thereby negating the need for more complex regulation.

To further assess the association of specificity with regulation, we quantified how frequently each reaction changes flux across all 15,051 media shifts. K-means clustering identified three dominant reaction clusters (Figure 4.4.A). Two clusters show frequent changes in flux, and these are enriched in absolute-specific enzymes (particularly associated with central and amino acid metabolism), while the reaction cluster with few flux changes is significantly enriched in multi-specific enzymes (Figure 4.4. B-C). Metabolic regulation is significantly enriched within the cluster experiencing the most change, but depleted from the cluster with few flux changes (Figure 4.4.D-E). Thus, by inference, there exists a pressure to enhance enzyme specificity for reactions that require more careful regulation of enzyme activity in order to control fluxes that are more sensitive to dynamic nutritional environments.

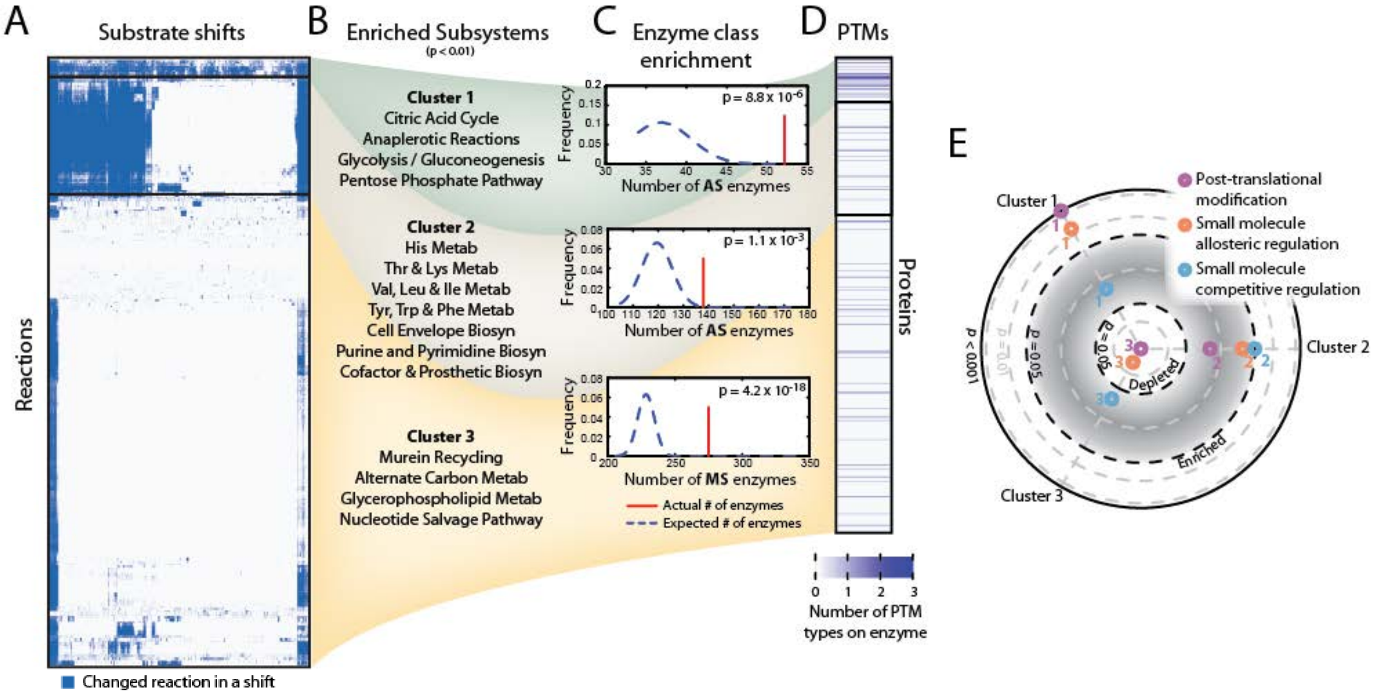


Figure 4.4. Specificity is associated with more regulation. (A) Clustering reactions that change (blue) or not (white) across 15,051 different media shifts yields three distinct clusters. (B) Each cluster is enriched in unique metabolic subsystems. (C) Absolute-specific genes are enriched in more sensitive clusters, while multi-specific genes are enriched in the cluster with few flux changes. (D) The number of PTMs (acetylation, phosphorylation, and/or succinylation) on enzymes increases with sensitivity of clusters. (E) Post-translational modifications and metabolite-mediated allosteric regulation are enriched among the sensitive cluster, and depleted in the insensitive cluster.

Functional properties of enzyme specificity are conserved

The aforementioned results are properties of how the *E. coli* metabolic network functions as a whole. However, if these properties influence selection of enzyme specificity in protein evolution, one may expect these properties to be conserved. Thus, we examined conservation of these properties using carefully curated genome-scale metabolic models of microbes from the other major domains of life: the archeon *Methanosarcina barkeri* [49], and the eukaryotes *Saccharomyces cerevisiae* [38] and *Chlamydomonas reinhardtii* [232].

As in *E. coli*, the three organisms contain numerous multi-specific enzymes (Figure 4.5.A-C). Enzyme activities were subsequently estimated using MCMC sampling while simulating common growth conditions for each organism, including experimentally measured substrate uptake rates [49, 232, 233]. In each organism, absolute-specific enzymes maintained a higher flux on average than multi-specific reactions. Moreover, when environmental shifts were simulated for each organism, multi-specific enzymes were substantially less likely to change flux between growth conditions. Thus, through the diversification of microbes, higher flux and a need for regulation in varying environments remain as general features of selection for absolute-specificity of enzymes.

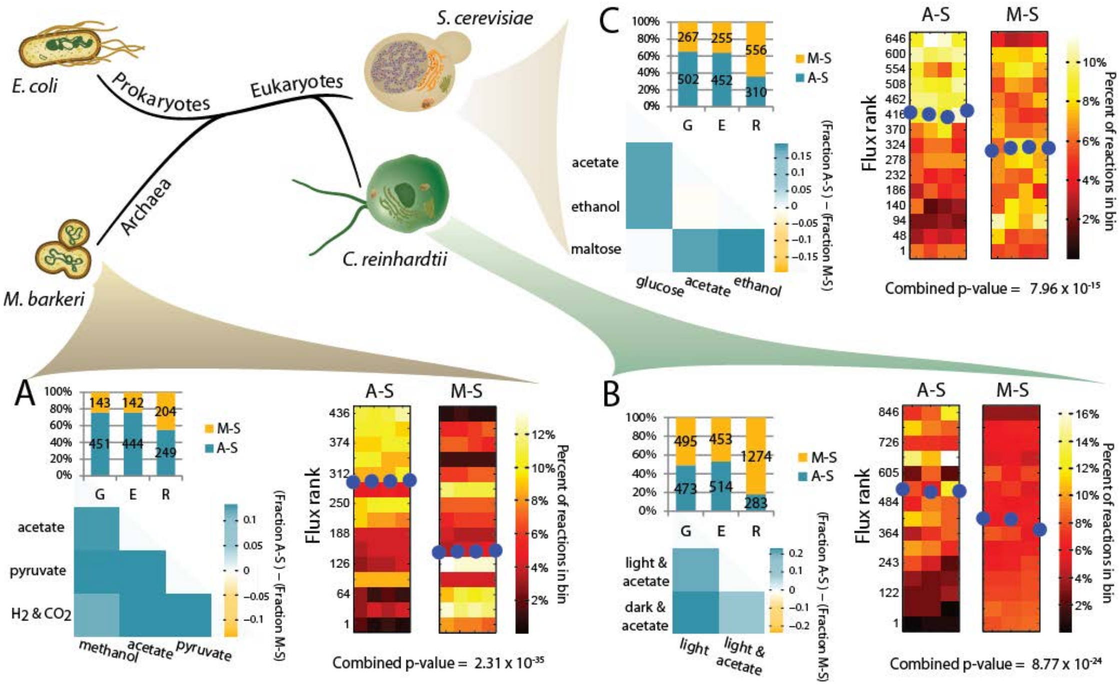


Figure 4.5. Enzyme-specificity characteristics hold for microbes in all domains of life, as shown here for (A) *M. barkeri*, (B) *C. reinhardtii*, and (C) *S. cerevisiae*. Multi-specificity is abundant, as shown by the gene, enzyme, and reaction (G/E/R) composition for each species. Moreover, absolute-specific reactions load higher magnitudes of flux, and are enriched in reactions that are predicted to be metabolically regulated for each change in nutritional environment.

The contribution of network context and cell needs to enzyme evolution

Absolute-specificity represents the textbook view of enzymes being "specific catalysts". However, evidence suggests that enzyme function may arise from the amplification of promiscuous activities that provide a fitness advantage in a new environment [224, 234]. Thus enzymes may be initially multi-specific [206, 228, 235, 236]. Only through mechanisms such as gene duplication [222-224] or horizontal gene transfer, do catalytic activities evolve beyond promiscuous intermediates toward absolute-specificity. While some multi-specificity may represent recently-amplified promiscuous activities, this work suggests that multi-specific enzymes are possibly widespread because they receive less selective pressure from their use in the network context. Their lower essentiality, smaller flux load, and reduced regulatory requirements may not provide adequate fitness advantages to offset the required costs of gene duplication and maintenance [229, 237] when catalytic functions are separated into several absolute-specific enzymes. However, if an environmental change elicits the right fitness challenge, the functions of the genome-scale network may cause these multi-specific enzymes to evolve towards absolute-specificity.

Awareness is increasing of how the functions of biomolecular networks influence evolution, as genome-scale network studies have demonstrated its role in gene essentiality, genome reduction, epistasis, and specific evolutionary trajectories. Similarly, the results presented here add to our understanding of evolutionary selection by showing that enzyme evolution is guided by the physiological functions that the metabolic network, as a whole, must generate to support organism survival. By analyzing the network function underlying cell physiology, and using it to integrate many disparate data types into a coherent whole, systems biology allows one to elucidate the sources and consequences of distal causation through intricate selection pressures that are not apparent at the level of a single enzyme.

Methods

***E. coli* culturing and phenotyping:** Phenotype information, including metabolite uptake rates and growth rates were previously obtained for aerobic growth on glucose, glycerol, and propylene glycol and anaerobic growth on glucose [114, 217, 238]. To complete the set of phenotypes for shifts in carbon, oxygen, and nitrogen metabolism, we obtained growth phenotyping data for different nitrogen conditions. This was done by taking glycerol stocks of *E. coli* K-12 MG1655 and inoculated 2 g/L glucose M9 minimal media to grow the culture at 37°C overnight. Aliquots were then grown anaerobically on 2 g/L glucose M9 minimal media supplemented with either ammonium or nitrate (20 mM) at 37°C. Cells were grown exponentially while sampling growth rates and media multiple times. Growth rates were determined by measuring the optical density of cultures at 600nm. Glucose uptake and acetate secretion were measured by HPLC. The nitrate metabolic rate was approximated from the *i*AF1260 model of *E. coli* using the measured growth rate and glucose uptake rate.

Gene-expression profiling: To provide support for the MCMC sampling-based computational flux change predictions, we compared these with gene expression changes. This was done using novel expression data sets published here and published data from glucose, glycerol, anaerobic, and nitrate conditions [114, 215, 217]. The glycerol-glucose shift provides support for the changes between substrates that are metabolized similarly, while we additional data we present here support our results for substrates that differ substantially in how they are metabolized, despite only differing by one oxygen in the molecular formulae (propylene glycol and glycerol).

Since *E. coli* cannot normally grow on propylene glycol, a strain of *E. coli* K-12 MG1655, adapted for glycerol growth was evolved to also metabolize propylene glycol [238]. This strain was subsequently grown and expression profiles on both 2g/L glycerol M9 minimal media and 2g/L propylene glycol M9 minimal media at 37°C. Affymetrix *E. coli* Antisense Genome Arrays were used for all transcriptional analyses. Each experimental condition was tested in triplicate in the respective carbon sources (i.e., glycerol or propylene glycol) using independent cultures and processed following the manufacturer-recommended protocols. Cultures were grown to mid-exponential growth phase aerobically (OD600 = 0.3) in minimal media supplemented with appropriate carbon source. Three ml of cultures were added to 2 volumes of RNAprotect Bacteria Reagent (Qiagen) and total RNA was then isolated using RNeasy columns (Qiagen) with DNase I treatment. Total RNA yields and quality were measured using a Nanodrop 1000 (Thermo Scientific) and agarose gels. cDNA synthesis, fragmentation, end-terminus biotin labeling, and array hybridization were performed as recommended by the Affymetrix standard protocol.

The Affymetrix CEL files were normalized using gcrma (version 2.20.0) implemented in R (version 2.11.1). Genes were considered not expressed if their median expression level across replicates was lower than the median value of intergenic (IG) probes, and removed from further analysis if they were not expressed in all conditions. Differentially expressed genes were determined using a two-tailed t-test followed by false discovery rate (FDR) p-value adjustment (FDR $\leq$ 0.01).

Designation of absolute- and multi-specificity. For this study we classified 1,147 enzymes from the E. coli genome-scale model (iAF1260) [189] as follows following the detailed process shown in

Figure 4.6.A-B. First, we selected 1,081 proteins which are reported as having enzymatic activity in the Ecocyc Database [212]. In this step, 66 proteins were removed since they did not have experimentally-validated catalytic activities. These included non-catalytic members of enzyme complexes (e.g., the electron transferring protein flavodoxin (b0684)) or predicted enzymes (e.g., predicted carbamate kinase (b0521)). Among these 1,081 enzymatic proteins, 671 and 410 proteins were classified as absolute- and multi-specific enzymes, respectively (Figure 4.6.A). These absolute-specific and multi-specific enzymes are encoded by 713 genes and 477 genes, respectively. We note that the majority of multi-specific enzymes in this study include enzymes exhibiting true multi-specificity (i.e., multiple enzymatic activities with physiological importance), but we anticipate that some reactions may represent examples of well-characterized catalytic promiscuity. While it is unlikely that any experimental technique could clearly differentiate between the two, this property should not substantially affect the conclusions in this work, since variations on the categorization led to qualitatively similar results.

Following enzyme classification, reactions associated with these enzymes were grouped into absolute- and multi-specific reaction classes. If a reaction is catalyzed by an absolute-specific enzyme, the reaction is classified as “absolute-specific reaction” (Figure 4.6.B). Otherwise it was classified as a “multi-specific reaction”. The reaction lists were also filtered to remove reactions with ambiguous classification. Specifically, 63 reactions associated with both of absolute- and multi-specific isozymes were filtered out from the further analysis. We also note that transport reactions were removed as they usually do not represent canonical metabolic catalysis beyond, for example, ATP hydrolysis for in ABC transporters. However, the presence of transporters did not qualitatively change the results in this work.

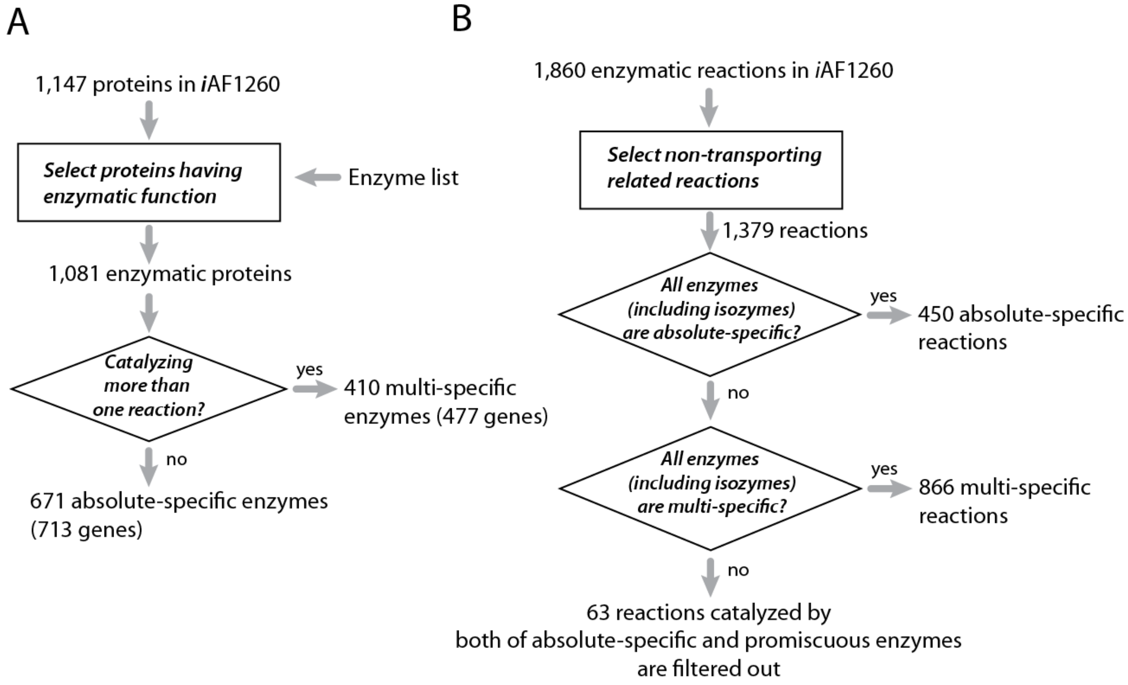

A
1,147 proteins in *i*AF1260
Select proteins having enzymatic function
Enzyme list
1,081 enzymatic proteins
Catalyzing more than one reaction?
yes
410 multi-specific enzymes (477 genes)
no
671 absolute-specific enzymes (713 genes)
B
1,860 enzymatic reactions in *i*AF1260
Select non-transporting related reactions
1,379 reactions
All enzymes (including isozymes) are absolute-specific?
yes
450 absolute-specific reactions
no
All enzymes (including isozymes) are multi-specific?
yes
866 multi-specific reactions
no
63 reactions catalyzed by both of absolute-specific and promiscuous enzymes are filtered out

Figure 4.6. Absolute-specificity and multi-specificity classification process in iAF1260. (A) Enzymes and genes classification steps. (B) Reaction classification steps.

Markov chain Monte Carlo sampling: The distribution of feasible fluxes for each reaction in the models used here were determined using Markov chain Monte Carlo (MCMC) sampling [35], as previously described [42, 86], and was implemented with the COBRA Toolbox v2.0 [214]. Published uptake rates were used to constrain the models. To model more realistic growth conditions [87], sub-optimal growth was modeled. Specifically, the biomass objective function (a proxy for growth rate) was provided a lower bound of 90% of the optimal growth rate as computed by flux balance analysis [90]. Thus, the sampled flux distributions represented sub-optimal flux-distributions, while still modeling fluxes relevant to cell growth and maintenance.

MCMC sampling was used to obtain thousands of feasible flux distributions (referred to here as “points”) using the artificially centered hit-and-run algorithm with slight modifications, as described elsewhere [42, 86]. Briefly, a set of non-uniform points was generated. Each point was subsequently moved randomly, while remaining within the feasible flux space. To do this, a random direction is first chosen.

Second, the limit for how far the point can travel in the randomly-chosen direction is calculated. Lastly, a new random point on this line is selected. This process is repeated until the set of points approaches a uniform sample of the solution space, as measured using the mixed fraction metric described previously [101]. A mixed fraction of approximately 0.50 was obtained, suggesting that the space of all possible flux distributions is nearly uniformly sampled.

For each reaction, a distribution of feasible steady-state flux values is acquired from the uniformly sampled points, subject to the network topology and model constraints. For the *E. coli* model such distributions of feasible flux values could be determined for 2,314 of the 2,382 reactions. The remaining 68 reactions were involved in loops [100] and therefore reliable flux estimates were not available. Thus, sampling distributions for these 68 reactions were removed from all analysis in this work. Similar measures were taken for all other models in this work.

Model parameterization: In general, metabolic models were used in their published format with published uptake and secretion rates [49, 114, 189, 232, 233]. For the 174 simulated media formulations in *E. coli*, glucose uptake was set to zero in the iAF1260 model, and flux balance analysis was used to find which of all other carbon sources could support growth (all of these carbon sources were supported by documented assays in the reconstruction of iAF1260). For each of the 174 growth-supporting carbon sources, an uptake uptake rate was set, which was consistent with uptake rate of glucose in the published iAF1260 model (i.e., 8 mmol grDW^{-1} hr^{-1}), normalized by the number of carbons in the metabolite. For example, since glucose has 6 carbons, the uptake rate of lactate, with 3 carbons, was set as 16 mmol grDW^{-1} hr^{-1} (which is similar to the actual reported lactate uptake rate in M9 minimal media [217]). While this was used to standardize the media conditions, variations in carbon uptake rates did not significantly impact the results presented in this work.

All models were selected based on the availability of carefully curated genome-scale metabolic network reconstructions with measured metabolite uptake rates. Specific media conditions for the eukaryotic and archaea models included the following. *M. barkeri* growth was simulated on minimal media containing methanol, acetate, pyruvate, or H_2 and CO_2. *S. cerevisiae* growth was simulated with glucose, acetate, ethanol, and maltose minimal medium. For *C. reinhardtii*, three growth conditions were used: light with no acetate, light with acetate, dark with acetate. Details on media formulations are provided elsewhere [49, 232, 233].

Flux load ranking: We compared the flux magnitudes of absolute- and multi-specific reactions. In order to avoid biases resulting from variations between growth conditions, we used a rank-based metric to compare flux between conditions. This was done as follows. The median flux magnitude values were calculated from the MCMC-sampled flux loads for each of the 174 different media formulations. For each condition, reactions were filtered out if they were transporter related, involved in loops, non-enzymatic, or could not carry flux. The median flux loads for each reaction were then rank-ordered and the distributions of ranks were compared for absolute- and multi specific reactions. In comparing the relative flux loads between reactions, higher flux magnitudes correspond to higher rank in this study. The significance of higher flux magnitudes in absolute-specific reactions were evaluated by using one-tailed t-tests and Fisher's method.

Essentiality: Previously, 300 essential genes in *E. coli* were identified experimentally [218], and this list was used here. To complement this analysis, an *in silico* analysis was used to assess reaction essentiality with respect to the synthesis of biomass precursors, since we hypothesize that the selective pressure would exert its influence through the reactions themselves. The *in silico* approach used MCMC sampling to simulate growth (>90% of the in silico-predicted optimal growth rate). The distributions of feasible flux values of each reaction was used to assess the correlation of flux between it and the biomass reaction (a pseudo-reaction that simulates the consumption of all biomass precursor metabolites in order to produce biomass) [10]. Reactions that significantly contribute to or are essential for growth are identified by having a significant p-value from the computation of the Pearson's correlation coefficient. While we selected a p-value cutoff of $1x10^{-10}$, the results were consistent for any reasonable p-value cutoff. These correlated reactions contain no redundant pathways, and would therefore provide the most stringent selective pressures since they are the most essential reactions.

Prediction of flux changes between media conditions: To simulate changes in reaction flux occurring in a shift between two conditions, the sampled fluxes for each reaction were compared between two media conditions as follows. First, reactions that carried no flux in both conditions or that were involved in loops [100] were removed and not used in further analysis. Next, flux magnitudes were normalized between each pair of media conditions. To do this, the flux value of each sample point was divided by the sum of all flux magnitudes for the of a sample point.

$$normed\_flux_{ij} = flux_{ij} / \sum_{i=1}^{n} abs(flux_{ij}) \quad , n = \text{number of reactions}$$

Once the flux values were normalized, the changes of fluxes between two conditions were determined as previously described [86]. Briefly, differential reaction activity was determined by assuming that a reaction is differentially activated if the distributions of feasible flux states (obtained from MCMC sampling) under two different conditions do not significantly overlap. For each metabolic reaction, a p-value was obtained by computing the probability of finding a flux value for a reaction in one condition that is equal to or more extreme than a given flux value in the second condition. Significances of p-values were adjusted for multiple hypotheses (FDR ≤ 0.01).

Clustering of reaction changes: An m x n matrix with m absolute- and multi-specific reactions and n media shifts (n = 15,051) was made, detailing in which shifts each reaction significantly changed flux (FDR < 0.01). The reactions were subjected to k-means clustering (k = 3). Clustering was repeated 100 times with different seed values to find consensus clusters. Enrichment tests in the clusters were done using the hypergeometric test.

Enrichment of post-translational modifications and metabolic regulation: Lists of metabolic proteins with post-translational modifications (PTMs) were obtained from studies that identified sites of protein acetylation, phosphorylation, and succinylation in *E. coli* [175-178]. All reported occurrences of non-covalent metabolite-mediated metabolic regulation were obtained from Ecocyc [212]. Metabolic regulation events labeled in Ecocyc as allosteric, noncompetitive, uncompetitive, and competitive were used in this analysis to distinguish between different regulatory properties of these enzymes. Enrichment and depletion of PTMs and metabolite-mediated metabolic regulation events in the gene lists and reaction clusters were determined using the hypergeometric test.

Cosine similarity: The patterns of how multi-specific reactions change when sharing the same enzyme was estimated by using the cosine similarity metric. For each shift, the median flux magnitudes of a reaction in conditions x and y were represented as a vector (R_α(fx, fy)). The similarity score of two reactions, α and β, was then measured by the cosine similarity of the two vectors, R_α and R_β.

$$\cos(R_\alpha, R_\beta) = \frac{(R_\alpha \bullet R_\beta)}{|R_\alpha| \times |R_\beta|}$$

The similarity score of reactions catalyzed by the same enzyme (*e_i*) was calculated as the mean value of all pair-wise cosine similarity scores for reactions catalyzed by that multi-specific enzyme.

$$\text{similarity}(e_i) = \sum^{n} \text{abs}(\cos(R_\alpha, R_\beta)) / n$$

For example, for an enzyme e1 that catalyzes three reactions (r1, r2, and r3), the flux similarity score of e1 is calculated as an average value of cosine distances of three reaction pairs. Similarity scores for multi-specific enzymes were compared to randomized tests. The randomized tests were achieved by averaging 2,000 cosine similarity scores of randomly paired reactions.

Chapter 4, in part, is a reprint of the material as it appears in Nam, H.J., Lewis, N.E., Lerman, J.A., Lee, D.H., Chang, R.L., Kim, D., Palsson, B.Ø. Network context and selection in the evolution to enzyme specificity. *Submitted*. I was a joint-primary author and the corresponding author, while the remaining co-authors provided support in the research that served as the basis for this study.

Section III: Distal causation in action: experimental evolution shapes metabolism

Retrospective evolutionary analysis has in many instances demonstrated how distal causation has shaped proximal metabolic capabilities. However, can one see these changes in our lifetime? Fortunately, some evolutionary theories may be experimentally tested through adaptive laboratory evolution (ALE). ALE has successfully demonstrated changes to the transcriptional regulatory network [200] and the metabolic network topology [239]. These changes provide further support that through distal causation in biology, the interaction between cellular objectives and selection pressures guides the evolution of metabolic networks [239]. The subsequent chapters demonstrate this by showing that biomass production is so important to cells under a growth rate selective pressure, that the cells will optimize their gene and protein expression to enhance the efficiency of metabolism. The cells do this not only for nutritional sources they are accustomed to, but they also evolve their metabolic capabilities to optimize metabolism for novel nutritional sources. This optimization happens at both the levels of metabolic pathways and specific enzymes.

Chapter 5: Laboratory evolved *E. coli* optimize cellular metabolism

When prokaryotes are grown at low to mid-log phase for hundreds of generations through periodic serial passage, they acquire an increased growth rate [48, 200, 240-243]. This example of laboratory adaptive evolution is expected, since faster growing mutants quickly outgrow slower growing cells, even if the initial fitness difference is small [244]. Molecular changes that confer the growth improvement have been previously studied using fluxomics [245, 246], transcriptomics [110, 217, 247, 248], and whole genome resequencing [200, 201, 243, 249]. For example, whole genome resequencing of adapted strains demonstrated that only a small number of mutations arise after hundreds of generations [200, 201]. While each evolved strain acquired a different set of mutations, each set of mutations yielded a similar growth phenotype. When these mutations were introduced into the wild-type strain by allelic replacement, the wild-type cells acquired the evolved-strain growth rates [201]. However, the mechanism linking the mutations to the improved growth rate in most evolved strains has yet to be clearly identified, except for cases in which strains had a mutation in RNA polymerase (RNAP) or *glpK* [201], which altered activity of transcription and glycerol uptake.

Although the genetic changes have been identified and characterized, the resulting coordination of cellular processes that lead to the altered phenotypes have only been studied briefly from a network perspective. For example, it has been shown that after hundreds of generations of adaptive evolution at exponential growth, *Escherichia coli* grows as predicted using flux balance analysis (FBA) on genome-scale metabolic models (GEMs). However, the pathway usage has only been compared indirectly with optimal pathway usage predictions from FBA and other modeling approaches. Such studies of adaptively evolved strains have shown an activation of normally latent metabolic pathways [246], expression improvements to the strains that make them more consistent with a high growth rate for various minimal media conditions [110], improved respiration [250], optimization of a small growth-coupled circuit [185], and protection from the uptake of compounds that can be toxic to a specific species [48]. In addition, the measured growth rates of evolved strains were shown to be consistent with most growth rate predictions from an *in silico* genome-scale metabolic model (GEM) of *E. coli* [240, 251].

While all of these studies have elucidated some characteristics of the complex adaptation process, it is not known 1) if absolute genome-scale gene and protein expression levels and expression changes are consistent with optimal growth predictions from *in silico* GEMs, or 2) if measured expression changes can be linked to physiological changes that are based on known mechanisms or pathways. To begin to address these questions, we use constraint-based modeling of *E. coli* K-12 metabolism [5, 134] to analyze a compendium of “omics” data obtained from adaptive evolution experiments. First we show that the data are consistent with pathway usage from the computationally-predicted optimal growth states. We next show that expression changes during the adaptation process relative to wild type further converge to predicted enzyme usage from the optimal growth rate predictions (Figure 5.1). Lastly we demonstrate that changes in known regulatory processes acting on the metabolic network, but not accounted for in the GEMs, are consistent with the improved-growth phenotypes of the adapted strains.

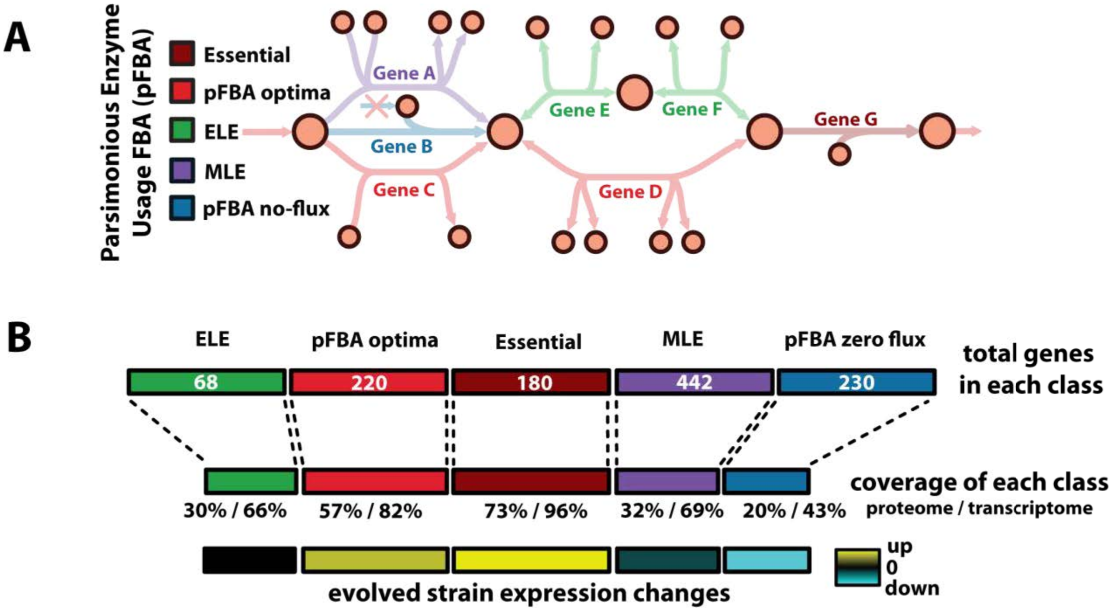


Figure 5.1. A variant of Flux Balance Analysis shows consistency with proteomic and transcriptomic data. Parsimonious Enzyme Usage FBA (pFBA) is used to label all metabolic genes based on simulation results. (A) pFBA classifies each gene based on its ability to contribute to the optimal growth rate predictions and flux level. These classes include: (i) Essential genes; (ii) pFBA optima, which includes genes that are predicted to be used for optimal growth *in silico*; (iii) ELE, which includes genes that will increase cellular metabolic flux if used; (iv) MLE, which includes genes predicted to decrease the growth rate if used; and (v) pFBA no-flux, which includes genes that cannot be used in the given growth conditions. (B) The omic data show good coverage of essential genes and the pFBA optima, and low coverage of the genes that are predicted to be non-functional. In addition, in laboratory evolution experiments, these optimal states are up-regulated, while non-functional genes are down-regulated. These results support predicted optimal growth states, and suggest that laboratory evolved strains further enhance these optimal growth states.

The proteomic and transcriptomic landscape of evolved E. coli

Multiple strains of *E. coli* were subjected to adaptive evolution via serial passaging in three different M9 minimal media conditions: lactate, glycerol, and glucose (glucose grown strains had the glycolytic gene *pgi* deleted to perturb the normal flux into glycolysis). For each growth condition, 3-6 replicates of the adaptive process were performed in parallel until each strain had reached and maintained a steady growth rate, which typically took 700-1000 generations (see [217, 246] for details). Through adaptive evolution, all strains improved their growth rate and efficiency in converting substrate to biomass (yield) within the exponential growth phase (Figure 5.2).

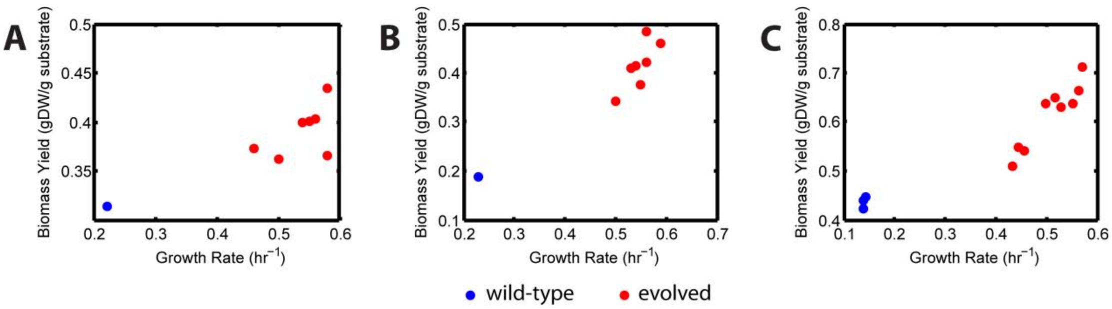


Figure 5.2. In adaptive evolution via serial passaging, *E. coli* evolves to a higher growth rate and biomass yield at exponential growth. Growth rates and substrate uptake rates were acquired for each strain prior to and following adaptive evolution (as reported previously [217, 249]. For growth on (A) glycerol, (B) lactate, and for (C) the Δ*pgi* strain grown on glucose, the evolved strains (red) all improved their growth rate and biomass yield at exponential growth, compared to the unevolved parent strains (blue).

Fifty quantitative proteomic data sets were obtained from the wild-type and evolved strains. Within these data sets, 983 proteins were identified with high confidence, of which 731 were identified in all strains. Transcriptomic data for strains corresponding to two of the three growth conditions (lactate and glycerol) have been previously published [217] and are also analyzed alongside the proteomic data in this study using the *E. coli* GEM as a context for the analysis.

In the omics datasets for the adaptation process, hundreds of genes and proteins are differentially expressed, representing 32-59% of the identified proteins and expressed genes in the data sets. The proteomic and transcriptomic data show significant agreement in the direction of differential expression for cases in which both the gene and protein significantly changed expression level (see [47] for details).

We first analyze the omics data with reference to enzyme usage in the computed optimal states from GEMs, then look at the changes that occur during evolution by analyzing the differential expression relative to the wild-type cells. Finally, we look at changes that correspond to the action of non-metabolic genes represented in the data sets.

Analysis of omics data in the context of computed optimal growth states

Both the omics data sets and the computed solutions can be compared in the context of network functions. The transcripts and proteins found in the omics data sets can be mapped onto the reconstructed genome-scale network. Computed optimal solutions can also be presented on the network map and compared to the omics data. A comparative analysis can then be performed.

To determine if gene and protein expression support properties of optimal predicted network function, we employed a variant of flux balance analysis (FBA), referred to as *Parsimonious enzyme usage FBA* (pFBA) (Figure 5.1.a). As described below, this method employs *in silico* simulations to identify functional properties of metabolic pathway genes under the given growth conditions. We applied pFBA to the omics data sets to determine if absolute expression and differential expression during adaptation supports the enzyme usage in computed optimal solutions. All reports of absolute expression coverage are a combination of WT and evolved strain data, since there are few proteins that are missing in the WT strains but identified the evolved strains, and vice versa (fewer than four for any single growth condition). To provide additional insight into the conclusions in this study, an alternative method, Flux Variability Analysis [252], was also used and yielded supportive results (see [47] for details).

pFBA (Figure 5.1.a), assumes that under exponential growth, there is a selection for the fastest growing strains and for strains that require the lowest overall flux through the metabolic network (a proxy for minimizing the total necessary enzyme mass to implement the optimal solution). This additional constraint introduces a small improvement over normal FBA. While these assumptions may not hold true in all growth conditions for all organisms [87, 253, 254], previous studies in *E. coli* [97, 240, 241] and data presented here support these assumptions under our experimental conditions.

pFBA finds the subset of genes and proteins that may contribute to the most efficient metabolic network topology under the given growth conditions, called here the pFBA optima. The genes contributing to pFBA solutions can be classified as follows:

1) Essential genes: metabolic genes necessary for growth in the given media.
2) pFBA optima: non-essential genes contributing to the optimal growth rate and minimum gene-associated flux.
3) Enzymatically less efficient (ELE): genes requiring more flux through enzymatic steps than alternative pathways that meet the same predicted growth rate.
4) Metabolically less efficient (MLE): genes requiring a growth rate reduction if used.

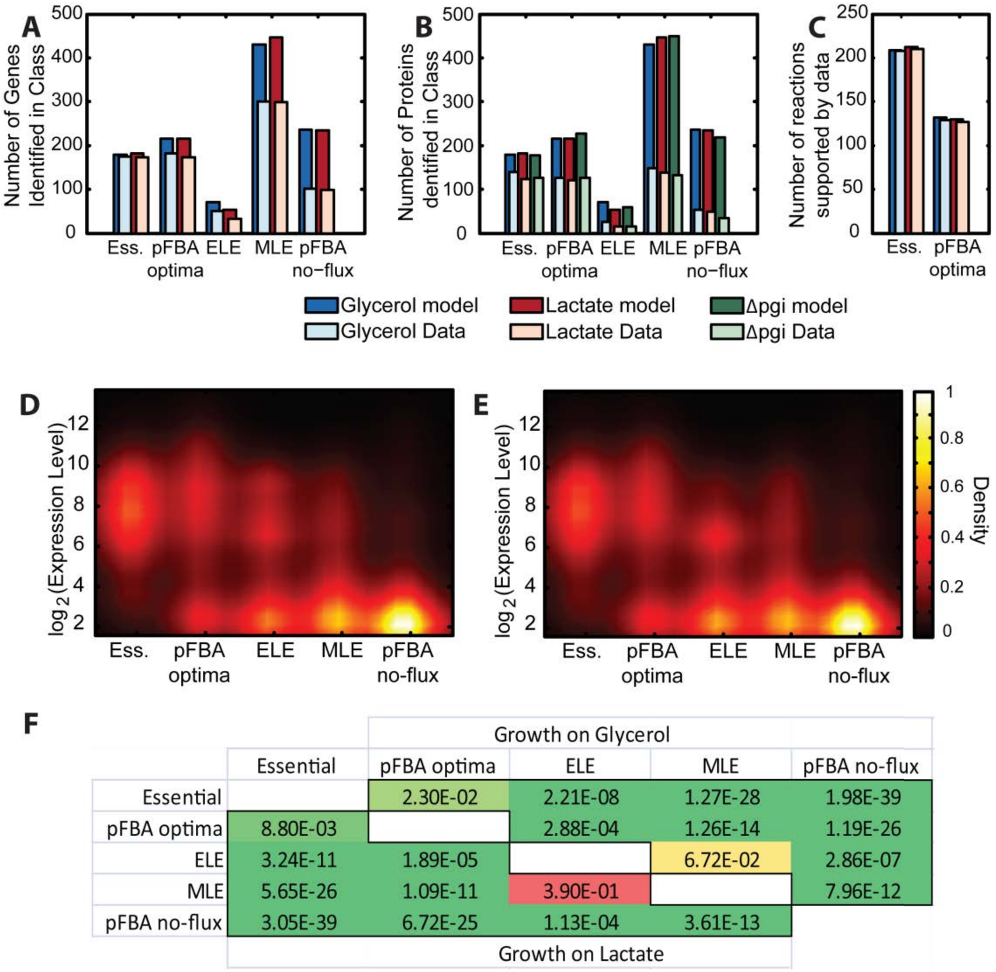
5) pFBA no-flux: genes that are unable to carry flux in the experimental conditions.

See Figure 5.1.b for average sizes of these classes.

Do omics data support pFBA optimal growth states? Computed pFBA solutions correspond well with the set of identified proteins and expressed genes, as well as gene expression levels. Almost all *in silico*-predicted essential genes are expressed. In addition, there is much higher omics data coverage of the genes and proteins in pFBA optima as compared to the less efficient classes (ELE and MLE) and the conditionally non-functional pFBA no-flux class (Figures 5.3 and 5.1.b). In the transcriptomic data, more than 82% of all genes that can contribute to the pFBA optima are expressed. Of the missing genes (mean of 38, representing about 18% of the pFBA optima), about 82% have known isozymes or redundant pathways in the pFBA optima that can replace their functions.

Coverage of proteins in the pFBA optima is less comprehensive than coverage from the transcriptomic data (Figure 5.3.b); however, about 40% of the missing proteins in the Essential and pFBA optima classes are members of the GO classes “membrane,” “integral to membrane,” or “transport.” These classes are significantly depleted from the proteomic, and commonly depleted in other proteomic data sets [255]. Moreover, more than 59% (>50 proteins) of the missing pFBA optima proteins have isozymes in the pFBA optima that could replace their function if these proteins are not expressed.

Neither proteomic nor transcriptomic data alone show expression of all genes or proteins that can contribute pFBA optima. Complete coverage, however, is not expected due to model alternate optima, inaccurate probes on the arrays, and hard-to-detect proteins. However, when the expressed genes and identified proteins are mapped back onto the metabolic network, the union of the proteomic and transcriptomic data corresponds to 97.7% of the non-essential active gene-associated reactions in the glycerol and lactate optimal solutions (Figure 5.3.c). Unsupported reactions include a few transporters (H_2O, NH_4^+) and reactions that are necessary for cofactor biosynthesis.



		Growth on Glycerol			
	Essential	pFBA optima	ELE	MLE	pFBA no-flux
Essential		2.30E-02	2.21E-08	1.27E-28	1.98E-39
pFBA optima	8.80E-03		2.88E-04	1.26E-14	1.19E-26
ELE	3.24E-11	1.89E-05		6.72E-02	2.86E-07
MLE	5.65E-26	1.09E-11	3.90E-01		7.96E-12
pFBA no-flux	3.05E-39	6.72E-25	1.13E-04	3.61E-13	
		Growth on Lactate			

Figure 5.3. pFBA classes are consistent with omic data. Simulations for each growth condition were used to classify each gene according to the efficiency of its associated reaction(s). The coverage of (A) expressed genes above a statistical cutoff and (B) identified proteins were determined. More genes and proteins in the essential and pFBA optima classes were expressed than genes and proteins in the less efficient (ELE and MLE) and conditionally non-functional (pFBA no-flux) pathways. Almost half of the missing essential and pFBA optima proteins are membrane associated, hence its lower coverage. (C) 99% and 98% of the active reactions associated with the essential genes and pFBA optima, respectively, are supported by the union of expressed genes and proteins. Gene expression levels are also consistent with the pFBA classes, as shown in density plots for growth on (D) glycerol and (E) lactate. (F) Pairwise comparisons of classes show the significant ordering of expression levels as Essential > pFBA optima > ELE > MLE > pFBA no-flux (one-sided Wilcoxon test) for growth on glycerol (upper triangle) and lactate (lower triangle).

Beyond presence and absence, the expression levels of genes are consistent with the various pFBA classes. That is, the expression levels are greatest for the essential genes and lowest for the pFBA no-flux genes (Figure 5.3.d-e), and is significant for almost all pairwise comparisons (Figure 5.3.f). All of the above results suggest that the pFBA optima are expressed and likely active in *E. coli* K12.

Many metabolic genes and proteins are differentially expressed with adaptation

There is high coverage of expressed genes and proteins in the optimal computed states. Since the efficient use of the metabolic network is presumed to underlie the optimal growth phenotype following adaptation, the question arises: Do metabolic genes dominate the differential changes during adaptation? Differential expression of proteins and genes in the adaptation process occur in many functional classes; however, a large fraction of these differentially expressed proteins and genes are associated with metabolic Clusters of Orthologous Groups (COGs) [256] (Figure 5.4). Specific metabolic COGs that show the highest enrichment include carbohydrate transport and metabolism for the lactate and glycerol evolved strains ($p < 0.009$) and amino acid and nucleotide metabolism in the *pgi* deletion strains ($p < 0.012$).

This high coverage of metabolism supports its important role in the evolved growth phenotype, and allows the analysis of the data in the context of the genome-scale metabolic network reconstruction [189]. The dominant contribution of metabolic genes to the changes in the omics data sets is further validated when the data are evaluated using singular value decomposition (SVD), which demonstrates that metabolic GO classes co-vary and separate evolved and unevolved strains.

Since some specific metabolic subsystems may change more than others, we mapped the differential gene and protein expression to the metabolic network using PathWave [154], a method that identifies groups of topologically close reactions that show concerted expression changes. Among the different data sets, this analysis shows significant changes in central carbon metabolism, tRNA charging, and/or the metabolism of specific amino acids. Changes in such regions of the metabolic network play a key role in providing the metabolic precursors for biomass production, and thus may contribute to an increased growth rate. However, for greater insight, changes in biomass-coupled pathways must be quantitatively associated to the actual growth state of the cell.

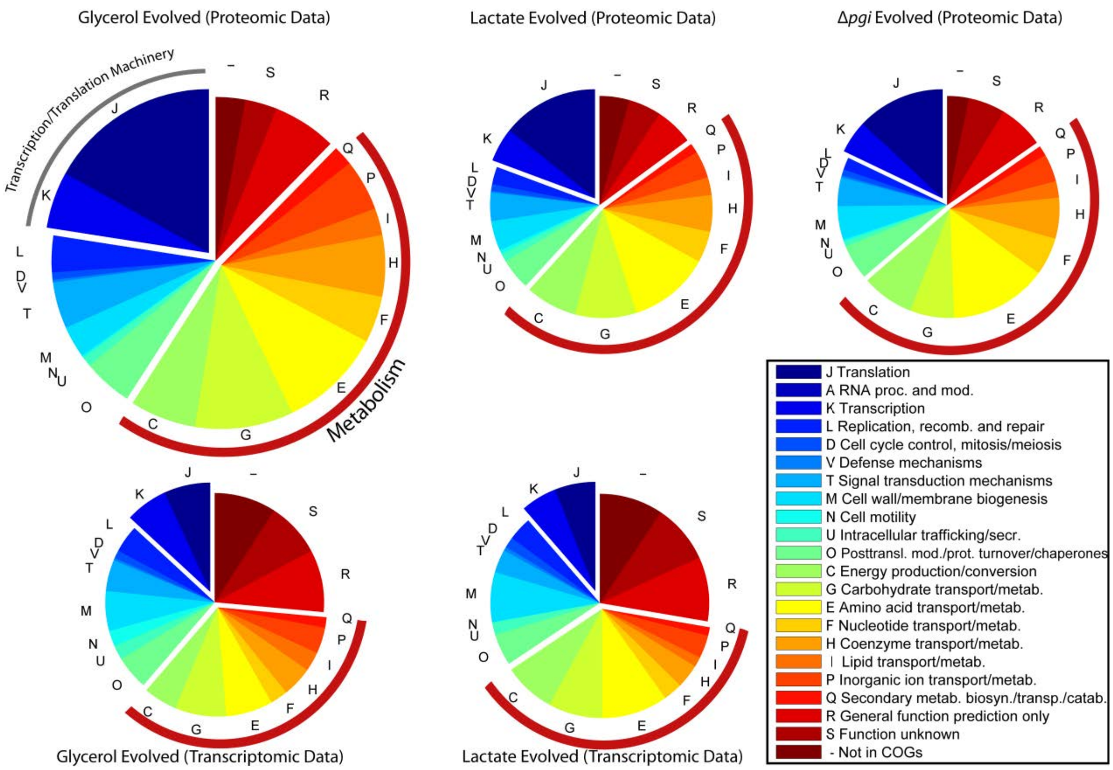


Figure 5.4. Proteins and genes associated with metabolic processes dominate differential expression in evolved strains. Adaptively evolved strains have hundreds of altered gene and protein expression levels covering a broad range of COGs. However, the relatively large coverage of genes and proteins in COGs associated with metabolism (red brackets) clearly suggests the importance of changes in the metabolic network. The most significantly enriched metabolic COG is "carbohydrate metabolism and transport". In addition, the proteomic data suggest that the transcription and translation machinery may also play an important role in the physiological changes witnessed in the adaptively evolved strains.

Adaptive evolution overcomes dosage limitations of essential genes

While pathways that produce key biomass precursors are significantly changed, it is not clear if necessary growth-coupled essential genes are consistently changed as would be needed for an increased growth rate. To address this question, we first used pFBA to identify all genes that are needed for growth *in silico* and compared these with experimental data (e.g., see Figures 5.5.a-b). Since *in silico* growth is dependent on these essential genes, they may be needed in higher abundances for higher growth rates. The adaptive evolution strains, with their improved growth rates, are consistent with this hypothesis. In the evolved strains, computationally predicted essential genes and proteins are significantly up-regulated (Figure 5.5.c) and have fewer down-regulated genes and proteins than expected (Table 5.1). Moreover, down-regulated essential proteins are more abundant in the WT strains than up-regulated proteins ($p < 2 \text{ x } 10^{-8}$). Thus, the down-regulation may be the result of tuning protein expression for over-expressed proteins in WT. This result, with the up-regulation of essential genes and proteins, suggests that the computationally predicted essential genes are indeed growth coupled as predicted *in silico*. Moreover this result suggests that these essential genes not only confer cellular viability, but they also may act as cellular bottlenecks due to dosage limitations. Expression changes during adaptive evolution allow these limitations to be overcome, thereby increasing the growth rate.

Table 5.1. p-values from hypergeometric tests involving the presence of pFBA classes in the up and down-regulated genes and proteins

h0	Glyc Prot	Lac Prot	Δpgi Prot	Glyc MA	Lac MA
Essential genes are not enriched in up-regulation	$8.88x10^{-5}$	$2.17x10^{-2}$	$2.05x10^{-7}$	$1.58x10^{-7}$	$4.52x10^{-2}$
Essential genes are not depleted in down-regulation	$4.08x10^{-5}$	$1.14x10^{-1}$	$1.76x10^{-5}$	$8.25x10^{-7}$	$5.63x10^{-7}$
pFBA optima are not enriched in up-regulation	$2.85x10^{-4}$	$9.88x10^{-5}$	$4.37x10^{-8}$	$4.01x10^{-11}$	$6.96x10^{-4}$
MLE is not depleted in up-regulation	$1.83x10^{-2}$	$2.01x10^{-2}$	$5.62x10^{-4}$	$6.12x10^{-6}$	$4.12x10^{-1}$
pFBA no-flux class is not depleted in up-regulation	$3.07x10^{-3}$	$8.70x10^{-2}$	$8.15x10^{-4}$	$2.18x10^{-7}$	$6.28x10^{-4}$
pFBA no-flux class is not enriched in down-regulation	$4.63x10^{-4}$	$2.96x10^{-1}$	$3.30x10^{-4}$	$6.98x10^{-9}$	$6.06x10^{-7}$

MLE = Metabolically Less Efficient, Glyc = glycerol, Lac = lactate,
MA = microarray, Prot = proteome

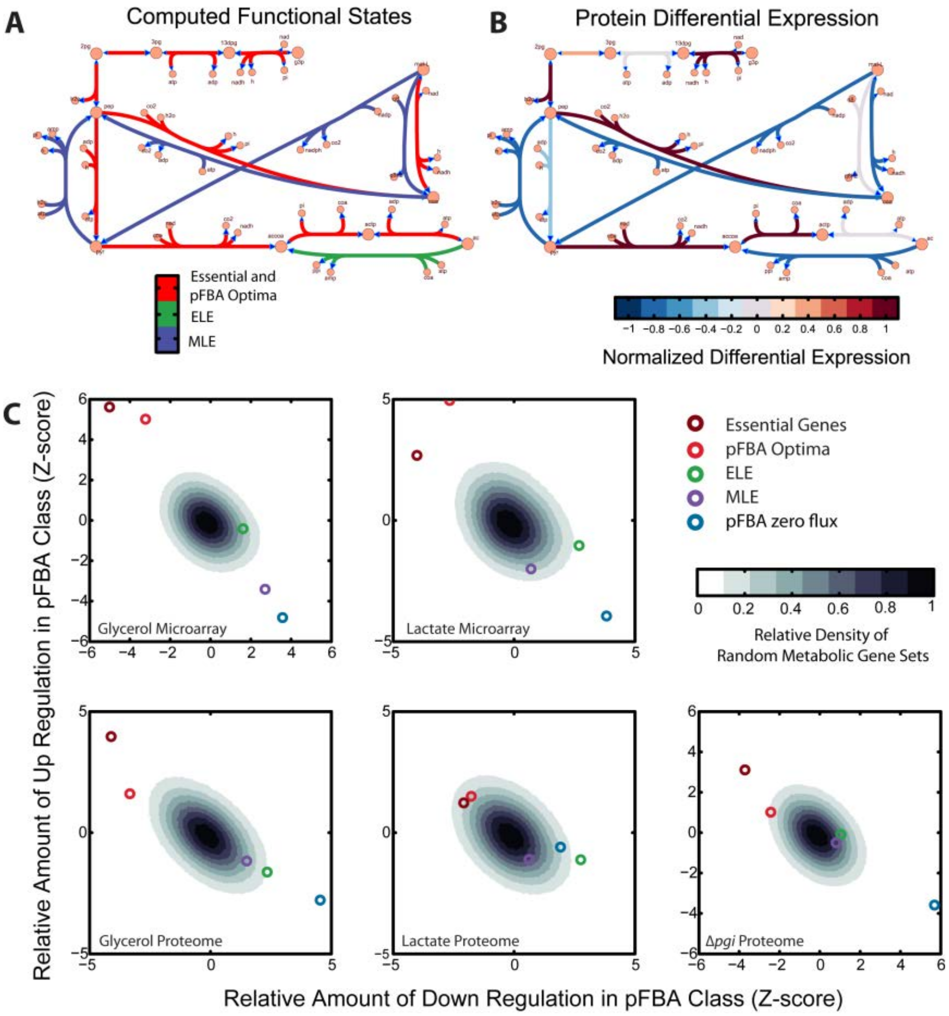


Figure 5.5. The sub-network providing optimal growth emerges in adaptively evolved strains. (A) For each growth condition, pFBA was used to classify all genes and their associated reactions. (B) Subsequently, these classifications were compared to differentially expressed genes or proteins for each growth condition (a representative portion of the metabolic network with glycerol strain proteomic data is shown). (C) A quantitative assessment was done, in which the sum of all down-regulated genes in each class (x axis) was plotted against the fold-change sum of all up-regulated genes (y axis), and then scaled by the variance of the sum of randomly selected differentially expressed genes. The cloud represents the normalized distribution of the summed up and down regulated genes or proteins of randomly chosen differentially expressed genes. This analysis shows that for all data sets, genes and proteins within the essential set and the pFBA optima demonstrate more up-regulation and much less down-regulation than expected from randomly selected differentially expressed metabolic genes and proteins. This emergence of the optimal pathways is enhanced by the lack of up-regulation and the significant down-regulation within the less efficient ELE and MLE pathways and the conditionally inactive pFBA no-flux pathways. ELE = enzymatically less efficient, MLE = metabolically less efficient.

The emergence of the optimal metabolic states in adaptive evolution

All evolved strains profiled here show improvements in both growth rate and yield (Figure 5.2). The up-regulation of essential genes may partially support the increased growth rate; however, it doesn't address the question as to if non-essential gene and protein expression is more consistent with the enzyme usage in computed optimal growth states. In addition, the highly interconnected nature of metabolic networks may preclude a growth improvement from up-regulated essential genes, if pathways that are up and downstream of the essential genes do not change accordingly. To answer these questions, we compared the differential gene and protein expression to computational simulations of genome-scale optimal growth states (Figure 5.5.a-b). Thus, all up and downstream pathways may be considered.

Using pFBA, we find that in all strains, the pFBA optima are significantly up-regulated in the transcriptomic and proteomic data. This up-regulation is significant for both the number of genes (Table 5.1), and the net fold change (Figure 5.5.c). Further support for the use of the pFBA optima comes from the findings that, in-general, the less-efficient MLE genes are not significantly up-regulated (Table 5.1), and that they are down-regulated in most data sets (Figure 5.5.c). Among those that are up-regulated, few contribute to any coherent functional metabolic pathways. Only one MLE gene is consistently up-regulated in all datasets and functional in the context of a non-down-regulated pathway. This protein, deoxyuridinetriphosphatase (E.C. 3.6.1.23), which dephosphorylates dUTP, is up-regulated in all datasets. While this process wastes resources, this enzyme is needed to preclude dUTP from being integrated into the genome, and the absence of this enzyme decreases the growth rate in *E. coli* [257]. A few other MLE genes were up-regulated in multiple, but not all datasets (see [47]).

The up-regulation of the pFBA optima, and the lack of up-regulation among less efficient pathways reveal that the adaptive evolution process leads to the further emergence of pathways that help to maximize the predicted growth rate. Thus the differential changes are consistent with the computed optimal growth state.

Adaptation suppresses conditionally inactive pathways

Since excess enzyme mass creates a large maintenance demand on cells [258], cells under growth selective pressure are expected to modulate expression levels of enzymes as needed for growth [185]. While we showed an up-regulation of optimal pathways, it is expected that genes and proteins associated with non-functional reactions should be down-regulated, thereby saving resources for improved growth performance.

Gene and protein expression changes in the pFBA conditionally nonfunctional class (pFBA no-flux) are consistent with this hypothesis. For all experimental conditions, there is a significant down-regulation of pFBA no-flux genes, except for the lactate strain proteomic data (Table 5.1). Moreover, when compared with the non-evolved strains, the mean abundances of expressed pFBA no-flux proteins and transcripts are significantly lower in all evolved strains ($p << 1 \times 10^{-16}$ and $p = 8.3 \times 10^{-8}$, respectively). Flux variability analysis further supports the suppression of conditionally non-functional metabolic reactions. Thus, during the process of adaptive evolution, computationally-predicted nonfunctional pathways are suppressed through a concerted down-regulation of genes associated with such pathways.

Only down-regulation is tied to known regulon structure

The analysis of the omics data shows that strains under growth pressure adjust their transcriptional program towards the *in silico* predicted optimal growth states in metabolism. However, the mechanisms controlling these changes are outside the scope of the reconstructed metabolic network, and their activities are not predicted. Thus, the question arises: are known transcriptional regulatory mechanisms consistent with the observed differential expression changes?

Across all conditions, the down-regulated transcripts and proteins correspond to several known regulons, and each condition has a unique set of differentially expressed regulons. For example, down-regulated molecular species in the glycerol evolved strains include the flagellar FlhC/FlhD regulon, the GatR regulon (transport and catabolism of galactitol), and Hns (chromosome organization). For lactate-evolved strains, the carbohydrate metabolism regulators Crp and DgsA regulons are enriched in the down-regulated genes and proteins, respectively. Among the Δ*pgi* strains, the four most significantly enriched regulons in the down-regulated proteins include Crp, IhfA/IhfB, MetJ, and ArcA. All of these are associated with carbon or nitrogen metabolism. Moreover, down-regulated members of these regulons account for a higher fraction of the expressed genes and proteins outside of the optimal growth solutions. Together, these results suggest that known regulatory programs may be employed in a condition-specific manner for the down-regulation of genes and proteins in the adaptation process.

Conversely, no data set reflects known regulons among the up-regulated transcripts or proteins. The only exception is for the glycerol-evolved strain microarrays, in which a few amino acid biosynthetic regulons are enriched (ArgR, LysR, MetJ), along with the purine synthesis regulon (PurR), and Fis. These results suggest that few known transcriptional regulatory programs are consistently used to up-regulate genes and ultimately proteins. Therefore it seems that there are unknown regulatory mechanisms at work, potentially due to mutations found in transcriptional regulators in the evolved strains [200, 201, 249]. Mutations in these regulators have previously led to drastic alterations in gene and protein expression [199, 200, 259, 260]. Further interrogation of these mutated regulators will aid in associating the expression changes to known regulatory pathways.

Adaptively evolved strains largely eliminate the stringent response

Changes in transcriptional regulation observed here lead to altered physiological responses associated with metabolism, such as the stringent response. All experiments here were performed in media without amino acids. Under such conditions, the stringent response increases transcription of amino acid biosynthesis genes needed for growth [261], and simultaneously decreases the growth rate; however, evolved strains manage to attain a higher growth rate, despite the stringent response.

To find a rebalancing of genes involved in the stringent response, we compared the microarray data from the glycerol and lactate-evolved strains to published data sets that profile the stringent response in *E. coli* K-12 MG1655 [261]. Out of the 170 differentially expressed stringent response genes, a total of 97 genes are also significantly differentially expressed in the evolved strains. In both evolved strain conditions, approximately 90% of the expression changes occur in the opposite direction as the stringent response. That is, after adaptation to minimal media, the *E. coli* strains show expression patterns consistent with a decreased stringent response during growth. Only eight genes show changes in the same direction in the evolved strains and the stringent response. Of these, half are amino acid biosynthetic genes (*ilvM*, *ilvD*, and *thrL*) or play a secondary role in amino acid biosynthesis (*folE*). Thus, there is a clear suppression of the stringent response in the evolved strains, but alternative mechanisms allow the needed up-regulation of amino acid biosynthesis genes normally activated by the stringent response.

Implications of the laboratory evolutions towards metabolic optimality

Wild-type lab strains of *E. coli* adapt to new growth conditions when placed under a growth rate selective pressure [48, 240-243]. The genetic and physiological characteristics of the adaptation have been described [200, 201, 243, 249]. The underlying genotype-phenotype relationship can be detailed using systems biology; namely the acquisition and analysis of omics data and the use of genome-scale models.

In this study we obtained a compendium of quantitative proteomic profiles of the evolved strains and used a similar set of previously published microarrays [215, 217]. The analysis of the data sets, using conventional statistical methods and GEM computations, yielded three key results. First, the proteomic and transcriptomic data are consistent with enzyme usage in optimal growth state computations using GEMs. Second, the essential and non-essential metabolic genes associated with the predicted optimal growth states are induced during the adaptive process. This is accompanied by a suppression of proteins and transcripts outside of the optimal growth solutions. Third, regulatory mechanisms, not accounted for in genome-scale metabolic network models, contribute to the altered metabolic states and the improved growth phenotype. Known transcriptional regulatory mechanisms contribute to the down-regulation of genes and proteins, and physiologically, there is a suppression of the stringent response. These results have three main implications.

First, in this work we found a high coverage of genes and proteins associated with the predicted optimal growth states. This result provides added support for the validity of predicted pathway utilization using GEMs and for the assumptions underlying their computation. More specifically, FBA pathway flux predictions are computed by relating uptake and secretion rates, given the stoichiometry of the metabolic network and a biomass objective function. The biomass function represents the stoichiometric balance of metabolites needed for growth. Thus, FBA allows the computation of the growth yield (the amount of biomass produced per mole of substrate), and predicts pathways that can be used to obtain this yield. FBA further computes the optimal growth rate, assuming the cell will optimize this growth yield, given the measured substrate uptake rate and cellular maintenance costs [262] (for discussion on the subtle differences between computed growth yields versus growth rates, see [48]).

The physiological relevance of the FBA optimal growth rate assumption has been discussed [87]. In particular, it has been proposed that two possible mechanisms can lead to improved growth rates: 1) the improved efficiency of converting substrate to biomass (consistent with FBA predictions) or 2) the speeding up of metabolism by increasing the expression level of any enzymes (efficient or less efficient) in order to speed up metabolism. Previous studies have presented evidence supporting both scenarios under the adaptive evolution experimental conditions by measuring growth rates, substrate uptake rates, and by-product secretion rates [48, 97, 240, 241]. The present study provides additional experimental support for both an improved efficiency and a speeding up of metabolism in adaptively evolved strains by showing the up-regulation of the pathways in the optimal growth rate solutions, and not in the less efficient pathways. The up-regulation of the essential genes allow for a higher growth rate since they are more tightly coupled to the *in silico* predicted growth rate. The up-regulation of the pFBA optima allows for improved efficiency in converting substrate to biomass (biomass yield). Thus the up-regulation of the essential and non-essential genes in the optimal pathways allows for both the "speeding up" of metabolism and increased efficiency, as the measured substrate uptake increases [217, 249]and is metabolized through the up-regulated optimal pathways.

The second implication of this work is that a few simple mutations may perturb the function of the entire network, and that the resulting phenotype can be better understood using GEMs. Previous studies have demonstrated that simple mutations in metabolic network enzymes produce a transient response that minimizes flux changes [55, 64]. However, in this work, each strain studied had ample time for more drastic changes in gene and protein expression as a result of the mutations in metabolic enzymes and global regulators attained in the adaptive time-course [200, 201, 249]. Even though the cellular biochemistry is tightly woven into a large network, the measured expression changes shifted towards the computed optimal growth predictions. This finding demonstrates that some physiological observations cannot be simply explained with a direct link to a single mutation. However, the genotype-phenotype link, which usually is complex, may be better identified by analyzing the data in the biomolecular network context.

The third implication of this work is that genome-scale models of other systems such as transcriptional regulation, transcription, and translation are needed for a more complete understanding of the genotype-phenotype link. This work demonstrated the successful model-based analysis of a large fraction of differentially expressed genes and proteins. However, we also witnessed changes beyond the scope of the model, such as in the transcription and translation machinery components (Figure 5.4). Many of these, such as tRNA charging enzymes, the ribosomal proteins, and subunits of the RNAP, were up-regulated in most strains (data not shown). Each of these could allow for faster growth by providing increased translation and transcription rates [263, 264]. Metabolic models do not directly account for these mechanisms. Thus, it is anticipated that genome-scale models of transcription and translation [265] will be useful in evaluating the functional consequences of changes in these systems. Moreover, efforts are also being made to address additional growth rate-associated parameters, such as changes to the cell surface to volume ratio and molecular crowding constraints [44, 254].

Metabolism, transcription, and translation are important for modulating growth rate. However, the expression changes for these systems are possibly controlled by alterations in transcriptional regulation [200]. In the evolved strains, there are mutations in several regulatory proteins, such as RNAP, Crp, Hfq, or AtoS [200, 201, 249]. Unfortunately, the normal wiring within these regulons is still not completely characterized. However, efforts are being made to identify the missing links in the *E. coli* transcriptional regulatory network (TRN) [216]. As genome-scale TRN models are completed and linked to the comprehensive transcription unit architecture for *E. coli* [266], greater insight into the scope of the regulatory changes in the evolved strains may be determined.

In conclusion, experimental adaptive evolution is a useful approach to develop an understanding of the metabolic genotype-phenotype relationship in bacteria and to aid in the identification of principles underlying evolution. To identify such principles, various types of data are being generated. For the strains in this study, these data types include the genome sequences, gene expression profiles, proteomic data, fluxomic data, and physiological data. The analysis of these omics data types using optimality properties of GEMs enables the elucidation of principles of distal causation and the identification of large-scale mechanisms that confer selected optimal phenotypes.

Methods

Parsimonious enzyme usage FBA (pFBA): pFBA is a bi-level linear programming optimization using the genome-scale constraint-based model of *E. coli* K-12 [189]. FBA was used to compute the optimal growth rate (using experimentally measured substrate uptake rates [217, 249], followed by a minimization of the sum of all gene-associated reaction fluxes while maintaining optimal growth. This proxy computes the pFBA optima, representing the set of genes associated with all maximum-growth, minimum-flux solutions, thereby predicting the most stoichiometrically efficient pathways. The idea underlying this method is similar to the "max biomass per unit flux" objective presented previously [97], but the mathematical implementation is different (see [47]).

Five classes of genes emerge, associated with reactions that 1) are essential for optimal and suboptimal growth, 2) are inside the pFBA optima, 3) are enzymatically less efficient (ELE), requiring more enzymatic steps than alternative pathways that meet the same cellular need, 4) are metabolically less efficient (MLE), requiring a reduction in growth rate if used, or 5) cannot carry a flux in the given environmental condition/genotype (pFBA no-flux).

Here, the pFBA optima were computed for wild-type *E. coli* under growth in lactate M9-minimal media, glycerol M9-minimal media, and a Δ*pgi* mutant on glucose M9-minimal media, using experimentally-measured substrate uptake rates (see [47]). The steps were as follow. First, each gene is removed from the model, and FBA was used to test gene essentiality. Second, Flux Variability Analysis (FVA) with no biomass constraint was conducted to identify reactions that cannot carry a flux. Third, FBA helped identify the optimal growth rate, which was subsequently set as a lower bound for the biomass function. Fourth, FVA was conducted again to find all metabolically less-efficient reactions. Fifth, flux through all gene-associated reactions was minimized using linear programming, and this flux was set as an upper bound for the summed network flux. Sixth, FVA was conducted on the model, holding the maximum predicted growth rate and minimum network flux constant, thereby identifying all reactions that are active in alternate optimal solutions [267]. Genes were assigned to the five categories as follow. All genes necessary for growth *in silico* were classified as "essential". Non-essential genes associated with reactions that were active when maximizing biomass and minimizing flux were classified as "pFBA optima" genes. Genes that were only associated with reactions that could not carry a flux were identified as "pFBA no-flux" genes. "ELE" genes were identified as those associated with reactions that could carry a flux while optimizing biomass, but not when minimizing flux (genes associated with the pFBA optima were filtered out). All remaining genes, which were associated with reactions that could carry a flux when not optimizing biomass, were classified at "MLE" genes.

The sets of pFBA genes and proteins were compared with all non-essential up-regulated proteins and mRNAs using the hypergeometric test to determine if there were more up-regulated proteins in the pFBA optima than expected by chance. A similar approach was used to find the enrichment and depletion of up- and down-regulated species in the essential, non-functional, and less efficient pathways; however, all genes were used for these tests. Significant tests are shown in Table 5.1. In addition, the significance of fold change within up- and down- regulation in the classes was tested by summing up all up- and all down-regulated genes within each class and then comparing to 10000 random sets of the same number of differentially expressed metabolic genes.

Regulon enrichment: Regulon structure was determined from RegulonDB 6.0 [268]. Significance of enrichment of regulons in up and down-regulated genes/proteins was determined using the hypergeometric test with a false discovery rate of 0.1. The results, however, were robust with FDR cutoff choice.

Gene expression profiling: Microarrays corresponding to the same glycerol and lactate-evolved strains in this study have been published previously as described in the corresponding studies [215, 217]. The arrays were re-normalized for this study using gcRMA. Genes which did not have a gene expression level significantly above a set of negative controls on the arrays (FDR = 0.05) were removed from the data set and were not considered in further analyses.

Cell preparation: *E. coli* K-12 MG1655 strains used for this study were prepared previously [217, 246]. Briefly, for the Δ*pgi* strains, *pgi* was removed as described in [269], and transferred to M9 glucose minimal media. Wild-type strains were also transferred to glycerol or lactate M9 minimal media. Adaptive evolution was conducted by growing the strains in batch culture until they reached mid-exponential growth. At this point the culture was diluted by serial passage into fresh media. The quantity of passaged cells was determined based on the growth rate from the previous day. Multiple replicates for each strain were evolved in parallel for about 700, 800, and 1000 generations for the glycerol, *pgi* deletion, and lactate evolved strains, respectively. Despite the different number of generations, all strains were evolved until they converged to a stable maximum growth rate which was maintained for at least 5 days [217, 246]. Growth rates and substrate uptake rates were determined and reported previously [217, 246]. Instantaneous steady-state biomass yields for the exponential growth phase (see [47]) were determined as similarly shown previously [72] by dividing the mid-exponential growth rate by the substrate uptake rate at that time:

$$Y_{B,E} = \frac{gr}{SUR},$$

where *gr* is the growth rate at mid-exponential phase (1/h) and *SUR* is the substrate uptake rate (g substrate / gDW biomass / h). This figure provides a measure of how efficiently the strains can convert substrate into biomass while in exponential growth.

For subsequent experiments, all strains were streaked out on solid media, and a single colony was then isolated, grown up, and frozen down. Glycerol stocks of each strain at day 1 and the evolution endpoint were placed in fresh media and grown up to an OD of 0.500 at 600nm. Cells were then pelleted, washed in PBS, and frozen prior to proteomic profiling.

Cell Lysis: Each cell pellet (~50 µL in size as measured in a microfuge tube) was resuspended in 1.5 mL of nanopure water. Lysis was achieved using pressure cycling technology with the Barocycler™ (Pressure BioSciences, West Bridgewater, MA) for 10 cycles going between ambient pressure for 20 sec and 2.4 x 10^5 kPa for 20 seconds. The lysate was collected and placed immediately on ice. Each lysate was concentrated down to about 500 µL using a speed vac (ThermoSavant, San

Jose, CA). The protein concentration of each cell lysate was measured using a Coomassie Plus protein assay (Pierce, Rockford, IL) using a bovine serum albumin standard.

Protein Reduction, Trypsin Digestion and Alkylation: Each lysate was dried down and 150 µL of 8M guanidine HCL, and 3 µL of Bond Breaker TCEP solution (Pierce, Rockford IL) was added. The samples were vortexed and incubated at 60 °C for 30 min. Iodoacetamide was added to a concentration of 20 mM and then each sample was incubated at room temperature for 30 min. The samples were diluted 10 fold with freshly prepared 50 mM ammonium bicarbonate solution, pH 7.8 and $CaCl_2$ was added to a final concentration of 1 mM. Finally, trypsin was added in a 1:50 (wt/wt) ratio of trypsin to sample protein, and the samples were digested at 37 °C for 4 hrs.

Peptide Concentration and Cleanup: Each digest was desalted using Supelco (St. Louis, MO) Supelclean C-18 tubes as described elsewhere [270]. Each sample was concentrated using vacuum centrifugation to adjust the concentration to be 1 mg/mL.

SCX Fractionation of Peptides and Data Preprocessing: 300 µg of a pooled sample of all glycerol adaptation samples, lactate adaptation samples, and Δ*pgi* study samples were fractionated separately into 25 SCX LC fractions for analysis using a LTQ iontrap mass spectrometer to obtain tandem MS (i.e. MS/MS) data for peptides as described previously [271]. The MS/MS spectra were analyzed using the peptide identification software SEQUEST [272] in conjunction with the annotated protein translations from the genome sequence of *E. coli*. 44,610 peptide identifications that met the criteria of: 1) a minimum XCorr value of 2; 2) a minimum discriminate score of 0.6 [273]; and 3) a Peptide Prophet Probability of at least 0.99 were used to build an Accurate Mass and Time (AMT) database with peptide sequences and normalized elution times.

Accurate Mass and Time Tag Analysis of Peptides: LC-MS spectra were analyzed using the accurate mass and elution time (AMT) tag approach [274]. A detailed description of this method is provided in the Supplementary Notes. The AMT tag approach, in the end, provides peptide identifications along with their abundances for all the data sets. The data for each peptide identified in each sample were represented by the median value obtained across the 3 LC-MS runs. These data were loaded into the software tool DAnTE [275] for further analysis. Peptide abundances were transformed to log base 2 and an outlier check was applied by observing the Pearson correlations between data sets. Any data sets with weak correlations were excluded from further analysis. A linear regression based normalization method available in DAnTE was then applied within each replicate category. The central tendency adjusted peptide abundances were used to infer the corresponding protein abundances via the 'Rrollup' algorithm in DAnTE [275]. During the Rrollup step, the Grubbs outlier test was applied with a p-value cutoff of 0.05 to further remove any outlying peptides. Protein expression values were computed with all data sets combined, and for individual growth conditions (see [47]) for differential expression analysis.

Computation of differential expression: Differential expression was computed for all identified proteins and all transcripts with a significantly higher expression than negative controls on the microarrays (FDR = 0.05). The grand mean was subtracted from the data sets of interest, and differentially expressed genes and proteins were determined with a two-sample t-test. False discovery rate cutoffs were determined as discussed in [276].

Chapter 5, in part, is a reprint of the material as it appears in 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., Palsson, B.Ø. Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.*, 6:390 (2010). I was the primary author, while the co-authors provided support in the research that served as the basis for this study.

Chapter 6: Evolved *E. coli* learns to optimize metabolism on a non-native substrate

A common assertion in evolution is that an organism evolves to improve its fitness. For example, studies have shown that the metabolic capabilities of microbes can evolve and that this plasticity likely facilitates the rapid adaptation to new and dynamic environments. Examples have surfaced showing how metabolism evolves both at the level of individual proteins and their organization in the metabolic network. At the level of a protein, new metabolic functions can be acquired as promiscuous enzyme activities are amplified [206, 235] or enhanced through mutation [225]. The recruitment of promiscuous activities for new environmental conditions is of particular interest, since enzyme promiscuity has allowed bacterial strains to metabolize synthetic compounds [277-280], and even grow on man-made antibiotics as the sole carbon source [281, 282]. In addition to these enzyme-level adaptations, the organization of enzymes in metabolism can evolve. Specific metabolic capacities encoded in the metabolic network can be lost through silencing and genome reduction if not needed [131] or they can be acquired through horizontal transfer [283]. Furthermore, adaptive laboratory evolution studies have shown that when bacteria is subjected to a growth-rate selective pressure, they can tune gene and protein expression to levels that are more consistent with model-predicted optimal metabolic efficiency phenotypes [47, 48, 185, 240].

Emergent metabolic functions from short-term laboratory evolution studies have been successfully associated with causal mutations on specific enzymes [198, 225, 226, 238], but have not addressed how the cell adapts to the downstream requirements of metabolizing the new substrates. After acquiring the ability to metabolize a new nutrient, can microbes also rapidly adapt the expression of enzymes throughout metabolism to optimally metabolize a new, non-native substrate? Also, if an emergent metabolic function is associated with specific improvements throughout the metabolic network, what are the mechanisms that underlie this adaptation?

To address these questions, we evolved *E. coli* to grow on a substrate upon which the wild-type strain cannot grow [238, 284]. Through two different evolutionary approaches, over the course of ~1400-1600 generations, *E. coli* learned to grow on minimal media with L-1,2-propandiol (L-1,2-pdo) as the sole carbon source. Expression profiles and proteomic data were obtained for these strains, and causal mutations were assessed. Through this study we show the following. Reproducible mutations are identified in all evolved strains, and these enhance L-1,2-pdo catabolism. Several lines of evidence affirm that the pathway for L-fucose metabolism is retooled for the first few steps of L-1,2-pdo metabolism. Moreover, changes in the transcriptome and proteome suggest that *E. coli* obtains the specific expression patterns throughout metabolism to efficiently metabolize L-1,2-pdo. Lastly, cAMP-receptor protein (CRP) is harnessed to help provide the expression state needed for optimal metabolism of L-1,2-pdo. Thus, as *E. coli* gains this new metabolic functionality, it can rapidly adapt both at the level of individual enzymes and the level of metabolic network expression, by modifying enzyme usage and transcription to optimize the utilization of the new substrate.

E. coli can learn to grow on L-1,2-propandiol

While *E. coli* K12 MG1655 is unable to metabolize L-1,2-propandiol, only a few genetic changes are required to grow with L-1,2-pdo as its sole carbon source [238, 285]. This adaptation was carried out using two different approaches. For the first approach, *E. coli* was grown on glycerol minimal media at mid-exponential phase for 44 days (~700 generations) [217] (strain GC). Subsequently, over 11 days (~250 generations) glycerol was gradually replaced with 1,2-propandiol until the glycerol had been completely replaced [238]. As previously reported, growth briefly stopped and then suddenly resumed after a couple days. Cultures were then maintained at mid-exponential growth on L-1,2-pdo for

46 days (~450 generations). At this point the cultures had reached and maintained a steady growth rate, which did not significantly increase for several days (strains PA and PB).

For the second approach, wild type *E. coli* was maintained at mid-exponential growth, initially in glycerol minimal media, supplemented with the mutagen N-Methyl-N′-nitro-N-nitrosoguanidine (NTG). Over the first 12 days of the evolution (~125 generations), glycerol was gradually removed as L-1,2-pdo was titrated in. Similar to the previous approach, growth briefly stopped, but resumed after a couple days. Exponential growth was maintained at this point for 105 days (~1500-1600 generations) [284] (strains PM1 and PM2). A strain was also grown exponentially in glycerol with mutagen for 59 days (1080 generations) for control purposes (GM1).

For both of these evolutionary approaches, strains were expression-profiled on glycerol M9 media prior to the complete removal of glycerol in the evolution (Day 45 and Day 8 for the non-mutagenized and the mutagenized evolutions, respectively). In like manner, at the end of the evolution experiment, strains were expression profiled on both L-1,2-pdo minimal media and glycerol minimal media. Proteomic data were also acquired at the same time points for one of the non-mutagenized evolutions. As the result of previous whole-genome resequencing results from a few L-1,2-pdo evolved strains [238, 284], a few genes were subjected to Sanger sequencing in all strains.

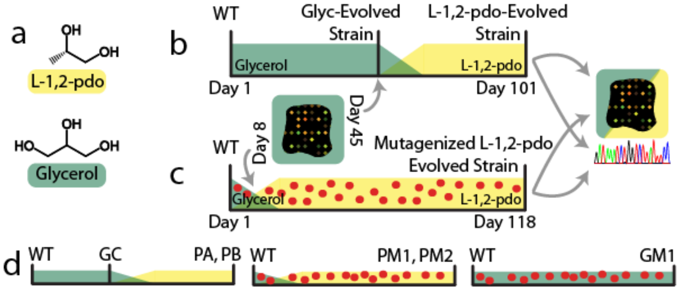


Figure 6.1. *E. coli* can be evolved to grow on a new substrate. *E. coli* was evolved to grow on L-1,2-propanediol (L-1,2-pdo) with two approaches. (a) Evolutions began on glycerol M9 minimal media, since L-1,2-pdo is structurally similar to glycerol. *E. coli* was maintained at exponential growth by passaging cells into fresh media each day. The cells evolved by (b) adapting to grow optimally on glycerol and then slowly replacing glycerol with L-1,2-pdo or (c) adding mutagen and rapidly replacing the glycerol with L-1,2-pdo. Strains were expression profiled before complete replacement of glycerol with L-1,2-pdo and at the end of the evolution experiment. In addition, the strains were further characterized through targeted resequencing, proteomic analysis, and phenotypic assays. (d) IDs by which the strains will be referenced throughout the text.

L-1,2-pdo is metabolized by reversing the anaerobic branch of L-fucose metabolism

The first laboratory evolution of an L-1,2-pdo metabolizing strain of *E. coli* K-12 was found to require the presence of a previously uncharacterized oxidoreductase that catalyzes the oxidation of L-1,2-pdo to L-lactaldehyde [285]. This enzyme was therefore named L-1,2-propanediol oxidoreductase (POR), encoded by the *fucO* gene. Subsequent studies found that in WT *E. coli*, POR participates in the anaerobic fermentation of L-fucose through an L-lactaldehyde intermediate to form L-1,2-pdo, which is secreted as a waste product (Figure 6.2.a). However, when L-fucose is metabolized aerobically, 1,2-pdo is not synthesized, since the Fe^{2+} ion is oxidized under aerobic conditions to avoid the wasteful secretion of L-1,2-pdo [286]. Instead L-lactaldehyde is oxidized to form L-lactate [287], which subsequently enters central metabolism (Figure 6.2.a). Thus, it has been suggested that as *E. coli* evolves to grow on this new substrate, existing enzymes are employed. However, the enzymes are retooled and combined to form a pathway for the new substrate. That is, the flux through POR is reversed and the anaerobic POR enzyme is used in aerobic metabolism.

While the pathway has been suggested previously and seems intuitive, it has only been indirectly validated and no tests have demonstrated that the network can recycle all of the cofactors required for its metabolism. Therefore, here we provide several additional lines of experimental evidence and systems-level support that this pathway is likely used for the evolved metabolic capability of L-1,2-pdo.

Mutations to the FucAO operon allow the metabolism of L-1,2-pdo

If a simple reversal of flux is needed to metabolize L-1,2-pdo, why cannot wildtype *E. coli* metabolize this substrate? Two independent studies reported that evolved L-1,2-pdo catabolizing strains required a constitutive aerobic expression of POR, and found that mutations occurred that enhanced POR's stability and kinetic properties with respect to aerobic L-1,2-pdo oxidation [238, 286]. Interestingly, similar mutations were found on all L-1,2-pdo-evolved strains presented here.

Constitutive expression of *fucO* was previously seen when an IS5 insertion element appeared in the *fucAO* promoter [238, 288]. Consistent with this finding, all of the end point strains reported here contain an IS5 element in the *fucAO* promoter. Interestingly, all 15 end-point clones tested from the evolutions in this study had the IS5 element inserted about 210 nucleotides upstream of the *fucAO* transcription start site.

Mutations associated with enhanced catalytic properties and protein stability were also found in all tested clones. Previously, a mutation in POR (L8M) in the PA strain was identified that enhanced the V_{max} by 10 fold with respect to reversing the native anaerobic reaction and encouraging the oxidation of L-1,2-pdo [238]. In addition, I7L and L8V mutations have been shown to reduce the occurrence of metal-catalyzed oxidation of the Fe^{2+} ions in POR, thereby improving its activity in aerobic conditions [288]. These substitutions reside in an 11 residue group that contributes to the POR dimerization interface [289] and is essential for POR function. Interestingly, all evolved strains contained mutations within this essential 11 residue stretch (I7L for the PB strain and R5L for PM1 and PM2).

Proposed uptake pathway is supported by gene and protein expression

While mutations clearly target the first enzyme of the proposed pathway, the question remains as to if the remaining steps of the pathway are used in the metabolism of L-1,2-pdo. qPCR results previously showed that *aldA* and *lldD* are up-regulated in the PA evolved strain [238]. To see if this occurs in the other end-point strains, we expression profiled all strains 1) on glycerol before they were adapted to L-1,2-pdo, 2) on L-1,2-pdo after the adaptation, and 3) after returning the end-point strains to glycerol (Figure 6.1). To differentiate between evolutionarily-important expression changes and those specific to NTG treatment, the GM1 strain (evolved with NTG on glycerol) was expression profiled as a control.

All L-1,2-pdo-evolved strains demonstrated significant up-regulation of *fucO*, *aldA*, and *lldD* when growing on L-1,2-pdo (Figure 6.2.b). The expression of *aldA* and *lldD* subsequently decreased when the strains were returned to glycerol minimal media. Similar gene expression levels were not seen in the glycerol-evolved GC and GM1 strains. The GC and PA strains were further assayed by quantitative proteomics. In comparison to the pre-L-1,2-pdo evolution GC strain, all three enzymes were strongly up-regulated in the L-1,2-pdo evolved PA strain while growing on L-1,2-pdo (Figure 6.2.c), and the constitutively-expressed POR was still highly expressed after being returned to glycerol-minimal media. Thus, gene and protein expression of the L-1,2-pdo evolved strains support the hypothesis that this pathway is used in L-1,2-pdo metabolism.

The proposed uptake pathway can maintain flux

The pathway detailed here seems simple and transcriptomic and proteomic data provide experimental support for it, but two questions remain with respect to how this pathway fits in the context of the entire metabolic network. First, this proposed pathway requires several cofactors that must be recycled. Can the cell balance the usage of all of these cofactors? Second, are there alternative pathways that can be employed? To investigate these questions, flux through all metabolic pathways was simulated for growth on L-1,2-pdo, using a genome-scale model of *E. coli* metabolism [189]. For these simulations, Markov chain Monte Carlo sampling [35] was used to look at all possible flux distributions that the cells could support at steady state, while maintaining 95% of the optimal growth rate on L-1,2-pdo minimal media. These simulations provided strong support for the usage of the proposed pathway. *In silico*, POR converts L-1,2-pdo to L-lactaldehyde, which is subsequently oxidized by lactaldehyde dehydrogenase (Figure 6.2.d). No known alternative pathways can catalyze this. L-lactate is subsequently oxidized primarily using quinone as a cofactor, and to a lesser degree menaquinone. Thus, as POR flux is reversed from its normal role in L-fucose fermentation, the rest of the proposed pathway can maintain flux and the metabolic network is able to adequately recycle all of the cofactors. Moreover, unless unknown enzymes are participating, this pathway is responsible for metabolizing L-1,2-pdo.

Gene deletions affirm the uptake pathway

Model simulations suggest that no known alternative pathways exist that metabolize L-1,2-pdo. However, can any promiscuous activities or unknown enzymes account for the metabolism of L-1,2-pdo? To test this, each gene in the proposed pathway was deleted from the PA strain. The growth phenotypes of these deletion strains were compared to flux balance analysis (FBA) predicted growth phenotypes, which predict the use of this pathway. Both the *ΔfucO* and *ΔaldA* mutants were unable to grow *in silico* or *in vivo*, suggesting that no other enzymes are contributing to the catalysis of these or supporting reactions (Figure 6.2.e). On the other hand, the *ΔlldD* mutants were able to grow. However, growth was lost with the subsequent deletion of the *ykgEFG* operon, an operon showing homology to a recently-discovered operon with L-lactate dehydrogenase activity in *Shewanella oneidensis* [290]. While the *ykgEFG* operon seems to provide lactate dehydrogenase activity to rescue the *ΔlldD* mutants, it probably does not provide most of the lactate dehydrogenase activity in the evolved strains, since the genes show lower expression in all of the evolved strains (data not shown), and little or no up-regulation through the laboratory evolution. Thus, these gene deletion mutants clearly demonstrate that *fucO*, *aldA* and *lldD* constitute the pathway that connects L-1,2-pdo metabolism to central metabolism.

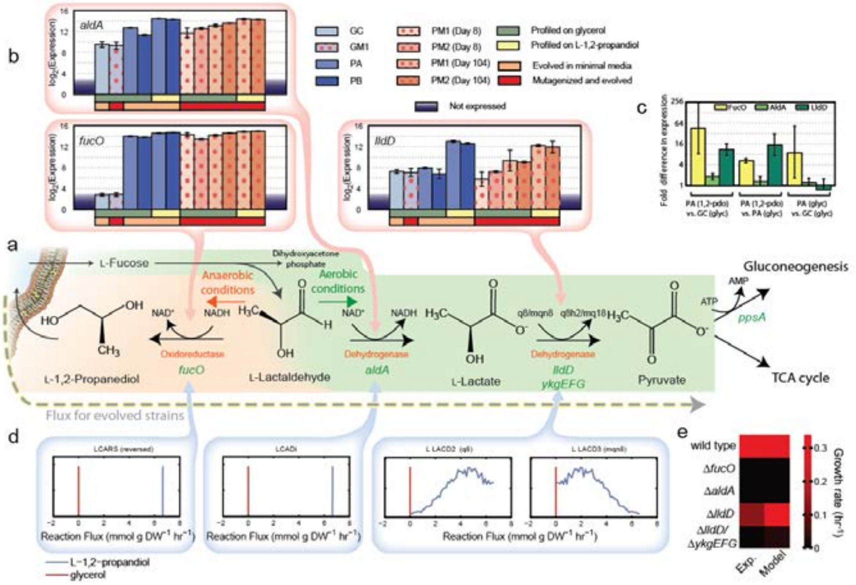


Figure 6.2. The pathway used to metabolize L-1,2-propanediol. (a) In wild type *E. coli* K12, L-lactaldehyde, an intermediate in L-fucose metabolism, enters central metabolism under aerobic conditions or is converted to L-1,2-pdo and secreted as a byproduct under anaerobic conditions. The evolved *E. coli* is able to reverse the anaerobic process and metabolize L-1,2-pdo under aerobic conditions. This is supported by several lines of evidence. (b) *fucO* mRNA is highly expressed in all strains that have been adapted on L-1,2-pdo, but is not expressed in strains with no such adaptation. The remaining genes in the pathway show slightly higher expression when evolved strains are grown on L-1,2-pdo (bars with yellow underneath). (c) Quantitative proteomic data shows a significant increase in the abundance of all enzymes in the pathway when the PA evolved strain is grown on L-1,2-pdo. (d) *In silico* flux through the pathway is simulated using MCMC sampling, and kernel density plots demonstrate that all feasible flux solutions use this pathway and no known alternative pathways exist that carry L-1,2-pdo to central metabolism. (e) This is further supported by *in vivo* and *in silico* gene deletion phenotypes.

Evolution enhances gene expression in many pathways to support growth on L-1,2-pdo

Each substrate an organism metabolizes requires a unique combination of reactions for its efficient metabolism. Even the nearest consumable substrates to L-1,2-pdo (i.e., L-fucose and L-lactate) require significantly different flux through many metabolic pathways, due in part to differences in redox balance and products from the unique reaction usage. Can *E. coli* quickly adapt to these substrate-specific requirements throughout the metabolic network when it encounters a new substrate, or does it only induce the immediate pathway needed to infuse the new substrate into central metabolism? (Figure 6.3.a)

To address this question, the gene expression data were compared with model simulations that predict which enzymatic reactions and pathways are needed for optimal L-1,2-pdo metabolism (Figure 6.3.b). This comparison was done as follows, using a method called Gene Inactivity Moderated by Metabolism and Expression (GIMME) [110]. GIMME first finds all reactions associated with highly expressed genes in the given growth condition (Figure 6.3.b.i). The metabolic network model is then used to identify the minimal reaction set with low gene expression that must be added to allow the cell to grow as witnessed experimentally (Figure 6.3.b.ii). Through this process, a score is assigned, called the normalized consistency score (NCS). The NCS is a quantitative measure of how fit the expression data are for achieving the measured phenotype (e.g., growth on L-1,2-pdo minimal media), given the predicted metabolic network requirements.

GIMME was used to compute the consistency of each dataset (before and after evolution) with the pathway requirements to maintain at least 90% of the optimal growth rate on different media conditions. If the cultures improve gene expression for a specific growth condition, the NCS for that condition should increase. To test if expression improved for growth on L-1,2-pdo, we compared the parent strains growing on glycerol (day 45 and day 8 for the evolved and NTG-evolved strains, respectively) to the L-1,2-pdo-evolved strains after being returned to glycerol. This was done because arrays from strains on L-1,2-pdo could not be reliably obtained for early L-1,2-pdo-evolved strains (i.e., the first few days following the replacement of glycerol with L-1,2-pdo in the media). Across gene expression data sets for all L-1,2-pdo-evolved strains, the NCS improved for optimal growth on L-1,2-

pdo, while there was no significant improvement for glycerol metabolism (Figure 6.3.c). Moreover, the improvement in NCS for L-1,2-pdo metabolism was significantly greater than glycerol ($p = 2.2 \times 10^{-7}$; one-sided Wilcoxon rank-sum test). This improved NSC resulted in part by improvements in specific pathways in amino acid metabolism, folate metabolism, and central metabolism (Figure 6.3.d). The improvements to amino acid metabolism may result from a previously-reported relaxation of the stringent response seen in laboratory-evolved *E. coli* [47, 199]. However, since the NCS is significantly higher for L-1,2-pdo and many other pathways are improved, it seems that the L-1,2-pdo evolution enhances a diverse range of metabolic pathways necessary specifically for L-1,2-pdo metabolism throughput the metabolic network.

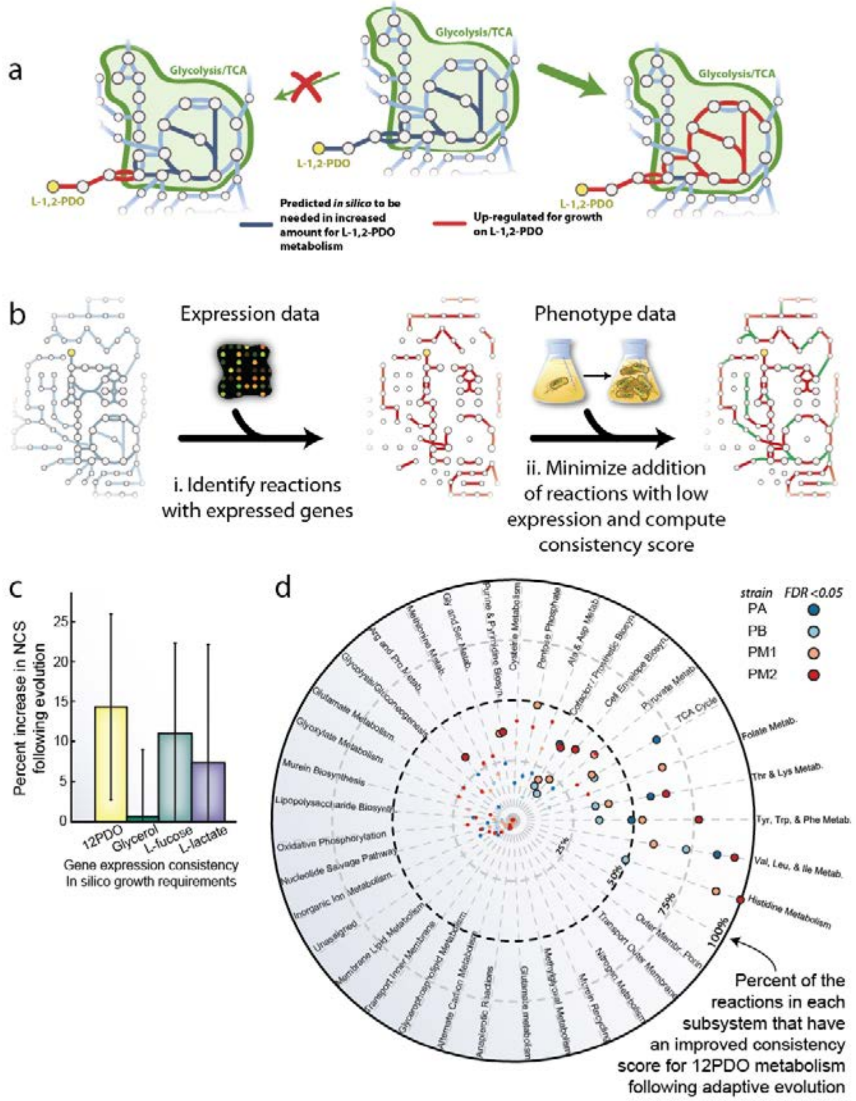


Figure 6.3. Evolution specifically enhances L-1,2-pdo metabolism throughout the metabolic network. The metabolism of each carbon substrate usually requires a unique balance of many different metabolic pathways. (a) Thus, as *E. coli* evolves to grow on L-1,2-pdo, it not only induces the immediate pathway that sends the new substrate into central metabolism, but it further adapts the expression of more distal metabolic pathways to improve metabolic efficiency for the new substrate. (b) This is assessed using the GIMME method, which (i) identifies all metabolic pathways with high gene expression values and then (ii) finds to minimal set of low-expression reactions that must be added to obtain the measured phenotype. Model flux and gene expression levels are used to assign a quantitative score of how well the data fit the measured phenotype (e.g., growth on L-1,2-pdo), called the normalized consistency score (NCS). (c) Through the different evolutions, the NCS improves for growth on L-1,2-pdo, with significantly less improvement for the similar substrates glycerol, L-fucose, and L-lactate. Mean NCS improvement values from jackknife resampling results are plotted here, with error bars representing ±1 S.D. of NCS for all strains. (d) For each evolution endpoint, improvements in the NCS for L-1,2-pdo growth can be attributed to several subsystems throughout the metabolic network, including amino acid metabolic pathways, and a significant improvement in TCA cycle fitness.

Evolved strains show L-1,2-pdo-specific expression regulation

The improved NCS for L-1,2-pdo metabolism suggests that *E. coli* shapes its gene expression throughout the metabolic network to meet the unique requirements of L-1,2-pdo. Moreover, when GIMME is used to compare the expression data to the metabolic requirement of the two nearest consumable metabolites, L-lactate and L-fucose, these substrates also yield lower NCS improvements than seen for L-1,2-pdo ($p = 0.018$ and $p = 0.047$, respectively; one-sided Wilcoxon rank-sum test; Figure 6.3.c). Thus, as *E. coli* adapts to this new substrate, its gene expression program migrates toward the expression state needed specifically for efficient L-1,2-pdo metabolism. However, an improvement in the NCS only represents a general metric for an improvement in gene expression for growth on L-1,2-pdo. Does this coarse-grained metric correspond to tangible improvements in the expression of specific genes needed for L-1,2-pdo metabolism?

To address this question, a different approach was employed to predict which specific enzymes were needed in particularly higher or lower quantities to grow specifically on L-1,2-pdo. While this approach has been described in detail previously [42, 78, 86], we briefly review it here. In this approach, all possible steady-state flux distributions are sampled using Markov chain Monte Carlo sampling [35]. This provides a distribution of feasible reaction flux for each reaction. When flux is simulated under two different conditions, the distributions of feasible flux values can be compared. If, for a given reaction, the flux magnitude significantly changes, its associated gene and protein expression usually changes accordingly. Using this approach, we predicted which metabolic genes and proteins are needed in increased or decreased quantities when the L-1,2-pdo-evolved strains are transferred to glycerol minimal media (Figure 6.4.a). When these predictions are compared with experimentally-measured transcriptomic and proteomic data (Figures 6.4.b-d), most differentially-expressed genes and proteins change as predicted *in silico* to support near-optimal metabolism on the respective substrates. Thus, the evolved strain has adapted its transcriptional program to more efficiently metabolize the new substrate.

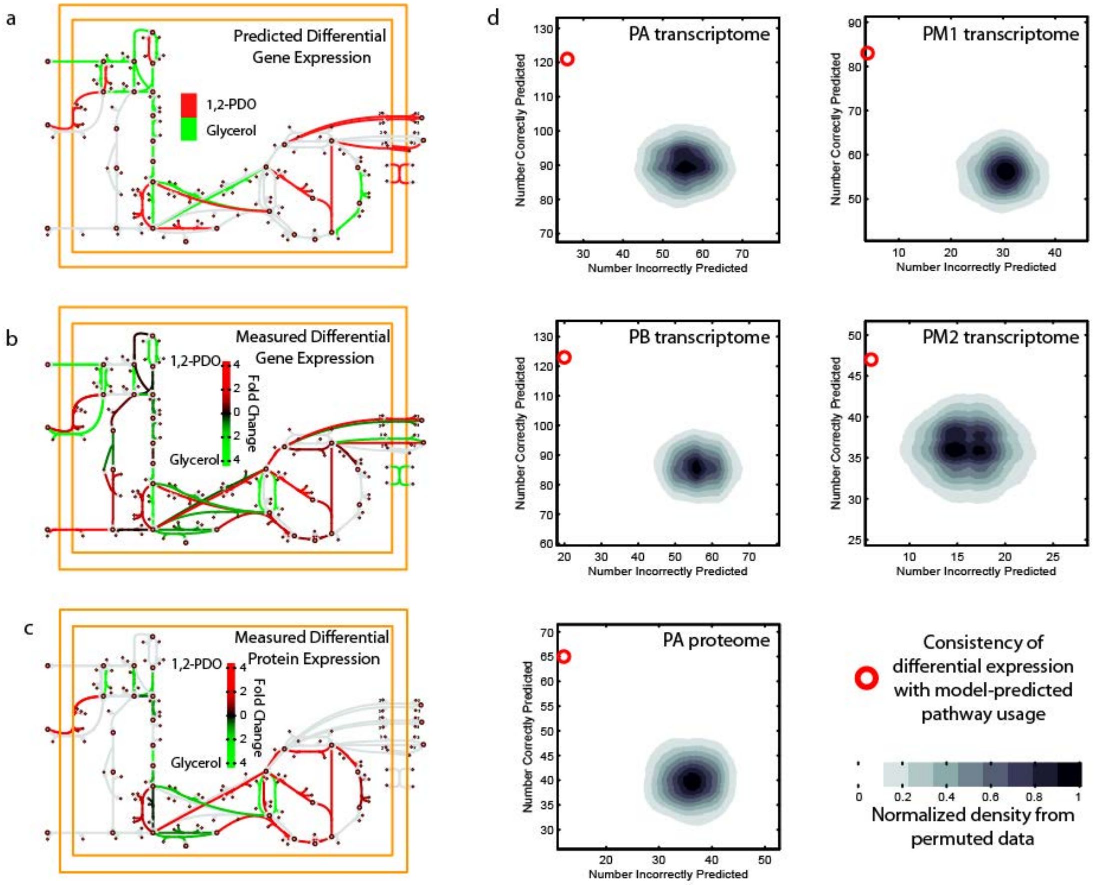


Figure 6.4. Expression is shaped to support model-predicted expression of specific genes and proteins for L-1,2-pdo metabolism. (a) By comparing randomly sampled flux loads for each reaction on glycerol and L-1,2-pdo minimal media, specific genes and proteins are identified that are required for efficient growth on either of the growth conditions. Enzymes that require significantly higher flux on glycerol or L-1,2-pdo are often consistent with differential (b) gene or (c) protein expression, as shown here for glycolysis and the TCA cycle for the PA strain. (d) All L-1,2-pdo-evolved strains show higher (or lower) gene expression for specific enzymes that are predicted to require a higher (or lower) flux load on L-1,2-pdo. Few genes deviate from this trend, and far fewer deviate than expected from permuted data.

Gene expression enhancements for L-1,2-pdo metabolism are facilitated by CRP

The transcriptomic and proteomic data show that *E. coli* gene expression evolves to meet the requirements for model-predicted optimal growth. At the same time, extant transcriptional programs are retained, since the evolved strains can mostly return to the glycerol growth expression state. These lines of evidence both suggest that as the cells evolve to grow on L-1,2-pdo, they are able to harness a preexisting transcriptional regulation program to help optimize the metabolic network for this new growth state. To identify such candidate transcription factors, known regulons were systematically screened against the expression profiles to identify regulons that were most consistent with the differential expression patterns seen when the L-1,2-pdo-evolved strains were returned to glycerol minimal media (Figure 6.5.a). Specifically, activator/repressor activities from all 165 regulons from RegulonDB v6.0 [268] were compared to statistically-significant up and down regulation in the transcriptomic data. For each evolution endpoint, gene expression was consistent with the activity of a few regulons. However, only CRP was significant for all L-1,2-pdo-evolved strains, but not activated in the control mutagenized glycerol-evolved strain. Thus, to enhance growth on L-1,2-pdo, the CRP regulon is activated.

CRP is activated when intracellular cAMP levels increase. Consistent with this concept, we found that cAMP levels decreased when the PA strain was transferred back to glycerol minimal media (Figure 6.5.b). It is anticipated that the increased cAMP level leads to the binding of cAMP to CRP, and then cAMP-CRP activates the genes needed for growth on L-1,2-pdo. However, cAMP-CRP activates several additional operons needed for the metabolism of other sub-optimal carbon substrates. Therefore, a number of unnecessary genes are up-regulated in the L-1,2-pdo evolved strains (e.g., genes for propionate metabolism).

The mutagenized strains show a weaker activation of the CRP regulon(Figure 6.5.c). Interestingly, CRP-regulated genes that are regulated in the non-mutagenized strains, but not in the mutagenized strains include genes that are predicted to not be necessary for growth on L-1,2-pdo and essential genes that are normally repressed by CRP. In addition, these genes are only known to be regulated by CRP, according to RegulonDB v6.0. On the other hand, genes that continue to be regulated by CRP are all regulated by other transcription factors. It is possible that these additional transcription factors could be still regulating these genes, even after CRP activity is relaxed.

The question remains as to why CRP is less active in the mutagenized strains. CRP gene expression does not vary significantly between the various strains before and after adaptation. However, both mutagenized strains acquired a G774S mutation in the regulatory region of adenylate cyclase. Current studies are now underway to see how this mutation affects the synthesis of cAMP, and therefore CRP activity. Preliminary analysis suggests that this mutation may decrease cAMP levels, thereby suppressing CRP's regulatory role. Thus, if the final experiments confirm this, it will seem that the mutagenized strains acquired an adaptive mutation that is able to further enhance expression in a manner that is specific to the metabolism of L-1,2-pdo.

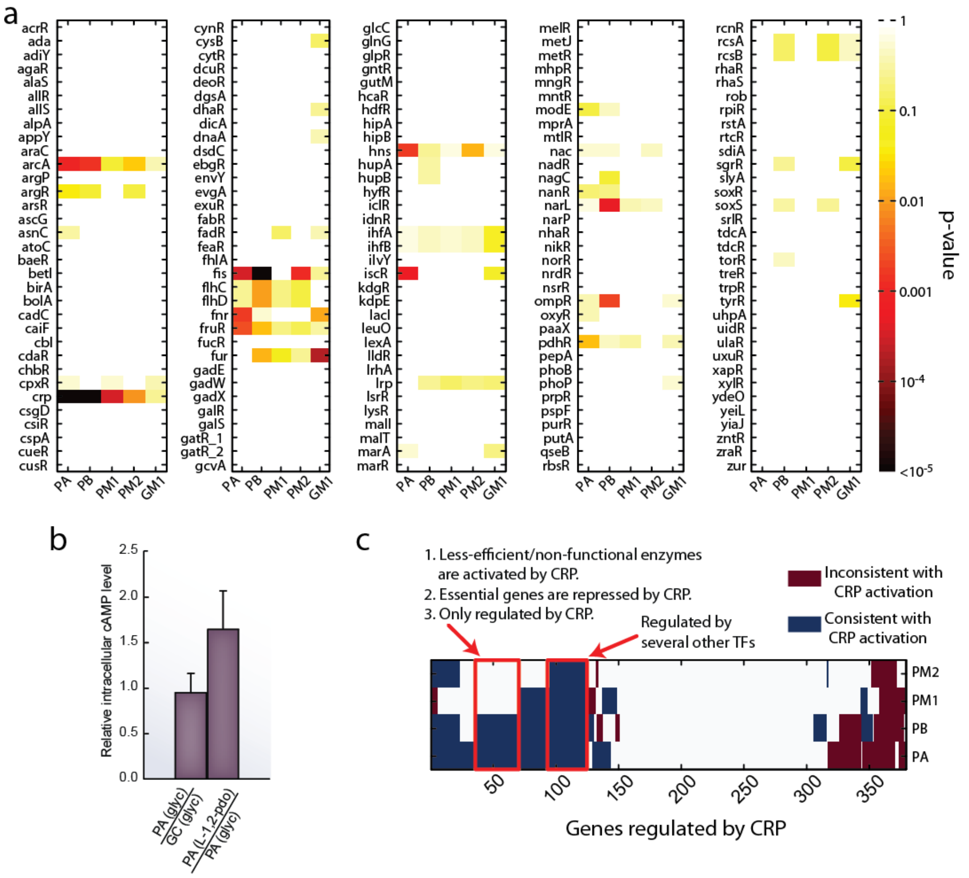


Figure 6.5. A systematic screen of regulon activity in the transcriptomes of evolved strains identifies CRP as an important regulator in aiding the evolution of *E. coli* to metabolize L-1,2-pdo. Activator/repressor knowledge for each transcription regulator was quantitatively compared to the differential gene expression when L-1,2-pdo-evolved strains were returned to glycerol. Differential gene expression over the course of evolution on mutagen and glycerol (GM1) was used as a control to account for the influence of mutagen. (b) The cAMP levels were measured for the PA and GC strains, and from this it is clear that cAMP levels increase when PA is growing on L-1,2-pdo. (c) Differential gene expression was compared with all genes that are reported to be regulated by CRP. Genes with higher (or lower) expression on L-1,2-pdo and known to be activated (or repressed) by CRP are "consistent" with CRP being active on L-1,2-pdo and shown in blue. Genes that are not significantly differentially expressed are white. It seems that CRP is activated in all conditions, though the activation is much lower for the mutagenized strains. Most genes that are still "consistent" with CRP activity in the mutagenized strains are known to be regulated by other transcription factors.

Experimentally-elucidated mechanisms of distal causation

A common assertion in evolution is that an organism evolves to improve its fitness, although it is often not immediately clear what mechanisms drive optimization and what the evolutionary optimal phenotype might be. However, genome-scale metabolic network models can connect the nutritional environment with optimal growth fitness by accounting for all of the known biochemical reactions that occur in the cell and linking these together to simulate cell growth [10]. These models thereby provide a mechanistic link between the metabolic genotype and the phenotype of a cell, given knowledge of a cell's environment. Thus, the use of such a model in this study provided insight into how a new environment influences the evolution of a new metabolic capacity. The growth rate selective pressure in the ALE experiments first enables the emergence of the ability to grow solely on a new carbon substrate and subsequently guides the metabolism of this new substrate toward a state of enhanced metabolic efficiency. These results have fundamental implications with respect to the role of optimality in the evolution of enzymes and the biomolecular networks in which they reside.

It has been proposed that through evolution, the introduction of a new environmental condition may induce the emergence of novel physiological metabolic functions through amplification and mutation of existing proteins [222, 224]. Specifically, if a fitness advantage is provided, an enzyme may be amplified when it contains a fold exhibiting a weak promiscuous catalytic activity for the new nutrient [291]. Several recent studies support this hypothesis. For example, several auxotrophic strains of *E. coli* were rescued when semi-random protein sequences (designed to form four helix bundles) were overexpressed [236]. Similarly, several native enzymes can be overexpressed in *E. coli* to rescue a mutant lacking 4-phosphoerythronate dehydrogenase [206] or transaldolase [292], and another study found that ~20% of 104 *E. coli* auxotrophs could be rescued by overexpressing unrelated *E. coli* proteins [293]. Furthermore, when any one of 61 native proteins are overexpressed in *E. coli*, resistance to a number of toxins and antibiotics is conferred [235]. Here we found that a mutation in the promoter region of *fucAO*, leads to the constitutive expression of the POR protein, which can break down L-1,2-pdo.

Models of enzyme evolution also suggest that the specificity of these amplified promiscuous activities is enhanced through mutation and selection. It has been proposed that mutations will randomly occur, and non-deleterious mutations can be retained. In particular, mutations providing a fitness advantage, without significantly destabilizing the native function, will be fixed in the population. The feasibility of this process has been demonstrated extensively using directed evolution approaches [228, 294], and ALE approaches in which the regulatory properties of GlpK [227] and the catalytic properties of ProA and ProB [225] were enhanced to meet the demands of a new environment. The mutations found in L-1,2-pdo-evolved strains also increased the catalytic efficiency for the reverse reaction in POR. Moreover, it has been suggested that these mutations stabilize the Fe^{2+} ion in POR, thereby protecting the ion from oxidation and allowing it to function in aerobic conditions [288].

A key component that is assumed by these evolutionary models, but never explicitly addressed, is that the products of these evolved enzymes must enter the metabolic network to generate energy and/or biomass. This assumption, however, is non-trivial. Metabolism is a complex and highly interconnected network, and its chemical reactions are constrained by stoichiometry and thermodynamics. In these networks, all cofactors must be balanced and ample energy must be generated to recycle the cofactors. Therefore, breaking down a new substrate to a metabolite that is recognized by another enzyme in the cell is not enough. Pathways must exist that will yield a net gain in energy or biomass production for the cell to use the metabolite. Furthermore, the combination of pathways needed for a new substrate may conflict with transcription regulation [295], which has often been optimized to enhance metabolism for other environmental conditions [26]. Previous ALE studies have shown that the expression of metabolic pathways can change to improve fitness on defined media by up-regulating important enzymes and suppressing unnecessary proteins [47, 185]. This study builds upon those previous studies by demonstrating that preexisting transcriptional regulators can actually respond to facilitate the improvement of the entire metabolic network for a new substrate.

We anticipate that this is facilitated by the fact that the metabolism of L-1,2-pdo leads to the synthesis of L-lactate, a substance that the organism already could consume. Moreover, L-lactate has previously shown a level of catabolite derepression [296], potentially through CRP activity, since L-lactate oxidation was previously inhibited in a CRP mutant [295]. Thus, the evolution of a new metabolic function may be facilitated by transcriptional regulatory mechanisms previously used for the metabolism of similar substrates. However, this study goes further to demonstrate that these transcriptional programs can be shaped for a new substrate. That is, the results from the GIMME analysis suggest that while CRP aided in the adaptive process, additional mechanisms provided further refinement of the transcriptional response to enhance the specific requirements for L-1,2-pdo metabolism. Moreover, adaptation with mutagen showed that the mutation to adenylate cyclase helped to soften the CRP response, presumably to decrease the expression of genes that are unnecessary for L-1,2-pdo metabolism. Thus, while the adaptation to L-1,2-pdo relied on existing mechanisms, this evolution provides examples of how cells will evolve to enhance the efficiency of metabolism, even for new metabolic functions that are more complicated to attain.

In conclusion, this study provides an example of how the use of a systems-level analysis of adaptive laboratory evolution [7, 47, 198, 297] can provide fundamental insights into the evolution and diversification of microbial species as they respond to variations in their environmental conditions [239]. Moreover, it demonstrates that in well-controlled conditions, distal causation can be estimated through the use of detailed genome-scale models of metabolism.

Methods

Strains and culture condition: For transcriptome and proteome analyses, the PA L-1,2-PDO-evolved *E. coli* K-12 MG1655 (eBOP12-6 strain) [238] and its parental glycerol-evolved *E. coli* (GC strain, F^{-} λ^{-} *ilvG*- *rfb*-50 *rph*-1 *glpK*$^{G\text{-}184\rightarrow T}$) [201] were grown from freezer stocks in M9 minimal media supplemented with 2 g/L of L-1,2-PDO or glycerol (Sigma Aldrich) at 37°C. M9 media contained (per liter of deionized water) 0.8 g of NH_4Cl, 0.5 g of NaCl, 7.5 g of $Na_2HPO_4{\cdot}2H_2O$, and 3.0 g of KH_2PO_4. The following components were sterilized separately and then added (per liter final volume of media): 2 mL of 1 M $MgSO_4$, 0.1 mL of 1 M $CaCl_2$, and 0.5 mL of a trace element solution containing (per liter) 1 g of $FeCl_3{\cdot}6H_2O$, 0.18 g of $ZnSO_4{\cdot}7H_2O$, 0.12 g of $CuCl_2{\cdot}2H_2O$, 0.12 g of $MnSO_4{\cdot}H_2O$, and 0.18 g of $CoCl_2{\cdot}6H_2O$. To determine the fitness gain in propionate, the PA eBOP12-6 strain was grown in M9 minimal media containing 2 g/L of sodium propionate using magnetic stir bars for aeration at 37°C. Growth rates were determined by measuring the optical density at 600 nm (OD_{600}) of triplicate cultures over several time points in which $0.05 < OD_{600} < 0.3$. Growth rates were defined as the slope of the linear best-fit line through a plot of ln(OD_{600}) versus time. During the knock-out process, the strains were cultured on Luria-Bertani media supplemented with 50 µg/mL of kanamycin or 100 µg/mL of ampicillin when necessary.

Generation of mutant strains: The knock-out *E. coli* mutants were generated by homologous recombination using the lambda Red recombinase system [269] on the PA strain. In short, the gene to be deleted was replaced by a kanamycin gene flanked by flippase recognition target sites and the insert was removed with a flippase recombination enzyme. In order to verify the genotypes of all mutants, colonies were isolated from solid media and tested with PCR. Wild-type *E. coli* colonies were tested in parallel as a negative control.

Transcriptome analysis: Affymetrix *E. coli* Antisense Genome Arrays were used for all transcriptional analyses. Each experimental condition was tested in triplicate on the carbon source used for evolution (glycerol or L-1,2-PDO) using independent cultures and processed following the manufacturer-recommended protocols. Cultures were grown to mid-exponential growth phase aerobically ($OD_{600} \approx 0.3$) in minimal media supplemented with appropriate carbon source. Three ml of cultures were added to 2 volumes of RNAprotect Bacteria Reagent (Qiagen) and total RNA was then isolated using RNeasy columns (Qiagen) with DNase I treatment. Total RNA yields and quality were measured using a Nanodrop 1000 (Thermo Scientific) and agarose gels. cDNA synthesis, fragmentation, end-terminus biotin labeling, and array hybridization were performed as recommended by the Affymetrix standard protocol. Raw CEL files were normalized using gcrma as previously described [215]. Genes that did not have an expression level above a set of negative controls on the arrays ($FDR \leq 0.05$) were removed from the dataset and not considered in further analyses. Significant differential expression was statistically assessed using the SAM test ($FDR \leq 0.05$). Fold change values for genes that are not statistically differentially expressed were obtained and a fold change cutoff for differential expression was set as 2 standard deviations beyond the mean fold change for these genes.

Proteome analysis: Frozen (-80°C) stocks of PA eBOP12-6 and GC strains were used to start cultures which were grown overnight in 3 mL of M9 minimal media containing 2 g/L of L-1,2-PDO and glycerol, respectively. Each was sub-cultured 1:100 into 50 mL of M9 minimal media and allowed to grow with shaking. Cells were harvested at mid-exponential phases of growth ($OD_{600} \approx 0.3$) by centrifugation for 30 min at 5,000 ×g, the supernatant discarded, and the remaining pellets were flash-frozen in an ethanol/dry-ice bath. The pellets were stored at -80°C until they were transported on dry-ice to the Pacific Northwest National Laboratory (PNNL; Richland, WA). Three independent biological samples were separately analyzed in triplicate (resulting in a total of nine datasets) for each condition for each strain. Samples were prepared as previously described [47]. Briefly, lysis of each sample was achieved using pressure cycling (via a barocycler), and whole cell lysate was reduced, alkylated, and tryptically digested. Desalted peptides were separated by reversed-phase liquid chromatography (LC) and detected online by a coupled Thermo Scientific LTQ Orbitrap Velos mass spectrometer. Peptides were identified by comparison with an established accurate mass and time (AMT) tag database as described previously [47]. Protein intensities for each sample were calculated by the RRollup feature of DAnTE [275], and changes in protein abundance were calculated from ratios of average intensities in corresponding samples. Analysis of variance (ANOVA) was also performed using DAnTE and used to determine the significance of individual peptide identifications. For identification of each protein, corresponding peptide ANOVA q-values were averaged to generate a protein level score. Proteins were considered differentially abundant if average ANOVA q-value ≤ 0.05, fold change ≥ 2.0, and proteins were identified in at least two of 3 biological replicates for each condition. Inherent to bottom-up proteomic technology, intensity values are not determined for all possible proteins. These "missing" values can result when a protein has low abundance or when a protein is not expressed. For those proteins for which an ANOVA could not be performed due to missing values across datasets, proteins were deemed significant if identification was made for each biological replicate in one condition and not for any biological replicates in the alternate condition. Specific details regarding sample preparation, LC-MS analysis, MS instrument settings, and a detailed description of the AMT tag methodology [274, 298, 299] are provided elsewhere [47]. RAW files for all datasets may be downloaded at http://omics.pnl.gov [300]

Constraint-based modeling: All constraint-based model simulations were done using a genome-scale metabolic model of *E. coli* K-12 MG1655 metabolism as previously published [189]. The *ykgEFG* operon was added to the model based on experimental evidence of this operon contributing lactate dehydrogenase activity [290]. Calculations were conducted in Matlab 7.6 using the COBRA Toolbox 2.0 [214] with the Tomlab CPLEX solver.

Markov chain Monte Carlo sampling: The distribution of feasible fluxes for each reaction in the models used here were determined using Markov chain Monte Carlo (MCMC) sampling [35], as previously described [42, 86], and was implemented with the COBRA Toolbox v2.0 [214]. Published uptake rates were used to constrain the models. To model more realistic growth conditions [87], sub-optimal growth was modeled. Specifically, the biomass objective function (a proxy for growth rate) was provided a lower bound of 95% of the optimal growth rate as computed by flux balance analysis [43]. Thus, the sampled flux distributions represented sub-optimal flux-distributions, while still modeling fluxes relevant to cell growth and maintenance.

MCMC sampling was used to obtain thousands of feasible flux distributions (referred to here as “points”) using the artificially centered hit-and-run algorithm with slight modifications, as described elsewhere [42, 86]. Briefly, a set of non-uniform points was generated. Each point was subsequently moved randomly, while remaining within the feasible flux space. To do this, a random direction is first chosen. Second, the limit for how far the point can travel in the randomly-chosen direction is calculated. Lastly, a new random point on this line is selected. This process is repeated until the set of points approaches a uniform sample of the solution space, as measured using the mixed fraction metric described previously [101]. A mixed fraction of approximately 0.50 was obtained, suggesting that the space of all possible flux distributions is nearly uniformly sampled.

For each reaction, a distribution of feasible steady-state flux values is acquired from the uniformly sampled points, subject to the network topology and model constraints. For the *E. coli* model such distributions of feasible flux values could be determined for 2,314 of the 2,382 reactions. The remaining 68 reactions were involved in loops [100] and therefore reliable flux estimates were not available. Thus, sampling distributions for these 68 reactions were removed from all analysis in this work. Similar measures were taken for all other models in this work.

GIMME: This method integrates genome-scale gene expression data to find the subset of metabolic reactions that are likely active in the given strain based on gene expression levels and the set of reactions that are needed to satisfy known network functions. Briefly, this analysis yields i) a network containing all metabolic reactions that are best supported by the expression data, ii) a report of the penalties assigned for having to add low expression genes, and iii) a normalized consistency score (NCS) that represents how well dataset supports the optimal growth predictions. These were obtained by setting the substrate and oxygen uptake rates to the experimentally measured values and then setting a lower limit on growth rate in the model (90% of the optimal growth rate as predicted by FBA). Gene expression data was then used by GIMME to build a model using all genes with an expression level above a given threshold. Any genes below this given threshold that were required to meet the 90% growth rate cutoff were subsequently assigned a penalty score. This penalty score is the product of the associated reaction flux and the difference between the threshold and the gene expression level. These penalty scores were summed up and then transformed into a NCS that quantitatively describes how well the data support a functional model of growth under the given condition [110]. The sensitivity of these NCS was determined using Jackknife re-sampling by removing each gene and then re-computing NCS scores. For the GIMME analysis in this study, a threshold is chosen as a cutoff $\log_2$ expression level on the Affymetrix data. While a threshold of 10 was chosen for this work, all NCSs from seven to 15 were tested and yielded similar results to those reported here (data not shown). Reactions that demonstrate improved consistency following adaptive evolution were determined by modeling L-1,2-PDO or glycerol growth using data from growth on glycerol because the parental GC strain cannot grow on L-1,2-PDO. The penalty scores from each reaction, before and after adaptive evolution, were subsequently compared. The hypergeometric test was used to find metabolic subsystems from iAF1260 that were enriched within the top 2.5% most significantly improved gene-associated metabolic reactions. To maintain consistency between conditions, substrate uptake rates were set such that the net carbon atom uptake was 30 mmol C gDW^{-1} hr^{-1} (e.g., for glycerol, L-1,2-pdo, and L-lactate had an uptake rate of 10 mmol C gDW^{-1} hr^{-1}, and L-fucose has an uptake rate of 5 mmol C gDW^{-1} hr^{-1}).

Assessment of regulon activity: Consistency of differential expression with the activity of each transcription regulator was determined by comparing differential expression changes with activator/repressor assignments in RegulonDB v6.0 [268]. That is, all genes known to be activated or repressed by a transcription factor were compared to up- and down-regulation assignments from the microarray data.

Predicting differential expression of metabolic genes: Genes that are specifically needed in higher abundance for growth on L-1,2-pdo (as opposed to glycerol) were predicted by simulating changes in reaction flux occurring in a shift between the two conditions. While this has been described in detail previously [78, 86], it was done here as follows. The distributions of MCMC-sampled fluxes for each reaction were compared between the two media conditions. First, flux magnitudes were normalized between each pair of media conditions (media *A* and *B*). To do this, a ratio of total flux through the metabolic network was computed and used to normalize each sample point. To compute this ratio, for each sample point, the magnitudes of all *n* non-loop-associated reaction fluxes were summed to acquire a value for the total network flux. For both media conditions, the median total network flux was taken and used to normalize each reaction flux for all sample points in media B, as follows:

$$v^*_{i,j,B} = v_{i,j,B}\frac{median(\{\sum_{r=1}^{n}|v_{r,1,A}|,\ldots,\sum_{r=1}^{n}|v_{r,j,A}|,\ldots,\sum_{r=1}^{n}|v_{r,p,A}|\})}{median(\{\sum_{r=1}^{n}|v_{r,1,B}|,\ldots,\sum_{r=1}^{n}|v_{r,j,B}|,\ldots,\sum_{r=1}^{n}|v_{r,p,B}|\})},$$

where $v^*_{i,j,B}$, is the normalized flux through reaction *i* in sample point *j* under media condition *B*, obtained after multiplying the sampled flux $v_{i,j,B}$, by the ratio of the median total flux magnitude for all *p* sample points under growth on medium *A* to the median total flux magnitude for all *p* sample points under growth on medium *B*.

Once the flux values were normalized, the changes of fluxes between two conditions were determined as previously described [86]. Calls on differential reaction activity were made when the distributions of feasible flux states (obtained from MCMC sampling) under two different conditions do not significantly overlap. For each metabolic reaction, a p-value was obtained by computing the probability of finding a flux value for a reaction in one condition that is equal to or more extreme than a given flux value in the second condition. P-values were adjusted for multiple hypotheses (FDR ≤ 0.01), and genes associated with reactions for which flux significantly increased or decreased were predicted to be up- or down-regulated, respectively.

Intracellular cAMP assay: The intracellular concentration of cAMP was determined as previously described [301]. Briefly, cells were cultured as described above and collected during logarithmic growth when reached OD_{600} of 0.2 - 0.3. Culture samples were rapidly vacuum-filtered through a triton-free nitrocellulose filter (Millipore, 25 mm diameter; pore size, 0.45 μm). The volume of culture filtered was based on the optical density, filtering the volume that would contain approximately the same number of cells as 1 mL of culture with an OD_{600} of 1 (i.e., 5 mL collected from a culture at OD_{600} 0.2). Filtration was immediately followed by washing with 10 mL of fresh media at 30°C to wash away extracellular cAMP. The cells collected on the filter were then immediately quenched by submersion in 5 mL ice-cold 65% ethanol, vortex-mixed vigorously, and stored at -20°C. Prior to assay, the 65% ethanol was evaporated by a speedvac and the dried residue was re-dissolved in cAMP assay buffer supplied with a kit. The cAMP levels were assayed using the enzyme-linked immunoassay kit (GE Healthcare) following the manufacturer's instructions for the non-acetylation protocol. One fifth of each sample was used per assay. Each strain was cultured in triplicate and each sample was assayed twice.

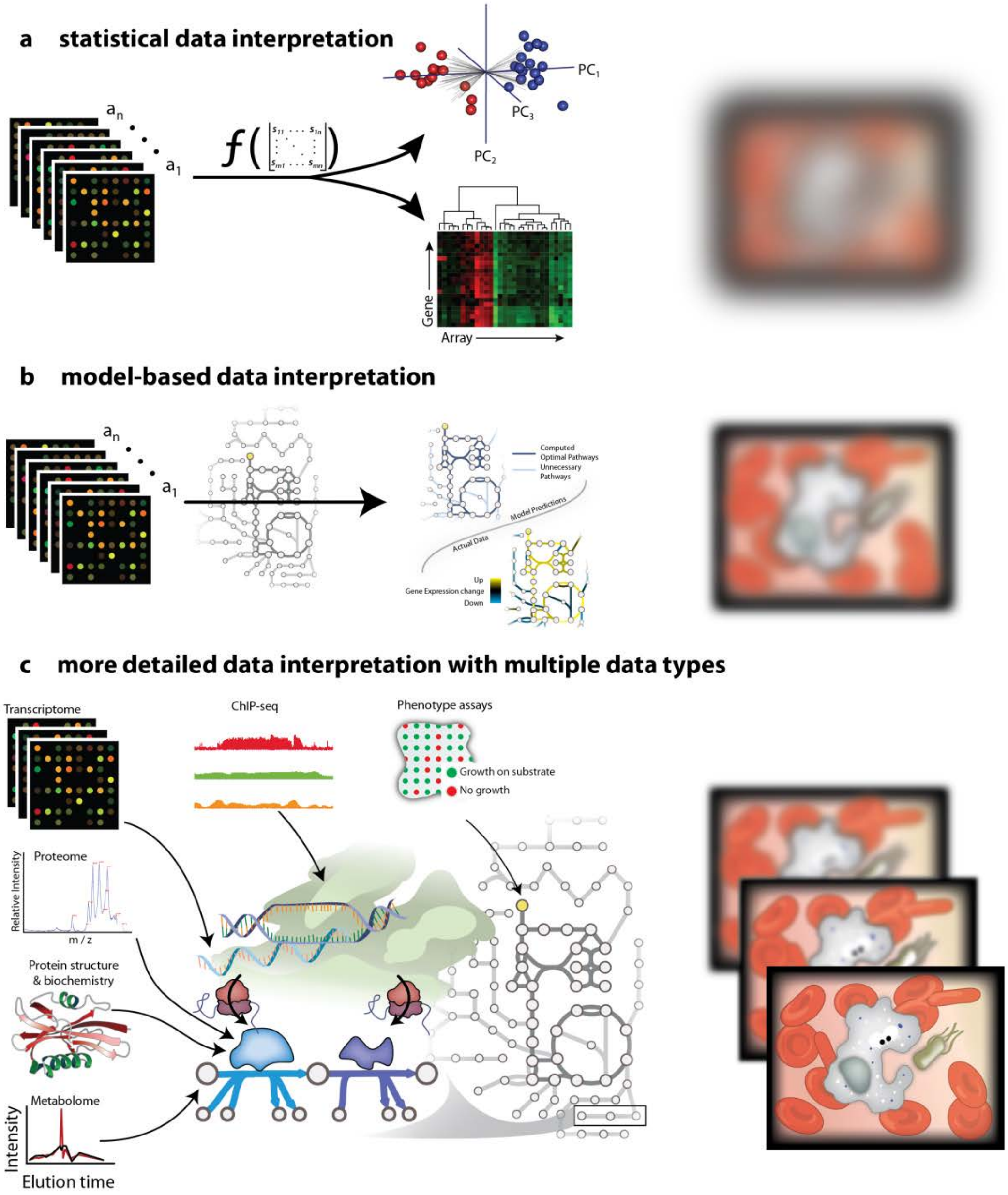
Chapter 6, in part, is a reprint of the material as it appears in Lewis, N.E., Lee, D.H., Rutledge, A., Conrad, T.M., Kim, D., Adkins, J.A., Smith, R.D., Palsson, B.Ø. *E. coli* learns to grow optimally on a non-native carbon substrate through laboratory evolution. *Under revision*. I was a joint-primary author, while the co-authors provided support in the research that served as the basis for this study.

Conclusion: Towards a predictive understanding of phenotype

Biological systems are subject to both proximal and distal causation, and to some degree, the mechanisms underlying both distal and proximal causation are being accounted for by *in silico* constraint-based models. Thus, these models can be used for simulation and analysis to understand biomolecular mechanisms responsible for proximal responses and some selective pressures guiding the evolution of metabolism.

Essentially, the study of causation in biology requires a detailed characterization of the molecular and chemical basis of biology. This is because the physicochemical properties of cell components (e.g., gene products and small molecules in a cell and its microenvironment) impose constraints on all possible cell phenotypes. Moreover, the cell genotype defines the repertoire of gene products that can be produced, which limits the phenotypes an organism can express. Ultimately, causation stems from an organism's genotype and the environment in which an organism is found. Thus, in order to build predictive models that can assess causal mechanisms, one must obtain the following detailed knowledge: 1) all of the relevant **components** in the cell and its microenvironment (e.g., genes, transcripts, proteins, metabolites, etc.), 2) the detailed **interactions** between these components and how these interactions can be described mathematically, and 3) **physical laws** constraining the function of these components (e.g., mass balance, thermodynamics, kinetics, etc.). For metabolism, decades of research has identified many cell components, and carefully studied the biochemical interactions between enzymes and metabolites. In like manner, the molecular biology of many components has been intensely studied, elucidating interactions between proteins, RNA, DNA, etc. In addition, over the centuries, deep theoretical understanding has been obtained with respect to physical laws. Moreover, the details of how these laws constrain the functions cellular components are being studied on an ongoing basis. The challenge remains as to how this knowledge can be expanded and leveraged to provide mechanistic insights into how organisms function.

Over the past couple decades, emerging technologies have enabled the systematic discovery of cellular components and some of their interactions. These include genome sequencing, transcriptomics, proteomics, metabolomics, chromatin immunoprecipitation, yeast 2 hybrid assays, etc. Advances in bioinformatics have enabled the interpretation of these data. Much focus has been given to statistical analyses, which have yielded great insights and guided numerous detailed mechanistic studies. However, on their own, these methods often only provide predictions of clusters of components that might contribute to biological processes (Figure 7.1.a). This is because these methods often only identify correlations between components as experimental conditions change. Thus, these analyses do not directly provide a clear picture of the biological system (Figure 7.1.a, right). If detailed mechanistic information is introduced (e.g., from metabolic or regulatory networks), this context can provide higher-resolution insights into how variations in cell components lead to a cell phenotype (Figure 7.1.b). However, detailed follow-up experiments are often needed to address causation. Even higher resolution insights can be attained as additional data sources are integrated into such analyses (Figure 7.1.c). Furthermore, modeling frameworks, such as constraint-based modeling, provide the mathematical language in which all of these data can be integrated with existing biochemical knowledge and employed to simulate phenotypes and provide mechanistic insights into causation.



a statistical data interpretation

b model-based data interpretation

c more detailed data interpretation with multiple data types

Figure 7.1. Increasing the resolution of knowledge with multi-omic data integration. (a) Exploratory analysis of large –omic datasets has successfully provided insight into large biological systems and guided more detailed studies. However, on their own, these approaches cannot give a detailed picture of the mechanisms underlying biological functions since they rely on correlations between biomolecular components. (b) Additional insights into the roles of variations in biological datasets can be obtained as these data are analyzed in the context of detailed biochemical networks. Such studies can provide higher-resolution and more mechanistic insights into biological functions. (c) Since the analysis of a single data type only captures one angle of the cellular biochemistry, more detailed and accurate mechanisms can be identified as multiple types of data are integrated and analyzed in the context of solid biochemical knowledge. Such analyses can potentially further clarify mechanisms, thereby providing a higher-resolution picture of real biological functions.

In this work I have focused on how detailed mechanistic predictions can be obtained by analyzing multiple data types in the context of detailed metabolic and regulatory networks. Each study required advances in model-based data integration, analysis, and interpretation. The ultimate goal of each of these studies was to gain insight into the mechanisms connecting genotype to phenotype (i.e., proximal causation), and the selective pressures causing the evolution of genotype (i.e., distal causation). Eventually, we hope to understand these complex molecular interactions and selective pressures. As we understand the basis for proximal causation in biological systems, we will be more able to predict phenotypes and individual-specific responses to perturbations. As we understand distal causation, we will be able to avoid problematic adaptation (e.g., drug resistance) [302] and more efficiently engineer biology [294]. While this extent of biological knowledge currently is beyond our reach for many cell processes, continued integrated efforts in 1) component identification, 2) interaction description, and 3) physicochemical constraint elucidation will together help us link genotypes with complex phenotypes, and thereby increase the resolution in our understanding of why biological systems do what they do.

References

1. Mayr, E. Cause and effect in biology. *Science* **134**, 1501-1506 (1961).

2. Feist, A. M., Herrgard, M. J., Thiele, I., Reed, J. L. & Palsson, B. O. Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.* **7**, 129-143 (2009).

3. Thiele, I. & Palsson, B. O. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protocols* **5**, 93-121 (2010).

4. Henry, C. S. *et al*. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* **28**, 977-982 (2010).

5. Feist, A. M. & Palsson, B. O. The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nat. Biotechnol.* **26**, 659-667 (2008).

6. Oberhardt, M. A., Palsson, B. O. & Papin, J. A. Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.* **5**, 320 (2009).

7. Papp, B., Notebaart, R. A. & Pal, C. Systems-biology approaches for predicting genomic evolution. *Nat. Rev. Genet.* **12**, 591-602 (2011).

8. Mahadevan, R., Palsson, B. O. & Lovley, D. R. In situ to in silico and back: elucidating the physiology and ecology of Geobacter spp. using genome-scale modelling. *Nat. Rev. Microbiol.* **9**, 39-50 (2011).

9. Palsson, B. in *Systems biology : properties of reconstructed networks* 322 (Cambridge University Press, Cambridge ; New York, 2006).

10. Feist, A. M. & Palsson, B. O. The biomass objective function. *Curr. Opin. Microbiol.* **13**, 344-349 (2010).

11. Fell, D. A. & Small, J. R. Fat synthesis in adipose tissue. An examination of stoichiometric constraints. *Biochem. J.* **238**, 781-786 (1986).

12. Watson, M. R. Metabolic maps for the Apple II. *Biochemical Society Transactions* **12**, 1093 (1984).

13. Savinell, J. M. & Palsson, B. O. Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism. *J. Theor. Biol.* **154**, 421-454 (1992).

14. Varma, A. & Palsson, B. O. Metabolic capabilities of Escherichia coli: I. synthesis of biosynthetic precursors and cofactors. *J. Theor. Biol.* **165**, 477-502 (1993).

15. Edwards, J. S. & Palsson, B. O. Systems properties of the Haemophilus influenzae Rd metabolic genotype. *J. Biol. Chem.* **274**, 17410-17416 (1999).

16. Edwards, J. S. & Palsson, B. O. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 5528-5533 (2000).

17. Reed, J. L., Famili, I., Thiele, I. & Palsson, B. O. Towards multidimensional genome annotation. *Nat. Rev. Genet.* **7**, 130-141 (2006).

18. Kim, T. Y., Kim, H. U. & Lee, S. Y. Data integration and analysis of biological networks. *Curr. Opin. Biotechnol.* **21**, 78-84 (2010).

19. Sauer, U. Metabolic networks in motion: 13C-based flux analysis. *Mol. Syst. Biol.* **2**, 62 (2006).

20. Papin, J. A. *et al*. Comparison of network-based pathway analysis methods. *Trends Biotechnol.* **22**, 400-405 (2004).

21. Trinh, C. T., Wlaschin, A. & Srienc, F. Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. *Appl. Microbiol. Biotechnol.* **81**, 813-826 (2009).

22. Llaneras, F. & Pico, J. Which metabolic pathways generate and characterize the flux space? A comparison among elementary modes, extreme pathways and minimal generators. *J. Biomed. Biotechnol.* **2010**, 753904 (2010).

23. Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S. & Gilles, E. D. Metabolic network structure determines key aspects of functionality and regulation. *Nature* **420**, 190-193 (2002).

24. Trinh, C. T., Unrean, P. & Srienc, F. Minimal Escherichia coli cell for the most efficient production of ethanol from hexoses and pentoses. *Appl. Environ. Microbiol.* **74**, 3634-3643 (2008).

25. Imielinski, M. & Belta, C. Exploiting the pathway structure of metabolism to reveal high-order epistasis. *BMC Syst. Biol.* **2**, 40 (2008).

26. Wessely, F. *et al*. Optimal regulatory strategies for metabolic pathways in Escherichia coli depending on protein costs. *Mol. Syst. Biol.* **7**, 515 (2011).

27. Schilling, C. H. & Palsson, B. O. Assessment of the metabolic capabilities of Haemophilus influenzae Rd through a genome-scale pathway analysis. *J. Theor. Biol.* **203**, 249-283 (2000).

28. Yeung, M., Thiele, I. & Palsson, B. O. Estimation of the number of extreme pathways for metabolic networks. *BMC Bioinformatics* **8**, 363 (2007).

29. Klamt, S. & Stelling, J. Combinatorial complexity of pathway analysis in metabolic networks. *Mol. Biol. Rep.* **29**, 233-236 (2002).

30. Kaleta, C., de Figueiredo, L. F. & Schuster, S. Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns. *Genome Res.* **19**, 1872-1883 (2009).

31. Rezola, A. *et al*. Exploring metabolic pathways in genome-scale networks via generating flux modes. *Bioinformatics* **27**, 534-540 (2011).

32. Ip, K., Colijn, C. & Lun, D. S. Analysis of complex metabolic behavior through pathway decomposition. *BMC Syst. Biol.* **5**, 91 (2011).

33. Chan, S. H. & Ji, P. Decomposing flux distributions into elementary flux modes in genome-scale metabolic networks. *Bioinformatics* **27**, 2256-2262 (2011).

34. Braunstein, A., Mulet, R. & Pagnani, A. Estimating the size of the solution space of metabolic networks. *BMC Bioinformatics* **9**, 240 (2008).

35. Schellenberger, J. & Palsson, B. O. Use of randomized sampling for analysis of metabolic networks. *J. Biol. Chem.* **284**, 5457-5461 (2009).

36. Almaas, E., Kovacs, B., Vicsek, T., Oltvai, Z. N. & Barabasi, A. L. Global organization of metabolic fluxes in the bacterium Escherichia coli. *Nature* **427**, 839-843 (2004).

37. Bordel, S., Agren, R. & Nielsen, J. Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes. *PLoS Comput. Biol.* **6**, e1000859 (2010).

38. Mo, M. L., Palsson, B. O. & Herrgard, M. J. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst. Biol.* **3**, 37 (2009).

39. Barrett, C. L., Herrgard, M. J. & Palsson, B. Decomposing complex reaction networks using random sampling, principal component analysis and basis rotation. *BMC Syst. Biol.* **3**, 30 (2009).

40. Thiele, I., Price, N. D., Vo, T. D. & Palsson, B. O. Candidate metabolic network states in human mitochondria. Impact of diabetes, ischemia, and diet. *J. Biol. Chem.* **280**, 11683-11695 (2005).

41. Price, N. D., Schellenberger, J. & Palsson, B. O. Uniform sampling of steady-state flux spaces: means to design experiments and to interpret enzymopathies. *Biophys. J.* **87**, 2172-2186 (2004).

42. Lewis, N. E. *et al*. Large-scale in silico modeling of metabolic interactions between cell types in the human brain. *Nat. Biotechnol.* **28**, 1279-1285 (2010).

43. Orth, J. D., Thiele, I. & Palsson, B. O. What is flux balance analysis? *Nat. Biotechnol.* **28**, 245-248 (2010).

44. Beg, Q. K. *et al*. Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 12663-12668 (2007).

45. Varma, A. & Palsson, B. O. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Appl. Environ. Microbiol.* **60**, 3724-3731 (1994).

46. Edwards, J. S., Ibarra, R. U. & Palsson, B. O. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.* **19**, 125-130 (2001).

47. Lewis, N. E. *et al*. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.* **6**, 390 (2010).

48. Teusink, B., Wiersma, A., Jacobs, L., Notebaart, R. A. & Smid, E. J. Understanding the adaptive growth strategy of Lactobacillus plantarum by in silico optimisation. *PLoS Comput. Biol.* **5**, e1000410 (2009).

49. Feist, A. M., Scholten, J. C., Palsson, B. O., Brockman, F. J. & Ideker, T. Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri. *Mol. Syst. Biol.* **2**, 2006.0004 (2006).

50. Wang, Z. & Zhang, J. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *Proc. Natl. Acad. Sci. U. S. A.* **108**, E67-76 (2011).

51. Goffin, P. *et al*. Understanding the physiology of Lactobacillus plantarum at zero growth. *Mol. Syst. Biol.* **6**, 413 (2010).

52. Lee, S., Palakornkule, C., Domach, M. M. & Grossmann, I. E. Recursive MILP model for finding all the alternate optima in LP models for metabolic networks. *Computers & Chemical Engineering,* **24**, 711-716 (2000).

53. Gudmundsson, S. & Thiele, I. Computationally efficient flux variability analysis. *BMC Bioinformatics* **11**, 489 (2010).

54. Burgard, A. P., Vaidyaraman, S. & Maranas, C. D. Minimal reaction sets for Escherichia coli metabolism under different growth requirements and uptake environments. *Biotechnol. Prog.* **17**, 791-797 (2001).

55. Segre, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 15112-15117 (2002).

56. Park, J. M., Kim, T. Y. & Lee, S. Y. Prediction of metabolic fluxes by incorporating genomic context and flux-converging pattern analyses. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 14931-14936 (2010).

57. Holzhutter, H. G. The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *Eur. J. Biochem.* **271**, 2905-2922 (2004).

58. Ponce de Leon, M., Cancela, H. & Acerenza, L. A strategy to calculate the patterns of nutrient consumption by microorganisms applying a two-level optimisation principle to reconstructed metabolic networks. *J. Biol. Phys.* **34**, 73-90 (2008).

59. Murabito, E., Simeonidis, E., Smallbone, K. & Swinton, J. Capturing the essence of a metabolic network: a flux balance analysis approach. *J. Theor. Biol.* **260**, 445-452 (2009).

60. Benyamini, T., Folger, O., Ruppin, E. & Shlomi, T. Flux balance analysis accounting for metabolite dilution. *Genome Biol.* **11**, R43 (2010).

61. Colijn, C. *et al*. Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. *PLoS Comput. Biol.* **5**, e1000489 (2009).

62. van Berlo, R. J. P. *et al*. Predicting Metabolic Fluxes Using Gene Expression Differences As Constraints. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* **8**, 206-216 (2011).

63. Zhuang, K., Vemuri, G. N. & Mahadevan, R. Economics of membrane occupancy and respiro-fermentation. *Mol. Syst. Biol.* **7**, 500 (2011).

64. Shlomi, T., Berkman, O. & Ruppin, E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 7695-7700 (2005).

65. Kim, P. J. *et al*. Metabolite essentiality elucidates robustness of Escherichia coli metabolism. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 13638-13642 (2007).

66. Chang, R. L., Xie, L., Xie, L., Bourne, P. E. & Palsson, B. O. Drug off-target effects predicted using structural analysis in the context of a metabolic network model. *PLoS Comput. Biol.* **6**, e1000938 (2010).

67. Shen, Y. *et al*. Blueprint for antimicrobial hit discovery targeting metabolic networks. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 1082-1087 (2010).

68. Kim, H. U. *et al*. Integrative genome-scale metabolic analysis of Vibrio vulnificus for drug targeting and discovery. *Mol. Syst. Biol.* **7**, 460 (2011).

69. Dobson, P. D., Patel, Y. & Kell, D. B. 'Metabolite-likeness' as a criterion in the design and selection of pharmaceutical drug libraries. *Drug Discov. Today* **14**, 31-40 (2009).

70. Fong, S. S. *et al*. In silico design and adaptive evolution of Escherichia coli for production of lactic acid. *Biotechnol. Bioeng.* **91**, 643-648 (2005).

71. Burgard, A. P., Pharkya, P. & Maranas, C. D. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* **84**, 647-657 (2003).

72. Feist, A. M. *et al*. Model-driven evaluation of the production potential for growth-coupled products of Escherichia coli. *Metab. Eng.* (2009).

73. Tepper, N. & Shlomi, T. Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways. *Bioinformatics* **26**, 536-543 (2010).

74. Patil, K. R., Rocha, I., Forster, J. & Nielsen, J. Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics* **6**, 308 (2005).

75. Lun, D. S. *et al*. Large-scale identification of genetic design strategies using local search. *Mol. Syst. Biol.* **5**, 296 (2009).

76. Yousofshahi, M., Lee, K. & Hassoun, S. Probabilistic pathway construction. *Metab. Eng.* **13**, 435-444 (2011).

77. Rodrigo, G., Carrera, J., Prather, K. J. & Jaramillo, A. DESHARKY: automatic design of metabolic pathways for optimal cell growth. *Bioinformatics* **24**, 2554-2556 (2008).

78. Bar-Even, A., Noor, E., Lewis, N. E. & Milo, R. Design and analysis of synthetic carbon fixation pathways. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 8889-8894 (2010).

79. Dueber, J. E. *et al*. Synthetic protein scaffolds provide modular control over metabolic flux. *Nat. Biotechnol.* **27**, 753-759 (2009).

80. Moon, T. S., Dueber, J. E., Shiue, E. & Prather, K. L. Use of modular, synthetic scaffolds for improved production of glucaric acid in engineered E. coli. *Metab. Eng.* **12**, 298-305 (2010).

81. Delebecque, C. J., Lindner, A. B., Silver, P. A. & Aldaye, F. A. Organization of intracellular reactions with rationally designed RNA assemblies. *Science* **333**, 470-474 (2011).

82. Yim, H. *et al*. Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. *Nat. Chem. Biol.* **7**, 445-452 (2011).

83. Pharkya, P., Burgard, A. P. & Maranas, C. D. OptStrain: a computational framework for redesign of microbial production systems. *Genome Res.* **14**, 2367-2376 (2004).

84. Szappanos, B. *et al*. An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nat. Genet.* **43**, 656-662 (2011).

85. Hsiao, T. L., Revelles, O., Chen, L., Sauer, U. & Vitkup, D. Automatic policing of biochemical annotations using genomic correlations. *Nat. Chem. Biol.* **6**, 34-40 (2010).

86. Bordbar, A., Lewis, N. E., Schellenberger, J., Palsson, B. O. & Jamshidi, N. Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions. *Mol. Syst. Biol.* **6**, 422 (2010).

87. Schuster, S., Pfeiffer, T. & Fell, D. A. Is maximization of molar yield in metabolic networks favoured by evolution? *J. Theor. Biol.* **252**, 497-504 (2008).

88. Milne, C. B. *et al*. Metabolic network reconstruction and genome-scale model of butanol-producing strain Clostridium beijerinckii NCIMB 8052. *BMC Syst. Biol.* **5**, 130 (2011).

89. Reed, J. L. *et al*. Systems approach to refining genome annotation. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 17480-17484 (2006).

90. Orth, J. D. & Palsson, B. O. Systematizing the generation of missing metabolic knowledge. *Biotechnol. Bioeng.* **107**, 403-412 (2010).

91. Satish Kumar, V., Dasika, M. S. & Maranas, C. D. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics* **8**, 212 (2007).

92. Kumar, V. S. & Maranas, C. D. GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. *PLoS Comput. Biol.* **5**, e1000308 (2009).

93. Suthers, P. F., Zomorrodi, A. & Maranas, C. D. Genome-scale gene/reaction essentiality and synthetic lethality analysis. *Mol. Syst. Biol.* **5**, 301 (2009).

94. Mintz-Oron, S., Aharoni, A., Ruppin, E. & Shlomi, T. Network-based prediction of metabolic enzymes' subcellular localization. *Bioinformatics* **25**, i247-52 (2009).

95. Burgard, A. P. & Maranas, C. D. Optimization-based framework for inferring and testing hypothesized metabolic objective functions. *Biotechnol. Bioeng.* **82**, 670-677 (2003).

96. Knorr, A. L., Jain, R. & Srivastava, R. Bayesian-based selection of metabolic objective functions. *Bioinformatics* **23**, 351-357 (2007).

97. Schuetz, R., Kuepfer, L. & Sauer, U. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Mol. Syst. Biol.* **3**, 119 (2007).

98. Gianchandani, E. P., Oberhardt, M. A., Burgard, A. P., Maranas, C. D. & Papin, J. A. Predicting biological system objectives de novo from internal state measurements. *BMC Bioinformatics* **9**, 43 (2008).

99. Zomorrodi, A. R. & Maranas, C. D. Improving the iMM904 S. cerevisiae metabolic model using essentiality and synthetic lethality data. *BMC Syst. Biol.* **4**, 178 (2010).

100. Beard, D. A., Liang, S. D. & Qian, H. Energy balance for analysis of complex metabolic networks. *Biophys. J.* **83**, 79-86 (2002).

101. Schellenberger, J., Lewis, N. E. & Palsson, B. O. Elimination of thermodynamically infeasible loops in steady-state metabolic models. *Biophys. J.* **100**, 544-553 (2011).

102. Price, N. D., Thiele, I. & Palsson, B. O. Candidate states of Helicobacter pylori's genome-scale metabolic network upon application of "loop law" thermodynamic constraints. *Biophys. J.* **90**, 3919-3928 (2006).

103. Henry, C. S., Broadbelt, L. J. & Hatzimanikatis, V. Thermodynamics-based metabolic flux analysis. *Biophys. J.* **92**, 1792-1805 (2007).

104. Fleming, R. M., Thiele, I., Provan, G. & Nasheuer, H. P. Integrated stoichiometric, thermodynamic and kinetic modelling of steady state metabolism. *J. Theor. Biol.* **264**, 683-692 (2010).

105. Kummel, A., Panke, S. & Heinemann, M. Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Mol. Syst. Biol.* **2**, 2006.0034 (2006).

106. Jankowski, M. D., Henry, C. S., Broadbelt, L. J. & Hatzimanikatis, V. Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys. J.* **95**, 1487-1499 (2008).

107. Henry, C. S., Jankowski, M. D., Broadbelt, L. J. & Hatzimanikatis, V. Genome-scale thermodynamic analysis of Escherichia coli metabolism. *Biophys. J.* **90**, 1453-1461 (2006).

108. Hoppe, A., Hoffmann, S. & Holzhutter, H. G. Including metabolite concentrations into flux balance analysis: thermodynamic realizability as a constraint on flux distributions in metabolic networks. *BMC Syst. Biol.* **1**, 23 (2007).

109. Fleming, R. M., Thiele, I. & Nasheuer, H. P. Quantitative assignment of reaction directionality in constraint-based models of metabolism: application to Escherichia coli. *Biophys. Chem.* **145**, 47-56 (2009).

110. Becker, S. A. & Palsson, B. O. Context-specific metabolic networks are consistent with experiments. *PLoS Comput. Biol.* **4**, e1000082 (2008).

111. Shlomi, T., Cabili, M. N., Herrgard, M. J., Palsson, B. O. & Ruppin, E. Network-based prediction of human tissue-specific metabolism. *Nat. Biotechnol.* **26**, 1003-1010 (2008).

112. Jerby, L., Shlomi, T. & Ruppin, E. Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Mol. Syst. Biol.* **6**, 401 (2010).

113. Jensen, P. A. & Papin, J. A. Functional integration of a metabolic network model and expression data without arbitrary thresholding. *Bioinformatics* **27**, 541-547 (2011).

114. Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J. & Palsson, B. O. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**, 92-96 (2004).

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

116. Covert, M. W., Xiao, N., Chen, T. J. & Karr, J. R. Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli. *Bioinformatics* **24**, 2044-2050 (2008).

117. Lee, J. M., Gianchandani, E. P., Eddy, J. A. & Papin, J. A. Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLoS Comput. Biol.* **4**, e1000086 (2008).

118. Folger, O. *et al*. Predicting selective drug targets in cancer through metabolic networks. *Mol. Syst. Biol.* **7**, 501 (2011).

119. Frezza, C. *et al*. Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase. *Nature* (2011).

120. Herrgard, M. J., Lee, B. S., Portnoy, V. & Palsson, B. O. Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae. *Genome Res.* **16**, 627-635 (2006).

121. Covert, M. W., Schilling, C. H. & Palsson, B. Regulation of gene expression in flux balance models of metabolism. *J. Theor. Biol.* **213**, 73-88 (2001).

122. Chandrasekaran, S. & Price, N. D. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 17845-17850 (2010).

123. Tyson, G. W. *et al*. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37-43 (2004).

124. Pal, C. *et al*. Chance and necessity in the evolution of minimal metabolic networks. *Nature* **440**, 667-670 (2006).

125. Stolyar, S. *et al*. Metabolic modeling of a mutualistic microbial community. *Mol. Syst. Biol.* **3**, 92 (2007).

126. Taffs, R. *et al*. In silico approaches to study mass and energy flows in microbial consortia: a syntrophic case study. *BMC Syst. Biol.* **3**, 114 (2009).

127. Klitgord, N. & Segre, D. Environments that induce synthetic microbial ecosystems. *PLoS Comput. Biol.* **6**, e1001002 (2010).

128. Wintermute, E. H. & Silver, P. A. Emergent cooperation in microbial metabolism. *Mol. Syst. Biol.* **6**, 407 (2010).

129. Zhuang, K. *et al*. Genome-scale dynamic modeling of the competition between Rhodoferax and Geobacter in anoxic subsurface environments. *ISME J.* **5**, 305-316 (2011).

130. Huthmacher, C., Hoppe, A., Bulik, S. & Holzhutter, H. G. Antimalarial drug targets in Plasmodium falciparum predicted by stage-specific metabolic network analysis. *BMC Syst. Biol.* **4**, 120 (2010).

131. Yizhak, K., Tuller, T., Papp, B. & Ruppin, E. Metabolic modeling of endosymbiont genome reduction on a temporal scale. *Mol. Syst. Biol.* **7**, 479 (2011).

132. Bonde, B. K., Beste, D. J., Laing, E., Kierzek, A. M. & McFadden, J. Differential Producibility Analysis (DPA) of Transcriptomic Data with Metabolic Networks: Deconstructing the Metabolic Response of M. tuberculosis. *PLoS Comput. Biol.* **7**, e1002060 (2011).

133. Breitling, R., Vitkup, D. & Barrett, M. P. New surveyor tools for charting microbial metabolic maps. *Nat. Rev. Microbiol.* **6**, 156-161 (2008).

134. Lewis, N. E., Jamshidi, N., Thiele, I. & Palsson, B. Ø. in *Encyclopedia of Complexity and Systems Science* 5535 (Springer, New York, 2009).

135. Duarte, N. C. *et al*. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 1777-1782 (2007).

136. Ponten, F. *et al*. A global view of protein expression in human cells, tissues, and organs. *Mol. Syst. Biol.* **5**, 337 (2009).

137. Chatziioannou, A., Palaiologos, G. & Kolisis, F. N. Metabolic flux analysis as a tool for the elucidation of the metabolism of neurotransmitter glutamate. *Metab. Eng.* **5**, 201-210 (2003).

138. Cakir, T., Alsan, S., Saybasili, H., Akin, A. & Ulgen, K. O. Reconstruction and flux analysis of coupling between metabolic pathways of astrocytes and neurons: application to cerebral hypoxia. *Theor. Biol. Med. Model.* **4**, 48 (2007).

139. Occhipinti, R., Puchowicz, M. A., LaManna, J. C., Somersalo, E. & Calvetti, D. Statistical analysis of metabolic pathways of brain metabolism at steady state. *Ann. Biomed. Eng.* **35**, 886-902 (2007).

140. Reiman, E. M. *et al*. Functional brain abnormalities in young adults at genetic risk for late-onset Alzheimer's dementia. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 284-289 (2004).

141. Fukui, H., Diaz, F., Garcia, S. & Moraes, C. T. Cytochrome c oxidase deficiency in neurons decreases both oxidative stress and amyloid formation in a mouse model of Alzheimer's disease. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 14163-14168 (2007).

142. Bubber, P., Haroutunian, V., Fisch, G., Blass, J. P. & Gibson, G. E. Mitochondrial abnormalities in Alzheimer brain: mechanistic implications. *Ann. Neurol.* **57**, 695-703 (2005).

143. Gibson, G. E. *et al*. Alpha-ketoglutarate dehydrogenase in Alzheimer brains bearing the APP670/671 mutation. *Ann. Neurol.* **44**, 676-681 (1998).

144. Casley, C. S., Canevari, L., Land, J. M., Clark, J. B. & Sharpe, M. A. Beta-amyloid inhibits integrated mitochondrial respiration and key enzyme activities. *J. Neurochem.* **80**, 91-100 (2002).

145. Hoshi, M. *et al*. Regulation of mitochondrial pyruvate dehydrogenase activity by tau protein kinase I/glycogen synthase kinase 3beta in brain. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 2719-2723 (1996).

146. Gorman, A. M., Ceccatelli, S. & Orrenius, S. Role of mitochondria in neuronal apoptosis. *Dev. Neurosci.* **22**, 348-358 (2000).

147. Ginsberg, S. D., Che, S., Counts, S. E. & Mufson, E. J. Single cell gene expression profiling in Alzheimer's disease. *NeuroRx* **3**, 302-318 (2006).

148. Lai, M. K. P., Ramirez, M. J., Tsang, S. W. Y. & Francis, P. T. in *Neurobiology of Alzheimer's Disease* (eds Dawbarn, D. & Allen, S. J.) 245-282 (Oxford, New York, 2007).

149. Santos, S. S. *et al*. Inhibitors of the alpha-ketoglutarate dehydrogenase complex alter [1-13C]glucose and [U-13C]glutamate metabolism in cerebellar granule neurons. *J. Neurosci. Res.* **83**, 450-458 (2006).

150. Hassel, B., Johannessen, C. U., Sonnewald, U. & Fonnum, F. Quantification of the GABA shunt and the importance of the GABA shunt versus the 2-oxoglutarate dehydrogenase pathway in GABAergic neurons. *J. Neurochem.* **71**, 1511-1518 (1998).

151. Liang, W. S. *et al*. Altered neuronal gene expression in brain regions differentially affected by Alzheimer's disease: a reference data set. *Physiol. Genomics* **33**, 240-256 (2008).

152. Stuhmer, T., Anderson, S. A., Ekker, M. & Rubenstein, J. L. Ectopic expression of the Dlx genes induces glutamic acid decarboxylase and Dlx expression. *Development* **129**, 245-252 (2002).

153. Ibanez, V. *et al*. Regional glucose metabolic abnormalities are not the result of atrophy in Alzheimer's disease. *Neurology* **50**, 1585-1593 (1998).

154. Schramm, G. *et al*. PathWave: discovering patterns of differentially regulated enzymes in metabolic pathways. *Bioinformatics* **26**, 1225-1231 (2010).

155. Ragozzino, M. E., Pal, S. N., Unick, K., Stefani, M. R. & Gold, P. E. Modulation of hippocampal acetylcholine release and spontaneous alternation scores by intrahippocampal glucose injections. *J. Neurosci.* **18**, 1595-1601 (1998).

156. Watson, G. S. & Craft, S. Modulation of memory by insulin and glucose: neuropsychological observations in Alzheimer's disease. *Eur. J. Pharmacol.* **490**, 97-113 (2004).

157. Cooper, J. R. Unsolved problems in the cholinergic nervous system. *J. Neurochem.* **63**, 395-399 (1994).

158. Karczmar, A. G. in *Exploring the vertebrate cholinergic nervous system* 686 (Springer, New York, N.Y., 2006).

159. Gibson, G. E., Jope, R. & Blass, J. P. Decreased synthesis of acetylcholine accompanying impaired oxidation of pyruvic acid in rat brain minces. *Biochem. J.* **148**, 17-23 (1975).

160. Abbott, N. J., Ronnback, L. & Hansson, E. Astrocyte-endothelial interactions at the blood-brain barrier. *Nat. Rev. Neurosci.* **7**, 41-53 (2006).

161. Thompson, M. D., Knee, K. & Golden, C. J. Olfaction in persons with Alzheimer's disease. *Neuropsychol. Rev.* **8**, 11-23 (1998).

162. Schryer, D. W., Peterson, P., Paalme, T. & Vendelin, M. Bidirectionality and compartmentation of metabolic fluxes are revealed in the dynamics of isotopomer networks. *Int. J. Mol. Sci.* **10**, 1697-1718 (2009).

163. Serres, S., Raffard, G., Franconi, J. M. & Merle, M. Close coupling between astrocytic and neuronal metabolisms to fulfill anaplerotic and energy needs in the rat brain. *J. Cereb. Blood Flow Metab.* **28**, 712-724 (2008).

164. Lee, D. S. *et al*. The implications of human metabolic network topology for disease comorbidity. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 9880-9885 (2008).

165. Mishra, G. R. *et al*. Human protein reference database--2006 update. *Nucleic Acids Res.* **34**, D411-4 (2006).

166. Fujii, Y., Imanishi, T. & Gojobori, T. H-Invitational Database: integrated database of human genes. *Tanpakushitsu Kakusan Koso* **49**, 1937-1943 (2004).

167. Reidegeld, K. A. *et al*. The power of cooperative investigation: summary and comparison of the HUPO Brain Proteome Project pilot study results. *Proteomics* **6**, 4997-5014 (2006).

168. Lying-Tunell, U., Lindblad, B. S., Malmlund, H. O. & Persson, B. Cerebral blood flow and metabolic rate of oxygen, glucose, lactate, pyruvate, ketone bodies and amino acids. *Acta Neurol. Scand.* **62**, 265-275 (1980).

169. Lying-Tunell, U., Lindblad, B. S., Malmlund, H. O. & Persson, B. Cerebral blood flow and metabolic rate of oxygen, glucose, lactate, pyruvate, ketone bodies and amino acids. *Acta Neurol. Scand.* **63**, 337-350 (1981).

170. Tischfield, M. A. *et al*. Human TUBB3 Mutations Perturb Microtubule Dynamics, Kinesin Interactions, and Axon Guidance. *Cell* **140**, 74-87 (2010).

171. Kim, K. K., Adelstein, R. S. & Kawamoto, S. Identification of neuronal nuclei (NeuN) as Fox-3, a new member of the Fox-1 gene family of splicing factors. *J. Biol. Chem.* **284**, 31052-31061 (2009).

172. De Camilli, P., Cameron, R. & Greengard, P. Synapsin I (protein I), a nerve terminal-specific phosphoprotein. I. Its general distribution in synapses of the central and peripheral nervous system

demonstrated by immunofluorescence in frozen and plastic sections. *J. Cell Biol.* **96**, 1337-1354 (1983).

173. Olave, I., Wang, W., Xue, Y., Kuo, A. & Crabtree, G. R. Identification of a polymorphic, neuron-specific chromatin remodeling complex. *Genes Dev.* **16**, 2509-2517 (2002).

174. Benjamini, Y. & Yekutieli, D. The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics* **29**, 1165-1188 (2001).

175. Zhang, J. *et al*. Lysine acetylation is a highly abundant and evolutionarily conserved modification in Escherichia coli. *Mol. Cell. Proteomics* **8**, 215-225 (2009).

176. Yu, B. J., Kim, J. A., Moon, J. H., Ryu, S. E. & Pan, J. G. The diversity of lysine-acetylated proteins in Escherichia coli. *J. Microbiol. Biotechnol.* **18**, 1529-1536 (2008).

177. Zhang, Z. *et al*. Identification of lysine succinylation as a new post-translational modification. *Nat. Chem. Biol.* **7**, 58-63 (2011).

178. Macek, B. *et al*. Phosphoproteome analysis of E. coli reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation. *Mol. Cell. Proteomics* **7**, 299-307 (2008).

179. Soufi, B. *et al*. The Ser/Thr/Tyr phosphoproteome of Lactococcus lactis IL1403 reveals multiply phosphorylated proteins. *Proteomics* **8**, 3486-3493 (2008).

180. Wang, Q. *et al*. Acetylation of metabolic enzymes coordinates carbon source utilization and metabolic flux. *Science* **327**, 1004-1007 (2010).

181. Ravichandran, A., Sugiyama, N., Tomita, M., Swarup, S. & Ishihama, Y. Ser/Thr/Tyr phosphoproteome analysis of pathogenic and non-pathogenic Pseudomonas species. *Proteomics* **9**, 2764-2775 (2009).

182. Aivaliotis, M. *et al*. Ser/Thr/Tyr protein phosphorylation in the archaeon Halobacterium salinarum--a representative of the third domain of life. *PLoS One* **4**, e4777 (2009).

183. Price, N. D., Reed, J. L. & Palsson, B. O. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* **2**, 886-897 (2004).

184. Zaslaver, A. *et al*. Just-in-time transcription program in metabolic pathways. *Nat. Genet.* **36**, 486-491 (2004).

185. Dekel, E. & Alon, U. Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436**, 588-592 (2005).

186. Holms, W. H. & Bennett, P. M. Regulation of isocitrate dehydrogenase activity in Escherichia coli on adaptation to acetate. *J. Gen. Microbiol.* **65**, 57-68 (1971).

187. Shinar, G., Rabinowitz, J. D. & Alon, U. Robustness in glyoxylate bypass regulation. *PLoS Comput. Biol.* **5**, e1000297 (2009).

188. Pearlman, S. M., Serber, Z. & Ferrell, J. E.,Jr. A mechanism for the evolution of phosphorylation sites. *Cell* **147**, 934-946 (2011).

189. Feist, A. M. *et al*. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* **3**, 121 (2007).

190. Lienhard, G. E. Non-functional phosphorylations? *Trends Biochem. Sci.* **33**, 351-352 (2008).

191. Hurley, J. H., Dean, A. M., Sohl, J. L., Koshland, D. E.,Jr & Stroud, R. M. Regulation of an enzyme by phosphorylation at the active site. *Science* **249**, 1012-1016 (1990).

192. Vivoli, M. *et al*. Role of a conserved active site cation-pi interaction in Escherichia coli serine hydroxymethyltransferase. *Biochemistry* **48**, 12034-12046 (2009).

193. Cai, K., Schirch, D. & Schirch, V. The affinity of pyridoxal 5'-phosphate for folding intermediates of Escherichia coli serine hydroxymethyltransferase. *J. Biol. Chem.* **270**, 19294-19299 (1995).

194. Spring, T. G. & Wold, F. The purification and characterization of Escherichia coli enolase. *J. Biol. Chem.* **246**, 6797-6802 (1971).

195. Dannelly, H. K., Duclos, B., Cozzone, A. J. & Reeves, H. C. Phosphorylation of Escherichia coli enolase. *Biochimie* **71**, 1095-1100 (1989).

196. Schorken, U. *et al*. Identification of catalytically important residues in the active site of Escherichia coli transaldolase. *Eur. J. Biochem.* **268**, 2408-2415 (2001).

197. Zheng, J. & Jia, Z. Structure of the bifunctional isocitrate dehydrogenase kinase/phosphatase. *Nature* **465**, 961-965 (2010).

198. Conrad, T. M., Lewis, N. E. & Palsson, B. O. Microbial laboratory evolution in the era of genome-scale science. *Mol. Syst. Biol.* **7**, 509 (2011).

199. Conrad, T. M. *et al*. RNA polymerase mutants found through adaptive evolution reprogram Escherichia coli for optimal growth in minimal media. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 20500-20505 (2010).

200. Conrad, T. M. *et al*. Whole-genome resequencing of Escherichia coli K-12 MG1655 undergoing short-term laboratory evolution in lactate minimal media reveals flexible selection of adaptive mutations. *Genome Biol.* **10**, R118 (2009).

201. Herring, C. D. *et al*. Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. *Nat. Genet.* **38**, 1406-1412 (2006).

202. Tenaillon, O. *et al*. The molecular diversity of adaptive convergence. *Science* **335**, 457-461 (2012).

203. Raser, J. M. & O'Shea, E. K. Noise in gene expression: origins, consequences, and control. *Science* **309**, 2010-2013 (2005).

204. Papp, B., Pal, C. & Hurst, L. D. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* **429**, 661-664 (2004).

205. Deutscher, D., Meilijson, I., Kupiec, M. & Ruppin, E. Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nat. Genet.* **38**, 993-998 (2006).

206. Kim, J., Kershner, J. P., Novikov, Y., Shoemaker, R. K. & Copley, S. D. Three serendipitous pathways in E. coli can bypass a block in pyridoxal-5'-phosphate synthesis. *Mol. Syst. Biol.* **6**, 436 (2010).

207. Kaern, M., Elston, T. C., Blake, W. J. & Collins, J. J. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* **6**, 451-464 (2005).

208. Yu, J., Xiao, J., Ren, X., Lao, K. & Xie, X. S. Probing gene expression in live cells, one protein molecule at a time. *Science* **311**, 1600-1603 (2006).

209. Mitchell, A. *et al*. Adaptive prediction of environmental changes by microorganisms. *Nature* **460**, 220-224 (2009).

210. Savageau, M. A. Demand theory of gene regulation. II. Quantitative application to the lactose and maltose operons of Escherichia coli. *Genetics* **149**, 1677-1691 (1998).

211. Savageau, M. A. Escherichia coli Habitats, Cell Types, and Molecular Mechanisms of Gene Control. *Am. Nat.* **122**, pp. 732-744 (1983).

212. Keseler, I. M. *et al*. EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Res.* **33**, D334-7 (2005).

213. Hua, Q., Joyce, A. R., Fong, S. S. & Palsson, B. O. Metabolic analysis of adaptive evolution for in silico-designed lactate-producing strains. *Biotechnol. Bioeng.* **95**, 992-1002 (2006).

214. Schellenberger, J. *et al*. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protoc.* **6**, 1290-1307 (2011).

215. Lewis, N. E., Cho, B. K., Knight, E. M. & Palsson, B. O. Gene expression profiling and the use of genome-scale in silico models of Escherichia coli for analysis: providing context for content. *J. Bacteriol.* **191**, 3437-3444 (2009).

216. Cho, B. K., Barrett, C. L., Knight, E. M., Park, Y. S. & Palsson, B. O. Genome-scale reconstruction of the Lrp regulatory network in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* (2008).

217. Fong, S. S., Joyce, A. R. & Palsson, B. O. Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. *Genome Res.* **15**, 1365-1372 (2005).

218. Baba, T. *et al*. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, 2006.0008 (2006).

219. Hoyt, J. C. & Reeves, H. C. In vivo phosphorylation of isocitrate lyase from Escherichia coli D5H3G7. *Biochem. Biophys. Res. Commun.* **153**, 875-880 (1988).

220. Dean, A. M., Lee, M. H. & Koshland, D. E.,Jr. Phosphorylation inactivates Escherichia coli isocitrate dehydrogenase by preventing isocitrate binding. *J. Biol. Chem.* **264**, 20482-20486 (1989).

221. Jensen, R. A. Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30**, 409-425 (1976).

222. Khersonsky, O. & Tawfik, D. S. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **79**, 471-505 (2010).

223. Des Marais, D. L. & Rausher, M. D. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* **454**, 762-765 (2008).

224. Bergthorsson, U., Andersson, D. I. & Roth, J. R. Ohno's dilemma: evolution of new genes under continuous selection. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 17004-17009 (2007).

225. Veeravalli, K., Boyd, D., Iverson, B. L., Beckwith, J. & Georgiou, G. Laboratory evolution of glutathione biosynthesis reveals natural compensatory pathways. *Nat. Chem. Biol.* **7**, 101-105 (2011).

226. Boronat, A., Caballero, E. & Aguilar, J. Experimental evolution of a metabolic pathway for ethylene glycol utilization by Escherichia coli. *J. Bacteriol.* **153**, 134-139 (1983).

227. Applebee, M. K., Joyce, A. R., Conrad, T. M., Pettigrew, D. W. & Palsson, B. O. Functional and metabolic effects of adaptive glycerol kinase (GLPK) mutants in Escherichia coli. *J. Biol. Chem.* **286**, 23150-23159 (2011).

228. Aharoni, A. *et al*. The 'evolvability' of promiscuous protein functions. *Nat. Genet.* **37**, 73-76 (2005).

229. Wagner, A. Energy costs constrain the evolution of gene expression. *J. Exp. Zool. B. Mol. Dev. Evol.* **308**, 322-324 (2007).

230. Gur, E., Biran, D. & Ron, E. Z. Regulated proteolysis in Gram-negative bacteria--how and when? *Nat. Rev. Microbiol.* **9**, 839-848 (2011).

231. Gerosa, L. & Sauer, U. Regulation and control of metabolic fluxes in microbes. *Curr. Opin. Biotechnol.* **22**, 566-575 (2011).

232. Chang, R. L. *et al*. Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism. *Mol. Syst. Biol.* **7**, 518 (2011).

233. Daran-Lapujade, P. *et al*. Role of transcriptional regulation in controlling fluxes in central carbon metabolism of Saccharomyces cerevisiae. A chemostat culture study. *J. Biol. Chem.* **279**, 9125-9138 (2004).

234. Sandegren, L. & Andersson, D. I. Bacterial gene amplification: implications for the evolution of antibiotic resistance. *Nat. Rev. Microbiol.* **7**, 578-588 (2009).

235. Soo, V. W., Hanson-Manful, P. & Patrick, W. M. From the Cover: Artificial gene amplification reveals an abundance of promiscuous resistance determinants in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 1484-1489 (2011).

236. Fisher, M. A., McKinley, K. L., Bradley, L. H., Viola, S. R. & Hecht, M. H. De novo designed proteins from a library of artificial sequences function in Escherichia coli and enable cell growth. *PLoS One* **6**, e15364 (2011).

237. Holloway, A. K., Palzkill, T. & Bull, J. J. Experimental evolution of gene duplicates in a bacterial plasmid model. *J. Mol. Evol.* **64**, 215-222 (2007).

238. Lee, D. H. & Palsson, B. O. Adaptive evolution of Escherichia coli K-12 MG1655 during growth on a Nonnative carbon source, L-1,2-propanediol. *Appl. Environ. Microbiol.* **76**, 4158-4168 (2010).

239. Nam, H., Conrad, T. M. & Lewis, N. E. The role of cellular objectives and selective pressures in metabolic pathway evolution. *Curr. Opin. Biotechnol.* (2011).

240. Ibarra, R. U., Edwards, J. S. & Palsson, B. O. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* **420**, 186-189 (2002).

241. Fong, S. S., Marciniak, J. Y. & Palsson, B. O. Description and interpretation of adaptive evolution of Escherichia coli K-12 MG1655 by using a genome-scale in silico metabolic model. *J. Bacteriol.* **185**, 6400-6408 (2003).

242. Lenski, R. E. & Travisano, M. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 6808-6814 (1994).

243. Barrick, J. E. *et al*. Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature* **461**, 1243-1247 (2009).

244. Applebee, M. K., Herrgard, M. J. & Palsson, B. O. Impact of individual mutations on increased fitness in adaptively evolved strains of Escherichia coli. *J. Bacteriol.* **190**, 5087-5094 (2008).

245. Hua, Q., Joyce, A. R., Palsson, B. O. & Fong, S. S. Metabolic characterization of Escherichia coli strains adapted to growth on lactate. *Appl. Environ. Microbiol.* **73**, 4639-4647 (2007).

246. Fong, S. S., Nanchen, A., Palsson, B. O. & Sauer, U. Latent pathway activation and increased pathway capacity enable Escherichia coli adaptation to loss of key metabolic enzymes. *J. Biol. Chem.* **281**, 8024-8033 (2006).

247. Le Gac, M. *et al*. Metabolic changes associated with adaptive diversification in Escherichia coli. *Genetics* **178**, 1049-1060 (2008).

248. Kinnersley, M. A., Holben, W. E. & Rosenzweig, F. E Unibus Plurum: genomic analysis of an experimentally evolved polymorphism in Escherichia coli. *PLoS Genet.* **5**, e1000713 (2009).

249. Charusanti, P. *et al*. Genetic basis of growth adaptation of Escherichia coli after deletion of pgi, a major metabolic gene. *PLoS Genet.* **6**, e1001186 (2010).

250. Ferea, T. L., Botstein, D., Brown, P. O. & Rosenzweig, R. F. Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 9721-9726 (1999).

251. Fong, S. S. & Palsson, B. O. Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. *Nat. Genet.* **36**, 1056-1058 (2004).

252. Mahadevan, R. & Schilling, C. H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* **5**, 264-276 (2003).

253. Teusink, B. *et al*. Analysis of growth of Lactobacillus plantarum WCFS1 on a complex medium using a genome-scale metabolic model. *J. Biol. Chem.* **281**, 40041-40048 (2006).

254. Molenaar, D., van Berlo, R., de Ridder, D. & Teusink, B. Shifts in growth strategies reflect tradeoffs in cellular economics. *Mol. Syst. Biol.* **5**, 323 (2009).

255. Ferguson, P. L. & Smith, R. D. Proteome analysis by mass spectrometry. *Annu. Rev. Biophys. Biomol. Struct.* **32**, 399-424 (2003).

256. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631-637 (1997).

257. Hochhauser, S. J. & Weiss, B. Escherichia coli mutants deficient in deoxyuridine triphosphatase. *J. Bacteriol.* **134**, 157-166 (1978).

258. Kurland, C. G. & Dong, H. Bacterial growth inhibition by overproduction of protein. *Mol. Microbiol.* **21**, 1-4 (1996).

259. Cooper, T. F., Remold, S. K., Lenski, R. E. & Schneider, D. Expression profiles reveal parallel evolution of epistatic interactions involving the CRP regulon in Escherichia coli. *PLoS Genet.* **4**, e35 (2008).

260. Ansong, C. *et al*. Global systems-level analysis of Hfq and SmpB deletion mutants in Salmonella: implications for virulence and global protein translation. *PLoS One* **4**, e4809 (2009).

261. Traxler, M. F. *et al*. The global, ppGpp-mediated stringent response to amino acid starvation in Escherichia coli. *Mol. Microbiol.* **68**, 1128-1148 (2008).

262. Varma, A. & Palsson, B. O. Metabolic Capabilities of Escherichia coli II. Optimal Growth Patterns. *Journal of Theoretical Biology,* **165**, 503-522 (1993).

263. Vogel, U. & Jensen, K. F. The RNA chain elongation rate in Escherichia coli depends on the growth rate. *J. Bacteriol.* **176**, 2807-2813 (1994).

264. Klumpp, S., Zhang, Z. & Hwa, T. Growth rate-dependent global effects on gene expression in bacteria. *Cell* **139**, 1366-1375 (2009).

265. Thiele, I., Jamshidi, N., Fleming, R. M. & Palsson, B. O. Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS Comput. Biol.* **5**, e1000312 (2009).

266. Cho, B. K. *et al*. The transcription unit architecture of the Escherichia coli genome. *Nat. Biotechnol.* **27**, 1043-1049 (2009).

267. Reed, J. L. & Palsson, B. O. Genome-scale in silico models of E. coli have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. *Genome Res.* **14**, 1797-1805 (2004).

268. Gama-Castro, S. *et al*. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.* **36**, D120-4 (2008).

269. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 6640-6645 (2000).

270. Masselon, C. *et al*. Targeted comparative proteomics by liquid chromatography-tandem Fourier ion cyclotron resonance mass spectrometry. *Anal. Chem.* **77**, 400-406 (2005).

271. Qian, W. J. *et al*. Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and SEQUEST analysis: the human proteome. *J. Proteome Res.* **4**, 53-62 (2005).

272. Eng, J. K., McCormack, A. L. & Yates, J. R. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom* **5**, 976 (1994).

273. Strittmatter, E. F. *et al*. Application of peptide LC retention time information in a discriminant function for peptide identification by tandem mass spectrometry. *J. Proteome Res.* **3**, 760-769 (2004).

274. Zimmer, J. S., Monroe, M. E., Qian, W. J. & Smith, R. D. Advances in proteomics data analysis and display using an accurate mass and time tag approach. *Mass Spectrom. Rev.* **25**, 450-482 (2006).

275. Polpitiya, A. D. *et al*. DAnTE: a statistical tool for quantitative analysis of -omics data. *Bioinformatics* **24**, 1556-1558 (2008).

276. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 9440-9445 (2003).

277. Hartley, C. J. *et al*. Amplification of DNA from preserved specimens shows blowflies were preadapted for the rapid evolution of insecticide resistance. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 8757-8762 (2006).

278. Janssen, D. B., Dinkla, I. J., Poelarends, G. J. & Terpstra, P. Bacterial degradation of xenobiotic compounds: evolution and distribution of novel enzyme activities. *Environ. Microbiol.* **7**, 1868-1882 (2005).

279. Raushel, F. M. & Holden, H. M. Phosphotriesterase: an enzyme in search of its natural substrate. *Adv. Enzymol. Relat. Areas Mol. Biol.* **74**, 51-93 (2000).

280. Copley, S. D. Evolution of efficient pathways for degradation of anthropogenic chemicals. *Nat. Chem. Biol.* **5**, 559-566 (2009).

281. Sommer, M. O., Dantas, G. & Church, G. M. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* **325**, 1128-1131 (2009).

282. Dantas, G., Sommer, M. O., Oluwasegun, R. D. & Church, G. M. Bacteria subsisting on antibiotics. *Science* **320**, 100-103 (2008).

283. Pal, C., Papp, B. & Lercher, M. J. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat. Genet.* **37**, 1372-1375 (2005).

284. Lee, D. H., Feist, A. M., Barrett, C. L. & Palsson, B. O. Cumulative number of cell divisions as a meaningful timescale for adaptive laboratory evolution of Escherichia coli. *PLoS One* **6**, e26172 (2011).

285. Sridhara, S., Wu, T. T., Chused, T. M. & Lin, E. C. Ferrous-activated nicotinamide adenine dinucleotide-linked dehydrogenase from a mutant of Escherichia coli capable of growth on 1, 2-propanediol. *J. Bacteriol.* **98**, 87-95 (1969).

286. Cocks, G. T., Aguilar, T. & Lin, E. C. Evolution of L-1, 2-propanediol catabolism in Escherichia coli by recruitment of enzymes for L-fucose and L-lactate metabolism. *J. Bacteriol.* **118**, 83-88 (1974).

287. Hacking, A. J. & Lin, E. C. Disruption of the fucose pathway as a consequence of genetic adaptation to propanediol as a carbon source in Escherichia coli. *J. Bacteriol.* **126**, 1166-1172 (1976).

288. Lu, Z. *et al*. Evolution of an Escherichia coli protein with increased resistance to oxidative stress. *J. Biol. Chem.* **273**, 8308-8316 (1998).

289. Montella, C. *et al*. Crystal structure of an iron-dependent group III dehydrogenase that interconverts L-lactaldehyde and L-1,2-propanediol in Escherichia coli. *J. Bacteriol.* **187**, 4957-4966 (2005).

290. Pinchuk, G. E. *et al*. Genomic reconstruction of Shewanella oneidensis MR-1 metabolism reveals a previously uncharacterized machinery for lactate utilization. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 2874-2879 (2009).

291. Tokuriki, N. & Tawfik, D. S. Protein dynamism and evolvability. *Science* **324**, 203-207 (2009).

292. Nakahigashi, K. *et al*. Systematic phenome analysis of Escherichia coli multiple-knockout mutants reveals hidden reactions in central carbon metabolism. *Mol. Syst. Biol.* **5**, 306 (2009).

293. Patrick, W. M., Quandt, E. M., Swartzlander, D. B. & Matsumura, I. Multicopy suppression underpins metabolic evolvability. *Mol. Biol. Evol.* **24**, 2716-2722 (2007).

294. Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **10**, 866-876 (2009).

295. Gosset, G., Zhang, Z., Nayyar, S., Cuevas, W. A. & Saier, M. H.,Jr. Transcriptome analysis of Crp-dependent catabolite control of gene expression in Escherichia coli. *J. Bacteriol.* **186**, 3516-3524 (2004).

296. MANDELSTAM, J. The repression of constitutive beta-galactosidase in Escherichia coli by glucose and other carbon sources. *Biochem. J.* **82**, 489-493 (1962).

297. Portnoy, V. A. *et al*. Deletion of genes encoding cytochrome oxidases and quinol monooxygenase blocks the aerobic-anaerobic shift in Escherichia coli K-12 MG1655. *Appl. Environ. Microbiol.* **76**, 6529-6540 (2010).

298. Monroe, M. E. *et al*. VIPER: an advanced software package to support high-throughput LC-MS peptide identification. *Bioinformatics* **23**, 2021-2023 (2007).

299. Jaitly, N. *et al*. Robust algorithm for alignment of liquid chromatography-mass spectrometry analyses in an accurate mass and time tag data analysis pipeline. *Anal. Chem.* **78**, 7397-7409 (2006).

300. Auberry, K. J. *et al*. Omics.pnl.gov: A Portal for the Distribution and Sharing of Multi-Disciplinary Pan-Omics Information. *J. Proteomics Bioinform* **3**, 1-4 (2010).

301. Death, A. & Ferenci, T. Between feast and famine: endogenous inducer synthesis in the adaptation of Escherichia coli to growth with limiting carbohydrates. *J. Bacteriol.* **176**, 5101-5107 (1994).

302. Holland, S. M. *et al*. STAT3 mutations in the hyper-IgE syndrome. *N. Engl. J. Med.* **357**, 1608-1619 (2007).