UNIVERSITY OF CALIFORNIA, SAN DIEGO

Proteome allocation trade-offs in bacterial evolution and regulation

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

in

Bioinformatics & Systems Biology

by

Edward John O'Brien

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Philip E. Bourne, Co-Chair
Professor Suckjoon Jun
Professor Andrew D. McCulloch
Professor Milton H. Saier

2015

The dissertation of Edward John O'Brien is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2015

DEDICATION

To my parents, for their endless interest and support in my education and research.

To Hilary, my best friend – we made it!

EPIGRAPH

Nothing in biology makes sense except in the light of evolution.

—Theodosius Dobzhansky

Optimization models help us to test our insight into the biological constraints that influence the outcome of evolution.

—G. A. Parker & J. Maynard Smith

Evolution is cleverer than you are.

—Francis Crick

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank Bernhard Palsson and everyone who was a part of the Systems Biology Research Group (SBRG) while I was at UCSD. SBRG is full of bright, creative, and motivated scientists. It was a great atmosphere to work in for the past 5 years.

I would like to thank everyone with whom I directly worked with on my thesis, including Josh Lerman, Jose Utrilla, Douglas McCloskey, Zak King, Jon Monk, Roger Chang, Daniel Hyduke, Ke Chen, Ryan LaCroix, Troy Sandberg, Adam Feist, and Bernard Palsson. I would especially like to thank Josh Lerman and Jose Utrilla. Josh Lerman, you were an invaluable mentor as I began my graduate research; in some ways, all of my thesis work was influenced by ideas we talked about while creating the ME-Model. Jose Utrilla, I could not have asked for more from an experimental collaborator; you are enthusiastic and excited to test and expand upon model predictions. I would also like to thank others whom I have had the pleasure to work with on projects and publications not included in my thesis including, Ali Ebrahim, Joanne Liu, Steve Federowicz, Liz Brunk, Alex Thomas, Laurence Yang, Aarash Bordbar, Daniel Zielinski, Nathan Lewis, Justin Tan, and Colton Lloyd. Also, a big thanks to Marc Abrams, Helder Balelo, Yana Campen, and Kathy Andrews for essential administrative support and manuscript editing.

I would finally like to thank the funding sources supporting my thesis research: the ARCS Foundation and the generous donation from Peter Ellsworth for my award, the National Institute of Health (NIH) training grant to the Bioinformatics and Systems Biology program at UCSD, NIH grant GM057089 to develop genome-scale models for *E. coli*, the Novo Nordisk Foundation, and the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the US Department of Energy (DOE) under Contract No. DE-AC02-05CH11231 for essential computational

resources.

The text of Chapter 1 is a full reprint of the material as it appears in: OBrien E.J.*, Monk J.A.*, Palsson B.O. Using genome-scale models to predict biological capabilities, Cell, 161(5):971-987. (2015). * indicates equal contribution. The dissertation author was the primary author of the manuscript. The other authors were Jon A. Monk (equal contributor) and Bernard Ø. Palsson.

The text of Chapter 2 is a full reprint of the material as it appears in: O'Brien E.J.*, Lerman J.A.*, Chang R.L., Hyduke D.R., Palsson B.O. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol. Syst. Biol., 9:693. (2013). * indicates equal contribution. The dissertation author was the primary author of the manuscript. The other authors were Joshua A. Lerman (equal contributor), Roger L. Chang, Daniel R. Hyduke, and Bernard Ø. Palsson.

The text of Chapter 3 is a full reprint of the material as it appears in: OBrien E.J., Utrilla J., Palsson B.O. Proteome allocation constraints determine cellular growth rates and demand fitness trade-offs, Submitted. The dissertation author was the primary author of the manuscript. The other authors were Jose Utrilla and Bernard Ø. Palsson.

The text of Chapter 4 is a full reprint of the material as it appears in: OBrien E.J.*, McCloskey D.*, Utrilla J., King Z.A., LaCroix R.A., Sandberg T.E., Feist A.M., Palsson B.O. Tradeoffs in microbial adaptation are determined by proteome complexity, Submitted. * indicates equal contribution. The dissertation author was the primary author of the manuscript. The other authors were Douglas McCloskey (equal contributor), Jose Utrilla, Zachary A. King, Ryan A. LaCroix, Troy E. Sandberg, Adam M. Feist, and Bernard Ø. Palsson.

The text of Chapter 5 is a full reprint of the material as it appears in: Utrilla J.*, OBrien E.J.*, Chen K., McCloskey D., Cheung J., Wang H., Armenta-Medina D., Feist A.M., Palsson B.O. Proteome and Energy Re-allocation by Adaptive Regulatory Mutations Reveals a Fitness Trade-off, Submitted. * indicates equal contribution. The dissertation author was the primary author of the manuscript. The other authors were Jose Utrilla (equal contributor), Ke Chen, Douglas McCloskey, Jacky Cheung, Harris Wang, Dagoberto Armenta-Medina, Adam M. Feist, and Bernard Ø. Palsson.

The text of Chapter 6 is a full reprint of the material as it appears in: OBrien E.J, Palsson B.O. Computing the functional proteome: recent progress and future prospects for genome-scale models. Curr. Opin. Biotechnol. 34C:125-134. (2015). The dissertation author was the primary author of the manuscript. The other author was Bernard Ø. Palsson.

VITA

2010	B. S. in Mathematics *summa cum laude*, Tufts University
2010	B. S. in Engineering Science, Tufts University
2015	Ph. D. in Bioinformatics & Systems Biology, University of California, San Diego

PUBLICATIONS

OBrien E.J.*, McCloskey D.*, Utrilla J., King Z.A., LaCroix R.A., Sandberg T.E., Feist A.M., Palsson B.O. Tradeoffs in microbial adaptation are determined by proteome complexity, Submitted

OBrien E.J., Utrilla J., Palsson B.O. Proteome allocation constraints determine cellular growth rates and demand fitness trade-offs, Submitted

Utrilla J.*, **OBrien E.J.***, Chen K., McCloskey D., Cheung J., Wang H., Armenta-Medina D., Feist A.M., Palsson B.O. Proteome and Energy Re-allocation by Adaptive Regulatory Mutations Reveals a Fitness Trade-off, Submitted

Yang L., Tan J., **OBrien E.J.**, Monk J., Kim D., Li H.J., Charusanti P., Ebrahim A., Lloyd C., Yurkovich J.T., Du B., Drager A., Thomas A., Sun Y., Saunders M.A., Palsson B.O. A model-predicted core proteome of metabolism and expression is consistent with high-throughput data, Proc. Natl. Acad. Sci. (2015).

Seo S.W., Kim D., **OBrien E.J.**, Szubin R., Palsson B.O. Decoding genome-wide GadEWX transcriptional regulatory networks reveals a multifaceted cellular response to acid stress in Escherichia coli, Nat. Comm., 6:7970. (2015).

OBrien E.J.*, Monk J.A.*, Palsson B.O. Using genome-scale models to predict biological capabilities, Cell, 161(5):971-987. (2015)

OBrien E.J, Palsson B.O. Computing the functional proteome: recent progress and future prospects for genome-scale models. Curr. Opin. Biotechnol. 34C:125-134. (2015).

LaCroix R.A., Sandberg T.E., **OBrien E.J.**, Utrilla J., Ebrahim A., Guzman G.I., Szubin R., Palsson B.O., Feist A.M. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. Appl. Envrion. Microbiol. 81(1):17-30. (2015).

Joanne J.K., **OBrien E.J.**, Lerman J.A., Zengler K., Palsson B.O., Feist A.M. Reconstruction and modeling protein translocation and compartmentalization in Escherichia coli at the genome-scale. BMC Syst. Biol., 8:110. (2014).

Seo S.W., Kim D., Latif H., **OBrien E.J.**, Szubin R., Palsson B.O. Deciphering Fur transcriptional regulatory network highlights its complex role beyond iron metabolism in Escherichia coli. Nat. Comm., 5:4910. (2014).

O'Brien E.J.*, Lerman J.A.*, Chang R.L., Hyduke D.R., Palsson B.O. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol. Syst. Biol., 9:693. (2013).

Lewis N.E., Liu X., Li Y., Nagarajan H., Yerganian G., **OBrien E.J.**, Bordbar A., Roth A.M., Rosenbloom J., Bian C., Xie M., Chen W., Li N., Baycin-Hizal D., Latif H., Forster J., Betenbaugh M.J., Famili I., Xu X., Wang J., Palsson B.O. The genomic divergence of CHO cell lines is revealed by the genomic sequence of the Chinese Hamster, Cricetulus griseus. Nat. Biotech., 31:759-765. (2013).

Hyduke D.R., Lerman J.A., Palsson B.O., **O'Brien E.J.**, Method for in silico modeling of gene product expression and metabolism. U.S. Patent Application. Application number: 14/399,129. Filed November 2014.

ABSTRACT OF THE DISSERTATION

Proteome allocation trade-offs in bacterial evolution and regulation

by

Edward John O'Brien

Doctor of Philosophy in Bioinformatics & Systems Biology

University of California, San Diego, 2015

Professor Bernhard Ø. Palsson, Chair
Professor Philip E. Bourne, Co-Chair

The abundance of proteins expressed in a particular environment are primary determinants of an organism's phenotypic and fitness properties. However, protein synthesis is costly and proteome size is limited; thus, the benefit of expressing proteins also comes with costs. In this thesis, I interrogate the evolutionary and regulatory trade-offs resulting from these proteome allocation constraints. Throughout the thesis I employ a genome-scale model of metabolism and protein synthesis for *Escherichia coli*, which can compute condition-specific proteome allocation requirements and limitations. First, I show that microbial growth rates are quantitatively determined by the expression of

unused protein. Rather than supporting growth in the current environment, large fractions of the expressed proteome enable readiness for environmental change and stress. The expression of these different proteome segments is regulated by global transcription factors and results in fitness trade-offs. Second, I show that after selecting for growth through experimental evolution, several adaptive regulatory mutations increase fitness through proteome and energy resource re-allocation. These pleiotropic mutations in the RNA Polymerase systematically re-allocate the proteome towards growth and away from stress resistance, showing that fitness trade-offs are readily modulated by global regulators during evolution. Finally, I show that the diversity present in evolving populations is predictable and due to proteome allocation trade-offs. Rather than evolving to a unique optimum, a range of near-optimal proteomic and metabolic phenotypes is apparent when strains are independently evolved in the same environment. The diversity of alternative phenotypes reflects a rate-yield trade-off due to the varying protein cost of metabolic pathways in central carbon metabolism. Thus, proteome allocation constraints have a pervasive and predictable effect on bacterial ecology, regulation, and evolution.

Chapter 1

Using Genome-Scale Models to Predict Biological Capabilities

Prediction is very difficult, especially about the future.
—Nils Bohr

1.1 Abstract

Constraint-based reconstruction and analysis (COBRA) methods at the genome-scale have been under development since the first whole genome sequences appeared in the mid-1990s. A few years ago this approach began to demonstrate the ability to predict a range of cellular functions including cellular growth capabilities on various substrates and the effect of gene knockouts at the genome-scale. Thus, much interest has developed in understanding and applying these methods to areas such as metabolic engineering, antibiotic design, and organismal and enzyme evolution. This primer will get you started.

1.2 Introduction

Bottom-up approaches to systems biology rely on constructing a mechanistic basis for the biochemical and genetic processes that underlie cellular functions. Genome-scale network reconstructions of metabolism are built from all known metabolic reactions and metabolic genes in a target organism. Networks are constructed based on genome annotation, biochemical characterization, and the published scientific literature on the target organism; the latter is sometimes referred to as the bibliome. DNA sequence assembly provides a useful analogy to the process of network reconstruction (Figure 1.1). The genome of an organism is systematically assembled from many short DNA stubs, called reads, using sophisticated computer algorithms. Similarly, the reactome of a cell is assembled, or reconstructed, from all the biochemical reactions known or predicted to be present in the target microorganism. Importantly, network reconstruction includes an explicit genetic basis for each biochemical reaction in the reactome as well as information about the genomic location of the gene. Thus, reconstructed networks, or an assembled reactome, for a target organism represents biochemically, genetically, and genomically structured knowledge bases, or BiGG k-bases. Network reconstructions have different biological scope and coverage. They may describe metabolism, protein-protein interactions, regulation, signaling, and other cellular processes, but they have a unifying aspect: an embedded, standardized biochemical and genetic representation amenable to computational analysis.

A network reconstruction can be converted into a mathematical format and thus lend itself to mathematical analysis and computational treatment. Genome-scale models, called GEMs, have been under development for nearly 15 years and have now reached a high level of sophistication. The first GEM was created for Haemophilus influenza and appeared shortly after this first genome was sequenced [1], and GEMs have now

grown to the level where they enable predictive biology [2, 3, 4]. Here, we will focus on reconstructions of metabolism and the process of converting them into GEMs to produce computational predictions of biological functions.

The fundamentals of the Constraints-Based Reconstruction and Analysis (COBRA) approach and its uses are also described in this Primer, which lays out the constraint-based methodology out at four levels. First, there is a textual description of the methods and their applications. Second, visualization is presented in the form of detailed figures to succinctly convey the key concepts and applications. Third, the figure captions contain more detailed information about the computational approaches illustrated in the figures. Fourth, the primer provides a table of selected detailed resources to enable an in-depth review for the keenly interested reader. The text is organized into six sections, each one addressing a grand challenge in todays world of big data biology:

1.3 Network Reconstructions Assemble Knowledge Systematically

A large library of scientific publications exists that describe different model organisms specific molecular features. Molecular biologys focus on knowing much about a limited number of molecular events changed once annotated genome sequences became available, leading to the emergence of a genome-scale point of view. Now, putting all available knowledge about the molecular processes of a target organism in context and linked to its genome sequence has emerged as a grand challenge. Genome-scale network reconstructions were a response to this challenge.

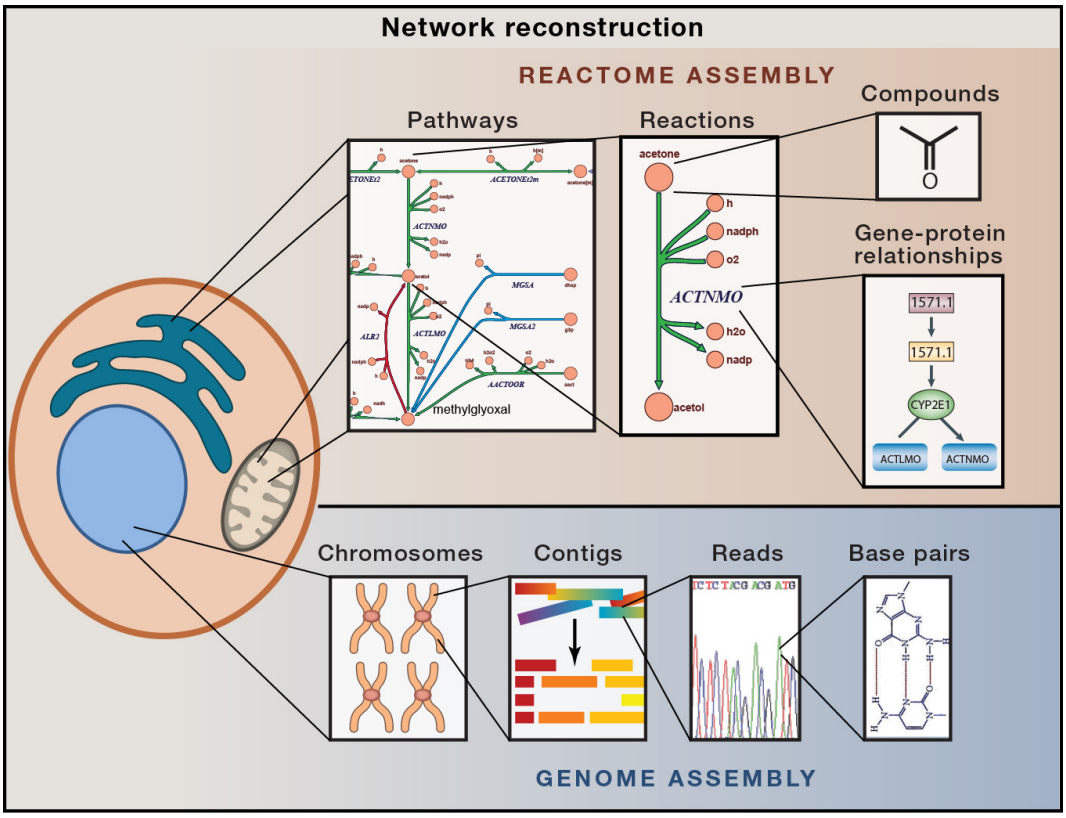


Figure 1.1: **Network reconstruction.** An organisms reactome can be assembled in a way that is analogous to DNA sequencing assembly. From right to left: first the interacting compounds must be identified. Then, the reactions acting on these compounds are tabulated and the protein that catalyzes the reaction and the corresponding open reading frame is identified in the organism of interest. These reactions are assembled into pathways that can be laid out graphically to visualize a cell's metabolic map at the genome-scale. Several tools for reactome assembly and curation exist including the COBRA Toolbox [5, 6], KEGG [7], EcoCyc [8], ModelSeed [9], BiGG [10], Rbionet [11], Subliminal [12], Raven toolbox [13] and others.

1.3.1 Network reconstructions organize knowledge into a structured format

The reconstruction process treats individual reactions as the basic elements of a network, somewhat similar to a base pair being the smallest element in an assembled DNA sequence (Figure 1.1). To implement the metabolic reconstruction process, a series of questions need to be answered for each of the enzymes in a metabolic network: 1) What are the substrates and products? 2) What are the stoichiometric coefficients for each metabolite that participates in the reaction (or reactions) catalyzed by an enzyme? 3) Are these reactions reversible? 4) In what cellular compartment does the reaction occur? 5) What gene(s) encode for the protein (or protein complex) and what is (are) their genomic location(s)? Genes are linked to the proteins they encode and the reactions they catalyze using the gene-protein-reaction relationship (GPR). All of this information is assembled from a range of sources including organism specific databases, high-throughput data, and primary literature. Establishing a set of the biochemical reactions that constitute a reaction network in the target organism culminates in a database of chemical equations. Reactions are then organized into pathways, pathways into sectors (such as amino acid synthesis), and ultimately into genome-scale networks, akin to reads becoming a full DNA sequence. This process has been described in the form of a 96-step standard operating procedure [14].

Today, after many years of hard work by many researchers, there exist collections of genome-scale reconstructions (sometimes called GENREs) for a number of target organisms [15, 16] and established protocols for reconstruction exist [14] that can be partially automated [9, 13].

1.3.2 Recapitulation.

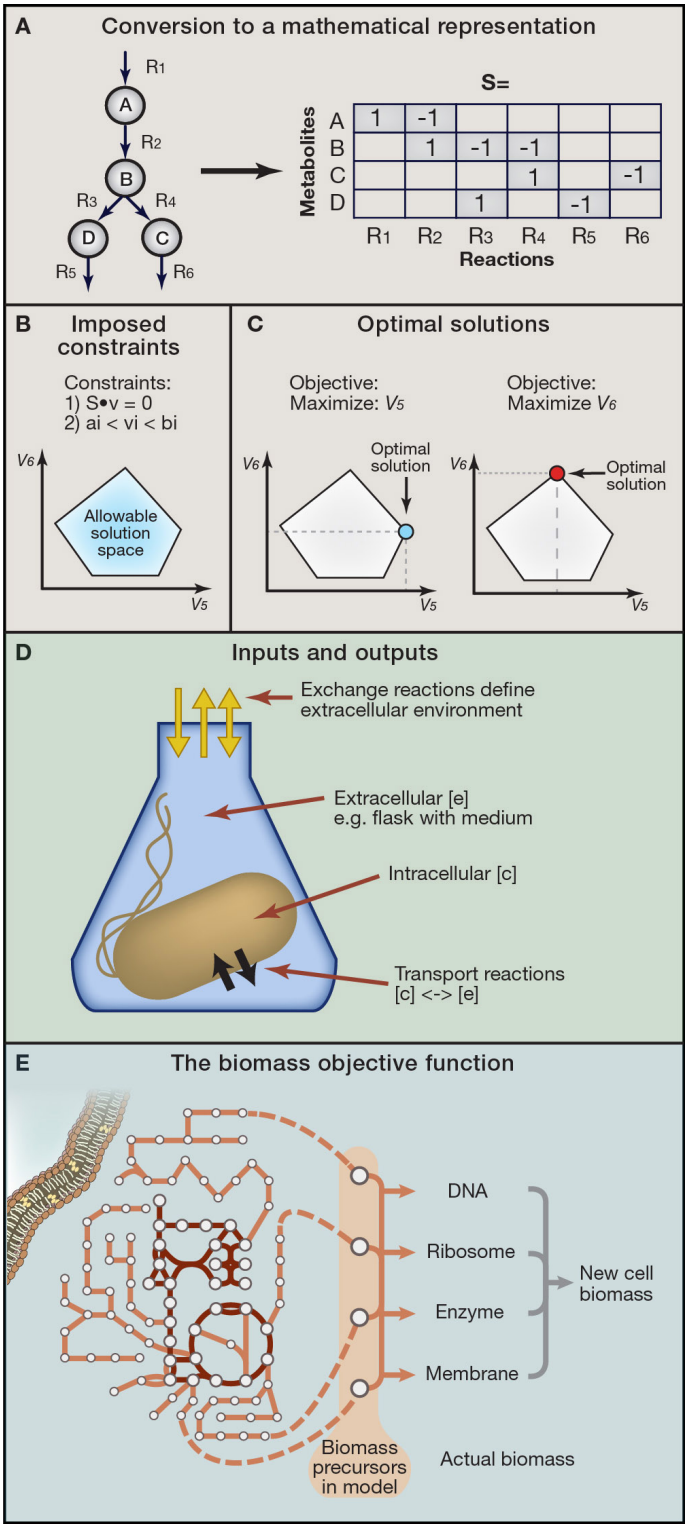
Network reconstructions represent an organized process for genome-scale assembly of disparate information about a target organism. All this information is put into context with the annotated genome to form a coherent whole that, through computations, is able to recapitulate whole cell functions. The grand challenge of disparate data integration into a coherent whole is achieved through the formulation of a GEM. A GEM can then compute cellular states such as an optimal growth state. This process is further explored in the next section. A detailed reading list is available in Supplemental Table 1 of [17] on the network reconstruction process and software tools used to facilitate it.

1.4 Converting a Genome-scale Reconstruction to a Computational Model

Before a reconstruction can be used to compute network properties, a subtle, but crucial step must be taken in which a network reconstruction is mathematically represented. This conversion translates a reconstructed network into a chemically accurate mathematical format that becomes the basis for a genome-scale model (Figure 1.2A). This conversion requires the mathematical representation of metabolic reactions. The core feature of this representation is tabulation, in the form of a numerical matrix, of the stoichiometric coefficients of each reaction (Figure 1.2B). These stoichiometries impose systemic constraints on the possible flow patterns (called a flux map, or flux distribution) of metabolites through the network. These concepts are detailed below. Imposition of constraints on network functions fundamentally differentiates the COBRA approach from models described by biophysical equations, which require many difficult-to-measure kinetic parameters.

Constraints are mathematically represented as equations that represent balances or as inequalities that impose bounds (Figure 1.2C). The matrix of stoichiometries imposes flux balance constraints on the network, ensuring that the total amount of any compound being produced must be equal to the total amount being consumed at steady state. Every reaction can also be given upper and lower bounds, which define the maximum and minimum allowable fluxes through the reactions, that in turn are related to the turnover number of the enzyme and its abundance. Once imposed on a network reconstruction, these balances and bounds define a space of allowable flux distributions in a network; the possible rates at which every metabolite is consumed or produced by every reaction in the network. The flux vector, a mathematical object, is a list of all such flux values for a single point in the space. The flux vector represents a state of the network that is directly related to the physiological function that the network produces. Many other constraints such as substrate uptake rates, secretion rates, and other limits on reaction flux can also be imposed, further restricting the possible state that a reconstructed network can take [18]. The computed network states that are consistent with all imposed constraints are thus candidate physiological states of the target organisms under the conditions considered. The study of the properties of this space thus becomes an important subject.

Figure 1.2: **Formulation of a computational model.** A. After the metabolic network has been assembled it must be converted into a mathematical representation. This conversion is performed using a stoichiometric (S) matrix where the stoichiometry of each metabolite involved in a reaction is enumerated. Reactions form the columns of this matrix and metabolites the rows. Each metabolite's entry corresponds to its stoichiometric coefficient in the corresponding reaction. Negative coefficient substrates are consumed (reactants), and positive coefficients are produced (products). Converting a metabolic network reconstruction to a mathematical formulation can be achieved with several of the toolboxes listed in Supplemental Table 1 of [17]. B. Constraints can be added to the model such as 1) enforcement of mass balance and 2) reaction flux (v) bounds. The blue polytope represents different possible fluxes for reactions 5 and 6 consistent with stated constraints. Those outside the polytope violate the imposed constraints and are thus infeasible. C. Constraint-based models predict the flow of metabolites through a defined network. The predicted path is determined using linear programming solvers and termed Flux Balance Analysis (FBA). FBA can be used to calculate the optimal flow of metabolites from a network input to a network output. The desired output is described by an objective function. If the objective is to optimize flux through reaction 5, the optimal flux distribution would correspond to the levels of flux 5 and flux 6 at the blue point circled in the figure. The objective function can be a simple value or draw on a combination of outputs, such as the biomass objective shown in Fig 2E. It is important to note that alternate optimal flux distributions may exist to reach the optimal state as discussed in Figure 1.4C. D. Once a network reconstruction is converted to a mathematical format, the inputs to the system must be defined by adding consideration of the extracellular environment. Compounds enter and exit the extracellular environment via exchange reactions. The GEM will not be able to import compounds unless a transport reaction from the external environment to the inside of the cell is present. E. In addition to exchange reactions, the biomass objective function acts as a drain on cellular components in the same ratios as they are experimentally measured in the biomass. In FBA simulations the biomass function is used to simulate cellular growth. The biomass function is composed of all necessary compounds needed to create a new cell including DNA, amino acids, lipids and polysaccharides. This is not the only physiological objective that can be examined using COBRA tools.


A
Conversion to a mathematical representation
R1
A
R2
B
R3
R4
D
C
R5
R6
S=
Metabolites
A 1 -1
B 1 -1 -1
C 1 -1
D 1 -1
R1 R2 R3 R4 R5 R6
Reactions
B
Imposed constraints
Constraints:
1) S•v = 0
2) ai < vi < bi
V6
Allowable solution space
V5
C
Optimal solutions
Objective:
Maximize: V5
Objective:
Maximize V6
V6
Optimal solution
V5
V6
Optimal solution
V5
D
Inputs and outputs
Exchange reactions define extracellular environment
Extracellular [e]
e.g. flask with medium
Intracellular [c]
Transport reactions
[c] <-> [e]
E
The biomass objective function
DNA
Ribosome
Enzyme
Membrane
New cell biomass
Biomass precursors in model
Actual biomass

1.4.1 Flux balance analysis (FBA) calculates candidate phenotypes

FBA is the oldest COBRA method. It is a mathematical approach for analyzing the flow of metabolites through a metabolic network [19]. This approach relies on an assumption of steady-state growth and mass balance (all mass that enters the system must leave). The constraints discussed above take the form of equalities and inequalities to define a polytope (blue area within the illustration in Figure 1.2C) that represents all possible flux states of the network given the constraints imposed. Thus, many network states are possible under the given constraints and multiple solutions exist that satisfy the governing equations. The blue area is therefore often called the solution space to denote a mathematical space that is filled with candidate solutions to the network equations given the governing constraints. FBA uses the stated objective to find the solution(s) that optimize the objective function. The solution is found using linear programming, and, as indicated in Figure 1.2D, the optimal solution lies at the edges of the solution space impinging up against governing constrains.

The utility of FBA has been increasingly recognized due to its simplicity and extensibility: it requires only the information on metabolic reaction stoichiometry and mass balances around the metabolites under pseudo-steady state assumption. It computes how the flux map must balance to achieve a particular homeostatic state. However, FBA has limitations. It balances fluxes, but cannot predict metabolite concentrations. Except in some modified forms, FBA does not account for regulatory effects such as activation of enzymes by protein kinases or regulation of gene expression. More details are found in the caption of Figure 1.2, and computational resources are summarized below that can be deployed to find the optimal state and to study its characteristics.

1.4.2 Models impose constraints and allow prediction.

One of the most basic constraints imposed on genome-scale models of metabolism is that of substrate, or nutrient, availability and its uptake rate (Figure 1.2E). Metabolites enter and leave the systems through what are termed exchange reactions (i.e., active or passive transport mechanisms). These reactions define the extracellular nutritional environment and are either left open (to allow a substrate to enter the system at a specified rate) or closed (the substrate can only leave the system). Measurements of the rate of exchange with the environment are relatively easy to perform and they prove to be some of the more important constraints placed on the possible functions of reaction networks internal to the cell. More biological- and data-derived constraints can also be imposed on a model. These advanced constraints are detailed in sections 4, 5 and 6.

The next step in converting a network reconstruction to a model is to define what biological function(s) the network can achieve. Mathematically, such a statement takes the form of an objective function. For predicting growth, the objective is biomass production, that is, the rate at which the network can convert metabolites into all required biomass constituents such as nucleic acids, proteins, and lipids needed to produce biomass. The objective of biomass production is mathematically represented by a biomass reaction that becomes an extra column of coefficients in the stoichiometric matrix. One can formulate a biomass objective function at an increasing level of detail: Basic, Intermediate, and Advanced [20]. The biomass reaction is scaled so that the flux through it represents the growth rate () of the target organism.

It is important to note that the biomass objective function is determined from measurements of biomass composition, the uptake and secretion rates from measuring the nutrients in the medium, and the model formulation is based on a network reconstruction that is knowledge-based. Thus, the growth rate optimization problem represents big data integrated into a structured format and the hypothesis of a biological objective; grow as

fast as possible with the resources available. This is a well-defined optimization problem.

1.4.3 GEMs are input-output flow models.

The inner workings of a GEM are readily understood conceptually. In a given environment (i.e., where the nutritional inputs are defined) GEMs can be used to compute network outputs. Flux balance analysis (FBA) can computationally trace a fully balanced path through the reactome from the available nutrients to the prerequisite output metabolite. Such calculations are performed as detailed above with an objective function that describes the removal of the target metabolite from the network. The synthesis of biomass in a cell requires the simultaneous removal of about 60-70 different metabolites. Using FBA, a GEM can also compute the balanced use of the reactome to produce all the prerequisite metabolites for growth simultaneously, and does so in the correct relative amounts while accounting for all the energetic, redox, and chemical interactions that must balance to enable such biomass synthesis. This exercise is one of genome-scale accounting of all molecules flowing through the reactome.

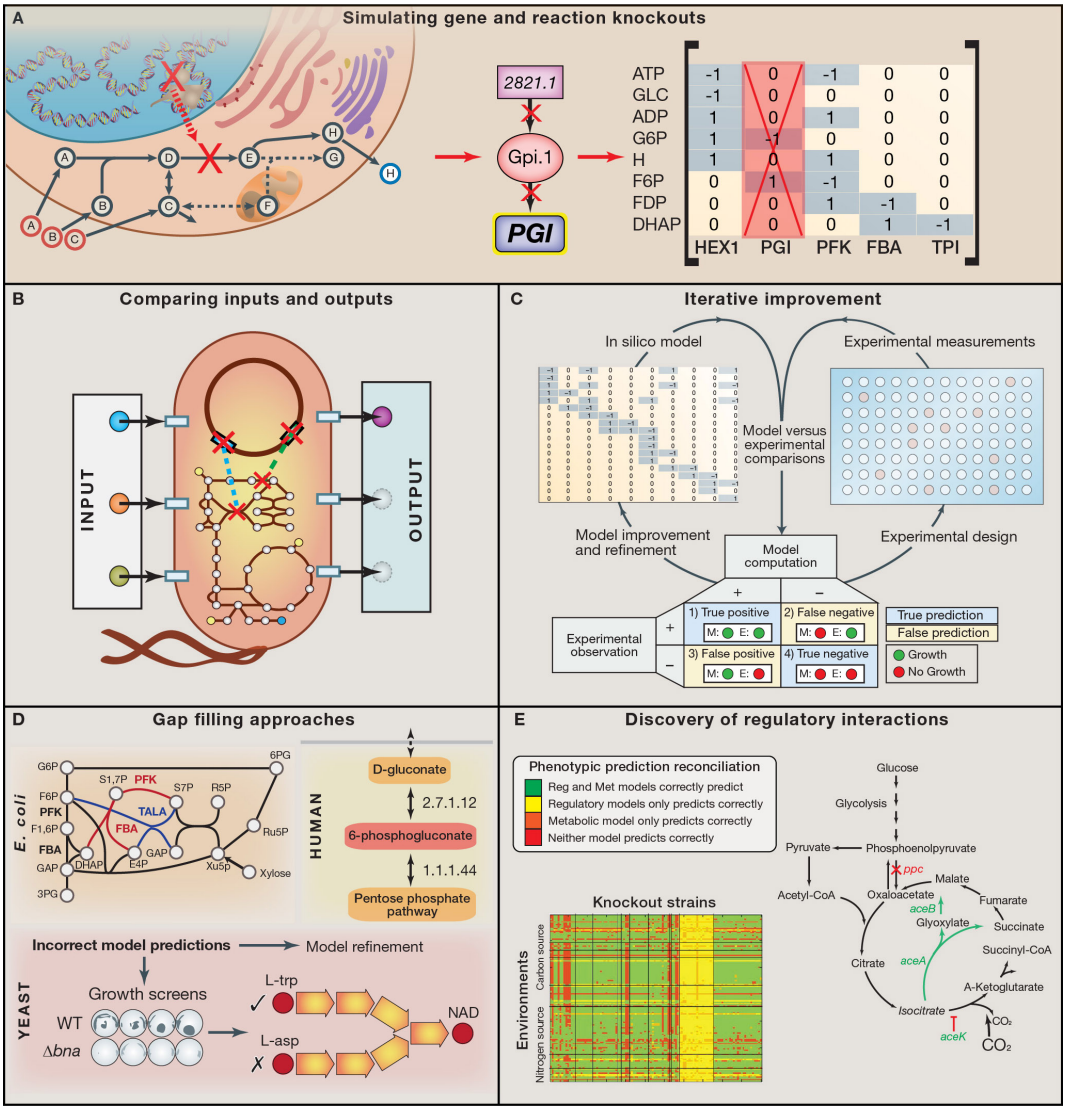
1.4.4 Recapitulation.

Given its simplicity and utility, FBA has become one of the most widely employed computational techniques for the systems-level analysis of living organisms [4, 21]. It has been successfully applied to a multitude of species for modeling their cellular metabolisms [2, 22, 3], and therefore, enabled a variety of applications such as metabolic engineering for the over-production of biochemicals [23, 24], identification of anti-microbial drug-targets [25], and the elucidation of cellcell interactions , [26]. Further reading and detailed descriptions of FBA and sources for existing genome-scale models are available in Supplemental Table 1 of [17].

1.5 Validation and reconciliation of qualitative model predictions

Ensuring the consistency and accuracy of all the information available for a target organism is a grand challenge of genome-scale biology. Since model predictions are based on a network reconstruction that represents the totality of what is known about a target organism, such predictions are a critical test of our comprehensive understanding of the metabolism for the target organism. Incorrect model predictions can be used for biological discovery by classifying them and understanding their underlying causes. Performing targeted experiments to understand failed predictions is a proven method for systematic discovery of new biochemical knowledge [27]. This section will focus on evaluating qualitative model predictions, their outcomes, underlying causes of incorrect predictions, and how to go about correcting them. Section 4 discusses the same process for quantitative model predictions.

Figure 1.3: **Using models for qualitative predictions and iterative improvement** A. Each reaction in the network is linked to a protein and encoding gene through the gene-protein-reaction (GPR) relationship. Because each reaction in the network corresponds to a column in the stoichiometric matrix, simply removing the column association with a particular reaction can simulate gene knockouts. Thus, multiple KO simulations can be performed. For example, it is easy to delete every pairwise combination of 136 central carbon metabolic E. coli genes to find double gene knockouts that are essential for survival of the bacteria. B. The simplicity of altering inputs to change cellular growth environments and removing genes in silico allows one to perform simulations in millions of experimental conditions quickly. Even on a modest laptop computer a single FBA calculation runs in a fraction of a second, thus simulating the effect of all gene knockouts in E. coli central metabolism can be run in less than 10 seconds. C. Incorrect model predictions are an opportunity for biological discovery because they highlight where knowledge is missing. Targeted experiments can be performed to discover new content that can then be added back to a model to improve its predictive accuracy. Missing model content can be discovered using automated approaches known as 'gap-filling' [27] that query a universal database of potential reactions to restore in silico growth to a model. D. Gap-filling approaches have been used to discover new metabolic reactions in several organisms. E. coli: Two new functions for two classical glycolytic enzymes phosphofructokinase (PFK) and fructose-bisphosphate aldolase (FBA) were discovered (red) [28]. Human: Gluconokinase (EC 2.7.1.12) activity was discovered based on the known presence of the metabolite 6-phosphogluconolactonate in the human reconstruction [29] (red). Yeast: Automated model refinement suggested modifications in the NAD biosynthesis pathway. Experiments demonstrated that a parallel pathway from aspartate thought to exist in yeast was not present [30]. E. False positive predictions can be reconciled by adding regulatory rules derived from high throughput data [31], for example, a recent study was able to reconcile 2,442 false model predictions from the E. coli GEM by updating the function of just 12 genes [32]. Additionally, a false positive growth inconsistency in the metabolic model of S. Typhimurium was reconciled by updating regulatory rules for the iclR gene products transcriptional repression of aceA encoding isocitrate lyase. Transcriptional repression can also often be relieved via adaptive laboratory evolution. Such evolution drives experimental phenotypes to achieve model predictions. Several experimental studies have shown that an organism can evolve to achieve model-predicted optimal growth state [33].


A
Simulating gene and reaction knockouts
2821.1
Gpi.1
PGI
ATP -1 0 -1 0 0
GLC -1 0 0 0 0
ADP 1 0 1 0 0
G6P 1 -1 0 0 0
H 1 0 1 0 0
F6P 0 1 -1 0 0
FDP 0 0 1 -1 0
DHAP 0 0 0 1 -1
HEX1 PGI PFK FBA TPI
B
Comparing inputs and outputs
INPUT
OUTPUT
C
Iterative improvement
In silico model
Experimental measurements
Model versus experimental comparisons
Model improvement and refinement
Model computation
Experimental design
Experimental observation
1) True positive
2) False negative
3) False positive
4) True negative
True prediction
False prediction
Growth
No Growth
D
Gap filling approaches
E. coli
G6P
F6P
PFK
F1,6P
FBA
GAP
DHAP
3PG
S1,7P
S7P
R5P
TALA
E4P
6PG
Ru5P
Xu5P
Xylose
HUMAN
D-gluconate
2.7.1.12
6-phosphogluconate
1.1.1.44
Pentose phosphate pathway
Incorrect model predictions
Model refinement
YEAST
Growth screens
WT
Δbna
L-trp
L-asp
NAD
E
Discovery of regulatory interactions
Phenotypic prediction reconciliation
Reg and Met models correctly predict
Regulatory models only predicts correctly
Metabolic model only predicts correctly
Neither model predicts correctly
Knockout strains
Environments
Carbon source
Nitrogen source
Glucose
Glycolysis
Pyruvate
Phosphoenolpyruvate
ppc
Malate
Acetyl-CoA
Oxaloacetate
aceB
Fumarate
Glyoxylate
Succinate
Citrate
aceA
Succinyl-CoA
A-Ketoglutarate
Isocitrate
aceK
CO2

1.5.1 Genetic and environmental parameters.

Genome-scale models have many genetic and environmental parameters that can be experimentally varied. Altering the composition of the growth media changes environmental parameters. Alteration of genetic parameters is achieved through genome editing methods. Both environmental and genetic parameters are explicit in GEMs and thus the consequence of both types of perturbations can be computed, predicted, and analyzed. The scale of such predictions has grown steadily since the first genome-scale model of E. coli appeared in 2000 [34].

Genome-scale gene essentiality data are available from specific projects or organism-specific databases. One can systematically remove genes from a reconstruction, and thus the corresponding reactions from the reactome, and repeat the growth computation to predict gene essentiality; i.e., if a growth state cannot be computed without a particular gene, the GEM predicts it to be essential (Figure 1.3A). Such growth rate predictions of gene deletion strains have gone from a hundred predictions in the year 2000 [34], to over 100,000 predictions in 2012 [35], and may be heading for over a million predictions in just a few years [36].

Both environmental and genetic parameters can be varied when performing FBA. The simplicity of computing growth states (i.e., an output) as a function of media composition (i.e., the nutritional inputs) with the selective removal of genes (Figure 1.3B) has led to a number of studies that cross environmental parameters with gene deletions. The explicit relationship between a gene and a reaction makes the deletion of genes and their encoding reactions straightforward. You can readily do this for your target organism, provided that you can construct a library of gene deletion strains. Improved molecular tools for generating knockout collection libraries (Tn-seq, CRISPR systems, etc.) and improved high-throughput methods for measuring knockout phenotypes have enabled a massive scale-up in the number of phenotypes that can be measured.

1.5.2 Classification of model predictions.

Computational predictions of outcomes fall into four categories: true-positives, true-negatives, false-positives and false-negatives. The true-positive and true-negative predictions, where computational predictions and experimental outcomes agree, have generally exceeded 80

FBA based models are highly precise because they are good at predicting impossible states (such as when a gene knockout leads to death). This assumes that the network structure is complete, an assumption that can be a problem when promiscuous enzyme activity arises, leading to a reaction with an encoding gene that is not captured in the model. Models have lower accuracy because FBA assumes that all reactions can happen at maximum rates. Model false positives often occur because an enzyme is either transcriptionally repressed or does not catalyze the designated reaction at a high enough rate (Supplemental Table 2 in [17], Evaluation of Model Predictions). Predictive failure is perhaps of more interest than success as it represents an opportunity for biological discovery. False negative predictions occur when a GEM predicts the inability to grow in a given environment without the deleted gene, but the experiments show growth. This discrepancy indicates that the reconstructed reactome is incomplete. In contrast, false positive predictions occur when a GEM predicts growth but the experiment results in no growth. This outcome indicates possible errors in the knowledge on which the reactome was based, or that a regulatory process is missing that prevents the use of a gene product factored in the computed solution. An example would be regulation that either represses gene expression or a metabolite-enzyme interaction that inhibits the function of an enzyme that the GEM used to compute the predicted growth state. Prediction failures can be used to systematically (i.e., algorithmically) generate hypotheses addressing the failures. Such hypotheses have been shown to direct experimentation to improve our knowledge base for the target organism. Computations that vary environmental

and genetic parameters become part of a workflow (Figure 1.3C). The outcome of the workflow is a set of qualitative model predictions of growth or no growth that are then compared to the experimental outcome of a growth screen. Correct predictions align with experimental results, while incorrect predictions do not. The two are then compared and classified into four categories as shown in Figure 1.3C. The failure modes lead to systematic experimentation.

1.5.3 Discovery using model false negatives.

Reconciling such discrepancies between predicted and observed growth states is now a proven approach for biological discovery. A series of algorithms have been developed that have been shown to compute the most likely reasons for failure of prediction that in turn led to a model-guided experimental inquiry and discovery. Furthermore, high-throughput tools such as phenotypic microarrays and robotic instruments are becoming available to screen cells at high rates. Such discoveries are then incorporated into the reconstruction, leading to its iterative improvement.

The discrepancies between GEM predictions and experimental data have been used to design targeted experiments that correct inaccuracies in metabolic knowledge. In this subsection we provide three illustrative examples that detail how reconciliation of model errors led to the discovery of new metabolic capabilities in three model organisms.

Human

The activity of open reading frame 103 on chromosome 9 (C9orf103) of the human genome was discovered [29] using established gap-filling protocols [27, 37]. The authors focused on unconnected, dead end metabolites in the human metabolic network reconstruction, Recon 1 [38]. Dead end metabolites lead to model errors by creating blocked reactions due to a violation of mass balance. Any flux leading to them cannot

leave the network. In an attempt to connect these dead end metabolites, a universal database of metabolic reactions was used to predict the fewest reactions required to fully connect all metabolites in the network. Focusing on gluconate, which is a disconnected metabolite, the authors experimentally characterized (C9orf103), previously identified as a candidate tumor suppressor gene, as the gene that encodes gluconokinase, thereby consuming this metabolite and connecting it to the rest of the human metabolic network.

E. coli

Gap-filling methods combined with systematic gene knockouts in E. coli [28], were used to discover new metabolic functions for the classic glycolytic enzymes phosphofructokinase and aldolase. Single, double, and triple knockout strains of central metabolic genes were grown on 13 different carbon sources. Concurrently, the same gene knockouts and growth conditions were simulated using the E. coli GEM. Several discrepancies between model predictions and experimental results were related to talAB interactions in the pentose phosphate pathway and could not be reconciled. A metabolomic analysis identified a new metabolite, sedoheptulose-1,7-bisphosphate, that had not been previously characterized. Using metabolic flux analysis and in vitro enzyme assays, the investigators confirmed that phosphofructokinase carries out the reaction and that glycolytic aldolase can split the seven-carbon sugar into three- and four-carbon sugars, glyceraldehyde-3-phosphate (G3P) and D-erythrose 4-phosphate (E4P) respectively.

Yeast

An analysis of synthetic lethal screens and gap-filling methods were used to correct incorrect pathways leading to NAD+ synthesis in yeast [30]. The study compared an experimental set of genetic interactions for metabolic genes against interactions that

were predicted by FBA. Using machine-learning techniques, key changes to the metabolic network that improved model accuracy were identified. Model refinement identified one of the two NAD+ biosynthetic pathways from amino acids in the GEM as a source of inaccurate predictions. Using growth screens with mutant strains, the authors validated that the synthesis of NAD+ from amino acids was only possible from L-tryptophan (L-trp) but not from L-aspartate (L-asp).

1.5.4 Adaptive laboratory evolution in the discovery process.

In contrast to false negatives, false positives arise when the model predicts growth, but experiments show no growth (Figure 1.3D). False positives occur in cases where experimental data show a particular gene to be essential but model simulations do not. Metabolic models can be used to predict efficient compensatory pathways, after which cloning and overexpression of these pathways are performed to investigate whether they restore growth and to help determine why these compensatory pathways are not active in mutant cells.

Discovering context-specific regulatory interactions using false positive predictions

Cloning and overexpression of a false positive associated gene has been demonstrated for a ppc knockout of Salmonella enterica serovar Typhimurium [39]. A metabolic model of S. Typhimurium predicted that the cells could route flux through the glyoxylate shunt when ppc is removed due to the backup function of isocitrate lyase encoded by aceA. However, the ppc cells were nonviable experimentally. The protein IclR is a transcription factor that regulates the transcription of genes involved in the glyoxylate shunt, including aceA. Therefore a dual knockout ppciclR mutant was constructed. Growth was restored in this double mutant at 60

Adaptive laboratory evolution can also be used to reconcile false positive predic-

tions. Often, cell populations may need time to adapt to a genetic change or shift in media conditions, giving them the appearance of slow or no growth, despite a model prediction of growth. However, it has been shown that incorrect predictions of in silico models based on optimal performance criteria may be incorrect due to incomplete adaptive laboratory evolution under the conditions examined. It has been shown that E. coli K-12 grown on glycerol over 40 days (or about 700 generations) and subjected to a growth rate selection pressure (passing a small fraction of the fastest growers) achieves a final growth rate that is predicted by the GEM [33]. The quantitative prediction of growth rates is discussed in section 4. Thus, a false positive result may indicate that the model is in fact correct and a researcher should be patient while the cell adapts to achieve the model-predicted growth.

1.5.5 Recapitulation

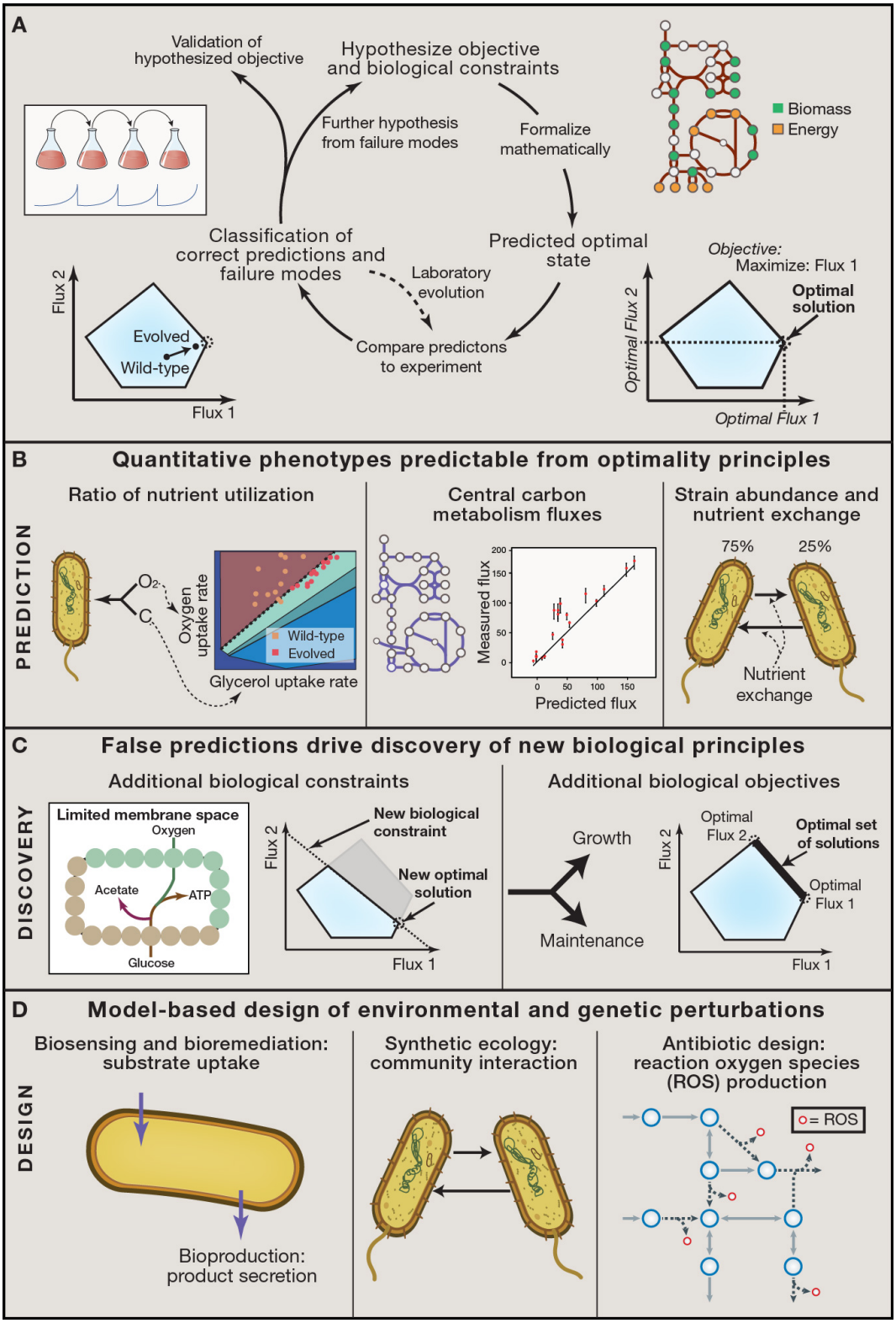
Given that our knowledge of any target organism is incomplete, its network reconstruction will also be incomplete. Thus, failures in GEM prediction of qualitative outcomes of growth capability are informative about the completeness of a network reconstruction and the consistency of its content. Furthermore, these approaches can be extended beyond model improvement. As genome editing techniques improve, in silico prediction of the effect of multiple gene-knockouts will be vital for contextualizing results of knockout studies and engineering genomes to achieve a desired phenotype [40]. Additionally, reconciliation of model false negatives have been used to explore the role that underground metabolism plays in adapting to alternate nutrient environments [41]. The algorithmic procedures that have been developed to address failure of prediction have led to some computer-generated hypotheses resulting in productive experimental undertaking. Further reading about the gap-filling process and algorithms for its implementation are available in Supplemental Table 1 of [17].

1.6 Quantitative phenotype prediction through optimality principles

The previous section treated qualitative predictions that relate to the presence or absence of parts from a reconstruction. Quantitative predictions of phenotypic functions are more challenging, but possible. The ability to compute quantitative organism functions from a genome-scale model represents a grand challenge in systems biology. Quantitative predictions are achievable with GEMs (even if they are based on incomplete reconstructions) by deploying cellular optimality principles. Evolutionary arguments underlie the deployment of optimality-based hypotheses. Phenotypes maximizing a hypothesized fitness function (as represented by an objective function) can be computed with constrained-optimization methods [19].

As for qualitative binary predictions of possible growth states, incorrect quantitative predictions often lead to new biological hypotheses and understanding. However, the discoveries arising from quantitative phenotype predictions are typically of a different nature than qualitative predictions. Rather than relating to missing reconstruction content (Section 3), the discoveries from quantitative phenotype prediction often relate to broad, fundamental organismal constraints [42, 43] and evolutionary objectives and trade-offs [44]. Quantitative phenotype prediction has also proven to be a useful capability for bioengineering applications. By optimizing an engineering (instead of evolutionary) objective, the best possible performance of an engineered biological system can be determined. Furthermore, the specific flux states needed to achieve high performance can guide engineering design [45].

Figure 1.4: **Quantitative phenotype prediction using optimization** A. Quantitative phenotype prediction is an iterative workflow. First, hypothesized biological constraints and objectives are formulated mathematically, and computational optimization is used to determine optimal phenotypic states (see Section 2). The predicted phenotypic states can then be compared to experimental measurements to identify where predictions are consistent. When consistent, the hypothesized evolutionary objective and constraints are validated. When inconsistent, laboratory evolution can be used to gain further insight as to why the computed and measured states differ. Examples of validation of quantitative phenotypes are detailed in 4B and further hypotheses derived from incorrect predictions are detailed in 4C. B. The generic workflow in 4A has been successfully applied to several classes of phenotypes. i) Nutrient utilization ratios can be predicted by maximizing biomass flux [46]. ii) Central carbon metabolism fluxes can be predicted; for some organisms, much of the variability in flux can be attributed to biomass flux maximization [47]. iii) The ratio of organism abundances and nutrient exchanges can be predicted for both natural and synthetic communities. Note that one important feature of quantitative phenotype predictions is that optimal flux solutions are often not unique. To address this, flux variability analysis (FVA) [48] can be used to identify the ranges of possible fluxes. It should be noted that non-uniqueness is not necessarily a handicap of COBRA as biological evolution can come up with alternate solutions [49]. C. Inconsistencies with model predictions have led to the appreciation of new constraints and objectives underlying cellular phenotypes. i) Inconsistent predictions in by-product secretion have led to the hypothesis that membrane space limits membrane protein abundance and metabolic flux [43]. ii) The range of metabolic fluxes observed across different environments have led to the realization that fluxes can be understood as simultaneously satisfying multiple competing objectives, such as growth and cellular maintenance. Multi-objective optimization algorithms find solutions that maximize multiple competing objectives. D. Accurate prediction of quantitative phenotypes has led to prospective design of biological functions. A number of algorithms have been developed that predict genetic and/or environmental perturbations required to achieve a bioengineering objective. Relevant bioengineering objectives have included biosensing, bioremediation, bioproduction, the creation of synthetic ecologies, and the intracellular production of reaction oxygen species (ROS) to potentiate antibiotic effects.


A
Validation of hypothesized objective
Hypothesize objective and biological constraints
Further hypothesis from failure modes
Formalize mathematically
Biomass
Energy
Classification of correct predictions and failure modes
Laboratory evolution
Predicted optimal state
Compare predictons to experiment
Flux 2
Evolved
Wild-type
Flux 1
Objective:
Maximize: Flux 1
Optimal solution
Optimal Flux 2
Optimal Flux 1
B
Quantitative phenotypes predictable from optimality principles
PREDICTION
Ratio of nutrient utilization
O2
C
Oxygen uptake rate
Wild-type
Evolved
Glycerol uptake rate
Central carbon metabolism fluxes
Measured flux
Predicted flux
Strain abundance and nutrient exchange
75%
25%
Nutrient exchange
C
False predictions drive discovery of new biological principles
DISCOVERY
Additional biological constraints
Limited membrane space
Oxygen
Acetate
ATP
Glucose
New biological constraint
New optimal solution
Flux 2
Flux 1
Additional biological objectives
Growth
Maintenance
Optimal Flux 2
Optimal set of solutions
Optimal Flux 1
Flux 2
Flux 1
D
Model-based design of environmental and genetic perturbations
DESIGN
Biosensing and bioremediation: substrate uptake
Bioproduction: product secretion
Synthetic ecology: community interaction
Antibiotic design: reaction oxygen species (ROS) production
= ROS

1.6.1 Workflow for quantitative phenotype prediction

Quantitative phenotypes can be predicted through the same computational procedures used for qualitative growth predictions (Figure 1.4A). An objective (either evolutionary or engineering) is assumed, and maximized computationally (subject to flux balance and other constraints). The flux state(s) that maximize the objective are then the predicted quantitative fluxes. These predictions can then be compared to experimental measurements. In cases of agreement, the evolutionary hypothesis is supported. In cases of a disagreement between experimental and theoretical predictions, either the biological system has not been exposed to the selection pressure to reach the theoretical optimum (i.e., the assumed evolutionary objective is incorrect or partially correct), or there are missing biological constraints that affect the theoretical predictions (i.e., the relevant biological constraints are incomplete). Experimental evolution can discriminate these alternatives [33, 47] by exposing the biological system to the appropriate selection pressure, leading it to evolve towards the stated optimum. For example, in one study, strains carrying deletions of one of six metabolic genes were evolved on four different carbon sources. A total of 78

1.6.2 Flux variability analysis (FVA) calculates possible flux states

Flux balance analysis computes an optimal objective value and a flux state that is consistent with that objective (and all of the imposed constraints). While the objective value is unique, multiple flux states can typically support the same objective value in genome-scale models. For this reason, flux variability analysis (FVA) is used to determine the possible ranges for each reaction flux [48]. With FVA, the objective value is set to be equal to its maximum value, and each reaction is maximized and minimized. For some fluxes, their maximum value will be equal to their minimum, enabling a

specific prediction. For others, there may be a wide range of possible values due to alternative pathways. Often, a parsimonious flux state is also assumed and computed with parsimonious-FBA (pFBA) [50]. With pFBA, the sum of fluxes across the entire network is minimized (again, subject to the optimal objective value determined); pFBA will eliminate some alternative pathways. Typically, many reaction fluxes can by uniquely predicted with optimality and parsimony assumptions. Additional biological constraints in next-generation models (Section 6) reduce the possible flux states further [51].

1.6.3 Types of possible (evolutionarily optimal) quantitative predictions

The simplest type of quantitative phenotype predictable with constraint-based models is nutrient utilization. While metabolic models do not predict absolute rates of nutrient uptake, they predict the optimal ratios at which nutrients are utilized. For example, metabolic models predict an optimal oxygen uptake rate relative to the carbon source uptake rate (resulting in a predicted optimal ratio between the two nutrients). In an early study, the ratios of oxygen and carbon uptake were shown to be predictable for a number of carbon sources in E. coli [46]. In a later study, E. coli was evolved in the laboratory on a carbon source (glycerol) for which the wild-type strain did not match the predicted nutrient utilization; after evolution, the strain exhibited the optimal uptake rates predicted theoretically (Figure 1.4B) [33]. Comparison of experimental and predicted phenotypes therefore reveals the environments to which an organism has been evolutionary exposed.

Metabolic fluxes for central carbon metabolism can be estimated with 13C carbon labeling experiments, making them candidates for quantitative prediction (Figure 1.4B). Since the dimensionality of carbon labeling data is larger than that for nutrient uptake, there is more opportunity to dissect the differences in computed and measured

fluxes to better understand the multiple objectives and constraints underlying microbial metabolism. Impressively, the biomass objective function can explain a large amount of the variability of fluxes [52]. Failure modes in prediction have led to the appreciation of the importance of protein cost [53], and membrane [43] and cytoplasmic spatial constraints [42], which affect the optimal flux state (Figure 1.4C). Furthermore, failure modes have led to the understanding that metabolism is simultaneously subject to multiple competing evolutionary objectives, resulting in trade-offs (e.g., growth versus maintenance) employed by different species (Figure 1.4C). In this way, outliers in quantitative predictions can improve the understanding of constraints and objectives underlying a particular organisms metabolism. Optimality principles from stoichiometric models have also been expanded from single populations of cells to microbial communities. To model microbial communities, multiple species are linked together through the exchange of nutrients extra-cellularly [54] or through direct electron transfer [55]. The secretion rate from one species limits the uptake rate for others, resulting in balanced species interactions. For a number of cases of communities composed of two or three members, the optimal rate of nutrient exchange and the ratio of the species in the population [56] can be predicted. The effects of spatial organization of community members are also being uncovered [57]. The constraints on nutrient flow between organisms (e.g., diffusion) have proven to be important for predicting community composition and behavior, highlighting the importance of abiotic constraints and community structure in the behavior of biological communities.

Evolution is a natural counterpart to optimality-based predictions with constraint-based methods. Constraint-based optimality predictions have focused on predicting the endpoints of short-term experimental evolution. However, this scope of application has increased in recent years to study long-term phenotypic and enzyme evolution [58, 59].

1.6.4 From optimality principles to prospective design

Quantitative phenotype prediction via optimization is also commonly used for bioengineering applications (Figure 1.4D). For example, in metabolic engineering, optimal pathway yields are used to prioritize pathways to be built into a production strain and to benchmark their performance. Furthermore, the flux states required to achieve these optima (and how they differ from wild-type growth states) can guide strain design [60].

A number of design algorithms have been built to work with metabolic models and predict the genetic and environmental modifications to increase performance [61, 62]. While many design algorithms and applications have been focused on metabolite production (e.g., for production of fuels and chemicals), metabolic models have also been utilized for the design of biosensors [63] and biodegradation [64, 65]. Also, design has expanded beyond single populations to microbial communities/ecosystems [66].

1.6.5 Recapitulation

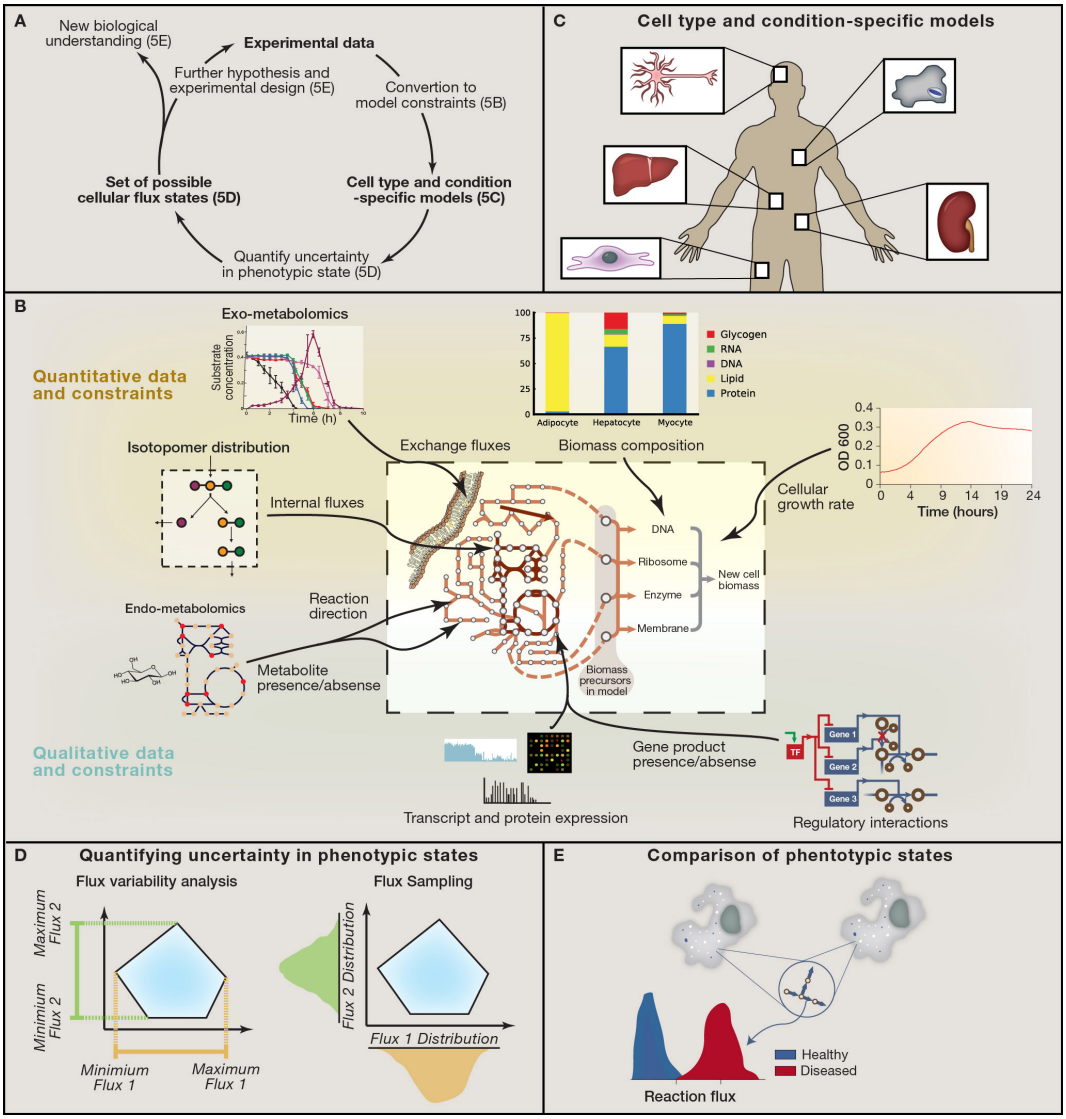
Quantitative phenotype predictions initially focused on simple physiological predictions and are still expanding to more complex phenotypes, biological systems [67], and environments. Although there have been notable successes of quantitative phenotype prediction, certain phenotypes are still difficult to predict. Historically, difficult predictions have led to the development of new computational methods and an appreciation of new biological constraints. Supplemental Table 2 (Evaluation of Model Capabilities) in [17] summarizes several types of predictions and the approximate performance of constraint-based methods utilized to date. The expansion in the scope and accuracy of predictions continues today with models of increased scope [53, 68], discussed in section 6. Thus far, quantitative phenotypes have been limited primarily to microbial systems and, more recently, plants [69, 70]. For multi-cellular organisms, specialized

cell types support the fitness of the entire organism. Cell-type specific objectives have been constructed [71], though they typically are used for qualitative (Section 3) instead of quantitative phenotype prediction. Instead, quantitative phenotypes in multi-cellular organisms are typically determined through model-driven analysis of experimental data, discussed in Section 5.

1.7 Multi-omic data integration: constraining and exploring possible phenotypic states

With the expanding quantity of omics and other phenotypic data, there is an increasing need to integrate these datasets to drive further understanding and hypothesis generation. Phenotypic data types can be integrated with metabolic GEMs to determine condition-specific capabilities and flux states in the absence of assumed objectives (Section 4). Computational methods that identify the possible range of phenotypic states given the measured data allow one to quantify the degree of (un)certainty in metabolic fluxes. Some types of data are quantitative and directly indicative of metabolic fluxes, whereas other data are qualitative or indirectly related to metabolic fluxes. By layering different data types, the true state of a biological system can be determined with increased precision. The need for formal integration of disparate data types represents a grand challenge that has been termed Big Data to Knowledge (BD2K, bd2k.nih.gov).

Figure 1.5: **Data integration and exploration of possible cellular phenotypes**A. The general workflow for multi-omic data integration begins with the conversion of the experimental data into model constraints (see Figure 1.5B). This procedure results in cell-type (e.g. neuron, macrophage) and condition-specific (e.g. healthy vs. diseased) models that represent the metabolic capabilities of those specific cells (see Figure 1.5C). Several computational procedures can then be used to explore the metabolic capabilities and determine achievable phenotypes systematically (see Figure 1.5D). Evaluation of these phenotypic capabilities and comparison of different cells or environments leads to identification of their molecular differences (see Figure 1.5E). Additionally, if the original experimental data cannot precisely distinguish between certain metabolic states, additional targeted experiments can be designed and integrated as further constraints. B. Numerous data types can be integrated into metabolic models. Some directly affects model structure and variables (e.g. growth rate, biomass composition, exchange fluxes, internal fluxes and reaction directionality). Standard processing of these data types allows for integration into the model. Other data types affect metabolic fluxes more indirectly. As such, different computational methods exist for formulating the appropriate constraints (Table 1). C. Experimental data is integrated to construct cell-type and/or condition-specific models. These models represent the metabolic capabilities in a certain state, and are then used for further inquiry (see Figures 5D,E). Specific algorithms for building cell-type specific models from gene expression data include MBA [72] and GIMME [73]. D. After adding constraints to the model, computational procedures are used to assess the implication of the experimental data on metabolic fluxes. The two main methods for querying the consequences of the measured data on a cells phenotypes are flux variability analysis (FVA) and Markov-chain Monte-Carlo (MCMC) sampling. i) FVA determines the maximum and minimum values of all metabolic fluxes. ii) MCMC sampling randomly samples feasible metabolic flux vectors (usually resulting in tens to hundreds of thousands of flux vectors). These sampled flux vectors can then be used to derive the distribution of possible flux values for a given metabolic reaction. E. Often a comparative approach is employed in which experimental data from two conditions are used to generate two condition-specific models. Then, the achievable phenotypes of the two states are compared (e.g. though MCMC sampling, see Figure 1.5D).


A
New biological understanding (5E)
Experimental data
Further hypothesis and experimental design (5E)
Convertion to model constraints (5B)
Set of possible cellular flux states (5D)
Cell type and condition -specific models (5C)
Quantify uncertainty in phenotypic state (5D)
C
Cell type and condition-specific models
B
Exo-metabolomics
Substrate concentration
Time (h)
Quantitative data and constraints
Isotopomer distribution
Internal fluxes
Endo-metabolomics
Reaction direction
Metabolite presence/absense
Exchange fluxes
Biomass composition
Glycogen
RNA
DNA
Lipid
Protein
Adipocyte
Hepatocyte
Myocyte
OD 600
Time (hours)
Cellular growth rate
DNA
Ribosome
Enzyme
Membrane
New cell biomass
Biomass precursors in model
Qualitative data and constraints
Gene product presence/absense
Transcript and protein expression
TF
Gene 1
Gene 2
Gene 3
Regulatory interactions
D
Quantifying uncertainty in phenotypic states
Flux variability analysis
Maximum Flux 2
Minimium Flux 2
Minimium Flux 1
Maximum Flux 1
Flux Sampling
Flux 2 Distribution
Flux 1 Distribution
E
Comparison of phentotypic states
Healthy
Diseased
Reaction flux

1.7.1 Workflow for multi-omic data integration

The overall procedure for multi-omic integration with genome-scale models is an iterative workflow (Figure 1.5A). Once experimental data from the particular biological system under study is obtained, it is converted into constraints on model function (Figure 1.5B). The successive application of experimentally derived constraints to the reaction network results in the generation of a cell-type and condition-specific model (Figure 1.5C). Several computational procedures can then be used to explore the metabolic capabilities and achievable phenotypes of the experimentally constrained model (Figure 1.5D). Evaluation of these phenotypic capabilities and comparison of different cells or environments leads to identification of their molecular differences (Figure 1.5E), providing biological insight and driving further hypotheses.

1.7.2 Converting data to model constraints

Successive imposition of constraints is a basic principle of COBRA [74]. Some data types can be directly converted into constraints on model variables. Biomass composition and growth rate affect the metabolic demands of cellular growth [20]. Time-course exo-metabolomics can be used to set the uptake and secretion rates of nutrients [75]. Intracellular quantitative metabolomics combined with reaction free energies can discern condition-specific reaction directionalities [76]. Isotopomer distributions from cellular biomass or metabolite pools can be used to infer and constrain intracellular fluxes [77]. These data can be used separately or combined to identify with increasing precision the true state of the cell.

Other data types affect metabolism more qualitatively. In theory, quantitative metabolite, transcript, and protein levels can be used to constrain metabolism quantitatively, but in practice this requires many parameters that are hard-to-measure and are

organism-specific. Instead, these data types can be used as qualitative constraints relating to gene product or metabolite presence/absence; that is, if a metabolite is present, a reaction must be active that produces it [78], and if a gene product is absent, its catalyzed reactions cannot carry flux [72, 79]. Similarly, regulatory interactions can be added to affect the presence/absence of a gene product based on condition-specific activity of a transcription factor [80].

1.7.3 Cell-type and condition-specific models

Starting from a large reconstructed reaction network (e.g., representing all metabolic reactions encoded in the human genome [81]), the imposition of experimental data results in the generation of cell-type and condition-specific models. Experimentally derived constraints pare down the achievable phenotypes from those encoded by the totality of the cells genome. By eliminating phenotypes that cannot be achieved, this new model represents the capabilities of the particular cell-type and environment assayed. This model summarizes the experimental data in a self-consistent and integrated format, and forms the starting point for further computational and biological inquiry [78, 82] (see Figure 1.5D,E).

1.7.4 Quantifying uncertainty

Once a cell-type and condition-specific model is created, computational methods are used to determine the possible flux states of the cell. Flux variability analysis (FVA, which is described in section 4) [48] can be used to determine the range of fluxes that are consistent with the experimental data. A more refined approach is flux sampling [83] (typically with Markov Chain Monte Carlo, MCMC, methods), which determines the distribution of fluxes for all reactions (instead of simply the range). When no cellular

objective is assumed, the feasible flux space is very unconstrained and a particular reaction could be operating at nearly any flux value. As more data is layered, the feasible flux space decreases. When no objective is assumed, fluxes are rarely precisely known, and many will remain completely unknown. However, an imprecisely known flux space is often sufficient to discern differences between two environments/states as discussed in the following subsection.

1.7.5 Using computed states to drive discovery

Once the range of possible phenotypic states is quantified, they must be analyzed to gain biological insights. Often a comparative approach is employed, in which two experimental states (e.g., neurons from Alzheimers disease patients compared to healthy controls [84]) are compared. Reactions that have a non-overlapping FVA range must be different between the two states, and can be indicative of important metabolic changes. In cases where the FVA ranges are overlapping, the flux distributions from MCMC sampling can still be different that is, the reactions are likely different between the two states, but the current experimental data is insufficient to guarantee it.

Pathway visualization is also helpful in gaining insight into changes in cell statesfluxes (or flux ranges) are most comprehensible in a network context. A few tools exist for the visualization of metabolic fluxes; some are based on static maps [10], whereas others create auto-generated layouts and new tools allow for the drawing of maps based on flux solutions [85]. Finally, identifying reactions or subsystems that remain partially identified (e.g., based on a large FVA range) can guide further experimentation, resulting in an iterative computational and experimental elucidation of a cells state.

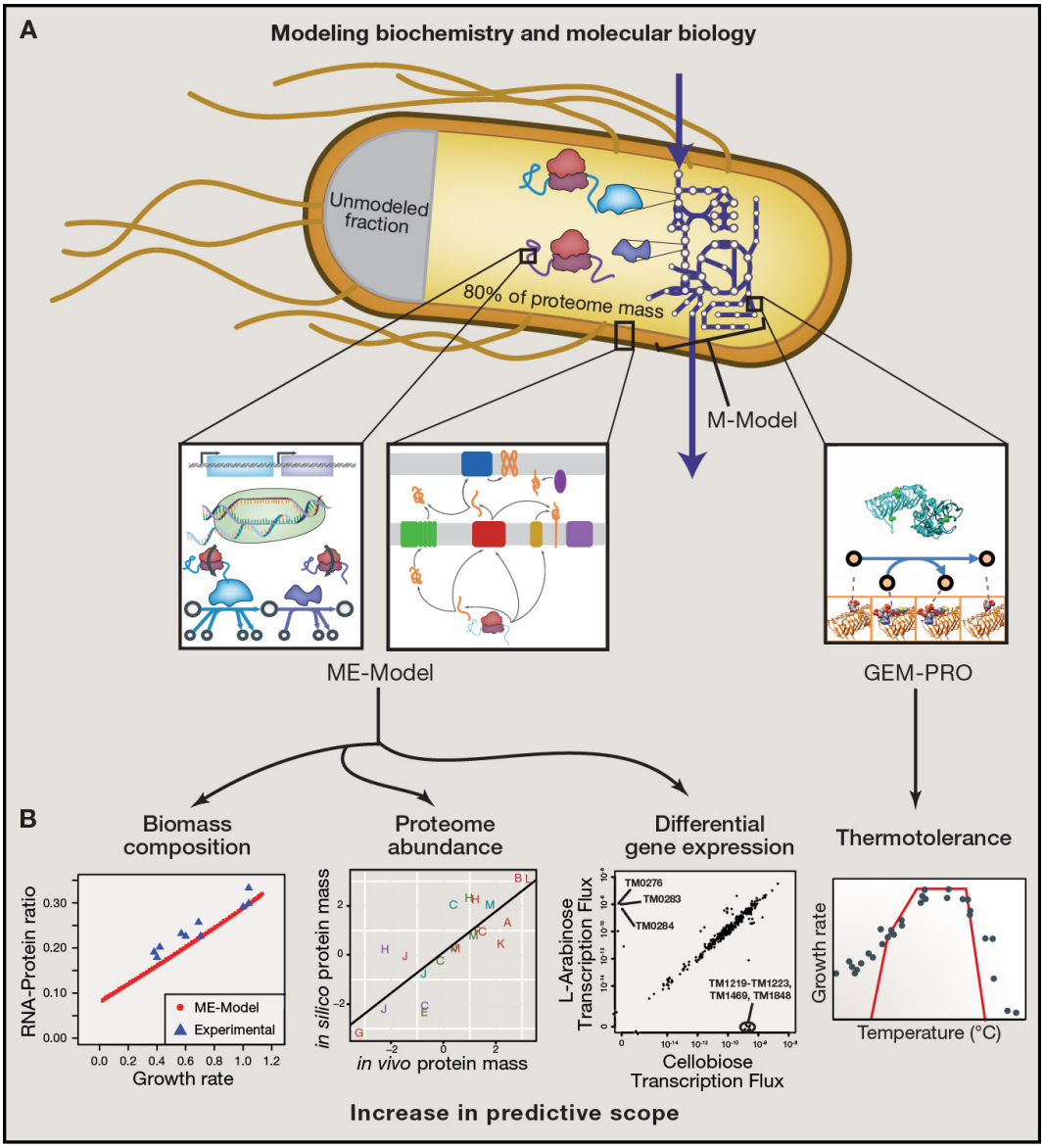
1.7.6 Recapitulation

GEMs can be used to integrate numerous data types. In fact, as more experimentally derived constraints are successively imposed, analysis often becomes easier (as the range of possible solutions shrinks [18]), instead of more challenging as often occurs with statistically based data integration procedures. A current challenge with metabolic GEMs is the explicit integration of data types that do not directly reflect metabolic fluxes (e.g., transcriptomics, proteomics, and regulatory interactions). This challenge is primarily due to the fact that these processes are not explicitly described in metabolic models. Expansions of metabolic models to encompass gene expression hold promise to address this challenge and are discussed in section 6.

1.8 Moving beyond metabolism to molecular biology

Up to this point, this Primer has focused on metabolic models, or M-Models. M-models have reached a high degree of sophistication after 15 years of development, resulting in standard operating procedures for their construction [14] and use [5]. However, M-Models are limited in their explicit coverage to metabolic fluxes. Thus, a grand challenge in the field has been to expand the concepts of constraint-based models of metabolism to other cellular processes to formally include more disparate data types in genome-scale models [86].

Figure 1.6: **Expansion of genome-scale models to encompass molecular biology** A. Metabolic models have been expanded to encompass the processes of proteome synthesis and localization as well as data on protein structures. Models including protein synthesis and localization are referred to as ME-Models, which stands for metabolism and gene expression. GEM-PRO refers to genome-scale models integrated with protein structures. For GEM-PRO, a combination of structural data directly references the GPRs in the metabolic reconstruction; structures can be obtained from experimental databases or homology modeling. The E. coli ME-Model mechanistically accounts for 80B. Addition of cellular processes vastly increases the predictive scope of models. ME-Models can predict biomass composition, abundances of protein across subsystems, and differential gene expression in certain environmental shifts (in addition to the predictions possible with M-Models); like FBA these were predicted by assuming growth maximization as an evolutionary objective, though the specific optimization algorithm differs due to the addition of coupling constraints. GEM-PRO has been used to predict the metabolic bottlenecks and growth defects of changes in temperature on protein stability and catalysis; protein stability is predicted with structural bioinformatics methods and then used to limit the catalyzed metabolic flux. The uses of these integrated models are just beginning to be explored.


A
Modeling biochemistry and molecular biology
Unmodeled fraction
80% of proteome mass
M-Model
ME-Model
GEM-PRO
B
Biomass composition
Proteome abundance
Differential gene expression
Thermotolerance
RNA-Protein ratio
Growth rate
ME-Model
Experimental
in silico protein mass
in vivo protein mass
L-Arabinose Transcription Flux
Cellobiose Transcription Flux
TM0276
TM0283
TM0284
TM1219-TM1223, TM1469, TM1848
Growth rate
Temperature (°C)
Increase in predictive scope

1.8.1 Computing properties of the proteome

The process of addressing this grand challenge has begun (Figure 1.6A). Recently, genome-scale network reconstructions have expanded to encompass aspects of molecular biology. Two significant expansions are genome-scale models integrated with protein structures, GEM-PRO, and integrated models of metabolism and protein expression, ME-Models. GEM-PRO allows for structural bioinformatics analysis to be performed from a systems-level perspective, and have those results in turn affect network simulations. ME-Models allow for the simulation of proteome synthesis, and account for the capacity and metabolic requirements of gene expression.

1.8.2 A structural biology view of cellular networks

GEM-PRO reconstructions can have varying degrees of detail, which affects the types of analysis possible. So far, GEM-PRO reconstructions have been created for T. maritima [87] and E. coli [68, 88]. Initial reconstructions have focused on single peptide chains [87], and utilized homology modeling to fill in gaps where organism-specific structures have not been identified. Further reconstruction detail has included protein-ligand complexes [68] and quaternary protein assemblies [88]. To link the structures to the metabolic model, structural data directly references the GPRs in the metabolic reconstruction. For cases of protein-metabolite complexes, the metabolites also need to be properly annotated in the structural data. The structural reconstruction therefore provides a physical embodiment of the gene-protein-reaction relationship.

There are a few notable cases demonstrating the unique analysis possible with the combination of protein structures and network models. In T. maritima, network context and protein fold annotations were combined to test alternative models for pathway evolution [87]. The T. maritima GEM-PRO supported the patchwork model for genesis

of new metabolic pathways. In E. coli, the effect of temperature on protein stability and enzyme activity was simulated at the systems level, recapitulating the effects of temperature on growth [68]. Also in E. coli, protein-ligand interactions were combined with gene essentiality predictions to discover new antibiotic leads and off-targets [88]. These examples just scratch the surface of analyses made possible with the integration of network and structural biology.

1.8.3 Modeling molecular biology and metabolism

ME-Models formalize all of the requirements for biosynthesis of the functional proteome (Figure 1.6B). They compute the proteome composition and its integrated function to produce phenotypic states and all the metabolic processes needed for its synthesis. This represents an integrated view of metabolic biochemistry and the core processes of molecular biology. As with GEM-PRO, the first ME-Models were formulated for T. maritima [51] and E. coli [53, 89].

The reconstruction of a ME-Model starts with the formation of reactions for gene expression and enzyme synthesis [90]. The processes explicitly accounted for in ME-Models are very detailed, including transcription units and initiation and termination factors for transcription, tRNAs and chaperones needed for translation and protein folding, and metal ion and prosthetic group requirements for catalysis. In other words, the reconstructions strive to match as closely as possible all the biochemical processes required to synthesize fully functional enzymes. To create a ME-Model, the reactions for enzyme synthesis are coupled to the totality of metabolic reactions with pseudo-kinetic constraints, termed coupling constraints [91, 51]. These constraints relate the abundance of an enzyme (or any recyclable chemical species, e.g., mRNA, tRNA), to its degradation rate and catalytic capacity.

ME-Models thus significantly expand the scope of phenotype predictions possible

to include aspects of transcription and translation. RNA and protein biomass composition are variables in ME-Models, and are no longer set a priori (as in the biomass objective function of M-Models). ME-Models predict the experimentally observed linear changes in the ratio of RNA-to-protein mass fractions as a consequence of changes in protein synthesis demands [53]. Furthermore, the mass fractions of protein subsystems agree well with those predicted by the ME-Model. This shows that the broad distribution of protein subsystem abundance is predictable using optimality principles and the comparison reveals that some subsystems were under-predicted, thus identifying them as gaps in knowledge and targets for further reconstruction and model refinement [92]. While the quantitative prediction of individual protein abundances is currently out of scope of the ME-Model (as these demands depend on enzyme-specific kinetics) the ME-Model has been shown to accurately predict differential expression across certain environmental shifts, due to the differential requirements of proteins across conditions (a more qualitative than quantitative prediction) [51].

A recent expansion to the ME-Model includes the addition of protein translocation, allowing for the localization of protein to be computed [92] (i.e., into cytoplasm, periplasm, inner and outer membrane). Translocase abundances and compartmentalized proteome mass was accurately predicted from the bottom-up based on optimality principles. Addition of compartmentalization also allows for membrane area and cytoplasmic volume constraints to be formalized, which, if combined with GEM-PRO, approaches a digital embodiment of a three-dimensional cell.

1.8.4 Recapitulation

Metabolic models are limited in their predictive ability dictated by the scope of the reconstruction. Nearly all of the predictions of metabolic models outlined in the previous sections can be refined and expanded with GEM-PRO or ME-Models. Advances

to include protein structures and protein synthesis open new vistas for constraint-based modeling.

The scope of genetic perturbations (Section 2) that can be simulated is significantly larger due to the inclusion of genes for gene expression (and accounting for protein cost) and the effects of coding mutations on protein structures; GEM-PRO also expands the scope of environmental perturbation to enable simulation of changes in temperature. GEM-PRO allows for new gap-filling approaches (Section 3) based on structural bioinformatics methods. ME-Models expand the scope of quantitative molecular phenotypes to include transcript and protein levels (Section 4), and transcriptomics and proteomics can be analyzed in mechanistic detail (Section 5).

With the added capabilities of GEM-PRO and ME-Models also come additional computational challenges. While single optimization calculations with M-Models take less than a second on a modest laptop computer, growth-maximization with a ME-Model can take over an hour. The ME-Model also requires specialized high-precision solvers. Many promising applications of GEM-PRO will require simulation of protein dynamics with molecular dynamics (MD) and hybrid quantum mechanics/molecular mechanics (QM/MM) simulations on protein structures. High-performance computing environments are required for such simulations, and there is a pervasive trade-off between the precision of simulations and the scope of structural coverage. However, advances in high-precision solvers for ME-Models [93] and structural simulations for GEM-PRO are rapid and are likely to ameliorate these challenges. Like discoveries enabled by comparing M-Model predictions to experimental data, we anticipate much biology can be learned from comparing in silico and in vivo proteome allocation [94], leading to increasingly predictive models. The E. coli ME-Model currently encompasses many key cellular functions, covering 80

1.9 Perspective

Genome-scale models have been under development since the first annotated genome-sequences appeared in the mid to late 1990s. For most of this history, the focus of GEMs has been on metabolism. After initial successes with metabolic GEMs it became clear that the same approach could be applied to other cellular process that could be reconstructed in biochemically accurate details. Thus, a vision was laid out in 2003 that the path to whole cell models was conceptually possible and that such models could be used as a context for mechanistically integrating disparate omic data types [86]. This vision is now being realized. This Primer shows how six grand challenges in cell and molecular and systems biology can be addressed using GEMs. A surprising range of cellular functions and phenotypic states can be now dealt with.

We now have the tools at hand to develop quantitative genotype-phenotype relationships from first principles and at the genome-scale. Current models of prokaryotes account for metabolism, transcription, translation, protein localization, and protein structure. Processes not described in the current ME models will be systematically reconstructed over the coming years to gain a more and more comprehensive description of cellular functions. Biology can thus look forward to the continued development and use of a mechanistic framework for the study of biological phenomena as physics and chemistry have enjoyed for over a century.

1.10 Acknowledgements

EJO, JMM, and BOP conceived, wrote, and edited the manuscript. This work was supported by National Institutes of Health grant no. R01 GM057089.

The text of Chapter 1 is a full reprint of the material as it appears in: OBrien E.J.*, Monk J.A.*, Palsson B.O. Using genome-scale models to predict biological capabilities,

Cell, 161(5):971-987. (2015). * indicates equal contribution. The dissertation author was the primary author of the manuscript. The other authors were Jon A. Monk (equal contributor) and Bernard Ø. Palsson.

Chapter 2

Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction

All models are wrong, but some are useful.
—George E. P. Box

2.1 Abstract

Growth is a fundamental process of life. Growth requirements are well characterized experimentally for many microbes; however, we lack a unified model for cellular growth. Such a model must be predictive of events at the molecular scale and capable of explaining the high-level behavior of the cell as a whole. Here, we construct an ME-Model for *Escherichia coli*—a genome-scale model that seamlessly integrates metabolic and gene product expression pathways. The model computes ∼80% of the functional proteome (by mass), which is used by the cell to support growth under a given condition. Metabolism and gene expression are interdependent processes that affect and constrain

each other. We formalize these constraints and apply the principle of growth optimization to enable the accurate prediction of multi-scale phenotypes, ranging from coarse-grained (growth rate, nutrient uptake, by-product secretion) to fine-grained (metabolic fluxes, gene expression levels). Our results unify many existing principles developed to describe bacterial growth.

2.2 Introduction

The genotype-phenotype relationship is fundamental to biology. Historically, and still for most phenotypic traits, this relationship is described through qualitative arguments based on observations or through statistical correlations. Understanding the genotype-phenotype relationship demands vantage points at multiple scales, ranging from the molecular to the cellular. Reductionist approaches to biology have produced 'parts lists', and successfully identified key concepts (e.g., central dogma) and specific chemical interactions and transformations (e.g., metabolic reactions) fundamental to life. However, reductionist viewpoints, by definition, do not provide a coherent understanding of whole cell functions. For this reason, modeling whole biological systems (or subsystems) has received increased attention.

A number of modeling approaches have been developed to predict systems-level phenotypes. What distinguish these models from each other are the underlying assumptions they make, the input data they require, and the scope and precision of their predictions [95]. The type of modeling formalism employed is influenced by all of these distinguishing characteristics [96]. Genome-scale optimality models of metabolism (termed as M-Models) have made much progress in recent years as they require only basic knowledge of reaction stoichiometry, are genome-scale in scope, and have fairly accurate predictive power. Recently, M-Models have been extended to include the process of

gene expression (termed as ME-Models) [51, 89], opening up completely new vistas in the development of microbial systems biology. On the heels of these developments, a whole-cell model (WCM) of the human pathogen *Mycoplasma genitalium* appeared [97]. The WCM integrates many more cellular processes and can be used to simulate dynamic cellular states; however, it depends on detailed molecular measurements of an initial state (e.g., growth rate, biomass composition, and gene expression). While the model described by Karr et al is a major advance toward whole-cell computation, many practical applications rely on the ability to compute optimal phenotypic states. The WCM does not have this ability owing to the disparate mathematical formalisms it employs. The WCM and genome-scale optimality models thus have different capabilities and will find use to predict and explain different biological phenomena.

Here, we construct an ME-Model for *E. coli* K-12 MG1655. The ME-Model is a microbial growth model that computes the optimal cellular state for growth in a given steady-state environment. It takes as input the availability of nutrients to the cell and produces experimentally testable predictions for: (1) the cell's maximum growth rate (μ*) in the specified environment, (2) substrate uptake/by-product secretion rates at μ*, (3) metabolic fluxes at μ*, and (4) gene product expression levels at μ*. The creation of this model required the development of a new modeling formalism and optimization procedure to couple gene expression with metabolism, which provided new insight into growth rate- and nutrient limitation-dependent changes in enzymatic efficiency. The model predicts three distinct regions of microbial growth, defined by the factors (nutrient and/or proteome) limiting growth. We show that proteomic constraints improve predictions of metabolism itself, rectifying dominant failure modes in M-Models. Finally, we compute gene expression changes as the cell transitions through and between the different growth regions. The ME-Model computes measurable coarse- and fine-grained cellular and molecular phenotypes, and provides unity in the field by reconciling a variety

of principles related to cellular growth at various scales of complexity.

2.3 Results

2.3.1 Integration of genome-scale reaction networks of protein synthesis and metabolism

To create an ME-Model for *E. coli*, we started from two previous network reconstructions. The first reaction network includes all known metabolic pathways as of late 2011 [98] and is referred to as the M-Model throughout. The second accounts for reactions that describe gene expression and the synthesis of functional macromolecules in a mechanistically detailed manner [90]. The two reaction networks were integrated (see Materials and methods), and reactions and gene functions in both networks were updated to reflect gaps in knowledge that have been filled since their creation. We updated subunit stoichiometries for hundreds of multiprotein complexes and expanded the types of prosthetic groups, cofactors, and post-translational modifications required for catalytic activity (Materials and methods; Supplementary Table S1 in [53]). The scope and coverage of cellular processes in the integrated network is extensive. The integrated network mechanistically links the functions of 1541 unique protein-coding open reading frames (ORFs) and 109 RNA genes; it thus accounts for $\sim$35% of the 4420 protein-coding ORFs, $\sim$65% of the functionally well-annotated ORFs [99], and 53.7% of the non-coding RNA genes identified in *E. coli* K-12 [8]. In total, 1295 unique functional protein complexes are produced. Taken together, these complexes account for 80-90% of *E. coli*'s expressed proteome by mass (Supplementary Table S2 in [53]).

The integrated reaction network covers and accurately predicts a large proportion of essential cellular functions. It includes 223 of the 302 (73.8%) genes classified as

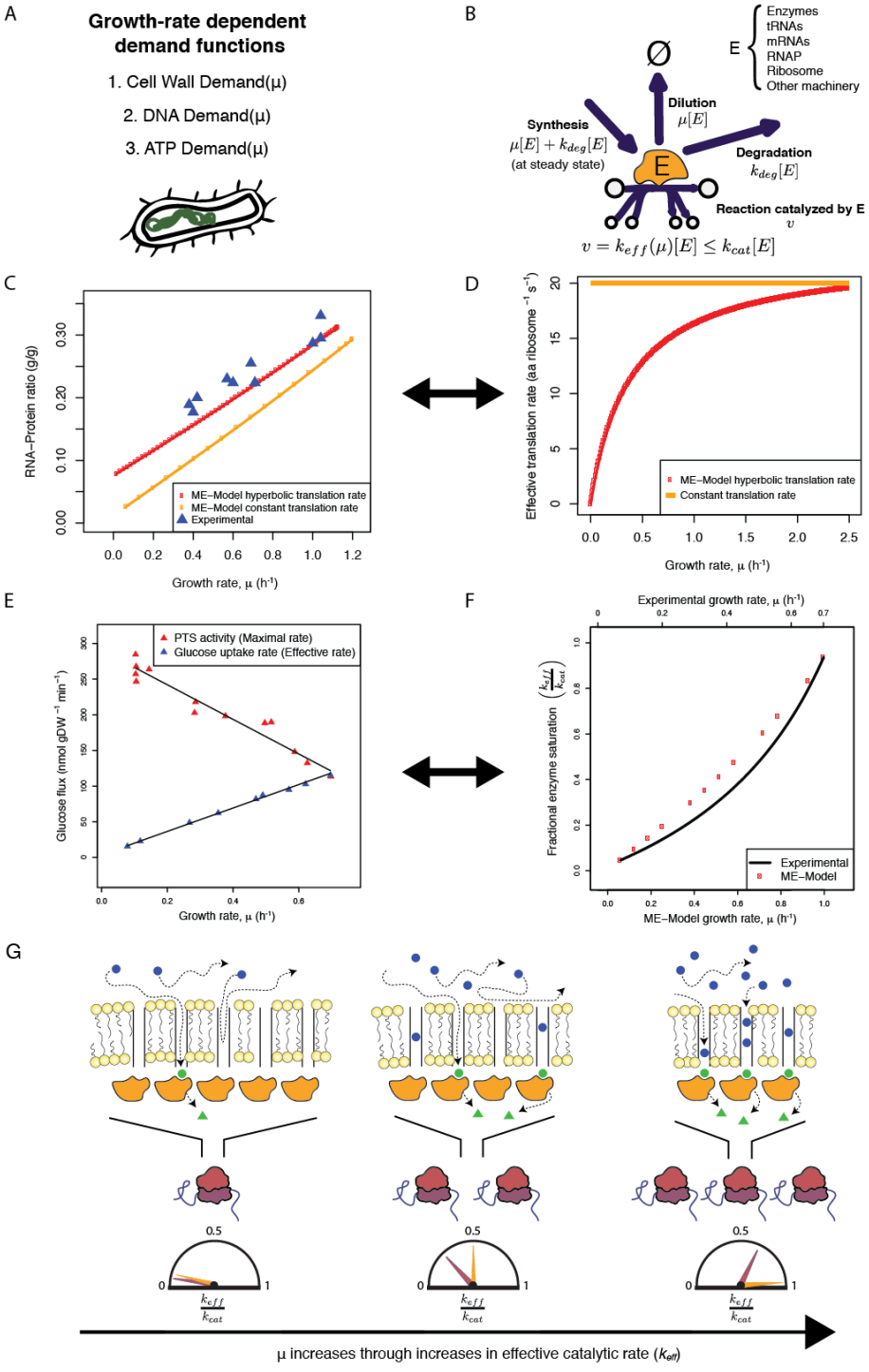
essential for cell growth under any condition [100] (Supplementary Table S3A in [53]), and 166 of the 206 functions (80.6%) estimated as essential for a minimal organism [101] (Supplementary Table S3B in [53]). *In silico* prediction of gene essentiality in glucose M9 minimal media results in an accuracy of 88.8% (precision=60.4%, recall=75%, Supplementary Table S4 in [53]). One of the dominant failure modes of essentiality predictions is due to the assumption that all tRNA and rRNA modifications are essential; removing these genes from predictions increases performance notably (accuracy=92.3%, precision=75.3%, recall=75%, Supplementary Table S4 in [53]). This accuracy is on par with previous approaches using the metabolic reaction network alone (accuracy=91.2%, precision=81%, recall=68%) [98]. Many of the key differences between the M-Model and the ME-Model essentiality predictions are due to the mechanistic treatment of cofactor and prosthetic group synthesis and utilization in the ME-Model. Specifically, for a protein complex to be functional in the ME-Model it has to contain the embedded prosthetic groups required for function; while this change in model structure results in some false predictions of essentiality compared with M-Models (which include all prosthetic groups in a biomass objective function that does not change across conditions), the essentiality predictions in the ME-Model can be directly related to the essential enzymes requiring the prosthetic group.

2.3.2 Growth demands and general constraints on molecular catalysis

To compute functional states of the integrated network, growth demands are first imposed. Growth requires the replication of the organism's genome and synthesis of a new cell wall to contain the replicated DNA. In the ME-Model, growth rate-dependent DNA and cell wall demand functions formalize these requirements (Figure 2.1A; Supplementary information in [53]). We derived these demand functions from growth

rate-dependent trends in cell size [102] and DNA content [103, 104] (Supplementary information in [53]). In addition, as in previous models, we imposed growth-associated and non-growth-associated ATP utilization demands [105] as the ostensible energy requirements [106, 43].

Figure 2.1: **Growth demands and coupling constraints leading to growth rate-dependent changes in enzyme and ribosome efficiency.** (A) Three growth rate-dependent demand functions derived from empirical observations determine the basic requirements for cell replication (detailed in Supplementary information in [53]). (B) Coupling constraints link gene expression to metabolism through the dependence of reaction fluxes on enzyme concentrations. (C, D) RNA:protein ratio predicted by the ME-Model with two different coupling constraint scenarios, one for variable translation rate versus growth rate (red lines) and one for constant translation rate (orange lines). Experimental data in (C) obtained from Scott et al (2010). (E) Phosphotransferase system (PTS) transient activity following a glucose pulse in a glucose-limited chemostat culture (red) and glucose uptake before the glucose pulse (blue) is plotted as a function of growth rate. The data shown were obtained from O'Brien et al (1980)). Data from $\mu > 0.7\ h^{-1}$ were omitted. (F) Data from (E) are used to plot glucose uptake as a fraction of PTS activity. The resulting value is the fractional enzyme saturation (black line). The fractional enzyme saturation predicted by the ME-Model is plotted as a function of growth rate under carbon limitation (red dots). (G) The cartoon depicts changes in extra- (blue) and intra- (green) cellular substrate (circle) and product (triangle) concentrations and metabolic enzyme (orange) and ribosome (purple/maroon) levels as the concentration of a growth-limiting nutrient (and growth rate) increases. The dials show k_{eff}/k_{cat}, the effective catalytic rate over the maximum for metabolic enzymes (orange) and ribosomes (purple/maroon).


A
Growth-rate dependent demand functions
1. Cell Wall Demand(μ)
2. DNA Demand(μ)
3. ATP Demand(μ)
B
Ø
E
Enzymes
tRNAs
mRNAs
RNAP
Ribosome
Other machinery
Dilution
$\mu[E]$
Synthesis
$\mu[E] + k_{deg}[E]$
(at steady state)
Degradation
$k_{deg}[E]$
Reaction catalyzed by E
v
$v = k_{eff}(\mu)[E] \leq k_{cat}[E]$
C
RNA–Protein ratio (g/g)
ME–Model hyperbolic translation rate
ME–Model constant translation rate
Experimental
Growth rate, μ (h-1)
D
Effective translation rate (aa ribosome −1 s−1)
ME–Model hyperbolic translation rate
Constant translation rate
Growth rate, μ (h-1)
E
Glucose flux (nmol gDW −1 min−1)
PTS activity (Maximal rate)
Glucose uptake rate (Effective rate)
Growth rate, μ (h-1)
F
Experimental growth rate, μ (h-1)
Fractional enzyme saturation $\left(\frac{k_{eff}}{k_{cat}}\right)$
Experimental
ME–Model
ME–Model growth rate, μ (h-1)
G
0.5
0
1
$\frac{k_{eff}}{k_{cat}}$
μ increases through increases in effective catalytic rate (k_{eff})

One large improvement is that RNA and protein are not included as demand functions (as they are in M-Models; [20]; instead, expression of specific RNA and protein molecules are free variables determined during ME-Model simulations. 'Coupling constraints' [91, 51] relate the synthesis of RNA- and protein-based molecules to their catalytic functions in the cell (Figure 2.1B). The coupling constraints are based on parameters that define the effective catalytic rate (k_{eff}) and degradation rate constant (k_{deg}) of molecular machines (Supplementary information in [53]).

A nutritional environment is then defined by setting constraints on the availability and uptake of nutrients. For a particular nutritional environment, there is a maximum growth rate at which the cell can no longer produce enough RNA and protein machinery to meet the demands of growth. The computed cellular state (biomass composition, substrate uptake and by-product secretion, metabolic flux, and gene expression) at this maximum growth rate is the predicted optimal response of the cell to the specified nutritional environment.

2.3.3 Derivation of constraints on molecular catalytic rates

Previous studies disagree as to if ribosomes translate with the same efficiency (amino acids per ribosome per second) across growth conditions [107, 108]. Here, we use the ME-Model and available data to determine an appropriate constraint for ribosomal efficiency as a function of growth rate. We find that if a constant translation rate of 20 amino acids per second is imposed as a constraint in the ME-Model, the model predicts a linear growth rate-dependent RNA-to-protein ratio (Figure 2.1C), consistent with the previous measurements [108]; however, the predicted RNA content does not quantitatively match measured values. In particular, a constant translation rate results in no RNA production in the limit of no growth. We therefore hypothesized that ribosomal translation rate systematically varies with growth rate, and back-calculated a growth

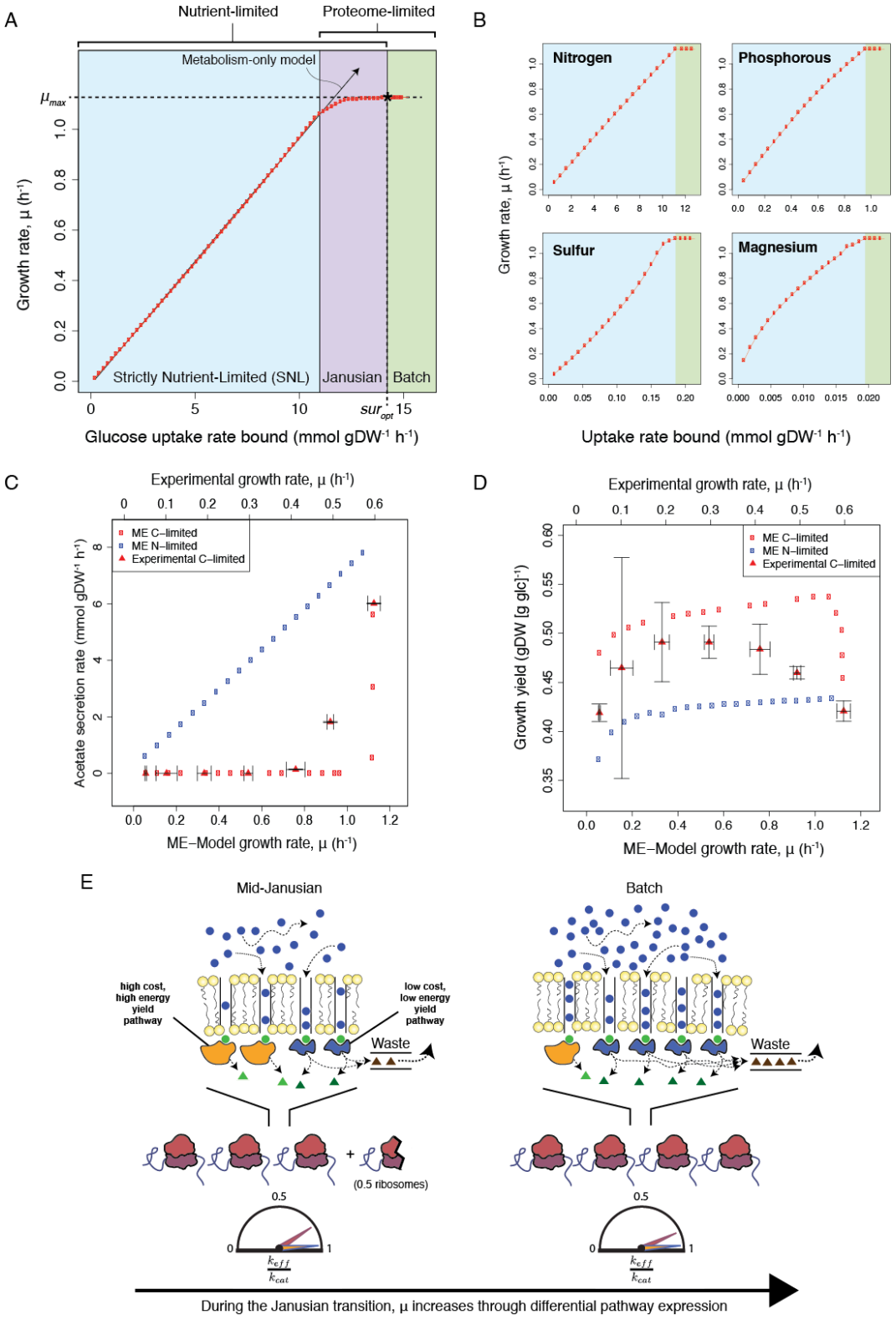
rate-dependent translation rate using measured growth rate-dependent RNA content (Supplementary information in [53]). Ultimately, we recovered a Michaelis-Menten-type rate law (Figure 2.1D) with a maximal rate (Vmax) of $\sim$20 amino acids per second, consistent with previous findings for maximal ribosomal speed [104]; the rate law results in a quantitative match of RNA content compared with experimental data (Figure 2.1C, Pearson's r=0.96). This rate law causes translation efficiency to increase under nutrient-richer conditions, which recent experimental evidence supports [109, 110]. Interestingly, when we applied the same Michaelis-Menten-type equations to constrain tRNA and mRNA catalytic rates, we recovered maximal turnover rates highly consistent with previous estimates (Supplementary information in [53]). The catalytic rates of metabolic enzymes are variable as well, and tend to decrease when nutrients are limited. Both metabolomics [111] and proteomics [110] data sets suggest a large-scale scaling of enzyme efficiencies under nutrient limitation. We approximate these changes in metabolic catalysis in the ME-Model with two minimal assumptions: (1) when the cell is nutrient-limited, protein content is maximized (at a given growth rate) and (2) this protein content specifically is metabolic enzymes not operating at their maximal catalytic rate [110] (i.e., $k_{eff}/k_{cat} < 1$, see Figure 2.1G and Supplementary information, Optimization procedure in [53]). These two assumptions allow us to predict average catalytic rates of metabolic enzymes under nutrient limitation. The nutrient limitation-dependent shape of our computed catalytic rates matches assays for glucose transporters under glucose limitation [112] (Figures 2.1E and F), LacZ under lactose limitation [113] Supplementary Figure S1A in [53]), and the enzyme efficiency in a small-scale optimality model accounting for substrate concentrations with Michaelis-Menten kinetics [114] (Supplementary Figure S1B in [53]). However, because the current ME-Model simulation procedure assumes that k_{eff} decreases uniformly across metabolism, the model does not capture the importance of specific enzymes for particular nutrient limitations; recent data sets [110] and kinetic

models [115] can help us understand and model these trends better at the genome-scale.

2.3.4 Growth regions under varying nutrient availability

Upon derivation of the growth demands and molecular efficiencies, we investigate high-level model behavior to variable nutrient availability. Unlike previous genome-scale models [98, 89], growth rate in the ME-Model is a non-linear function of the substrate uptake rate bound (Figure 2.2A), and eventually reaches a maximum. This behavior is consistent with long-standing empirical models of microbial growth [116, 117], in which growth is first nutrient-limited, but then limited by some intra-organismal bound.

Figure 2.2: **Predicted growth, yield, and secretion.** (A) Predicted growth rate is plotted as a function of the glucose uptake rate bound imposed in glucose minimal media. Three regions of growth are labeled Strictly Nutrient-Limited (SNL), Janusian, and Batch (i.e., excess of substrate) based on the dominant active constraints (nutrient and/or proteome limitation). The proteome-activity constraint inherent in the ME-Model results in a maximal growth rate and substrate uptake rate. The behavior of a genome-scale metabolic model (M-Model) is depicted with an arrow. (B) Predicted growth rates as a function of uptake of a limiting nutrient with glucose in excess. The shaded regions correspond to those as labeled in (A). (C) Experimental (triangle) and ME-Model-predicted (circle) acetate secretion in Nitrogen- (blue) and Carbon- (red) limited glucose minimal medium are plotted as a function of growth rate. Data were obtained from Zhuang et al (2011). The root-mean-square error (RMSE) between data and the ME-Model is 0.12 (for comparison, RMSE=0.40 for the M-Model). (D) Experimental (triangle) and ME-Model-predicted (circle) carbon yield (gDW Biomass/g Glucose) in Carbon- (red) and Nitrogen- (blue) limited glucose minimal medium are plotted as a function of growth rate. Data were obtained from Zhuang et al (2011). RMSE between data and the ME-Model is 0.04 (for comparison, RMSE=0.07 for the M-Model). (E) The cartoon depicts changes in extra- (blue) and intra- (green) cellular substrate (circle) and product (triangle) concentrations and metabolic enzyme (blue/orange) and ribosome (purple/maroon) levels during the Janusian region. Metabolic enzymes are saturated throughout the entire Janusian region. To increase the growth rate, the cell expresses metabolic pathways that have lower operating costs. (Pathways with the smaller blue proteins taken to be 0.25 the cost of the pathways with larger orange proteins.) A higher glucose uptake and turnover results, but energy yield is lower and some carbon is 'wasted' and secreted (brown triangles). The dials show k_{eff}/k_{cat}, the effective catalytic rate over the maximum for metabolic enzymes (blue/orange) and ribosomes (purple/maroon).


A
Nutrient-limited
Proteome-limited
Metabolism-only model
μ_{max}
Growth rate, μ (h⁻¹)
Strictly Nutrient-Limited (SNL)
Janusian
Batch
sur_{opt}
Glucose uptake rate bound (mmol gDW⁻¹ h⁻¹)
B
Nitrogen
Phosphorous
Sulfur
Magnesium
Growth rate, μ (h⁻¹)
Uptake rate bound (mmol gDW⁻¹ h⁻¹)
C
Experimental growth rate, μ (h⁻¹)
ME C–limited
ME N–limited
Experimental C–limited
Acetate secretion rate (mmol gDW⁻¹ h⁻¹)
ME–Model growth rate, μ (h⁻¹)
D
Experimental growth rate, μ (h⁻¹)
ME C–limited
ME N–limited
Experimental C–limited
Growth yield (gDW [g glc]⁻¹)
ME–Model growth rate, μ (h⁻¹)
E
Mid-Janusian
Batch
high cost, high energy yield pathway
low cost, low energy yield pathway
Waste
(0.5 ribosomes)
$\frac{k_{eff}}{k_{cat}}$
During the Janusian transition, μ increases through differential pathway expression

Under nutrient-excess conditions, growth in the ME-Model is limited by internal constraints on protein production and catalysis—the cell is 'proteome-limited'—resulting in a corresponding maximal growth rate (Figure 2.2A). This feature allows Batch culture growth to be simulated without specifying nutrient uptake bounds; instead, the ME-Model predicts a maximum batch growth rate and optimal substrate uptake rate.

Supporting the validity of the proteomic constraints limiting growth in Batch culture, optimal Batch growth rates, substrate uptake rates, and biomass yields correlate with experimental data for growth on different carbon sources (Supplementary Table S5 in [53]). The ME-Model predicted substrate uptake and biomass yield closely matches laboratory evolved strains (Pearson's r=0.89 and r=0.91, respectively) (Supplementary Table S5C in [53], sensitivity analysis in Supplementary Table S6 in [53]). Though less accurate, predicted growth rates by the ME-Model correlate with measured growth rates in batch culture better than standard M-Models, in which growth rate is maximized subject to a specified nutrient uptake, and the correlation increases when compared with laboratory evolved strains (M-Model Pearson's r=0.49, ME-Model Pearson's r=0.61) as opposed to wild-type strains (M-Model Pearson's r=0.30, ME-Model Pearson's r=0.39). Other methods that include various approximate constraints on the total flux through the metabolic network also show an increased performance in growth rate prediction, though all computational methods [42, 118] still correlate better with each other than with the experimental data (Supplementary Table S5B in [53]).

When the uptake of glucose is restricted below the amount required for optimal growth in batch culture, the cell's growth is carbon-limited. Growth rate linearly increases with glucose uptake when glucose availability is low. In this region (termed as the Strictly Nutrient-Limited (SNL) region in Figure 2.2A), the capabilities of the proteome are not fully utilized as the proteome could process more incoming glucose if it was available (Figures 2.1E-G). By varying the glucose availability, we find that a region exists in

which the cell is both nutrient- and proteome- limited; we refer to this transition region as the Janusian region [119]. ME-Model computations thus reveal three distinct regions of microbial growth (Figure 2.2A; see Supplementary information, Optimization procedure, Computational definition, and identification of growth regions in [53]).

When the uptake of non-carbon sources is restricted below the amount required for optimal growth in batch culture, the cell's growth is limited by that nutrient. Unlike carbon-source limitation, we find the nutrient- and proteome-limited regions to be distinct (Figure 2.2B). However, in the SNL region, growth is sometimes non-linear as a function of uptake rate, due to changing biomass requirements (e.g., Sulfur and Magnesium).

2.3.5 Effect of proteome limitation on secretion phenotypes

To understand the proteome-limited growth regions in the ME-Model, we first investigate trends in secretion phenotypes and biomass yield. Under glucose limitation, different metabolic pathways are utilized in the Janusian region than in the SNL region, resulting in acetate secretion (Figure 2.2C, red). This metabolic switch, combined with growth rate-dependent ATP requirements, results in a concave biomass yield as a function of growth rate (Figure 2.2D, red). Both the biomass yield and secretion trends have repeatedly been experimentally observed [43]. The example of glucose limitation provides an illustrative example for the general behavior in the Janusian growth region. In the Janusian region, the cell increases its growth rate through differential expression of pathways, as illustrated in Figure 2.2E. Due to proteome limitations, the cell switches to pathways that require less protein mass but are lower in nutrient yield (defined as energy and/or biomass precursors produced per molecule of limiting nutrient consumed). This behavior is in contrast to that in the SNL region, in which high-yield pathways are optimal (as in M-Models) and growth rate increases through changes in the effective catalytic rate of metabolic enzymes (Figure 2.1G). These results provide further support that 'overflow'

metabolism can be understood in terms of proteomic constraints, as suggested with a small-scale model [114].

The ME-Model also predicts that acetate will be secreted at all growth rates when *E. coli* is Nitrogen (Ammonium)-limited (Figure 2.2C, blue). Experimentally, acetate is secreted under nitrogen limitation even at low growth rates [120]. This secretion phenotype is explained by the ME-Model as follows: protein 'saved' by utilizing low-yield carbon metabolism is diverted to synthesize other enzymes that are not operating at their maximal catalytic capacity.

No Janusian region is observed under non-carbon limitation. In the ME-Model, this is likely due to reaction network topology—while there are many alternative pathways for energy, redox, and biomass precursor generation in carbon metabolism, non-carbon nutrient assimilation is often achieved using more linear pathways. As a result, there are fewer opportunities for trade-offs between uptake rate and biomass yield. However, perhaps including variable substrate affinities for alternative pathways would reveal Janusian regions corresponding to non-carbon limitations.

2.3.6 Central carbon fluxes reflect growth optimization subject to catalytic constraints

Further supporting the importance of proteomic constraints on metabolic phenotypes is the prediction of central carbon fluxes by the ME-Model. When glucose availability is varied, the ME-Model predicts changes in central carbon metabolism consistent with the changes from ^{13}C fluxomic data sets (Figure 2.3; Supplementary Figure S2 in [53], Pearson's r=0.93, 0.90, 0.86) [121, 52, 47]. Importantly, the ME-Model predicts the dominant changes in pathway splits as the glucose availability is varied (Figure 2.3, insets).

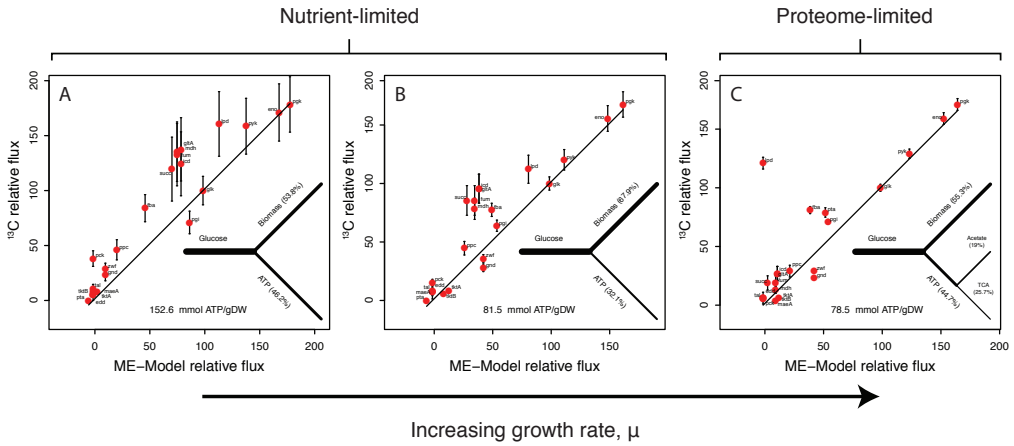


Figure 2.3: **Central carbon metabolic flux patterns under glucose-limited and glucose-excess conditions.** (A-C) Relative fluxes from ^{13}C experiments are plotted versus the fluxes predicted by the ME-Model. (A, B) Comparison of nutrient-limited model solutions with chemostat culture conditions and (C) comparison of the batch ME-Model solution with batch culture data. All simulations and experiments correspond to growth in glucose minimal media. Fluxes are normalized so that glucose uptake is 100. Insets show the main flux changes under increasing glucose concentrations. The only model parameter that is modulated is the glucose uptake rate bound. Data were obtained from Nanchen et al (2006) and Schuetz et al (2007). The ME-Model flux for the reaction 'pyk' is taken to include phosphoenolpyruvate (PEP) to pyruvate (PYR) conversion via the phosphotransferase system (PTS). Flux splits shown as insets were computed using the ME-Model. The percentages indicate the percent carbon (Glucose) converted to CO_2 (for branch labeled 'TCA'), acetate, and biomass. Both the TCA and acetate branches contribute to ATP production. The total mmol ATP per gDW biomass produced is indicated.

Previous studies have evaluated the ability of M-Models together with assumed optimality principles to predict metabolic fluxes [52, 47]. These studies concluded that no single objective function applied to M-Models can accurately represent fluxomic data from all environmental conditions studied [52]. Instead, metabolic fluxes can be understood as being Pareto optimal: multiple objectives are simultaneously optimized and their relative importance varies depending on the environmental condition [47]. The three objectives needed to explain most of the variations in the data from Schuetz et al were (1) maximum ATP yield, (2) maximum biomass yield, and (3) minimum sum of absolute fluxes (which is a proxy for minimum enzyme investment). These three objectives formed a Pareto optimal surface that was valuable for interpreting fluxomic data; however, the surface was large and it was not possible to predict the importance of each of the objectives a priori.

By explicitly accounting for variable growth demands, enzyme expression, and constraints on enzymatic activity, the ME-Model eliminates the need for multiple objectives; growth rate optimization alone is sufficient to predict the fluxes through central carbon metabolism (Figure 2.3; Supplementary Figure S2 in [53]; Supplementary Table S7 in [53]). The three original objectives chosen by Schuetz et al are biologically meaningful dimensions and required for interpreting fluxomic data when using an M-Model. In contrast, the ME-Model accounts for all three of these dimensions implicitly during growth rate maximization without adjusting any model parameters (see Supplementary information in [53] and Supplementary Table S7 in [53]). Accordingly, ME-Models can determine, at least qualitatively, the importance and weighting of the objectives for growth in a given environment. Ultimately, the primary changes in flux through central carbon metabolism can be understood as responses to the same constraints causing the observed relationship in biomass yield (Figure 2.2D): at low growth rates under carbon limitation, the dominant changes are due to a changing ATP demand, and in the transition

from carbon-limited to carbon-excess (proteome-limited) conditions, the primary changes are due to the switch to lower yield carbon catabolism (Figure 2.3, insets).

2.3.7 *In silico* gene expression profiling from nutrient-limited to batch growth conditions

We now use the ME-Model to predict groups of proteins that change in expression under various degrees of glucose limitation. Under glucose limitation, the optimal proteome changes due to shifting growth demands and proteomic constraints. The groups of functionally related proteins that shift in our simulations match those previously reported experimentally [122, 123], but the model predictions of quantitative differential expression (at the level of single genes) are weak. We separate the analysis of the SNL region (Figure 2.4; Supplementary Table S8A in [53]) from the Janusian region (Figure 2.5; Supplementary Table S8B in [53]), due to the different dominant constraints and phenotypic responses specific to each region.

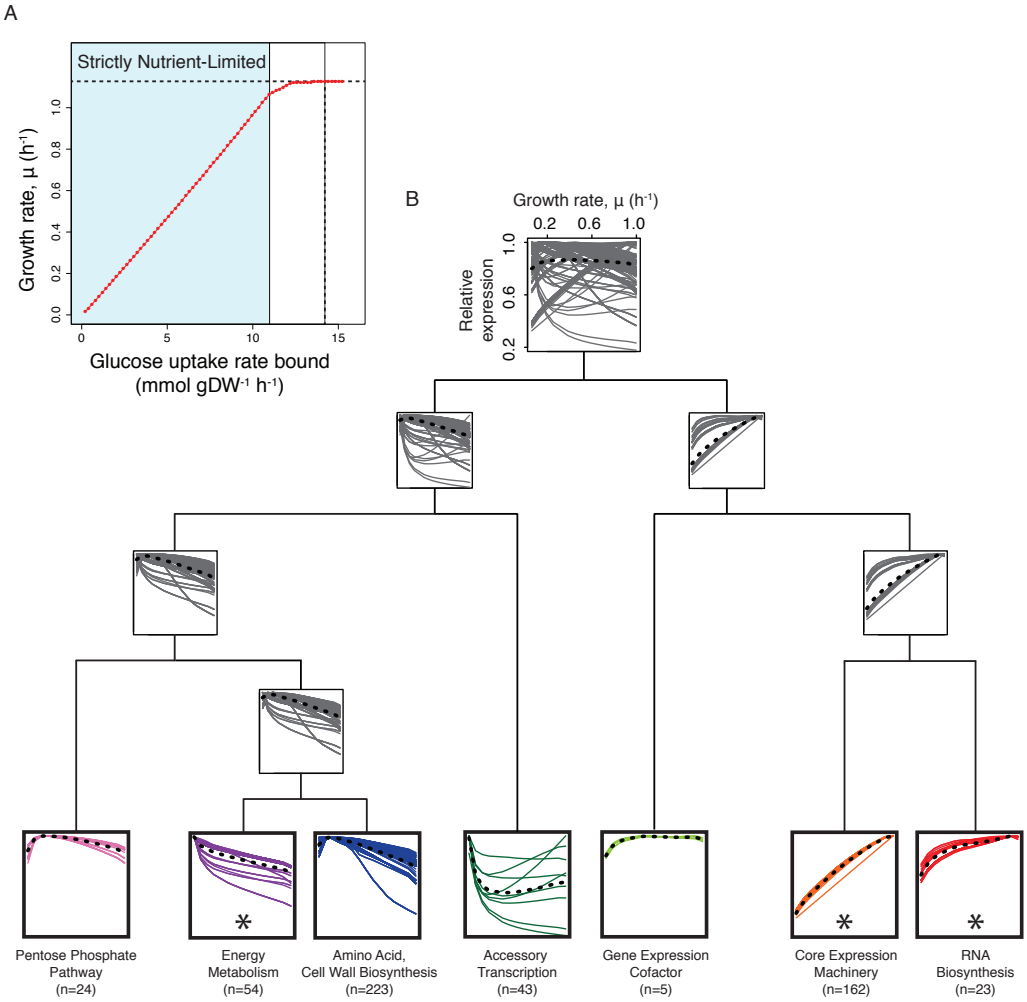


Figure 2.4: **Growth rate-dependent gene expression under glucose limitation.** (A) Gene expression changes predicted by the ME-Model to occur in the Strictly Nutrient-Limited (SNL) growth region indicated in light blue under glucose limitation in minimal media are analyzed. (B) ME-Model-computed relative gene-enzyme pair expression is plotted as a function of growth rate; the normalized *in silico* expression profiles are clustered hierarchically (see Materials and methods). Solid lines are expression profiles of individual gene-enzyme pairs and dotted black lines are the centroid of each cluster. Each leaf node is colored and qualitatively labeled by function. The number of genes in each leaf node is indicated and listed in Supplementary Table S8A in [53]. Asterisks indicate clusters with monotonic expression changes that significantly match the directionality observed in expression data (Wilcoxon signed-rank test, $P < 1 \times 10^{-4}$). Expression data were obtained from a previous study [123], in which *E. coli* was cultivated in a chemostat at dilution rates $0.3\,h^{-1}$ and $\sim 0.5\,h^{-1}$.

In the SNL region, the expression of most proteins decreases as growth rate increases (Figure 2.4B, left side of tree, Supplementary Figure S3 in [53]). The largest group of proteins includes those responsible for amino-acid and cell wall synthesis; the growth rate-dependent decrease in expression of these proteins is due to the combined effects of a decrease in cell wall and protein biomass (g/gDW) and an increase in the effective catalytic rate of enzymes (Figures 2.1E-G). Proteins involved in energy metabolism also decrease in expression with increasing growth rate due to changes in catalytic rate and growth rate-dependent demands. Surprisingly, the predicted expression levels of several accessory transcription proteins, including four stress-associated sigma factors (RpoS, RpoH, RpoE, and RpoN), are elevated at very low growth rates, reflecting an association with metabolic proteins needed for slow growth.

A smaller number of proteins show increases in their relative expression levels at higher growth rates (Figure 2.4B, right side of tree, Supplementary Figure S3 in [53]). These proteins include those responsible for protein synthesis (ribosome, RNAP, and accessory proteins such as elongation factors) and proteins involved in RNA biosynthesis. The increase in expression of RNA biosynthetic machinery is necessary for de novo synthesis of ribonucleotides and to ensure flux through nucleotide salvage pathways (mainly to support an increase in rRNA biomass). Finally, the expression profile of the pentose phosphate pathway reflects the interplay between the increasing demand for ribonucleotide precursors and the decreasing demand for amino-acid precursors.

To validate our predicted expression changes, we compared gene clusters with expression data from *E. coli* grown at $0.3\,h^{-1}$ and $\sim 0.5\,h^{-1}$ in a glucose-limited chemostat [123]. In this data set, genes in Energy Metabolism (purple), Core Expression Machinery (orange), and RNA Biosynthesis (red) all significantly change in the predicted direction (Wilcoxon signed-rank test, $P < 1 \times 10^{-4}$), supporting our predicted expression profiles. The other clusters showed no significant changes in the data set; these clusters are either

small in size or do not change monotonically, hindering direct comparison with this data set. The ME-Model is not yet predictive of quantitative gene expression changes (at the level of single genes); the correlation over the entire data set is statistically significant ($P < 0.005$), but weak (Pearson's r=0.14). Our approach is at present limited to qualitative predictions of the direction of change of small groups of functionally related proteins.

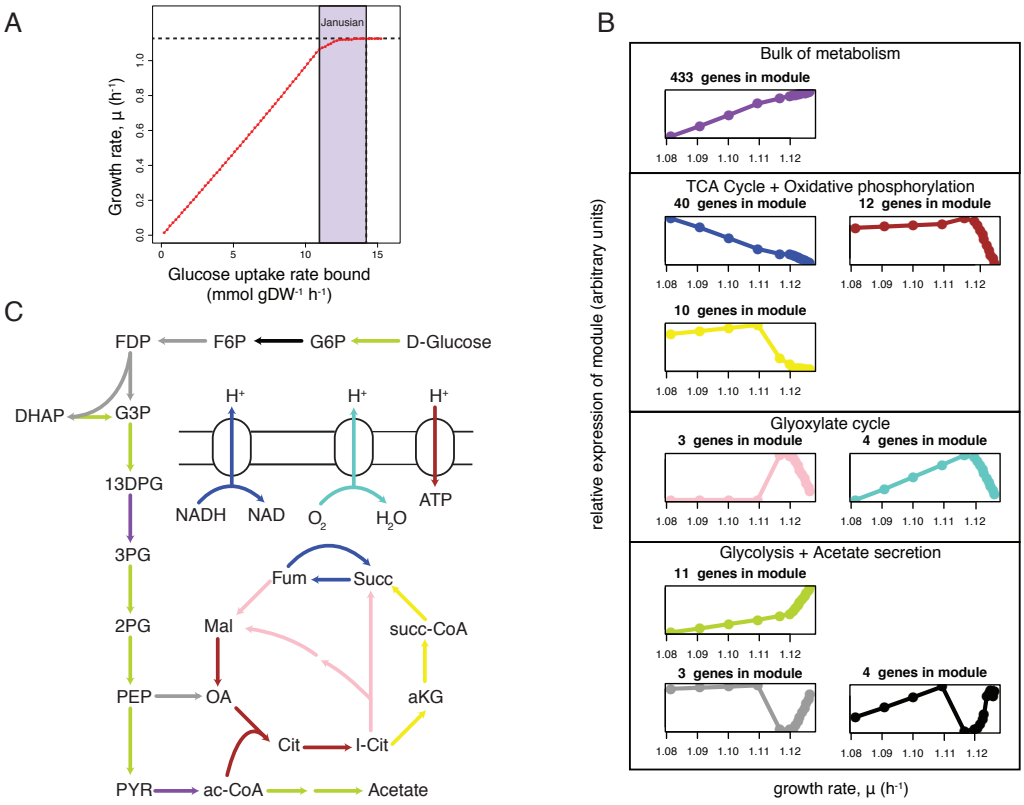


Figure 2.5: **Gene expression during the Janusian region.** (A) Gene expression changes predicted by the ME-Model to occur in the Janusian growth region indicated in purple under glucose limitation in minimal media are analyzed. (B) Simulated expression profiles are clustered using signed power ($\beta = 25$) correlation similarity and average agglomeration. A freely available R package was used (Langfelder and Horvath, 2008). Eleven clusters resulted. Two small clusters were removed because they represented stochastic expression of alternative isozymes. The first principal component of the remaining nine clusters is displayed and grouped qualitatively by function. (C) Many of the expression modules correspond to genes of central carbon energy metabolism. Reactions are colored according to the module color in (B).

In the Janusian region of growth (Figure 2.5), the cell transitions from carbon-limited to proteome-limited constraints, resulting in a distinct transcriptional response. At the beginning of this transition, the cell has reached a nutrient level where enzymes are saturated (Figure 2.1G); as growth rate increases, the total demand of anabolic processes increases, causing a global increase in the bulk of metabolism and gene expression machinery (Figure 2.5B). To meet these proteome demands, energy metabolism is altered to favor lower yield catabolic pathways that require less protein (so that the protein can instead be used for anabolic processes); this is accomplished through a decrease in TCA Cycle and Oxidative Phosphorylation expression in favor of a transient increase in the Glyoxylate Cycle followed by a large increase in Glycolysis and acetate secretion (Figures 2.5B and C), consistent with previously observed changes in gene expression in the transition to glucose-excess environments [122].

The ME-Model predicts intricate expression changes as glucose availability changes by employing relatively simple constraints on molecular catalysis and biomass composition. This study is the first to attempt genome-scale prediction of gene expression levels under changing growth rate and/or nutrient limitation from optimality principles alone. Systematic consideration of transcriptional regulation and inclusion of missing constraints and parameters impacting optimality (e.g., kinetic constraints and parameters) are future endeavors necessary to extend the predictive power to the level of single genes (see Discussion).

2.4 Discussion

The ME-Model is a microbial growth model that computes the optimal cellular state for growth in a given steady-state environment. It takes as input the availability of nutrients to the cell and produces experimentally testable predictions for: (1) the cell's

maximum growth rate (μ^*) in the specified environment, (2) substrate uptake/by-product secretion rates at μ^*, (3) metabolic fluxes at μ^*, and (4) gene product expression levels at μ^*.

Important to the predictions of the ME-Model is the proper coupling between metabolism and gene product expression. Through comparison of model simulations with experimental data, we derived two general classes of molecular efficiencies that vary based on the growth rate and the degree of nutrient limitation. For ribosomes (and tRNA and mRNA), we propose a growth rate-dependent Michaelis-Menten-type model for polymerization speed, which has preliminary experimental evidence [109], though we have not seen it previously proposed. We furthermore show that two simple assumptions allow us to approximate the effect of nutrient limitation on metabolic enzyme catalysis. While enzyme-specific trends in catalytic rates depend on the limiting nutrient [124, 111], our formulation is a first step toward modeling genome-scale effects of nutrient limitation and suggests that simple principles may underlie these trends. Both of these molecular efficiency variables are essential for genome-scale modeling of gene expression and warrant future studies to validate and refine them further. Paired proteomic and metabolomic data sets under nutrient-limited conditions will allow for a deeper understanding of nutrient limitation-dependent effective catalytic rates, and new data sets [125] and models [126] on the processes of gene expression can help to refine model parameters and determine their genome-scale effects.

The proteomic constraints inherent to the ME-Model result in qualitatively different growth predictions compared with previous genome-scale models. In the ME-Model, growth rate is not a simple linear function of substrate uptake bounds; instead, the ME-Model predicts a maximal growth rate and optimal substrate uptake rates, which better reflects empirical growth models and better predicts experimentally measured growth rates and substrate uptake rates. The ME-Model reveals three distinct growth regions,

which we term SNL, Janusian, and Batch; while nutrient-limited (chemostat culture) and nutrient-excess (batch culture) conditions are commonplace, the Janusian region (where the cell is limited by both nutrient availability and proteome capacity) is rarely considered in microbiology. Interestingly, we observe the Janusian region to occur under carbon limitation but not under various non-carbon limitations. We take this to mean that Janusian regions may exist for non-carbon limitations, but the constraints that may cause them to arise are outside the scope of the current ME-Model.

The proteomic constraints in the ME-Model also improve predictions of by-product secretion and metabolic flux under both nutrient-excess and nutrient-limited conditions. By accounting for the metabolic cost of proteins and limitations of protein production capacity, the ME-Model accurately decouples substrate uptake, growth rate, and growth yield, allowing for important rate-yield trade-offs to be predicted. In particular, we show that seemingly inefficient metabolism in batch culture and under nitrogen limitation (both when carbon is in excess), can be explained and predicted through proteomic trade-offs. This capability rectifies the dominant failure mode in predicting metabolic flux previously reported for M-Models [47], and suggests that a single objective of growth rate (if the proper constraints are included) may be able to predict metabolic fluxes. This result shows that proteomic constraints are necessary to accurately predict metabolic responses—optimal growth and metabolic phenotypes cannot be fully understood without taking gene expression into account. From a practical standpoint, the natural parsimony present in ME-Model simulations [51] strongly reduces the optimal solution space, allowing for more precise predictions, an important feature in diverse applications. The effect of proteomic constraints on secretion phenotypes is of particular importance for applications in systems metabolic engineering, and will be necessary for simulating behavior in complex media and predicting nutrient preferences.

At the level of gene expression, the ME-Model predicts detailed behavior in each

growth region. In the SNL and Janusian growth regions, gene modules have distinct nutrient limitation-dependent profiles. A number of the gene modules change in the correctly predicted direction compared with expression data from *E. coli* in a chemostat at different growth rates [122, 123], supporting our predicted expression profiles. By predicting optimal gene expression profiles, the ME-Model aids in understanding the factors shaping the evolution of gene expression patterns (e.g., proteomic constraints and changing biomass composition).

Modeling optimal transcriptional responses is complementary to the elucidation and modeling of specific regulatory mechanisms [127, 128, 129]. It is tempting to relate the expression profiles predicted by the ME-Model to molecular mechanisms underlying the control gene expression *in vivo* [127, 129, 130]. For example, constitutively expressed genes display growth rate-dependent expression trends [131, 127], which might provide the cell with an economical way of responding to global changes in metabolic efficiency [110]. Also, PurR could be responsible for regulating the increase in expression of nucleotide biosynthesis genes at higher growth rates (as PurR is an autorepressor, this could be accomplished through mechanisms described in [127]. Finally, though the primary role of ArcA is to respond oxygen availability [132], it also represses many of the genes in the TCA cycle and Oxidative Phosphorylation that decrease during the glucose-limited to glucose-excess (Janusian) transition [122, 133]. However, as regulatory mechanisms are not explicitly considered in the ME-Model, the relation between regulatory mechanisms and simulated expression profiles is indirect; while this comparison can assist in explaining and expanding upon the functional roles of cellular regulators, much further work is required to validate the resulting hypotheses.

As it is an optimality model, the ME-Model is particularly suited for studies related to adaptive laboratory evolution (ALE). Recently, it was reported that it is not possible to predict some changes that occur during ALE in Batch culture using an M-

Model [134]. This is because M-Models only take biomass yield optimization into account; these results are consistent with the rate-yield trade-offs present in the ME-Model under nutrient-excess conditions. In the ME-Model, a number of inherent factors can limit cellular growth (e.g., translation rate and metabolic catalysis); the ME-Model can thus provide alternative hypotheses for the mechanisms of growth increase and aid in understanding the results of ALE.

The ME-Model can simulate coarse- to fine-grained cellular and molecular phenotypes with an improved accuracy and scope compared with previous genome-scale models. The ME-Model shows complex behavior as a result of linear constraints applied to an integrated network. The ME-Model thus shows that intricate and seemingly unintuitive phenotypes can be modeled at a genome-scale with simple enough assumptions to understand their underlying cause. Due to the richness of the model simulations, we primarily focused on *E. coli* growing in glucose minimal media at different growth rates by modulating the availability of glucose; there are therefore many future opportunities to investigate model predictions under many environmental and genetic conditions.

A whole-cell *E. coli* model has been desired for some time [135] as such a model would have profound impacts for basic microbiology, the study of microbial communities, antibiotic discovery, the elucidation of regulatory networks, and systems metabolic engineering. We hope the ME-Model will serve as a scaffold for continued model development toward these practical applications.

2.5 Materials and methods

2.5.1 Network reconstruction

The two primary reaction networks used to create the ME-Model were the most recent metabolic reconstruction [98], and a network detailing the reactions of gene

expression and functional enzyme synthesis [90]. The gene expression reconstruction is formalized as a set of 'template reactions' that can be applied to different components (e.g., gene, peptide, and set of peptides) to generate balanced reactions. Merging the *E. coli* metabolic network reconstruction with the gene expression reconstruction required a conversion of the Boolean Gene-Protein-Reaction associations (GPRs) into protein complexes. We utilized EcoCyc's annotation to map gene sets to functional enzyme complexes. The content of the final reconstruction is detailed in Supplementary Tables S1, S9, and S10 in [53].

2.5.2 Coupling constraint formulation and imposition

Coupling constraints provide a mechanism for linking the flux values of one or more reactions in the ME-Model. For example, they were used to bound the number of proteins that may be translated from an mRNA before the mRNA decays or is transmitted to a daughter cell. They are also the mechanism through which we related enzyme abundance and activity. Often, the coupling constraints are a function of the organism's growth rate (μ). The coupling constraints are a set of inequality constraints appended to the stoichiometric matrix as additional rows. Assumptions and literature citations for all parameters used can be found in Supplementary information in [53].

2.5.3 Optimization procedure

As the demand reactions and coupling constraints are functions of the organism's growth rate (μ), growth-rate optimization is not a linear program (LP) as in metabolic models, which rely on a linear biomass objective function. Instead, to optimize for growth rate, we solve a sequence of LPs to search for the maximum growth rate, μ*, that still results in a feasible LP. This search for μ* is accomplished through a binary search; the

search procedure is slightly different depending on whether the cell is proteome-limited (Janusian and Batch growth modes) or SNL. Detailed traces of the execution of the optimization procedures can be found in Supplementary information in [53].

2.5.4 Hierarchical clustering

For Figure 2.4B, relative fractional proteome mass was calculated for each gene-enzyme pair. If a gene is present in multiple enzyme complexes, then it is represented twice, and all subunits of an enzyme complex are counted separately. To filter out the stochastic expression of alternative isozymes (to make the observed trends clear), we eliminated gene-enzyme pairs that were not expressed across all growth rates and filtered gene-enzyme pairs that changed in relative expression by >0.3 across more than one pair of consecutive growth rates. Hierarchical clustering was performed on the resulting expression profiles; we used a signed power ($\beta = 6$) correlation similarity (as in [136]) and average agglomeration.

2.5.5 File formats and accessibility

The model is freely available as part of the openCOBRA Project (http://opencobra.sourceforge.net).

2.6 Acknowledgements

We thank Thorsten Koch, Matthias Miltenberger, Ambros Gleixner, Michael Saunders, and Martina Ma for help solving linear programming problems. We thank Adam Feist, Harish Nagarajan, and Pep Charusanti for invigorating discussions. EJO and RLC were supported by NIH R01 GM057089. JAL was supported by NIH U01 GM102098. This work was supported in part by the US National Institute of Allergy

and Infectious Diseases and the US Department of Health and Human Services through interagency agreement Y1-AI-8401-01. DRH is supported in part by Y1-AI-8401-01. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231.

Author Contributions: JAL, EJO, and BOP conceived the study. JAL and EJO performed the computational analysis. JAL, EJO, and RLC performed model updates. RLC, DRH, and BOP supervised the study. JAL and EJO wrote the manuscript. All authors helped edit the final manuscript.

The text of Chapter 2 is a full reprint of the material as it appears in: O'Brien E.J.*, Lerman J.A.*, Chang R.L., Hyduke D.R., Palsson B.O. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol. Syst. Biol., 9:693. (2013). * indicates equal contribution. The dissertation author was the primary author of the manuscript. The other authors were Joshua A. Lerman (equal contributor), Roger L. Chang, Daniel R. Hyduke, and Bernard Ø. Palsson.

Chapter 3

Proteome allocation constraints determine cellular growth rates and demand fitness trade-offs

The dream of every cell is to become two cells.
—Francois Jacob

Exponential growth is miraculous no matter where it happens.
—Michael Saunders

3.1 Summary

Protein synthesis is costly and the proteome size is constrained. Using a genome-scale computational model of proteome allocation together with absolute proteomics data sets from many growth environments, we determine how these fundamental limitations constrain growth and fitness in Escherichia coli. First, we show that the observed variation in growth rates across environments is largely determined by the expression of protein not utilized for growth in a given environment. We then elucidate the overall transcriptional

regulatory logic that underlies the expression of unused protein. We systematically classify the unused proteome into segments devoted to environmental readiness and stress resistance functions. While expression of these proteome segments incurs a fitness cost of decreased growth in a fixed environment, they provide fitness benefits in a changing environment. Thus, the systems biology of the prokaryotic proteome can be quantitatively understood based on resource allocation to growth, environmental readiness, and stress resistance functions.

3.2 Introduction

Cellular proteomes are limited in size and expensive to synthesize. Allocation of proteome resources to the various functions enabling organism growth and survival is therefore a significant evolutionary selection pressure [137]. Studying the proteome allocation of an organism is thus crucial to understanding its fitness characteristics, ecological strategy, and evolutionary history.

To relate proteome allocation to organismal fitness and evolution, protein expression must first be related to physiological functions. Several quantitative relationships between protein expression levels and cellular physiology have been identified [108, 138]. It is increasingly recognized that these patterns reflect evolutionary selection pressures that balance the cost and benefit of protein expression [44, 139, 140]. With recent advances in proteomic data [141] and modeling [53, 94], we can now link proteome allocation to cellular physiology and underlying evolutionary pressures at a genome-scale and a protein-level resolution.

Here, we study how the model bacterium Escherichia coli allocates its proteome in different environments. With our quantitative systems biology approach, we can now elucidate, on a genome-scale, the determinants of microbial growth rates and the

underlying regulatory logic and proteomic basis of fitness trade-offs. This systems biology approach results in an unprecedented mechanistic basis for proteome allocation. Fundamental understanding of the systems biology of the bacterial proteome is thus revealed.

3.3 Results

3.3.1 Defining the un-utilized and under-utilized ME proteome

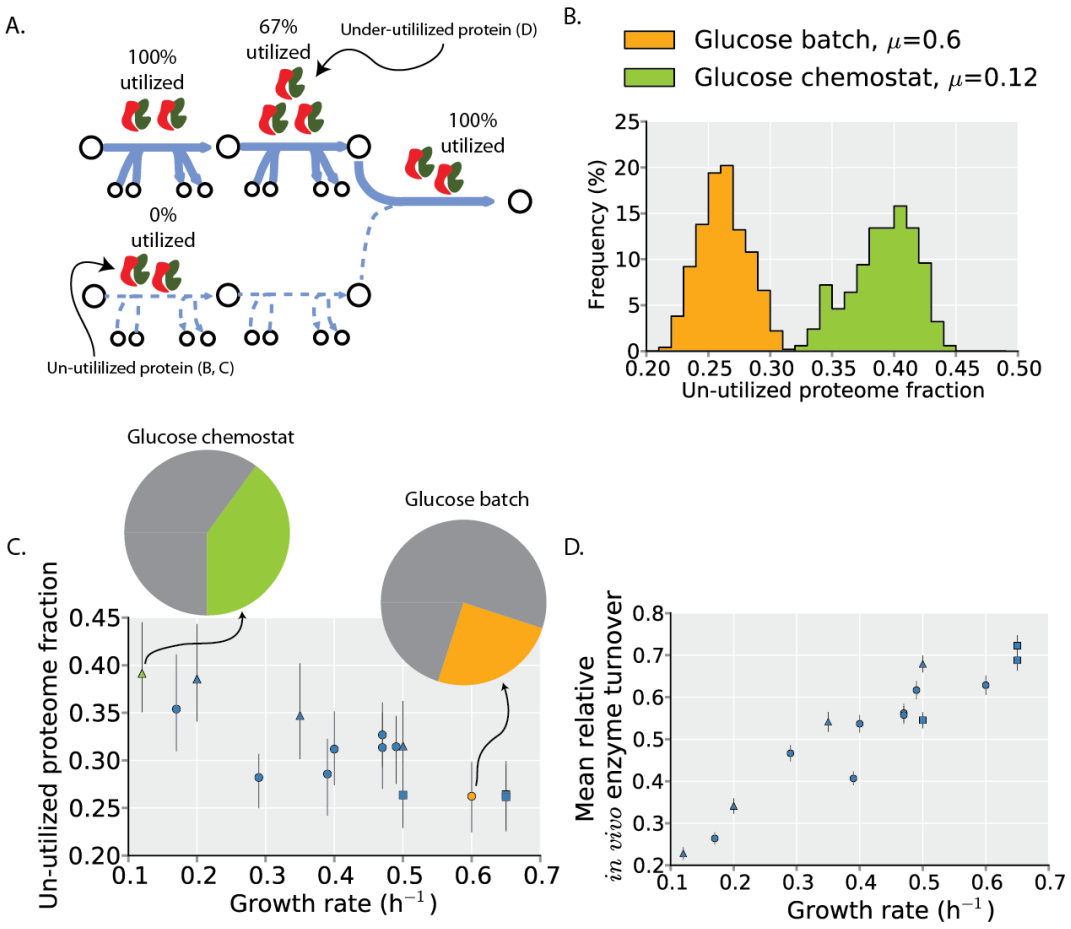
We distinguish between two classes of unused protein (Figure 3.1A). The first class is the un-utilized protein. This is protein that, in the specified environment, is not utilized for cellular growth. For example, in glucose minimal media, the glycerol transporter is un-utilized, but it might be expressed.

The second class of unused protein is the under-utilized protein. This is protein that is catalytically active, but is present in excess and thus operating under its maximal capacity. By combining known demands for biosynthesis with measured protein expression levels we identify enzymes operating below their maximal capacity (see Methods).

We use a genome-scale model of proteome allocation in Escherichia coli, termed a ME-Model [53, 94] to aid in the quantification of un- and under-utilized protein. The ME-Model formalizes the function and synthesis of proteins involved in metabolism and protein expression, which we refer to as the ME proteome. The ME protome encompasses much of the proteome required for cellular growth and accounts for 80% of the proteome by mass in conditions of exponential growth. The remaining 20% of the proteome is the non-ME proteome, which largely have functions outside of metabolism and protein expression; the non-ME proteome is analyzed after we detail the computable ME-proteome.

With ME-Model simulations we can identify proteins that can be utilized for growth in a particular environment. By combining these simulations with quantitative proteomics data, we can identify proteome-wide changes in both un- and under-utilized protein abundances (see Methods). If the abundances of un- and under-utilized protein vary significantly across environments, the unused proteome expression will be an important determinant of growth rate variation.

Figure 3.1: **Unused protein abundances are not constant across environments** A. Graphical illustration of the two classes of unused protein defined in the text. The first class is protein that is expressed but completely un-utilized (i.e., 0% utilized); this is protein that is catalytically inactive (i.e., carrying zero flux; first step in lower pathway). The second class is protein that is under-utilized (i.e., ¡100% but ¿0% utilized; second step in upper pathway); this protein is catalytically active, but not operating at its maximal turnover rate (see Methods). The percent utilization indicated is based on the upper pathway having flux of 2, the lower pathway having flux of 0, and all enzymes having a maximal rate of 1; thick blue lines indicate the active pathway and dotted line indicates the inactive pathway. B. The un-utilized proteome fractions for batch and chemostat culture on glucose minimal media. The un-utilized proteome fraction is a distribution rather than a specific value because of different potential alternative pathways and enzymes that can be used to support cellular growth. The distributions are non-overlapping indicating that the un-utilized proteome fraction is not the same in these two environments. As the nutrient source is the same in these two environments, the proteins that can be utilized for growth in the ME-Model are identical, and variation in the proteomics data determines the two distributions. C. The un-utilized proteome fraction across all profiled environments–8 different carbon source batch cultures (circles), 4 glucose-limited chemostat cultures (triangles), and 3 stress conditions (squares)–is plotted as a function of the growth rate measured in that condition. Error bars indicate 2.5 and 97.5 percentiles of the un-utilized proteome distributions for each condition (e.g. Figure 3.1B). The orange and green points and pie charts correspond to the environments in B of the same color; the pie charts show the change in proteome allocation to un-utilized protein. D. Enzyme turnover rates tend to increase at higher growth rates. Changes in turnover rates indicate under-utilized protein. For proteins used across all environments, turnover rates are determined by taking the ratio of the computed demand of the protein and the measured protein abundance; relative values are found by normalizing by the maximum turnover rate for that protein across all environments (see Methods). The mean relative turnover (across all proteins) is plotted with error bars indicating the 95% confidence interval for the mean. Point shape indicates environment type as in C (carbon source batch cultures = circles, glucose-limited chemostat cultures = triangles, stress conditions = squares).


A.
100% utilized
67% utilized
Under-utililized protein (D)
100% utilized
0% utilized
Un-utililized protein (B, C)
B.
Glucose batch, μ=0.6
Glucose chemostat, μ=0.12
Frequency (%)
Un-utilized proteome fraction
C.
Glucose chemostat
Glucose batch
Un-utilized proteome fraction
Growth rate (h⁻¹)
D.
Mean relative in vivo enzyme turnover
Growth rate (h⁻¹)

3.3.2 Un-utilized and under-utilized ME proteome abundance varies across environments

In a given environment, several different alternative pathways and enzymes can be used to support cellular growth [48, 142, 17]. Thus, the computed un-utilized proteome fraction needs to be assessed across all such alternate solutions. Randomized sampling the alternate network states leads to a distribution of possible un-utilized proteome fractions rather than a single specific value (Figure 3.1B, see Methods).

We first compare the un-utilized proteome fraction during growth in glucose batch culture and chemostat culture. As the nutrient source is the same in these two environments, the set of proteins that can be utilized for growth in the ME-Model are identical; variation in the proteomics data will determine differences in the un-utilized proteome abundances. As the two distributions of un-utilized proteome abundance are non-overlapping, the un-utilized proteome fraction does vary significantly between these two environments (Figure 3.1B).

When data from multiple growth conditions (8 different carbon sources in batch cultures, 4 glucose-limited chemostat cultures, and 3 stress conditions–acid, osmotic, and temperature; see Methods) are analyzed in a similar manner, a clear general trend emerges in which environmental conditions resulting in higher growth rates tend to have lower un-utilized proteome fractions (Figure 3.1C). This correlation suggests that un-utilized proteome fraction is an important source of growth rate variation.

Next we consider the under-utilized proteome under the same growth conditions by computing the variation in in vivo enzyme turnover (flux per protein; see Methods). We find that the abundance of the under-utilized proteome is not constant across environments; rather, the average enzyme turnover tends to increase in environments with higher growth rates (Figure 3.1D). This trend has been observed for several individual proteins

using lower throughput methods [53, 110, 143].

Thus, we find that the amount of both un-utilized protein and under-utilized protein is reduced with increasing cellular growth rate.

3.3.3 Growth rate is determined by the unused proteome expression

How much does the variation in un- and under-utilized proteome contribute to the variation in growth rates across environments? Un-utilized proteome fraction and in vivo enzyme turnover are variables that are formalized in the ME-Model. Setting these parameters to measured values can determine how they quantitatively affect growth rates.

Even if un- and under-utilized protein abundances are constant across environments, different nutrient sources are expected to result in different growth rates (e.g., due to differences in biomass yield). As a baseline for comparison, we first predict maximum growth rates with the ME-Model while holding the un-utilized proteome fraction and in vivo enzyme turnover rates constant across environments; this assumption allows us to assess the contribution of nutrient quality (i.e., biomass yield) and proteome capacity (i.e., the maximal catalytic capacity of the proteome) to growth rate variation independent of changes in un- and under-utilized protein. Comparing predicted and measured growth rates across the 8 carbon sources shows that, in general, the predictions are poor. Some particular carbon sources are well predicted (e.g. glucose, fumarate, succinate) but the correlation between predicted and measured growth rates is relatively low (Pearsons r=0.39); furthermore, the variation in growth rates across the experimental data is much larger than that across the computational data (s=0.16 compared to s=0.7; s is the sample standard deviation). Thus, there are significant contributors to growth rate variation other than the necessary changes in metabolic pathway and proteome usage when the primary carbon source for growth is varied.

We next sequentially assess the contribution of changes in un- and under-utilized

protein to growth rate variation. In addition to the 8 carbon sources (circles), we can also assess the 4 carbon-limitation (triangles) and 3 environmental stress (squares) conditions that were proteomically profiled. In these additional environments, glucose is the carbon source, and other environmental changes affect proteome expression and growth. Accounting for un-utilized protein increases the correlation between predicted and measured growth rates to r=0.82 and increases the predicted growth rate variance to s=0.09. Finally, accounting for the variation in in vivo turnover rates increases the correlation to r=0.98 and results in a predicted growth rate variance that is similar to the experimental variance (s=0.15).

Thus, accounting for changes in the un- and under-utilized proteome explains much of the variation in growth rates across environments (Figure 3.2B, C)–microbial growth rates are largely determined by unused protein.

3.3.4 Defining the core and conditionally-utilized ME proteome segments

Why does E. coli allocate its proteome in a way that detracts from achieving its maximal growth rate in a given environment? To answer this question, we first systematically segment the proteome by protein function to gain insight into proteome allocation. We classify the ME proteome based on its condition-specific utility. While much of the expressed proteome is un-utilized in a given nutritional environment (Figure 3.1C), the un-utilized expressed proteome can be utilized in other environments the organism may encounter. To identify these conditionally-utilized proteins, we simulated growth under all growth-supporting Carbon, Nitrogen, Phosphorous, and Sulfur sources.

Comparing the totality of growth-supporting proteomes reveals a common core proteome that is utilized across all (minimal media) environments [144] (Figure 3.3A). This core proteome is largely involved in anabolism and protein synthesis (including

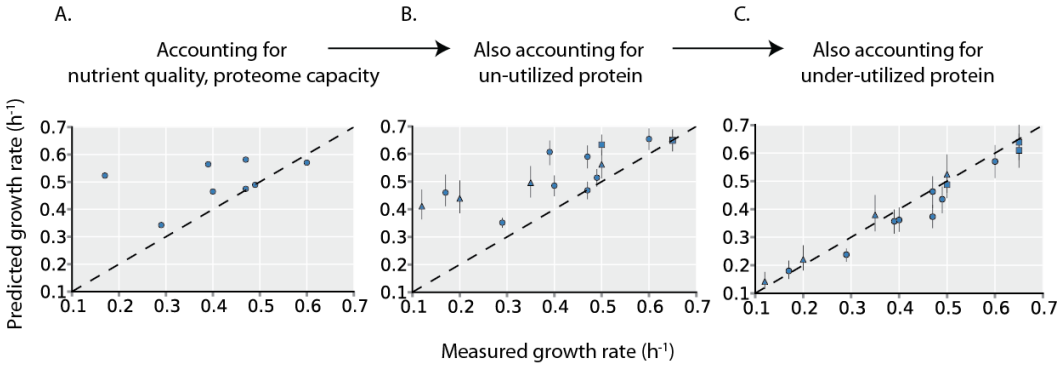


Figure 3.2: **Growth rate is determined by unused protein** A. Predicted growth rates are plotted versus measured growth rates during batch growth on 8 different carbon sources. Predicted growth rates are the computed maximal growth rates by the ME-Model, assuming the un-utilized proteome fraction and in vivo turnover rates are the same across all environments (see Methods, this assumption is eliminated in panels B and C to assess the effects of un- and under-utilized protein on growth). B. Predicted maximal growth rates are computed with the ME-Model with the un-utilized proteome fraction set to the values inferred from proteomics data (see Methods and Figure 3.1C). In addition to the 8 carbon sources (circles), 4 glucose-limited chemostat cultures (triangles) and 3 stress conditions (squares) are also shown. These additional growth conditions are not included in panel A as growth rate predictions would require further information. C. Predicted maximal growth rates are computed with the ME-Model with both the un-utilized proteome fraction and the in vivo turnover rate (indicative of under-utilized protein, Figure 3.1D) set to the values inferred from proteomics data. Point shape indicates environment type as in B (carbon source batch cultures = circles, glucose-limited chemostat cultures = triangles, stress conditions = squares).

the necessary transcription, translation, and protein folding enzymes). Importantly, the core proteome constitutes the vast majority of the utilized proteome in any particular environment.

In addition to the core proteome, there are conditionally-utilized proteins that we classify by element source (Carbon, Nitrogen, Phosphorous, and Sulfur) (Figure 3.3A). These proteome segments are largely catabolic. Due to the environmental specificity of these proteome segments, they are largely un-utilized in a particular minimal environment.

Thus, simulations with the ME-Model reveals a set of protein that constitute a core proteome utilized under all minimal media conditions [144], and comprising the vast majority of the utilized proteome mass in these environments. In contrast, the conditionally-utilized ME proteome is largely un-utilized in a particular environment, though certain proteins within these proteome segments are important for utilization of particular elemental sources. In the following sections, these proteome segments form the basis for understanding the fitness benefits and regulatory logic of expression of un- and under-utilized protein.

3.3.5 Regulatory logic of the under-utilized core ME proteome

The core proteome comprises 30-50% of the proteome mass in the conditions examined, and its abundance increases linearly with growth rate, consistent with higher biosynthetic demands (Figure 3.4A). This linear relation with growth rate has been observed for individual proteins within the core proteome [108].

However, while the core proteome abundance does increase with growth rate, it is over-expressed compared to biosynthetic demands (Figure 3.4B, as is evidenced by the non-zero y-intercept is Figure 3.4A). This overexpression of utilized protein was previously observed for ribosomal proteins, leading to the realization that translation rates are growth rate dependent [53, 110, 143]. At higher growth rates, the core proteome

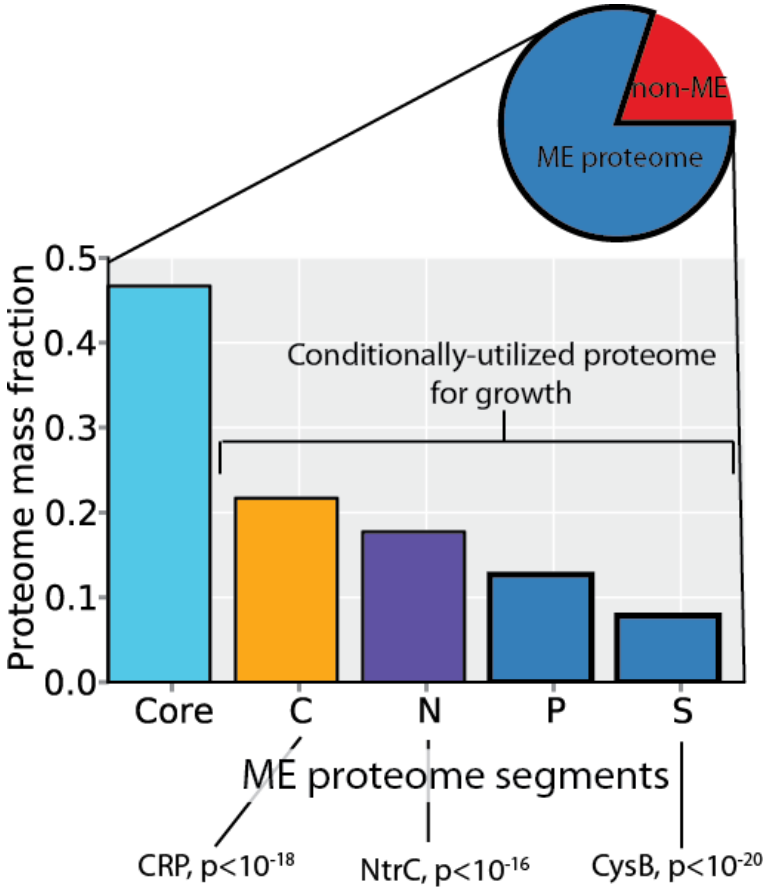


Figure 3.3: **Classification of the ME proteome into functional segments** The ME proteome encompasses 80% of the proteome by mass in glucose minimal media (pie chart). ME-Model growth simulations are used to define proteome segments (see Methods). The core proteome is comprised of proteins that are used in all minimal media environments simulated. The conditionally-utilized C-, N-, P-, S-proteome segments contain proteins that are expressed under at least one alternative carbon, nitrogen, phosphorous or sulfur source environments. A protein may belong to multiple conditionally-utilized proteome (C-, N-, P-, S-) segments (i.e., these segments are overlapping), but the proteins in the core proteome are unique to that segment. Abundance of these proteome segments in glucose minimal media batch culture is shown. Several global transcription factors have targets that are highly enriched ($p < 10^{-15}$) in the segments shown.

abundance approaches its demand, resulting in an increase in in vivo enzyme turnover at higher growth rates (Figure 3.1D). Thus, the over-expression of the core proteome underlies the previously observed under-utilized protein (Figure 3.1D).

While over-expression of the core proteome incurs a fitness cost on steady-state growth rates (Figure 3.2C), are there other fitness benefits to the expression of an under-utilized proteome? The over-expression of the core proteome may enable a fitness benefit upon encountering more favorable growth environments (Figure 3.4C). To demonstrate this effect, we simulate growth upon shifting from the lowest growth carbon source profiled (galactose) to the highest (glucose). If the core proteome is expressed in excess when growing in galactose, the organism grows faster upon the environmental up-shift; otherwise, the maximum instantaneous growth rate on glucose will be the same as that on galactose. This fitness benefit upon environmental up-shifts is consistent protein over-expression incurring a higher fitness cost upon environmental up-shifts (where the over-expressed core proteome becomes important for growth in the new environment) than down-shifts [145].

The over-expression of the core proteome, while a seemingly inefficient use of cellular resources under static environments, confers a fitness benefit upon environmental changes to alternate substrates. Thus, the allocation to the core proteome results in a trade-off between growth in the current environment and readiness for improved environmental conditions. As organisms vary quantitatively in how much the core proteome is overexpressed [146, 147], the regulatory logic of the core proteome reflects an organisms ecological strategy, that in turn reflects the organisms evolutionary history.

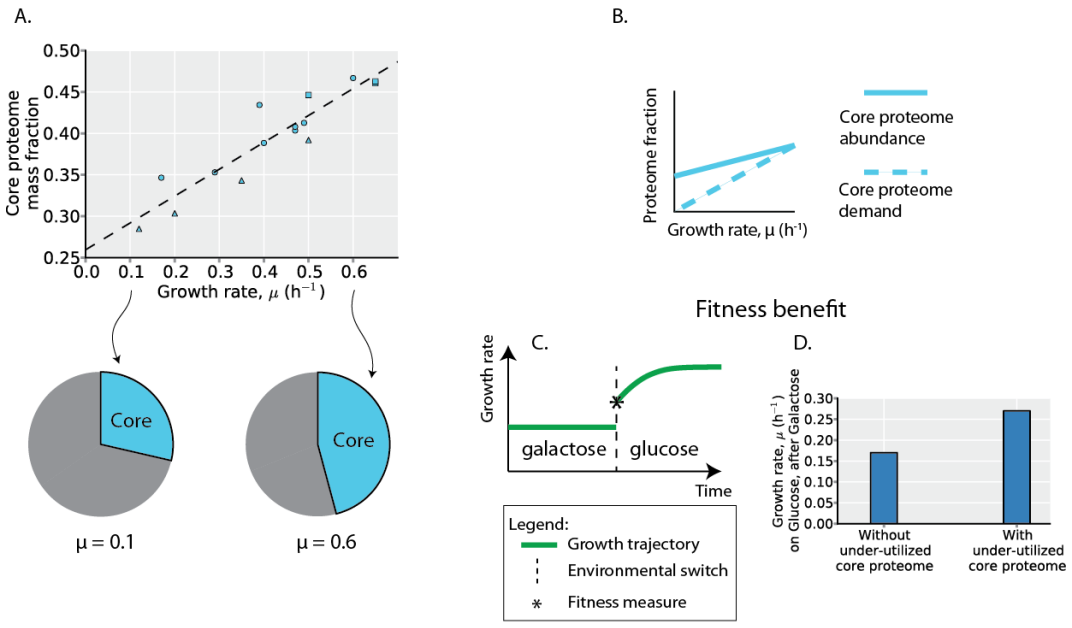


Figure 3.4: **Regulatory logic of the under-utilized core ME proteome** A. The core proteome mass fraction plotted as a function of growth rate in the profiled environments. Point shape indicates environment type (carbon source batch cultures = circles, glucose-limited chemostat cultures = triangles, stress conditions = squares). The dashed line is a linear regression (y = 0.33 x + 0.26, r2 = 0.85, p ¡ 10-5). The allocation to the core proteome at different growth rates is shown in the pie charts based on the regression. B. Depicted is the regulatory logic where the core proteome abundance is expressed at a level above its demands for growth at lower growth rates, resulting in under-utilized protein. C. Green line indicates organismal growth rate through an environmental shift. After an environmental shift, the instantaneous growth rate (asterisk; which is limited by the expressed proteome prior to the shift) is the measure of fitness used in D. D. The regulatory logic depicted in B confers a fitness benefit under changing environments. When shifting from galactose (the lowest growth carbon source profiled) to glucose (the highest growth carbon source profiled), having the core proteome initially in excess beyond the needs for biosynthesis enables a higher growth rate upon the nutrient up-shift (see Methods).

3.3.6 Regulatory logic of the conditionally-utilized, but un-utilized, ME proteome

Several of the conditionally-utilized proteome segments, show a large enrichment of transcriptional regulatory targets. The conditionally-utilized C-proteome is predominantly regulated by CRP, the N-proteome by NtrC, and the S-proteome by CysB (Figure 3.3). These global transcription factors are known to respond to changes in the availability of the corresponding nutrient sources [148, 149, 150], and are therefore likely important regulators of these proteome segments in response to environmental change.

In the conditions examined with quantitative proteomics, which predominantly comprise shifts and limitations in carbon sources, the conditionally-utilized C-proteome (regulated by CRP), decreases in expression at higher growth rates (Figure 3.5A) (as does the un-utilized proteome overall, Figure 3.1C). Thus, under growth limitation by carbon, the C-proteome becomes induced by CRP, as has been observed for individual proteins regulated by CRP [138].

The up-regulation of the conditionally-utilized proteome segments will depend on environments that affect the identified global regulators. To experimentally examine this effect, we obtained RNA-seq data from different degrees of carbon- (C-) and nitrogen- (N-) limitation. Under C-limitation, the C-proteome (regulated by CRP) is up-regulated, but the N-proteome is not (Figure 3.5B, top), as was observed in the proteomics data. In contrast, under N-limitation, the N-proteome (regulated by NtrC) is up-regulated but the C-proteome is not (Figure 3.5B, bottom).

Expression of the conditionally-utilized (but largely un-utilized) proteome segments incur a fitness cost to steady-state growth (Figure 3.2B), but the expression patterns (Figure 3.5C) point to a fitness benefit. The expression of the conditionally-utilized proteome, like the over-expression of the core proteome, likely enables readiness for

environmental change. To demonstrate this effect, we simulate nutrient supplementations under C- and N-limitation with the ME-Model.

Under C-limitation, supplementation with nitrogen sources provides no fitness benefit. Similarly, under N-limitation, supplementation with carbon sources provides no fitness benefit (Figure 3.5D). On the other hand, supplementation with alternative sources of the limiting element provides growth advantages (Figure 3.5D). Thus, expression of the conditionally-utilized C-proteome provides a fitness benefit under C-limitation but not under N-limitation (and vice versa for the conditionally-utilized N-proteome). While these proteome segments are largely un-unutilized in a current environment, they do provide a fitness advantage upon environmental shifts.

Thus, the regulatory logic for the conditionally-utilized C- and N-proteomes can be understood as providing a fitness benefit other than for steady-state growth; namely, that proteome allocation constraints result in a trade-off between growth in stable and variable environments. Such a trade-off should change under a selection pressure to compete in a stable environment. In a companion study [151], we show that this is indeed the case: evolutionary selection results in a substantial reduction in the expression of the un-utilized proteome coinciding with an increased fitness in stable environments and a reduced fitness under environmental shifts. The genetic basis for such a shift in proteome composition is surprisingly simple.

3.3.7 Functional composition and regulatory logic of the non-ME proteome

We have thus far focused on the function, composition, and regulation of the ME proteome. The non-ME proteome contains functions that are not as well classified as those in the ME proteome. Therefore, to understand the function and regulation of the non-ME proteome, we first manually classify the proteins by function using annotations

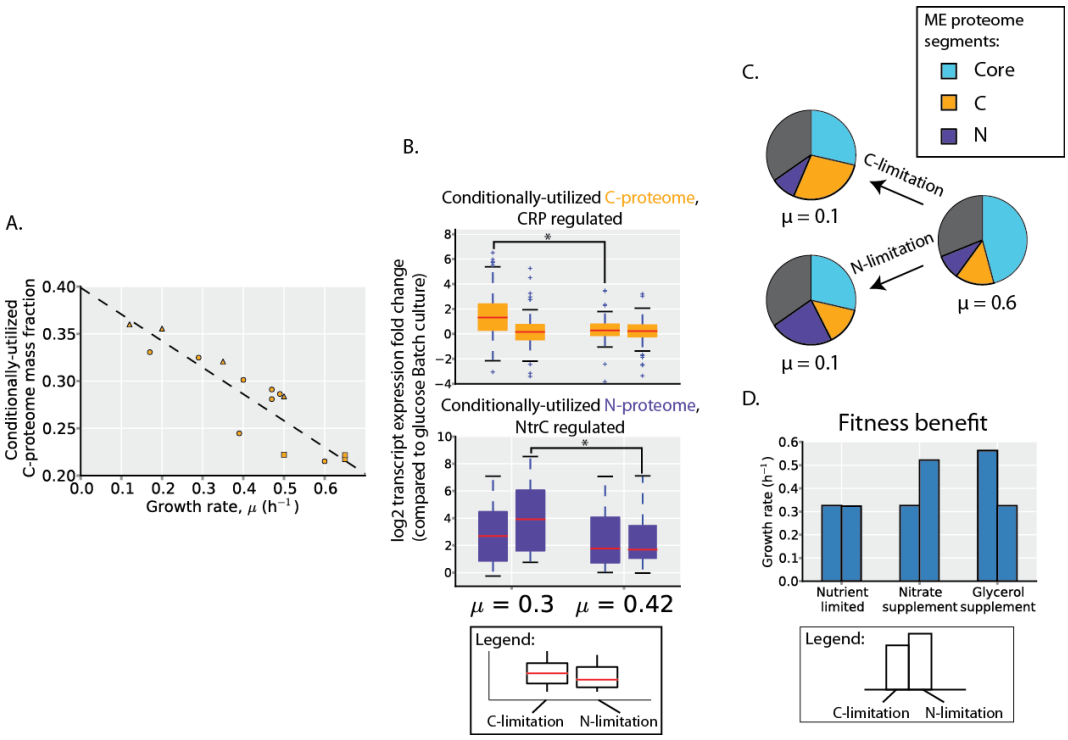


Figure 3.5: **Regulatory logic of the conditionally-utilized, but un-utilized, ME proteome** A. The proteome mass fraction of the conditionally-utilized C-proteome is plotted as a function of growth rate in the profiled environments (which predominantly comprise carbon source shifts and carbon limitation). The dashed line is a linear regression (y = -0.28 x + 0.40, r2 = 0.81, p ¡ 10-4). B. Boxplots indicate transcript expression fold changes under carbon-limited (left) and nitrogen-limited (right) growth at two different growth rates. The conditionally-utilized C-proteome regulated by CRP is up-regulated under stronger carbon-limitation (top), and the conditionally-utilized N-proteome regulated by NtrC is up-regulated under stronger nitrogen-limitation (bottom). Asterisks indicate a significantly different average fold change, p¡0.01. C. Depicted is a regulatory logic where the conditionally-utilized C-proteome (orange) is up-regulated under carbon-limitation (top), but not nitrogen-limitation (bottom); the conditionally-utilized N-proteome (purple) is up-regulated under nitrogen-limitation, but not carbon-limitation (as in Figure 3.5B). The size of the proteome segments are determined based on the linear regression of the core proteome (Figure 3.4A) and conditionally-utilized C-proteome (Figure 3.5A); at =0.6, the mass of proteins that are in both the conditionally-utilized C- and N-proteomes is evenly divided across the two segments; under N-limitation, the N-proteome is assumed to have the same quantitative trend as does the C-proteome under C-limitation, based on the differential expression observed in the transcriptomics data (Figure 3.5B). D. The regulatory logic of the conditionally-utilized C- and N-protomes (Figure 3.5C) confers a fitness benefit under environmental shifts: when C-limited, readiness for alternative carbon sources confers a fitness benefit whereas readiness for alternative nitrogen sources does not (and vice versa when the organism is N-limited; see Methods).

and descriptions present in EcoCyc [8]. Much of the non-ME proteome mass can be classified by focusing on its most abundant proteins.

We therefore classify the most abundant non-ME proteins that together comprise at least 80% of the non-ME proteome by mass across all 15 conditions examined. The abundant non-ME functions include replication, regulation, stress responses, and proteins of unknown function (encoded by so called y-genes). In glucose minimal media, the most abundant non-ME proteome functions are regulatory proteins and proteins of unknown function (Figure 3.6A).

Overall, the non-ME proteome is slightly more abundant at lower than higher growth rates (Figure 3.6B). However, as the non-ME proteome contains a variety of cellular functions, some non-ME proteome functions are positively correlated with growth rate, whereas others are negatively correlated with growth rate (Figure 3.6C). Broadly speaking, the functions that are positively correlated with growth rate are those related to growth, including cell division, replication, proteostasis, and protein translocation. The functions that are negatively correlated with growth rate are those related to stress resistance and survival, including osmotic, acid, and oxidative stresses.

Thus, at higher growth rates, the growth-related functions (though they are not yet formalized in the ME-Model) are more highly expressed, and at lower growth rates, several stress resistance functions are more highly expressed. Indeed, several studies have shown that slower growing cells are more resistant to cellular stresses [152, 153, 154]. Therefore, proteome allocation constraints result in a trade-off between growth and stress resistance.

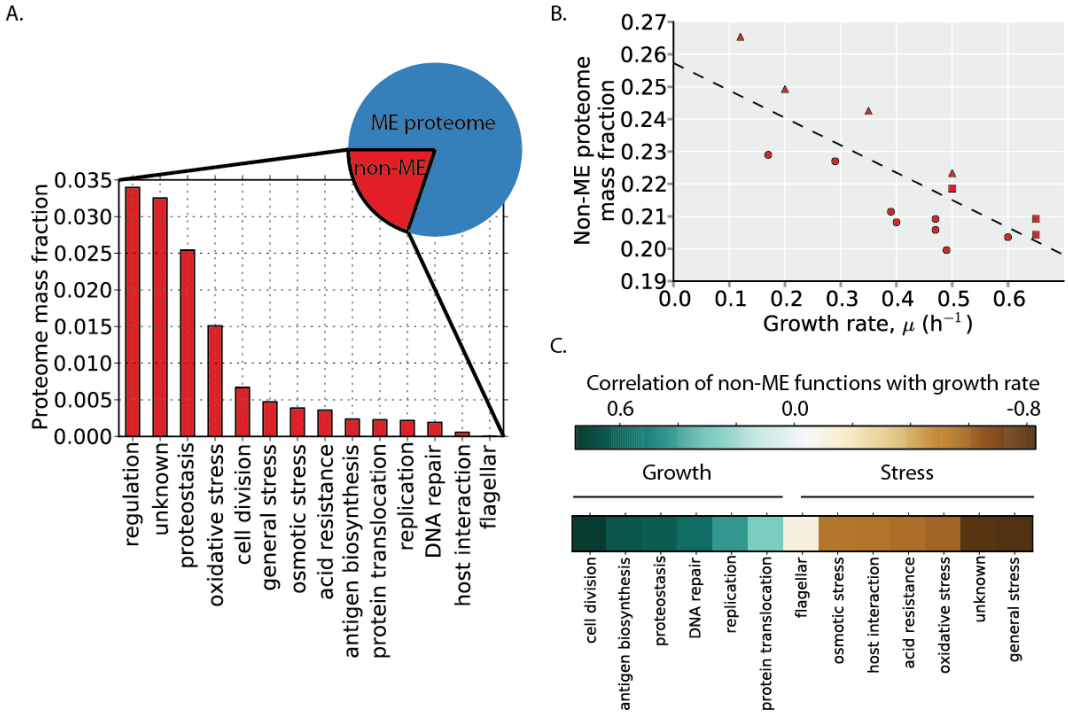


Figure 3.6: **Growth versus stress regulatory logic** A. The non-ME proteome encompasses 20% of the proteome by mass in glucose minimal media (pie chart). The functional composition and abundance of the non-ME proteome is shown based on proteomics data in glucose minimal media. B. The proteome mass fraction of the non-ME proteome is plotted as a function of growth rate in the profiled environments. The dashed line is a linear regression (y = -0.09 x + 0.26, r2 = 0.53, p ¡ 0.01). Point shape indicates environment type (carbon source batch cultures = circles, glucose-limited chemostat cultures = triangles, stress conditions = squares). C. The median correlation of the proteins in each non-ME function is shown in the heatmap in rank order. While the overall non-ME proteome fraction is larger at higher growth rates, the trend depends on the specific function. Generally, functions positively correlated with growth rate (blue) are associated with biosynthetic and growth functions whereas functions negatively correlated with growth rate (brown) are associated with stress resistance.

3.4 Discussion

The composition of an expressed proteome reflects a microbes resource allocation. Deep coverage proteome data sets under multiple growth conditions and ME-Model enabled analysis yields fundamental insights into the regulatory logic of proteome allocation. Here, we have taken the first steps to elucidate the systems biology of the prokaryotic proteome and the evolutionary determinants of its composition.

3.4.1 Proteome allocation constraints demand fitness trade-offs

By classifying the proteome not actively contributing towards growth by function, we gain insight into the ecological and evolutionary forces shaping the composition of the proteome. Lower growth states have a lower growth-associated proteome, but a higher environmental readiness and stress resistance proteome (Figure 3.7). In addition to growth in a given environment, environmental readiness and stress resistance are important functions for organismal fitness. As the proteome is limited in size, its allocation results in a trade-off between optimal growth performance and readiness to deal with environmental change.

The evolutionary trade-off between growth rate, environmental readiness, and stress resistance is consistent with several studies that have documented an anti-correlation in these fitness measures across strains [155, 156, 157]. Now, we can understand the origin of these phenotypic trade-offs at a mechanistic and transcriptional regulatory level on a genome-scale basis as arising from constraints on proteome allocation.

Indeed, in an accompanying paper [151], we show how adaptive regulatory mutations in a constant environment result in a reallocation of the proteome towards growth and away from hedging functions. The evolved strain is specialized for growth in the constant environment, exhibiting faster growth rates in steady-state environments but

impaired fitness upon environmental shift and stress related shocks.

Species will vary in their relative weightings of the different components of organismal fitness (growth, environmental readiness, stress resistance), resulting in different regulatory logic and ecological strategies. For example, while the strain of E. coli profiled here regulates its core proteome to decrease at lower growth rates (though still at a level greater than is demanded for biosynthesis), other organisms have a constant, un-regulated expression of ribosomal proteins [146, 147]. The constant expression of the core proteome would likely result in higher environmental readiness at the cost of lower maximal growth rates. These results give a strong impetus for comparative proteome allocation studies.

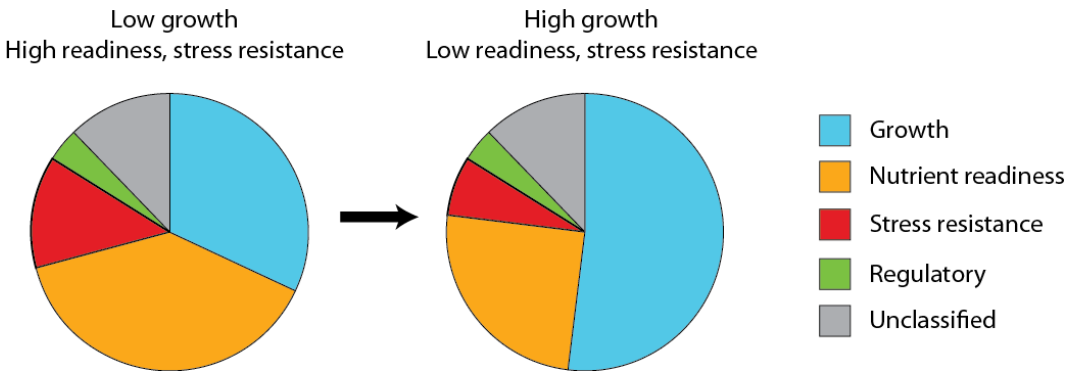


Figure 3.7: **Proteome allocation constraints result in fitness trade-offs** The pie charts summarize the global proteome classification and how allocation to the proteome segments varies across environments. In environments with lower growth rates, the proteome allocated towards growth is lower, but the proteome allocated to nutrient readiness and stress resistance is higher. As the proteome is a limited resource, proteome allocation to the different segments results in fitness trade-offs between growth, nutrient readiness, and stress resistance.

3.4.2 Evolutionary history is a primary determinant of microbial growth rates

An important implication of the identified fitness trade-offs is that cellular growth rates may be primarily determined by environmental history, rather than nutrient quality (i.e., the maximum growth rate possible with a specified nutrient, Figure 3.2A). As growth rates are determined by the proteome allocated towards growth, evolutionary and ecological factors may be more important than the identity and quality of the substrates themselves. For example, while glucose and galactose have similar inherent qualities as sole carbon sources (chemically and as defined by the growth potential of E. coli on these substrates), glucose is the highest growth substrate in this dataset and galactose is the lowest. Galactose may be a more rarely or transiently encountered carbon source or perhaps it is often associated with harsher environments, which would make environmental readiness and stress resistance comparatively more important components of overall cellular fitness [158, 159, 160].

3.4.3 The proteome burden of a generalist species

While we broadly classify the proteome here based on nutrient readiness and stress resistance functions, it is important to realize that within these proteome segments is a variety of distinct functions enabling readiness for specific nutrients and stresses. Concerted up-regulation of these segments results in a general nutritional readiness and stress resistance. However, as an organism becomes prepared for a wider array of nutrients or stresses, these proteome segments must become larger. Therefore, in addition to a trade-off between the different components of organismal fitness identified here (growth, environmental readiness, stress resistance), proteome allocation constraints also result in a trade-off in specialist versus generalist ecological strategies. E. coli

is a generalist species, capable of growing in a variety of environments. Its broad environmental niche results in large proteome burden. On the other hand, specialist species (capable of growing on a narrower range of substrates) would require a smaller proteome allocation to be equally ready for environmental change. Proteome allocation constraints will therefore also result in a trade-off between specialist and generalist strategies [161].

3.5 Experimental Procedures

3.5.1 Proteomics dataset and normalization

The proteomics data was obtained from Schmidt et al. and growth rates were obtained from [162]. The dataset contains protein counts per cell for Escherichia coli K-12 BW25113 grown in 15 different environments. The environments include batch culture with 8 different carbon sources (glucose, galactose, acetate, glycerol, glucosamine, fumarate, succinate, pyruvate) as the sole carbon substrate, 4 glucose-limited chemostat cultures (=0.12, =0.20, =0.35, =0.5), and 3 stress environments (high temperature [42C], acid stress [pH 6], and osmotic stress [50 mM NaCl]); 2024 proteins are quantified. For all analysis here, the protein copy numbers were transformed to mass fractions using the protein molecular weights. While the proteomics data does not cover all proteins in the E. coli proteome; using previously published ribosome profiling data [163], we estimate that the proteomics data in glucose minimal media batch culture covers 94% of the proteome mass.

3.5.2 Quantifying the utilized and un-utilized proteome

To identify sets of proteins that can be utilized under a particular environment, we sample ME-Model enzymatic rate parameters; we then identify growth rate optimizing proteomes (for each parameter set) based on the growth-maximizing procedure outlined in OBrien et al. 2013. We independently sampled all enzymatic rates in the ME-Model based on the global distribution of kcat across all enzymes; we used a (base 10) lognormal distribution with mean =1.11 and 2=1.31, based on data from [164]. For each carbon source present in the proteomics dataset, we performed 100 samples with the defined nutrient availability in the ME-Model; these simulations result in 100 sets of proteins that are predicted to be utilized in that environment [144], and the abundance of these protein sets are interrogated in the proteomics data to obtain utilized and un-utilized proteome fractions (Figure 3.1B, 3.1C).

3.5.3 Quantifying the under-utilized proteome

For all proteins in the core proteome (see ME proteome classification), the ME-Model was used to predict the protein demand (for cell growth). Protein demand is defined as the protein abundance predicted by setting the growth rate to its measured value and maximizing the expression of an unmodeled protein. Taking the ratio of the protein demand (i.e., model-predicted protein abundance) and the measured protein abundance, then gives a measure of the protein utilization. For each protein, to get the relative in vivo turnover for that protein (on a scale from 0 to 1), this ratio was then normalized by the maximum value (of the ratio) for that protein across all profiled environments (Figure 3.1D).

3.5.4 Growth rate predictions

Maximum growth rates are determined with the computational procedure described in OBrien et al. 2013. Unused protein fraction and mean in vivo enzyme activity are changeable variables in the ME-Model that affect predicted growth rates. The values of these 2 variables inferred from the proteomics data (Figure 3.1) are set in the ME-Model to assess their effect on growth rates (Figure 3.2). When the unused protein fraction and mean in vivo enzyme are kept constant across all environments (Figure 3.2A), they are set to the values inferred from glucose batch culture.

3.5.5 ME proteome classification

All growth-supporting minimal media were simulated with the ME-Model. The minimal media were defined by starting from the default glucose M9 medium (with ammonium as a nitrogen source, phosphate as a phosphorus source, and sulfate as a sulfur source) and individually changing all carbon, nitrogen, phosphorous, and sulfur sources. In total, 333 environments were simulated, corresponding to 180 carbon, 93 nitrogen, 49 phosphorus, and 11 sulfur sources. Isozymes were required to be used in equal abundance. All expressed proteins were identified as those with non-zero translation fluxes. Proteins expressed across all simulated environments were considered the core proteome. Proteins not in the core, but expressed under certain alternative Carbon, Nitrogen, Phosphorous, or Sulfur sources are considered in the conditionally-utilized C-, N-, P-, and S-proteome segments, respectively. A protein may belong to more than one conditionally-utilized proteome segment.

3.5.6 non-ME proteome classification

A subset of the proteins outside of the scope of the ME-Model (i.e., the non-ME proteome) were manually classified by function. For each environment, the most abundant proteins, comprising at least 80% of the non-ME proteome mass were annotated based on descriptions from EcoCyc [8].

3.5.7 Fitness benefit simulations for the under-utilized core proteome

Steady-state growth was first simulated with the ME-Model in the initial environment (galactose batch culture). Two scenarios were considered: one in which the in vivo enzyme turnover was equal to that measured in the proteomics data in galactose batch culture and the other in which the in vivo enzyme turnover was equal to that measured in the second environment (glucose batch culture). The predicted protein abundance to support growth for all proteins in the core proteome were obtained from the simulation output. Then, the maximal growth rate after an environmental shift (to glucose batch culture) was computed subject to the expression level of the core proteome expression prior to the shift for both scenarios.

3.5.8 Fitness benefit simulations for conditionally-useful ME proteome

The uptake rate of glucose and ammonium was limited in glucose minimal media to simulate carbon (C-) and nitrogen (N-) limited growth by limiting the uptake reaction flux. Then, subject to the glucose and ammonium uptake limitations, additional carbon (glycerol) and nitrogen (nitrate) sources were supplied in excess and maximal growth rates predicted.

3.5.9 Chemostat cultivation

Carbon and nitrogen limited chemostats were carried out in a 1.3 L Bioflo 110 fermentor (New Brunswick Scientific, NJ) with a 0.7L of working volume. For carbon limitation M9 media was supplemented with 4 g/L of glucose. For Nitrogen limitation same media was limited in nitrogen by adding a reduced amount of ammonium chloride (10 mM). Dilution rates (0.26, 0.31, 0.44 and 0.56 h-1) were controlled by adding and removing media at the same rate with a peristaltic pump. Steady state was achieved after 3-5 residence times and was verified by biomass measurements.

3.5.10 RNA-seq libraries

Samples for RNA-sequencing were taken in mid log phase of batch cultures or during the steady-state in chemostats. Cells were collected with Qiagen RNA-protect Bacteria Reagent and pelleted for storage at -80C prior to RNA extraction. Cell pellets were thawed and incubated with Readylyse Lysozyme, SuperaseIn, Protease K, and 20% SDS for 20 minutes at 37C. Total RNA was isolated and purified using the Qiagen RNeasy Mini Kit columns and following vendor procedures. An on-column DNase-treatment was performed for 30 minutes at room temperature. RNA was quantified using a Nano drop and quality assessed by running an RNA-nano chip on a bioanalyzer. Paired-end, strand-specific RNA-seq was performed following a modified dUTP method [165]. The rRNA was isolated using Epicentres Ribo-Zero rRNA removal kit for Gram Negative Bacteria. RNA-seq was performed using a modified dUTP method [165].

3.5.11 Transcriptome analyses

The obtained reads were mapped to the E. coli MG1655 genome (NC_000913.2) using the short-read aligner Bowtie [166] with two mismatches allowed per read align-

ment. To estimate gene expression FPKM values were calculated using cufflinks tool and differential expression analysis was carried out using cuffdiff feature of the same package using the upper quartile normalization [167].

3.6 Acknowledgements

We thank and Laurence Yang, Justin Tan, Zak King, Ali Ebrahim, Daniel Zielinski, and Joshua Lerman for valuable discussions. EJO was supported by NIH GM057089. The Novo Nordisk Foundation supported this work. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Author contributions: EJO and BOP conceived the study. EJO performed computational simulations and analysis. JUC performed chemostat culture, RNA-seq experiments, and phenotypic assays. EJO, JUC, and BOP wrote and edited the manuscript.

The text of Chapter 3 is a full reprint of the material as it appears in: OBrien E.J., Utrilla J., Palsson B.O. Proteome allocation constraints determine cellular growth rates and demand fitness trade-offs, Submitted. The dissertation author was the primary author of the manuscript. The other authors were Jose Utrilla and Bernard Ø. Palsson.

Chapter 4

Observed fitness plateaus in microbial adaptation result from tradeoffs in proteome complexity

Problems with many solutions are the rule rather than the exception in living systems.
—Andreas Wagner

4.1 Abstract

In 1932 Sewall Wright introduced the now familiar fitness landscape comprised of peaks and valleys [168]. Since then, theoretical developments have suggested that the fitness landscape should include plateaus across which equally fit alternatives exist [169, 170, 48, 44, 171]. Using a recently developed genome-scale in silico model of metabolism and proteome synthesis for Escherichia coli [53], we uncover a fitness plateau that is found across all environments examined. Well-defined and experimentally testable alternative optimal physiological, metabolic, and proteome states are predicted to exist on the fitness plateau. Adaptive laboratory evolution (ALE) shows that strains evolve

towards the predicted fitness plateaus in many environments. Furthermore, using C13 flux and mRNA expression profiling, we show that the phenotypic diversity observed on a plateau matches model predictions. Fundamentally, the range of equivalent evolutionary outcomes is characterized by the complexity of the expressed proteomes, reflected in a trade-off between metabolic rate and biomass yield. At the extremes, nutrient-efficient strategies (that optimize for metabolic yield) require complex diversified proteomes, in contrast to proteome-efficient strategies (that optimize for metabolic rate), which require simple and streamlined proteomes. Thus, predicted alternative equivalent optima exist and a fundamental principle emerges describing the relationship between proteome complexity and organismal physiology.

4.2 Results and Discussion

While a large number of genetic perturbations are neutral [172, 173], it is not clear how pervasive are alternative phenotypic states of the same fitness. While several theoretical models have predicted alternative optimal phenotypes to be pervasive [170, 48, 171], empirical examples with a mechanistic basis for equivalence in organism function are lacking. Here, we reveal a family of alternative optimal growth phenotypes and show that they arise due to a trade-off in proteome complexity.

We use growth rate (μ) as a selection pressure and thus a measure of fitness. This evolutionary objective is a readily selectable trait by ALE, enabling theoretical predictions to be directly tested. Growth predictions are made with a genome-scale model of metabolism and protein expression for E. coli, termed a ME-Model [53]. The ME-Model is based on reconstructed genome-scale metabolic [98] and proteome synthesis [90] networks and it allows computational prediction of optimal physiological, metabolic, and proteomic phenotypes subject to genetic and environmental parameters.

The ME model thus computes proteome composition and allocation on a genome-wide basis.

We first investigated the physiological phenotypes of substrate uptake rate, q (i.e., the rate carbon substrates are consumed by a cell), and biomass yield, Y (i.e., the efficiency of conversion of carbon into biomass). By varying the substrate uptake rate in the ME-Model a range of values for q appear that can support an optimal growth rate (Fig. 4.1A). The computed fitness maximum is a broad plateau rather than a sharp peak.

The fitness plateau is found to be a pervasive characteristic across all environments examined. The features of the plateau change with the genetic background and environmental conditions considered (Fig. 4.1A). For example, in a strain with the electron transfer system (ETS) removed, the optimal growth rate decreases and the range of substrate uptake rate across the plateau increases (Fig. 4.1A, triangles). These shifts were experimentally observed [174]. Plateaus of optimal growth fitness are a property of the genome-scale ME-model.

We used growth rate selection with ALE to determine if strains evolve toward the predicted fitness plateaus. The data set spans a variety of environmental and genetic perturbations, including carbon source shifts, gene knockouts, and both aerobic and anaerobic growth [49, 175, 176, 177, 33, 174]. In un-evolved strains, experimental uptake rates, q, correlate with computed uptakes with moderate accuracy (PCC=0.7, p¡10-2, Fig. 4.1B) and are quantitatively low (RMSD=9.05). However, after growth-rate selection with ALE, q approaches the predicted optimum (PCC=0.83, p¡10-4, RMSD=4.03, Fig. 4.1C). Biomass yields, Y, are well predicted for both un-evolved and evolved strains. Thus, we observe adaptive evolution towards the predicted fitness plateau and quantitative agreement with model predictions.

The fitness plateau, opposed to a fitness peak, implies the existence of alternative optimal phenotypes. For the same environment, different physiological, metabolic,

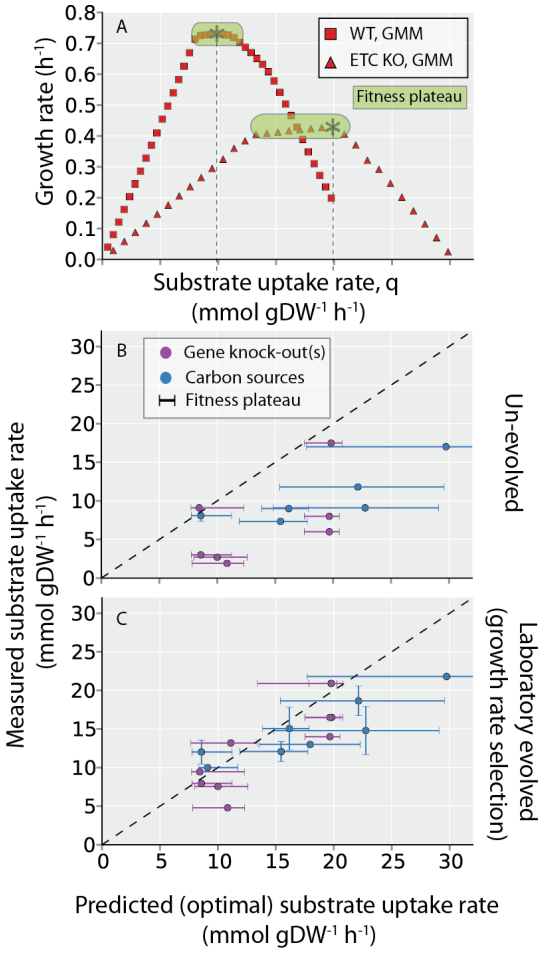


Figure 4.1: **Evolution to predicted optimal metabolic rate.** A) The maximum growth rate as a function of the substrate uptake rate (q) as computed with an E. coli ME-Model [53] is shown, demonstrating the presence of a fitness plateau. The simulated fitness plateau results in a predicted range of optimal q (shown in green). The fitness plateaus for wild-type (WT, squares) and an electron transport chain knockout strain [174] (ETC KO, cydAB-cyoABCD-cbdAB-ygi) on glucose minimal media are shown; growth rate is predicted to decrease and uptake rate predicted to increase as observed in this strain [174] and in anaerobic conditions. B) Measured substrate uptake rates for un-evolved E. coli after genetic (gene knockouts, purple dots) and environmental (carbon source shifts, blue dots) perturbations are predicted with moderate correlation for carbon source shifts and poor correlation for gene knockouts. C) The same plot as B) after adaptive laboratory evolution (ALE) to select for growth rate maximizing phenotypes; measured substrate uptake rates match predicted optimal rates. Error bars on the y-axis indicate standard deviations across independently evolved strains where reported. Error bars on the x-axis indicate the predicted optimal q range in the fitness plateau (defined to be within 95% of the maximum growth rate achieved). WT=wild-type, GMM=glucose minimal media, ETC KO=electron transport chain knockout. Asterisk indicates maximum growth phenotype. Dashed line indicates where prediction equals measurement.

and proteomic states are predicted to support the same growth rate. Is the diversity of predicted phenotypic states experimentally observed? To answer this question, we investigate the diversity of phenotypes observed across strains independently selected for maximum growth rates in constant environments.

On the fitness plateau, a range of q is predicted to support optimal growth rate. Across the optimal q values, a specific range of Y values is predicted that is inversely correlated with q (Fig. 4.2A). Thus, a q-Y tradeoff is computed. A q-Y tradeoff is repeatedly observed across replicate endpoint strains independently evolved in a constant nutritional environment (Fig. 4.2A). In addition to strains evolved in glucose minimal media [178] and at an elevated temperature [179], we evolved several replicate strains with the ETS eliminated and also observed a q-Y tradeoff.

The fitness plateau (and the q-Y tradeoff) is associated with changes in by-product secretion, commonly referred to as overflow metabolism [180, 181, 182]. In glucose minimal media, a higher glucose uptake is associated with higher acetate production with a relationship that quantitatively matches that predicted by the ME-Model (Fig. 4.2B); in the ETS knockout strain, a higher glucose uptake is associated with higher lactate production. Thus, a series of experimental evolutions give organismal-level phenotypes that are consistent with what was computed with the ME-model.

We then looked at the systems-level properties of the alternative optimal endpoints. Across the fitness plateau in glucose minimal media [178], the ME-Model predicts systematic shifts in proteome allocation corresponding to well-defined changes in metabolic pathway utilization (Fig. 4.3A,C). At low q (and high Y), complete oxidation of glucose through the oxidative tricarboxylic acid (TCA) cycle and the ETS are predicted; this pathway is nutrient efficient (i.e., high Y), but also has a large protein requirement (due to the many catalytic steps and large protein complexes). As q increases (and Y decreases), TCA cycle use is predicted to decrease, and use of EmbdenMeyerhofParnas (EMP)

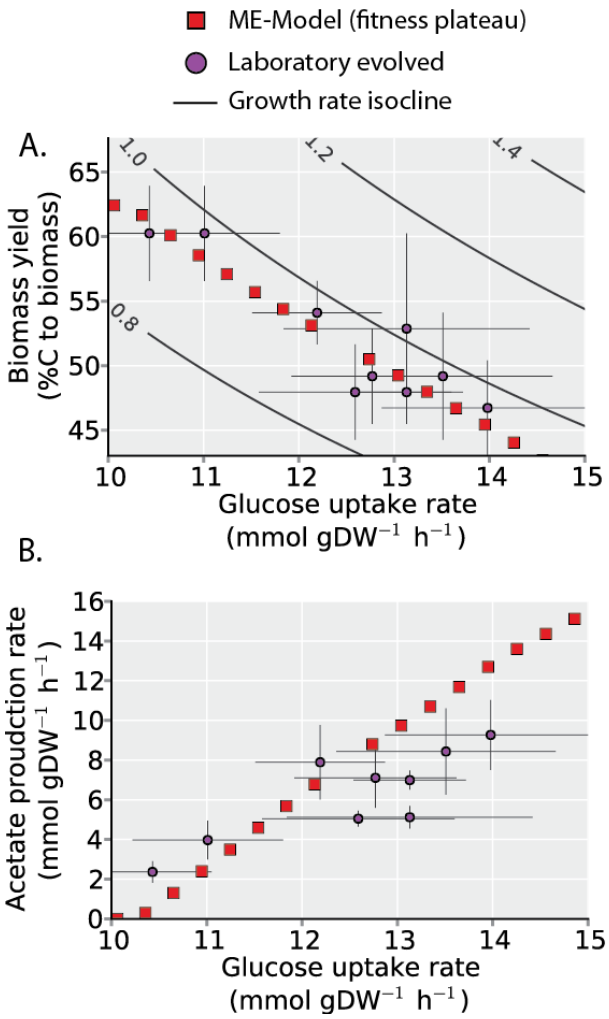


Figure 4.2: **Rate-yield tradeoff across the fitness plateau in a fixed environment.** A) The predicted tradeoff between biomass yield (Y) and glucose uptake rate (q) in the ME-Model fitness plateau (red squares) is evident in independent laboratory evolved strains (purple circles) in glucose minimal media [178]. B) Across the fitness plateau, the linear increase in acetate production is also evident in laboratory-evolved strains. Error bars indicate 95% confidence intervals across 3 biological replicate measurements.

glycolytic pathway is predicted to increase.

Though these are the dominant pathway shifts, the ME-Model additionally predicts that several other pathways can support maximal growth through higher q and lower Y, including the EntnerDoudoroff (ED) glycolytic pathway, alternative uptake systems (hexokinase), alternative acetate production reactions (pyruvate oxidase), and the glyoxylate cycle and non-oxidative pentose phosphate pathway (Fig. 4.3A,C).

Generally, the pathway shifts predicted across the fitness plateau (in glucose minima media) progress from more complex proteomes (used in high Y pathways; ETS and TCA cycle) to simpler proteomes (used in high q pathways; ED and EMP). Thus, constraints on proteome allocation underlie the model-predicted shifts.

We then sought to determine if the predicted shifts in pathway use are experimentally observed. We examined the transcriptome and fluxome of the two evolved strains with the highest q and highest Y (Fig. 4.3). In the transcriptome, ED and EMP pathway genes are up-regulated and TCA cycle and oxidative phosphorylation genes are down-regulated in the high q strain compared to the high Y strain (Fig. 4.3B). The primary metabolic flux differences between the strains are higher EMP fluxes and lower TCA fluxes (Fig. 4.3C-E), and a large change in the flux split at acetyl-CoA. The ED pathway and several of the alternative predicted pathways were found to be active but not highly utilized in either strain. Therefore, molecular expression and pathway flux phenotypic data from evolved strains are consistent with the computationally predicted shifts in proteome allocation across the fitness plateau.

A key prediction by the ME-Model across the fitness plateau is that both substrate-level phosphorylation and oxidative phosphorylation can support the same cellular growth and neither is strictly required. It has been shown that overflow metabolism can be nearly eliminated without any effect on growth [183]. The additional observation that the same growth rate can be achieved with very little flux through the oxidative TCA cycle in

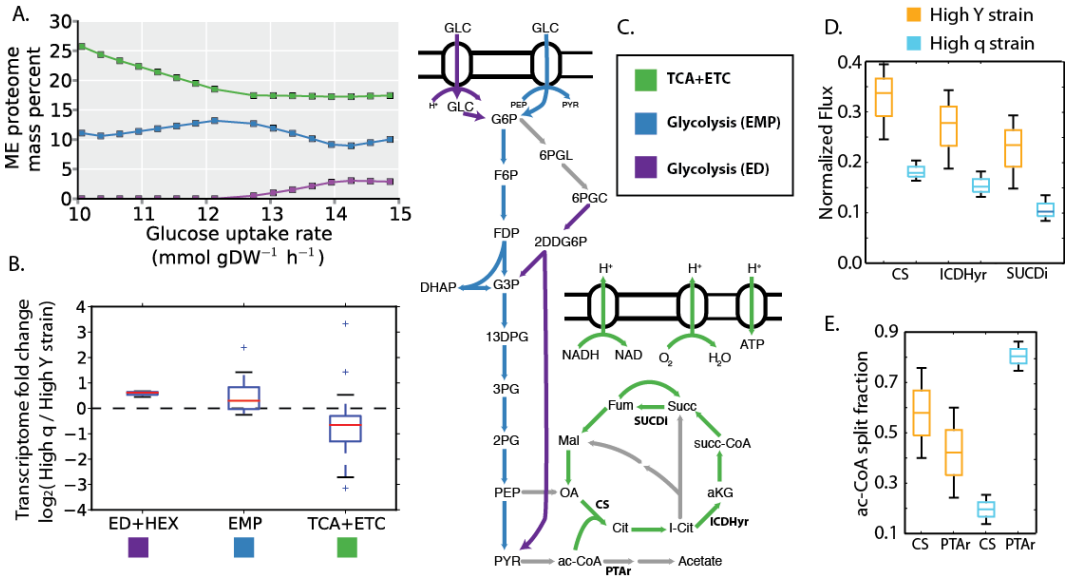


Figure 4.3: **Alternative proteomic and pathway use across the fitness plateau.** A) The pathway shifts predicted to support maximal growth across the fitness plateau are shown in terms of proteome mass percentage. B) The fold change in transcript expression in the evolved strain with highest q (and lowest Y) compared to the strain with the highest Y (and lowest q) are shown. Genes are grouped according to the identified pathways and are consistent with the direction in change predicted by the ME-Model. C) The identified pathways predicted to shift (with colors corresponding to 3A,B) are shown. D) Metabolic fluxes inferred from C13-MFA for key reactions in the TCA cycle in the high Y and high q strains are consistent with the predicted proteome and measured transcriptome shifts. E) The flux differences between the two strains are also reflected in the divergent flux split around acetyl-CoA. Little difference is observed in ED flux for these particular strains. For the box plots, whiskers are 95% confidence intervals, boxes are 68% confidence interval, and the lines are the mean values. TCA=tricarboxylic acid cycle, ETC=electron transport chain, EMP=EmbdenMeyerhofParnas, ED=EntnerDoudoroff, HEX=hexokinase.

evolved strains (Fig. 4.3) confirms the broad range of alternative optimal pathways.

Finally, we sought to determine the generality of the shifts in proteome complexity and allocation and how they relate to cellular physiology. In addition to a q-Y tradeoff in alternative optimal phenotypes in the same environment, the ME-Model predicts there is a q-Y tradeoff across different environments (Fig. 4.4A, left). While wild-type strains do not show a clear anti-correlation between q and Y across environments, the q-Y tradeoff clearly emerges once growth-optimizing phenotypes have been selected with ALE (Fig. 4.4B, right).

Like the fitness plateau in a given environment, the q-Y tradeoff is due to the inherent relationship between pathway yield and proteome requirements. We predict a negative correlation between q and catabolic proteome mass (rs = -0.77, p¡10-61; Fig. 4.4B). In general, higher Y proteomes have several complicating features (Fig. 4.4C). These include a higher number of genes expressed, more sophisticated enzyme complexes, more extensive use of prosthetic groups, and more complex mechanisms for transcriptional regulation. Conversely, they require a lower average expression level of the proteins involved. The complex proteomes are more evolutionarily sophisticated.

Taken together, these results show that there is a fundamental relationship between cell physiology and the underlying catabolic proteome. Proteomes that optimize for metabolic yield (i.e., nutrient efficiency) require more complex and costly proteomes. On the other hand, strategies that optimize for metabolic rate (i.e., proteome efficiency) utilize simpler, more streamlined proteomes.

In a constant environment, variation in proteome complexity and allocation defines a class of alternative phenotypes of the same fitness, resulting in a plateau in the fitness landscape. Like neutral genotypes [172, 173], alternative optimal phenotypes will affect both short- and long-term evolutionary dynamics. The ALE experiments detailed here show that short-term evolutionary diversification is indeed affected. It is intriguing

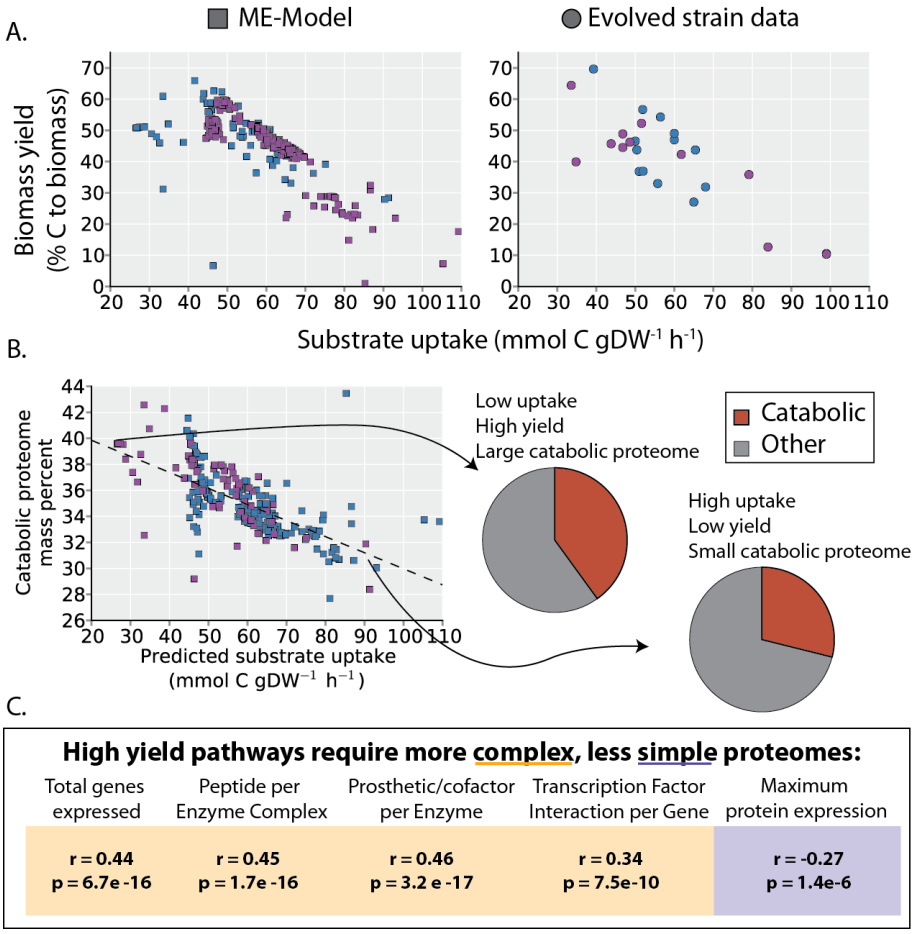


Figure 4.4: **Proteome complexity underlies the rate-yield tradeoff across environments.** A) Biomass yield (Y) and substrate uptake rate (q) are inversely correlated across simulated (left, squares) and ALE (right, circles) data. All growth-supporting single carbon sources and non-essential single and double gene knockouts in central carbon metabolism are simulated with the ME-Model; only non-redundant data points are plotted (see Methods). As in Figure 4.1, carbon source shifts are shown in blue and gene knockouts are shown in purple. B) The catabolic proteome fraction required to support each simulated genetic and environmental perturbation is inversely correlated with substrate uptake rateslow carbon uptake and high carbon yield phenotypes use a large catabolic proteome, whereas high carbon uptake and low yield phenotypes use a smaller catabolic proteome. C) High yield (low uptake rate) phenotypes require a more complex (orange) and less simple (purple) proteome according to a variety of metrics computed from the proteome required in model simulations. Spearman correlation coefficient and p-value between yield and the labeled property are shown.

to note that proteins from more complex pathways are thought to have appeared more recently in evolutionary history [184, 185, 186]; complex proteomes may prove to enable long-term evolvability [187, 188].

4.3 Methods

4.3.1 Prediction of optimal metabolic rates and ranges

Growth-optimizing substrate uptake rates are found by computationally maximizing for growth rate as in OBrien et al. Carbon sources present in the media are allowed in excess (i.e., infinite bounds on uptake rate). Translation fluxes for genes that are knocked out are set to zero. After determining optimal substrate uptake rates, the range of near-optimal substrate uptake rates (x-axis error bars in Figure 4.1C,D) is determined by setting growth rate to 95% of its maximum determined value, only allowing fermentation products to be secreted which are secreted in the growth-maximizing state (setting other secretion to zero), and maximizing and minimizing the uptake rate reaction with linear programming.

4.3.2 Calculating catabolic proteome mass fraction

Genes were considered to be a part of carbon catabolism based on reaction subsystem labels and gene-protein-reaction relationships (GPRs) from the most recent E. coli metabolic reconstruction [98]. Subsystems considered to be a part of carbon catabolism are: Oxidative Phosphorylation, Pentose Phosphate Pathway, Alternate Carbon Metabolism, Citric Acid Cycle, Anaplerotic Reactions, Glycolysis/Gluconeogenesis, Pyruvate Metabolism, Methylglyoxal Metabolism, Glyoxylate Metabolism. Genes that can catalyze any of these reactions (i.e., are present in the GPR) are considered to be

catabolic. The catabolic proteome mass fraction is then computed based on protein synthesis fluxes (in ME-Model simulations) and molecular weights.

4.3.3 Statistical analysis

For substrate uptake rate predictions, Pearsons correlation coefficient (PCC) is used as the relationship is expected to be linear. All other correlations are reported as Spearmans correlation coefficient (rs) as they are not necessarily expected to be linear.

4.3.4 Assembly of ALE data compendium

We gathered physiological data from endpoint strains after ALE from the literature. We only considered short-term (30-60 day) ALEs performed at 37 degrees with E. coli K12 MG1655 as the starting strain. ALE experiments that do not measure substrate uptake rates were not included.

4.3.5 Physiological characterizations

Growth rates of clones isolated from the primary ALE experiments were screened by inoculating cells from an overnight culture to a low optical density (OD) and sampling the OD600nm until stationary phase was reached. A linear regression of the log-linear region was computed using polyfit in MATLAB and the growth rate (slope) was determined. Growth rates of populations were determined by the output of the interpolated cubic spline used, unless stated otherwise. Extra-Cellular by-products were determined by HPLC. Cell cultures were first sampled and then sterile filtered. The filtrate was injected into an HPLC column (Aminex HPX-87H Column #125-0140). Concentrations of detected compounds were determined by comparison to a normalized curve of known concentrations. Substrate uptake and secretion rates were calculated from the product of

the growth rate and the slope from a linear regression of gDW vs substrate concentration. Biomass Yield (Y) was calculated as the quotient of the growth rate and glucose uptake rates during the exponential growth phase.

4.3.6 RNA sequencing

RNA sequencing for two strains was obtained from LaCroix et al. and available in the Gene Expression Omnibus (GEO) database under accession number GSE61327. Reads were mapped with bowtie2 [166]. The gene expression fold change between the high Y and high r strain was found using cuffdiff2 [167].

4.3.7 13C-MFA

Triplicate cultures were grown on labeled glucose M9 minimal media with trace elements [49] and sampled from a heat block that was maintained at 37 C and fully aerated with tumble stir magnets. The labeled tracer consisted of 20/80 mixture of 13C glucose and 1-13C glucose purchased from Cambridge Isotope Laboratories, Inc. (Tewksbury, MA). MFA simulations were conducted using MATLAB and INCA v1.3 [189].

4.4 Acknowledgements

We thank Elizabeth Brunk, Daniel Zielinski, Joshua Lerman, Steve Federowicz for discussions. EJO was supported by NIH R01 GM057089. The Novo Nordisk Foundation supported this research. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Author contributions: EJO and BOP conceived the study. EJO performed simulations, data processing, and analysis. DM performed gene knockouts, physiological characterization, LC-MS/MS analytics, and metabolic flux analysis. RAL and TES performed additional ALE experiments. All co-authors provided feedback and suggestions. BOP supervised the study. EJO and BOP wrote the manuscript.

The text of Chapter 4 is a full reprint of the material as it appears in: OBrien E.J.*, McCloskey D.*, Utrilla J., King Z.A., LaCroix R.A., Sandberg T.E., Feist A.M., Palsson B.O. Tradeoffs in microbial adaptation are determined by proteome complexity, Submitted. * indicates equal contribution. The dissertation author was the primary author of the manuscript. The other authors were Douglas McCloskey (equal contributor), Jose Utrilla, Zachary A. King, Ryan A. LaCroix, Troy E. Sandberg, Adam M. Feist, and Bernard Ø. Palsson.

Chapter 5

Proteome and Energy Re-allocation by Adaptive Regulatory Mutations Reveals a Fitness Trade-off

You get what you select for.
—Frances Arnold

5.1 Summary

Adaptive laboratory evolution (ALE) with genome re-sequencing of endpoint strains can identify the genetic basis for new phenotypes. Causation is established by introducing mutations found in endpoints into the starting strain. This approach, augmented with omics data and systems analysis, reveals multi-scale mechanistic genotype-phenotype relationships. This process is detailed for ALE-selected variants in Escherichia coli RNA polymerase. We show that these mutants perturb the transcriptional regulatory network to rebalance proteome and energy allocation towards growth and away from several hedging functions. These findings highlight the resource allocation constraints

organisms face and suggests how regulatory structure enhances evolvability.

5.2 Introduction

Many causal genetic variants across all forms of life are found in regulatory regions [190, 191, 192, 193, 194, 195]. In addition to cis regulatory variation, causal mutations are often found in trans-acting transcriptional regulators [178, 196, 179, 197, 198]. Here, we detail the multi-scale mechanism underlying several trans-acting adaptive regulatory mutations of E. colis RNA polymerase (RNAP) [178, 199, 200]. Though these mutations are not physically close in sequence or structure, we find that they share a common molecular mechanism. Detailed phenotypic assays show consistent fitness benefits of the mutations in static environments and fitness detriments in variable environments (i.e., nutrient shifts and stress shocks). A multi-omic approach with key environmental controls reveals a systematic and consistent modulation of the transcriptional regulatory network (TRN) towards growth functions and away from functions that hedge against environmental change. Econometric analysis using a genome-scale model reveals that the resulting resource re-allocation can quantitatively explain the fitness effects. Finally, structural dynamics of RNA polymerase (RNAP) provide insight as to how these mutations result in strikingly similar effects. Though RNAP is typically not considered a transcription factor, these results show that it lies at the top of the TRN hierarchy, regulating cellular growth and various hedging functions [201].

Thus, these mutations in RNAP result in a broad form of antagonistic pleiotropy (growth versus hedging) based on resource re-allocation. As protein synthesis and energy are limited resources, we can conclude that the pleiotropic effects reflect an inherent trade-off between growth and hedging functions. Similar antagonistic pleiotropy has been observed in other trans regulatory variants [155, 156, 157, 202]. This study moves the

field forward by detailing the multi-scale mechanism underlying the pleiotropic effects of adaptive regulatory mutations. It provides insight into the evolutionary constraints and the mechanisms that govern resource allocation in simple organisms.

5.3 Results

5.3.1 Adaptive mutations in RNA polymerase reveal growth versus hedging phenotypes

A recent adaptive laboratory evolution (ALE) experiment of E. coli in glucose minimal media (MM) identified recurring mutations in rpoB (the β subunit of RNAP), including rpoB E546V and rpoB E672K [178]. We introduced these two ALE-selected mutations into the starting strain (i.e., the wild type strain) and observed consistent physiological effects. Growth rate increased (by 25%) resulting from increases in both biomass yield (by 11%) and substrate uptake rate (by 14%). The use of an automated plate reader to obtain frequent measurements revealed a diauxic shift of the mutant strains in glucose M9 mineral media (Fig. 5.1A).

As mutations often have positive and negative fitness effects across several environments (referred to as pleiotropy), we then assessed the growth rate of the rpoB E546V and rpoB E672K mutants under a variety of single carbon sources, mixtures of carbon sources, rich media, and stress conditions. Additionally we performed, motility, acid shock, and antibiotic persistence phenotypic tests (Fig. 5.1B). These RNAP mutations show consistent fitness effects: they enable faster growth in several carbon sources, in low pH, and in the presence of erythromycin. However, they lead to lower motility, lower survival under acid shock, reduced antibiotic persistence, longer diauxic shifts, and lower growth rates in complex media.

Therefore, the mutants show increased fitness in conditions of steady-state growth, but a decreased fitness in changing environments. They show strong, consistent antagonistic pleiotropy for growth versus hedging functions.

5.3.2 Mutations in RNA polymerase are highly specific

To assess whether other amino acid substitutions in the RNAP ALE-selected loci affect growth phenotypes, we generated a series of additional variants using multiplex automated genome engineering (MAGE) [203]. Two amino acid substitutions with similar chemical properties as those discovered by ALE resulted in an increase in growth rate (i.e., E546K and E672R), whereas all other amino acid substitutions generated by MAGE did not affect growth rate significantly. MAGE selected mutants that grow faster than the wild type also exhibit longer diauxic shifts, showing similar pleiotropic effects as the ALE selected mutants.

Therefore, the mutations in RNAP affecting fitness are specific. All faster growing RNAP mutants showed antagonistic pleiotropy for growth versus hedging.

5.3.3 Genome-scale transcript profiling reveals conserved growth versus hedging response

To reveal the systems-level mechanism of the pleiotropic effects of the RNAP mutations, we obtained RNA-seq and metabolomics data from mid-logarithmic growth phase in glucose minimal media for the wild-type, rpoB E546V, and rpoB E672K mutant strains. Metabolite concentrations that changed significantly compared to the wild-type include pyrimidine, glycolytic, and TCA intermediates, but overall, the metabolome remained fairly stable. On the other hand, the expression profiling data revealed 243 consistently differentially expressed genes. Like the pleiotropic fitness effects of the mu-

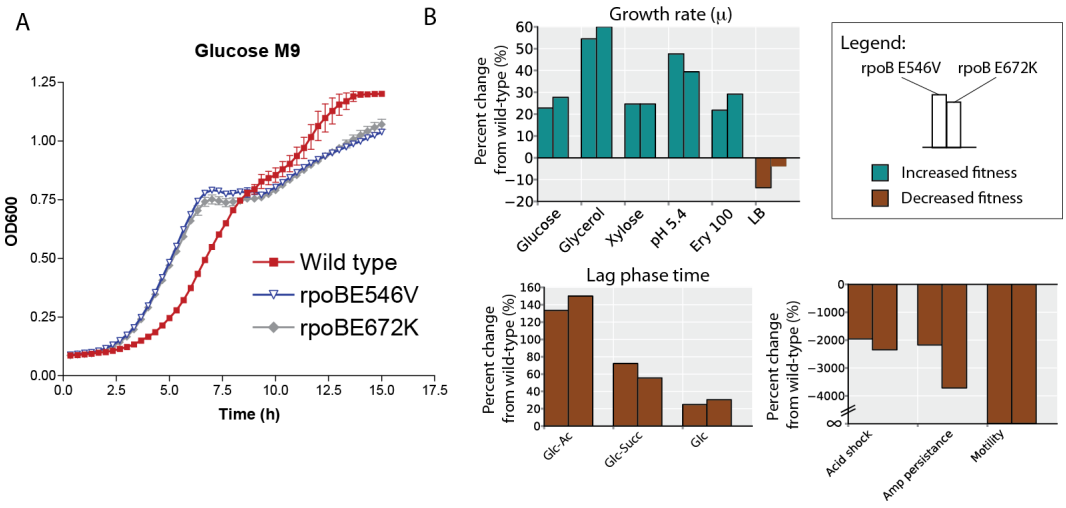


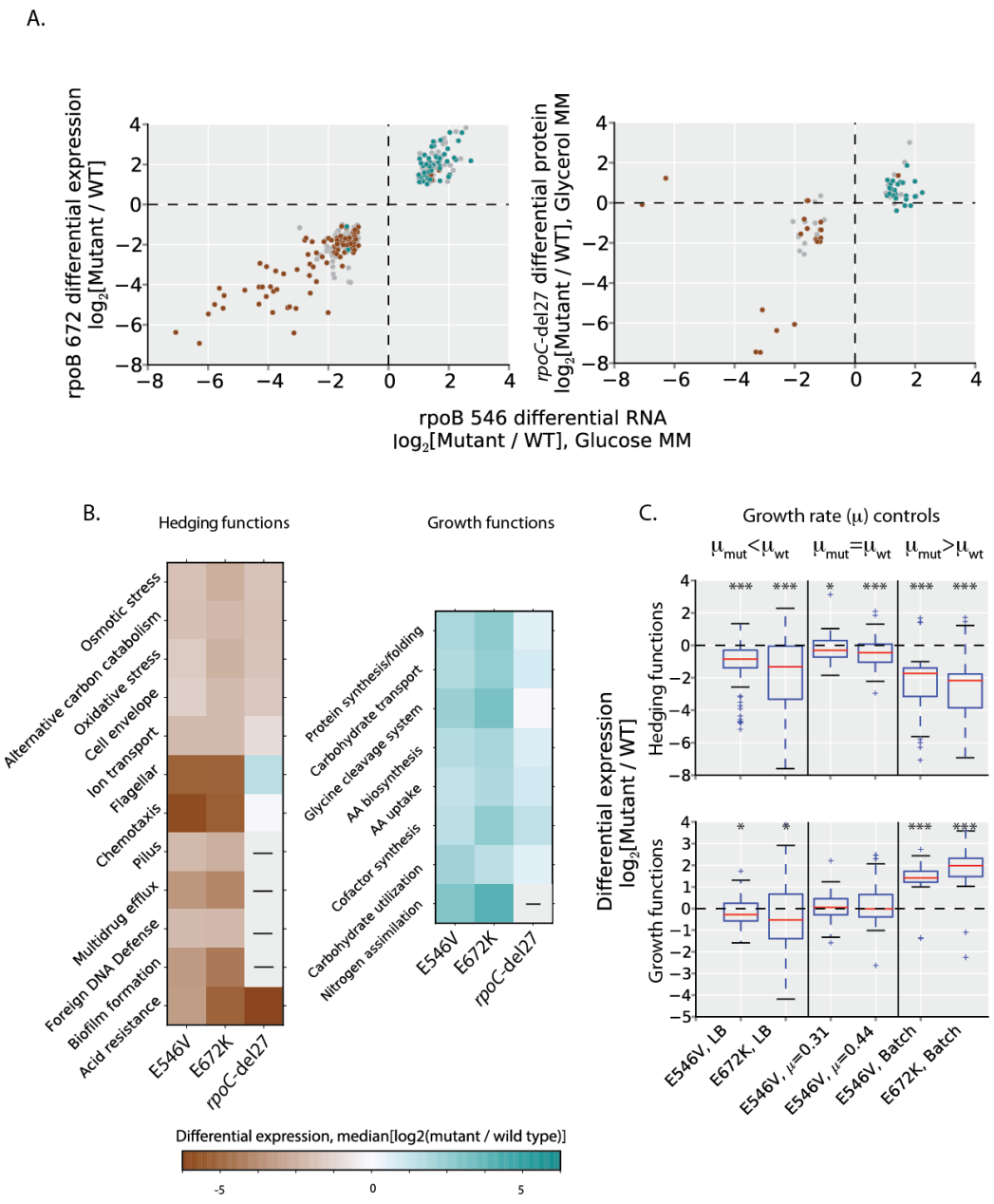
Figure 5.1: **Growth versus hedging antagonistic pleiotropy in organismal phenotypes.** A) Adaptive Laboratory Evolution (ALE) -selected rpoB mutations (E546V blue, E672K gray) grow faster in the glucose consumption phase but have a longer diauxic shift to grow on acetate than the wild type (red) In addition to growth on glucose (the environment in which the mutants were selected), several additional organismal phenotypes are affected by the rpoB mutations. Bar charts show the percent change in measured phenotypes compared to the wild type. Steady-state growth rates increases (cyan) and growth rate in LB medium as well as fitness in environmental shifts and shocks decreases (brown). LB: Luria Broth, Glc: Glucose, Succ: Succinate, Ac: Acetate, Ery 100: 100 g/mL erythromycin, Amp: Ampicillin.

tants, the differential gene expression is strikingly conserved (Fig. 5.2A, left), indicating a common underlying mechanism at the systems level.

Interestingly, we also find that the differential expression of the two rpoB mutants is similar to a previously profiled 27 amino acid deletion mutant in the β' subunit of the RNAP (rpoC-del27, identified by ALE on glycerol) [199, 200, 204]. The changes in expression of the rpoC-del27 mutant [200] (compared to wild-type) grown in glycerol match those of the rpoB mutants grown in glucose (Fig. 5.2A, right).

To obtain insight into the processes perturbed by the RNAP mutations, we classified the 243 consistently differentially expressed genes by function. We found that the genes in the same functional category are often differentially expressed in a consistent direction. We used this observation to define up-regulated and down-regulated functions. The up-regulated functions (defined as ¿80% of the genes being up-regulated) are broadly related to cellular growth, including protein synthesis and folding, amino acid biosynthesis and uptake, and carbohydrate transport and utilization. On the other hand, the down-regulated functions (defined as ¿80% of the genes being down-regulated) broadly hedge against environmental change and stress, including osmotic and oxidative stress, flagella, chemotaxis, acid resistance, and biofilm formation. Two categories of genes are not consistently up or down-regulated; these are DNA repair and genes with unknown function. Thus, at the molecular level, the differentially expressed genes reflect the growth versus hedging phenotypes observed at the organismal level.

Figure 5.2: **Conserved molecular growth versus hedging response.** A) The differential RNA expression in the ALE-selected rpoB mutants (E546V, E672K) is conserved (left). The differential RNA expression in glucose is also concordant with the differential protein expression in glycerol of an ALE-selected 27 aa deletion in β' (rpoC-del27) (right). B) Functional classification of differentially expressed genes reveals that genes with common functions are often differentially expressed in the same direction, segregating growth (up-regulated, cyan) and hedging (down-regulated, brown) functions. Gray dots are genes with functions that are not consistently differentially expressed. Median differential expression of genes in the functional categories is shown in the heatmap; dashes indicate genes not detected in proteomics data [200]. C) Environmental controls disentangle direct effects of the mutations and indirect effects of changes in growth. Box plots show differential expression of identified growth and hedging functions across environments, showing that hedging functions are consistently down-regulated and the expression of growth functions depends on the growth rate. Stars indicate if the mean differential expression of the group of genes is significantly different than zero, based on a two-sided t-test ($p < 0.05$, *; $p < 0.0001$, ***).


A.
rpoB 672 differential expression
log2[Mutant / WT]
rpoC-del27 differential protein
log2[Mutant / WT], Glycerol MM
rpoB 546 differential RNA
log2[Mutant / WT], Glucose MM
B.
Hedging functions
Growth functions
Osmotic stress
Alternative carbon catabolism
Oxidative stress
Cell envelope
Ion transport
Flagellar
Chemotaxis
Pilus
Multidrug efflux
Foreign DNA Defense
Biofilm formation
Acid resistance
Protein synthesis/folding
Carbohydrate transport
Glycine cleavage system
AA biosynthesis
AA uptake
Cofactor synthesis
Carbohydrate utilization
Nitrogen assimilation
E546V
E672K
rpoC-del27
Differential expression, median[log2(mutant / wild type)]
C.
Growth rate (μ) controls
μmut<μwt
μmut=μwt
μmut>μwt
Differential expression
log2[Mutant / WT]
Hedging functions
Growth functions
E546V, LB
E672K, LB
E546V, μ=0.31
E546V, μ=0.44
E546V, Batch
E672K, Batch

5.3.4 Environmental controls disentangles cause versus effect of mutations

As growth rate itself has a strong effect on gene expression [205], we sought to identify the differential expression caused only by the mutation from that indirectly caused by increased growth. To disentangle these effects we obtained RNA-seq data under conditions where the wild-type and mutant strains grow at the same rate (glucose limited chemostat culture) and under conditions where the mutants grow slower than the wild-type (LB rich media). Regardless of the growth rate and environment, the hedging functions are down-regulated in the mutant strain compared to the wild-type (Fig. 5.2C). Differential expression of the growth functions, however, is dependent on the growth rate: growth genes are not differentially expressed in chemostat and are down-regulated in LB. Thus, these environmental controls disentangle the cause and effect of the mutations: the mutations directly result in the down-regulation of hedging genes whereas the growth-related genes are coupled to the cells growth rate.

5.3.5 Structural dynamics of RNAP suggests a common allosteric mechanism

Both mutations, rpoB E546V and E672K, are located approximately 25 away from the catalytic site of RNAP, and about 25 from each other. How do they result in such similar patterns in transcriptional reprogramming to down-regulate hedging functions?

To answer this question, we performed molecular dynamics simulations aiming to propose a common putative molecular mechanism for the pleiotropic fitness effects of the rpoB mutations. Interestingly, we found a strong correlation between the extent of increase in interaction energy between the β and β' subunits, and the increase in cell fitness for various E672 mutations generated by MAGE (both beneficial and neutral, Fig.

5.3A). Such destabilization of subunit interaction is consistent with a previous study that showed a decrease in open complex half-life of the rpoC-del27 mutation, which has similar growth and transcriptional effects [199].

To further explore the functional correlation among different mutations, we decomposed the RNAP complex into structural communities within which the molecular motions of residues are strongly correlated [206]. In spite of the large spatial separation between E672 and E546, they belong to the same dynamical community (Fig. 5.3B). Furthermore, many mutations detected in RNAP in other ALE experiments [178, 197, 207] can also be found in this and neighboring communities (Fig. 5.3B). This structural community consists of 250 residues in rpoB, the bridge helix in rpoC, and nucleotides on the template DNA strand. Because the bending motion of bridge helix has been shown to coordinate catalysis and DNA translocation in the nucleotide addition reaction [208, 209, 210], the collective motion of this community may be directly related to nucleotide elongation. In fact, we observe a strong correlation between the bending angle of the bridge helix (a motion known to be directly involved in elongation [208, 209, 210]) and the relative motion between neighboring communities along the direction of DNA translocation. Again, the relation between the community dynamics and transcriptional elongation is consistent with the increased elongation rate observed in the related rpoC-del27 mutation.

The observed destabilization of subunit interaction and its role in elongation are both reminiscent of the effects of (p)ppGpp and dksA on the stringent response [211, 212]. The allosteric regulator, (p)ppGpp, modulates transcription by destabilizing the intrinsically short lived open complexes [213] and affecting sigma factors use [214]. Interestingly, we observed a conserved optimal path linking E564/E672 and the (p)ppGpp binding site in the ω subunit (Fig. 5.3C), showing a common effective allosteric communication between these distantly located functional residues. The ALE-selected mutations

may therefore modulate transcription in a similar manner as (p)ppGpp [201].

In summary, several features of RNAP structural dynamics and function suggest a common allosteric mechanism of these mutations. The ALE-selected mutations are capable of modulating RNAP complex interactions and nucleotide elongation at the molecular level, which in turn, modulates global transcriptional regulation.

5.3.6 Transcriptional regulatory network perturbation explains observed molecular response

Consistent with the perturbed structural properties of the mutated RNAP, the differentially expressed growth and hedging functions have sigma factor biases. Even though the sigma factors are not detectably differentially expressed, the down-regulated (hedging) genes tend to have promoters utilizing stress related sigma factors (S, F) and the up-regulated (growth) genes tend to have promoters utilizing growth related sigma factors (D, N, H) (Fig. 5.4A).

However, the observed differential expression is more specific than that caused by sigma factors alone. There are 10 transcription factors (TFs) and regulatory small RNAs (sRNAs) that are differentially expressed in the mutant strains (Fig. 5.4B). Each of these regulators can be associated with one or more of the differentially expressed functional categories identified. Furthermore, across all of the strains (wild-type, rpoB E546V, and rpoB E672K) and environments (glucose excess, glucose limitation, and rich media) examined with RNA-seq, the differential expression of the identified growth and hedging functions is in a direction consistent with the differential expression of their regulators (based on known activation or repression relationships; Fig. 5.4B).

Thus, the balance between growth and hedging functions is achieved through global modulation of the TRN. The structure of the TRN enables E. coli to rebalance its proteome in response to evolutionary pressures with single point mutations in RNAP.

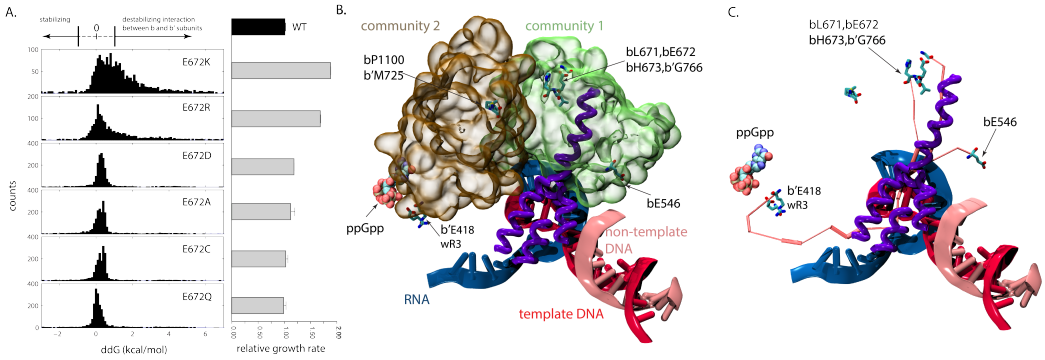


Figure 5.3: **ALE-selected rpoB mutations modulate structural dynamic of the E. coli RNAP.** A) Change in interaction energy between the β & β′ subunits across six different E672 mutations, compared with their corresponding growth rates. To reduce bias from a single static crystal structure, interaction energy is calculated every 25 ps over a 60 ns molecular dynamic trajectory starting from the RNAP open complex. B) Dynamical community structures encompassing the ALE-selected mutations. Community 1 (green), as discussed in the text, includes the bridge helix in β′ subunit (purple), βE672, βE546, and a few other ALE-selected mutations in contact with βE672. Community 2 (brown) spans the interface between the β & β′ subunits, interacting with community 1 on one side, and the (p)ppGpp binding site on the other. C) Effective allosteric communication between distantly located residues can be resolved from optimal path calculated based on a dynamical correlation network. The result shows that βE672 and βE546 share the same optimal dynamical path (orange) towards the ppGpp binding site in the ω subunit. Structural elements are shown from the same perspective, and color-coded the same as in B).

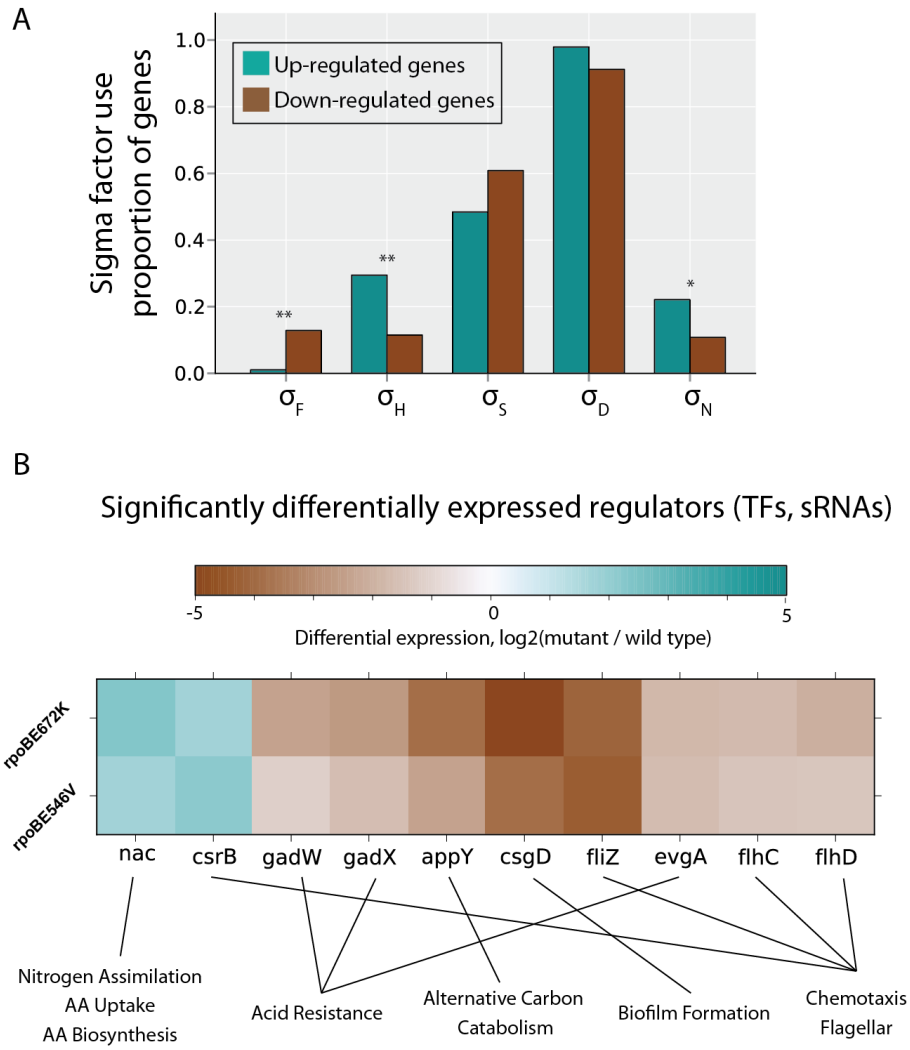


Figure 5.4: **Reprograming of the regulatory network.** A) The factor usage of differentially expressed genes in mutant strains is shown. Bars indicate the fraction of up-regulated (cyan) and down-regulated (brown) genes that have a promoter that is regulated by a given factor. Only factors with greater than 10% of promoters regulated among either up-regulated or down-regulated genes are shown. Significant differences in the proportion between factor use in up-regulated and down-regulated genes are indicated with asterisks; one asterisk indicates p¡0.05 and two asterisks indicate p¡0.005. B) The fold change for transcription factors and sRNA that are significantly differentially expressed in both mutant strains compared to the wild type are shown.

5.3.7 Econometric analysis of proteome and energy resource allocation explains fitness trade-off

The molecular and regulatory effects of the rpoB mutations reveal that resource allocation underlies the observed growth versus hedging fitness effects. A recently developed genome-scale computer model of microbial growth [53], called a ME-model [51, 89, 53] (for metabolism and expression) can quantify the fitness effects associated with proteome and energy re-allocation (Fig. 5.5A).

The ME-model allows global energy accounting based on the physiological data from wild-type and RNAP mutant strains. The results show that the RNAP mutations eliminate about a third (28-37%) of the unaccounted for energy (i.e., processes not involved in metabolism and protein synthesis, often referred to as the maintenance energy [215], Fig. 5.5B). Then, using the gene expression data we estimate a 2-5% reduction of the transcriptome allocated to non-ME genes (i.e., not included in the ME-model, non-growth functions) and a commensurate increase in ME gene (i.e., modeled, growth) allocation in the RNAP mutants (Fig. 5.5B). ME-model analysis thus shows a clear shift to a more growth-supporting proteome as a result of the observed RNAP mutations.

We used the ME-model to understand how these changes in resource allocation affect cellular physiology (i.e., growth rate, biomass yield, and uptake rate). The non-ME proteome and energy allocation are adjustable model variables. Indeed, when varied in the model, the measured changes in non-ME energy and transcriptome use can quantitatively account for the measured physiological changes (biomass yield and uptake rate) in the mutant strains (Fig. 5.5C). Therefore, the growth increase can be accounted for by the measured change in resource allocation. The expression of hedging functions restrains growth rate in the wild-type strain.

The ME-model allows us to quantitatively elucidate the relationship between

changes in overall physiological measures (i.e., growth rate, substrate uptake rate, and yield) and the changes in allocation of protein and energy (Fig. 5.5). This quantitative relationship allows us to conclude that the pleiotropic effects of the rpoB mutation are due to a fundamental constraint of limited proteome and energy resources, leading to an inherent trade-off in resource allocation.

5.4 Discussion

Here, we elucidate the mechanistic multi-scale basis of adaptive regulatory mutations. Single amino acid changes in the RNAP reprogram the TRN to re-allocate resources towards growth and away from hedging functions. The mutations result in antagonistic pleiotropy where the organism is more fit in stable environments but less fit in environmental shifts and shocks [216].

5.4.1 Antagonistic pleiotropy due to a fundamental trade-off

Mutations that are beneficial or neutral in one environment often have negative fitness effects in other environments, referred to as pleiotropy. Pleiotropy shapes the evolution of organisms and is thought to underlie the evolution of specialist species [216]. Several mechanisms can give rise to pleiotropy and some have been demonstrated [217, 218, 219].

Fundamental biological constraints can result in antagonistic pleiotropy, though examples of these cases are lacking. Using a systems biology approach, we show that the growth rate difference in wild-type and mutant strains can be quantitatively explained by changes in proteome and energy allocation. These resources are limited, resulting in an inherent trade-off between growth and hedging functions. Such proteome and energy allocation constraints likely result in pervasive evolutionary trade-offs and likely underlie

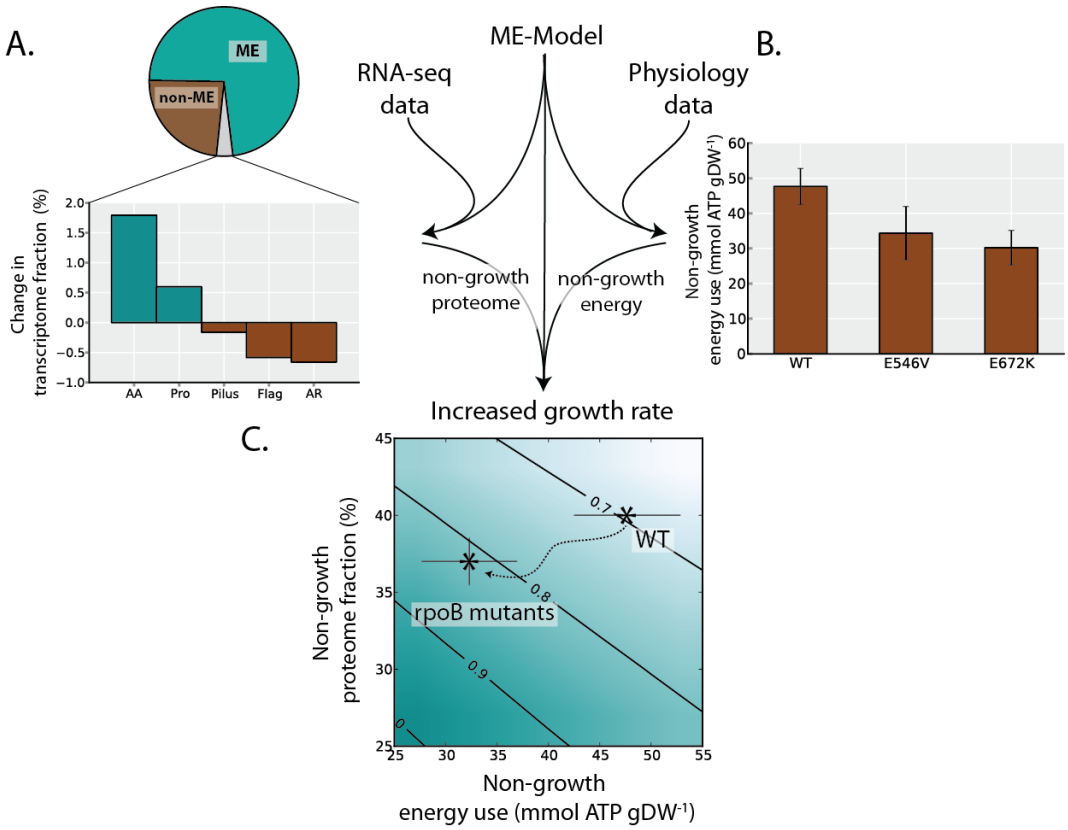


Figure 5.5: **The changes and effects of proteomic and energetic resource allocation.** A) A genome-scale model of Metabolism and gene Expression (ME-Model) is used to integrate the RNA-sequencing and physiological data. The transcriptome fraction devoted to ME and non-ME (i.e., not included in the ME-Model) genes is calculated for the wild-type and mutant strains. Grey area of the pie chart indicates the fraction of the transcriptome reallocated from non-ME to ME genes. Bar chart shows the functional categories that reduced or increased in expression by more than 0.1% of the total transcriptome. Abbreviations for the functional categories are: amino acid biosynthesis (AA), protein synthesis/folding (Pro), acid resistance (AR), and flagellar (Fla). All percentages are shown as the average for E546V and E672K. B) The physiological data was used to calculate the energy use not accounted for by the ME-Model (see Methods, Computation of maximum unaccounted for energy), showing a reduction in unaccounted for energy use in rpoB mutants compared to the wild-type. Error bars indicate standard error across biological replicates. C) The effects of non-ME protein and energy use on maximal growth rates in the ME-Model are computed and shown in the contour plot (see Methods). The wild-type and mutant strains are indicated on the plot, showing how lower non-ME protein and energy use can cause increased growth.

several recent examples of antagonistic pleiotropy [155, 156, 157].

5.4.2 Evolvability through regulatory network structure

Mounting evidence supports that much of the functional divergence between organisms occurs in regulatory regions [195, 190, 191, 192, 193, 194]. The detailed example of the RNAP mutations here suggests why (in part) this may be the case.

As regulatory networks are aligned with particular functional subsystems, mutations that perturb them change phenotypes in a functionally coherent manner [220, 221, 198]. The regulatory rebalancing detailed here occurs along a coherent growth versus hedging trajectory. On the other hand, mutations that are inconsistent or imbalanced in the molecular changes they cause would likely not be selected. Therefore, in addition to enabling proximal response to environmental change, the structure of the regulatory network also enables productive evolutionary change. Remarkably, single, but non-unique, point mutations allow such adaptation.

5.4.3 Multi-scale characterization of genotype to phenotype

Sequencing of many individual genomes has led to the identification of genomic regions under selection [222] and enabled the association of variants with organismal [223] and molecular [224] phenotypes. However, there is a large gap between identifying causal variants and mechanistically understanding their phenotypic consequences. The mutations studied here are some of the most comprehensively phenotyped to date, with environmental controls to separate cause and effect. We employ state-of-the-art structural and systems biology modeling approaches to help bridge the gap between genotype and phenotype. Together, these analysis approaches enable us to step from mutation to biophysical effects on protein function to systems-level molecular and regulatory

response, and finally to organismal phenotype (Fig. 5.6). Therefore, this study outlines how we might begin to understand the multi-scale genotype-phenotype relationship at a true systems level.

5.5 Experimental procedures

5.5.1 Strains and cultivations

E. coli MG1655 was used as wild-type. The ALE selected rpoBE564V and rpoBE672K knock in strains were previously constructed by allelic replacement [178] . To generate additional variants of rpoB546 and 672 positions, MAGE was performed on the wild-type strain by first transformation of recombineering plasmid pKD46 [225], then inactivation of mutS with two nonsense mutations at residues 189 and 191 using an oligo (mutS_MUT). Two oligos (rpoB_E546X and rpoB_E672X) that resulted in NNS codon mutations at rpoB residues 546 and 672 were introduced into the strain through 8-12 rounds of MAGE, followed by colony isolation of mutants, PCR verification, and Sanger sequencing. To perform each cycle of MAGE, the l-Red system was induced with 0.5%-arabinose 45 minutes prior to generation of electrocompetent cells and oligo. Batch cultures were done in flask with M9 minimal media and 4 g/L of glucose at 37C or LB rich media. Glucose limited chemostats were carried out in a Bioflo 110 fermentor (New Brunswick Scientific, NJ). Glucose supplemented M9 was added to the reactor at 0.31 and 0.44 h-1 dilution rates controlled by a peristaltic pump. Steady state was achieved after 3-5 residence times and was verified by biomass measurements. Phenotypic tests were performed by inoculation of media with an overnight pre-culture of glucose M9 media for all cases. Erythromycin was added to the media to the indicated concentration. The pH of M9 was adjusted to the indicated value with 6M HCl. Different substrates and mixtures were added to M9 to test growth in the indicated conditions. All growth

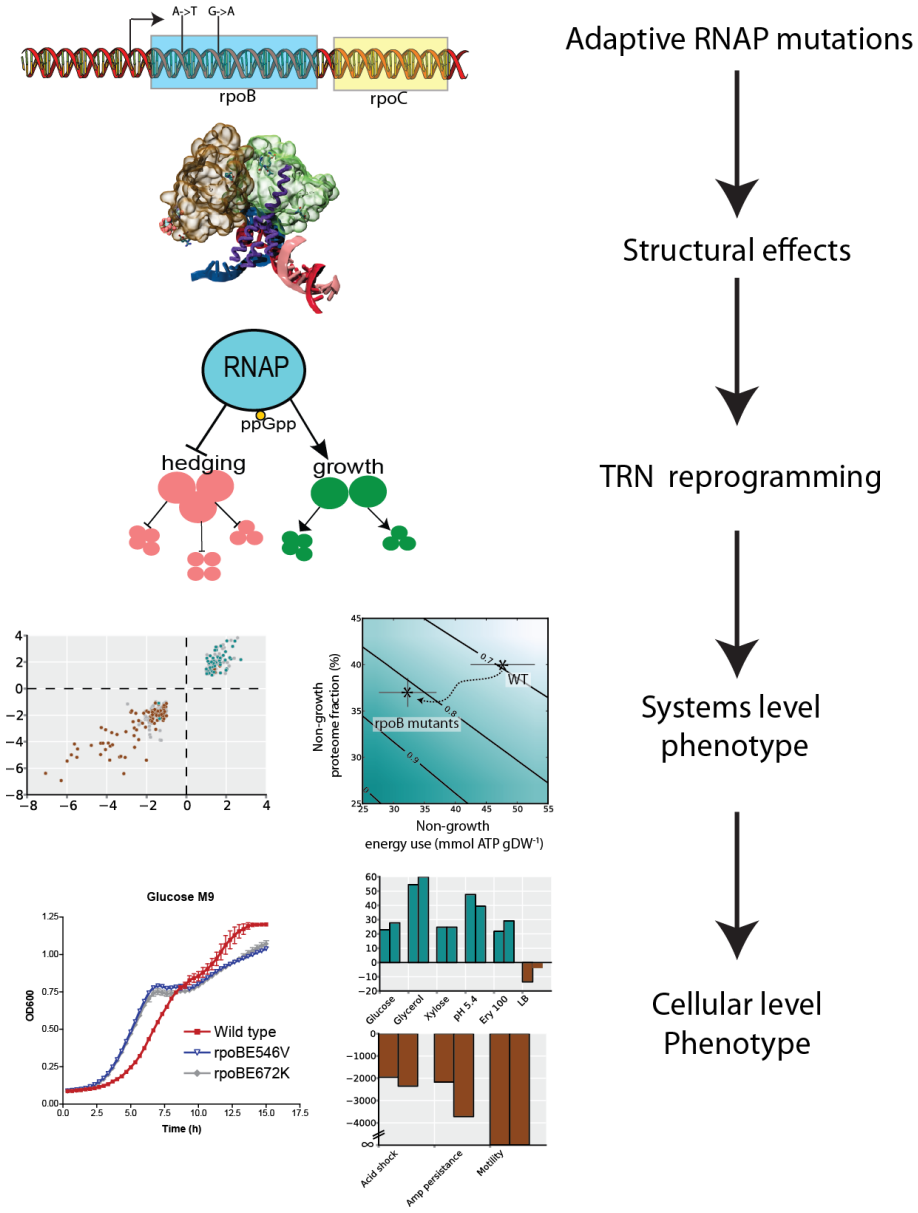


Figure 5.6: **Multi-scale characterization from genotype to phenotype.** The multi-scale effects of the studied adaptive regulatory mutations in RNAP are summarized. The mutations alter the structural dynamics of the RNAP, perturbing the TRN through the action of key transcription factors. The decrease in expression of hedging functions lowers the proteome and energy allocation towards hedging functions and increases cellular growth. In turn, the cell can grow faster in conditions of steady-state growth, but is less fit under environmental shifts and shocks.

curves were inoculated to a 0.02 OD and 200 L were cultured by triplicate in a Bioscreen C device at 37C for 15- 24 h

5.5.2 Motility test

Cells were grown to mid log phase and 10 microliters of cell suspension were spotted onto 0.3% agar plate with glucose M9 media, plates were photographed motility was determined by halo expansion between 24 and 48h

5.5.3 Acid shock

Cells were harvested in mid log phase and normalized to 1X108 cells/mL, 50 L of cells suspension were resuspended in 950 L of pH 2.6 glucose M9 media. After 3 hours of incubation cells were diluted and plated in LB agar plates for cell counts [226].

5.5.4 Antibiotic persistence

Cells were harvested in mid log phase and normalized to 1X108 cells/mL, different dilutions were plated in LB ampicillin plates after 24h a sterile solution of 25 U of penicinillase was plated and plates were re-incubated for 24h. Appearance of colonies was determined and persistence frequency determined in base of initial cell counts [227].

5.5.5 Analytics

Biomass was determined by measuring the absorbance of the culture at 600nm using an equivalence of 0.429gDW/L per OD600 unit. Glucose, and acetate were measured by HPLC using refractive index (RI) detection by high-performance liquid chromatography (HPLC) (Waters, MA) with a Bio-Rad Aminex HPX87-H ion exclusion column (injection volume, 10 l) and 5mM H2SO4 as the mobile phase (0.5 ml/min, 45C).

Metabolomic sampling, extraction and analysis was carried out as described earlier by our group [228].

5.5.6 RNA-seq libraries

Samples for RNA-sequencing were taken in mid log phase of batch cultures or during the steady-state in chemostats. Cells were collected with Qiagen RNA-protect Bacteria Reagent and pelleted for storage at -80C prior to RNA extraction. Cell pellets were thawed and incubated with Readylyse Lysozyme, SuperaseIn, Protease K, and 20% SDS for 20 minutes at 37C. Total RNA was isolated and purified using the Qiagen RNeasy Mini Kit columns and following vendor procedures. An on-column DNase-treatment was performed for 30 minutes at room temperature. RNA was quantified using a Nano drop and quality assessed by running an RNA-nano chip on a bioanalyzer. Paired-end, strand-specific RNA-seq was performed following a modified dUTP method [165] . The rRNA was isolated using Epicentres Ribo-Zero rRNA removal kit for Gram Negative Bacteria.RNA-seq was performed using a modified dUTP method [165]

5.5.7 Transcriptome analyses

The obtained reads were mapped to the E. coli MG1655 genome (NC_000913.2) using the short-read aligner Bowtie (http://bowtie-bio.sourceforge.net) [166] with two mismatches allowed per read alignment. To estimate gene expression FPKM values were calculated using cufflinks tool and differential expression analysis was carried out using cuffdiff feature of the same package using the upper quartile normalization (http://cufflinks.cbcb.umd.edu/) [167]. Gene set enrichment analysis on differentially expressed genes was performed using GO annotations from EcoCyc [8]. A hypergeometric test and p-value cutoff of 0.01 was used.

5.5.8 Regulatory network

Sigma factor use at promoters was obtained by combining annotations in Cho et al. [229] and EcoCyc [8]. The list of all transcription factors and sRNAs was obtained from RegulonDB [230]. A two-proportion z-test with two-tailed comparisons was used to determine significant differences in sigma factor usage among up-regulated and down-regulated genes.

5.5.9 Computation of maximum non-growth energy use

The E. coli ME-Model with all parameters as published in OBrien et al. was used [53]. For all replicate cultivations, the measured growth rate, glucose uptake rate, and acetate secretion rate were fixed in the model. The maximum unaccounted for energy use was then computed by maximizing the flux through ATP maintenance reaction, which hydrolyzes ATP. For a given strain, the unaccounted for energy use is reported as the average across biological replicates.

5.5.10 Computation of non-ME transcriptome

The (protein coding) ME and non-ME transcriptome fractions were estimated using FPKM and gene length. A genes transcriptome fraction was taken to be the product of FPKM and the gene length, divided by the sum of this product over all genes. The ME and non-ME transcriptome fractions were then calculated by summing the transcriptome fractions of all ME and non-ME genes, respectively. Ranges are determined from the estimated lower and upper FPKM values across different samples.

5.5.11 Computation of the effects of changes in resource allocation

Protein and energy that are not used towards cell growth are changeable variables in the ME-Model. These are varied to determine the growth rate, biomass yield, and substrate uptake rate contours (Fig. 5.5C). The points and error bars for wild-type and rpoB mutants are placed according to the unaccounted for energy (Fig. 5.5C) and change in non-ME transcriptome (Fig. 5.5B). As we do not explicitly know the proteome fraction devoted to growth in each strain, we determine these values with two assumptions. First, we assume the change in non-growth proteome is equal to the change in the non-ME transcriptome. Second, we infer the non-growth proteome in the wild-type strain based on its measured growth (which is why there is no y-axis error bar for the wild-type), resulting in a value consistent with previous estimates [108].

5.5.12 Molecular dynamics simulations

Molecular model of the E. coli RNAP elongation complex (EC) were created using the crystal structure of the E. coli RNAP core enzymes (PDB code: 3LU0 [231]), the template and non-template DNA strands, and the DNA:RNA hybrid helix (PDB code: 2O5J [232]). The system were neutralized with Mg2+ and K+ ions, initially placed in positions occupied by metal ions in the crystal structure or according to the electrostatic potential. The complex was then solvated by well-equilibrated water molecules with periodic boundary conditions. 200mM KCl was added to the final solution. Molecular dynamics simulations were run for 60ns (1-fs time steps) under constant pressure (1atm) and constant temperature (25C) using NAMD2.9 [233] with the CHARMM36 force field [234] Community analysis and optimal path calculation were done using algorithms described in [206] with the software VMD [235].

5.5.13 Interaction energy calculation

Change in the interaction energy between the β and β′ subunits upon mutations were calculated with the alanine scan script using PyRosetta [236], originally distributed by the Gray lab. We applied modifications of the score function parameterized according to recently reported protocols [237, 238]. To reduce the bias introduced by a single static crystal structure, we performed the computational alanine scan every 25 ps through the entire trajectory, resulting in a broad distribution of the ddG values. Although such ddG value was taken to be qualitative conventionally (with ddG $>$ 1 kcal/mol to be destabilizing), we emphasized that it was the observed trend over the dynamical trajectory that correlated with phenotypic fitness of the MAGE mutants.

5.6 Acknowledgments

Richard Szubin for technical assistance, Elizabeth Brunk for discussions. The Novo Nordisk Foundation for funding this study. EJO was supported by NIH GM057089. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Author contributions: JU and BOP conceived the study. JU performed the physiological and RNA-seq experiments. DM performed metabolomics experiments. JC and HW generated mutation variants by MAGE. JU and EJO performed data processing and computational analysis. KC performed structural analysis. BOP supervised the study. AMF provided critical feedback. JU, EJO, KC, and BOP wrote the manuscript.

The text of Chapter 5 is a full reprint of the material as it appears in: Utrilla J.*, OBrien E.J.*, Chen K., McCloskey D., Cheung J., Wang H., Armenta-Medina D., Feist A.M., Palsson B.O. Proteome and Energy Re-allocation by Adaptive Regulatory

Mutations Reveals a Fitness Trade-off, Submitted. * indicates equal contribution. The dissertation author was the primary author of the manuscript. The other authors were Jose Utrilla (equal contributor), Ke Chen, Douglas McCloskey, Jacky Cheung, Harris Wang, Dagoberto Armenta-Medina, Adam M. Feist, and Bernard Ø. Palsson.

Chapter 6

Computing the functional proteome: recent progress and future prospects for genome-scale models

6.1 Abstract

Constraint-based models enable the computation of feasible, optimal, and realized biological phenotypes from reaction network reconstructions and constraints on their operation. To date, stoichiometric reconstructions have largely focused on metabolism, resulting in genome-scale metabolic models (M-Models). Recent expansions in network content to encompass proteome synthesis have resulted in models of metabolism and protein expression (ME-Models). ME-Models advance the predictions possible with constraint-based models from network flux states to the spatially resolved molecular composition of a cell. Specifically, ME-Models enable the prediction of transcriptome and proteome allocation and limitations, and basal expression states and regulatory needs. Continued expansion in reconstruction content and constraints will result in an

increasingly refined representation of cellular composition and behavior.

6.2 Introduction

Building computational whole cell models has been a long-standing goal of theoretical biology. In the 1980s, serious attempts to build large-scale models of a whole bacterium were undertaken [239]. A few years later, an attempt to build whole cell models for the human red cell represented a culmination of decades of work [240, 241, 242, 243, 244]. Perhaps the most comprehensive whole organism model appeared in the mid 1990s for the lambda-bacteriophage [245, 246]. Time scale decomposition of these early models showed that their effective dynamic order was low [247] and that their dynamic structure was relatively condition invariant [248], motivating the development of constraint-based models that minimized the need for kinetic information [181].

After the first full genome sequences appeared, constraint-based models could be scaled up to the genome-scale [34]. As the first genome-scale models (GEMs) proved their ability to predict biological functions [34, 46], a vision was laid out in 2003 [86] that all cellular functions could be reconstructed in biochemical terms and seamlessly integrated. A decade later, some of this vision has been realized [51, 89, 53, 92, 97]. With these achievements, we can now assess what might be ahead with genome-scale models over the coming decade. We lay out some of our thoughts in this commentary.

6.3 The expanding scope of reconstructions: synthesis and function of the proteome

To date, stoichiometric reconstructions have largely focused on metabolism, resulting in genome-scale metabolic models, M-Models. The processes of enzyme

synthesis including transcription, translation, protein folding, complex formation, and prosthetic group integration were formalized in a gene expression reconstruction [91]. Protein translocation and localization pathways [92, 249] and DNA replication, repair, and cell division have also been reconstructed [97]. These networks have been merged with metabolic reconstructions to create integrated reaction networks [51, 89, 53, 92, 97] that formalize the primary chemical transformations that occur in cell (Figure 6.1A). Models integrating metabolism with protein expression are called ME-models.

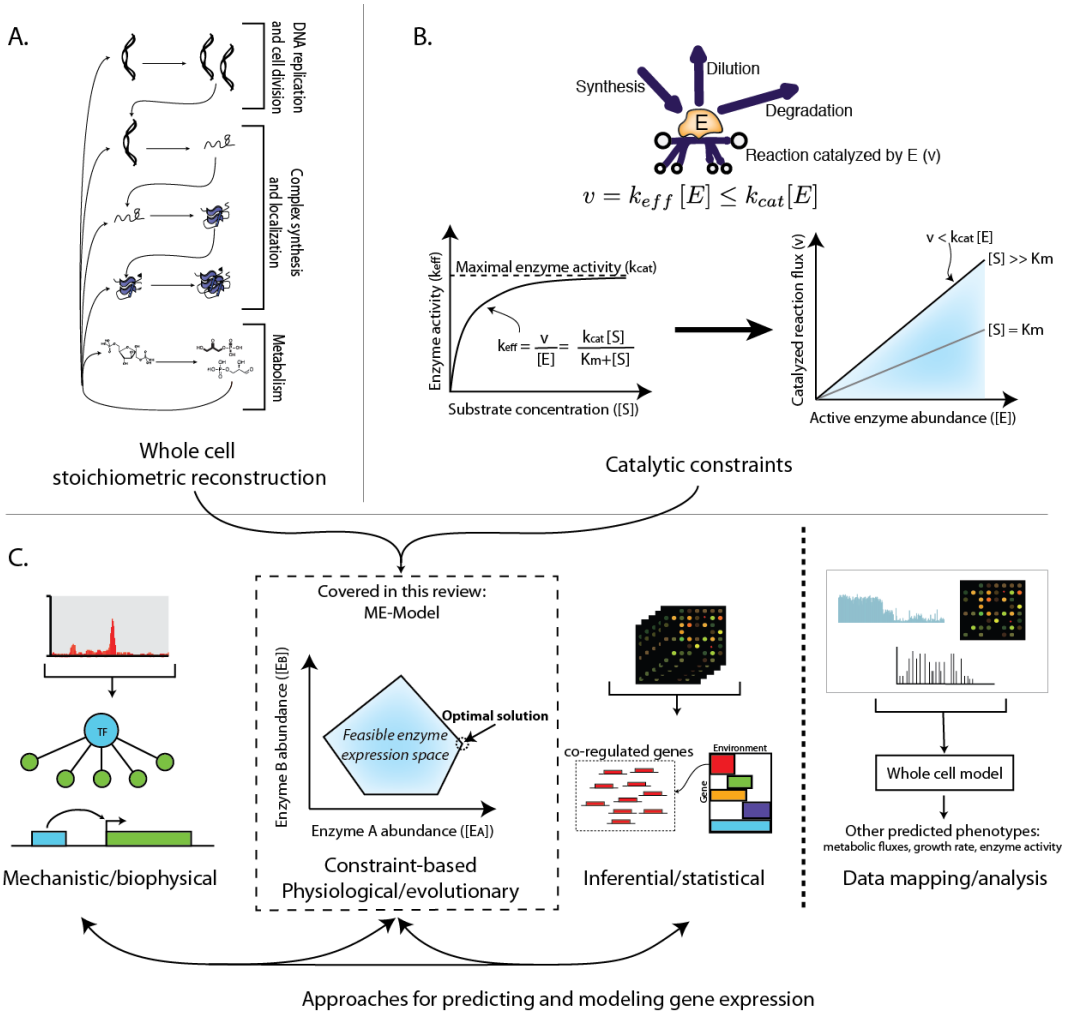
To enable prediction of biological phenotypes, stoichiometric reconstructions are combined with constraints on their operation. Stoichiometric networks are subject to (dynamic or steady-state) mass balance constraints on the production and consumption of molecules. For ME-Models, enzyme catalytic constraints are also necessary. In contrast to the typical use of kinetic equations to simulate system dynamics, the catalytic constraints in ME-Models are approximate stoichiometric relationships between enzyme abundance and catalyzed flux (Figure 6.1B). Adding these catalytic coupling constraints [91] enables the computation of feasible and optimal proteome and transcriptome states.

Most models encompassing gene expression have used measured expression states as a prerequisite for simulations. Often, gene expression measured under a particular condition is used to predict other molecular and physiological phenotypes [97, 250]. Alternatively, some approaches utilize gene expression data under environmental or genetic perturbations to build regulatory models [251]. These two approaches can be combined to predict molecular and physiological phenotypes subject to a transcriptional regulatory model [252, 80]. These are undoubtedly invaluable types of models and predictions; similar methods will likely be applied to ME-Models (Figure 6.1C).

ME-Models can predict gene expression with no previous input expression measurements: they can compute protein abundances that are required to (optimally) achieve integrated physiological functions (Figure 6.1C). Enzymes have an optimal expression

level subject to their (biosynthetic) cost and (physiological) benefit [137]. The ME-Model solves this cost-benefit optimization to compute genome-scale proteome states. Thus, compared to other methods for prediction and analysis of gene expression, predictions of gene expression in the ME-Model are based on fundamental constraints and optimality principles (as are the predictions of flux states in M-Models). ME-models can therefore be used to predict optimal expression and regulatory states.

Figure 6.1: **The expanding scope of reconstructions: synthesis and function of the proteome.**A) Stoichiometric reconstructions represent the chemical transformations that can occur in a cell and form the base of whole cell models. Recent reconstructions represent all of the major steps in the central dogma of molecular biology in biochemical detail [51, 89, 53, 92, 97]. B) Constraints on network operation are utilized to predict functional states. For reconstructions that encompass enzyme synthesis and function, catalytic constraints are necessary [91, 51]. Catalytic constraints relate enzyme abundance to its dilution (to daughter cells), degradation, and catalyzed flux with kinetic and/or thermodynamic relationships. C) Several general approaches exist to predict and model gene expression. Mechanistic/biophysical approaches first start from bottom-up reconstructions of transcription factor interactions and promoter architectures [253], aided by high-throughput data types [254]; models of regulatory logic can then be reconstructed and imposed [31, 255]. Constraint-based models are built from reconstructions of biochemical networks and constraints on their operation and can then be used to compute feasible and optimal physiological states [51, 89, 53, 92]. Importantly, the constraint-based approach enables prediction of gene expression states without any previous gene expression measurements. Inferential/statistical approaches are based on large gene expression datasets across environmental and genetic perturbations to identify co-regulated gene sets and their expression under novel perturbations. These general approaches can also be combined into hybrid models [80, 252]. Finally, we distinguish approaches that predict gene expression from those that use gene expression data from a particular state to predict other phenotypes [97, 250]another important capability of genome-scale models.


A.
DNA replication and cell division
Complex synthesis and localization
Metabolism
Whole cell stoichiometric reconstruction
B.
Synthesis
Dilution
Degradation
E
Reaction catalyzed by E (v)
$v = k_{eff}[E] \leq k_{cat}[E]$
Enzyme activity (keff)
Maximal enzyme activity (kcat)
keff = v/[E] = kcat[S]/Km+[S]
Substrate concentration ([S])
Catalyzed reaction flux (v)
v < kcat [E]
[S] >> Km
[S] = Km
Active enzyme abundance ([E])
Catalytic constraints
C.
TF
Mechanistic/biophysical
Covered in this review:
ME-Model
Enzyme B abundance ([EB])
Optimal solution
Feasible enzyme expression space
Enzyme A abundance ([EA])
Constraint-based
Physiological/evolutionary
co-regulated genes
Environment
Gene
Inferential/statistical
Whole cell model
Other predicted phenotypes:
metabolic fluxes, growth rate, enzyme activity
Data mapping/analysis
Approaches for predicting and modeling gene expression

6.4 Prediction of the molecular composition of a cell

An important distinction between M- and ME-Models is the prediction of cellular biomass composition. Instead of having protein and RNA biomass composition as an input (in the form of a biomass objective function [20]), biomass composition is an output that is predicted by ME-models. The expressed molecular machinery, such as the proteome, must support the integrated physiological functions of the cell. Rather than processes being coupled stoichiometrically through the biomass function, demands for vitamins and cofactors, chaperones, amino acids, nucleotides, tRNAs, etc. are derived directly from the computed proteome state.

Furthermore, a recent expansion of the ME-Model to include protein translocation enables predictions of a cells coarse-grained spatial organization [92]. Protein complexes are localized in cellular compartments required for enzyme function. With this expansion in scope, aspects of compartmentalized proteome abundance and molecular crowding can be assessed [92].

Thus, ME-Models advance the predictions possible with constraint-based models from network flux states to the spatially resolved molecular composition of a cell (Figure 6.2A).

6.5 Phenotypic effects of proteome allocation constraints

In addition to satisfying flux balance constraints, ME-Models are subject to proteome allocation constraints. While M-Models account for the operating expenses (i.e., metabolic requirements) to carry flux through pathways, ME-Models also account for the capital expenses (i.e., enzyme machinery) needed to catalyze all network reactions.

Therefore, in addition to cellular functions being limited by nutrients, they can also be limited by properties of the proteome (i.e., due to limited protein synthesis capacity and enzyme catalytic rates). Proteome allocation constraints govern integrated cell functions and, combined with growth-optimality assumptions, can explain several aspects of cell behavior not encompassed by previous models (Figure 6.2B).

First, the change in ribosomal protein abundance can be explained by growth-optimization subject to proteome allocation constraints [51, 53, 91, 90, 139]. At faster growth rates, more ribosomes are required to sustain the faster dilution of protein to daughter cells. Previous models have taken this growth rate dependent relationship as an observed (and subsequently assumed and fixed) phenomenological relationship [108], rather than a prediction.

Second, specific pathway shifts in central carbon metabolism from carbon-limited to carbon-excess environments (i.e., batch culture or non-carbon limitations), can be explained as a consequence of proteome allocation constraints. Specific pathway shifts from carbon-limited to proteome-limited growth are consistent with pathway shifts observed between chemostat and batch cultures [53]. Whereas previous approaches required the invocation of multiple competing objectives [47], now a single objective of maximal growth rate subject to proteome allocation constraints can explain the same phenomenon.

Third, the constraints limiting absolute growth rates and substrate uptake rates have remained elusive. Proteome-limitations in the ME-Model result in a maximal growth rate and optimal substrate uptake rate that is consistent with experimental data when nutrients are available in excess [53]. The limitations placed on substrate uptake by the proteome significantly expand the scope of environments that can be simulated with constraint-based models to include nutrient-excess and complex media conditions.

Fourth, spatial limitations on the membrane proteome further refine predictions

of pathway shifts and substrate uptake rates in nutrient-excess environments [92]. Limitations on protein synthesis and protein space result in similar phenotypic responses, but have some differences in enzyme utilization; membrane proteomics data and experimental evolution can help to illuminate which constraints are dominant.

The phenotypic effects of proteome allocation constraints are just beginning to be uncovered and will likely change our conception of optimal behavior and pathway use [256].

6.6 Gene expression states and molecular phenotypes can now be computed

The prediction of proteome composition is an ambitious endeavor. To date, a few predictions of absolute gene expression have been validated. The ME-Model accurately predicts ribosomal [51, 53] and translocase [92] protein abundances, which have well-known catalytic rates. In general, however, catalytic rates of specific enzymes in vivo are not known. Nonetheless, the relatively accurate prediction of overall proteome abundance of different cellular compartments and functional subsystems is possible [92]. Furthermore, genome-scale predicted and measured mRNA abundance correlate significantly [51]. These early predictions provide support for the genome-scale prediction of absolute gene expression from evolutionary optimality principles (Figure 6.2B).

As with false predictions from M-Models [27], discrepancies between predicted and observed expression levels have led to discovery (Figure 6.3A). First, the quantitative difference between predicted and measured gross RNA and protein biomass composition has led to the realization that translation rate is a hyperbolic function of growth rate [53], which has been independently validated [143]. Second, comparing predicted and measured abundance of functional subsystems identified the processes of

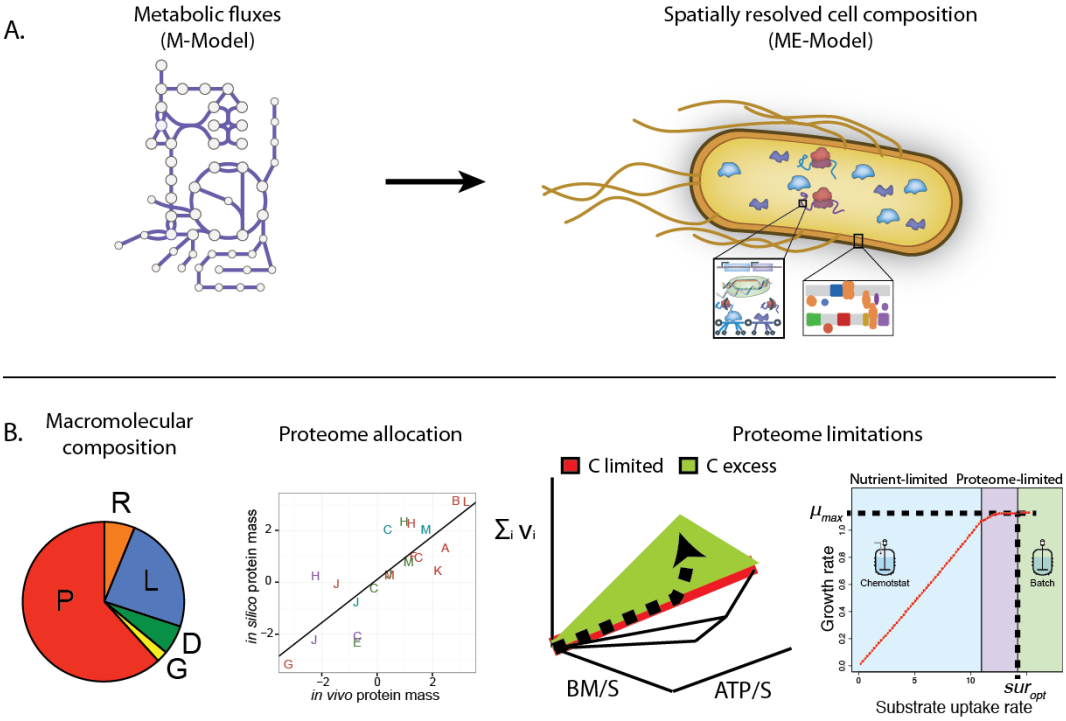


Figure 6.2: **Prediction of spatially resolved proteome allocation and limitations.** A) The expanded scope of reconstruction content advances the predictions possible with constraint-based models from network flux states (with M-Models) to the spatially resolved molecular composition of a cell (with ME-Models). B) ME-Models predict the gross macromolecular composition of the cell and the detailed allocation of the proteome. Additionally, the effects of proteome limitations can be accounted for, including the prediction of optimal substrate uptake rates and specific pathway shifts from carbon-limited to carbon-excess environments.

protein folding and metal ion and prosthetic group integration as under-predicted [92]. These under-predictions are consistent with known knowledge gaps of chaperone targets [257] and metal ion usage by proteins [258] and prioritizes these processes for further reconstruction.

Given the discordance between measured RNA and protein abundances [259], the moderate correlation between genome-scale predicted and measured gene product abundance is unsurprising. The factors contributing to the discrepancy between RNA and protein abundances are beginning to be uncovered [260, 261], aided by diverse data types on the various steps of gene expression, including promoter activity [262, 263], RNA abundance, RNA degradation rates, ribosome occupancy [163], and protein abundance [264]. These data types and gene-specific rates on the steps of gene expression can readily be integrated into ME-Models. Parameterizing the steps of gene expression with these data types, biophysical models [265, 266] or synthetic parts characterization [267, 268, 269, 270] will help understand the gap between RNA and protein abundance as well as in silico and in vivo gene expression levels.

Precise prediction of protein abundance is also limited by knowledge of enzyme catalytic rates. However, even though data on individual enzyme rates is noisy and sparse [271, 272], statistics on the distributions of catalytic rates are robust [164], enabling confident distributions of expression levels to be computed. Furthermore, model-driven approaches can be used to infer catalytic rates that are consistent with in vivo data [273, 274, 275]. These efforts will iteratively result in more precise predictions of protein expression.

Bottom-up prediction of gene expression will truly test our understanding of the biological demand and activity of enzymes. We believe that quantitative proteome levels will not fully be understood until we are able to predict them from the bottom-up with GEMs. Given the early successful uses of ME-models it seems clear that there is much

more discovery that lies ahead. Just as the community has become accustomed to flux balances, and thus the uses of metabolic networks, ME-models are likely to help us understand how the composition of the proteome is optimally balanced.

6.7 Defining and understanding regulatory needs

Prediction of regulatory needs during shifts in homeostatic states is another important challenge for ME-Models. Differential expression data is more abundantly available than absolute expression data, and will aide in ME-Model validation and model-driven discovery. We anticipate that this comparison and, more generally, a physiological needs perspective on gene expression will help reveal the principles underlying transcriptional regulation.

Recent examples of bottom-up prediction of differential expression include the use of an M-Model to predict transcriptional changes after redox shifts [253] and the use of a ME-Model to predict differential expression after a shift in carbon sources [51]. Furthermore, the principle of simplest pathway structure can predict gene co-expression and transcriptional regulatory relationships [4]. These examples provide evidence that transcriptional regulation is somewhat predictable based on optimality principles.

There are various reasons as to why transcriptional regulation may seem non-optimal [276]. First, there could be errors in the reaction network reconstruction, which can be rectified by systematic comparison of computational predictions and experimental data [27]. Second, discrepancies could be due to constraints or optimality principles that are not yet modeled or understood (such as proteome constraints added in moving from M- to ME-Models). Third, the environmental history of the organism may have coupled seemingly unrelated biological processes [159], or be optimized for fluctuating rather than static environments [277, 278, 279, 280]. Finally, it is likely that transcriptional regulation

is moderately efficient rather than perfectly optimal. Enumerating and classifying these discrepancies can drive biological discovery as has occurred through classifying the false predictions of gene essentiality [27, 36]. Identified discrepancies can provide insight into organismal physiology and prioritize the development of explicit transcriptional regulatory models.

Parallel to the prediction of transcriptional regulation with constraint-based models, a physiological perspective has revealed striking simplicity and optimality in transcriptional regulation [149, 253]. In several studies, the mass fractions of large protein subsystems and the activity of transcription factors have been shown to change linearly with growth rate or specific metabolic fluxes [131, 127, 138, 281]. Furthermore, linear models covering several genes have been shown to capture the variation in their expression with relative accuracy [263, 282]. The simplicity of these regulatory relationships (despite the complex topology and biophysical relationships [255] underlying regulatory networks) provides promise for accurate genome-scale regulatory models. However, the cross-talk and competition between transcription factors is still not generally understood; in the meantime, top-down approaches [251] may be necessary to capture the essence of these more complex relationships.

Importantly, the explicit representation of transcription in the ME-Model allows for the molecular details of transcription factor targets to be combined with the physiological principles underlying transcription factor activity. This will enable new approaches to model transcriptional regulation that move beyond binary representations of transcription factor activity [80]. Regulatory model development can be prioritized by the physiological importance of regulatory shifts and failure modes identified through comparison of predicted and measured differential expression.

Though we have focused on transcriptional regulation here, optimality and physiological principles will likely apply to translational (e.g., by sRNAs) and post-translational

regulation (e.g., by post-translational modifications and allosteric interactions) as well. These regulatory networks have received less attention, partially due to the difficulty in identifying the underlying interactions networks (compared to transcriptional regulatory networks [230]). However, new computational [283, 284] and experimental [285, 266] methods are emerging, and optimality principles are being uncovered [286, 287] to elucidate these regulatory networks. Like transcriptional regulation, we anticipate that the explicit representation of enzyme abundance and activity in ME-Models will aid in the genome-scale modeling of post-transcriptional regulation.

6.8 Seeking a comprehensive biophysical representation of cellular composition

The conceptual change in GEMs to enable the prediction of proteome abundance, localization, and limitations affords numerous opportunities for model application and expansion. The E. coli ME-Model currently encompasses 80% of the proteome and transcriptome by mass in environments of exponential growth [53]; this equates to 60% of the cells entire mass. While the requirements for biosynthesis of a whole cell are encompassed by the model, not all molecular abundances are predicted; gaps include: 1) the non-ME proteome, 2) the cell envelope, 3) metabolite concentrations, 4) DNA replication and gene copy number, and 5) glycogen. We briefly cover factors that may enable prediction of these molecular abundances deemed most important (Figure 6.3B).

Metabolite concentrations are beginning to be predicted with genome-scale models using thermodynamic considerations. By extending a method to ensure metabolic fluxes are thermodynamically feasible [76], thermodynamic constraints were used to predict steady-state metabolite concentrations that are consistent with a given flux state [288]. Later, an objective to minimize metabolite concentrations over the thermody-

namically feasible space was shown to increase prediction accuracy [289]. Additional constraints on osmolarity, metabolite toxicity, and correlations observed between metabolite concentrations and their enzyme affinities [124] and chemical properties [164] may improve predictions further. Then, the effects of these concentrations on enzyme and transcription factor activity may be accounted for in future genome-scale models.

A first requirement for the prediction of cell envelope composition is the prediction of cell size and shape, which determines the surface area of the cell that must be covered. The consistent growth rate dependence of cell size Donachie1987 suggests that simple principles may underlie the determination of cell size (for example, the balance between cytosolic and membrane proteome abundance [114]). However, the constraints underlying the exact composition of membrane lipid, glycans, LPS, and murein (together accounting for 15% of cell dry weight) are not well understood. Perhaps data on cell envelope composition will aide in understanding when and how it varies across environments and strains (an important characteristic of particular E. coli serotypes).

Expanding the proteome coverage to the remaining 20% of the proteome not encompassed by the ME-Model will require expansion of reconstruction content. Proteome abundance can be used to prioritize model expansion [290]. Many of these non-ME proteins can be broadly categorized as non-growth and stress response genes (e.g., biofilm and flagella formation and pH, osmolarity, and temperature responses). Therefore, modeling of stress responses is important to increase the coverage of the proteome and environments that can be simulated. Importantly, the ME-Model already accounts for the biosynthetic costs of synthesizing stress response proteins; however, the constraints imposed by environmental stresses will be needed to understand the proteins physiological benefit.

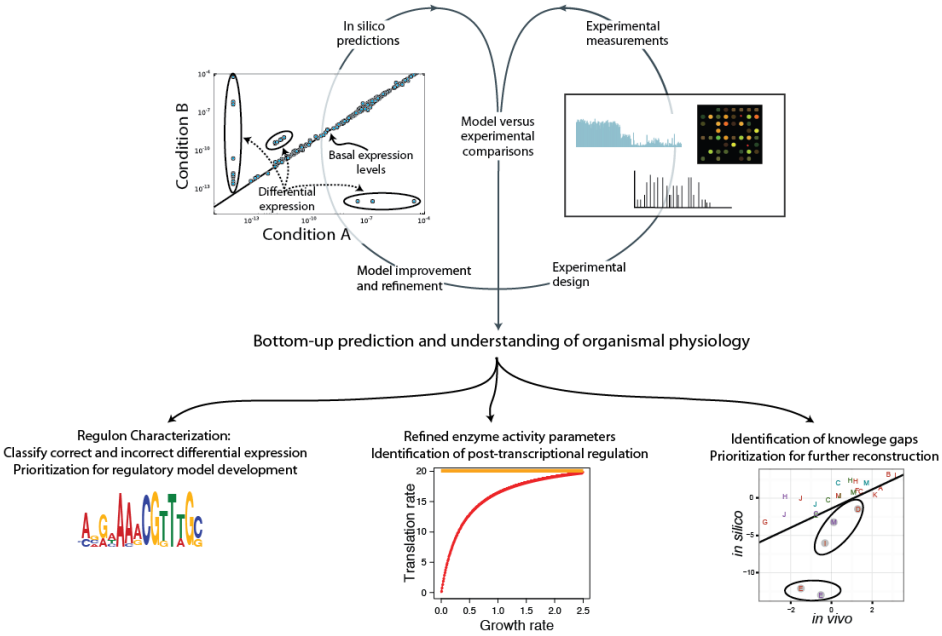
Protein structures will aide in formalizing the constraints on the proteome. Genome-scale models integrated with protein structures (GEM-PRO) enable simula-

tion of the effects of temperature: structure-based predictions of protein thermostability and the subsequent limitations on metabolic fluxes result in accurate predictions of growth and nutrient supplementations at high temperatures [68]. As other cellular stresses (e.g., pH) also affect protein catalytic capacity, protein structures may enable the simulation of other physiochemical stresses as well. Protein structures combined with ME-Models will also approach a more detailed biophysical representation of a cell. Spatial resolution can be refined further with protein-protein interaction data [291] (or prediction of protein-protein interactions with protein structures themselves [292]). Spatial considerations may be important for understanding co-localization of sequential catalytic steps [293, 294] and the effects of molecular crowding [295, 296] in the cytosol and membranes.

Figure 6.3: **Iterative model validation and biological discovery enabled by expanded scope.** A) The prediction of basal gene expression states and regulatory needs upon environmental shifts enables comparisons to gene expression datasets. Like M-Model predictions of gene essentiality [36] and metabolic flux [52], comparison of in silico and in vivo gene expression states will enable model validation and biological discovery. B) To increase the scope and resolution of predicted cellular composition and organization, there are several prioritized areas for model expansion. These include metabolite concentrations, the non-ME proteome, cell envelope composition, and the spatial organization and physiochemical constraints on the proteome (aided by protein structures).

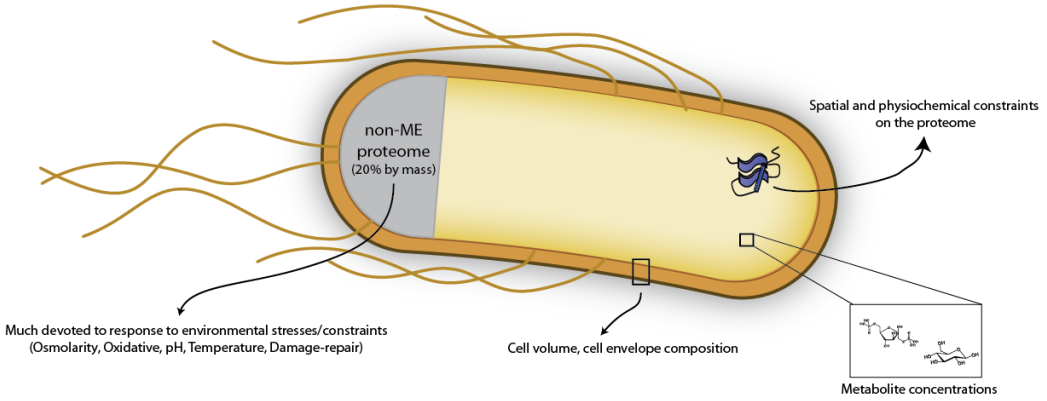
A.

Iterative comparison to gene expression and regulatory data
for model validation and biological discovery.



B.

Expanding the scope and resolution of cellular composition and constraints



6.9 Conclusion

Building whole cell computational models has been a long-standing goal. Genome-scale metabolic models, M-Models, have become widely used due to the numerous actionable predictions they can make [4], and the ease of draft model construction from readily available genome sequences and annotations [9]. Here we reviewed recent advancements expanding the scope of whole cell computational models to encompass the synthesis and localization of the proteome. The constraint-based philosophy underlying ME-Models parallels that of M-Models. However, the expanded scope of components and constraints enables the prediction of enzyme abundance and activity. Already, ME-Models have revealed how constraints on proteome allocation explain aspects of cell behavior that have remained elusive or require invocation of phenomenological relationships. Furthermore, several cases demonstrate that ME-Models enable accurate prediction of protein abundance and differential expression. As the basic capabilities of M-Models to predict flux states have led to numerous applications, future work will capitalize on the new capabilities of GEMs to compute proteome allocation and limitations. Optimality-based predictions will no doubt be imperfect, but they form a strong conceptual base to drive biological discovery, bioengineering, and further model development.

6.10 Acknowledgements

We thank Ali Ebrahim, Daniel Zielinski, Zak King, and Joshua Lerman for insightful discussions and critical feedback on the manuscript. EJO was supported by National Institute of Health R01 GM057089.

The text of Chapter 6 is a full reprint of the material as it appears in: OBrien E.J, Palsson B.O. Computing the functional proteome: recent progress and future prospects for genome-scale models. Curr. Opin. Biotechnol. 34C:125-134. (2015). The dissertation

author was the primary author of the manuscript. The other author was Bernard Ø. Palsson.

Bibliography

[1] J. S. Edwards and B. O. Palsson. Systems properties of the haemophilus influenzae rd metabolic genotype. *J Biol Chem*, 274(25):17410–6, 1999.

[2] M. A. Oberhardt, B. O. Palsson, and J. A. Papin. Applications of genome-scale metabolic reconstructions. *Molecular systems biology*, 5:320, 2009.

[3] D. McCloskey, B. O. Palsson, and A. M. Feist. Basic and applied uses of genome-scale metabolic network reconstructions of escherichia coli. *Molecular systems biology*, 9:661, 2013.

[4] A. Bordbar, J. M. Monk, Z. A. King, and B. O. Palsson. Constraint-based models predict metabolic and associated cellular functions. *Nat Rev Genet*, 15(2):107–20, 2014.

[5] J. Schellenberger, R. Que, R. M. Fleming, I. Thiele, J. D. Orth, A. M. Feist, D. C. Zielinski, A. Bordbar, N. E. Lewis, S. Rahmanian, J. Kang, D. R. Hyduke, and B. O. Palsson. Quantitative prediction of cellular metabolism with constraint-based models: the cobra toolbox v2.0. *Nature protocols*, 6(9):1290–307, 2011.

[6] A. Ebrahim, J. A. Lerman, B. O. Palsson, and D. R. Hyduke. Cobrapy: Constraints-based reconstruction and analysis for python. *BMC systems biology*, 7:74, 2013.

[7] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. Data, information, knowledge and principle: back to metabolism in kegg. *Nucleic Acids Res*, 42(Database issue):D199–205, 2014.

[8] I. M. Keseler, A. Mackie, M. Peralta-Gil, A. Santos-Zavaleta, S. Gama-Castro, C. Bonavides-Martinez, C. Fulcher, A. M. Huerta, A. Kothari, M. Krummenacker, M. Latendresse, L. Muniz-Rascado, Q. Ong, S. Paley, I. Schroder, A. G. Shearer, P. Subhraveti, M. Travers, D. Weerasinghe, V. Weiss, J. Collado-Vides, R. P. Gunsalus, I. Paulsen, and P. D. Karp. Ecocyc: fusing model organism databases with systems biology. *Nucleic Acids Res*, 41(Database issue):D605–12, 2013.

[9] C. S. Henry, M. DeJongh, A. A. Best, P. M. Frybarger, B. Linsay, and R. L. Stevens. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology*, 28(9):977–82, 2010.

[10] J. Schellenberger, J. O. Park, T. M. Conrad, and B. O. Palsson. Bigg: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC bioinformatics*, 11:213, 2010.

[11] S. G. Thorleifsson and I. Thiele. rbionet: A cobra toolbox extension for reconstructing high-quality biochemical networks. *Bioinformatics*, 27(14):2009–10, 2011.

[12] N. Swainston, K. Smallbone, P. Mendes, D. Kell, and N. Paton. The subliminal toolbox: automating steps in the reconstruction of metabolic networks. *Journal of integrative bioinformatics*, 8(2):186, 2011.

[13] R. Agren, L. Liu, S. Shoaie, W. Vongsangnak, I. Nookaew, and J. Nielsen. The raven toolbox and its use for generating a genome-scale metabolic model for penicillium chrysogenum. *PLoS computational biology*, 9(3):e1002980, 2013.

[14] I. Thiele and B. O. Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1):93–121, 2010.

[15] J. Monk, J. Nogales, and B. O. Palsson. Optimizing genome-scale network reconstructions. *Nature biotechnology*, 32(5):447–52, 2014.

[16] M. A. Oberhardt, J. Puchalka, V. A. Martins dos Santos, and J. A. Papin. Reconciliation of genome-scale metabolic reconstructions for comparative systems analysis. *PLoS computational biology*, 7(3):e1001116, 2011.

[17] Edward J OBrien, Jonathan M Monk, and Bernhard O Palsson. Using genome-scale models to predict biological capabilities. *Cell*, 161(5):971987, May 2015.

[18] J. L. Reed. Shrinking the metabolic solution space using experimental datasets. *PLoS computational biology*, 8(8):e1002662, 2012.

[19] J. D. Orth, I. Thiele, and B. O. Palsson. What is flux balance analysis? *Nature biotechnology*, 28(3):245–8, 2010.

[20] A. M. Feist and B. O. Palsson. The biomass objective function. *Current opinion in microbiology*, 13(3):344–9, 2010.

[21] N. E. Lewis, H. Nagarajan, and B. O. Palsson. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat Rev Microbiol*, 10(4):291–305, 2012.

[22] A. M. Feist and B. O. Palsson. The growing scope of applications of genome-scale metabolic reconstructions using escherichia coli. *Nature biotechnology*, 26(6):659–67, 2008.

[23] H. Yim, R. Haselbeck, W. Niu, C. Pujol-Baxley, A. Burgard, J. Boldt, J. Khandurina, J. D. Trawick, R. E. Osterhout, R. Stephen, J. Estadilla, S. Teisan, H. B. Schreyer, S. Andrae, T. H. Yang, S. Y. Lee, M. J. Burk, and S. Van Dien. Metabolic engineering of escherichia coli for direct production of 1,4-butanediol. *Nat Chem Biol*, 7(7):445–52, 2011.

[24] J. Adkins, S. Pugh, R. McKenna, and D. R. Nielsen. Engineering microbial chemical factories to produce renewable "biomonomers". *Front Microbiol*, 3:313, 2012.

[25] H. U. Kim, S. Y. Kim, H. Jeong, T. Y. Kim, J. J. Kim, H. E. Choy, K. Y. Yi, J. H. Rhee, and S. Y. Lee. Integrative genome-scale metabolic analysis of vibrio vulnificus for drug targeting and discovery. *Molecular systems biology*, 7:460, 2011.

[26] A. Bordbar, N. E. Lewis, J. Schellenberger, B. O. Palsson, and N. Jamshidi. Insight into human alveolar macrophage and m. tuberculosis interactions via metabolic reconstructions. *Molecular systems biology*, 6:422, 2010.

[27] J. D. Orth and B. O. Palsson. Systematizing the generation of missing metabolic knowledge. *Biotechnology and bioengineering*, 107(3):403–12, 2010.

[28] K. Nakahigashi, Y. Toya, N. Ishii, T. Soga, M. Hasegawa, H. Watanabe, Y. Takai, M. Honma, H. Mori, and M. Tomita. Systematic phenome analysis of escherichia coli multiple-knockout mutants reveals hidden reactions in central carbon metabolism. *Molecular systems biology*, 5:306, 2009.

[29] O. Rolfsson, B. O. Palsson, and I. Thiele. The human metabolic reconstruction recon 1 directs hypotheses of novel human metabolic functions. *BMC systems biology*, 5:155, 2011.

[30] B. Szappanos, K. Kovacs, B. Szamecz, F. Honti, M. Costanzo, A. Baryshnikova, G. Gelius-Dietrich, M. J. Lercher, M. Jelasity, C. L. Myers, B. J. Andrews, C. Boone, S. G. Oliver, C. Pal, and B. Papp. An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nat Genet*, 43(7):656–62, 2011.

[31] M. W. Covert, E. M. Knight, J. L. Reed, M. J. Herrgard, and B. O. Palsson. Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, 429(6987):92–6, 2004.

[32] D. Barua, J. Kim, and J. L. Reed. An automated phenotype-driven approach (geneforce) for refining metabolic and regulatory models. *PLoS Comput Biol*, 6(10):e1000970, 2010.

[33] R. U. Ibarra, J. S. Edwards, and B. O. Palsson. Escherichia coli k-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature*, 420(6912):186–9, 2002.

[34] J. S. Edwards and B. O. Palsson. The escherichia coli mg1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences of the United States of America*, 97(10):5528–33, 2000.

[35] N. Yamamoto, K. Nakahigashi, T. Nakamichi, M. Yoshino, Y. Takai, Y. Touda, A. Furubayashi, S. Kinjyo, H. Dose, M. Hasegawa, K. A. Datsenko, T. Nakayashiki, M. Tomita, B. L. Wanner, and H. Mori. Update on the keio collection of escherichia coli single-gene deletion mutants. *Molecular systems biology*, 5:335, 2009.

[36] J. Monk and B. O. Palsson. Genetics. predicting microbial growth. *Science*, 344(6191):1448–9, 2014.

[37] J. L. Reed, T. R. Patel, K. H. Chen, A. R. Joyce, M. K. Applebee, C. D. Herring, O. T. Bui, E. M. Knight, S. S. Fong, and B. O. Palsson. Systems approach to refining genome annotation. *Proceedings of the National Academy of Sciences of the United States of America*, 103(46):17480–4, 2006.

[38] N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, and B. O. Palsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America*, 104(6):1777–82, 2007.

[39] N. L. Fong, J. A. Lerman, I. Lam, B. O. Palsson, and P. Charusanti. Reconciling a salmonella enterica metabolic model with experimental data confirms that overexpression of the glyoxylate shunt can rescue a lethal ppc deletion mutant. *FEMS Microbiol Lett*, 342(1):62–9, 2013.

[40] C. Pal, B. Papp, and G. Posfai. The dawn of evolutionary genome engineering. *Nat Rev Genet*, 15(7):504–12, 2014.

[41] R. A. Notebaart, B. Szappanos, B. Kintses, F. Pal, A. Gyorkei, B. Bogos, V. Lazar, R. Spohn, B. Csorgo, A. Wagner, E. Ruppin, C. Pal, and B. Papp. Network-level architecture and the evolutionary potential of underground metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, 111(32):11762–7, 2014.

[42] Q. K. Beg, A. Vazquez, J. Ernst, M. A. de Menezes, Z. Bar-Joseph, A. L. Barabasi, and Z. N. Oltvai. Intracellular crowding defines the mode and sequence of substrate uptake by escherichia coli and constrains its metabolic activity. *Proceedings of the National Academy of Sciences of the United States of America*, 104(31):12663–8, 2007.

[43] K. Zhuang, G. N. Vemuri, and R. Mahadevan. Economics of membrane occupancy and respiro-fermentation. *Molecular systems biology*, 7:500, 2011.

[44] O. Shoval, H. Sheftel, G. Shinar, Y. Hart, O. Ramote, A. Mayo, E. Dekel, K. Kavanagh, and U. Alon. Evolutionary trade-offs, pareto optimality, and the geometry of phenotype space. *Science*, 336(6085):1157–60, 2012.

[45] Z. A. King, C. J. Lloyd, A. M. Feist, and B. O. Palsson. Next-generation genome-scale models for metabolic engineering. *Current opinion in biotechnology*, 35C:23–29, 2015.

[46] J. S. Edwards, R. U. Ibarra, and B. O. Palsson. In silico predictions of escherichia coli metabolic capabilities are consistent with experimental data. *Nature biotechnology*, 19(2):125–30, 2001.

[47] R. Schuetz, N. Zamboni, M. Zampieri, M. Heinemann, and U. Sauer. Multidimensional optimality of microbial metabolism. *Science*, 336(6081):601–4, 2012.

[48] R. Mahadevan and C. H. Schilling. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic engineering*, 5(4):264–76, 2003.

[49] S. S. Fong, A. R. Joyce, and B. O. Palsson. Parallel adaptive evolution cultures of escherichia coli lead to convergent growth phenotypes with different gene expression states. *Genome research*, 15(10):1365–72, 2005.

[50] N. E. Lewis, K. K. Hixson, T. M. Conrad, J. A. Lerman, P. Charusanti, A. D. Polpitiya, J. N. Adkins, G. Schramm, S. O. Purvine, D. Lopez-Ferrer, K. K. Weitz, R. Eils, R. Konig, R. D. Smith, and B. O. Palsson. Omic data from evolved e. coli are consistent with computed optimal growth from genome-scale models. *Molecular systems biology*, 6:390, 2010.

[51] J. A. Lerman, D. R. Hyduke, H. Latif, V. A. Portnoy, N. E. Lewis, J. D. Orth, A. C. Schrimpe-Rutledge, R. D. Smith, J. N. Adkins, K. Zengler, and B. O. Palsson. In silico method for modelling metabolism and gene product expression at genome scale. *Nature communications*, 3:929, 2012.

[52] R. Schuetz, L. Kuepfer, and U. Sauer. Systematic evaluation of objective functions for predicting intracellular fluxes in escherichia coli. *Molecular systems biology*, 3:119, 2007.

[53] E. J. O'Brien, J. A. Lerman, R. L. Chang, D. R. Hyduke, and B. O. Palsson. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Molecular systems biology*, 9:693, 2013.

[54] S. Stolyar, S. Van Dien, K. L. Hillesland, N. Pinel, T. J. Lie, J. A. Leigh, and D. A. Stahl. Metabolic modeling of a mutualistic microbial community. *Molecular systems biology*, 3:92, 2007.

[55] H. Nagarajan, M. Embree, A. E. Rotaru, P. M. Shrestha, A. M. Feist, B. O. Palsson, D. R. Lovley, and K. Zengler. Characterization and modelling of interspecies electron transfer mechanisms and microbial community dynamics of a syntrophic association. *Nature communications*, 4:2809, 2013.

[56] E. H. Wintermute and P. A. Silver. Emergent cooperation in microbial metabolism. *Molecular systems biology*, 6:407, 2010.

[57] W. R. Harcombe, W. J. Riehl, I. Dukovski, B. R. Granger, A. Betts, A. H. Lang, G. Bonilla, A. Kar, N. Leiby, P. Mehta, C. J. Marx, and D. Segre. Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell reports*, 7(4):1104–15, 2014.

[58] H. Nam, N. E. Lewis, J. A. Lerman, D. H. Lee, R. L. Chang, D. Kim, and B. O. Palsson. Network context and selection in the evolution to enzyme specificity. *Science*, 337(6098):1101–4, 2012.

[59] G. Plata, C. S. Henry, and D. Vitkup. Long-term phenotypic evolution of bacteria. *Nature*, 517(7534):369–72, 2015.

[60] M. Cvijovic, S. Bordel, and J. Nielsen. Mathematical models of cell factories: moving towards the core of industrial biotechnology. *Microbial biotechnology*, 4(5):572–84, 2011.

[61] S. Ranganathan, P. F. Suthers, and C. D. Maranas. Optforce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. *PLoS computational biology*, 6(4):e1000744, 2010.

[62] A. P. Burgard, P. Pharkya, and C. D. Maranas. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and bioengineering*, 84(6):647–57, 2003.

[63] N. Tepper and T. Shlomi. Computational design of auxotrophy-dependent microbial biosensors for combinatorial metabolic engineering experiments. *PloS one*, 6(1):e16274, 2011.

[64] T. D. Scheibe, R. Mahadevan, Y. Fang, S. Garg, P. E. Long, and D. R. Lovley. Coupling a genome-scale metabolic model with a reactive transport model to describe in situ uranium bioremediation. *Microbial biotechnology*, 2(2):274–86, 2009.

[65] K. Zhuang, M. Izallalen, P. Mouser, H. Richter, C. Risso, R. Mahadevan, and D. R. Lovley. Genome-scale dynamic modeling of the competition between rhodoferax and geobacter in anoxic subsurface environments. *The ISME journal*, 5(2):305–16, 2011.

[66] N. Klitgord and D. Segre. Environments that induce synthetic microbial ecosystems. *PLoS computational biology*, 6(11):e1001002, 2010.

[67] R. Levy and E. Borenstein. Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. *Proceedings of the National Academy of Sciences of the United States of America*, 110(31):12804–9, 2013.

[68] R. L. Chang, K. Andrews, D. Kim, Z. Li, A. Godzik, and B. O. Palsson. Structural systems biology evaluation of metabolic thermotolerance in escherichia coli. *Science*, 340(6137):1220–3, 2013.

[69] T. C. Williams, M. G. Poolman, A. J. Howden, M. Schwarzlander, D. A. Fell, R. G. Ratcliffe, and L. J. Sweetlove. A genome-scale metabolic model accurately predicts fluxes in central carbon metabolism under stress conditions. *Plant physiology*, 154(1):311–23, 2010.

[70] E. Collakova, J. Y. Yen, and R. S. Senger. Are we ready for genome-scale modeling in plants? *Plant science : an international journal of experimental plant biology*, 191-192:53–70, 2012.

[71] R. L. Chang, L. Xie, P. E. Bourne, and B. O. Palsson. Drug off-target effects predicted using structural analysis in the context of a metabolic network model. *PLoS computational biology*, 6(9):e1000938, 2010.

[72] L. Jerby, T. Shlomi, and E. Ruppin. Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Molecular systems biology*, 6:401, 2010.

[73] S. A. Becker and B. O. Palsson. Context-specific metabolic networks are consistent with experiments. *PLoS computational biology*, 4(5):e1000082, 2008.

[74] B. Palsson. The challenges of in silico biology. *Nature biotechnology*, 18(11):1147–50, 2000.

[75] M. L. Mo, B. O. Palsson, and M. J. Herrgard. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC systems biology*, 3:37, 2009.

[76] C. S. Henry, L. J. Broadbelt, and V. Hatzimanikatis. Thermodynamics-based metabolic flux analysis. *Biophysical journal*, 92(5):1792–805, 2007.

[77] N. Zamboni, S. M. Fendt, M. Ruhl, and U. Sauer. (13)c-based metabolic flux analysis. *Nature protocols*, 4(6):878–92, 2009.

[78] T. Shlomi, M. N. Cabili, M. J. Herrgard, B. O. Palsson, and E. Ruppin. Network-based prediction of human tissue-specific metabolism. *Nature biotechnology*, 26(9):1003–10, 2008.

[79] B. J. Schmidt, A. Ebrahim, T. O. Metz, J. N. Adkins, B. O. Palsson, and D. R. Hyduke. Gim3e: condition-specific models of cellular metabolism developed from metabolomics and expression data. *Bioinformatics*, 29(22):2900–8, 2013.

[80] S. Chandrasekaran and N. D. Price. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in escherichia coli and mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences of the United States of America*, 107(41):17845–50, 2010.

[81] I. Thiele, N. Swainston, R. M. Fleming, A. Hoppe, S. Sahoo, M. K. Aurich, H. Haraldsdottir, M. L. Mo, O. Rolfsson, M. D. Stobbe, S. G. Thorleifsson, R. Agren, C. Bolling, S. Bordel, A. K. Chavali, P. Dobson, W. B. Dunn, L. Endler, D. Hala, M. Hucka, D. Hull, D. Jameson, N. Jamshidi, J. J. Jonsson, N. Juty, S. Keating, I. Nookaew, N. Le Novere, N. Malys, A. Mazein, J. A. Papin, N. D. Price, Sr. Selkov, E., M. I. Sigurdsson, E. Simeonidis, N. Sonnenschein, K. Smallbone, A. Sorokin, J. H. van Beek, D. Weichart, I. Goryanin, J. Nielsen, H. V. Westerhoff, D. B. Kell, P. Mendes, and B. O. Palsson. A community-driven global reconstruction of human metabolism. *Nature biotechnology*, 31(5):419–25, 2013.

[82] R. Agren, S. Bordel, A. Mardinoglu, N. Pornputtapong, I. Nookaew, and J. Nielsen. Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using init. *PLoS computational biology*, 8(5):e1002518, 2012.

[83] J. Schellenberger and B. O. Palsson. Use of randomized sampling for analysis of metabolic networks. *The Journal of biological chemistry*, 284(9):5457–61, 2009.

[84] N. E. Lewis, G. Schramm, A. Bordbar, J. Schellenberger, M. P. Andersen, J. K. Cheng, N. Patel, A. Yee, R. A. Lewis, R. Eils, R. Konig, and B. O. Palsson. Large-scale in silico modeling of metabolic interactions between cell types in the human brain. *Nature biotechnology*, 28(12):1279–85, 2010.

[85] Z. A. King and A. Ebrahim. escher: Escher 1.0.0 beta 3. *ZENODO*, 2014.

[86] J. L. Reed and B. O. Palsson. Thirteen years of building constraint-based in silico models of escherichia coli. *Journal of bacteriology*, 185(9):2692–9, 2003.

[87] Y. Zhang, I. Thiele, D. Weekes, Z. Li, L. Jaroszewski, K. Ginalski, A. M. Deacon, J. Wooley, S. A. Lesley, I. A. Wilson, B. Palsson, A. Osterman, and A. Godzik. Three-dimensional structural view of the central metabolic network of thermotoga maritima. *Science*, 325(5947):1544–9, 2009.

[88] R. L. Chang, L. Xie, P. E. Bourne, and B. O. Palsson. Antibacterial mechanisms identified through structural systems pharmacology. *BMC systems biology*, 7:102, 2013.

[89] I. Thiele, R. M. Fleming, R. Que, A. Bordbar, D. Diep, and B. O. Palsson. Multiscale modeling of metabolism and macromolecular synthesis in e. coli and its application to the evolution of codon usage. *PloS one*, 7(9):e45635, 2012.

[90] I. Thiele, N. Jamshidi, R. M. Fleming, and B. O. Palsson. Genome-scale reconstruction of escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS computational biology*, 5(3):e1000312, 2009.

[91] I. Thiele, R. M. Fleming, A. Bordbar, J. Schellenberger, and B. O. Palsson. Functional characterization of alternate optimal solutions of escherichia coli's transcriptional and translational machinery. *Biophysical journal*, 98(10):2072–81, 2010.

[92] J. K. Liu, E. J. OBrien, J. A. Lerman, K. Zengler, B. O. Palsson, and Feist A. M. Reconstruction and modeling protein translocation and compartmentalization in escherichia coli at the genome-scale. *BMC Systems Biology*, 2014.

[93] Y. Sun, R. M. Fleming, I. Thiele, and M. A. Saunders. Robust flux balance analysis of multiscale biochemical reaction networks. *BMC bioinformatics*, 14:240, 2013.

[94] E. J. O'Brien and B. O. Palsson. Computing the functional proteome: recent progress and future prospects for genome-scale models. *Current opinion in biotechnology*, 34C:125–134, 2015.

[95] D W Selinger, M A Wright, and G M Church. On the complete determination of biological systems. *Trends Biotechnol*, 21:251–254, 2003.

[96] D Machado, R Costa, M Rocha, E Ferreira, B Tidor, and I Rocha. Modeling formalisms in Systems Biology. *AMB Expr*, 1:1–14, 2011.

[97] J. R. Karr, J. C. Sanghvi, D. N. Macklin, M. V. Gutschow, J. M. Jacobs, Jr. Bolival, B., N. Assad-Garcia, J. I. Glass, and M. W. Covert. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401, 2012.

[98] Jeffrey D Orth, Tom M Conrad, Jessica Na, Joshua A Lerman, Hojung Nam, Adam M Feist, and Bernhard ØPalsson. A comprehensive genome-scale reconstruction of Escherichia coli metabolism 2011. *Molecular Systems Biology*, 7:535, 2011.

[99] M Riley, T Abe, M B Arnaud, M K Berlyn, F R Blattner, R R Chaudhuri, J D Glasner, T Horiuchi, I M Keseler, T Kosuge, H Mori, N T Perna, Gr Plunkett, K E Rudd, M H Serres, G H Thomas, N R Thomson, D Wishart, and B L Wanner. Escherichia coli K-12: a cooperatively developed annotation snapshot–2005. *Nucleic Acids Res*, 34:1–9, 2006.

[100] J Kato and M Hashimoto. Construction of consecutive deletions of the Escherichia coli chromosome. *Mol Syst Biol*, 3:132, 2007.

[101] R Gil, F J Silva, J Pereto, and A Moya. Determination of the core of a minimal bacterial gene set. *Microbiol. Mol. Biol. Rev.*, 68:518–537, 2004.

[102] W D Donachie and A C Robinson. Cell division: parameter values and the process. *Escherichia coli and Salmonella typhimurium. Cellular and Molecular Biology*, Vol. I and II:1578–1593, 1987.

[103] K V Meyenburg and F G Hansen. Regulation of chromosome replication. *Escherichia coli and Salmonella typhimurium. Cellular and Molecular Biology*, Vol. I and II:1555–1577, 1987.

[104] H Bremer and P P Dennis. Modulation of chemical composition and other parameters of the cell by growth rate. *Escherichia coli and Salmonella*, pages 1553–1569, 1996.

[105] S J Pirt. The maintenance energy of bacteria in growing cultures. *Proc R Soc Lond B Biol Sci*, 163:224–231, 1965.

[106] O M Neijssel, MJTd Mattos, and D W Tempest. Growth Yield and Energy Distribution. *Escherichia coli and Salmonella*, pages 1683–1692, 1996.

[107] R Young and H Bremer. Polypeptide-chain-elongation rate in Escherichia coli B/r as a function of growth rate. *Biochem J*, 160:185, 1976.

[108] M Scott, C W Gunderson, E M Mateescu, Z Zhang, and T Hwa. Interdependence of cell growth and gene expression: origins and consequences. *Science*, 330:1099–1102, 2010.

[109] S Proshkin, A R Rahmouni, A Mironov, and E Nudler. Cooperation between translating ribosomes and RNA polymerase in transcription elongation. *Science*, 328:504–508, 2010.

[110] K Valgepea, K Adamberg, A Seiman, and R Vilu. Escherichia coli achieves faster growth by increasing catalytic and translation rates of proteins. *Mol Biosyst*, 9:2344–2358, 2013.

[111] V M Boer, C A Crutchfield, P H Bradley, D Botstein, and J D Rabinowitz. Growth-limiting intracellular metabolites in yeast growing under diverse nutrient limitations. *Mol Biol Cell*, 21:198–211, 2010.

[112] R W O/'Brien, O M Neijssel, and D W Tempest. Glucose phosphoenolpyruvate phosphotransferase activity and glucose uptake rate of Klebsiella aerogenes growing in chemostat culture. *J Gen Microbiol*, 116:305–314, 1980.

[113] R W Smith and A C Dean. Beta-galactosidase synthesis in Klebsiella aerogenes growing in continuous culture. *J Gen Microbiol*, 72:37–47, 1972.

[114] D Molenaar, R van Berlo, D de Ridder, and B Teusink. Shifts in growth strategies reflect tradeoffs in cellular economics. *Mol Syst Biol*, 5:323, 2009.

[115] M Kim, Z Zhang, H Okano, D Yan, A Groisman, and T Hwa. Need-based activation of ammonium uptake in Escherichia coli. *Mol Syst Biol*, 8:616, 2012.

[116] J Monod. The growth of bacterial cultures. *Annu Rev Microbiol*, 3:371–394, 1949.

[117] A L Koch. Microbial physiology and ecology of slow growth. *Microbiol Mol Biol Rev*, 61:305–318, 1997.

[118] R Adadi, B Volkmer, R Milo, M Heinemann, and T Shlomi. Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. *PLoS Comput Biol*, 8:e1002575, 2012.

[119] D K Button. Biochemical basis for whole-cell uptake kinetics: specific affinity, oligotrophic capacity, and the meaning of the Michaelis constant. *Appl Environ Microbiol*, 57:2033–2038, 1991.

[120] Qiang Hua, Chen Yang, Taku Oshima, Hirotada Mori, and Kazuyuki Shimizu. Analysis of gene expression in escherichia coli in response to changes of growth-limiting nutrient in chemostat cultures. *Applied and environmental microbiology*, 70(4):2354–2366, 2004.

[121] A Nanchen, A Schicker, and U Sauer. Nonlinear dependency of intracellular fluxes on growth rate in miniaturized continuous cultures of Escherichia coli. *Appl Environ Microbiol*, 72:1164–1172, 2006.

[122] G N Vemuri, E Altman, D P Sangurdekar, A B Khodursky, and M A Eiteman. Overflow metabolism in Escherichia coli during steady-state growth: transcriptional regulation and effect of the redox ratio. *Appl Environ Microbiol*, 72:3653–3661, 2006.

[123] R Nahku, K Valgepea, P-J Lahtvee, S Erm, K Abner, K Adamberg, and R Vilu. Specific growth rate dependent transcriptome profiling of Escherichia coli [bsol]{K12[bsol]} [bsol]{MG1655[bsol]} in accelerostat cultures. *J Biotechnol*, 145:60–65, 2010.

[124] B D Bennett, E H Kimball, M Gao, R Osterhout, S J Van Dien, and J D Rabinowitz. Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. *Nat Chem Biol*, 5:593–599, 2009.

[125] G W Li, E Oh, and J S Weissman. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature*, 484:538–541, 2012.

[126] T Tuller, A Carmi, K Vestsigian, S Navon, Y Dorfan, J Zaborske, T Pan, O Dahan, I Furman, and Y Pilpel. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, 141:344–354, 2010.

[127] S Klumpp, Z Zhang, and T Hwa. Growth rate-dependent global effects on gene expression in bacteria. *Cell*, 139:1366–1375, 2009.

[128] B K Cho, S A Federowicz, M Embree, Y S Park, D Kim, and B O Palsson. The PurR regulon in Escherichia coli K-12 MG1655. *Nucleic Acids Res*, 39:6456–6464, 2011.

[129] S Berthoumieux, H de Jong, G Baptist, C Pinel, C Ranquet, D Ropers, and J Geiselmann. Shared control of gene expression in bacteria by transcription factors and global physiology of the cell. *Mol Syst Biol*, 9:634, 2013.

[130] L Gerosa, K Kochanowski, M Heinemann, and U Sauer. Dissecting specific and global transcriptional regulation of bacterial gene expression. *Mol Syst Biol*, 9:658, 2013.

[131] S Klumpp and T Hwa. Growth-rate-dependent partitioning of RNA polymerases in bacteria. *Proc Natl Acad Sci USA*, 105:20245–20250, 2008.

[132] B K Cho, E M Knight, and B O Palsson. Transcriptional regulation of the fad regulon genes of Escherichia coli by ArcA. *Microbiology*, 152:2207–2219, 2006.

[133] B R Haverkorn van Rijsewijk, A Nanchen, S Nallet, R J Kleijn, and U Sauer. Large-scale 13C-flux analysis reveals distinct transcriptional control of respiratory and fermentative metabolism in Escherichia coli. *Mol Syst Biol*, 7:477, 2011.

[134] W R Harcombe, N F Delaney, N Leiby, N Klitgord, and C J Marx. The ability of flux balance analysis to predict evolution of central metabolism scales with the initial distance to the optimum. *PLoS Comput Biol*, 9:e1003091, 2013.

[135] F Crick. Project K: The Complete Solution of E. coli. *Perspect. Biol. Med.*, 17:67–70, 1973.

[136] P Langfelder and S Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559, 2008.

[137] E. Dekel and U. Alon. Optimality and evolutionary tuning of the expression level of a protein. *Nature*, 436(7050):588–92, 2005.

[138] C. You, H. Okano, S. Hui, Z. Zhang, M. Kim, C. W. Gunderson, Y. P. Wang, P. Lenz, D. Yan, and T. Hwa. Coordination of bacterial proteome with metabolism by cyclic amp signalling. *Nature*, 500(7462):301–6, 2013.

[139] M. Scott, S. Klumpp, E. M. Mateescu, and T. Hwa. Emergence of robust growth laws from optimal regulation of ribosome synthesis. *Mol Syst Biol*, 10:747, 2014.

[140] Andrea Y Weie, Diego A Oyarzn, Vincent Danos, and Peter S Swain. Mechanistic links between cellular trade-offs, gene expression, and growth. *Proceedings of the National Academy of Sciences of the United States of America*, 112(9):E103847, Mar 2015.

[141] Alexander Schmidt, Silke Vedelaar, Erik Ahrnem, Benjamin Volkmer, Karl Kochanowski, Luciano Callipo, Kvin Knoops, Manuel Bauer, Ruedi Aebersold, and Matthias Heinemann. The quantitative and condition-dependent escherichia coli proteome. Submitted.

[142] Jennifer L Reed and Bernhard Palsson. Genome-scale in silico models of e. coli have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. *Genome research*, 14(9):17971805, Sep 2004.

[143] S. Klumpp, M. Scott, S. Pedersen, and T. Hwa. Molecular crowding limits translation and cell growth. *Proc Natl Acad Sci U S A*, 110(42):16754–9, 2013.

[144] Laurence Yang, Justin Tan, Edward J OBrien, Jonathan M Monk, Donghyuk Kim, Howard J Li, Pep Charusanti, Ali Ebrahim, Colton J Lloyd, James T Yurkovich, and et al. A systems biology definition of the core proteome of metabolism and expression is consistent with high-throughput data. *Proceedings of the National Academy of Sciences*, 2015.

[145] Irit Shachrai, Alon Zaslaver, Uri Alon, and Erez Dekel. Cost of unneeded proteins in e. coli is reduced after several generations in exponential growth. *Molecular cell*, 38(5):758767, Jun 2010.

[146] Anisha Goel, Thomas H Eckhardt, Pranav Puri, Anne de Jong, Filipe Branco Dos Santos, Martin Giera, Fabrizia Fusetti, Willem M de Vos, Jan Kok, Bert Poolman, and et al. Protein costs do not explain evolution of metabolic strategies and regulation of ribosomal content: does protein investment explain an anaerobic bacterial crabtree effect? *Molecular microbiology*, 97(1):7792, Jul 2015.

[147] Olga T Schubert, Christina Ludwig, Maria Kogadeeva, Michael Zimmermann, George Rosenberger, Martin Gengenbacher, Ludovic C Gillet, Ben C Collins, Hannes L Rst, Stefan H E Kaufmann, and et al. Absolute proteome composition and dynamics during dormancy and resuscitation of mycobacterium tuberculosis. *Cell host & microbe*, Jun 2015.

[148] Tomohiro Shimada, Nobuyuki Fujita, Kaneyoshi Yamamoto, and Akira Ishihama. Novel roles of camp receptor protein (crp) in regulation of transport and metabolism of carbon sources. *PloS one*, 6(6):e20081, Jun 2011.

[149] V. Chubukov, L. Gerosa, K. Kochanowski, and U. Sauer. Coordination of microbial metabolism. *Nature reviews. Microbiology*, 12(5):327–40, 2014.

[150] Larry Reitzer. Nitrogen assimilation and global regulation in escherichia coli. *Annual review of microbiology*, 57:155176, May 2003.

[151] Jose Utrilla, Ke Chen Edward J. OBrien, Douglas McCloskey, Jacky Cheung, Harris Wang, Dagoberto Armenta-Medina, Adam M Feist, and Bernhard O Palsson. Proteome and energy re-allocation by adaptive regulatory mutations reveals a fitness trade-off. Submitted.

[152] Michael Berney, Hans-Ulrich Weilenmann, Julian Ihssen, Claudio Bassin, and Thomas Egli. Specific growth rate determines the sensitivity of escherichia coli to thermal, uva, and solar disinfection. *Applied and environmental microbiology*, 72(4):25862593, Apr 2006.

[153] Roland Lindqvist and Gunilla Barmark. Specific growth rate determines the sensitivity of escherichia coli to lactic acid stress: implications for predictive microbiology. *BioMed research international*, 2014:471317, Jul 2014.

[154] Keith Poole. Stress responses as determinants of antimicrobial resistance in gram-negative bacteria. *Trends in microbiology*, 20(5):227234, May 2012.

[155] Jue Wang, Esha Atolia, Bo Hua, Yonatan Savir, Renan Escalante-Chong, and Michael Springer. Natural variation in preparation for nutrient depletion reveals a costbenefit tradeoff. *PLOS Biology*, 13:e1002041–e1002041, 2015.

[156] Ana Solopova, Jordi van Gestel, Franz J. Weissing, Herwig Bachmann, Bas Teusink, Jan Kok, and Oscar P. Kuipers. Bet-hedging during bacterial diauxic shift. *Proceedings of the National Academy of Sciences of the United States of America*, 2014.

[157] Ophelia S. Venturelli, Ignacio Zuleta, Richard M. Murray, and Hana El-Samad. Population diversification in a yeast metabolic program promotes anticipation of environmental shifts. *PLOS Biology*, 13(1):e1002042–e1002042, 2015.

[158] Ilias Tagkopoulos, Yir-Chung Liu, and Saeed Tavazoie. Predictive behavior within microbial genetic networks. *Science*, 320(5881):13131317, Jun 2008.

[159] A. Mitchell, G. H. Romano, B. Groisman, A. Yona, E. Dekel, M. Kupiec, O. Dahan, and Y. Pilpel. Adaptive prediction of environmental changes by microorganisms. *Nature*, 460(7252):220–4, 2009.

[160] Theodore J Perkins and Peter S Swain. Strategies for cellular decision-making. *Molecular systems biology*, 5:326, Nov 2009.

[161] R Kassen. The experimental evolution of specialists, generalists, and the maintenance of diversity. *Journal of evolutionary biology*, 15(2):173190, Mar 2002.

[162] Benjamin Volkmer and Matthias Heinemann. Condition-dependent cell volume and concentration of escherichia coli to facilitate data conversion for systems biology modeling. *PloS one*, 6(7):e23126, Jul 2011.

[163] G. W. Li, D. Burkhardt, C. Gross, and J. S. Weissman. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, 157(3):624–35, 2014.

[164] A. Bar-Even, E. Noor, Y. Savir, W. Liebermeister, D. Davidi, D. S. Tawfik, and R. Milo. The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry*, 50(21):4402–10, 2011.

[165] H. Latif, J. A. Lerman, V. A. Portnoy, Y. Tarasova, H. Nagarajan, A. C. Schrimpe-Rutledge, R. D. Smith, J. N. Adkins, D. H. Lee, Y. Qiu, and K. Zengler. The genome organization of thermotoga maritima reflects its lifestyle. *PLoS Genet*, 9(4):e1003485, 2013.

[166] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357359, Apr 2012.

[167] Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with rna-seq. *Nature biotechnology*, 31(1):4653, Jan 2013.

[168] Sewall Wright. *The roles of mutation, inbreeding, crossbreeding, and selection in evolution*, volume 1, page 356366. 1932.

[169] Tanguy Chouard. Darwin 200: Beneath the surface. *Nature*, 456(7220):300303, Nov 2008.

[170] S Ciliberti, O C Martin, and A Wagner. Innovation and robustness in complex regulatory gene networks. *Proceedings of the National Academy of Sciences of the United States of America*, 104(34):1359113596, Aug 2007.

[171] A Varma and B O Palsson. Metabolic capabilities of escherichia coli: I. synthesis of biosynthetic precursors and cofactors. *Journal of theoretical biology*, 165(4):477502, Dec 1993.

[172] Adam Eyre-Walker and Peter D Keightley. The distribution of fitness effects of new mutations. 8:610618, Aug 2007.

[173] Motoo Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, 1985.

[174] Vasiliy A Portnoy, David A Scott, Nathan E Lewis, Yekaterina Tarasova, Andrei L Osterman, and Bernhard Palsson. Deletion of genes encoding cytochrome oxidases and quinol monooxygenase blocks the aerobic-anaerobic shift in escherichia

coli k-12 mg1655. *Applied and environmental microbiology*, 76(19):65296540, Oct 2010.

[175] Stephen S Fong, Jennifer Y Marciniak, and Bernhard Palsson. Description and interpretation of adaptive evolution of escherichia coli k-12 mg1655 by using a genome-scale in silico metabolic model. *Journal of bacteriology*, 185(21):64006408, Nov 2003.

[176] Stephen S Fong, Annik Nanchen, Bernhard O Palsson, and Uwe Sauer. Latent pathway activation and increased pathway capacity enable escherichia coli adaptation to loss of key metabolic enzymes. *The Journal of biological chemistry*, 281(12):80248033, Mar 2006.

[177] S. S. Fong and B. O. Palsson. Metabolic gene-deletion strains of escherichia coli evolve to computationally predicted growth phenotypes. *Nat Genet*, 36(10):1056–8, 2004.

[178] Ryan A LaCroix, Troy E Sandberg, Edward J OBrien, Jose Utrilla, Ali Ebrahim, Gabriela I Guzman, Richard Szubin, Bernhard O Palsson, and Adam M Feist. Discovery of key mutations enabling rapid growth of escherichia coli k-12 mg1655 on glucose minimal media using adaptive laboratory evolution. *Applied and environmental microbiology*, Oct 2014.

[179] Troy E Sandberg, Margit Pedersen, Ryan A LaCroix, Ali Ebrahim, Mads Bonde, Markus J Herrgard, Bernhard O Palsson, Morten Sommer, and Adam M Feist. Evolution of escherichia coli to 42 c and subsequent genetic engineering reveals adaptive mechanisms and novel mutations. *Molecular biology and evolution*, 31(10):26472662, Oct 2014.

[180] M G Vander, L C Cantley, and C B Thompson. Understanding the warburg effect: the metabolic requirements of cell proliferation. *Science*, 2009.

[181] A Varma and B O Palsson. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type escherichia coli w3110. *Applied and environmental microbiology*, 60(10):37243731, Oct 1994.

[182] Alan J Wolfe. The acetate switch. *Microbiology and molecular biology reviews: MMBR*, 69(1):1250, Mar 2005.

[183] Karl Peebo, Kaspar Valgepea, Ranno Nahku, Gethe Riis, Mikk Oun, Kaarel Adamberg, and Raivo Vilu. Coordinated activation of pta-acs and tca cycles strongly reduces overflow metabolism of acetate in escherichia coli. *Applied microbiology and biotechnology*, 98(11):51315143, Jun 2014.

[184] Jason Raymond and Daniel Segr. The effect of oxygen on biochemical networks and the evolution of complex life. *Science*, 311(5768):17641767, Mar 2006.

[185] A H Romano and T Conway. Evolution of carbohydrate metabolic pathways. *Research in microbiology*, 147(6-7):448455, Jul 1996.

[186] Minglei Wang, Ying-Ying Jiang, Kyung Mo Kim, Ge Qu, Hong-Fang Ji, Jay E Mittenthal, Hong-Yu Zhang, and Gustavo Caetano-Anolls. A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation. 28:567582, Jan 2011.

[187] M Kirschner and J Gerhart. Evolvability. *Proceedings of the National Academy of Sciences of the United States of America*, 95(15):84208427, Jul 1998.

[188] Gunter P Wagner and Lee Altenberg. Perspective: Complex adaptations and the evolution of evolvability. *Evolution; international journal of organic evolution*, 50(3):967976, Jun 1996.

[189] Jamey D Young. Inca: a computational platform for isotopically non-stationary metabolic flux analysis. 30:13331335, May 2014.

[190] F. C. Jones, M. G. Grabherr, Y. F. Chan, P. , E. Mauceli, J. Johnson, R. Swofford, M. Pirun, M. C. Zody, S. White, E. Birney, S. Searle, J. Schmutz, J. Grimwood, M. C. Dickson, R. M. Myers, C. T. Miller, B. R. Summers, A. K. Knecht, S. D. Brady, H. Zhang, A. A. Pollen, T. Howes, C. Amemiya, J. Baldwin, T. Bloom, D. B. Jaffe, R. Nicol, J. Wilkinson, E. S. Lander, F. Di Palma, K. Lindblad-Toh, and D. M. Kingsley. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, 484(7392):55–61, 2012.

[191] H. B. Fraser. Gene expression drives local adaptation in humans. *Genome Research*, 23(7):1089–96, 2013.

[192] G. A. Wray. The evolutionary significance of cis-regulatory mutations. *Nature reviews. Genetics*, 8(3):206–16, 2007.

[193] B. Prud'homme, N. Gompel, and S. B. Carroll. Emerging principles of regulatory evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 104 Suppl 1:8605–12, 2007.

[194] D. Enard, P. W. Messer, and D. A. Petrov. Genome-wide signals of positive selection in human evolution. *Genome Research*, 24(6):885–95, 2014.

[195] M. C. King and A. C. Wilson. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–16, 1975.

[196] T. Ferenci. The spread of a beneficial mutation in experimental bacterial populations: the influence of the environment and genotype on the fixation of rpos mutations. *Heredity*, 100(5):446–52, 2008.

[197] Jeffrey E. Barrick, Mark R. Kauth, Christopher C. Strelioff, and Richard E. Lenski. Escherichia coli rpob mutants have increased evolvability in proportion to their fitness defects. *Molecular biology and evolution*, 27(6):1338–47, 2010.

[198] Gerda Saxer, Michael D. Krepps, Eric D. Merkley, Charles Ansong, Brooke L. Deatherage Kaiser, Marie-Thrse Valovska, Nikola Ristic, Ping T. Yeh, Vittal P. Prakash, Owen P. Leiser, Luay Nakhleh, Henry S. Gibbons, Helen W. Kreuzer, and Yousif Shamoo. Mutations in global regulators lead to metabolic selection during adaptation to complex environments. *PLoS Genetics*, 10(12):e1004872–e1004872, 2014.

[199] T. M. Conrad, M. Frazier, A. R. Joyce, B. K. Cho, E. M. Knight, N. E. Lewis, R. Landick, and B. O. Palsson. Rna polymerase mutants found through adaptive evolution reprogram escherichia coli for optimal growth in minimal media. *Proceedings of the National Academy of Sciences of the United States of America*, 107(47):20500–20505, 2010.

[200] Kian-Kai Cheng, Baek-Seok Lee, Takeshi Masuda, Takuro Ito, Kazutaka Ikeda, Akiyoshi Hirayama, Lingli Deng, Jiyang Dong, Kazuyuki Shimizu, Tomoyoshi Soga, Masaru Tomita, Bernhard O. Palsson, and Martin Robert. Global metabolic network reorganization by adaptive mutations allows fast growth of escherichia coli on glycerol. *Nature Communications*, 5:1–9, 2014.

[201] V. Hauryliuk, G. C. Atkinson, K. S. Murakami, T. Tenson, and K. Gerdes. Recent functional insights into the role of (p)ppgpp in bacterial physiology. *Nat Rev Microbiol*, 13(5):298–309, 2015.

[202] Thea King, Akira Ishihama, Ayako Kori, and Thomas Ferenci. A regulatory trade-off as a source of strain variation in the species escherichia coli a regulatory trade-off as a source of strain variation in the species escherichia coli . 186(17):5614–5620, 2004.

[203] H. H. Wang, F. J. Isaacs, P. A. Carr, Z. Z. Sun, G. Xu, C. R. Forest, and G. M. Church. Programming cells by multiplex genome engineering and accelerated evolution. *Nature*, 460(7257):894–8, 2009.

[204] C. D. Herring, A. Raghunathan, C. Honisch, T. Patel, M. K. Applebee, A. R. Joyce, T. J. Albert, F. R. Blattner, D. van den Boom, C. R. Cantor, and B. O. Palsson. Comparative genome sequencing of escherichia coli allows observation of bacterial evolution on a laboratory timescale. *Nat Genet*, 38(12):1406–12, 2006.

[205] Stefan Klumpp and Terence Hwa. Bacterial growth: global effects on gene expression, growth feedback and proteome partition. *Current opinion in biotechnology*, 28C:96–102, 2014.

[206] A. Sethi, J. Eargle, A. A. Black, and Z. Luthey-Schulten. Dynamical networks in trna:protein complexes. *Proc Natl Acad Sci U S A*, 106(16):6620–5, 2009.

[207] O. Tenaillon, A. Rodriguez-Verdugo, R. L. Gaut, P. McDonald, A. F. Bennett, A. D. Long, and B. S. Gaut. The molecular diversity of adaptive convergence. *Science*, 335(6067):457–61, 2012.

[208] G. Bar-Nahum, V. Epshtein, A. E. Ruckenstein, R. Rafikov, A. Mustaev, and E. Nudler. A ratchet mechanism of transcription elongation and its control. *Cell*, 120(2):183–93, 2005.

[209] R. O. Weinzierl. The nucleotide addition cycle of rna polymerase is controlled by two molecular hinges in the bridge helix domain. *BMC Biol*, 8:134, 2010.

[210] R. O. Weinzierl. The bridge helix of rna polymerase acts as a central nanomechanical switchboard for coordinating catalysis and substrate movement. *Archaea*, 2011:608385, 2011.

[211] M. Jishage, K. Kvint, V. Shingler, and T. Nystrom. Regulation of sigma factor competition by the alarmone ppgpp. *Genes Dev*, 16(10):1260–70, 2002.

[212] Y. N. Zhou and D. J. Jin. The rpob mutants destabilizing initiation complexes at stringently controlled promoters behave like "stringent" rna polymerases in escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 95(6):2908–13, 1998.

[213] M. M. Barker, T. Gaal, C. a Josaitis, and R. L. Gourse. Mechanism of regulation of transcription initiation by ppgpp. i. effects of ppgpp on transcription initiation in vivo and in vitro. *Journal of molecular biology*, 305(4):673–88, 2001.

[214] S. Osterberg, T. del Peso-Santos, and V. Shingler. Regulation of alternative sigma factor use. *Annual review of microbiology*, 65:37–55, 2011.

[215] S. J. Pirt. Maintenance energy: a general model for energy-limited and energy-sufficient growth. *Archives of Microbiology*, pages 300–302, 1982.

[216] D. J. Futuyma and G. Moreno. The evolution of ecological specialization. *Annual Review of Ecology and Systematics*, 19:207–233, 1988.

[217] S. Remold. Understanding specialism when the jack of all trades can be the master of all. *Proceedings. Biological sciences / The Royal Society*, 279(1749):4861–9, 2012.

[218] V. S. Cooper and R. E. Lenski. The population genetics of ecological specialization in evolving escherichia coli populations. *Nature*, 407(6805):736–9, 2000.

[219] N. Leiby and C. J. Marx. Metabolic erosion primarily through mutation accumulation, and not tradeoffs, drives limited evolution of substrate specificity in escherichia coli. *PLoS biology*, 12(2):e1001789, 2014.

[220] P. Innocenti and S. F. Chenoweth. Interspecific divergence of transcription networks along lines of genetic variance in drosophila: dimensionality, evolvability, and constraint. *Molecular biology and evolution*, 30(6):1358–67, 2013.

[221] G. P. Wagner, M. Pavlicev, and J. M. Cheverud. The road to modularity. *Nature reviews. Genetics*, 8(12):921–31, 2007.

[222] S. R. Grossman, K. G. Andersen, I. Shlyakhter, S. Tabrizi, S. Winnicki, A. Yen, D. J. Park, D. Griesemer, E. K. Karlsson, S. H. Wong, M. Cabili, R. A. Adegbola, R. N. Bamezai, A. V. Hill, F. O. Vannberg, J. L. Rinn, E. S. Lander, S. F. Schaffner, and P. C. Sabeti. Identifying recent adaptations in large-scale genomic data. *Cell*, 152(4):703–13, 2013.

[223] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis, and J. N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews. Genetics*, 9(5):356–69, 2008.

[224] W. Cookson, L. Liang, G. Abecasis, M. Moffatt, and M. Lathrop. Mapping complex disease traits with global gene expression. *Nature reviews. Genetics*, 10(3):184–94, 2009.

[225] K. A. Datsenko and B. L. Wanner. One-step inactivation of chromosomal genes in escherichia coli k-12 using pcr products. *Proc Natl Acad Sci U S A*, 97(12):6640–5, 2000.

[226] D. L. Tucker, N. Tucker, Z. Ma, J. W. Foster, R. L. Miranda, P. S. Cohen, and T. Conway. Genes of the gadx-gadw regulon in escherichia coli. *Journal of bacteriology*, 185(10):3190–201, 2003.

[227] Shaleen B. Korch, Thomas A. Henderson, and Thomas M. Hill. Characterization of the hipa7 allele of escherichia coli and evidence that high persistence is governed by (p)ppgpp synthesis. *Molecular Microbiology*, 50(4):1199–1213, 2003.

[228] Douglas McCloskey, Jose Utrilla, Robert K. Naviaux, Bernhard O. Palsson, and Adam M. Feist. Fast swinnex filtration (fsf): a fast and robust sampling and extraction method suitable for metabolomics analysis of cultures grown in complex media. *Metabolomics*, 2014.

[229] Byung-Kwan Cho, Donghyuk Kim, Eric M. Knight, Karsten Zengler, and Bernhard O. Palsson. Genome-scale reconstruction of the sigma factor network in escherichia coli: topology and functional states. *BMC biology*, 12(1):4–4, 2014.

[230] H. Salgado, M. Peralta-Gil, S. Gama-Castro, A. Santos-Zavaleta, L. Muniz-Rascado, J. S. Garcia-Sotelo, V. Weiss, H. Solano-Lira, I. Martinez-Flores, A. Medina-Rivera, G. Salgado-Osorio, S. Alquicira-Hernandez, K. Alquicira-Hernandez, A. Lopez-Fuentes, L. Porron-Sotelo, A. M. Huerta, C. Bonavides-Martinez, Y. I. Balderas-Martinez, L. Pannier, M. Olvera, A. Labastida, V. Jimenez-Jacinto, L. Vega-Alvarado, V. Del Moral-Chavez, A. Hernandez-Alvarez, E. Morett, and J. Collado-Vides. Regulondb v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Research*, 41(Database issue):D203–13, 2013.

[231] Natacha Opalka, Jesse Brown, William J. Lane, Kelly-Anne F. Twist, Robert Landick, Francisco J. Asturias, and Seth a Darst. Complete structural model of escherichia coli rna polymerase from a hybrid approach. *PLoS biology*, 8(9), 2010.

[232] D. G. Vassylyev, M. N. Vassylyeva, J. Zhang, M. Palangat, I. Artsimovitch, and R. Landick. Structural basis for substrate loading in bacterial rna polymerase. *Nature*, 448(7150):163–8, 2007.

[233] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten. Scalable molecular dynamics with namd. *J Comput Chem*, 26(16):1781–802, 2005.

[234] R. B. Best, X. Zhu, J. Shim, P. E. Lopes, J. Mittal, M. Feig, and Jr. Mackerell, A. D. Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi(1) and chi(2) dihedral angles. *J Chem Theory Comput*, 8(9):3257–3273, 2012.

[235] W. Humphrey, A. Dalke, and K. Schulten. Vmd: visual molecular dynamics. *J Mol Graph*, 14(1):33–8, 27–8, 1996.

[236] S. Chaudhury, S. Lyskov, and J. J. Gray. Pyrosetta: a script-based interface for implementing molecular modeling algorithms using rosetta. *Bioinformatics*, 26(5):689–91, 2010.

[237] T. Kortemme and D. Baker. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A*, 99(22):14116–21, 2002.

[238] J. Gavenonis, B. A. Sheneman, T. R. Siegert, M. R. Eshelman, and J. A. Kritzer. Comprehensive analysis of loops at protein-protein interfaces for macrocycle design. *Nat Chem Biol*, 10(9):716–22, 2014.

[239] M. M. Domach, S. K. Leung, R. E. Cahn, G. G. Cocks, and M. L. Shuler. Computer model for glucose-limited growth of a single cell of escherichia coli b/r-a. *Biotechnology and bioengineering*, 26(3):203–16, 1984.

[240] A. Joshi and B. O. Palsson. Metabolic dynamics in the human red cell. part i–a comprehensive kinetic model. *Journal of theoretical biology*, 141(4):515–28, 1989.

[241] A. Joshi and B. O. Palsson. Metabolic dynamics in the human red cell. part ii–interactions with the environment. *Journal of theoretical biology*, 141(4):529–45, 1989.

[242] A. Joshi and B. O. Palsson. Metabolic dynamics in the human red cell. part iii–metabolic reaction rates. *Journal of theoretical biology*, 142(1):41–68, 1990.

[243] A. Joshi and B. O. Palsson. Metabolic dynamics in the human red cell. part iv–data prediction and some model computations. *Journal of theoretical biology*, 142(1):69–85, 1990.

[244] R. Heinrich, S. M. Rapoport, and T. A. Rapoport. Metabolic regulation and mathematical models. *Progress in biophysics and molecular biology*, 32(1):1–82, 1977.

[245] H. H. McAdams and L. Shapiro. Circuit simulation of genetic networks. *Science*, 269(5224):650–6, 1995.

[246] H. H. McAdams and A. Arkin. Simulation of prokaryotic genetic circuits. *Annual review of biophysics and biomolecular structure*, 27:199–224, 1998.

[247] A. Joshi and B. O. Palsson. Escherichia coil growth dynamics: A three-pool biochemically based description. *Biotechnology and bioengineering*, 31(2):102–16, 1988.

[248] B. O. Palsson and A. Joshi. On the dynamic order of structured escherichia coli growth models. *Biotechnology and bioengineering*, 29(6):789–92, 1987.

[249] A. Feizi, T. Osterlund, D. Petranovic, S. Bordel, and J. Nielsen. Genome-scale modeling of the protein secretory machinery in yeast. *PLoS One*, 8(5):e63284, 2013.

[250] D. Machado and M. Herrgard. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Comput Biol*, 10(4):e1003580, 2014.

[251] A. N. Brooks, D. J. Reiss, A. Allard, W. J. Wu, D. M. Salvanha, C. L. Plaisier, S. Chandrasekaran, M. Pan, A. Kaur, and N. S. Baliga. A system-level model for the microbial regulatory genome. *Mol Syst Biol*, 10(7):740, 2014.

[252] J. Carrera, R. Estrela, J. Luo, N. Rai, A. Tsoukalas, and I. Tagkopoulos. An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of escherichia coli. *Mol Syst Biol*, 10(7):735, 2014.

[253] S. Federowicz, D. Kim, A. Ebrahim, J. Lerman, H. Nagarajan, B. K. Cho, K. Zengler, and B. Palsson. Determining the control circuitry of redox metabolism at the genome-scale. *PLoS genetics*, 10(4):e1004264, 2014.

[254] H. S. Rhee and B. F. Pugh. Comprehensive genome-wide protein-dna interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–19, 2011.

[255] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips. Transcriptional regulation by the numbers: models. *Current opinion in genetics & development*, 15(2):116–24, 2005.

[256] A. Flamholz, E. Noor, A. Bar-Even, W. Liebermeister, and R. Milo. Glycolytic strategy as a tradeoff between energy yield and protein cost. *Proc Natl Acad Sci U S A*, 110(24):10039–44, 2013.

[257] E. Oh, A. H. Becker, A. Sandikci, D. Huber, R. Chaba, F. Gloge, R. J. Nichols, A. Typas, C. A. Gross, G. Kramer, J. S. Weissman, and B. Bukau. Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell*, 147(6):1295–308, 2011.

[258] A. Cvetkovic, A. L. Menon, M. P. Thorgersen, J. W. Scott, 2nd Poole, F. L., Jr. Jenney, F. E., W. A. Lancaster, J. L. Praissman, S. Shanmukh, B. J. Vaccaro, S. A. Trauger, E. Kalisiak, J. V. Apon, G. Siuzdak, S. M. Yannone, J. A. Tainer, and M. W. Adams. Microbial metalloproteomes are largely uncharacterized. *Nature*, 466(7307):779–82, 2010.

[259] T. Maier, M. Guell, and L. Serrano. Correlation of mrna and protein in complex biological samples. *FEBS letters*, 583(24):3966–73, 2009.

[260] C. Vogel and E. M. Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet*, 13(4):227–32, 2012.

[261] B. Schwanhausser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–42, 2011.

[262] A. Zaslaver, S. Kaplan, A. Bren, A. Jinich, A. Mayo, E. Dekel, U. Alon, and S. Itzkovitz. Invariant distribution of promoter activities in escherichia coli. *PLoS Comput Biol*, 5(10):e1000545, 2009.

[263] L. Keren, O. Zackay, M. Lotan-Pompan, U. Barenholz, E. Dekel, V. Sasson, G. Aidelberg, A. Bren, D. Zeevi, A. Weinberger, U. Alon, R. Milo, and E. Segal. Promoters maintain their relative activity levels under different growth conditions. *Mol Syst Biol*, 9:701, 2013.

[264] K. Valgepea, K. Adamberg, A. Seiman, and R. Vilu. Escherichia coli achieves faster growth by increasing catalytic and translation rates of proteins. *Molecular bioSystems*, 9(9):2344–58, 2013.

[265] P. Shah, Y. Ding, M. Niemczyk, G. Kudla, and J. B. Plotkin. Rate-limiting steps in yeast protein translation. *Cell*, 153(7):1589–601, 2013.

[266] N. T. Ingolia. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat Rev Genet*, 15(3):205–13, 2014.

[267] S. Kosuri, D. B. Goodman, G. Cambray, V. K. Mutalik, Y. Gao, A. P. Arkin, D. Endy, and G. M. Church. Composability of regulatory sequences controlling transcription and translation in escherichia coli. *Proc Natl Acad Sci U S A*, 110(34):14024–9, 2013.

[268] V. K. Mutalik, J. C. Guimaraes, G. Cambray, Q. A. Mai, M. J. Christoffersen, L. Martin, A. Yu, C. Lam, C. Rodriguez, G. Bennett, J. D. Keasling, D. Endy, and A. P. Arkin. Quantitative estimation of activity and quality for collections of functional genetic elements. *Nature methods*, 10(4):347–53, 2013.

[269] G. Cambray, J. C. Guimaraes, V. K. Mutalik, C. Lam, Q. A. Mai, T. Thimmaiah, J. M. Carothers, A. P. Arkin, and D. Endy. Measurement and modeling of intrinsic transcription terminators. *Nucleic Acids Research*, 41(9):5139–48, 2013.

[270] Y. J. Chen, P. Liu, A. A. Nielsen, J. A. Brophy, K. Clancy, T. Peterson, and C. A. Voigt. Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nature methods*, 10(7):659–64, 2013.

[271] K. van Eunen, J. A. Kiewiet, H. V. Westerhoff, and B. M. Bakker. Testing biochemistry revisited: how in vivo metabolism can be understood from in vitro enzyme kinetics. *PLoS Comput Biol*, 8(4):e1002483, 2012.

[272] R. Garcia-Contreras, P. Vos, H. V. Westerhoff, and F. C. Boogerd. Why in vivo may not equal in vitro - new effectors revealed by measurement of enzymatic activities under the same in vivo-like assay conditions. *FEBS J*, 279(22):4145–59, 2012.

[273] J. C. Sanghvi, S. Regot, S. Carrasco, J. R. Karr, M. V. Gutschow, Jr. Bolival, B., and M. W. Covert. Accelerated discovery via a whole-cell model. *Nature methods*, 10(12):1192–5, 2013.

[274] C. Cotten and J. L. Reed. Mechanistic analysis of multi-omics datasets to generate kinetic parameters for constraint-based metabolic models. *BMC Bioinformatics*, 14:32, 2013.

[275] A. Khodayari, A. R. Zomorrodi, J. C. Liao, and C. D. Maranas. A kinetic model of escherichia coli core metabolism satisfying multiple sets of mutant flux data. *Metab Eng*, 2014.

[276] M. N. Price, A. M. Deutschbauer, J. M. Skerker, K. M. Wetmore, T. Ruths, J. S. Mar, J. V. Kuehl, W. Shao, and A. P. Arkin. Indirect and suboptimal control of gene expression is widespread in bacteria. *Mol Syst Biol*, 9:660, 2013.

[277] A. M. New, B. Cerulus, S. K. Govers, G. Perez-Samper, B. Zhu, S. Boogmans, J. B. Xavier, and K. J. Verstrepen. Different levels of catabolite repression optimize growth in stable and variable environments. *PLoS biology*, 12(1):e1001764, 2014.

[278] G. M. de Hijas-Liste, E. Klipp, E. Balsa-Canto, and J. R. Banga. Global dynamic optimization approach to predict activation in metabolic pathways. *BMC systems biology*, 8:1, 2014.

[279] M. Y. Pavlov and M. Ehrenberg. Optimal control of gene expression for fast proteome adaptation to environmental change. *Proc Natl Acad Sci U S A*, 110(51):20527–32, 2013.

[280] M. Bartl, M. Kotzing, S. Schuster, P. Li, and C. Kaleta. Dynamic optimization identifies optimal programmes for pathway regulation in prokaryotes. *Nature communications*, 4:2243, 2013.

[281] K. Kochanowski, U. Sauer, and V. Chubukov. Somewhat in control–the role of transcription in regulating microbial metabolic fluxes. *Curr Opin Biotechnol*, 24(6):987–93, 2013.

[282] D. Rothschild, E. Dekel, J. Hausser, A. Bren, G. Aidelberg, P. Szekely, and U. Alon. Linear superposition and prediction of bacterial promoter activity dynamics in complex conditions. *PLoS Comput Biol*, 10(5):e1003602, 2014.

[283] H. Link, K. Kochanowski, and U. Sauer. Systematic identification of allosteric protein-metabolite interactions that control enzyme activity in vivo. *Nat Biotechnol*, 31(4):357–61, 2013.

[284] S. R. Modi, D. M. Camacho, M. A. Kohanski, G. C. Walker, and J. J. Collins. Functional characterization of bacterial srnas using a network biology approach. *Proc Natl Acad Sci U S A*, 108(37):15522–7, 2011.

[285] X. Li, T. A. Gianoulis, K. Y. Yip, M. Gerstein, and M. Snyder. Extensive in vivo metabolite-protein interactions revealed by large-scale systematic analyses. *Cell*, 143(4):639–50, 2010.

[286] S. Goyal, J. Yuan, T. Chen, J. D. Rabinowitz, and N. S. Wingreen. Achieving optimal growth through product feedback inhibition in metabolism. *Plos Computational Biology*, 6(6), 2010.

[287] V. Chubukov, I. A. Zuleta, and H. Li. Regulatory architecture determines optimal regulation of gene expression in metabolic pathways. *Proc Natl Acad Sci U S A*, 109(13):5127–32, 2012.

[288] Y. Tan, J. G. Rivera, C. A. Contador, J. A. Asenjo, and J. C. Liao. Reducing the allowable kinetic space by constructing ensemble of dynamic models with the same steady-state flux. *Metab Eng*, 13(1):60–75, 2011.

[289] N. Tepper, E. Noor, D. Amador-Noguez, H. S. Haraldsdottir, R. Milo, J. Rabinowitz, W. Liebermeister, and T. Shlomi. Steady-state metabolite concentrations reflect a balance between maximizing enzyme efficiency and minimizing total metabolite load. *PLoS One*, 8(9):e75370, 2013.

[290] W. Liebermeister, E. Noor, A. Flamholz, D. Davidi, J. Bernhardt, and R. Milo. Visual account of protein investment in cellular functions. *Proc Natl Acad Sci U S A*, 111(23):8488–93, 2014.

[291] S. V. Rajagopala, P. Sikorski, A. Kumar, R. Mosca, J. Vlasblom, R. Arnold, J. Franca-Koh, S. B. Pakala, S. Phanse, A. Ceol, R. Hauser, G. Siszler, S. Wuchty, A. Emili, M. Babu, P. Aloy, R. Pieper, and P. Uetz. The binary protein-protein interaction landscape of escherichia coli. *Nat Biotechnol*, 32(3):285–90, 2014.

[292] Q. C. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter, T. Maniatis, A. Califano, and B. Honig. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, 490(7421):556–60, 2012.

[293] X. Huang, H. M. Holden, and F. M. Raushel. Channeling of substrates and intermediates in enzyme-catalyzed reactions. *Annual review of biochemistry*, 70:149–80, 2001.

[294] C. M. Agapakis, P. M. Boyle, and P. A. Silver. Natural strategies for the spatial optimization of metabolism in synthetic biology. *Nature chemical biology*, 8(6):527–35, 2012.

[295] R. J. Ellis. Macromolecular crowding: obvious but underappreciated. *Trends in biochemical sciences*, 26(10):597–604, 2001.

[296] K. A. Dill, K. Ghosh, and J. D. Schmit. Physical limits of cells and proteomes. *Proc Natl Acad Sci U S A*, 108(44):17876–82, 2011.