UNIVERSITY OF CALIFORNIA, SAN DIEGO

Expanding the Scope of Genome-scale Models of Metabolism and Gene Expression

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

in

Bioengineering

by

Colton Joseph Lloyd

Committee in charge:

Bernhard O. Palsson, Chair
Adam M. Feist
Terence Hwa
Andrew D. McCulloch
Christian M. Metallo
Larry Smarr

2019

The dissertation of Colton Joseph Lloyd is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

_____

Chair

University of California San Diego

2019

DEDICATION

To:

To my family and friends for their unrelenting support and patience.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I express my deepest gratitude toward Bernhard Palsson for his guidance and support over the past five and half years. I'm constantly surprised by the personal interest he takes in the professional development of his students, despite a litany of other obligations vying for his time and headspace.

To have a mentor who not only provides staunch feedback on the science underlying your work but also reminds you why your research makes a difference is invaluableand a testament to how the SBRG provides an unmatched environment for graduate research.

My greatest thanks, as well, to Adam Feist. He was the first person I contacted about joining the SBRG, and he mentored me through what developed into Chapter 5 of this dissertation. On a personal level, it also was enjoyable to have someone else in the research group with a Midwest background, a Big10 alma mater, and a shared birthday.

I was lucky to have a number of fantastic mentors in the SBRG, starting with Zak King. He helped me find my bearings after joining the lab and exuded only patience in entertaining my misguided questions. Since then, hes consistently been a great resource. I also thank Laurence Yang, who has steadily been willing to discuss and talk through challenging, intimidating questions. Besides being one of the smartest people I know, Laurence is incredibly generous with his time. My gratitude, too, to Jonathan Monk, who worked with me on Chapter 3 of this dissertation and taught me how to approach impactful problems in health and biology. For his collaboration and guidance on Chapter 2 of this dissertation and, more importantly, training me how to write robust academic software, I express my appreciation to Ali Ebrahim.

I also must acknowledge David Heckman, JC LaChance, Anand Sastry, Patrick Phaneuf, and Yara Seif and others lab mates for the coffee breaks and discussions that kept me sane. To

everyone else who has crossed paths with me during my time at in the SBRG, thank you. There are few academic settings where you will find more brilliant peopleall of whom are so kind with their time and genuinely want you to see you succeed.

Next, thank you to my family, who instilled in me an excitement to learn early in life. Their support has been unrelenting, and I would not have made it to this point without them.

Finally, Sarah Pfledderer has been a constant source of positivity and joy in my life. There are plenty of highs and lows in graduate school but having her as a loving companion makes it feel easy. I adore you, and I'm forever grateful for the support from you and your family.

I would also like to thank my funding sources that have supported this work. These include the National Science Foundation, the Novo Nordisk Foundation, the National Institutes of Health the US Department of Energy.

Chapter 1 in part is a reprint of material published in:

- **CJ Lloyd**, N Mih, L Yang and BO Palsson. 2019. "Fundamentals of Metabolic Systems Biology." *Encyclopedia of Microbiology*, The dissertation author was the primary author.

Chapter 2 in part is a reprint of material published in:

- **CJ Lloyd***, A Ebrahim*, L Yang, ZA King, E Catoiu, EJ OBrien, JK Liu, and BO Palsson. 2018. "COBRAme: A computational framework for genome-scale models of metabolism and gene expression." *PLoS Computational Biology* 14(7): e1006302. The dissertation author was one of the primary authors.

Chapter 3 in part is a reprint of material published in:

- JM Monk*, **CJ Lloyd***, E Brunk, N Mih, A Sastry, Z King, R Takeuchi, W Nomura, Z

Zhang, H Mori, AM Feist, BO Palsson. 2017. "*i*ML1515 , a knowledgebase that computes *Escherichia coli* traits." *Nature Biotechnology* 35 (10): 9048. The dissertation author was one of the two primary authors.

Chapter 4 in part is a reprint of material published in:

- **CJ Lloyd**, JM Monk, L, Yang, A Ebrahim, and BO Palsson. "Genome-scale models reveal the intricate relationships between proteome cofactor requirements and growth environments." 2019. *In Preparation.* The dissertation author is the primary author.

Chapter 5 in part is a reprint of material published in:

- **CJ Lloyd**, ZA King, TE Sandberg, Y Hefner, CA Olson, EJ OBrien, JG Sanders, RA Salido, K Sanders, C Brennan, G Humphrey, R Knight, and AM Feist. 2019 "The genetic basis for adaptation of model-designed syntrophic co-cultures." *PLOS Computational Biology*, 15(3): e1006213. The dissertation author was the primary author.

Chapter 6 in part is a reprint of material published in:

- **CJ Lloyd**, N Mih, L Yang and BO Palsson. 2019. "Fundamentals of Metabolic Systems Biology." *Encyclopedia of Microbiology.* The dissertation author was the primary author.

VITA

2013	Bachelor of Science in Biomedical Engineering, The Ohio State University
2019	Doctor of Philosophy in Bioengineering, University of California, San Diego

PUBLICATIONS

ZA King, **CJ Lloyd**, AM Feist, BO Palsson. 2015. "Next-generation genome-scale models for metabolic engineering." *Current Opinion in Biotechnology*, 35: 2329.

L Yang*, J Tan*, EJ O'Brien, JM Monk, D Kim, HJ Li, P Charusantia, A Ebrahim, **CJ Lloyd**, JT Yurkovich, B Du, A Dräger A Thomas, Y Sun, MA Saunders, and BO Palsson. 2015. "Systems biology definition of the core proteome of metabolism and expression is consistent with high-throughput data." *Proceedings of the National Academy of Science of the United States of America*, 112(34):10810-10815.

L Yang, JT Yurkovich, **CJ Lloyd**, A Ebrahim, MA Saunders, and BO Palsson. 2016. "Principles of proteome allocation are revealed using proteomic data and genome-scale models." *Scientific reports*, 6, 36734

L Yang, D Ma, A Ebrahim, **CJ Lloyd**, MA Saunders, and BO Palsson. 2016. "solveME: fast and reliable solution of nonlinear ME models" *BMC Bioinformatics*, 17 (1): 391

L Yang*, JT Yurkovich*, **CJ Lloyd**, A Ebrahim, MA Saunders, and BO Palsson. 2016. "Principles of proteome allocation are revealed using proteomic data and genome-scale models." *Scientific Reports*, 6:36734.

X Fang*, A Sastry*, N Mih, D Kim, J Tan, JT Yurkovich, **CJ Lloyd**, Y Gao, L Yang, and BO Palsson. 2017. "Global transcriptional regulatory network for Escherichia coli robustly connects gene expression to transcription factor activities." *Proceedings of the National Academy of Sciences of the United States of America*, 114(38):10286-10291.

JM Monk*, **CJ Lloyd***, E Brunk*, N Mih, A Sastry, Z King, R Takeuchi, W Nomura, Z Zhang, H Mori, AM Feist, BO Palsson. 2017."*i*ML1515, a knowledgebase that computes *Escherichia coli* traits." *Nature Biotechnology* 35 (10): 9048.

TE Sandberg, **CJ Lloyd**, BO Palsson, and AM Feist. 2017. "Laboratory evolution to alternating substrate environments yields distinct phenotypic and genetic adaptive strategies" *Applied Environmental Microbiology*, 83 (13): e00410-17

CJ Lloyd*, A Ebrahim*, L Yang, ZA King, E Catoiu, EJ OBrien, JK Liu, and BO Palsson. 2018. "COBRAme: A computational framework for genome-scale models of metabolism and gene expression." *PLoS Computational Biology*, 14(7): e1006302.

D Heckmann, **CJ Lloyd**, N Mih, Y Ha, DC Zielinski, ZB Haiman, AA Desouki, MJ Lercher, and BO Palsson. 2018. "Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models." *Nature communications*, 9 (1): 5252

CJ Lloyd, ZA King, TE Sandberg, Y Hefner, CA Olson, EJ OBrien, JG Sanders, RA Salido, K Sanders, C Brennan, G Humphrey, R Knight, and AM Feist. 2019 "The genetic basis for adaptation of model-designed syntrophic co-cultures." *PLOS Computational Biology*, 15(3): e1006213. https://doi.org/10.1371/journal.pcbi.1006213.

CJ Lloyd, N Mih, L Yang and BO Palsson. 2019. "Fundamentals of Metabolic Systems Biology." *Encyclopedia of Microbiology*

* equal contribution

ABSTRACT OF THE DISSERTATION

Expanding the Scope of Genome-scale Models of Metabolism and Gene Expression

by

Colton Joseph Lloyd

Doctor of Philosophy in Bioengineering

University of California, San Diego, 2019

Bernhard O. Palsson, Chair

The cost of whole genome sequencing has declined precipitously over the past two decades. This reduced cost of data collection has contributed to a deluge of multi-omics data types for *Escherichia coli* and other commonly studied microbes. As a result, methods to obtain actionable information from whole genome sequencing and other omics data have become increasingly valuable. Here, we broaden the scope of genome-scale models, thus allowing them to add context to multi-omics data and add insight into *E. coli* metabolism. First, we outline a new computational framework, COBRAme, that empowers the use and facilitates the reconstruction of models of

metabolism and gene expression (ME-models). These models offer a comprehensive method to study protein use in microbes and how their metabolism is impacted by the evolutionary pressures to allocate proteome most efficiently. Previously, ME-models were prohibitively difficult to use, but the development of COBRAme has optimized the ME-model reconstruction process making them smaller, easier to understand, and quicker to solve. Second, the next-generation *E. coli* metabolic modelcomposing the metabolic core of the ME-modelis detailed. It was further demonstrated that such models can be applied to contextualize multi-omics data types and broaden our understanding of *E. coli* as a species. Third, adding to the species characterization of *E. coli*, we leveraged the *E. coli* ME-model to study how enzyme cofactor availability can shape condition-dependent metabolism. Given that some strains of *E. coli* are auxotrophic for cofactors, this information provides insight into the consequence and evolutionary drivers of auxotrophy. Lastly, a community *E. coli* ME-model was constructed to study the adaptation of syntrophy in co-cultures of *E. coli* auxotrophs. The model provided predictions of how the proteome efficiency of strains in co-culture could affect community characteristics. The totality of this work demonstrates that the scope and applications of these models can be expanded to obtain valuable information about the characteristics of *E. coli* as a single strain, a species, and in community.

Chapter 1

The promise of systems biology

Some 60 years ago, the promise of molecular biology held that if we knew and understood the function of the molecules that comprise cells, then we could understand cells and their functions. Although this was true in principle, the sheer number of molecules made it very difficult to comprehend so many simultaneous functions. Thus, to fulfill this promise, systems analysis is needed.

The simultaneous measurement of members of whole classes of biomolecules became possible over the past two decades (metabolomics, lipidomics, proteomics, etc) . In addition, methods (ChIP-Exo, Ribo-Seq, PPI, etc) have been developed to measure interactions between large numbers of biomolecules. Consequently, there are a growing number of datasets available that give us the molecular composition of cells under a certain condition. Many of the chemical interactions (i.e., mechanisms) between many of these components are now known and this knowledge gives rise to reconstructed biochemical reaction networks on a genome-scale that underlie various cellular functions. Structured representations of this information can be converted into a mathematical form enabling computation and model building. Thus the formulation of *in silico* cells

became possible that represent their *in vivo* counterpart based on our current state of knowledge. *In silico* cells became a foundation of the bottom-up approach to systems biology.

1.1 Reconstructing the *In Silico* Cell

Systems biology is not focused so much on the biomolecules themselves, but rather on the activity of the links (i.e., their chemical and physical interactions) that connect them, and the computation of functional states of reconstructed networks. Functional states of networks correspond to observable physiological or homeostatic states. Completing quantitative relationships between the chemical components of a cell (with their genetic bases) and their physiological functions is the promise of (molecular) systems biology. This undertaking represents the *de facto* construction of a high-dimensional mechanistic genotypephenotype relationship.

As an example of the systems biology approach, we illustrate computational prediction of essential genes, and synthetically lethal gene pairs, based on genome-scale metabolic models. Predictions of gene essentiality are made by computing the growth capabilities (i.e. the synthesis of all biomass components in the right ratios) of an in silico cell by removing the activities of a single gene. Synthetic lethals can be predicted by simultaneously removing two genes from an *in silico* cell and computing its growth capabilities. These predictions represent perhaps the largest scale and most intricate computational predictions of phenotypes performed to date, reaching hundreds of thousands of predicted experimental outcomes. The comparison of computed lethality and experimental measurement in a number of studies is shown in Figure 1.1 . The remainder of this Chapter illustrates the concept of a network reconstruction, the formation of an *in silico* cell (i.e., a computable knowledge base), and its uses to understand biological functions.

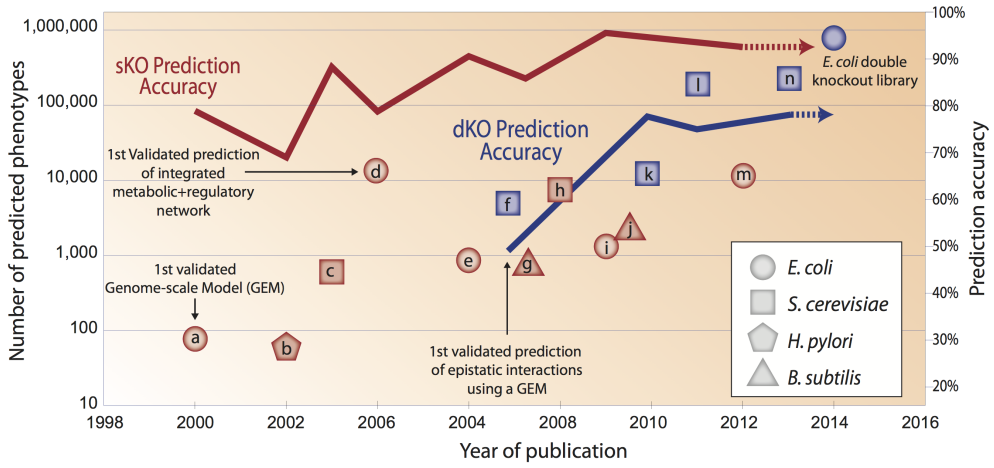

Figure 1.1: The number of predictions made on growth screens that cross environmental conditions with gene knockouts has grown steadily over the past 15 years. Over this time, both single-gene knockout (SKO) predictions (red line) and double-gene knockout (DKO) predictions (blue line) have become increasingly accurate. Figure is taken from [1]

Fundamentals of Constraints Based Reconstruction and Analysis (COBRA)

The underlying COnstraints Based Reconstruction and Analysis (COBRA) methods are based on relatively simple concepts, but their application can result in non-intuitive, novel predictions. COBRA entails two fundamental steps (Figure 1.2):

- The first step is the network reconstruction process described above. A reconstruction, once mathematically represented, can distinguish between the possible and impossible phenotypic states. Basically, one can find the fundamental constraints that come with all the interactions between the biomolecular components of a cell and form what is called a solution space. This space contains all the allowable network states. It is analogous to the term reaction norm used in ecology and genetics.

- The second step is to find the states within the solution space that are likely to represent the homeostatic state of an organism. These states are found by using constraint-based opti-

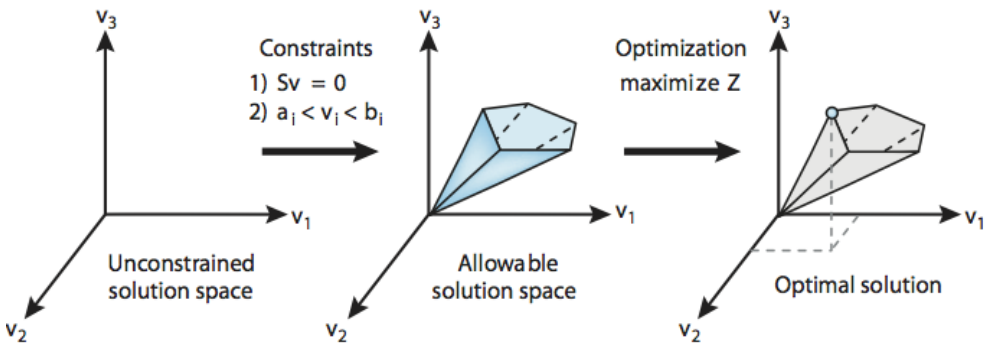

Figure 1.2: With no constraints, the flux distribution of a biological network may lie at any point in a solution space. When mass balance constraints (Sv=0) imposed by the stoichiometric matrix S (labeled 1) and capacity constraints imposed by the lower and upper bounds (ai and bi) (labeled 2) are applied to a network, it defines an allowable solution space. The network may acquire any flux distribution within this space, but points outside this space are denied by the constraints. Through optimization of an objective function, one can identify an optimal flux distribution that lies on the edge of the allowable solution space. Taken from [2]

mization methods. Optimization requires an objective function that describes the desirable properties of the homeostatic state. Constraint-based optimization finds the best points within the solution space based on the stated objective function. The most commonly used objective function is growth rate, although many others have been studied. Ultimately, the objective function describes distal causation, thus representing fundamental biological causation.

Illustrative example

An overview of the COBRA method is shown in Figure 1.3 for a simple toy network, though the same procedure is used to produce COBRA models at the genome scale. The key differences being the number of reactions in the model and a few additional considerations, which will be discussed below in Topic 3.

A COBRA model for the simple toy network shown in Figure 1.3 can be constructed using the following steps. First, the stoichiometries for all of the reactions in the toy network

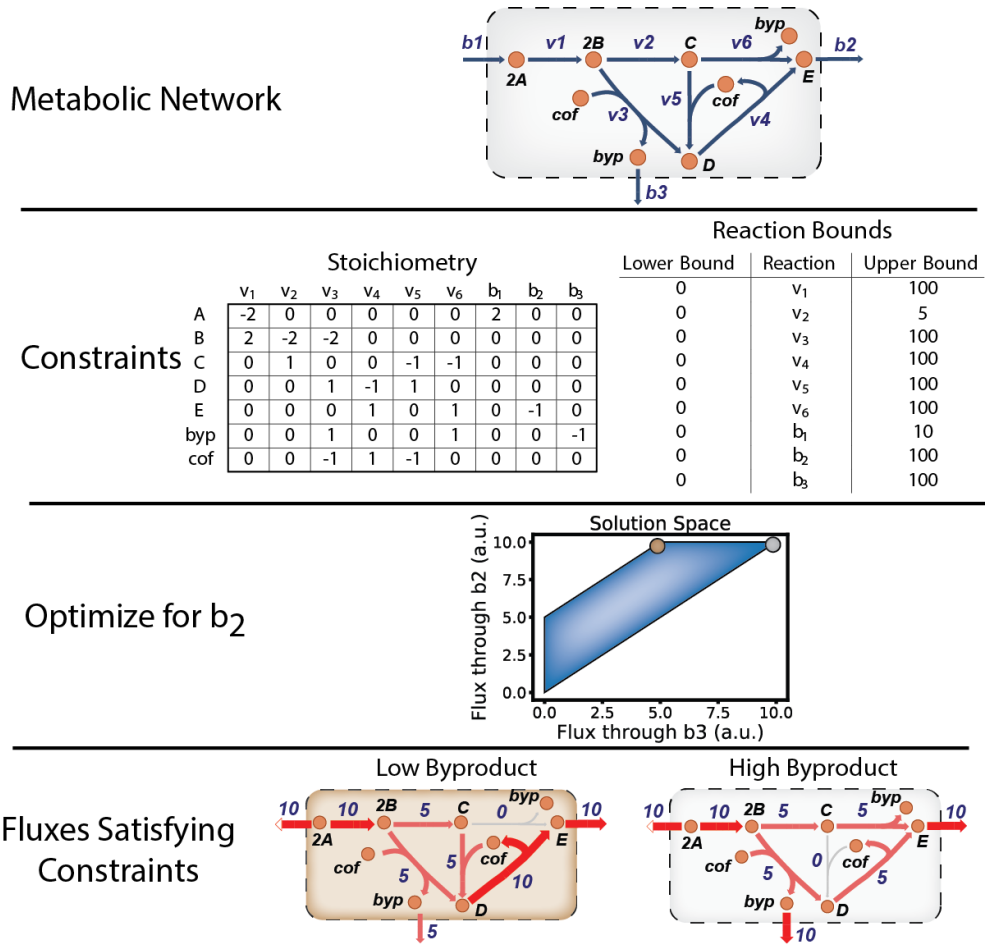

Figure 1.3: Overview of constraint-based modeling method. The toy metabolic network displayed in the top panel can be simulated by converting it into a mathematical format as shown in second panel. Doing so involved first converting the stoichiometry of all of the reactions in the network into a matrix representation where the columns correspond to reactions, the rows correspond to metabolites, and the matrix values correspond to the stoichiometry of the metabolite in the reaction. Further, the upper and lower limits on each reaction must be defined as reaction bounds, as shown. This representation can be converted into a linear programming problem using available software and solved. The solution to this problem however is not unique and the full range of possible solutions can be represented as a solution space, shown in the third panel. Two of the labeled vertices of the solution space are shown overlaid on the network in the bottom panel which are equally optimal solutions that produce the most reaction flux through b2. These can be considered low byproduct and high byproduct solutions based on the amount of metabolite byp excreted.

must be defined including all possible inputs (b1) and all outputs (b2 and b3) of the system. This definition is trivial for the toy network shown, but when executing this process on a genome-

scale, there are numerous additional considerations that must be addressed. Next, the catalog of reactions must be transformed into a mathematical format that lends itself to computation. This transformation is accomplished by constructing what is called a stoichiometric matrix, shown in Figure 1.3. In this format, each column represents a unique reaction and each row a unique metabolite. The numbers in the matrix thus indicate the stoichiometry of the metabolite in each reaction, with negative numbers representing reactants, and positive numbers representing products. The matrix entry is 0 if the metabolite does not participate in the reaction.

Constructing the stoichiometric matrix alone does not produce a useful COBRA model. Bounds must be placed on the activity of each individual reaction as well. Bounds provide the limits on the amount of flux each reaction can carry. In the toy example the units of the reaction flux is ambiguous, but at the genome scale reaction fluxes have units of mmol per gram dry weight of cell per hour. In the toy example, all reactions are irreversible and operate only in the forward direction, indicated by reaction bounds being greater than or equal to 0 for each reaction. If reactions operate only in the reverse direction and/or were reversible, the lower reaction bounds would be below zero. Reaction b1 defines the input of species into the system, which is metabolite A in the toy network and limited to 10 units of uptake flux (Figure 1.3). This flux limitation is analogous to the substrate uptake rate for an organisms genome-scale model.

After defining the stoichiometric matrix and reaction bound constraints, we can define a particular reaction to optimize (i.e., maximize or minimize its reaction rate, or flux) subject to the constraints inherent in the model. One of the several freely available software tools can then be used to compute with the model using a linear programming solver. Solutions to a COBRA metabolic model are non-unique, meaning there are an infinite number of flux states that can produce an optimal result. The full extent of fluxes that satisfy constraints can be represented

as a solution space, depicted as the polygon for the toy network in Figure 1.3. This example shows all of the fluxes possible through reactions b2 and b3 given the stoichiometric and reaction bound constraints. Two extremes of the solution space are shown. These are equally optimal solutions that produce a maximum amount of flux through the b2 reaction.

Escherichia coli as a model organism for systems biology

The organism with the most extensive, validated, and widely-used COBRA model is *E. coli* . As a model organism, many metabolic processes were first discovered in *E. coli* and are best characterized in this organism. This rich *E. coli* bibliome has culminated in the recent, most comprehensive COBRA model of *E. coli* K-12 MG1655, iML1515. Information regarding the pathways and processes modeled in iML1515 and other high-quality models can be found at bigg.ucsd.edu .

Scalability: from a strain to a species to infections on a national scale

Beyond predicted metabolic capabilities of the lab strain, the MG1655 model can help us understand the *E. coli* species as a whole. An initial study of 55 sequenced *E. coli* strains demonstrated the ability to predict auxotrophies and niche-specific nutritional requirements [3], that were subsequently repeated with over 1200 strains [4].

Prediction from sequence alone opens up new possibilities. For example, using a genome sequence of a clinical isolate, *i*ML1515 can be used to construct COBRA models tailored to that specific pathogen. With the number of sequenced clinical *E. coli* isolates, and those of other major pathogens, increasing at a staggering pace (Figure 1.4), genome-scale models offer a means to extract essential information that might lead to guidelines on how to treat a specific unique

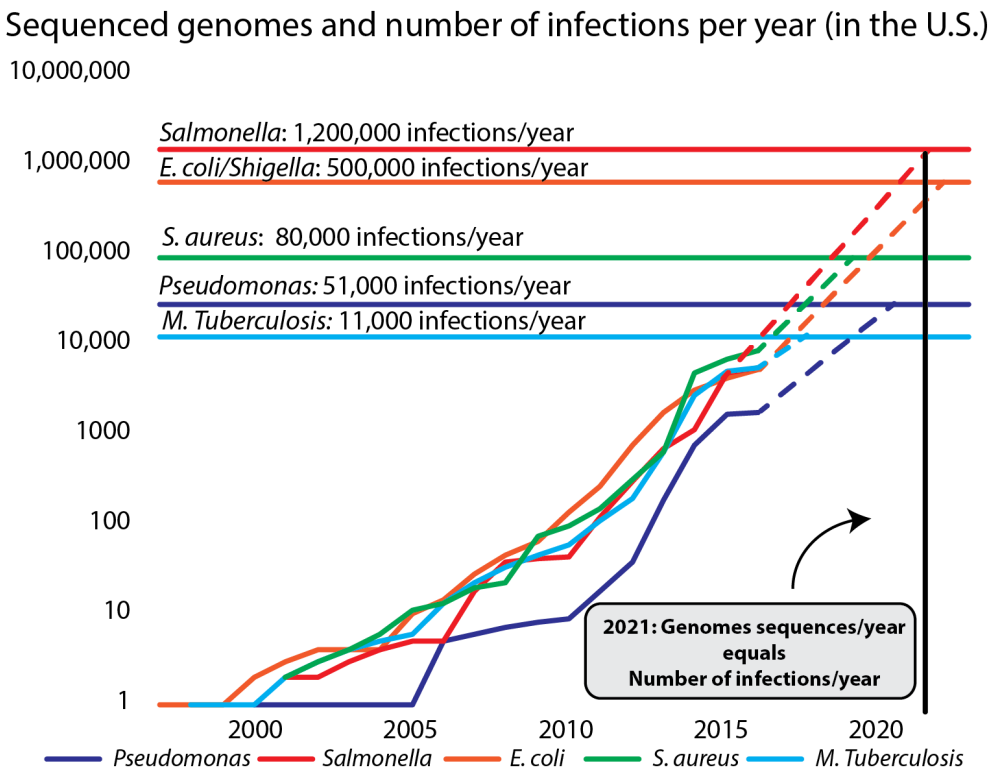

Figure 1.4: Increase in sequenced genomes for five common clinical pathogens. If current trends are maintained, the number of sequenced genomes for all five pathogens will equal the number of infections per year by 2022. The number of sequenced genomes per organism were obtained from the Patric database. The infections per year were obtained from the Center for Disease Control (www.cdc.gov) with the infections per year obtained for S. aureus from the Pew Charitable Trusts (www.pewtrusts.org).

infection. If such predictions materialize, we can move the field of infectious disease closer to truly personalized medicine. If the current rate of sequencing continues it becomes feasible by ca. 2021 to sequence all clinical isolates in the United States (Figure 1.4). A truly amazing prospect.

1.2 Modeling Metabolic Capabilities (M-models)

Pathway vs genome-scale networks viewpoints

Traditionally, microbiology is taught from a pathway or biochemical reaction perspective. In reality, however, cellular survival relies on the activity of many such interconnecting pathways

working in conjunction to produce the metabolites needed for growth in the proper amounts. Beyond this, these individual pathways vary among species and even to some extent within strains of the *E. coli* species itself. Obtaining actionable knowledge by mapping out the metabolic pathways within an organism is difficult. To this end, systems biology methods provide the tools that enable the use of this information to compute phenotypic and metabolic characteristics of an organism despite the complex reality of cellular metabolism.

The power of systems biology can be illustrated by analyzing the 11 reaction linear pathway to produce L-histidine from ribose. It is shown in Figure 1.5 as is it might be discussed in a traditional microbiology book or class. What is typically glossed over, however, is the role that cofactor or metabolic precursor availability may play in shaping the functioning of this pathway or how the activity of any byproducts that are synthesized may affect other metabolic activities in the cell. When using flux balance analysis to simulate the production of L-histidine from ribose, 72 total reactions must be active in order to enable the function of the 11 reaction pathway in a fully mass balanced way.

Why are these 61 extra reactions necessary?

L-histidine synthesis is an anabolic process relying on chemical or energetic contributions from various metabolites in the cell. For three reactions in the pathway, energy from ATP is required to fuel metabolic conversions or to transfer high energy phosphate bonds to L-histidine precursors to fuel future metabolic conversions. Synthesizing ATP thus requires that some of the substrate (glucose) is diverted from reactions in the L-histidine synthesis pathway to the pentose phosphate pathway and, eventually, for glycolysis to produce ATP directly and indirectly by producing charged NADH for oxidative phosphorylation. Further, L-histidine contains three

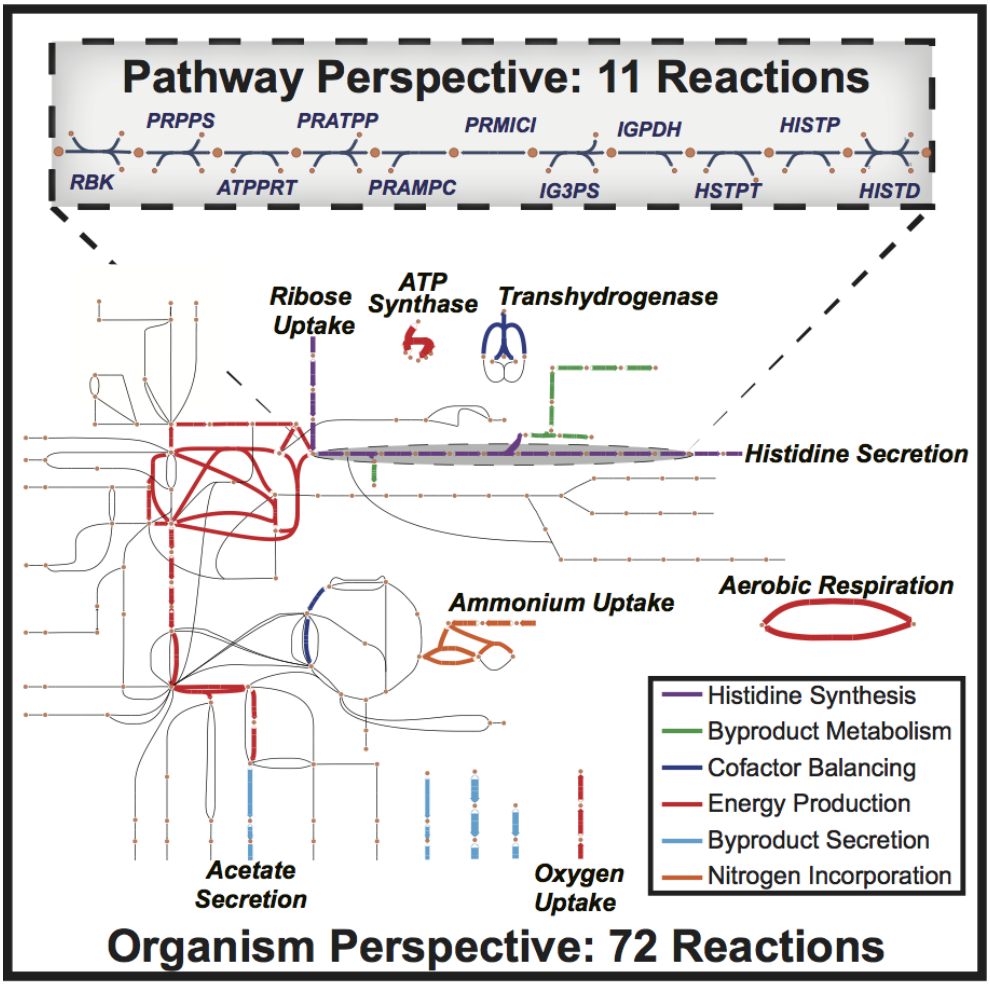

Figure 1.5: Contrasting the traditional pathway (i.e., textbook) view and the network view (i.e., systems biology) of the biochemical processes needed to produce a single biosynthetic precursor. The histidine biosynthetic pathway is considered 'long,' being composed of 11 enzyme-catalyzed reactions on 1 operon. When one examines all the biochemical reactions needed to synthesize histidine from external substrates in a way that is fully mass-balanced, one discovers that a total of 72 reactions are actually needed for a cell to synthesize histidine. Below we discuss the full proteome requirement for such biosynthesis.

nitrogen atoms that it must acquire given that ribose does not contain nitrogen. One nitrogen is incorporated from adenosine and the other two are added from L-glutamine and L-glutamate. The pathways must then be active to incorporate inorganic nitrogen into these metabolites. Side products of the pathway must also be degraded. One such byproduct of L-histidine synthesis is 5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxamide (AICAR) which is degraded to ADP

in this example, though in vivo AICAR is used for inosine biosynthesis. Lastly, all NAD/NADH cofactor usage must be balanced, meaning that NADH must be produced and consumed at equal levels (Figure 1.5). Synchronizing all these functions requires a total of 61 active biochemical reactions.

Networks are highly interconnected, and thus their characteristics can be non-intuitive

Many important properties of the network cannot be determined based on simple characteristics or intuition. For instance, there is no relationship between the predicted carbon yield (i.e. the fraction of carbons in a substrate that end up in the product metabolite) and the number of reactions separating the substrate and product metabolite (Figure 1.6). Ribose, glucose, and other metabolites with few enzymatic reactions separating them from L-histidine are actually capable of synthesizing L-histidine at a lower carbon yield than substrates separated from L-histidine by many more intermediate reactions, like acetate.

Acetate, however, is a poor-quality substrate for *E. coli* , meaning that it grows slowly when fed this metabolite. This is due to the fact that growth is dependent on the organism?s ability to produce all biosynthetic precursors and energy necessary to grow from a given substrate. Therefore, while acetate may be able to produce L-histidine efficiently, it is a highly oxidized carbon source meaning that it cannot produce ATP as efficiently as substrates like glucose, limiting its growth capabilities.

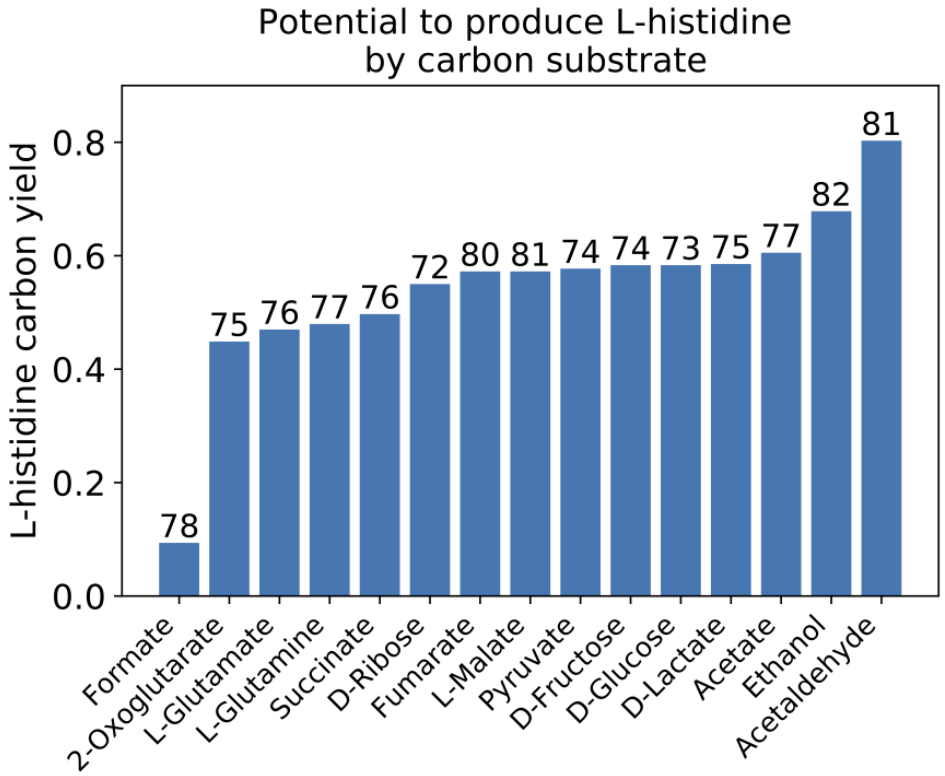

Figure 1.6: Maximal aerobic carbon yield of L-histidine from 15 *E. coli* carbon substrates. The number of active reactions required for each substrate to produce L-histidine is shown above each bar.

From a single amino acid to synthesis of all biomass constituents

This systems-level analysis has shown that even the activity of a short, linear pathway with few branch points relies on the interconnectivity of many different metabolic subsystems in order to operate. Beyond this, all of the considerations discussed for the L-histidine biosynthesis pathway hold true for all other essential biosynthetic precursors that need to be synthesized in a growing cell. For an organism to successfully synthesize the dozens of macromolecular precursors and produce an adequate amount of energy needed to grow, all of these pathways must be active and working in harmony.

In order to take the biosynthesis of these biomass constituents into account, COBRA

models include a reaction called the biomass objective function. This function includes all of the essential metabolites (i.e., structural components, biomass building blocks, cofactors and currency metabolites, etc.) with empirically derived coefficients corresponding to their concentrations in a growing cell. The flux through this reaction corresponds to the *in silico* growth rate of the cell under the environmental conditions imposed. Since the production of each metabolite in the biomass objective function must be produced in the proper amounts and in a fully mass balanced way, the interconnectivity of the pathways producing each essential metabolite can be fully accessed in model simulation. Remarkably, the power of systems science allows us to perform these computations.

1.3 Modeling protein-limited growth (ME-models)

Proteome size is limited

The availability of protein within a growing cell is limited, to about 2 to 3 million protein molecules per cell. Cells have therefore evolved to utilize the available proteome as efficiently as possible. This optimization of proteome investment has led to certain unintuitive and sometimes seemingly 'wasteful' growth phenotypes (e.g. acetate overflow in *E. coli* or the Warburg effect in cancer cells) that require the consideration of protein cost to be accurately described and understood. While the COBRA methods described in Topics 2 and 3 have proven effective in computing the metabolic potential and characteristics of an organism, they do not account for protein costs. In other words, metabolic network models do not consider the proteome investment required to synthesize the enzymes that catalyze flux through a metabolic network.

This lack of consideration of proteome cost means that any two pathways converting metabolite A to metabolite D are treated equally by the model regardless of the efficiency of

the enzymes catalyzing the reactions, the length of the pathways, or the metabolic investment required to synthesize the enzymes making the metabolic conversions. Using an economic analogy, the ?operating expense? of a metabolic reaction is accounted for in a metabolic model, but not the ?capital expense? required to build the protein that catalyzes the reaction. Ignoring proteome costs can lead to metabolic model predictions of incorrect or unrealistic pathway activity and the presence of alternative optimal reaction flux states (Figure 1.3).

Genome-scale models can compute proteome composition

COBRA models have been developed that explicitly describe the processes involved in gene expression and proteome synthesis. These genome-scale models of proteome synthesis impose a biosynthetic cost of catalyzing a metabolic reaction in the model by requiring the synthesis of the catalyzing enzyme in order for a reaction to carry flux. They are called ME-models, for Metabolism and Expression.

Proteome cost is imposed by considering the dilution of macromolecules to daughter cells as the organism grows and divides (Figure 1.7A). The faster the cell is growing, the more the macromolecules are passed on and therefore must be replaced through model synthesis. This relationship provides the basis for genome-scale models of proteome synthesis and their ability to link the synthesis of a macromolecule to the reaction which it catalyzes. For a more thorough explanation on this computational method, refer to [5, 6].

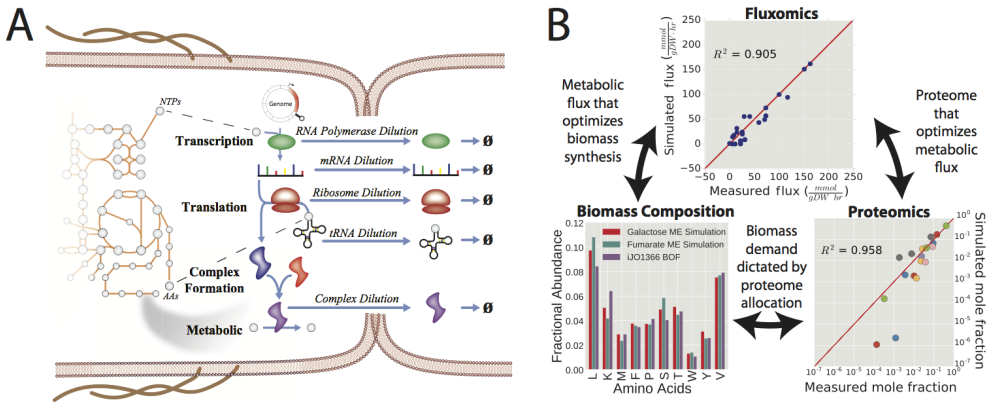

Figure 1.7: ME-model overview and capabilities. **A**. A depiction of each of the major cellular processes involved in gene expression, all of which are modeled in gene expression models. **B**. In order for a single genome-scale model of proteome synthesis (ME-model) to solve, multiple processes involved in cellular growth must be reconciled. For example, the cellular reaction flux state must be supported by the proteome catalyzing all of these processes. The proteome, in turn, is itself supported by biomass precursor synthesis. The biomass precursor demand for the proteome then informs the optimal metabolic flux state. Solutions of ME-models can then be validated against disparate OMICs data types (i.e., 13C fluxomics, proteomics, etc.). Panel A is from [6].

Proteome limitations lead to optimal proteome allocation that can be a governing constraint

How do ME-models improve the framework we introduced in Topic 2? First, the nutrient uptake and metabolic reaction fluxes are determined by optimal protein utilization as a consequence of limited proteome availability. By accounting for proteome constraints, and the assumption of efficient protein allocation, nutrient uptake rates are computed accurately. In addition, ME-model solutions more accurately estimate the metabolic flux distribution by eliminating biologically infeasible, alternative optimal flux distributions (described in Topic 2) that do not satisfy the proteome constraint. This optimal solution of a ME-model is often unique. Second, the relative abundance of proteins and other macromolecules required to support the metabolic phenotype is predicted by a ME-model. These macromolecule fluxes provide novel predictions of macromolecular expression that can be compared against proteomics data or other

numerous OMICs data types (Figure 1.7B). The unique predictions enabled by ME-models open entirely new avenues of biological questions and discovery.

1.3.1 The condition-dependent proteome

The genome-scale computation of the composition of the proteome is just the first step in computationally representing a functioning proteome. To become functional, many proteins require post-translational modification and engraftment of prosthetic groups. Additionally, there are constant degradative forces operating on a proteome. Remarkably, with the chemical basis for these functions, they can be explicitly reconstructed and added to ME-models that compute the composition of the proteome. Here we briefly describe how such in silico models of whole cells deal with, 1) the function of iron in the functional center of enzymes, and 2) how chaperones keep the proteome correctly folded in the face of thermally caused protein misfolding.

1) Functionalizing the proteome: metalloproteins

The catalytic activity of inorganic iron has played a central role in facilitating important metabolic processes in organisms dating back to the beginning of life on earth. As a result, many important processes in living cells still depend on the availability of iron in order to function (Figure 1.8A). Using ME-models, we can predict how reductions in the availability of iron causes the metabolic phenotype of *E. coli* to change. In agreement with experimental observations, decreasing iron availability causes the cell to grow much slower and shift to secreting lactate instead of acetate (Figure 1.8BC). The predicted causes of this shift is the reduced activity of the metabolic enzymes that have iron in their catalytic site. The reduction of flux through the iron-containing enzymes, causes the pathway shift.

Beyond studying growth and metabolism under iron-limited conditions, ME-models have been developed that incorporate damage and repair processes of iron-containing proteins in the presence of oxidative stress. In doing so, these models have the capability to correctly predict amino acid auxotrophies in *E. coli* when exposed to elevated oxidative stress in superoxide dismutase mutants [7].

These two examples highlight the importance of accounting for the metallo-proteome when fully modeling cell metabolism. Both the availability of iron and oxidation load on iron can be described by elementary processes and thus included in an *in silico* model of a cell.

2) The physical integrity of the proteome: proteostasis

Beyond incorporating vital coenzymes, prosthetic groups or metal ions, proteins must also be properly folded to enable their proper catalytic function. By incorporating chaperone activity into these models we can now compute the systems level consequences of protein unfolding and refolding (proteostasis). To begin to model the protein unfolding response, we must first understand that proteins themselves have intrinsic capabilities to keep their shape. In a recent study, we described how computing properties of individual proteins can determine when they are targeted by the chaperones of *E. coli* (Figure 1.9) (Chen et al., 2017).

How are these protein properties incorporated into genome-scale models? The recently introduced GEM-PRO pipeline (genome-scale models with protein structures) enables the curation and computation of protein sequences and structures within genome-scale models (Figure 1.9A) (Brunk et al., 2016; Mih et al., 2018). Thousands of 3D protein structures are solved experimentally every year, thanks to efforts of structural biologists worldwide and programs such as the Protein Structure Initiative (Montelione, 2012). However, a large gap remains between known

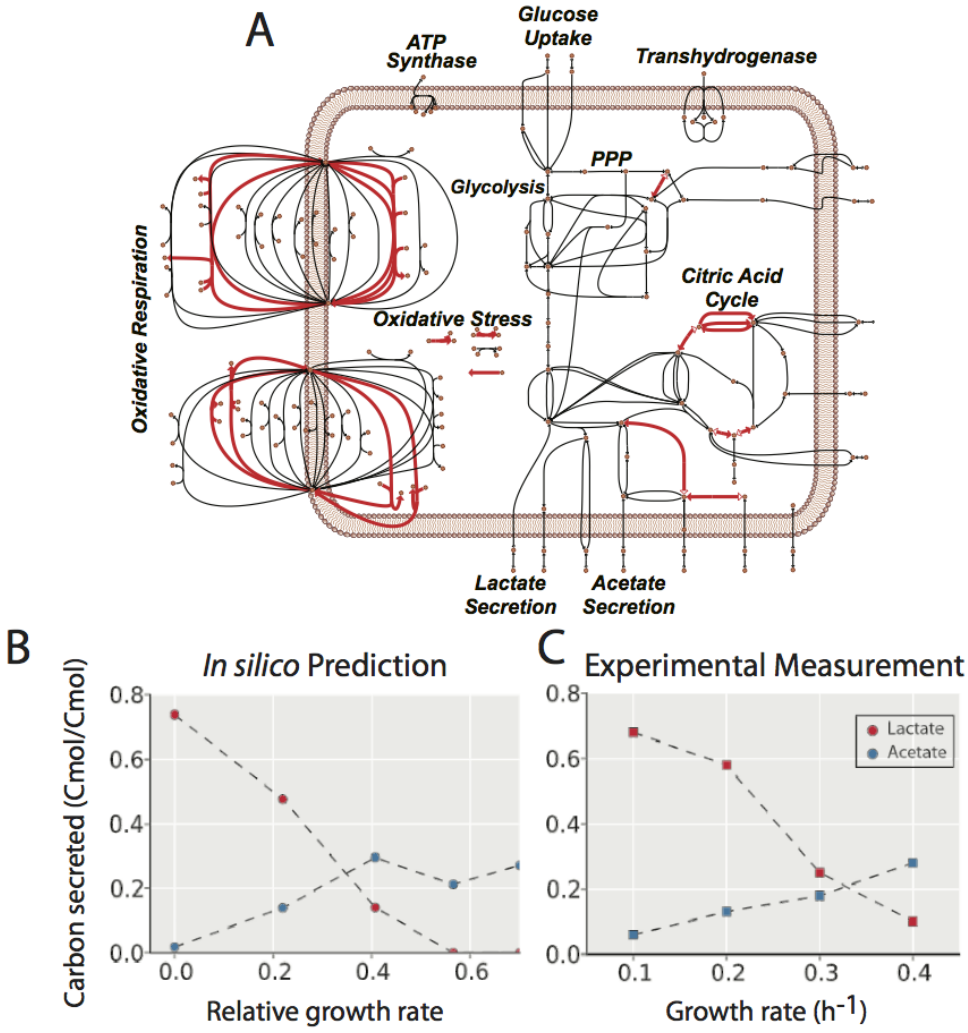

Figure 1.8: Modeling iron-limited growth. **A**. Central carbon metabolism with the reactions catalyzed by iron-containing enzymes in red **B**. *In silico* predictions of relative growth rate and metabolic byproduct secretion when under varying levels of iron limitation. **C**. These in silico predictions agree with experimental measurements of *E. coli* growing in iron limitation. Parts **B** and **C** adapted from [8].

structures and the total number of annotated open reading frames within sequenced genomes. Due to this gap, structural information remains incomplete for a number of reasons, such as the protein (membrane proteins are difficult to crystallize) or organism of interest (some organisms are not as well studied). Fortunately, for model organisms such as *E. coli* , we have now reached a point where structural information is available or can reliably be predicted for a large portion

of the expressed proteome. This enables what is broadly known as a structural systems biology for genome-scale models.

Once a GEM-PRO has been generated for *E. coli* , we can then predict or compute protein properties as a function of temperature from first principles (Figure 1.9B). Incorporating two major chaperone systems and their folding networks (Figure 1.9C) to the ME-model thus allows us to understand how changes in these protein properties affect the systems-level response to changes in temperature. This model, named foldME, thus enables a surprisingly accurate prediction of relative growth rates under different temperatures (Figure 1.9D). Furthermore, foldME allows us to inspect changes in the production levels of certain proteins (i.e., the cells proteome allocation). Under high temperatures, there is a dramatic shift toward the production of chaperones while shifting away from the production of cytoplasmic proteins (Figure 1.9E).

Thermostability has often been a property studied in the context of individual proteins, or as a global phenotypic response. FoldME exemplifies how incorporating atomic-level information can expand the systems biology paradigm, and it provides an understanding of how a cell balances its protein production and maintenance (folding) machinery under thermal stress.

1.4 Multi-strain reconstructions enable assessment of strain variation

Genomic sequences from multiple strains allow us to ponder the definition of a species. As shown in Figure 1.4, there has been a staggering increase in the number of sequenced strains of *E. coli* . These sequenced genomes have allowed us to adapt reconstructed COBRA models to inspect variation at both the gene-level (i.e., absent metabolic reactions that may cause aux-

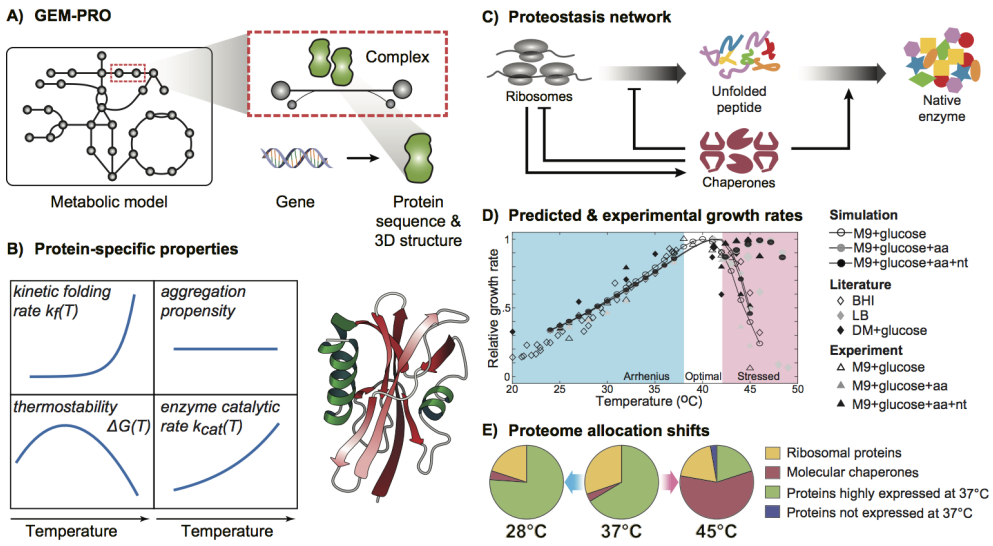

Figure 1.9: Utilization of protein structural properties within a genome-scale model for the purpose of modeling thermal stress adaptations. **A)** The formal integration is termed a GEM-PRO, or genome-scale model with protein structures. Enzymatic reactions are mapped to their corresponding protein sequences and structures. **B)** Protein-specific properties can be predicted from sequence and structure, and further computed to reflect protein property changes under stress. For example, four parameters are chosen to represent the influence of temperature on the status of the structural proteome (Chen et al., 2017)). **C)** Incorporation of chaperone folding networks adds the proteostasis network response to the model. **D)** Simulated growth rates of an integrated model match very closely to measured experimental rates at different temperatures. **E)** Proteome allocation of the cell under thermal stress can be inspected to find a dramatic shift towards chaperone production at high temperatures. Adapted from [9] and [10].

otrophies) and at the base-pair level (i.e., changes to a protein sequence and their impact on structure). This sequence data availability brings into focus the question: what is a strain and what is a species? A comprehensive comparison of the metabolic capabilities of 55 strains of *a*nd *Shigella* revealed that these strains can adapt to catabolize nutrients with alternate pathways. This result exemplifies how COBRA models provide utility by predicting a strains environmental niche. This approach has been applied to characterize differences in industrial strains of *E. coli* , the aquatic bacterium *Shewanella, Staphylococcus aureus, Leptospira, and Pseudomonas putida.*

Structures can provide a template for analyzing sequence variation. However, some pathways are often quite conserved and essential in many organisms. The L-histidine biosynthesis

pathway, for instance, is known to be essential across bacteria, fungi, plants, and archaea, and is present in most strains of *E. coli* . As such, differences in this pathway among strains must now be probed at a deeper level to begin to understand what enables the broad diversity within a single species. An initial analysis of sequence-level variation in over 1000 clinical isolates of *E. coli* revealed a large number of alleles (unique variant forms of the protein) for enzymes in L-histidine biosynthesis and a number of other pathways related to amino acid synthesis (Figure 1.10A). This analysis was then coarse-grained to consider protein domains, which can be thought of as the functional units of proteins. This domain-level analysis indicated that the aldehyde dehydrogenase (ALDH-like) domain has, on average, more variation across all studied strains compared to other domains (Figure 1.10B). The last enzyme in L-histidine biosynthesis, hisD, contains this ALDH-like domain and had one of the highest counts of alleles, as well as number of mutations conferring an allele (Figure 1.10C, 10D). The potential for further characterization of these enzymes due to these variations remains, as these are but clues to begin to understand the true relationship between genotype and phenotype.

Acknowledgements

All authors helped draft and edit the final review. The work was funded by the Novo Nordisk Foundation and by grant 1R01GM057089 from the NIH/NIGMS.

Chapter 1 in part is a reprint of material published in: **CJ Lloyd**, N Mih, L Yang and BO Palsson. 2019. "Fundamentals of Metabolic Systems Biology." *Encyclopedia of Microbiology*, The dissertation author was the primary author.

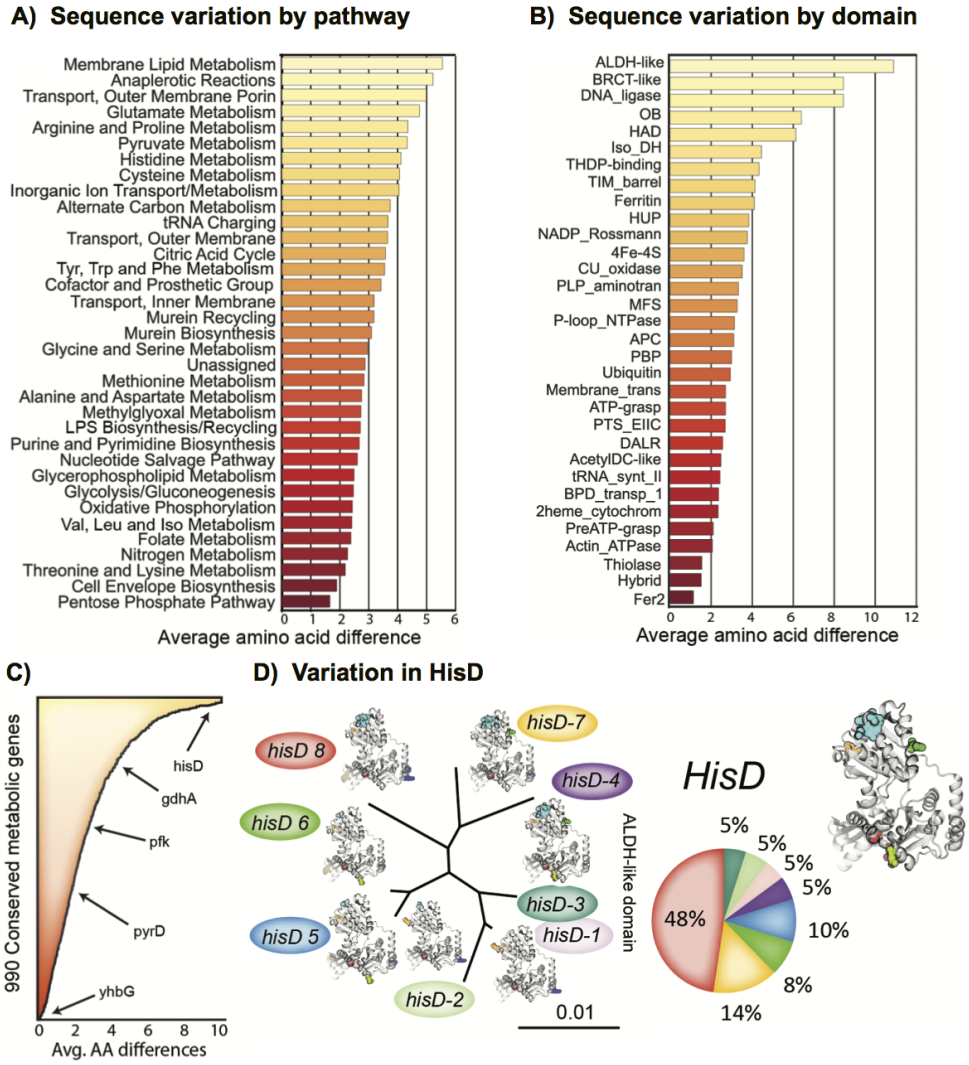

Figure 1.10: Analyzing strain variation in *E. coli* at multiple levels. **A)** Inspecting variation at the pathway level shows that membrane lipid metabolism and a number of amino acid synthesis pathways top the list of average number of changes in their catalyzing enzymes. **B)** Incorporating protein domains into the analysis points out specific domains that contain this variation. **C)** 976 genes with metabolic functions are conserved across 99% of *E. coli* strains, and this core set also has mutations. The bar chart shows the average number of amino acid mutations in these core genes for 1,122 strains of *E. coli* . D) Histidine pathways showed high levels of amino acid differences among genes involved. The pie chart represents the percentage of strains that contain unique hisD alleles. The hisD allele in *E. coli* K-12 MG1655 is present in only 19 (1.7%) of the clinical isolates. Adapted from [4].

1.5 References

1. Monk, J. & Palsson, B. O. Predicting microbial growth. *Science* **344,** 1448–1449 (2014).

2. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? en. *Nat. Biotechnol.* **28,** 245–248 (Mar. 2010).

3. Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., Feist, A. M. & Palsson, B. Ø. Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. en. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 20338–20343 (Dec. 2013).

4. Monk, J. M., Lloyd, C. J., Brunk, E., Mih, N., Sastry, A., King, Z., Takeuchi, R., Nomura, W., Zhang, Z., Mori, H., Feist, A. M. & Palsson, B. O. iML1515, a knowledgebase that computes Escherichia coli traits. en. *Nat. Biotechnol.* **35,** 904–908 (Oct. 2017).

5. O'Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R. & Palsson, B. Ø. Genom-efffdfffdfffdscale models of metabolism and gene expression extend and refine growth phenotype prediction. *Molecular Systems Biology* (2013).

6. Lloyd, C. J., Ebrahim, A., Yang, L., King, Z. A., Catoiu, E., O'Brien, E. J., Liu, J. K. & Palsson, B. O. *COBRAme: A Computational Framework for Building and Manipulating Models of Metabolism and Gene Expression* en. Feb. 2017.

7. Yang, L., Mih, N., Yurkovich, J. T., Park, J. H., Seo, S., Kim, D., Monk, J. M., Lloyd, C. J., Tan, J., Gao, Y., Broddrick, J. T., Chen, K., Heckmann, D., Feist, A. M. & Palsson, B. O. *Multi-scale model of the proteomic and metabolic consequences of reactive oxygen species* en. Dec. 2017.

8. O'Brien, E. J., Utrilla, J. & Palsson, B. O. Quantification and Classification of E. coli Proteome Utilization and Unused Protein Costs across Environments. en. *PLoS Comput. Biol.* **12,** e1004998 (June 2016).

9. Mih, N., Brunk, E., Chen, K., Catoiu, E., Sastry, A., Kavvas, E., Monk, J. M., Zhang, Z. & Palsson, B. O. ssbio: A Python Framework for Structural Systems Biology. en. *Bioinformatics* **34,** 2155–2157 (Feb. 2018).

10. Chen, K., Gao, Y., Mih, N., O'Brien, E. J., Yang, L. & Palsson, B. O. Thermosensitivity of growth is determined by chaperone-mediated proteome reallocation. *Proceedings of the National Academy of Sciences* **114,** 11548–11553 (Oct. 2017).

Chapter 2

A computational framework to empower ME-model development

Genome-scale models of metabolism and macromolecular expression (ME-models) explicitly compute the optimal proteome composition of a growing cell. ME-models expand upon the well-established genome-scale models of metabolism (M-models), and they enable a new fundamental understanding of cellular growth. ME-models have increased predictive capabilities and accuracy due to their inclusion of the biosynthetic costs for the machinery of life, but they come with a significant increase in model size and complexity. This challenge results in models which are both difficult to compute and challenging to understand conceptually. As a result, ME-models exist for only two organisms (Escherichia coli and Thermotoga maritima) and are still used by relatively few researchers. To address these challenges, we have developed a new software framework called COBRAme for building and simulating ME-models. It is coded in Python and built on COBRApy, a popular platform for using M-models. COBRAme streamlines

computation and analysis of ME-models. It provides tools to simplify constructing and editing ME-models to enable ME-model reconstructions for new organisms. We used COBRAme to reconstruct a condensed *E. coli* ME-model called iJL1678b-ME. This reformulated model gives functionally identical solutions to previous *E. coli* ME-models while using 1/5 the number of free variables and solving in less than 10 minutes, a marked improvement over the 6 hour solve time of previous ME-model formulations. Errors in previous ME-models were also corrected leading to 52 additional genes that must be expressed in iJL1678b-ME to grow aerobically in glucose minimal in silico media. This manuscript outlines the architecture of COBRAme and demonstrates how ME-models can be created, modified, and shared most efficiently using the new software framework.

2.1 Background

Genome-scale metabolic models (M-models) have shown significant success predicting various aspects of cellular metabolism by integrating all of the experimentally determined metabolic reactions taking place in an organism of interest [1–4]. These predictions are enabled based on the stoichiometric and thermodynamic constraints of the organisms metabolic reaction network and the metabolic interactions with the environment. M-models are capable of accurately predicting the metabolic capabilities of an organism, but they require defined substrate input constraints and empirical metabolite measurements to make predictions of its growth capabilities. Therefore, a focus of development in the field of genome-scale models has been to increase the scope and capabilities of M-models [5].

Recently, M-models have been extended to include the synthesis of the gene expression machinery which can be used to compute the entire metabolic and gene expression proteome

in a growing cell [6–9]. These ME-models integrate Metabolism and Expression on the genome scale (Figure 2.1), and they are capable of explicitly computing a large percentage (up to 80% in some cases) of the proteome by mass in enterobacteria [10]. In other words, ME-models not only compute optimal metabolic flux states, as with M-models, but they additionally compute the optimal proteome composition required to sustain the metabolic phenotype. ME-models enable a wide range of new biological questions that can be investigated including direct calculations of proteome allocation [11] to cellular processes, temperature dependent activity of the chaperone network [12], metabolic pathway usage, and the effects of membrane and volume constraints [7]. Furthermore, their ability to compute the optimal proteome abundances for a given condition make them ideal for mechanistically integrating transcriptomics and proteomics data.

So far ME-models have been constructed for only two organisms, Thermotoga maritima [8] and Escherichia coli K-12 MG1655 [6, 7, 9, 13]. The slow pace of ME-model construction can be attributed to two basic challenges. First, ME-models are much slower to numerically solve than M-models; it takes 5 orders of magnitude more CPU time to solve iOL1650-ME [6] than it does the corresponding iJO1366 M-model [14] (6 hrs for iOL1650-ME vs 100 ms for iJO1366). Therefore while M-models can be solved on personal computers, ME-models have required large clusters or supercomputers to parallelize simulations. Second, the large model sizes and complex structure have made analyzing and debugging the model difficult and time consuming. M-models can use generalized software tools [15–19], but each organisms ME-model has required its own dedicated codebase and database schema, which makes advances for one organisms model difficult to apply to another organism. Therefore, each organisms ME-model has required dedicated person-years of effort.

We addressed these challenges by developing a computational framework called CO-

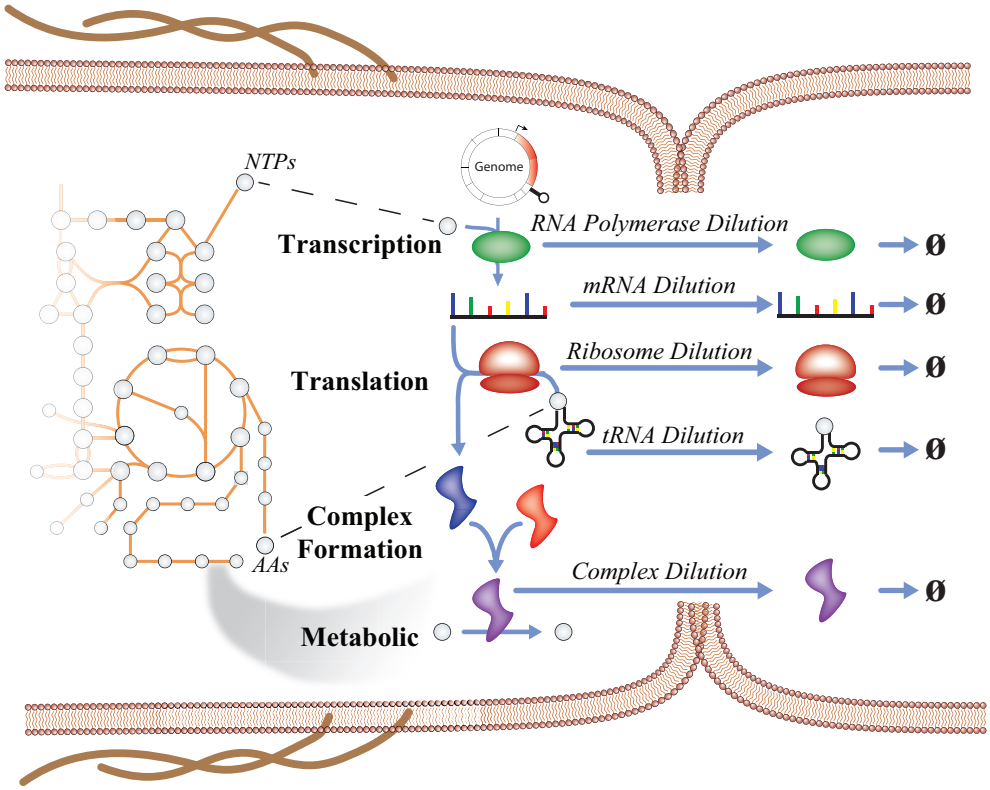

Figure 2.1: Multi-scale processes modeled in a ME-model depicted in a dividing *E. coli* cell. ME-models expand upon underlying M-models by explicitly accounting for the reactions involved in expressing genes which are required to catalyze enzymatic processes. The synthesis of each major macromolecule is coupled to the reaction that it is involved in by accounting for its dilution to daughter cells during cell division. Each dilution is a function of growth rate (μ).

BRAme for building, editing, simulating, and interpreting ME-models. COBRAme is written in Python and extends the widely used COBRApy software that only supports M-models [18]. COBRAme is designed to: 1) support any organism with an existing M-model; 2) use protocols and commands familiar to current users of COBRApy; 3) represent ME-models with an intuitive collection of Python classes; and 4) solve FBA simulations orders of magnitude faster than previous ME-models [6]. As a result of the above considerations, we hope that COBRAme and its associated tools will accelerate the development and use of models of metabolism and expression.

2.2 Design and implementation

Software dependencies

The COBRAme software (S1 File) is written entirely in Python 2.7+/3.5+ and requires the COBRApy [18] software package to enable full COBRA model functionality. Additionally, COBRAme requires the SymPy Python module [20] in order to handle the symbolic variable representing cellular growth rate (), which participates as a member of many stoichiometric coefficients in the ME-model. The BioPython package [21] is used by COBRAme to construct transcription, translation, and tRNA charging reactions for each gene product in the organisms genbank genome annotation file. The ME-model is solved using the SoPlex [22, 23] or quadMINOS [24] solvers via APIs written in Python and included as part of this project. Further, the ECOLIme Python package is included in this work (S2 File) and contains information pertaining to *E. coli* gene expression and scripts to build iJL1678b-ME starting with the *E. coli* metabolic model, iJO1366 [14]. ECOLIme can further act as a blueprint for ME-model reconstructions of new organisms.

ME-model architecture

Constructing a ME-model requires assembling information pertaining to many different cellular processes. For instance, in order to construct a translation reaction for the ME-model, the sequence of the gene, the codon table for the organism, the tRNAs for each codon, ribosome translation rates, elongation factor usage, etc. must be incorporated. Further, several processes in the ME-model recur for many genes that are transcribed or translated due to their template-like nature [13]. To address these challenges, the COBRAme ME-model was structured to compartmentalize information for individual cellular processes. A key component of this ap-

proach was the separation of the ME-model into two major Python class types: the information storage vessels called ProcessData and the functional model reactions called MEReaction, which is analogous to the COBRApy Reaction.

ProcessData

COBRAme constructs ME-models that are composed of two major Python classes. The first of these is the ProcessData class, which is used to store information associated with a cellular process. The type of information contained in each ProcessData type is summarized in the COBRAme Documentation (http://cobrame.readthedocs.io/, S4 File). This method of information storage has several advantages over alternatives such as establishing a database to query information as it is needed, which was the approach used to build previous ME-model versions. For example, this method simplifies the dissemination of the information used to construct a ME-model given that the information can now be included as part of a published ME-model without requiring the user to install and populate a database. Further, this gives the ability to compartmentalize the information based on which cellular processes it elucidates. By storing this information in Python objects, methods can be implemented to further allow data contained in each ProcessData instance to be manipulated. This method also reduces error by enabling many features to be computed using defined inputs in a consistent way. For example, the amino acid sequence for a protein can be dynamically computed and used to construct a TranslationReaction instance using a genes nucleotide sequence and codon table (Table 2.1, S5 File).

Table 2.1: Overview of all ProcessData subclasses.

ProcessData Subclass	Information Contained	Example	Number in *i*JL1678b-ME
StoichiometricData	Metabolite stoichiometry of a metabolic reaction (often equivalent to M-model reaction)	HISTD	2282
ComplexData	Protein subunit stoichiometry of an enzyme complex as well as the modifications required for its activity	CPLX-153	1445
SubreactionData	Some processes occur in multiple steps (e.g. translation reactions) or require modifications. This class details the stoichiometry and catalytic enzyme associated with the process.	ala_addition_at_GCA or mod_2fe2s_c	353
TranscriptionData	Nucleotide sequence, RNA products, sigma factor usage, etc. for a given transcription unit	TU00001_from_RpoD_mono	1447
TranslationData	Subreactions (tRNA mediated amino acid additions), sequence of mRNA/protein, etc. for a given mRNA being translated	b2020	1569
tRNAData	Codon, amino acid, tRNA, and modifications required to make a functioning tRNA	tRNA_b0202_AUU	158
TranslocationData	Keff, enzymes, metabolite stoichiometry of a particular protein translocation pathway	srp_translocation	9
PostTranslationData	Details the translocation pathways, protein modifications (for lipoproteins), etc. required to produce a functioning protein.	translocation_protein_b0733	682
GenericData	List of complexes or metabolites that are redundant and represented as generics	generic_Tuf	11

MEReactions

ME-models are multiscale in nature, meaning they contain reactions that operate on dramatically different scales in time and space whose rates span 15 orders of magnitude [25]. Fast reactions (e.g., metabolic) are coupled to slow reactions (e.g., complex formation) through coupling coefficients that determine the amount of macromolecule needed to catalyze particular reactions. To facilitate this coupling and to handle the unique characteristics of each major reaction type found in cell biology, the MEReaction Python class is used. The MEReaction classes inherit all of the methods of a COBRApy Reaction. In addition to the functionality of COBRApy Reactions, however, MEReactions contain methods to read and process the information contained in ProcessData objects and to update this information into a complete, functional reaction. In many cases, part of compiling a ME-model reaction also includes imposing the appropriate growth rate dependant coupling constraints (coupling constraints detailed in

Table 2.2: ProcessData types used to construct each MEReaction type.

MEReaction Type	ProcessData Information Used	Number in *i*JL1678b-ME
MEReaction	None	2021
SummaryVariable	None	22
MetabolicReaction	**StoichiometericData**, SubreactionData, ComplexData	5266
ComplexFormation	**ComplexData**, SubreactionData	1445
TranslationReaction	**TranslationData**, SubreactionData	1569
TranscriptionReaction	**TranscriptionData**, SubreactionData	1447
PostTranslationReaction	**PostTranslationData**, TranslocationData, SubreactionData	682
tRNAChargingReaction	**tRNAData**, SubreactionData	158
GenericFormationReaction	**GenericData**	44

the COBRAme Documentation and Supplemental Text (S8 File)). These coupling constraints are imposed directly as part of the MEReactions update method and can vary depending on the reaction type. Since MEReactions are directly linked to the information used to construct them through the associated ProcessData, this codebase has the ability to easily query, edit, and update the information and metabolite stoichiometry constituting the MEReaction and therefore the model (Table 2.2, S5-7 Files). Examples of how this ME-model architecture can be leveraged to query and edit reaction information can be found in the COBRAme Documentation.

Most MEReaction types in COBRAme must be linked to at least one ProcessData instance that defines the core information underlying the reaction being represented. The required ProcessData for each reaction is listed in bold.

ME-model reconstruction workflow

ME-models of *E. coli* are reconstructed using the two Python packages presented here, COBRAme and ECOLIme. COBRAme contains the class definitions and necessary methods to facilitate building and editing a working ME-model. COBRAme is written to be organism-agnostic so that it can be applied to ME-models for any organism. ECOLIme contains the *E. coli* specific information (e.g., the *E. coli* ribosome composition) as well as functions required to process files containing *E. coli* reaction information (e.g., the text file containing transcription

unit definitions) and associate them with the ME-model being constructed. Therefore, ECOLIme is required to assemble the reaction and gene expression information that comprises iJL1678b-ME. COBRAme, on the other hand, supplies the computational framework underlying the ME-model. The composition along with further demonstrations of the utility of each of these packages is outlined in the COBRAme Documentation.

The procedure used to build iJL1678b-ME using COBRAme and ECOLIme is presented in the building script, build_me_model (Figure 2.2). This script goes through each of the major gene expression processes modeled in iJL1678b-ME and uses ECOLIme to load all the relevant information. Once the information is loaded, it is used to create and populate ProcessData instances associated with the information. Each of the ProcessData instances are then linked to the appropriate MEReaction instance and updated to form a functioning ME-model (Figure 2.2).

Reformulating the *E. coli* ME-model

Significant efforts were made to simplify the ME-model while also optimizing the model size, modularity, and time required to solve. These included: 1) reformulating the implementation of explicit coupling constraints (metabolites) and 2) lumping major cellular processes such as transcription and translation into single ME-model reactions. Further, a number of updates, changes, and corrections have been made to the *E. coli* ME-model reconstruction which are detailed below.

Macromolecular Coupling

The largest mathematical difference between the original ME-model formulation [6] and COBRAme is the change in the macromolecular coupling implementation. Coupling coefficients

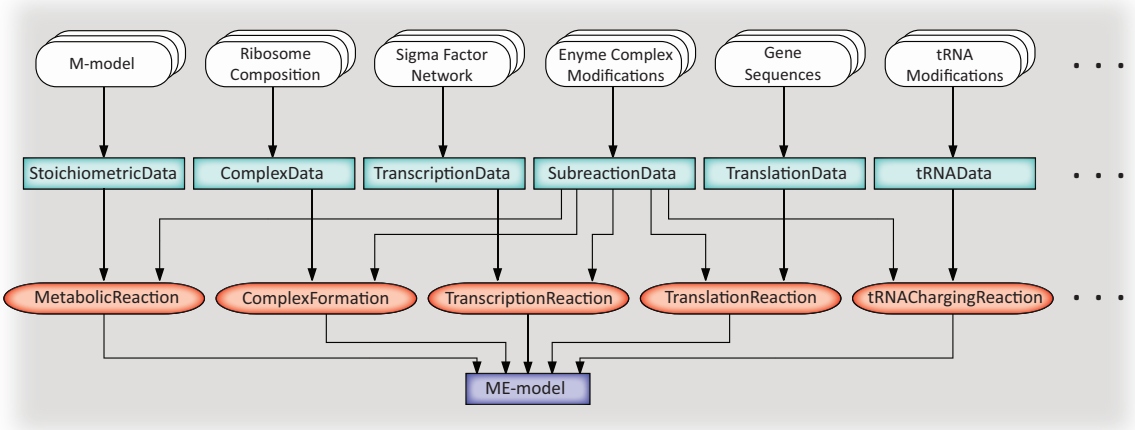

Figure 2.2: The flow of information from input data to the ME-model, as facilitated using the build_me_model script. The build_me_model workflow uses the ECOLIme package to load and process the *E. coli* M-model along with all supplied files containing information defining gene expression processes/reactions. This information is then used to populate the different ProcessData classes (shown in turquoise boxes) and link them to the appropriate MEReaction classes (shown in red ovals), all of which are defined in the COBRAme package. The entirety of the MEReactions comprise a working ME-model. Not all input data, ProcessData classes, and MEReaction classes are shown. For a complete list, reference the COBRAme Documentation.

dictate the amount of macromolecule synthesis flux that is required for the reaction catalyzed by that macromolecule to carry flux. They are derived based on the fact that, as a cell grows and divides, it must dilute macromolecules to its daughter cells. Therefore, coupling constraints have a general form of $\frac{\mu}{k_{eff}}$ [6] (Figure 2.3). While these are essential in a ME-model to couple together the various reaction types, in previous model versions they inflated the number of metabolites and reactions contained in the ME-model (ME-matrix), resulting in longer solve times. COBRAme improves coupling constraint implementation by directly embedding macromolecule dilution coupling into its catalytic reaction (Figure 2.3).

A more thorough description of coupling constraints reformulations and their implementation can be found in the online COBRAme Documentation.

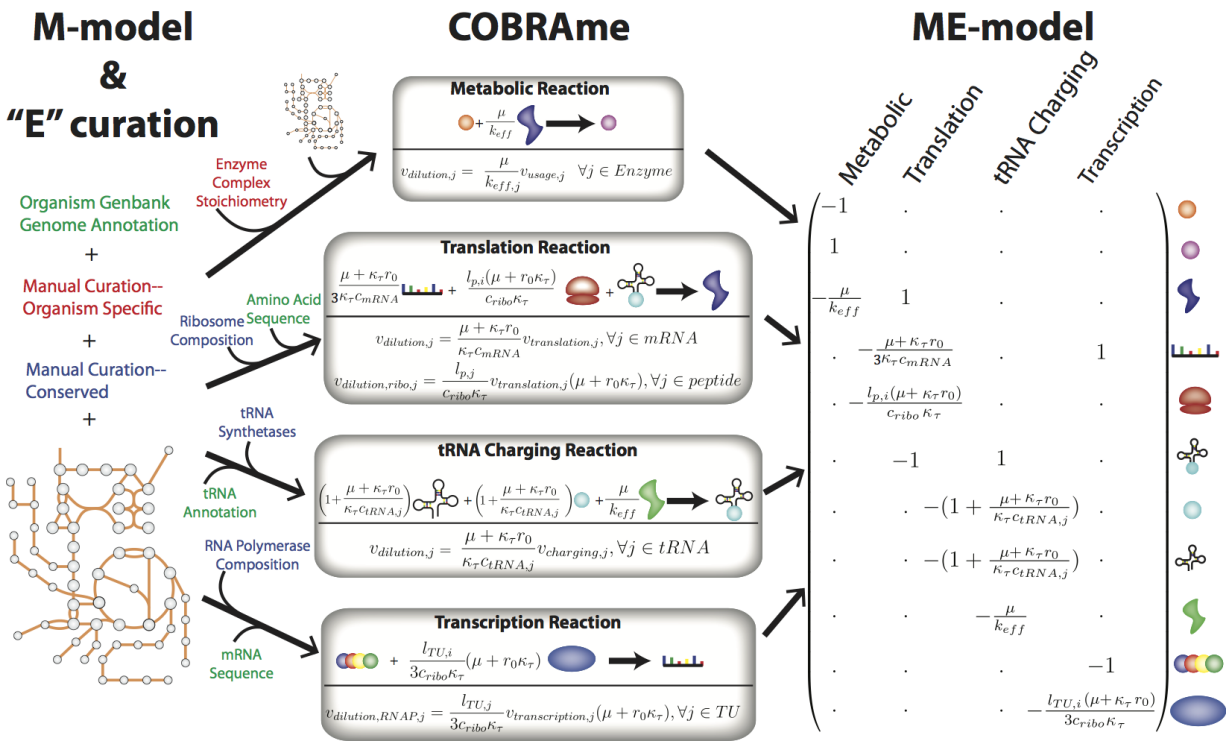

Figure 2.3: An overview of the COBRAme ME-model formulation. The previous ME-models implemented coupling constraints explicitly as model pseudo-metabolites. With COBRAme, instead of using explicit coupling constraints (metabolites), dilution of coupled macromolecules to the daughter cell is accounted for by embedding them directly in the reaction in which they are used. For example, for the metabolic reaction shown above, a small amount ($\frac{\mu}{k_{eff}}$) of the catalyzing enzyme is consumed by the reaction in which it is involved. In other words, for a given amount of flux carried by the metabolic reaction, $\frac{\mu}{k_{eff}} \cdot v_{metabolic\_reaction}$ of the catalyzing enzyme must be synthesized. A subset of the major macromolecular coupling that is applied in iJL1678b-ME is also shown, along with their representation in the ME-matrix. Reference the COBRAme Documentation for derivations and further explanation of the coupling coefficients.

Reaction Lumping

Using equality constraints in the COBRAme formulation and splitting the model into ProcessData and MEReactions allows for a variety of model simplifications. One major simplification is that reactions which occur in a number of individual steps or sub-reactions (i.e., ribosome formation, translation, etc.) can be lumped into a single reaction. The single lumped MEReaction can be constructed by associating it with the multiple ProcessData instances that

detail the individual sub-reactions involved in the overall reaction. All sub-reaction information is further accessible through the MEReaction instance itself which allows the information to be queried, edited, and updated throughout the reaction. If the sub-reaction participates in many different reactions, the changes can be further be applied throughout the entire model. This lumping has the obvious benefit of reducing the number of model reactions, thus shortening the solve time. Lumping complex reactions has the added benefit of making the ME-model much more modular in nature. This simplifies the process of adding or removing new processes associated with the ME-model reaction. Examples of accessing and editing ProcessData through MEReactions can be found in the COBRAme Documentation.

Nonequivalent Changes

Unlike the reformulations described above, some of the changes made in the COBRAme formulation purposefully changed the model in a nonequivalent way. One of the most significant differences was assigning a dummy complex monomer with a representative amino acid composition as the catalytic enzyme for orphan reactions. These are non-spontaneous reactions which do not have a known enzymatic catalyst. The previous ME-model formulations modeled these orphan reactions as spontaneous which resulted in a slight bias toward using these reactions, given that they did not have an associated protein expression cost. This was corrected in iJL1678b-ME. Additionally, in iJL1678b-ME, protein carriers (e.g., acyl carrier protein) are assigned as catalysts to their transfer reactions. Therefore, iJL1678b-ME will require translation of these carriers in order for them to participate in the reactions in which they are involved, thus resulting in the expression of 52 more genes when simulating on glucose minimal media compared to iJL1678-ME.

Further, membrane surface area constraints imposed in iJL1678-ME were removed. This constraint limited the number of membrane proteins that could be expressed at a given growth rate. Protein competition for membrane space may play an important role in shaping *E. coli* s metabolic phenotype, particularly when growing aerobically. Despite this, the constraint was removed to prevent the model from being over constrained when growing in non-glucose aerobic conditions, leading to unrealistic behavior. Removing this constraint makes iJL1678b-ME more flexible and applicable to more in silico conditions. Similarly, growth-dependent surface area calculations were used when imposing lipid demands, therefore they were also removed and replaced with demands identical to those defined in the iJO1366 biomass objective function. The protein translocation genes and pathways added when reconstructing iJL1678-ME, however, remain in iJL1678b-ME.

Additional corrections and changes made when reconstructing iJL1678b-ME are outlined in the Supplemental Text (S8 File).

Optimization Procedure

Unlike M-models, the stoichiometric matrix for each ME-model consists of numerous growth rate (μ) dependent metabolite coupling coefficients and variable bounds (Figs 1 and 3). This makes the ME-model nonlinear, meaning it cannot be solved as a normal LP like M-models. The ME-matrix, however, is quasi-convex [25], meaning that, for any feasible substituted μ, all smaller μ values will also be feasible. Therefore, the maximal feasible μ value can be determined by a binary search or bisection algorithm wherein successive linear programs are solved at different values of μ to find the largest value of μ that gives a feasible flux state, as done for iJL1678-ME and iOL1650-ME. For each optimization, the production of a representative dummy protein is

maximized. In doing so, it allows the same algorithm to be used for both batch and nutrient limited growth, which required different procedures in iJL1678-ME and iOL1650-ME [6].

While any linear programming solver supported by COBRApy [18] could technically have been used, ME-models are very ill-scaled [6], unlike M-models [26]. Therefore, two specialized solvers are used due to their extended numerical precision, thus ensuring acceptable numerical error: 1) qMINOS [23, 24], which supports quad (128-bit) numerical precision, and 2) SoPlex [22], which supports "long double" (80-bit) numerical precision as well as iterative refinement in rational arithmetic to further reduce numerical error.

2.3 Results and discussion

Model Overview

The COBRAme framework was used to reconstruct a mass-balance checked, reformulated version of the *E. coli* K-12 MG1655 ME-model iJL1678-ME, called iJL1678b-ME (S3 File). This produced a model with 12,655 reactions and 7,031 metabolites (S6 and S7 Files), a marked improvement over iJL1678-ME which contained 79,871 reactions and 70,751 metabolites. As a result, iJL1678b-ME has a matrix with 80% fewer columns than iOL1650-ME. This dramatically speeds up the solving procedure and allows processes such as iterative refinement, which uses rational arithmetic and is unsuited for fast vectorized SIMD operations, to become feasible for fast and accurate solutions (Figure 2.4, S8 File).

iOL1650-ME, constructed using COBRAme, was simulated in glucose aerobic minimal media in silico conditions and compared against simulations from the previous iOL1650-ME version. Both simulations were run using a selection of keff parameters that were fit to proteomics

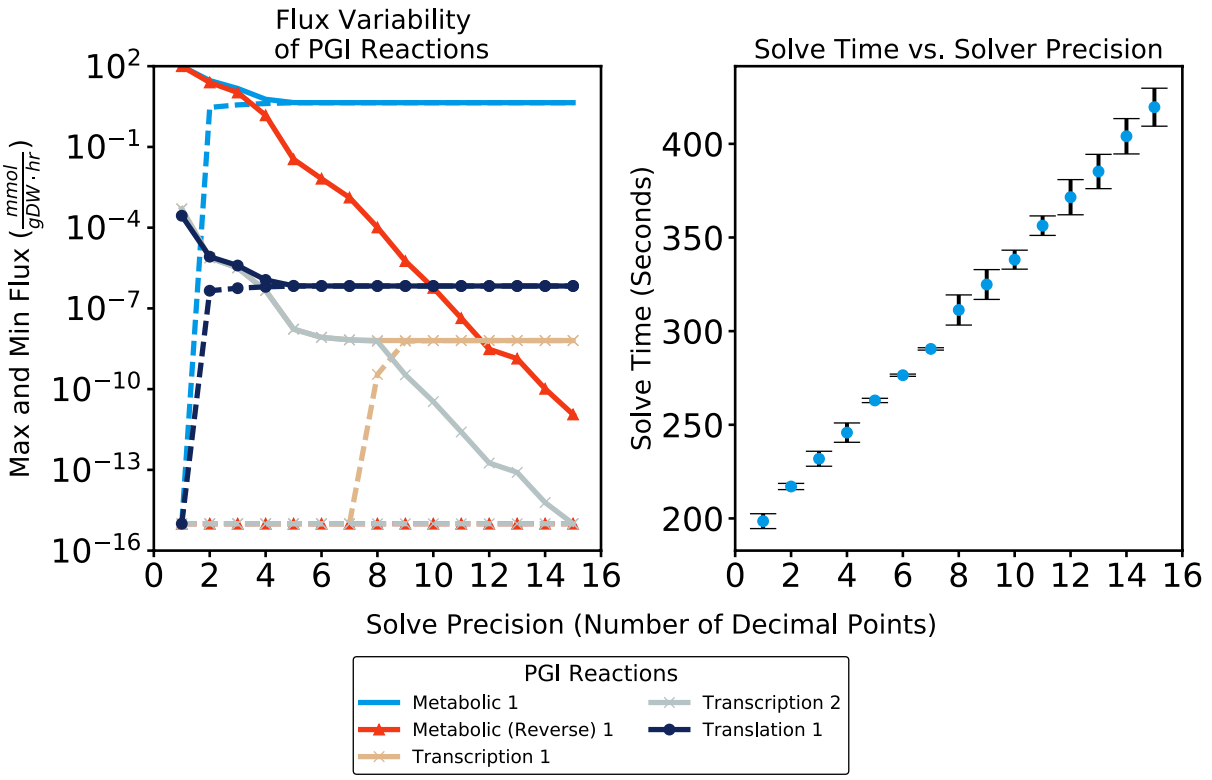

Figure 2.4: Flux variability analysis of reactions representing the expression of the Pgi enzyme and the PGI metabolic reaction. The variability becomes negligible (the max and min possible fluxes converge) for metabolic and translation fluxes when using a μ precision of 10^{-5} and for transcription fluxes when using a μ precision of 10^{-5}. There are two transcription reactions for pgi to model transcription of this gene using two different sigma factors. The lower limit of reaction flux values is set to 10^{-5} mmol $\cdot$ gDW^{-1} $\cdot$ hr^{-1} as this is close to the lowest value that can be accurately represented in double-precision floating-point in Python. Note the maximum reaction flux for the reverse direction of PGI does not drop to 10^{-5} mmol $\cdot$ gDW^{-1} $\cdot$ hr^{-1} by this μ precision. However, considering the general scale of metabolic reaction fluxes (see Figure 2.5), the maximum flux effectively drops to zero for practical purposes. High μ precision can be achieved without sizeable increases in total solve time, using qMINOS. The ME-model simulations were repeated nine times for each precision and the error bars represent the standard deviation of the solve times.

data obtained from *E. coli* grown in multiple conditions [27]. The new model version gave very similar ($R^2 > .98$) fluxes when comparing model solutions on a transcription, translation, and metabolic level (Figure 2.5) suggesting that the two models are practically identical, computationally. The reformulated ME-model cannot be expected to give completely identical solutions

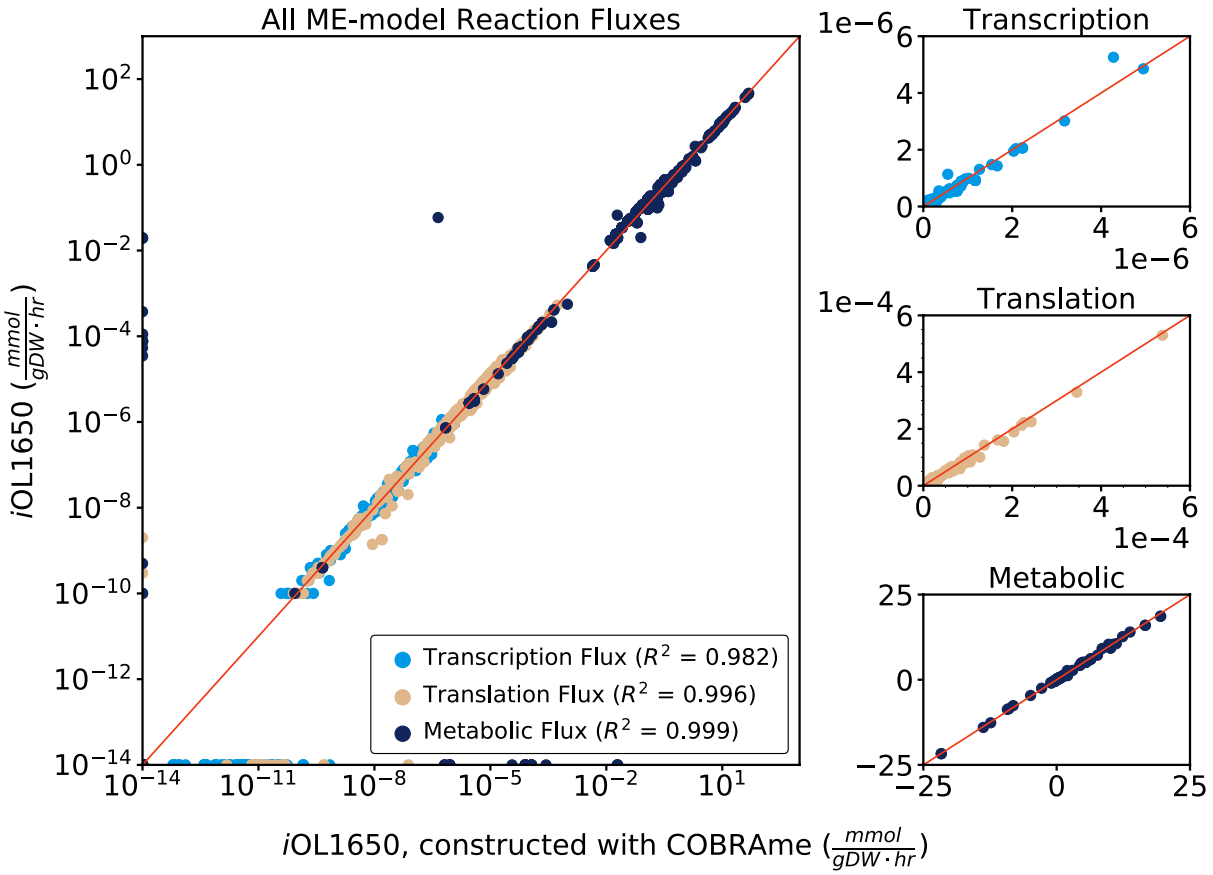

Figure 2.5: Comparison of the simulated fluxes of iOL1650-ME to the COBRAme generated version of the same model at transcription, translation, and metabolic flux scales. All fluxes are shown in pairwise comparison on the left using a log scale axis and separated into the major flux scales to be shown on a linear axis on the right. In order for fluxes of 0 $\text{mmol} \cdot \text{gDW}^{-1} \cdot \text{hr}^{-1}$ to appear, 0 fluxes have been replaced with 10^{-14} on the left plot. At each level, the models provided comparable flux predictions ($R^2 > 0.98$). The models cannot be expected to give completely identical flux predictions due to the ME-model updates outlined in Nonequivalent Changes. Since iJL1678b-ME does not contain membrane surface area constraints, iOL1650-ME was used for comparison.

to iOL1650-ME due to some of the nonequivalent changes and model corrections described in Nonequivalent Changes. Particularly, the RNA degradosome and RNA excision machinery was slightly under expressed due to the change in stable RNA excision handling described in the Supplemental Text (S8 File).

Computational essentiality predictions for both iJL1678b-ME and iOL1650-ME were com-

Table 2.3: Essentiality predictions between iJL1678b-ME and iOL1650-ME.

	Experimentally essential	**Experimentally nonessential**
iJL1678b-ME essential	1070 (69.5%)	109 (7.1%)
iJL1678b-ME non essential	84 (5.5%)	276 (17.9%)
iOL1650-ME essential	1092 (71.0%)	87 (5.7%)
iOL1650-ME nonessential	119 (7.7%)	241 (15.4%)

pared against a genome-wide essentiality screen of single gene knockouts grown in glucose M9 minimal media [28]. Due to the corrections described above and in the Supplemental Text (S8 File), iJL1678b-ME displayed improved gene essentiality predictions when comparing essentiality for the 1539 proteins also modeled in iOL1650. The bulk of these improvements stem from modeling the expression of enzyme carriers as mentioned in Nonequivalent Changes. This correction led to a 35 gene decrease in the number of false positive predictions made by iJL1678b-ME, but also led to a 22 gene increase in true positives. Overall, the accuracy of the model improved from 86.6% to 87.5%. Further, the Matthews Correlation Coefficient [29], a machine learning metric to gauge the performance of binary classifiers, saw an increase of 7% from 0.616 to 0.659 (Table 2.3).

Predictions of essentiality are from a genome wide screen of Keio collection [30] knockouts grown on glucose M9 minimal media [28].

Beyond performance and predictive capabilities, the reformulations and reduced size make iJL1678b-ME more understandable to the user. By lumping cellular processes into individual model reactions, the structure of the ME-model reactions is able to more closely resemble the central dogma of biology. For instance, the translation of a given gene, ¡gene_id¿, occurs in a single model reaction, translation_¡gene_id¿, where all components and coupling constraints are applied in one place (Figure 2.3) as opposed to occurring in multiple, separate reactions. In addition to being more easily understandable by the user, the reformulation makes the model more amenable to visualization tools like Escher [19], further easing the process of interpreting

simulation results.

2.4 Software availability and future directions

Both the COBRAme and ECOLIme software packages are required to construct iJL1678b-ME and are currently available on the Systems Biology Research Groups Github page (https://github.com/SBRG). Installation procedures as well as all necessary documentation required to build, simulate, and edit ME-models are present in the repository READMEs. The qMINOS solver [24] is also freely available for academic use. Instructions for installing and using the solver can be found as part of the solveme package [25]. Alternatively, the SoPlex solver can be found at (http://soplex.zib.de/) and is freely available to all academic institutions. The soplex_cython package contains instructions to compile the soplex solver with 80-bit precision capabilities along with the necessary code required to solve iJL1678b-ME with SoPlex. Builds of COBRAme, ECOLIme, the qMINOS solver, and all dependencies can be further obtained from Docker Hub (https://hub.docker.com/r/sbrg/cobrame/). The scripts and instructions for locally building Docker images that include the above software as well as SoPlex can be found on the COBRAme GitHub repository. This allows researchers to easily install and use ME-models regardless of platform and enables cloud computing platforms for ME-model simulations. These software packages will be actively maintained and improved. The COBRAme documentation can be found on readthedocs (https://cobrame.readthedocs.io/). The scripts, data, and instructions needed to reproduce the presented results can be found in the S3 File.

Enable New ME-model Reconstructions

We anticipate that the presented software tools will facilitate the reconstruction of many new ME-models beyond iJL1678b-ME for Escherichia coli K-12 MG1655. While the COBRAme code was constructed to be readily applicable to many different organisms, it is likely that some organisms will require additional features for their ME-model reconstruction that we did not originally anticipate. It is our priority to continue to update and improve the code to enhance its utility to model new, diverse organisms. Future efforts will be also be made to create standards to govern how ME-models are reconstructed, structured, and shared within the scientific community. This will include working with the systems biology community to develop SBML [31, 32] standards capable of encoding the information required to reproducibly build and simulate ME-models.

Acknowledgements

The following individuals contributed to this work: C.J. Lloyd, A. Ebrahim,and E. OBrien conceived the study; C.J. Lloyd, A. Ebrahim, L. Yang, Z. King, and E. OBrien developed the methodology; C.J. Lloyd, A. Ebrahim, L. Yang, Z. King, E. Catoiu, and J. Liu wrote the software; C.J. Lloyd, A. Ebrahim, and B.O. Palsson wrote the manuscript; B.O. Palsson supervised the work; all authors edited and revised the manuscript. The authors would like to thank Joshua Lerman, Aarash Bordbar, Justin Tan and Bin Du for informative discussions.

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. Funding for this work was provided by the Novo Nordisk Foundation through the Center for Biosustainability at the Technical University of Denmark

[NNF10CC1016517] and the National Institute of General Medical Science of the National Institute of Health (award U01GM102098). CJL was supported by the National Science Foundation Graduate Research Fellowship under Grant no. DGE-1144086. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Chapter 2 in part is a reprint of material published in: **CJ Lloyd***, A Ebrahim*, L Yang, ZA King, E Catoiu, EJ OBrien, JK Liu, and BO Palsson. 2018. "COBRAme: A computational framework for genome-scale models of metabolism and gene expression." *PLoS Computational Biology* 14(7): e1006302. The dissertation author was one of the primary authors.

2.5 References

1. Bordbar, A., Aarash, B., Monk, J. M., King, Z. A. & Palsson, B. O. Constraint-based models predict metabolic and associated cellular functions. *Nat. Rev. Genet.* **15,** 107–120 (2014).

2. O'Brien, E. J., Monk, J. M. & Palsson, B. O. Using Genome-scale Models to Predict Biological Capabilities. *Cell* **161,** 971–987 (May 2015).

3. Lewis, N. E., Nagarajan, H. & Palsson, B. O. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. en. *Nat. Rev. Microbiol.* **10,** 291–305 (Apr. 2012).

4. McCloskey, D., Palsson, B. Ø. & Feist, A. M. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol. Syst. Biol.* **9,** 661 (2013).

5. Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B., Assad-Garcia, N., Glass, J. I. & Covert, M. W. A Whole-Cell Computational Model Predicts Phenotype from Genotype. *Cell* **150,** 389–401 (July 2012).

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

7. Liu, J. K., O Brien, 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 (Sept. 2014).

8. Lerman, J. A., Hyduke, D. R., Latif, H., Portnoy, V. A., Lewis, N. E., Orth, J. D., Schrimpe-Rutledge, A. C., Smith, R. D., Adkins, J. N., Zengler, K. & Palsson, B. O. In silico method

for modelling metabolism and gene product expression at genome scale. *Nat. Commun.* **3,** 929 (July 2012).

9. Thiele, I., Fleming, R. M. T., Que, R., Bordbar, A., Diep, D. & Palsson, B. O. Multiscale Modeling of Metabolism and Macromolecular Synthesis in E. coli and Its Application to the Evolution of Codon Usage. *PLoS One* **7,** e45635 (Sept. 2012).

10. O'Brien, E. J. & Palsson, B. O. Computing the functional proteome: recent progress and future prospects for genome-scale models. en. *Curr. Opin. Biotechnol.* **34,** 125–134 (Aug. 2015).

11. LaCroix, R. A., Sandberg, T. E., O'Brien, 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. Environ. Microbiol.* **81,** 17–30 (Jan. 2015).

12. Chen, K., Gao, Y., Mih, N., O'Brien, E. J., Yang, L. & Palsson, B. O. Thermosensitivity of growth is determined by chaperone-mediated proteome reallocation. *Proceedings of the National Academy of Sciences* (Oct. 2017).

13. Thiele, I., Jamshidi, N., Fleming, R. M. T. & Palsson, B. Ø. Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS Comput. Biol.* **5,** e1000312 (Mar. 2009).

14. Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M. & Palsson, B. Ø. A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011. *Mol. Syst. Biol.* **7,** 535 (Oct. 2011).

15. Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., Bordbar, A., Lewis, N. E., Rahmanian, S., Kang, J., Hyduke, D. R. & Palsson, B. Ø. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protoc.* **6,** 1290–1307 (Sept. 2011).

16. Gelius-Dietrich, G., Desouki, A. A., Fritzemeier, C. J. & Lercher, M. J. sybil – Efficient constraint-based modelling in R. *BMC Syst. Biol.* **7,** 1–8 (2013).

17. Agren, R., Liu, L., Shoaie, S., Vongsangnak, W., Nookaew, I. & Nielsen, J. The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for Penicillium chrysogenum. *PLoS Comput. Biol.* **9,** e1002980 (Mar. 2013).

18. Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst. Biol.* **7,** 74 (Aug. 2013).

19. King, Z. A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N. E. & Palsson, B. O. Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. *PLoS Comput. Biol.* **11,** e1004321 (Aug. 2015).

20. Joyner, D., Čertík, O., Meurer, A. & Granger, B. E. Open source computer algebra systems: SymPy. *ACM Commun. Comput. Algebra* **45,** 225–234 (Jan. 2012).

21. Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. & de Hoon, M. J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. en. *Bioinformatics* **25,** 1422–1423 (June 2009).

22. Wunderling, R. *SOPLEX: the sequential object-oriented simplex class library* 1997.

23. Yang, L., Ma, D., Ebrahim, A., Lloyd, C. J., Saunders, M. A. & Palsson, B. O. solveME: fast and reliable solution of nonlinear ME models. en. *BMC Bioinformatics* **17,** 391 (Sept. 2016).

24. Ma, D., Yang, L., Fleming, R. M. T., Thiele, I., Palsson, B. O. & Saunders, M. A. Reliable and efficient solution of genome-scale models of Metabolism and macromolecular Expression. en. *Sci. Rep.* **7,** 40863 (Jan. 2017).

25. Laurence Yang, Ding Ma, Ali Ebrahim,Colton J. Lloyd, Michael A. Saunders, Bernhard O. Palsson. *solveME: fast and reliable solution of nonlinear ME models* Under Review.

26. Ebrahim, A., Almaas, E., Bauer, E., Bordbar, A., Burgard, A. P., Chang, R. L., Dräger, A., Famili, I., Feist, A. M., Fleming, R. M., Fong, S. S., Hatzimanikatis, V., Herrgård, M. J., Holder, A., Hucka, M., Hyduke, D., Jamshidi, N., Lee, S. Y., Le Novère, N., Lerman, J. A., Lewis, N. E., Ma, D., Mahadevan, R., Maranas, C., Nagarajan, H., Navid, A., Nielsen, J., Nielsen, L. K., Nogales, J., Noronha, A., Pal, C., Palsson, B. O., Papin, J. A., Patil, K. R., Price, N. D., Reed, J. L., Saunders, M., Senger, R. S., Sonnenschein, N., Sun, Y. & Thiele, I. Do genome-scale models need exact solvers or clearer standards? *Mol. Syst. Biol.* **11,** 831 (Oct. 2015).

27. Ebrahim, A., Brunk, E., Tan, J., O'Brien, E. J., Kim, D., Szubin, R., Lerman, J. A., Lechner, A., Sastry, A., Bordbar, A., Feist, A. M. & Palsson, B. O. Multi-omic data integration enables discovery of hidden biological regularities. en. *Nat. Commun.* **7,** 13091 (Oct. 2016).

28. Monk, J. M., Lloyd, C. J., Brunk, E., Mih, N., Sastry, A., King, Z., Takeuchi, R., Nomura, W., Zhang, Z., Mori, H., Feist, A. M. & Palsson, B. O. iML1515, a knowledgebase that computes Escherichia coli traits. en. *Nat. Biotechnol.* **35,** 904–908 (Oct. 2017).

29. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* **405,** 442–451 (Oct. 1975).

30. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L. & Mori, H. Construction of Escherichia coli Kfffdfffdfffd12 infffdfffdfffdframe, singlefffdfffdfffdgene knockout mutants: the Keio collection. en. *Mol. Syst. Biol.* **2,** 2006.0008 (Jan. 2006).

31. Refizul & Dräger, A. *draeger-lab/SBML-ME: SBMLme converter* Mar. 2018.

32. Dräger, A., Rodriguez, N., Dumousseau, M., Dörr, A., Wrzodek, C., Le Novère, N., Zell, A. & Hucka, M. JSBML: a flexible Java library for working with SBML. en. *Bioinformatics* **27,** 2167–2168 (Aug. 2011).

Chapter 3

The next generation *E. coli* M-model

The *big-data-to-knowledge* grand challenge can be addressed using mechanistic bottom-up reconstructions to productively integrate diverse data types at the genome-scale. Here we: 1) Present a genome-scale reconstruction of the metabolic network in Escherichia coli K-12 MG1655, named *i*ML1515 , that accounts for 1,515 open reading frames, 2,719 metabolic reactions involving 1,183 unique metabolites, and 1,515 protein structures; 2) Validate *i*ML1515 using growth/no growth predictions from 23,620 growth conditions, obtained using a genome-wide knockout strain collection grown on 16 different carbon sources, achieving greater than 93.4% prediction accuracy; and 3) Demonstrate: (i) how inclusion of protein structures allows the formulation of new domain-based associations that enables multi-level multi-strain functional analysis of sequence variations and is demonstrated through comparative structural proteome analysis of 1,122 *E. coli* strains; (ii) how *i*ML1515 can be used to build metabolic models of fresh *E. coli* clinical

isolates that predict their metabolic capabilities; and (iii) how *i*ML1515 can be used to build metabolic models of *E. coli* strains present in the microbiome from metagenomic sequencing data. Thus, *i*ML1515 represents a Resource that enables a broad range of new computational and big data analytic studies that can systematically generate fundamental knowledge about *E. coli* as a species.

3.1 The *E. coli* metabolic model at the cutting edge of systems biology

Genome-scale network reconstructions of metabolism form a common denominator for bottom-up systems biology studies [1, 2]. A network reconstruction represents a biochemically, genetically, and genomically (BiGG) structured knowledge-base that contains detailed information about the target organism [3]. Reconstructions must be of high-coverage and of high quality [4–7] to obtain the highest accuracy of data interpretation and physiological predictions. Further, the more disparate data types that are represented in a reconstruction, the more explanatory and predictive capability it has [3, 8]. The knowledge-base that a reconstruction represents provides a way to meet the so-called big-data-to-knowledge grand challenge that now faces the life sciences [9, 10] (Figure 3.1A).

Even though *E. coli* K-12 MG1655 is perhaps the best characterized organism in molecular and genetic terms, new functions and capabilities that it possesses continue to be discovered [11–13]. As the quote above suggests, *E. coli* is a widely-used model organism and a reference for biological research, and that is certainly true for its role in the development of the bottom-up approach to systems biology. Newly discovered biochemical functions and metabolic capabilities

demand that the *E. coli* metabolic reconstruction be updated to account for them.

We present an updated and expanded version of the *E. coli* metabolic network reconstruction. This new version, named *i*ML1515 , includes: 1) newly characterized genes and reactions, the majority of which have been discovered since 2011; 2) new structural information about the proteins in the reconstruction including a link to domains within the structure; 3) metabolism of reactive oxygen species (ROS); 4) metabolite repair pathways; 5) evaluation of growth maintenance coefficients; 6) validation against new extensive data sets presented here; 7) customization for use under commonly used growth conditions; and 8) detailed comparison to the metabolic gene portfolio of 1122 sequenced *E. coli* strains.

Previously, protein structures have been integrated into a genome-scale metabolic model of *E. coli* to compute growth rate as a function of temperature [14]. Here, we have followed a curated, standardized procedure [15] to expand the number of high-quality enzyme structures represented in *i*ML1515 . Structural systems biology is a field of growing importance that needs new tools and resources. *i*ML1515 can be used as a tool to explore the 3-dimensional structural diversity of the proteome across different strains of *E. coli* by using just its genome sequence. On a structural (molecular) level, *i*ML1515 provides insight into molecular properties through atomistic representations of proteins and their ligands, whereas, on a systems level, it constitutes a powerful tool for the characterization of complex biological systems. The representation of protein structures also allows for the development of a finer grained view of the structural proteome and the creation of a new domain-Gene-Protein-Reaction (dGPR) relationship. This addition allows for detailed strain-to-strain comparative analysis using *i*ML1515 as a tool to study *E. coli* as species at multiple biological scales.

*i*ML1515 is a Computational Resource that allows a range of new questions to be ad-

dressed computationally, including in-depth strain comparisons, structural biology analyses, systematic studies of enzyme promiscuity, high-resolution interpretation of Tn-seq data, large-scale evaluation of sequence diversity and evolution across a species in mechanistic detail. *i*ML1515 , like its predecessors [16], will likely enable and aid numerous scientific and engineering studies by expanding the range of model coverage to include phylogenetics, structural biology, physiological properties, and new metabolic capabilities.

3.2 Results

We describe this Computational Resource in three steps: first, the description of the *i*ML1515 reconstruction and its content, second, its validation, and third, address a range of new questions that were not addressable with previous metabolic reconstructions of *E. coli* .

3.2.1 Reconstruction

Rebuilding an expanded *E. coli* metabolic reconstruction with new information

All content of previous *E. coli* metabolic reconstructions [17–19] was re-evaluated by assigning new quality metrics based on evidence for a particular functional assignment, ranging from enzymatic assays (highest confidence) to genetic perturbations to computational inferences (lowest evidence) (Figure 3.1A, Supplementary Data File 1). This process led to 54 changes in this reconstruction compared to its most recent predecessor (see supplementary material for detailed discussion these model changes). Additionally, new gene functions discovered via model driven gap-filling studies[20] were also added to the reconstruction [21–23]along with other newly discovered metabolic functions in *E. coli* including those for sulphoglycolysis [11], phosphonate metabolism, [12] and the degradation of curcumin [13] (the active ingredient in tumeric). A

particular emphasis was placed on reconstructing metabolite damage and repair pathways in *E. coli* , whose importance in all organisms is increasingly appreciated [24]. Also, reactive oxygen species (ROS) generating reactions based on a recent study [25] were updated and re-curated, leading to the addition of 64 new ROS generating reactions (Supplementary Data File 2). A version of *i*ML1515 with reactions coupled to ROS generation, called *i*ML1515 -ROS is presented in Supplementary Data File 4 and available in the BiGG database [26]. Furthermore, we reconstructed known connections between a gene to its known transcriptional regulators [27–30] in the form a promoter barcode for each gene (Figure 3.1B, Supplementary Data File 1). This barcode indicates whether a metabolic gene is known to be regulated by a given transcription factor and the type of regulation (activator, repressor, or unknown).

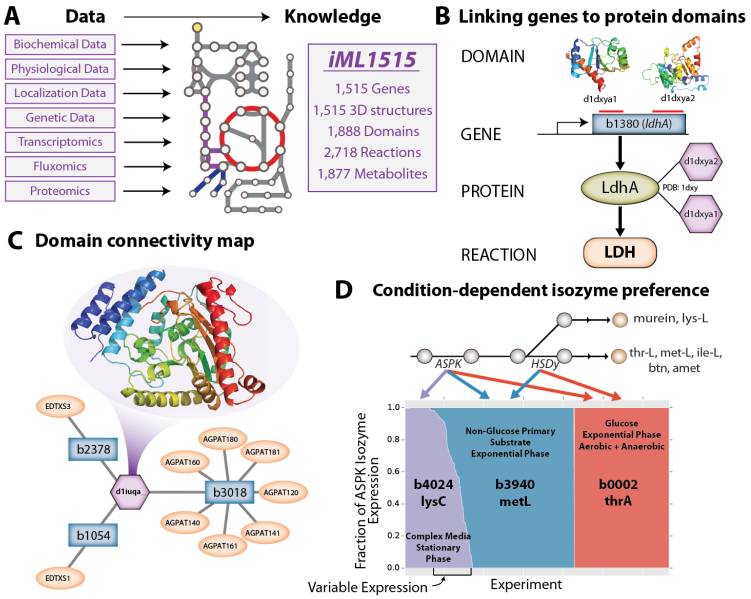

Figure 3.1: The properties and content of the *i*ML1515 knowledge base. **A)** The *i*ML1515 genome-scale reconstruction covers 1515 open reading frames encoding enzymes that catalyze 2719 reactions involving 1183 unique metabolites. It also includes 1515 protein structures. All reconstruction content is linked to external databases, including KEGG, PDB, and CHEBI. *i*ML1515 is capable of performing flux-balance analysis to integrate and interpret a variety of emerging data types including linking mutations identified from re-sequencing and/or transcriptomics data to fluxomics [31] . **B)** All reactions are directly linked to their catalyzing protein and encoding gene(s). For the first time, a network reconstruction update is released with connections to PDB structures and homology models, thus forming a domain-gene-protein-reaction relationship, or dGPR. **C)** A clustering of domain architecture and metabolite usage provides new tools to explore promiscuity and underground metabolism [21, 32]. The domain connectivity can be explored across the entire reconstruction to identify shared domains responsible for catalyzing reactions. The domain-connectivity network can be laid out using Cytoscape [33] and is available as a network file (Supplementary Data File 6) for interactive investigation. The acetyltransferase domain within panel C highlights a specific example of domain connectivity. The acyltransferase domain (d1iuqa_) is present in three genes (b3018, b1054, and b2378). The encoded proteins catalyze different but related reactions in glycerophospholipid metabolism and endotoxin synthesis. All reactions are ACP dependent acyltransferases. **D)** A database consisting of 334 normalized transcriptomics datasets [34] was contextualized using the GPRs of *i*ML1515 . Relative expression for all three genes catalyzing the Aspartate Kinase (ASPK) reaction are plotted across all experimental conditions, thus displaying condition-specific preferences for a genes usage. Listed are experimental conditions that favor particular isozymes use. At the top of the panel we depict the two reactions (ASPK, HSDy) and the two isozymes can catalyze these two reactions (thrA, metL). The third isozyme (lysC) can only catalyze APSK. ASPK and HSDy activity must be present to synthesize L-threonine, L-methionine, L-isoleucine, biotin, and S-adenosyl-L-methionine. Only ASPK activity must be present to synthesize murein derivatives and L-lysine (further discussion can be found in supplementary text).

Taken together, 184 new genes and 196 new reactions were added to the previous reconstruction [19] to form *i*ML1515 , and confidence scores were re-evaluated for all other genes (see Supplementary Text). In addition to new genes, reactions, and metabolites, the growth and non-growth associated ATP maintenance values were recalculated using measurements from strains of *E. coli* evolved to grow on different substrates and conditions (Supplementary Figures 2). The model content is presented in Supplementary Table 1 and Supplementary Figures 1 and 2.

Bridging systems and structural molecular biology

The scope of *i*ML1515 was extended to a new orthogonal data type by linking all metabolic gene products to their three-dimensional (3D) structural representations in the Protein Data Bank (PDB) (Supplementary Data File 3). A total of 716 of the 1515 proteins had crystallized structures available; the remaining 799 proteins required homology modeling [35]. Of the 184 new genes added to the model, 97% had available crystal structures. Using these 1515 protein structures, we identified the catalytic domains within them to form a complete, spatially resolved connection from encoding gene to protein product to catalyzing domain to enzymatic transformation (Figure 3.1C). This additional data type allows the use of *i*ML1515 for a new range of structural systems biology studies (see below).

Connecting each enzyme to its 3D structure allows for a fine-grained characterization of the classical gene-protein-reaction (GPR) relationship [17]. The GPR provides an explicit and formal connection between the genotype and the phenotype in a genome-scale reconstruction; it links the gene coding (G) to the protein (P) that catalyzes a reaction (R) in the network. With the inclusion of 3D structures of proteins, we obtain a detailed insight into the catalytic process

by identifying the specific domains involved in enzymatic transformations. This new data-type allows us to characterize genes by comparing and contrasting how structural motifs are linked to their phenotypic properties and enables a new relationship to be formed, termed the dGPR, or domain-gene-protein-reaction (Figure 3.1C).

The connection of 3D crystal structures to proteins in the network enables accounting of all conserved domains in proteins of the network and an analysis of how many genes share characteristic domains. We identified 1,888 unique domains within the structural proteome of *i*ML1515 . On the domain level, the maximum number of occurrences of any given domain was 17, but most often a given domain was found in 1 or 2 PDB structures. On the protein level, the maximum number of domains in a given PDB structure was 4 (e.g., b2444 (PDB: 2IPO), Aspartate carbamoyltransferase), but most often a given structure contained 1-4 domains (Supplementary Data File 3). We used this data to examine the domain-connectivity of the network, with a focus on the types of domains that were linked to each other (Figure 3.1D). This expansion in the scope of the reconstruction will allow the mapping of sequence variation to structure and provide additional information about many organism properties, such as enzyme promiscuity and underground metabolism, [32] and will enable a deeper understanding of the relationship between the structural proteome and the reactome.

Thus, *i*ML1515 represents more than an updated and re-curated metabolic reconstruction. It contains new dimensions in reconstruction coverage, content, and capabilities. *i*ML1515 can be converted into a genome-scale model (GEM)[36], and the knowledge-base it represents can be computationally characterized, such as through the use of constraint-based [37, 38] protein structure,[39] and genetic variation [40, 41] methods.

3.2.2 Validation: Computing the outcomes of high-throughput growth screens

Genome-scale models can be used to computationally predict the effect of genetic perturbations on a genome-scale [42]. They have proven to be particularly useful for predicting the condition-dependent growth effect of gene knockouts39. Furthermore, the inconsistencies between prediction and experimentation can lead to discovery [20, 21, 42].

To validate *i*ML1515 we performed experimental genome-wide gene-knockout screens for the entire KEIO collection [43, 44] grown on 16 different carbon sources that spanned different substrate entry points into central carbon metabolism (e.g., via glycolysis, TCA cycle intermediates, etc.). The screens were conducted in triplicate using scanning and image processing techniques that allow measurements beyond binary growth calls [45] (Figure 3.2A). This approach facilitated the collection of growth profiles and subsequent evaluation of lag-time, maximum growth rate, and yield for 3,892 gene knockouts across 16 conditions (total 61,904 data points, Supplementary Data File 7) (Figure 3.2C). This dataset was used to evaluate gene essentiality of the 1515 ORFs in the reconstruction. Of the 345 identified essential genes in at least one of the 16 conditions, 188 were universal to all conditions, while 157 were specific to different sets of nutrient sources. *i*ML1515 could predict gene essentiality across these 16 conditions with accuracy of 93.4%. The previous version of this reconstruction had an accuracy of 89.9%, thus this represents an increase in predictive accuracy of 3.5% compared to GEMs based on previous versions of *E. coli* s metabolic reconstruction [18, 19].

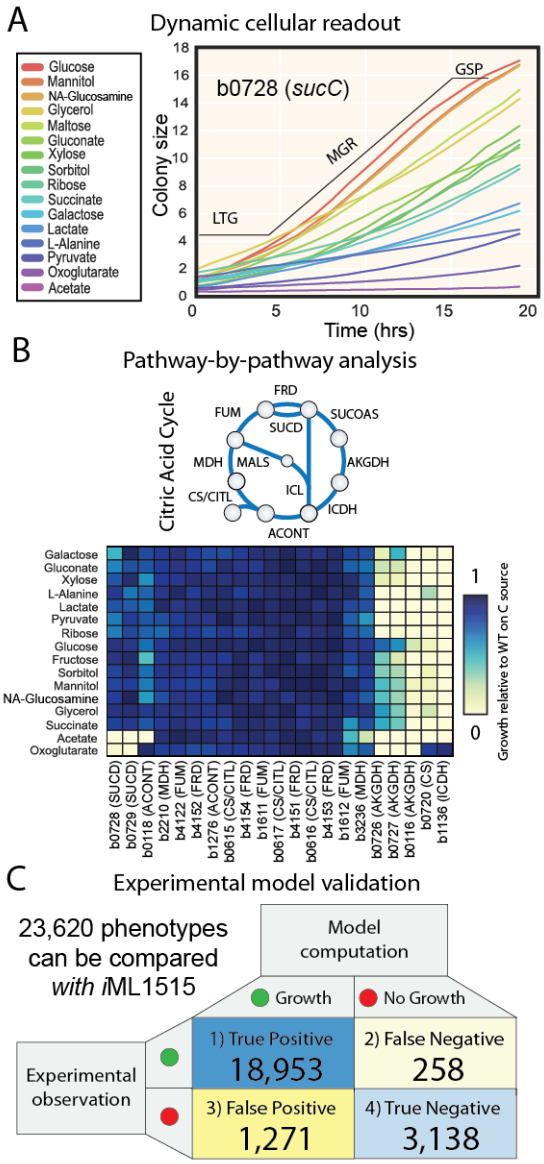

Figure 3.2: Model validation with high-throughput growth screens. A) The colonyLive platform [45] was used to perform experimental growth screens in triplicate to measure growth capabilities of 3,869 single-knockout mutant *E. coli* strains on minimal medias with 16 different carbon sources forming a total of 62,272 measured phenotypes. ColonyLive provides specific values for lag-time (LTG), maximum growth rate (MGR), and saturation point (GSP) for each gene knock-out and condition. B) Subset of knockout data highlighting growth rates for gene knockouts in the TCA cycle. C) The 1,515 genes with metabolic functions accounted for in *i*ML1515 can be directly compared to model predictions. The model is 93.5% accurate in predicting the effect of gene knockouts, an increase in accuracy of 3.6% over the previous version of the *E. coli* GEMs accuracy of 89.9% [46].

3.2.3 New questions addressable with *i*ML1515

Using gene expression data to analyze isozyme usage and underground metabolism

The *i*ML1515 reconstruction was utilized to contextualize a database of normalized transcriptomics data from 334 unique experiments [34] to address specific questions. The database contains transcriptomic data from growth in different phases of growth, on various nutrient sources, under experimental perturbations (i.e., nutrient shifts, pH shock, etc.), and with varying oxygenation conditions. Thus, the *i*ML1515 resource can be used to analyze large datasets allowing for both broad and specific questions to be addressed through computational analyses.

Are highly promiscuous enzymes more differentially expressed than specialist enzymes? We used *i*ML1515 to identify promiscuous enzymes (based on degrees of promiscuity or genes in the reconstruction that act on more than one substrate) to answer this question. Gene expression variability across the dataset showed that there is no major difference in the condition-specific differential expression for a given gene based solely on degree of promiscuity (Supplementary Figure 3.4). To address this topic further, all measured, normalized values from the gene expression database were mapped to dGPRs and overlaid to visualize each genes level of transcriptional variation across all conditions in the data set (Figure 3.1D, Supplementary Data File 8). Using the reconstruction to identify reactions catalyzed by isozymes, it was found that 383 genes in *i*ML1515 had expression patterns that were dependent on growth phase, carbon source, medium, or the experimental perturbation. Aspartate kinase (APSK), for instance, has three isozymes in *i*ML1515 (lysC, metL, and thrA) that can each be dominantly expressed depending on the culture conditions (Figure 3.1D). LysC is preferred under nutrient rich media conditions or during the stationary growth phase, metL gene is dominantly expressed in

culture conditions where glucose is not the primary carbon source, and thrA is preferentially expressed in anaerobic and aerobic glucose M9 minimal media conditions (see Supplement for further discussion).

Can proteomics data be used to further improve model predictions? Network reconstructions represent the totality of an organism's metabolic capabilities; therefore, as genome-scale network reconstructions grow in scope and scale, the number of false positive predictions may increase due to a mismatch between the regulation of gene expression and computation, which assumes that all reactions can be used under any condition. To address such false positive predictions, condition-specific models can be formed using transcriptomics or proteomics data to remove reactions catalyzed by gene products that are not active in a particular condition. We used this approach with proteomics data for *E. coli* K-12 MG1655 grown on 8 carbon sources [47]. The data was used to remove reactions and alter GPRs associated with non-expressed genes under the given condition (Supplementary Data File 9). Models tailored using this approach have on average a 13.0% decrease in false positive predictions and a 2.1% increase in essentiality predictions (MCC score). Thus, condition-specific models are suitable for designing and interpreting experiments in conditions of aerobic growth on minimal media using the corresponding carbon source for relatively short experimental observation windows. These cases include those in which secondary isozymes will not be upregulated to physiologically active levels.

Defining a core metabolic network: conservation of metabolic capabilities amongst *E. coli* strains

*i*ML1515 is specific for the *E. coli* strain K-12 MG1655. This was the first *E. coli* strain whose genome was sequenced [48]. Since then, many more genome sequences of *E. coli* strains

have become available. Most sequenced *E. coli* strains have 15-20% larger genomes than the MG1655 strain [49, 50].

We compared the metabolic genes in *i*ML1515 across 1122 sequenced strains of *E. coli* and Shigella (Figure 3.3, Supplementary Data File 5). We found that 978 metabolic genes were shared among >99% of the strains. *i*ML1515 was stripped of those genes not present in greater than 1111 strains (99% of strains) to form a model representing conserved or core *E. coli* metabolic capabilities, named iML978. This reconstruction contained 978 genes, 1,864 reactions, and 1,169 unique metabolites, making it similar in size to a previous conserved metabolic reconstruction formed from 55 strains of *E. coli* [51].

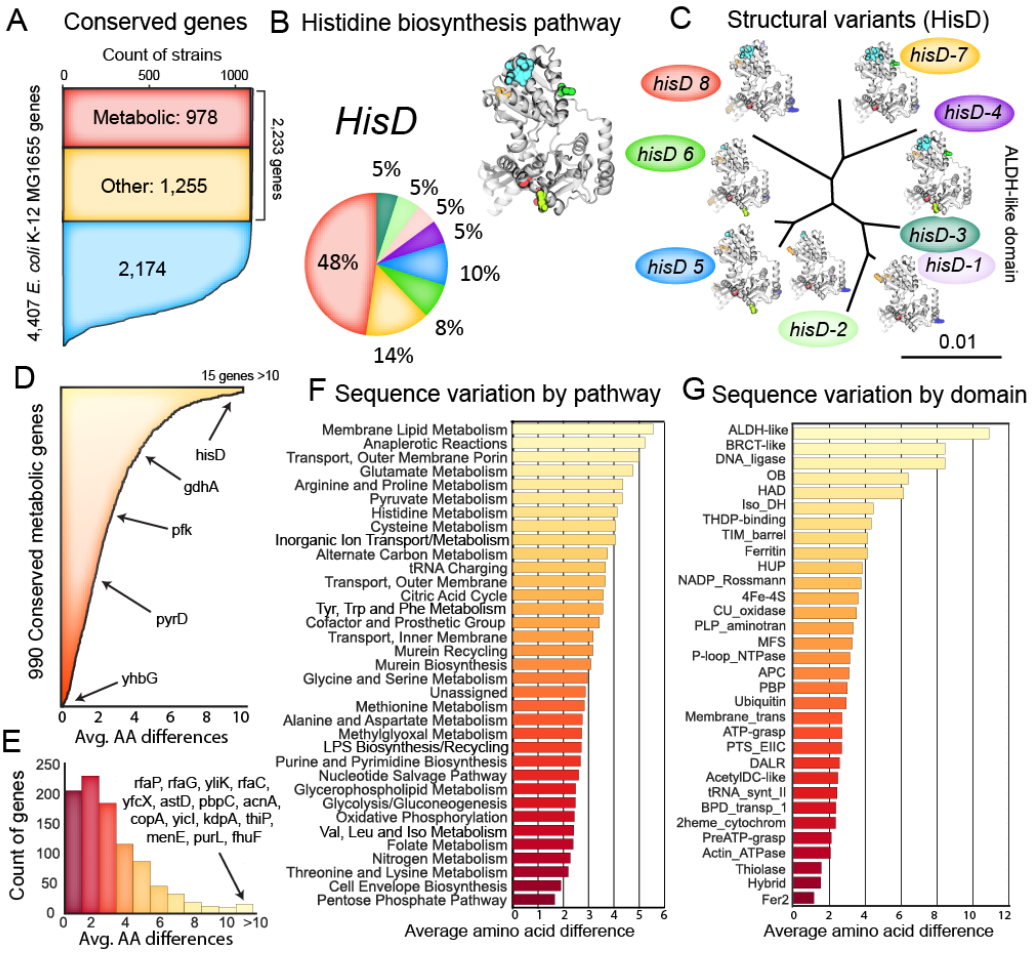

Figure 3.3: Overview of the *i*ML1515 reconstruction and its comparison to sequence variations across 1122 strains of *E. coli* . A) The count of each *E. coli* K-12 MG1655 genes presence across 1,122 sequenced strains of *E. coli* . B) The histidine pathways showed high levels of amino acid differences among genes involved. Histidinol dehydrogenase was found to have many unique sequences across the 1104 orthologs examined. The pie chart represents the percentage of strains that contain unique hisD alleles. *E. coli* K-12 MG1655 possesses the hisD-116 allele which is only present in 19 (1.7%) of the sequenced strains. C) Unique genetic mutations can be examined using structural biology methods. For example, all of the unique alleles can be compared to the *E. coli* K-12 MG1655 gene sequence to hypothesize the effect of mutations. D) Even the 978 genes with metabolic functions that are conserved across 99% of *E. coli* strain possess unique genetic mutations. The bar chart displays the average number of amino acid mutations across these highly-conserved genes for all 1,122 strains of *E. coli* . E) The histogram shows how many genes have a given number of average mutations. F) Amino acid mutations can be compared on a pathway-basis using structural biology techniques. G) Amino acid changes can also be compared on a protein domain basis. For example, genes that encode the aldehyde dehydrogenase (ALDH-like) domain present in hisD have, on average, more genetic mutations across all 1,122 strains of *E. coli* than genes encoding other domains.

Analysis of iML978 using constraint-based modelling revealed phenotypic differences from the full *i*ML1515 *E. coli* K-12 MG1655 model. For example, iML978 is auxotrophic for nutrients including L-phenylalanine, L-tryptophan, L-arginine, L-tyrosine, L-glutamine, biotin, thiamine, and tetrahydrofolate, indicating that the ability to synthesize these molecules is not conserved across all strains of *E. coli* or alternate routes for their synthesis exist. When the *in silico* minimal media is supplemented with these nutrients, iML978 is able to compute growth on 115/187 C sources, 75/94 N sources, 6/11 S sources, and 41/50 P sources (compared to the full *i*ML1515 model, Supplementary Table 2). iML978 is provided with this manuscript for use in applications studying conserved *E. coli* metabolic functions and as a starting point for developing a metabolic reconstruction from a freshly sequenced *E. coli* strain (Supplementary Data File 4).

Using the core metabolic network to analyze *E. coli* clinical isolates

Can we build informative metabolic models solely based on genomic sequence from clinical isolates? To address this question, we analyzed 552 sequenced clinical isolates from two recent studies of pathogenic *E. coli* [52, 53]. Strain-specific genome-scale models (GEMs) of the clinical isolates were constructed using *i*ML1515 (Figure 3.4A) by mapping metabolic capabilities of each respective clinical isolate (by using the sequence from K-12 MG1655 to search for orthologs in each respective strain). The average strain-specific GEM was based on 1404 30 genes (Figure 3.4C). We used the core model, iML978 to evaluate the assembly quality of each genome (any strains missing a core gene were excluded from the following analyses).

We predicted growth capabilities on all growth supporting carbon, nitrogen, phosphorous, and sulfur sources for each of the 552 clinical isolate models (Figure 3.4C). The predicted growth capabilities were sufficient to distinguish strains of extra-intestinal *E. coli* (ExPEC) from

those of intestinal strains (InPEC). We also compared the conservation of the 176 new metabolic reactions in *i*ML1515 across the 552 clinical isolates (Figure 3.4B). The new reactions that were not conserved across all strains included the curcumin catabolism pathway (NADPH-dependent curcumin reductase, present in 232/552 GEMs) as well as the pathway for degradation of sulpho-quinovose (6-deoxy-6-sulfofructose-1-phosphate aldolase, present in 501/552 GEMs), showing that not all clinical isolates of *E. coli* may be capable of utilizing these nutrients. Thus, the *i*ML1515 resource enables identification of important metabolic differences in clinical isolates. If such differences are found to link to different treatments and clinical outcomes, future diagnosis may be possible from sequence alone.

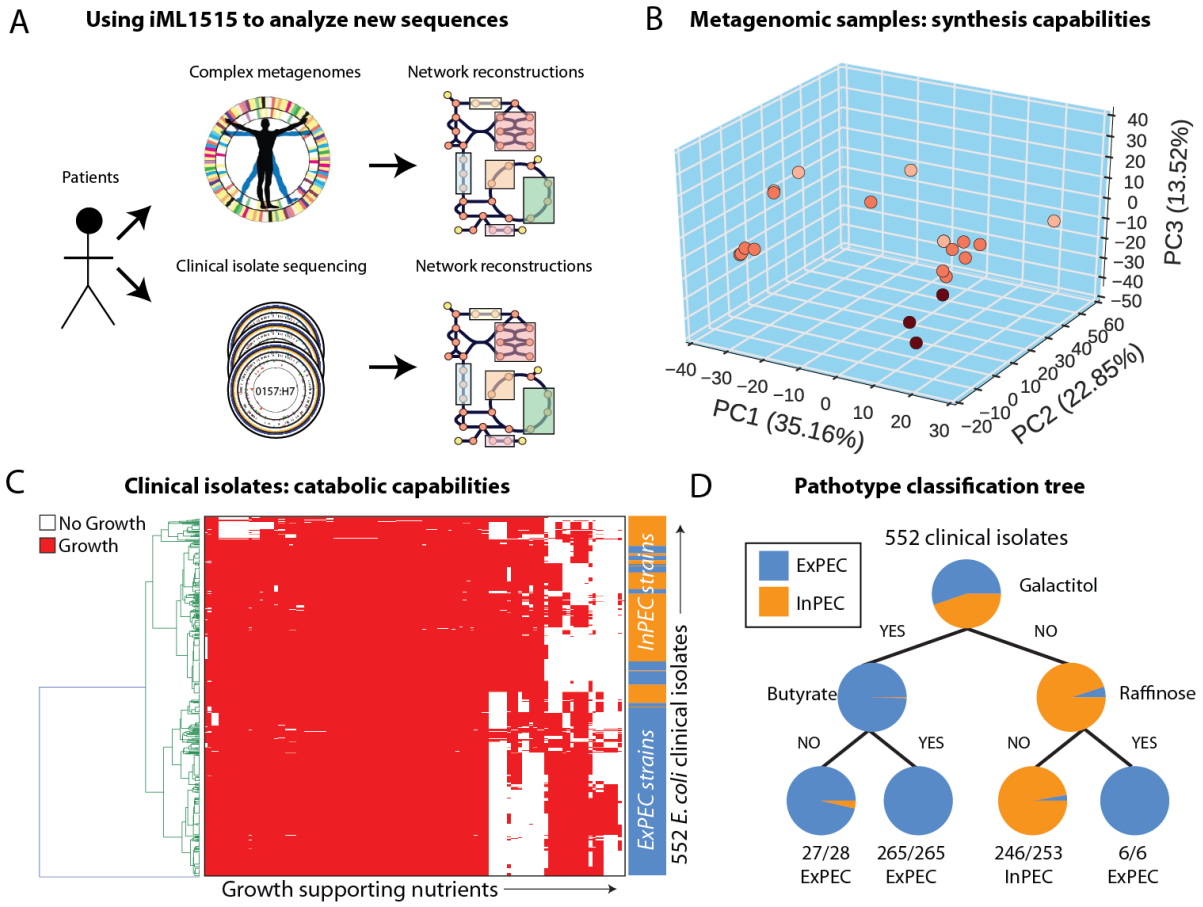

Figure 3.4: Using *i*ML1515 to investigate freshly-sequencing clinical isolates and metagenomic samples. A) Workflow for construction of strain-specific models from sequenced clinical isolates and complex metagenomics samples. The *i*ML1515 resource can be used to rapidly construct strain-specific models of metabolism from sequenced clinical isolates and complex metagenomics samples. Genes that are part of the *i*ML1515 model are identified and extracted for comparison across each of the metagenomics samples. B) Model-predicted metabolite synthesis capabilities from metagenomics samples. Sample-specific models of *E. coli* metabolism were constructed for 22 metagenomic samples by evaluating shared content from *i*ML1515 . Metabolite synthesis capabilities and yields were calculated for each model and evaluated using PCA to illustrate a separation in sample-specific metabolite synthesis capabilities. Points are colored based on model-predicted max autoinducer-2 yield. C). Heatmap of model-predicted catabolic capabilities for clinical isolates. Strain-specific models were constructed for 552 *E. coli* clinical isolates from two recent studies [52, 53]. Models were used to predict the ability to grow on over 300 different carbon, nitrogen, phosphorous, and sulfur sources. D) Machine learning techniques, such as a decision tree, can be applied to model predictions. For example, model-predicted catabolic capabilities can be used to classify clinical isolates between extra-intestinal pathotypes (ExPEC: isolated from blood or urine) and from intestinal pathotypes (isolated from feces) based solely on the model-predicted ability to catabolize three substrates (galactitol, butyrate, and raffinose).

Building models of *E. coli* strains in metagenome samples

With continued decreases in DNA sequencing costs, is strain-level resolution and modelling from metagenomics samples feasible?[54] The metagenomics field anticipates that in 2017 full Illumina read coverage of the DNA found in stool samples will be standard. To answer the above question, we deployed *i*ML1515 to analyze metagenomics sequencing data and evaluate whether strain-level resolution is achievable. Using the same procedure as above, we can build draft GEMs for *E. coli* strains using metagenomic data [55]. We built such draft models for 22 microbiome samples from two recently published studies [55, 56]. On average, sequencing data from each sample allowed the identification of 131194 metabolic genes contained in *i*ML1515 .

The gut microbiome plays a significant role in breaking down and synthesizing compounds important for human health. Sample-specific metabolic models from metagenomic sequences can be used to study this process by determining the dominant metabolic capabilities of species closely related *E. coli* K-12 MG1655 contained within the sample, and therefore the patients microbiome.

To examine how metabolic capabilities translate to possible metabolite secretion in the human gut, we used the sample-specific models to predict levels of maximum metabolite synthesis (see Methods). PCA analysis of the results shows that models cluster into discrete groups based on metabolites that can be synthesized (Figure 3.4B). For example, the models separate in principal component 1 based, in part, on the maximum capability to produce autoinducer-2 which could have implications on quorum sensing for strains present in each sample.

These GEMs enable an analysis of metabolic capabilities of the *E. coli* strains present in these metagenomic samples. A total of 350 metabolic genes from *i*ML1515 were variably present across the 22 samples. Interestingly, of the 184 new genes that were added to *i*ML1515 over

its predecessor [19], 3417 of them were variably present across the 22 samples, showing that the new content in *i*ML1515 provides valuable new information for analysis of metagenomic samples. We found that the core metabolic capabilities consisted of 2,326 reactions while 356 reactions were variable across the samples (Synopsis Figure 3.1A). Reactions variably present among the strains-specific models, included those involved in catabolism of sulphoquinovose and curcumin. Both of these metabolites are components of the human diet. Thus, commensal *E. coli* strains may play a different metabolic role in different microbiomes.

Protein structure-guided discovery of mutations across different *E. coli* strains

Does the introduction of protein structural information to metabolic reconstructions allow the evaluation of how strains differ in their metabolic function? *i*ML1515 not only allows for analysis of conserved metabolic capabilities but also for comparison of sequence variation among conserved metabolic genes. The presence of specific alleles in the *E. coli* K-12 MG1655 genome was evaluated across the 1122 different strains. All genes in *i*ML1515 were compared to their corresponding gene in each of the 1122 *E. coli* strains. The number of specific and unique alleles ranged from 20 (e.g. *pfkA*) to 249 (e.g. *hisD*, see Figure 3.4). Overall we found that MG1655 possessed the dominant allele (the allele present in a majority of strains) for only 30% of the 1122 strains (Supplementary Data File 5). For example, the K-12 MG1655 allele of *rph* was found to be present in less than 1% of strains (7 close K-12 derivatives including K-12 W3110 and BW25113). This mutation has been shown to result in reduced expression of pyrE and leads to pyrimidine starvation conditions where strains grow 10-15% slower in pyrimidine free media than in media supplemented with uracil [57]. Thus, *i*ML1515 can be used to develop fundamental understanding of the differences between laboratory and true wild-type strains through in-depth

comparison of their genetic composition.

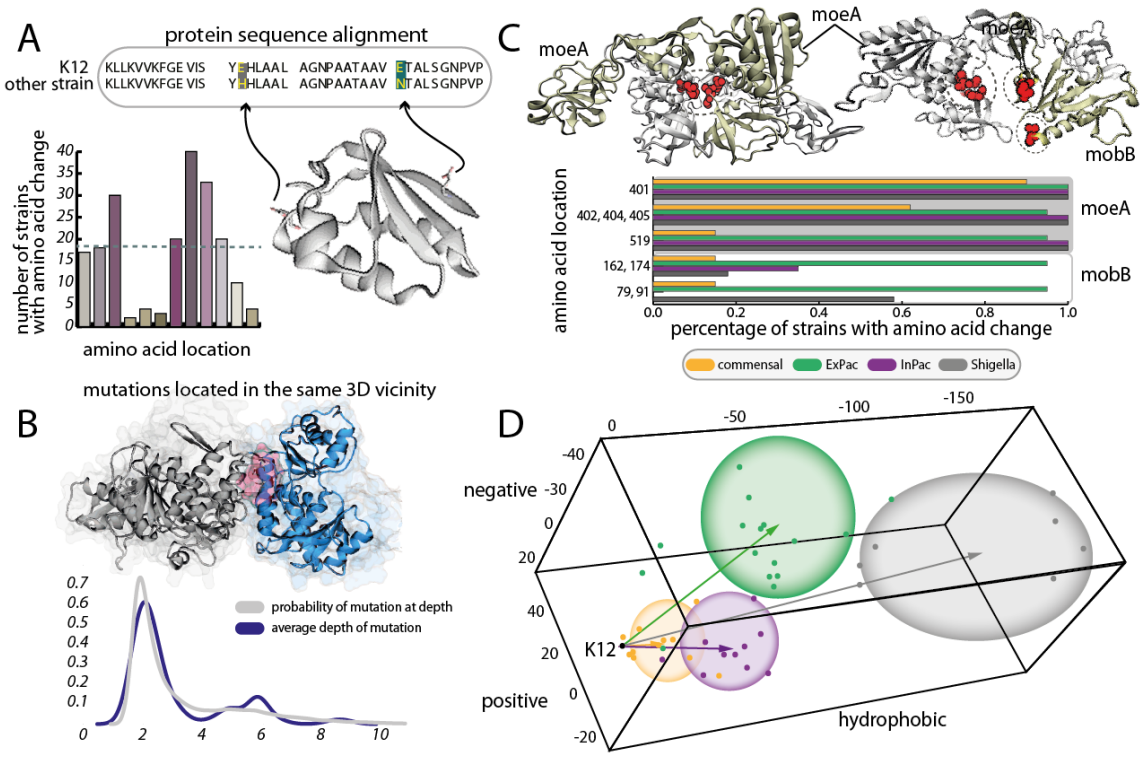

Figure 3.5: Analysis of the structural proteome across the *E. coli* species. A) We aligned all protein sequences from more than 50 different strains of *E. coli* (including MG1655, commensal, extra-intestinal and intra-intestinal pathogenic and Shigella strains). These alignments allowed for the identification of amino acid differences between the K-12 laboratory strain and the 50 strains studied here. We identified specific amino acid positions across all genes that were significantly different (more than 20 strains marked an amino acid difference with respect to the K-12 strain at a given position in a gene). B) For the set of genes that shared a high degree of dissimilarity from K-12, we further investigated how many of these amino acid changes (or mutations) cluster within a 10 vicinity of all other mutations. We found that the amino acid mutations that tend to cluster are commonly found on protein surfaces, while a select subset occur in buried regions (greater than 4-6 from the surface of the protein). C) The amino acid differences that clustered at the surface of proteins are in protein-protein interface regions. An illustrative example is that of homo and/or heterodimer complexes of genes involved in molybdenum cofactor biosynthesis. We find the co-evolution of mutations occurring at prime positions in the protein structure may influence the biological assembly of these subunits. D) A global analysis of all amino acid differences indicates that, in general, the landscape of the structural proteome variation across *E. coli* strains clusters based on distinct properties of the proteins themselves that are linked to strain differences. A 3D Principal Component Analysis (PCA) analysis shows three main axes, representing negative charge, positive charge, and hydrophobic properties of all proteins in each strain of *E. coli* .

Can analyzing the 3-dimensional structural diversity of the proteome across different strains of *E. coli* hint at their lifestyle? We compared the conservation of *E. coli* K-12 MG1655 amino acid sequence for genes shared among 55 strains of *E. coli* spanning a range of pathotypes [51] and identified the regions of the proteome that were significantly different (Figure 3.5A). Representing sequence variants in the context of their network-level and structural-level properties provides a unique extension of genome-scale reconstructions [56]. For 30% of the genes, we found that sequences were conserved in more than 20 of the 55 strains analyzed (e.g., 980 genetic alterations across 414 genes), indicating that the laboratory strain had either mutated spontaneously or evolved specific functionalities that were not necessarily needed by the other strains. We further investigated whether variants were co-located in the same 10 vicinity of other variants (Figure 3.5B) and found that the majority of cases occur on the surfaces of proteins. For these cases, we probed known protein-protein interactions and found that certain variants likely influence biological assembly, such as in the case of *moeA* and *mobB* genes where co-evolution of mutations occurs at prime positions in the protein structure and may influence their biological assembly (Figure 3.5C).

Finally, a global analysis of the proteome landscape for all 55 *E. coli* strains indicates that variation clusters by (i) physico-chemical properties and (ii) strain type. Commensal strains retain similar global proteome properties to that of K-12 MG1655 when compared to other strains (Figure 3.5D). This observation is interesting and potentially impactful, given that Shigella is known to withstand radical changes in environmental conditions (e.g., heat and pH shock).

3.3 Discussion

The *i*ML1515 knowledge base includes an unprecedented range of organism-specific information that has been organized and systematized, thus opening it to integrated computational assessment. Such assessment allows for extraction of knowledge through integrated and structured analysis of disparate data sets. Examples include the assessment of gene essentiality and physiological properties as well as novel computations of structure-derived protein properties and multi-strain sequence comparisons.

All of the metabolic genes in *i*ML1515 are linked to 3D representations of protein structure. This link expands the utility of the model to structural biologists and those interested in studying the effect of protein properties on the functions of an entire network. Furthermore, linking domains to their encoding genes (through the dGPR relationship) should help in effectively analyzing transposon-mutagenesis data where a given transposon integration may abolish a portion of a gene while leaving a domain-coding portion intact and potentially active.

*i*ML1515 allows for database interoperability, as all reconstruction content is linked to external databases, including KEGG, PDB, and CHEBI. *i*ML1515 can be converted into a genome-scale model, and thus the knowledge base it represents can be computationally interrogated and characterized. A mathematical format enables the use of multiple different computational tools. Constraint-based methods can be used (i.e., the COBRA Toolbox [58, 59]) to assess network properties. Protein structure tools can be deployed [39] to assess similarities and properties of protein structure and genetic variation can be examined using phylogenetic tools [60] to study gene evolution and transfer between organisms as well as its effect on species evolution.

In addition, we present three versions of *i*ML1515 tailored for specific use cases. These three versions can be used to augment *i*ML1515 s capabilities and simulate metabolic functions

under unique situations. 1) A version containing reactions known to produce reactive oxygen species coupled to ROS production (*i*ML1515 -ROS) allows for the simulation of ROS production and genetic manipulations that might increase such production for use in antibiotic design or potentiating activity25. 2) A version containing only conserved *E. coli* metabolism (iML978) can be used to probe the core metabolic capabilities of *E. coli* as a species, which also serves as a starting point for building metabolic network reconstructions of newly sequenced *E. coli* strains. 3) A version where genes that are unlikely to be expressed in the common growth conditions of M9-glucose minimal media has been removed and the growth objective has been modified (iML1400-glucose).

Thus, *i*ML1515 provides a Computational Resource for studying the metabolic state of *E. coli* strains accounting for protein structural features and genetic variation. For the past decade, *E. coli* metabolic reconstructions have been used in a wide range of studies16, from the discovery and characterization of new metabolic genes [21], to the design of new antibiotics [25, 61], to the construction of high-yield production strains for industrially valuable compounds [62–64]. The unique capabilities of *i*ML1515 will enable new categories of scientific pursuits and practical applications.

3.4 Methods

3.4.1 Network reconstruction procedure

The *i*ML1515 reconstruction was assembled by updating the *i*JO1366 *E. coli* metabolic reconstruction [65]. A 96-step procedure [66] for metabolic network reconstruction was followed when adding new genes, reactions, and metabolites to form *i*ML1515 . The reconstruction was

assembled using the SimPheny (Genomatica Inc., San Diego, CA) software platform. All new metabolites were checked against public databases (e.g. KEGG, PubChem) for correct structure and charge at a pH of 7.2. New reactions were mass and charge balanced and reversibility was assigned based on experimental studies, thermodynamic information, or the heuristic rules in the standard reconstruction protocol [66]. Reactions were associated with genes and functional proteins to form GPRs. The *i*ML1515 model was exported from SimPheny as an SBML file, and the COBRApy Toolbox [59] was used for additional model testing. The Gurobi linear programming solver (Gurobi Optimization, Inc., Houston, TX, USA) was used for all optimization procedures.

3.4.2 Updating the Biomass Objective Function GAM and NGAM

Growth-associated and non-growth-associated maintenances values were recalculated based on recent *E. coli* K-12 MG1655 adaptive laboratory evolution studies (ALE). For these calculations, the electron transport system NADH dehydrogenase reactions NADH16pp (nuo) and NADH5 (ndh) were constrained to carry identical fluxes by replacing these reactions with an equivalent flux split reaction. Additionally, TRN constraints were applied to turn off reactions not active under the specific experimental condition [67]. The model was adjusted to remove all GAM and NGAM. Next, it was constrained using end-point measured physiological data from the ALE studies (including growth rate, substrate uptake rate and byproduct secretion rates). Under these constraints, the model was used to predict maximum possible ATP generation by optimizing the ATPM reaction. This data was plotted (Supplementary Figure 7) and the slope and intercept of the experimental data were identified using linear regression. NGAM (the y-intercept) was determined to be 6.86 ATP $\frac{\text{mmol}}{\text{gDW}}$ and the GAM (slope) was 75.55 $\frac{\text{mmol}}{\text{gDW}\cdot\text{hr}}$

Statistics

The growth associated (GAM) and non-growth associated maintenance (NGAM) was calculated using wild type and evolved *E. coli* K-12 MG1655 strains. The substrate uptake, secretion and growth rates were presented using data from 3 technical replicates in each of the studies. The mean values for the measured substrate uptake and secretion rates are shown in Supplementary Figure 6 with the error bars depicting the standard deviation in the measured values as provided by the studies. To include a consideration of this error when calculating GAM and NGAM values, we included error bars in Supplementary Figure 7 to show the maximum range that *i*ML1515 computed ATP synthesis values could possibly vary for each *E. coli* culture. For each physiological measurement, the horizontal error bars represent the standard deviation in experimental growth rate measurements, as provided by the ALE studies. Vertical error bars represent the maximum and minimum ATP hydrolysis flux capable of being produced by *i*ML1515 for each experiment. This was done by determining the combination of measurements plus-or-minus their error that would result in the lowest and highest capacity for the model to produce ATP. For instance, for an aerobic glucose simulation, the highest ATP flux would be obtained by constraining the growth_rate = measured - error, the acetate_flux = measured - error, and glucose_uptake_rate = measured + error. In other words, the model has the highest capacity for ATP production when growing slowly while taking excess glucose and secreting minimal acetate. The lowest ATP flux would be calculated by setting these constraints to the opposite values and using the opposite logic. Each individual physiological measurement used to perform this analysis has an acceptably low amount of error (Supplementary Figure 6). This low amount of error, however, becomes compounded when running the simulations in the way described above. Further, it is possible, since the error bars represent the extremes of possible ATP production

capability for each measurement, that some of the physiological states considered are actually even infeasible in vivo.

3.4.3 Protein structure integration

Integration of protein-related information into the GEM involves four stages: (i) linking the genes to available experimental protein structures, found in publicly available databases, such as the Protein Data Bank (PDB); (ii) determining genes with and without available protein structures and performing homology modeling using the I-TASSER suite of programs when structures are not available [35]; (iii) performing QC/QA on all structures based on a set selection criteria (e.g., resolution, number of mutations, completeness); (iv) mapping GEM genes to other databases (e.g., BRENDA [68, 69], SwissProt [70], Pfam [71], SCOP [72]) for complementary protein-structural information. For the majority of gene to protein structure mapping, we used a previously generated GEM-PRO for iJO1366. For details about generation of GEM-PRO models, we directed the reader to Brunk *et al* [15]. Use of PfamScan and HMMER3 algorithms generated protein fold family annotations [73]. Open source software for protein structural predictions are available and are used in conjunction with the IPython framework.

3.4.4 *In vitro* phenotypic screens

Stock plates comprising all the KEIO SKO mutants in 384-well-format (24 columns and 16 rows) were thawed at room temperature for about 1 h before use. The liquid cells on the thawed plates were spotted onto fresh Luria-Bertani (LB) agar plates with 384-long pins. For the wild-type experiment, *E. coli* K-12 BW25113 was inoculated into 2 ml LB and grown for 20 h at 37°C with shaking. The liquid culture was spotted onto fresh LB agar plates with 384-

long pins. After overnight incubation at 37°C, the grown colonies were arrayed from 384-format to 1536-format plates (48 columns and 32 rows) with 384-short pins. These inoculations were performed using a Singer RoToR HDA machine with designated pins (Singer Instruments). After inoculation, colony growth at 37°C was monitored using the Colony-live system [45]. The Colony-live system produced three growth characteristic values Lag-Time Growth (LTG), Maximum Growth Rate (MGR) and Saturation Point of Growth (SPG). All experiments were done using a validated Keio collection for all SKO mutants [44, 74] and the wild-type *E. coli* K-12 BW25113 carrying kanamycin-resistant pXX563 (mini-F plasmid, single copy number; unpublished). All SKO mutants were stored in a total of twelve 384-well microtiter plates at -80 °C. Cells were grown in LB medium with 30 μ g/ml of kanamycin (Wako, Osaka, Japan). Agar plates were prepared by adding 1.5% agar (Mitsui Sugar, Tokyo, Japan) to M9 medium and autoclaving, and 50 ml of the medium was then poured onto a Singer PlusPlate (Singer Instruments, Somerset, United Kingdom). Before use, the agar plate was dried in a laminar flow cabinet for 10-30 min.

Statistics

Growth analysis of the Colony-live system was performed with R (http://www.r-project.org) and rpy2 package. After the image analysis, colony growth values less than 1 were set to 1, which is the lower limit quantification value. Colony growth values at all incubation times were normalized by the minimum growth value, and then the first 25 valid growth values exceeding 1 were regressed to the Gompertz growth model. All measurements were performed in triplicate. See Takeuchi *et. al.* [45] for full details on statistical analyses.

3.4.5 Constraint-based modeling

The *i*ML1515 model, constructed in SimPheny, was exported as an SBML file and used to perform simulations and constraint-based analyses using the COBRApy Toolbox and Gurobi linear programming solver. The constraint-based model consists of a stoichiometric matrix (S) with m rows and n columns, where m is the number of distinct metabolites and n is the number of reactions. Each of the n reactions has an upper and lower bound on the flux it can carry. Reversible reactions have an upper bound of 1000 $\frac{\text{mmol}}{\text{gDW}\cdot\text{hr}}$ and a lower bound of -1000 $\frac{\text{mmol}}{\text{gDW}\cdot\text{hr}}$, while irreversible reactions have a lower bound of zero. FBA can be used to identify optimal steady-state flux distributions of constraint-based models. Linear programming is used to find a solution to the equation Sv = 0 that optimizes an objective cT*v, given the set of upper and lower bound constraints. v is a vector of reaction fluxes of length n. Typically, c is a vector of 0s of length n with a 1 at the position of the reaction flux to be maximized or minimized. For a thorough description of FBA, see (Orth et al, 2010) [37]. For most growth simulations, the core biomass reaction is set as the objective to be maximized. One reaction, FHL, is not used under typical growth states and is by default constrained to carry zero flux. The NGAM constraint is imposed by setting a lower bound of 6.86 mmol gDW-1 on the reaction ATPM. The exchange reactions that allow for extracellular metabolites to pass in and out of the system are defined such that a positive flux indicates flow out. All exchange reactions have a lower bound of zero except for glucose (-10 $\frac{\text{mmol}}{\text{gDW}\cdot\text{hr}}$) and oxygen and all inorganic ions required by the biomass reaction (-1000 $\frac{\text{mmol}}{\text{gDW}\cdot\text{hr}}$).

3.4.6 *in silico* phenotypic screens

The *i*ML1515 *E. coli* K-12 MG1655 metabolic reconstruction was used to make computational gene essentiality predictions. The parent strain of the Keio Collection, BW25113, is derived from K-12 MG1655 and is missing several genes that are present in K-12 MG1655: araBAD, rhaBAD, and lacZ. Therefore, flux through the associated reactions without isozymes (ARAI, RBK_L1, RMPA, LYXI, RMI, RMK, and LACZ) was constrained by setting the upper and lower flux bounds of the reactions to zero. The lower bounds of exchange reactions were set to default values to simulate minimal media. For aerobic growth, oxygen uptake was allowed by setting the lower bound of the oxygen exchange reaction to -20 $\frac{\text{mmol}}{\text{gDW}\cdot\text{hr}}$. Anaerobic growth was modeled by setting the lower bound of this reaction to zero. The substrate uptake rate for all primary carbon sources experimentally screened (see in vitro phenotypic screens in Supplementary Methods) was set to a lower bound of -10 $\frac{\text{mmol}}{\text{gDW}\cdot\text{hr}}$. After setting the bounds for each condition, the predicted effect of the single deletion of each gene in *i*ML1515 for each condition was computed using the COBRApy Toolbox delete_model_genes function, which uses GPRs to constrain the appropriate reactions to carry zero flux and then predicts maximum growth using FBA. A gene was considered computationally essential for the simulated condition if deletion of the gene reduced optimal growth rate to less than 0.05 h-1. The newly identified essential genes were added to the lists of essential genes under each condition and the results of the essentiality prediction comparisons for iJO1366 and *i*ML1515 are represented in Supplementary Data 11.

3.4.7 Prediction of different carbon, nitrogen, phosphorus, and sulfur sources

The possible growth-supporting carbon, nitrogen, phosphorus, and sulfur sources of *E. coli* were identified using FBA. First, all exchange reactions for extracellular metabolites

containing the four elements were identified from the metabolite formulas. Every extracellular compound containing carbon was considered a potential carbon source. Next, to determine possible growth supporting carbon sources, the lower bound of the glucose exchange reaction was constrained to zero. Then the lower bound of each carbon exchange reaction was set, one at a time, to -10 $\frac{\text{mmol}}{\text{gDW}\cdot\text{hr}}$, and growth was maximized by FBA using the core biomass reaction. The target substrate was considered growth supporting if the predicted growth rate was above zero. While identifying carbon sources, the default nitrogen, phosphorus, and sulfur sources were ammonium (nh4), inorganic phosphate (pi), and inorganic sulfate (so4). Prediction of growth supporting sources for these other three elements was performed in the same manner as growth on carbon, with glucose as the default carbon source.

3.4.8 Mapping to other *E. coli* strains

The protein sequences of all available *E. coli* and Shigella strains (1122 strains total) were downloaded from the RAST database [75]. The RAST API [76] get_corresponding_genes function was used to identify orthologs between *E. coli* K-12 MG1655 and each of the other strains based on bi-directional best BLAST hits and genomic context. Genes were considered conserved if they were present in another organism at greater than 80% percentage identity with a best bidirectional hit (BBH). Those genes that were not shared at this cutoff in less than 99% of strains were then removed from the *i*ML1515 model to form *i*ML976. This model was investigated to determine which components of the biomass function could not be produced. When a strain is unable to synthesize a certain biomass component, it either has an alternate route to produce this biomass component or an auxotrophy requiring transport of this metabolite to sustain growth. This method can accurately determine known strain-specific auxotrophies [51].

To simulate growth on all possible C, N, P and S sources the model was provided with exchange reactions for those components that it was auxotrophic for (btn, glu-L, dttp, thmpp, 2ohph, trp-L, 10fthf, udcpdp, tyr-L, phe-L) with a lower bound of -1. The biomass function was modified by removing four components that could not be exchanged (pe161_p, pe160_p, kdo2lipid4_e, murein5px4p_p) to form a conserved biomass function: Ec_biomass_iML976_CONS_75p37M. This biomass function was used to optimize for growth of the model by individually replacing the sole source of C, P, N or S. Growth was considered to be present if the predicted growth rate was 5% greater than the growth rate with no exogenous source of C, N, P, or S (excluding the exchange reactions provided auxotrophies).

3.4.9 Mapping protein structures to other *E. coli* strains

Incorporating protein-related information into a GEM involves four stages of semi-automated curation: (i) link the genes of the organism to available experimental protein structures, found in publicly available databases, such as the Protein Data Bank (PDB); (ii) determine genes with and without available protein structures and perform homology modeling to create structures based on templates when none are available; (iii) perform ranking and filtering of PDB structures for each gene based on a set selection criteria (e.g., resolution, number of mutations, completeness); (iv) map GEM genes to other databases (e.g., BRENDA [68, 69]), SwissProt [70], Pfam [71], SCOP [72]) for complementary protein-structure derived data. More details on this workflow can be found in the protocol publication [64]. The quality of the structural reconstruction is further improved through a series of QC/QA verification steps during the ranking and filtering stage. To this end, each protein structure is assessed for gaps (non-resolved portions of the protein), changes amino acid sequence (mutations) and missing residues. The final structure

for every protein in *i*ML1515 has a complete, gap-less and correct amino acid sequence. This process was then carried out for 55 different strains of bacteria. We used the K-12 strain as a base strain and made modifications to the K-12 protein sequence and 3D structure based to reflect the individual differences of a given strain. Modifications to the 3D structure of the protein were carried out as previously described67.

Once we had a complete metabolic proteome for all 55 strains, we calculated 29 physical properties of the protein to construct a multidimensional data matrix. These properties include solvent-accessible surface area (SASA), number of total contacts, disulfide bond distance (SS-bond), percent of the protein that is buried, percent of the protein that is on the surface, secondary structure composition (-helical content, -strand content, 310 helix content, -helix content, hydrogen bonded turn content, bend content, disordered content), ovality (SASA/Nres2/3), residue depth (distance of the C atom from the protein surface), percent of the total structure that is nonpolar, polar, positively charged, or negatively charged, and percentage of the surface/buried residues that are nonpolar, polar, positively charged, or negatively charged. Details for each of these calculations are previously described (Brunk et al. 2016). To determine which protein properties were most important to strain differences, we performed multivariate analyses, such as Principal Component Analysis (PCA). All 3D visualization of strain variation was carried out using VMD (Visual Molecular Dynamics).

Context-specific model construction

The context-specific models were constructed using a proteomics dataset of *E. coli* BW25113 grown in 22 different culture conditions [47]. Of the 22 conditions, we built context-specific models using the proteomics data for *E. coli* grown on 7 substrates that were used for

the genome-wide essentiality screens presented here (Figure 3.2). This was done by mapping the copy numbers of each protein to the genes in *i*ML1515 for each of the 7 conditions. This data was then used to determine the dominant isozyme for each reaction in *i*ML1515 . For the model reactions capable of being catalyzed by two or more isozymes, the expression per isozyme subunit was averaged. Isozymes with average subunit expression of less than 10% of the maximum isozyme expression (i.e., the primary isozyme) were removed from the reaction. If a particular locus occurred was present in more than one AND or OR relationship in the GPR, the reaction was left unchanged. The updated GPR rules for each reaction in *i*ML1428-Glucose can be found in Supplementary Data 12. The *i*ML1428-Glucose models essentiality predictions were compared against a glucose aerobic essentiality screen of the Keio collection (see Gene Essentiality Predictions in Supplementary Methods). The remaining false positive and false negative model predictions for *i*ML1428-Glucose were tabulated and explained (Supplementary Data 12).

3.4.10 Reactive oxygen species producing reactions identification and inclusion

To include all ROS sources in our model, every enzyme with the capacity to lose electrons to O2 was incorporated from a previous study modeling ROS generation in *E. coli* K-12 MG1655 [25]. Additional ROS-generating enzymes were identified using the Ecocyc database [77]. These enzymes use flavins, quinones, and/or transition metal centers during catalysis, and are listed along with their intended, H2O2-generating and O2-generating reactions in Supplemental Data 3. In total, 164 (32 new reactions) have the capacity to generate ROS in *E. coli* K-12 MG1655 and were included in the model. We followed the same procedure as Brynildsen *et al.* [25] for coupling ROS generation to the identified ROS generating reactions. All enzymes were allowed to

produce both H2O2 and O2- simultaneously. Enzymes that use flavins or quinones derived both species from O2, while enzymes that only utilize transition metal centers derived O2- from O2, and H2O2 from O2-. This is in recognition of the fact that enzymes with only transition metal centers (e.g., Fe-S), such as aconitase, fumarase, and dihydroxy acid dehydratase, are readily oxidized by O2- [78], and that continuous recycling of these enzymes active sites occurs [79].

3.4.11 Genome-scale contextualization of transcriptomics data

A normalized database consisting of 2258 transcriptomic experiments was filtered by removing all experiments performed using mutant strains or in undefined media. An additional experiment (E70) was filtered out which had expression values inconsistent with the remaining date sets [34]. The remaining 333 experiments that met this criteria were further manually curated in order to achieve higher resolution in the experimental metadata, particularly more specific characterization of primary carbon sources and salts (i.e., MOPS, M9, etc.) in culture media, experimental perturbation, growth phase, aerobicity and strain (Supplementary Data 9). The absolute gene subunit expression was averaged for each isozyme in *i*ML1515 and the fractional expression for each of the isozymes within a reactions GPR was plotted across all experimental conditions [34] (shown for Aspartate Kinase (ASPK) in Supplementary Figure 8B).

Acknowledgements

JM Monk, CJ Lloyd, E Brunk, H Mori, AM Feist and BO Palsson designed the study. JM Monk, CJ Lloyd and AM Feist performed the model updates. E Brunk, N Mih and JM Monk performed the structure updates. JM Monk, R Takeuchi, W Nomura and H Mori performed the growth experiments. All authors helped draft and edit the final paper. The authors declare that

they have no conflict of interest. The work was funded by the Novo Nordisk Foundation and by grant 1R01GM057089 from the NIH/NIGMS.

Chapter 3 in full is a reprint of material published in: JM Monk*, **CJ Lloyd***, E Brunk, N Mih, A Sastry, Z King, R Takeuchi, W Nomura, Z Zhang, H Mori, AM Feist, BO Palsson. 2017. iML1515, a knowledgebase that computes *Escherichia coli* traits. *Nature Biotechnology* 35 (10): 9048. The dissertation author was one of the two primary authors.

3.5 References

1. Bordbar, A., Aarash, B., Monk, J. M., King, Z. A. & Palsson, B. O. Constraint-based models predict metabolic and associated cellular functions. *Nat. Rev. Genet.* **15,** 107–120 (2014).

2. Eddy, J. A., Funk, C. C. & Price, N. D. Fostering synergy between cell biology and systems biology. en. *Trends Cell Biol.* **25,** 440–445 (Aug. 2015).

3. O'Brien, E. J., Monk, J. M. & Palsson, B. O. Using Genome-scale Models to Predict Biological Capabilities. *Cell* **161,** 971–987 (May 2015).

4. Monk, J., Nogales, J. & Palsson, B. O. Optimizing genome-scale network reconstructions. en. *Nat. Biotechnol.* **32,** 447–452 (May 2014).

5. Ebrahim, A., Almaas, E., Bauer, E., Bordbar, A., Burgard, A. P., Chang, R. L., Dräger, A., Famili, I., Feist, A. M., Fleming, R. M., Fong, S. S., Hatzimanikatis, V., Herrgård, M. J., Holder, A., Hucka, M., Hyduke, D., Jamshidi, N., Lee, S. Y., Le Novère, N., Lerman, J. A., Lewis, N. E., Ma, D., Mahadevan, R., Maranas, C., Nagarajan, H., Navid, A., Nielsen, J., Nielsen, L. K., Nogales, J., Noronha, A., Pal, C., Palsson, B. O., Papin, J. A., Patil, K. R., Price, N. D., Reed, J. L., Saunders, M., Senger, R. S., Sonnenschein, N., Sun, Y. & Thiele, I. Do genome-scale models need exact solvers or clearer standards? *Mol. Syst. Biol.* **11,** 831 (Oct. 2015).

6. Heavner, B. D. & Price, N. D. Transparency in metabolic network reconstruction enables scalable biological discovery. en. *Curr. Opin. Biotechnol.* **34,** 105–109 (Aug. 2015).

7. Ravikrishnan, A. & Raman, K. Critical assessment of genome-scale metabolic networks: the need for a unified standard. en. *Brief. Bioinform.* **16,** 1057–1068 (Nov. 2015).

8. Machado, D., Herrgård, M. J. & Rocha, I. Modeling the Contribution of Allosteric Regulation for Flux Control in the Central Carbon Metabolism of E. coli. en. *Front Bioeng Biotechnol* **3,** 154 (Oct. 2015).

9. Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J., Guyer, M. & Green, E. D. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. en. *J. Am. Med. Inform. Assoc.* **21,** 957–958 (Nov. 2014).

10. Nussinov, R., Bonhoeffer, S., Papin, J. A. & Sporns, O. From "What Is?" to "What Isn't?" Computational Biology. en. *PLoS Comput. Biol.* **11,** e1004318 (July 2015).

11. Denger, K., Weiss, M., Felux, A.-K., Schneider, A., Mayer, C., Spiteller, D., Huhn, T., Cook, A. M. & Schleheck, D. Sulphoglycolysis in Escherichia coli K-12 closes a gap in the biogeochemical sulphur cycle. en. *Nature* **507,** 114–117 (Mar. 2014).

12. Kamat, S. S., Williams, H. J. & Raushel, F. M. Intermediates in the transformation of phosphonates to phosphate by bacteria. en. *Nature* **480,** 570–573 (Nov. 2011).

13. Hassaninasab, A., Hashimoto, Y., Tomita-Yokotani, K. & Kobayashi, M. Discovery of the curcumin metabolic pathway involving a unique enzyme in an intestinal microorganism. en. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 6615–6620 (Apr. 2011).

14. Chang, R. L., Andrews, K., Kim, D., Li, Z., Godzik, A. & Palsson, B. O. Structural systems biology evaluation of metabolic thermotolerance in Escherichia coli. en. *Science* **340,** 1220–1223 (June 2013).

15. Brunk, E., Mih, N., Monk, J., Zhang, Z., O'Brien, E. J., Bliven, S. E., Chen, K., Chang, R. L., Bourne, P. E. & Palsson, B. O. Systems biology of the structural proteome. en. *BMC Syst. Biol.* **10,** 26 (Mar. 2016).

16. McCloskey, D., Palsson, B. Ø. & Feist, A. M. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol. Syst. Biol.* **9,** 661 (2013).

17. Reed, J. L., Vo, T. D., Schilling, C. H. & Palsson, B. O. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). en. *Genome Biol.* **4,** R54 (Aug. 2003).

18. Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V. & Palsson, B. Ø. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. en. *Mol. Syst. Biol.* **3,** 121 (June 2007).

19. Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M. & Palsson, B. Ø. A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011. en. *Mol. Syst. Biol.* **7,** 535 (Oct. 2011).

20. Orth, J. D. & Palsson, B. Ø. Systematizing the generation of missing metabolic knowledge. en. *Biotechnol. Bioeng.* **107,** 403–412 (Oct. 2010).

21. Guzmán, G. I., Utrilla, J., Nurk, S., Brunk, E., Monk, J. M., Ebrahim, A., Palsson, B. O. & Feist, A. M. Model-driven discovery of underground metabolic functions in Escherichia coli. en. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 929–934 (Jan. 2015).

22. Orth, J. D. & Palsson, B. Gap-filling analysis of the iJO1366 Escherichia coli metabolic network reconstruction for discovery of metabolic functions. en. *BMC Syst. Biol.* **6,** 30 (May 2012).

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

24. Linster, C. L., Van Schaftingen, E. & Hanson, A. D. Metabolite damage and its repair or pre-emption. en. *Nat. Chem. Biol.* **9,** 72–80 (Feb. 2013).

25. Brynildsen, M. P., Winkler, J. A., Spina, C. S., MacDonald, I. C. & Collins, J. J. Potentiating antibacterial activity by predictably enhancing endogenous microbial ROS production. en. *Nat. Biotechnol.* **31,** 160–165 (Feb. 2013).

26. King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O. & Lewis, N. E. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. en. *Nucleic Acids Res.* **44,** D515–22 (Jan. 2016).

27. Gama-Castro, S., Jiménez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Peñaloza-Spinola, M. I., Contreras-Moreira, B., Segura-Salazar, J., Muñiz-Rascado, L., Martínez-Flores, I., Salgado, H., Bonavides-Martínez, C., Abreu-Goodger, C., Rodríguez-Penagos, C., Miranda-Ríos, J., Morett, E., Merino, E., Huerta, A. M., Treviño-Quintanilla, L. & Collado-Vides, J. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. en. *Nucleic Acids Res.* **36,** D120–4 (Jan. 2008).

28. Cho, B.-K., Federowicz, S., Park, Y.-S., Zengler, K. & Palsson, B. Ø. Deciphering the transcriptional regulatory logic of amino acid metabolism. en. *Nat. Chem. Biol.* **8,** 65–71 (Nov. 2011).

29. Cho, B.-K., Federowicz, S. A., Embree, M., Park, Y.-S., Kim, D. & Palsson, B. Ø. The PurR regulon in Escherichia coli K-12 MG1655. en. *Nucleic Acids Res.* **39,** 6456–6464 (Aug. 2011).

30. Cho, S., Cho, Y.-B., Kang, T. J., Kim, S. C., Palsson, B. & Cho, B.-K. The architecture of ArgR-DNA complexes at the genome-scale in Escherichia coli. en. *Nucleic Acids Res.* **43,** 3079–3088 (Mar. 2015).

31. Cardoso, J. G. R., Andersen, M. R., Herrgård, M. J. & Sonnenschein, N. Analysis of genetic variation and potential applications in genome-scale metabolic modeling. en. *Front Bioeng Biotechnol* **3,** 13 (Feb. 2015).

32. Notebaart, R. A., Szappanos, B., Kintses, B., Pál, F., Györkei, Á., Bogos, B., Lázár, V., Spohn, R., Csörgő, B., Wagner, A., Ruppin, E., Pál, C. & Papp, B. Network-level architecture and the evolutionary potential of underground metabolism. en. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 11762–11767 (Aug. 2014).

33. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. en. *Genome Res.* **13,** 2498–2504 (Nov. 2003).

34. Kim, M., Zorraquino, V. & Tagkopoulos, I. Microbial forensics: predicting phenotypic characteristics and environmental conditions from large-scale gene expression profiles. en. *PLoS Comput. Biol.* **11,** e1004127 (Mar. 2015).

35. Zhang, Y. I-TASSER: fully automated protein structure prediction in CASP8. en. *Proteins* **77 Suppl 9,** 100–113 (2009).

36. *Systems Biology: Constraint-based Reconstruction and Analysis* (Cambridge University Press, 2015).

37. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? en. *Nat. Biotechnol.* **28,** 245–248 (Mar. 2010).

38. Lewis, N. E., Nagarajan, H. & Palsson, B. O. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. en. *Nat. Rev. Microbiol.* **10,** 291–305 (Apr. 2012).

39. Ye, Y. & Godzik, A. Database searching by flexible protein structure alignment. en. *Protein Sci.* **13,** 1841–1850 (July 2004).

40. Kufareva, I. & Abagyan, R. Methods of protein structure comparison. en. *Methods Mol. Biol.* **857,** 231–257 (2012).

41. Manolio, T. A. Genomewide association studies and assessment of the risk of disease. en. *N. Engl. J. Med.* **363,** 166–176 (July 2010).

42. O'Brien, E. J., Monk, J. M. & Palsson, B. O. Using Genome-scale Models to Predict Biological Capabilities. en. *Cell* **161,** 971–987 (May 2015).

43. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L. & Mori, H. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. en. *Mol. Syst. Biol.* **2,** 2006.0008 (Feb. 2006).

44. Yamamoto, N., Nakahigashi, K., Nakamichi, T., Yoshino, M., Takai, Y., Touda, Y., Furubayashi, A., Kinjyo, S., Dose, H., Hasegawa, M., Datsenko, K. A., Nakayashiki, T., Tomita, M., Wanner, B. L. & Mori, H. Update on the Keio collection of Escherichia coli single-gene deletion mutants. en. *Mol. Syst. Biol.* **5,** 335 (Dec. 2009).

45. Takeuchi, R., Tamura, T., Nakayashiki, T., Tanaka, Y., Muto, A., Wanner, B. L. & Mori, H. Colony-live–a high-throughput method for measuring microbial colony growth kinetics–reveals diverse growth effects of gene knockouts in Escherichia coli. en. *BMC Microbiol.* **14,** 171 (June 2014).

46. Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M. & Palsson, B. Ø. A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011. en. *Mol. Syst. Biol.* **7,** 535 (Oct. 2011).

47. Schmidt, A., Kochanowski, K., Vedelaar, S., Ahrné, E., Volkmer, B., Callipo, L., Knoops, K., Bauer, M., Aebersold, R. & Heinemann, M. The quantitative and condition-dependent Escherichia coli proteome. en. *Nat. Biotechnol.* **34,** 104–110 (Jan. 2016).

48. Blattner, F. R., Plunkett 3rd, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B. & Shao, Y. The complete genome sequence of Escherichia coli K-12. en. *Science* **277,** 1453–1462 (Sept. 1997).

49. Hobman, J. L., Penn, C. W. & Pallen, M. J. Laboratory strains of Escherichia coli: model citizens or deceitful delinquents growing old disgracefully? en. *Mol. Microbiol.* **64,** 881–885 (May 2007).

50. Lukjancenko, O., Wassenaar, T. M. & Ussery, D. W. Comparison of 61 sequenced Escherichia coli genomes. en. *Microb. Ecol.* **60,** 708–720 (Nov. 2010).

51. Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., Feist, A. M. & Palsson, B. Ø. Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. en. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 20338–20343 (Dec. 2013).

52. Salipante, S. J., Roach, D. J., Kitzman, J. O., Snyder, M. W., Stackhouse, B., Butler-Wu, S. M., Lee, C., Cookson, B. T. & Shendure, J. Large-scale genomic sequencing of extraintestinal pathogenic Escherichia coli strains. en. *Genome Res.* **25,** 119–128 (Jan. 2015).

53. Von Mentzer, A., Connor, T. R., Wieler, L. H., Semmler, T., Iguchi, A., Thomson, N. R., Rasko, D. A., Joffre, E., Corander, J., Pickard, D., Wiklund, G., Svennerholm, A.-M., Sjöling, Å. & Dougan, G. Identification of enterotoxigenic Escherichia coli (ETEC) clades with long-term global distribution. en. *Nat. Genet.* **46,** 1321–1326 (Dec. 2014).

54. Marx, V. Microbiology: the road to strain-level identification. en. *Nat. Methods* **13,** 401–404 (Apr. 2016).

55. Scholz, M., Ward, D. V., Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., Truong, D. T., Tett, A., Morrow, A. L. & Segata, N. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. en. *Nat. Methods* **13,** 435–438 (May 2016).

56. Karlsson, F. H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C. J., Fagerberg, B., Nielsen, J. & Bäckhed, F. Gut metagenome in European women with normal, impaired and diabetic glucose control. en. *Nature* **498,** 99–103 (June 2013).

57. Jensen, K. F. The Escherichia coli K-12“ wild types” W3110 and MG1655 have an rph frameshift mutation that leads to pyrimidine starvation due to low pyrE expression levels. *J. Bacteriol.* **175,** 3401–3407 (1993).

58. Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., Bordbar, A., Lewis, N. E., Rahmanian, S., Kang, J., Hyduke, D. R. & Palsson, B. Ø. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protoc.* **6,** 1290–1307 (2011).

59. Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. en. *BMC Syst. Biol.* **7,** 74 (Aug. 2013).

60. Furnham, N., Sillitoe, I., Holliday, G. L., Cuff, A. L., Rahman, S. A., Laskowski, R. A., Orengo, C. A. & Thornton, J. M. FunTree: a resource for exploring the functional evolution of structurally defined enzyme superfamilies. en. *Nucleic Acids Res.* **40,** D776–82 (Jan. 2012).

61. Chang, R. L., Xie, L., Bourne, P. E. & Palsson, B. O. Antibacterial mechanisms identified through structural systems pharmacology. en. *BMC Syst. Biol.* **7,** 102 (Oct. 2013).

62. Yim, H., Haselbeck, R., Niu, W., Pujol-Baxley, C., Burgard, A., Boldt, J., Khandurina, J., Trawick, J. D., Osterhout, R. E., Stephen, R., Estadilla, J., Teisan, S., Schreyer, H. B., Andrae, S., Yang, T. H., Lee, S. Y., Burk, M. J. & Van Dien, S. Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. en. *Nat. Chem. Biol.* **7,** 445–452 (May 2011).

63. Chung, H., Yang, J. E., Ha, J. Y., Chae, T. U., Shin, J. H., Gustavsson, M. & Lee, S. Y. Bio-based production of monomers and polymers by metabolically engineered microorganisms. en. *Curr. Opin. Biotechnol.* **36,** 73–84 (Dec. 2015).

64. Xu, P., Ranganathan, S., Fowler, Z. L., Maranas, C. D. & Koffas, M. A. G. Genome-scale metabolic network modeling results in minimal interventions that cooperatively force carbon flux towards malonyl-CoA. en. *Metab. Eng.* **13,** 578–587 (Sept. 2011).

65. Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M. & Palsson, B. Ø. A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011. *Mol. Syst. Biol.* **7,** 535 (Oct. 2011).

66. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* **5,** 93–121 (2010).

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

68. Chang, A., Scheer, M., Grote, A., Schomburg, I. & Schomburg, D. BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. en. *Nucleic Acids Res.* **37,** D588–92 (Jan. 2009).

69. Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G. & Schomburg, D. BRENDA, the enzyme database: updates and major new developments. en. *Nucleic Acids Res.* **32,** D431–3 (Jan. 2004).

70. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. & Schneider, M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. en. *Nucleic Acids Res.* **31,** 365–370 (Jan. 2003).

71. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C. & Eddy, S. R. The Pfam protein families database. en. *Nucleic Acids Res.* **32,** D138–41 (Jan. 2004).

72. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. en. *J. Mol. Biol.* **247,** 536–540 (Apr. 1995).

73. Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J. & Punta, M. Pfam: the protein families database. en. *Nucleic Acids Res.* **42,** D222–30 (Jan. 2014).

74. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L. & Mori, H. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. en. *Mol. Syst. Biol.* **2,** 2006.0008 (Feb. 2006).

75. Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G. D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A. & Zagnitko, O. The RAST Server: rapid annotations using subsystems technology. en. *BMC Genomics* **9,** 75 (Feb. 2008).

76. Disz, T., Akhter, S., Cuevas, D., Olson, R., Overbeek, R., Vonstein, V., Stevens, R. & Edwards, R. A. Accessing the SEED genome databases via Web services API: tools for programmers. en. *BMC Bioinformatics* **11,** 319 (June 2010).

77. Karp, P. D., Weaver, D., Paley, S., Fulcher, C., Kubo, A., Kothari, A., Krummenacker, M., Subhraveti, P., Weerasinghe, D., Gama-Castro, S., Huerta, A. M., Muñiz-Rascado, L., Bonavides-Martinez, C., Weiss, V., Peralta-Gil, M., Santos-Zavaleta, A., Schröder, I., Mackie, A., Gunsalus, R., Collado-Vides, J., Keseler, I. M. & Paulsen, I. The EcoCyc Database. en. *EcoSal Plus* **6** (May 2014).

78. Keseler, I. M., Bonavides-Martínez, C., Collado-Vides, J., Gama-Castro, S., Gunsalus, R. P., Johnson, D. A., Krummenacker, M., Nolan, L. M., Paley, S., Paulsen, I. T., Peralta-Gil, M., Santos-Zavaleta, A., Shearer, A. G. & Karp, P. D. EcoCyc: a comprehensive view of Escherichia coli biology. en. *Nucleic Acids Res.* **37,** D464–70 (Jan. 2009).

79. Gardner, P. R. & Fridovich, I. Inactivation-reactivation of aconitase in Escherichia coli. A sensitive measure of superoxide radical. en. *J. Biol. Chem.* **267,** 8757–8763 (May 1992).

Chapter 4

Revealing the intricate relationships between proteome cofactor requirements and growth environments

Sustaining a robust metabolic network requires a balanced and fully functioning proteome. Many enzymes require cofactors (coenzymes and engrafted prosthetic groups) to function properly. The metabolic capabilities of *Escherichia coli* have been comprehensively described using extensively validated genome-scale metabolic and expression models (ME-models). ME-models have the unique ability to compute an optimal proteome composition underlying a metabolic phenotype, including the provision of all required cofactors. Here we use ME-models for 55 different *E. coli* strains to examine how environmental conditions change proteome usage. We

found that: (1) ME-model computations reveal variability in cofactor use depending on the specific metabolic phenotype; (2) Despite the critical importance of cofactors for healthy, robust growth, B vitamin (enzyme cofactors) auxotrophies are relatively common in wild-type *E. coli* strains; (3) The ME-model could describe how such auxotrophies affect the metabolic state of *E. coli* , revealing candidate evolutionary drivers of auxotrophy in terms of ROS stress mitigation, protein specialization, growth yield, and opportunism. Genome-scale models have reached a level of sophistication where they reveal intricate properties of functional proteomes and how they support different *E. coli* lifestyles.

4.1 Enzyme cofactor activity and metabolism are intrinsically linked

A key component of synthesizing a functional proteome–along with translating the proper amino acid sequence and folding the protein into the proper 3D structure–often includes equipping enzymes with the necessary prosthetic groups and coenzymes. These accessory enzyme cofactors often drive chemical conversions at the heart of the enzymes activity, making their presence essential for detectable catalytic activity [1, 2]. The essential functions of many of some cofactors such as flavins and iron sulfur clusters can be traced back to enzymatic functions found at the beginning of life [3]. Therefore, ensuring that all coenzymes and prosthetic groups are available to enzymes is essential for a robustly growing organism. Scarcity of one or more of the essential micronutrients can have a profound impact on the metabolic state of an organism, such as disruptions in energy metabolism and lactate secretion in iron-limited stress conditions [4]. Likewise, the presence of certain cofactors can shape the evolutionary trajectory of an organism.

For example, *E. coli* does not have the metabolic capability to synthesize tetrahydrobiopterin, a required cofactor for hydroxylase reactions. To circumvent this limitation, *E. coli* has evolved to synthesize tyrosine from chorismate as opposed to simply hydroxylating phenylalanine.

The cofactors essential in microorganisms are also required in organisms across the tree of life, either obtained through direct synthesis or through the organisms diet. Humans, for example, require the same cofactors but do not have the catabolic capabilities to produce them thus making cofactors an essential component of our diet as B vitamins (Table 4.1). Alternatively, *E. coli* K-12 MG1655 and many other microbe, are capable of *de novo* synthesizing most of the cofactors listed in Table 4.1. Many strains within the *E. coli* species, however, have evolved in a way rendering them unable synthesize some of these essential cofactors, making them auxotrophic. The ubiquity in the requirement of these cofactors for growth, along with variability in the ability of organisms to produce these metabolites, often give cofactors an important role in shaping community dynamics [5]. This can be seen in natural systems–such as the complex interactions between microbial communities and the human host in the gut microbiome–or in industrial settings [6].

It is well known how most cofactors enable the catalytic function of its associated enzymes. Additionally, the essential catalytic activity enabled by these cofactors have been tied to a growth-supporting function. However, a systems understanding of how cofactor use–or lack thereof–can shape the metabolism of an organism on a genome scale is missing.

Computing metabolic states and proteome composition

One established method for studying the metabolic capabilities of an organism is a genome-scale metabolic model (M-model). M-models have shown significant success predicting

Table 4.1: Summary of the relevant cofactors in *E. coli* K-12 MG1655

Cofactor Name (Bigg ID)	General Function	Cellular Role	Essentiality from Xavier *et al.*	Vitamin
Thiamin (thm)	Energy metabolism	Prosthetic group	Universal	B1
Riboflavin (ribflv)	Redox metabolism	Prosthetic group/redox coenzyme	Universal (FMN, FAD)	B2
Niacin (nac)	NAD(P) precursor, electron carrier	Coenzyme	Universal (NAD, NADP)	B3
Pantothenoic Acid (pnto_R)	CoA precursor, fatty acid biosynthesis	Coenzyme	Universal (CoA)	B5
Pyridoxine (pydxn)	Versatile coenzyme that participates in transamination, decarboxylation, etc.	Prosthetic group	Universal	B6
Biotin (btn)	Required for carboxylase activity	Prosthetic group	Conditional	B7
Folate (thf)	Carrier of single carbon moieties	Coenzyme	Universal	B9
Cobalamin (cbl1)	Certain isomerases and methyltransferases, not essential in K-12	Prosthetic group	Conditional	B12
Menaquinone 8(mqn8)	Electron carrier	Coenzyme	Conditional (quinones)	K2
Ubiquinone 8 (q8)	Electron carrier	Coenzyme	Conditional (quinones)	-
2-Demethylmenaquinone 8 (2dmmq8)	Electron carrier	Coenzyme	Conditional (quinones)	-

the metabolic capabilities of a cell by integrating all of the experimentally determined enzymatic reactions taking place in an organism [7–10]. These predictions are enabled simply based on the stoichiometric constraints of the organisms metabolic network and metabolic interactions with the environment. Therefore, a focus of development in the field of genome-scale models has been to increase the scope and capabilities of M-models [11].

Recently the capabilities of M-models have been leveraged by a multi-strain modeling workflow that can be used to construct strain-specific metabolic models for a variety of different species including Escherichia coli [12, 13], Staphylococcus aureus [14], and Salmonella [15]. This modeling approach has shown that the metabolic capabilities of a strain can be predicted with relative accuracy starting with a high-quality genome-scale model of an organism and whole-genome sequencing of individual strains [16]. Most of these 55 strains of *E. coli* have evolved the ability to produce all essential biomass constituents from glucose and a limited number of

inorganic salts. Despite the clear evolutionary benefit of maintaining an ability to grow in the presence of scarce resources, throughout their evolutionary history, certain strains of *E. coli* have shed–or possibly never evolved–the ability to synthesize some metabolites. The cause for this is can be logically inferred in some cases. For example, it is known that *E. coli* can become dependent on supplemented branched chain and aromatic amino acids when exposed to high levels of oxidative stress [17]. This is due to enzymes in the biosynthetic pathways of these metabolites that are sensitive to oxidative damage. Therefore, it is likely that auxotrophies for these amino acids can evolve as a method to mitigate high reactive oxygen species (ROS) stress. This could be particularly beneficial for pathogenic strains, which experience high oxidative stress levels during the immune response. Alternatively, auxotrophies can evolve as a mechanism for cells to adapt to growth in community. The specific ways that auxotrophies can shape cell metabolism, however, have not been studied in detail.

The 55 multi-strain reconstructions demonstrated that M-models were effective in predicting growth supporting nutrients of an organism. However, these models are not capable of mechanistically accounting for the unique proteome that supports growth on these nutrients, such as the unique cofactor requirement. In M-models the cofactors are either enzyme prosthetic groups and have no modeled metabolic function (pyridoxine, biotin, etc.) or are recycled (NAD, folates, etc.), meaning there is no metabolic process driving their biosynthesis. Thus they have often been incorporated into the biomass objective function to force the essential biosynthetic activity of synthesizing these cofactors [18]. The inclusion of cofactors into biomass objective functions has been studied across various bacterial and archaea species providing insight into the essentiality of individual cofactors in prokaryotes [19]. However, even when included in the biomass objective function, a negligible amount of each cofactor is required to be synthesized for

growth, causing cofactor synthesis to have little impact on metabolism overall. Furthermore, the specific requirement of the cofactors is largely condition independent. Modeling efforts have been made to assess how the biomass function composition (lipid and amino acid composition) affects metabolic fluxes [20]. A mechanistic model, however, has not been employed to relate cofactor availability to condition dependent metabolism.

To that end, M-models have been extended to include the synthesis of the gene expression machinery and its use to compute the entire metabolic and gene expression proteome [21–23]. These models integrate Metabolism and Expression on the genome-scale, termed ME-models, and they are capable of explicitly computing over 80% of the proteome by mass in enterobacteria. ME-models enable a wide range of new biological questions to be investigated including direct calculations of proteome allocation [24], metabolic pathway usage, and the effects of membrane and volume constraints [22]. Furthermore, their ability to compute the optimal proteome abundances for a given condition make them ideal for mechanistically integrating transcriptomics and proteomics data.

ME-models include a mechanistic accounting of all of the components required to produce a functioning proteome (Figure 4.1). This means that for a particular enzymatic reaction to carry flux in the model, not only must the amino acids be synthesized in the proper proportions, but enzyme cofactors must be produced as well. Further, given that many common auxotrophies include enzyme cofactor vitamins, ME-models are a suitable tool to address the consequences of these auxotrophies on a genome scale. Here we address these questions by applying the *E. coli* ME-model to examine the relationship between growth condition and cofactor demand along with the metabolic consequences of auxotrophy. This work presents the first effort to comprehensively study the role that essential cofactors play in defining the metabolic capabilities of *E. coli* .

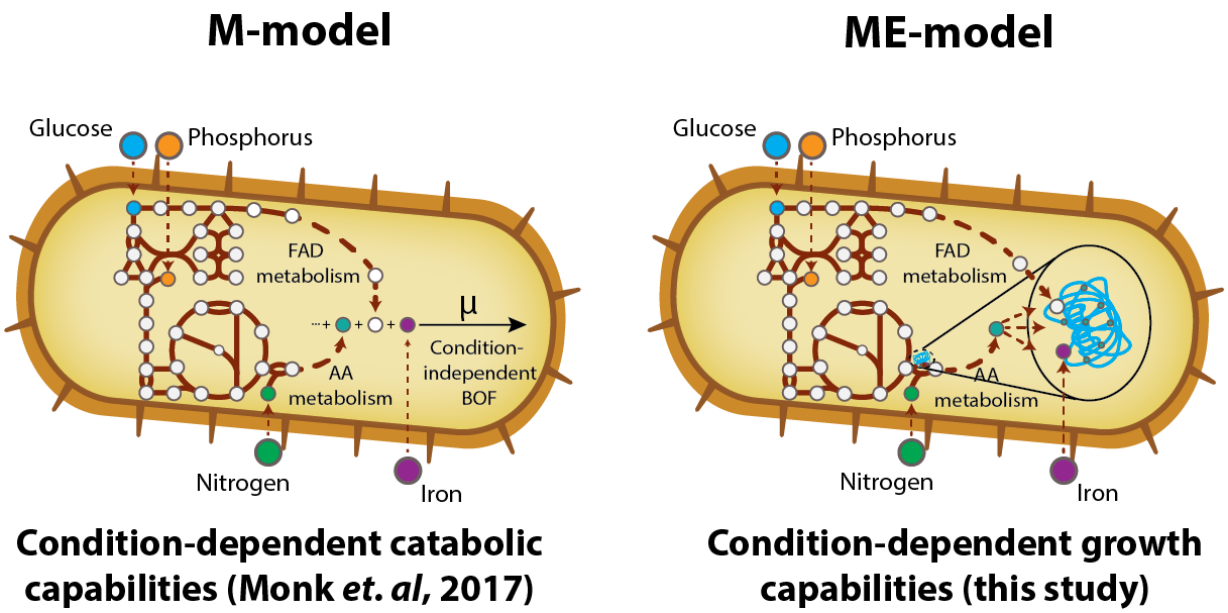

Figure 4.1: Difference in M- and ME- model scope. M-models offer a means to comprehensively probe the range of enzymatic conversions possible within an organism. This information can be used to predict possible growth supporting nutrient environments [12, 25]. ME-models add additional information about the proteome sustaining the growth state by including a mechanistic accounting of enzyme synthesis. This provides the ability to study how proteome allocation and cofactor use affects condition-dependent growth.

4.2 Results

4.2.1 Predicting demand for essential biomass components

ME-models include an explicit accounting of the steps required to produce a functioning enzyme, making them capable of predicting the condition-specific synthesis demand of most essential biomass components. This enables *i*JL1678b-ME , the most recent *E. coli* K-12 MG1655 ME-model reconstruction, to predict abundances of enzyme prosthetic groups. However, coenzymes such as NAD and folates are recycled throughout the network and are therefore not synthesized by default. *i*JL1678b-ME was thus modified to formally describe the activity of these coenzymes and to couple coenzyme synthesis to its metabolic function (see Methods). Following these modifications, the synthesis fluxes provided by *i*JL1678b-ME agreed relatively well

with the *i*JO1366 biomass objective function under aerobic and anaerobic *in silico* conditions (Figure 4.1). The amino acid synthesis fluxes agree best with the *i*JO1366 BOF since these have little dependence on the activity of individual enzymes and are more properties of global protein amino acid proportions. Likewise there is little change between aerobic and anaerobic simulations.

Unlike amino acids, the majority of the cofactor and micronutrient coefficients used in the *i*JO1366 core objective function are not derived from empirical data but rather included in low amounts to more correctly compute gene essentiality [19]. Therefore, the quantitative comparison of ME-model predicted *in silico* micronutrient demand to the M-model biomass objective function holds little meaning. As a result, the growth rate dependent synthesis of these micronutrients has a negligible effect on the computed fluxes overall.

There is a stark difference in cofactor demand between the aerobic and anaerobic computed fluxes. This stems from the fact that predictions provided by the ME-model are dependent on the activity of specific reactions as well as the kinetic parameters used to couple reaction flux to enzyme abundance [26]. Therefore, processes/reactions such as oxidative phosphorylation and pyruvate dehydrogenase that are not used in anaerobic conditions see a decrease in their accompanying cofactors, ubiquinone and thiamine diphosphate (vitamin B1), respectively.

4.2.2 Growth condition-dependent cofactor demand

*i*JL1678b-ME was used to simulate growth on 552 nitrogen, phosphorus, sulfur, and carbon sources under aerobic and anaerobic *in silico* conditions and the activity of each cofactor was found. A high degree of variability was observed for many of the cofactors depending on the particular nutrient source (Figure 4.3). These differences in cofactor demand stem from the fact

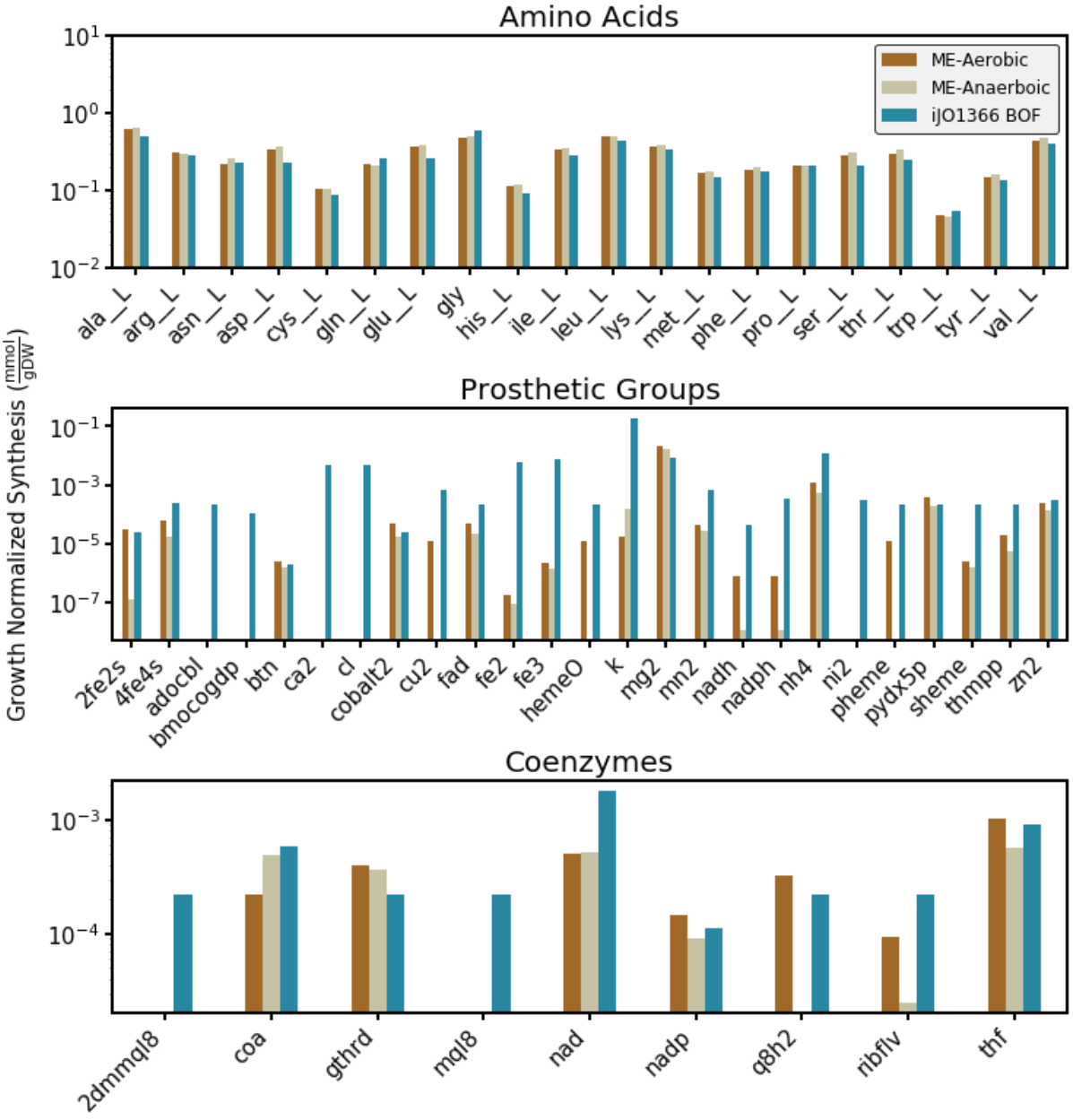

Figure 4.2: Comparison of ME-model and M-model predicted amino acid and cofactor growth-normalized synthesize rates. The ME-model predictions are a function of the predicted intracellular fluxes provided by the simulation, whereas the M-model values are provided by the biomass objective. ME-model predictions are shown for aerobic and anaerobic *in silico* conditions. Metabolites are in black or red if their M-model coefficient value was taken from the core objective function or the wild-type objective function, respectively.

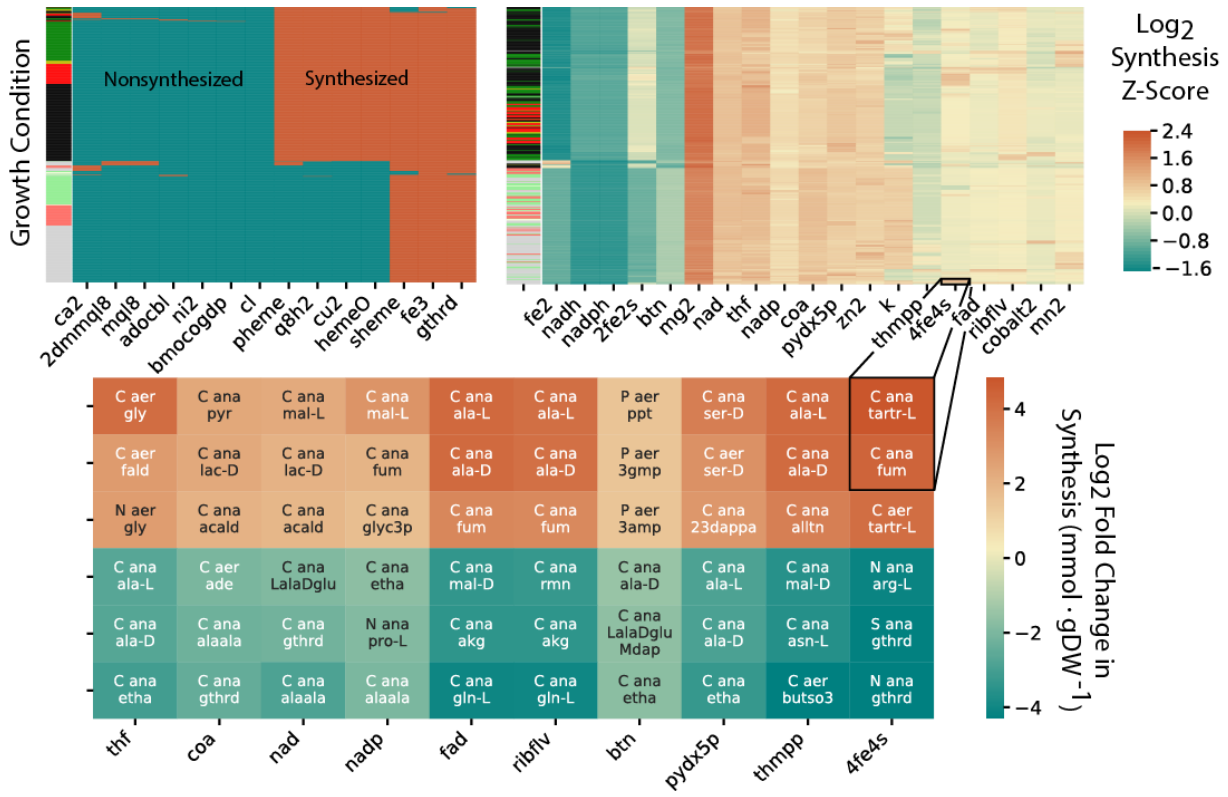

Figure 4.3: Condition-dependent synthesis demand of common enzyme cofactors. The upper left panel shows the cofactors that are conditionally required, clustered by growth condition. The nutrient sources are shown in red, green, yellow, and black for phosphorus, nitrogen, sulfur, and carbon sources, respectively. The light colors indicate anaerobic simulations while the dark colors indicate aerobic simulations. The top right panel shows a clustering of the Z score values of the growth normalized cofactor synthesis values computed from the ME-model. The bottom panel depicts the 6 carbon, nitrogen, phosphorus, and sulfur sources that most severely increase or decrease demand of the cofactor on the x axis is shown on below. The heatmap is colored based on the log2 fold change between the cofactor demand for the nutrient source listed and the median demand for all growth conditions.

that the simulation for each feasible growth supporting nutrient possesses a unique metabolic state sustained by a unique proteome. Thus, the cofactors required to support this proteome differ accordingly.

A subset for enzymatic cofactors was computationally predicted to be synthesized only in some growth conditions. The use of some of these cofactors (e.g., ubiquinone-8, pheme, and hemeO) differed largely based on whether simulated under aerobic or anaerobic conditions

(Figure 4.3). This is the expected behavior for these cofactors, as they are primarily required for aerobic respiration functions. The demand of other cofactors such as adenosylcobalamin and siroheme are specific to individual growth conditions. For example, adenosylcobalamin is computationally required for growth only when the carbon or nitrogen source is ethanolamine, since this adenosylcobalamin in an essential cofactor for ethanolamine ammonia-lyase, the first step of ethanolamine catabolism. Siroheme is required in most growth conditions as a prosthetic group for sulfite reductase, an essential stop in the reduction of sulfate to hydrogen sulfide for sulfur assimilation. Growth in other sulfur sources such as cysteine and cysteine derivatives is computationally predicted to alleviate the need for siroheme.

Likewise, the universally synthesized cofactors cluster based on the aerobicity of the simulations (Figure 4.3). The differential clustering is primarily driven by differences in the growth normalized demand of biotin, NAD(P)H, iron moieties, and CoA. The observed differences in NAD(P)H and iron demand are easily explainable by an increase in glycolytic/fermentation activity and a decrease in oxidative respiration for anaerobic metabolism relative to aerobic. The observed increase in biotin and CoA demand in anaerobic conditions suggest an increase in relative lipid metabolism under this growth condition. Other cofactors such as tetrahydrofolate, thiamine, and pyridoxine appear to be required at similar levels under aerobic and anaerobic conditions.

There are specific growth conditions, however, that significantly increase or decrease the demand of all cofactors. The three nutrient sources that most dramatically increase or decrease the computed growth normalized metabolic demand for each of the cofactors are also shown in Figure 4.3. The change in growth normalized cofactor demand ranged from 46 to 1/18 fold compared to the median demand for each cofactor across all conditions. L-alanine as

an anaerobic carbon substrate provides the highest increase and L-glutamine as an anaerobic carbon source provided the largest increase and decrease in thiamin and riboflavin demand, respectively. This information can be used to design cellular microenvironments to increase or decrease the susceptibility of *E. coli* to some antibiotic treatments. For example, antifolates that inhibit the synthesis of tetrahydrofolate are a commonly used antibiotic. This analysis would suggest that cells in a glycine rich microenvironment would be more susceptible to antifolate treatment, due to the 22 fold increase in folate demand (Figure 4.3).

Beyond prospective design, this information provides insight into possible factors underlying *E. coli* evolution. It is predicted that anaerobic growth on ethanolamine displays the lowest growth normalized cellular requirement for tetrahydrofolate, biotin, and pyridoxine of all the nutrient sources tested. While this substrate cannot grow in *E. coli* as a sole carbon source, ethanolamine is abundantly found in the animal gut as a byproduct of lipid degradation [27]. Ethanolamine has also been shown to be beneficial to growth in some strains of *E. coli* when provided as a supplemental nitrogen source in addition to ammonium [28]. The computational results shown in Figure 4.3 suggest that supplementing this metabolite could improve growth by decreasing the cellular demand of a subset of the essential cofactors. Furthermore, since this trend is only seen when ethanolamine, not nitrogen, is supplemented as a carbon source, this effect is likely the result of an increase in the acetaldehyde produced during ethanolamine catabolism.

4.2.3 Relating auxotrophy and *E. coli* metabolism

As demonstrated above, ME-models offer the unique ability to comprehensively study the ways that the metabolic demands of an organism can directly influence its use of essential cofac-

tors. Alternatively, ME-models can be applied to understand the opposite relationship: between cofactor availability and the metabolic state of the organism. These cofactors are involved in many important cellular functions, thus their activity can have a profound impact on the cellular phenotype [29]. Regardless of this fact, throughout their evolution, many strains of *E. coli* have lost the ability to synthesize some of these cofactors. To assess these metabolic consequences, the ME-model was applied to study *E. coli* auxotrophs in *in silico* conditions of excess and limited availability of the auxotrophic metabolite.

Auxotrophy in nutrient excess

Known [12] and computationally predicted [30] gene knockouts that result in auxotrophy in *E. coli* were individually imposed in *i*JL1678b-ME for 16 amino acids and cofactors. Figure 4.4 shows computed percent difference in four features of cell growth compared to the prototrophic ancestor strain when an excess of the auxotrophic nutrient is available. Essentially, this shows the computed growth benefit to the cell when the metabolic burden of producing an essential biomass building block or cofactor is alleviated.

This analysis shows a differential *in silico* metabolic response between the amino acid auxotrophs and cofactor auxotrophs. Amino acid auxotrophs in nutrient excess displayed a general decrease in the activity of proteins susceptible to ROS damage, particularly the arginine and glutamate auxotrophs. Protein susceptibility to oxidative stress was determined based on a previous analysis that considered protein structure information and residue 3D location [17]. Furthermore, the amino acid supplemented auxotrophs showed a notable increase in growth rate accompanied by a decrease in growth yield, while the cofactor supplemented auxotrophs showed a more modest increase in growth rate but no change in yield.

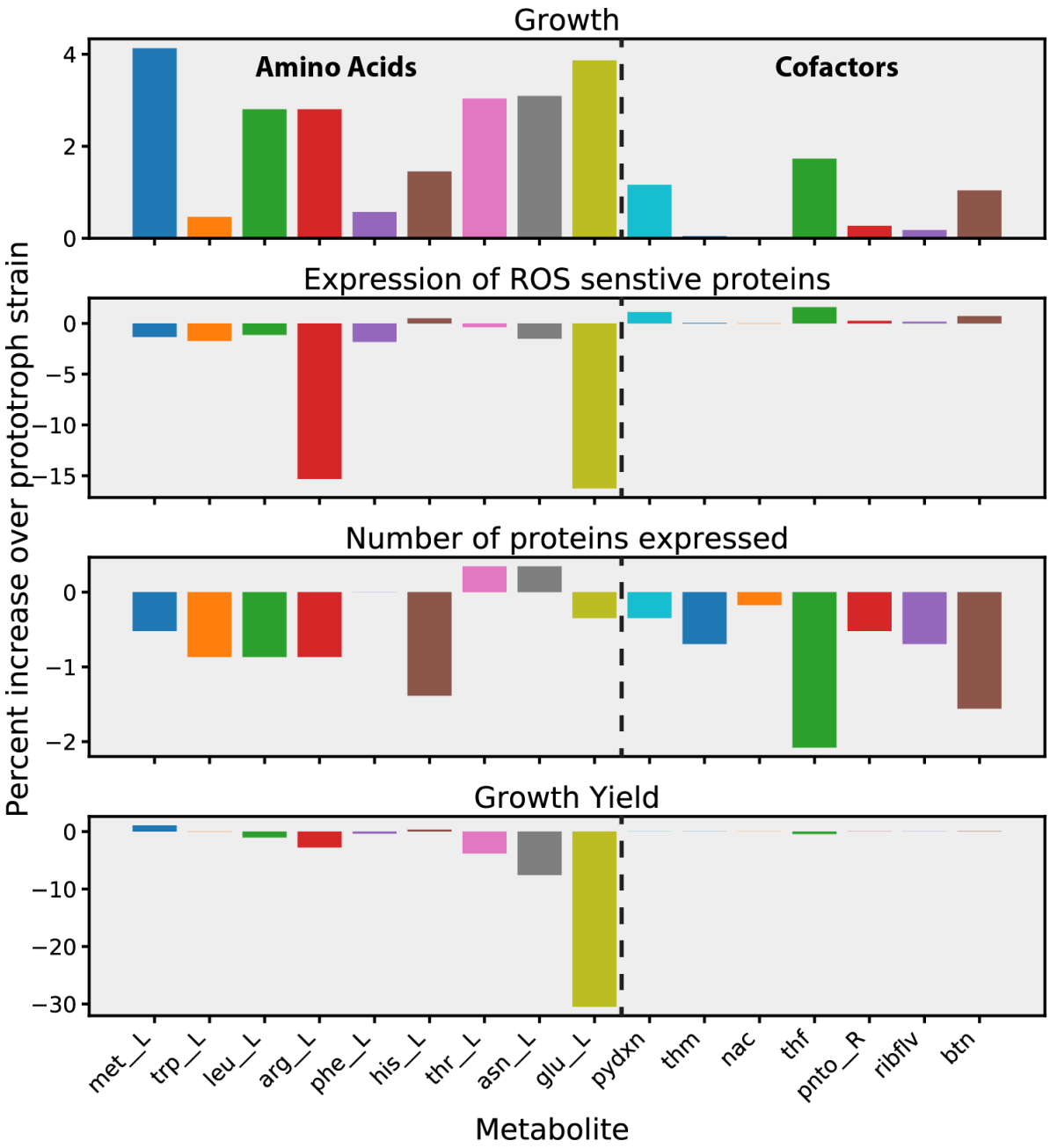

Figure 4.4: Growth characteristics in excess of the auxotrophic nutrient shown on the x-axis. There is a clear differential response to conditions of excess cofactors versus excess amino acids.

Auxotrophy in nutrient limitation

No microbes growth environment is constant throughout its lifespan. The same is true for any auxotrophic organism, meaning that any auxotroph will experience periods of auxotrophic

nutrient limitation along with excess. To gauge how nutrient limitation impacts growth, simulations were run with *in silico* metabolite exposure ranging from the computationally optimal uptake to 1/20th of the optimal value, and growth rate was optimized. The same features of the *in silico* growth state from Figure 4.4 were observed as a function of auxotroph metabolite uptake. As shown in Figure 4.5, there is notable variability in the sensitivity of the *in silico* cells depending on the identity of the auxotrophic metabolite being limited. Most notably, folate (thf) auxotrophs are computationally predicted to be particularly growth sensitive to drops in folate availability below the optimal amount. This drop in growth occurs in two phases, one in which growth drops sharply to 50% of the maximum as 15% of the folate availability decreases. The second phase displays a gradual decrease in growth from 50% of the maximum growth to 0. In the first region flux is predicted to be redirected from the oxidative PPP (Phosphogluconate dehydrogenase) and the TCA cycle toward the acetyl-CoA node, potentially as a sink for NADH. Further, in folate excess, the Phosphoribosylglycinamide formyltransferase (GARFT) reaction uses formyl-thf to produce N2-Formyl-N1-(5-phospho-D-ribosyl)glycinamide (fgam), an important intermediate metabolite in nucleotide biosynthesis. Folate limitation results in a transition to using GAR transformylase-T (GART), an ATP driven reaction that can produce fgam from free formate (see Figure 4.6). Given the rise in antimicrobial resistance to some antifolates antibiotics, understanding metabolic strategies for tolerating folate stress could provide clues for how to combat this resistance through possible combinatorial therapies.

Nicotinic acid (nac) limitation is predicted to result in an immediate 6% increase in the number of expressed proteins and up to a 30% increase in the activity of ROS sensitive proteins. In an nac limited growth state, the cellular polls of reduced and oxidized NADP+ and NAD+ would be highly depleted. Therefore, an optimally growing cell in this state would likely redirect

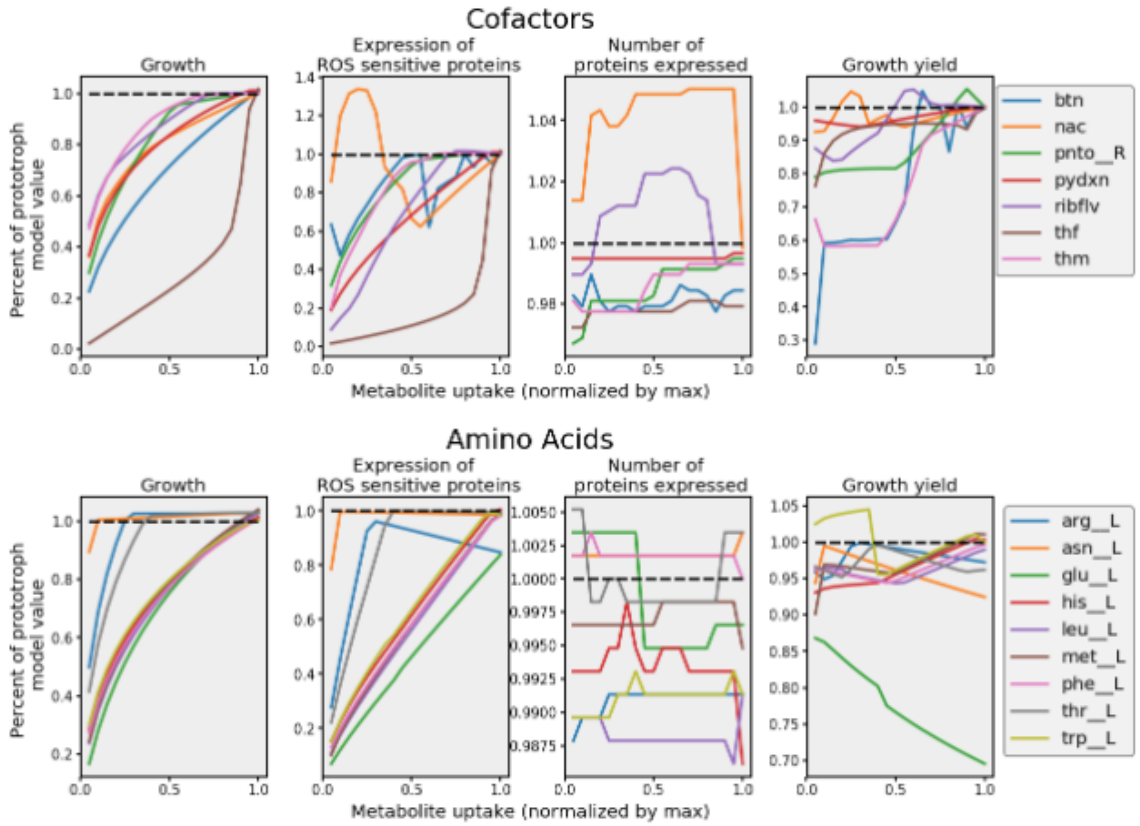

Figure 4.5: Growth characteristics of 16 computational *E. coli* auxotrophs as a function of availability of auxotrophic metabolites. For each metabolite shown in the legend, reactions were imposed into the model creating a specific auxotrophy for that metabolite. The growth rate, iron uptake, protein expression, and growth yield per mol carbon are plotted as a function of the availability of the metabolite indicated in the legend. The percent change of the plotted quantities compared to the default prototroph model is shown.

flux into pathways that still maximize growth, but do not require these two cofactors. *i*JL1678b-ME predicts that the optimal approach to accomplish optimize NAD+ and NADP+ use is to upregulate the enter-duteroff pathway (bypassing lower glycolysis), to increase activity of the glyoxylate shunt to donate electrons to the quinate pool via malate, and to donate electrons to the quinate pool via pyruvate formate lyase (PFL) and formate dehydrogenase (FDH) (Supplementary Figure 4.3). It is unclear whether the latter metabolic route is feasible since PFL is only expressed in anaerobic conditions in *E. coli* K-12 MG1655. Many of these computationally

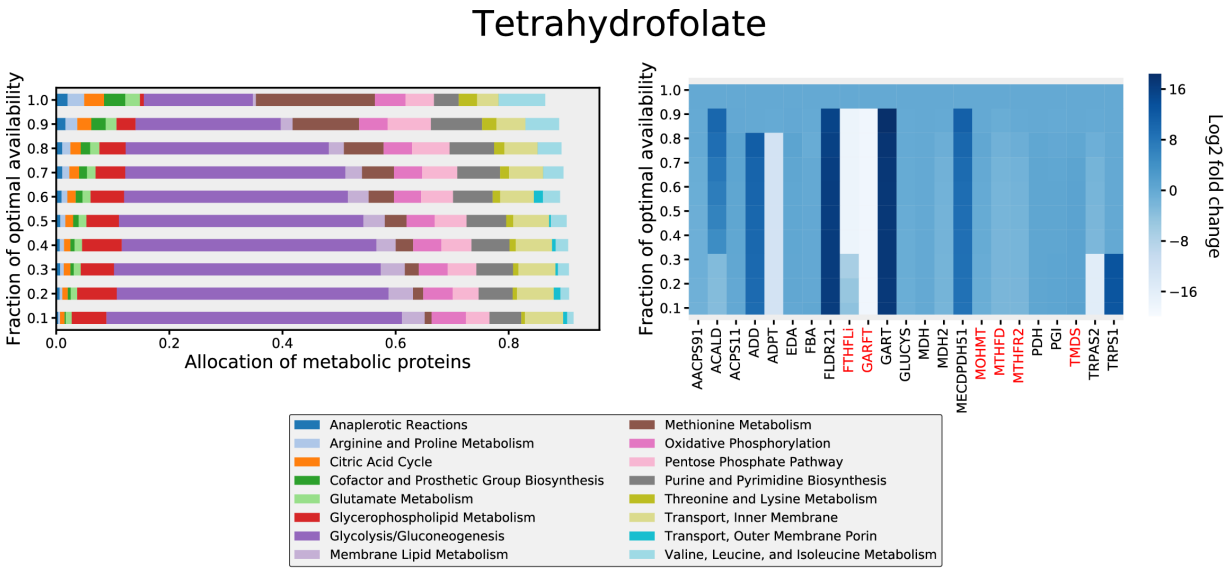

Figure 4.6: Model-predicted metabolic changes in response to folate limitation. Left panel: Fraction of protein allocated to each metabolic subsystem for varying folate availability (rows) Right panel: Heatmap of $\log_2$ fold changes in growth rate normalized reaction fluxes compared to fluxes computed with excess folate. The reactions with the highest standard deviation are shown and are highlighted in red if the reaction relies on folate activity. The top row depicts a simulation with the highest folate availability and bottom row depicting a simulation with the lowest folate availability.

upregulated pathways are enriched in iron cofactors (e.g., the citric acid cycle) largely accounting for the increase in iron uptake observed.

4.3 Discussion

The presented work provides the first computational study highlighting the interplay between *E. coli* s condition dependent growth state and it cofactor use. Simulations of model growth on a variety of different growth supporting nutrients suggest that notable variability can be observed in the demand of cofactors based on the specific growth environment. Such information can be used to boost the efficacy of antibiotics that target specific cofactors by suggesting ways to manipulate the cells microenvironment. Further, metabolic consequences accompanying the

loss of metabolic capability to synthesize one of these cofactors was examined. These results provide insight into the metabolic consequences and evolutionary drivers of auxotrophy.

This work provides a new look into the inherent coupling between *E. coli* synthesized small molecule cofactors and metabolism. A separate integral part of a functional proteome includes the metal ion cofactors that form the enzymatic center of many enzymes. Due to the interchangeability of some ion cofactors and intricate mechanisms underlying enzyme mismetallation, fully examining metal ion cofactors was out of the scope of this study [31]. Future work is warranted to exclusively study the metalloproteome and how metal ion availability shapes metabolism.

This ME-modeling method could be applied to answer additional questions regarding cofactor biosynthesis in *E. coli* . For example, *E. coli* cannot synthesize vitamin B12 (cobalamin) from glucose, but when this vitamin is supplemented, *E. coli* can synthesize various derivatives of vitamin B12 including adenosylcobalamin. Further, *E. coli* possesses a handful of enzymes that require vitamin B12 for catalytic activity. Future work could asses the evolutionary pressures driving the loss of this synthetic capability.

Lastly, the predictions from this computational study are well suited for future experimental validation. First, many of the vitamin and amino acid auxotrophs are the product of only single gene knockouts in *E. coli* K-12 MG1655, meaning these strains either already exist in single knockout libraries [32] or can be easily synthesized. Adaptive laboratory evolution of these auxotrophs in low concentrations of the essential nutrients could provide valuable insight into the way *E. coli* can adapt to re-invest its protein toward pathways that maximize growth, while minimizing cofactor use. Second, this work provide predictions of how manipulate the growth environment of *E. coli* to potentiate the effect of antibiotics. Testing these predictions *in vivo*

would be relatively trivial and could further solidify the utility of this modeling method.

4.4 Methods

4.4.1 Software

All constraint-based modeling analyses were performed using Python 3.6 and the COBRApy software [33], and ME-model operations were performed using the COBRAme framework [34]. Due to the fact that ME-models are ill-scaled [21], qMINOS [35, 36], which supports quad (128-bit) precision, was used for each ME-models simulations. M-model simulations were performed using the *i*JO1366 model of *E. coli* K-12 MG1655 metabolism [37], since this is the model that *i*JL1678b-ME was reconstructed from. All M-model optimizations were performed using the Gurobi (Gurobi Optimization, Inc., Houston, TX) linear programming (LP) solver.

4.4.2 ME-model parameterization

The keff coupling parameters [21] for each metabolic reaction in *i*JL1678b-ME were determined based on a machine learning approach that incorporated enzyme and network features to predict keffs [26] from a set of *in vivo* derived turnover rates [38]. The remaining keffs for expression machinery and transport reactions were set to 65 s-1 as these processes were out of the scope of the machine learning approach.

4.4.3 Coupling cofactor activity to biosynthesis demand

The *i*JL1678b-ME ME-model of *E. coli* K-12 MG1655 was used for all simulations in the presented work. The activity of enzyme prosthetic groups are inherent in the ME-model formulation [21], which uses coupling constraints to couple the activity of individual enzymes

(including their accessory groups) to the reaction they catalyze. Coenzymes (NAD, folates, etc.) have some of the same properties of enzymes in that they are recycled within the network in both M- and ME- models. These models ensure that the cofactors are balanced, but they do not account for the biosynthesis of these coenzymes. As a result, models have incorporated these coenzymes in a biomass function to force their biosynthesis in a way that is independent on the coenzymes activity in the model.

The *i*JL1678b-ME ME-model was thus modified to couple the biosynthesis of coenzymes to their activity, similar to other enzymes in the model. This is accomplished using a pseudo-kinetic term to relate the concentration of the coenzyme pool to its activity throughout the metabolic network, which we will simply call $k_{activity}$. This term represents a very rough estimation of the individual kinetics of all of the reaction involving the coenzyme in the network. This term was chosen as 1x104 hr^{-1} and applied to each reaction where the uncharged version of the coenzyme acts as a reactant:

$$(1 + \frac{\mathrm{mu}}{\mathrm{k_{activity}}})\ \mathrm{uncharged\_coenzyme} + \mathrm{met\_1} \rightarrow \mathrm{charged\_coenzyme} + \mathrm{met\_2}$$

For this study, we are interested in the relative activity of these coenzymes across varying growth conditions with decreases in relative coenzyme availability. It is important that the computed coenzyme abundances are within a reasonable range (Figure 4.2), but quantitative accuracy of the abundance predictions are not necessary. Therefore, accounting for the complex kinetics of the coenzymes throughout the reconstruction was outside the scope of this work. This approach effectively scales the rate of coenzyme biosynthesis linearly with its metabolic activity and growth rate.

4.4.4 Optimization Procedure

Due to non-linearities stemming from the cofactor and enzyme coupling coupling constraints ME-models cannot be optimized using a binary search algorithm. To perform the binary search, the following procedure was implemented. First, each symbolic coefficient or reaction bound was compiled into a function by sympy [39]. Then, a linear program was created for the linear programming solver, with all of these symbolic functions evaluated at 0. While the model will always be feasible at 0, starting with a known feasible point results in a basis which can be used to speed up the next run. Afterwards, for each instance of the binary search in μ, values in the linear program were replaced by recomputed ones, and the problem was resolved using the last feasible basis.

Acknowledgements

The following authors contributed to this reserach: CJ Lloyd, JM Monk, L Yang, A Ebrahim, and BO Palsson conceived the study; CJ Lloyd performed the analysis; L Yang provided the enzyme susceptibility analysis; JM Monk provided multi-strain analysis; CJ Lloyd and BO Palsson wrote the manuscript. The authors would like to thank Zachary King, Justin Tan, Bin Du, and Joshua Lerman for informative discussions. Funding for this work was provided by the Novo Nordisk Foundation through the Center for Biosustainability at the Technical University of Denmark under Grant NNF10CC1016517 and the NIH National Institute of General Medical Sciences under Grant no. NIH R01 GM057089. 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.

Chapter 4 in part is a reprint of material published in:**CJ Lloyd**, JM Monk, L, Yang, A

Ebrahim, and BO Palsson. "Genome-scale models reveal the intricate relationships between proteome cofactor requirements and growth environments." 2019. *In Preparation.* The dissertation author is the primary author.

4.5 References

1. Sprenger, G. A., Schörken, U., Sprenger, G. & Sahm, H. Transketolase A of Escherichia coli K12. Purification and properties of the enzyme from recombinant strains. en. *Eur. J. Biochem.* **230,** 525–532 (June 1995).

2. Barber, M. C., Price, N. T. & Travers, M. T. Structure and regulation of acetyl-CoA carboxylase genes of metazoa. en. *Biochim. Biophys. Acta* **1733,** 1–28 (Mar. 2005).

3. Weiss, M. C., Sousa, F. L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S. & Martin, W. F. The physiology and habitat of the last universal common ancestor. en. *Nat Microbiol* **1,** 16116 (July 2016).

4. Folsom, J. P., Parker, A. E. & Carlson, R. P. Physiological and proteomic analysis of Escherichia coli iron-limited chemostat growth. en. *J. Bacteriol.* **196,** 2748–2761 (Aug. 2014).

5. Zengler, K. & Zaramela, L. S. The social network of microorganisms — how auxotrophies shape complex communities. *Nat. Rev. Microbiol.* **16,** 383–390 (2018).

6. Ruiz-Barba, J. L. & Jimenez-Diaz, R. Availability of Essential B-Group Vitamins to Lactobacillus plantarum in Green Olive Fermentation Brines. en. *Appl. Environ. Microbiol.* **61,** 1294–1297 (Apr. 1995).

7. Bordbar, A., Aarash, B., Monk, J. M., King, Z. A. & Palsson, B. O. Constraint-based models predict metabolic and associated cellular functions. *Nat. Rev. Genet.* **15,** 107–120 (2014).

8. O'Brien, E. J., Monk, J. M. & Palsson, B. O. Using Genome-scale Models to Predict Biological Capabilities. *Cell* **161,** 971–987 (May 2015).

9. Lewis, N. E., Nagarajan, H. & Palsson, B. O. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. en. *Nat. Rev. Microbiol.* **10,** 291–305 (Apr. 2012).

10. McCloskey, D., Palsson, B. Ø. & Feist, A. M. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol. Syst. Biol.* **9,** 661 (2013).

11. Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B., Assad-Garcia, N., Glass, J. I. & Covert, M. W. A Whole-Cell Computational Model Predicts Phenotype from Genotype. *Cell* **150,** 389–401 (July 2012).

12. Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., Feist, A. M. & Palsson, B. Ø. Genome-scale metabolic reconstructions of multiple Escherichia coli

strains highlight strain-specific adaptations to nutritional environments. en. *Proceedings of the National Academy of Sciences,* 201307797 (Nov. 2013).

13. Monk, J. M., Koza, A., Campodonico, M. A., Machado, D., Seoane, J. M., Palsson, B. O., Herrgård, M. J. & Feist, A. M. Multi-omics Quantification of Species Variation of Escherichia coli Links Molecular Features with Strain Phenotypes. en. *Cell Syst* **3,** 238–251.e12 (Sept. 2016).

14. Bosi, E., Monk, J. M., Aziz, R. K., Fondi, M., Nizet, V. & Palsson, B. Ø. Comparative genome-scale modelling of Staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity. en. *Proc. Natl. Acad. Sci. U. S. A.* **113,** E3801–9 (June 2016).

15. Seif, Y., Kavvas, E., Lachance, J.-C., Yurkovich, J. T., Nuccio, S.-P., Fang, X., Catoiu, E., Raffatellu, M., Palsson, B. O. & Monk, J. M. Genome-scale metabolic reconstructions of multiple Salmonella strains reveal serovar-specific metabolic traits. en. *Nat. Commun.* **9,** 3771 (Sept. 2018).

16. Monk, J. & Bosi, E. Integration of Comparative Genomics with Genome-Scale Metabolic Modeling to Investigate Strain-Specific Phenotypical Differences. en. *Methods Mol. Biol.* **1716,** 151–175 (2018).

17. Yang, L., Mih, N., Anand, A., Park, J. H., Tan, J., Yurkovich, J. T., Monk, J. M., Lloyd, C. J., Sandberg, T. E., Seo, S. W., Kim, D., Sastry, A. V., Phaneuf, P., Gao, Y., Broddrick, J. T., Chen, K., Heckmann, D., Szubin, R., Hefner, Y., Feist, A. M. & Palsson, B. O. *Cellular responses to reactive oxygen species can be predicted on multiple biological scales from molecular mechanisms* 2017.

18. Feist, A. M. & Palsson, B. O. The biomass objective function. en. *Curr. Opin. Microbiol.* **13,** 344–349 (June 2010).

19. Xavier, J. C., Patil, K. R. & Rocha, I. Integration of Biomass Formulations of Genome-Scale Metabolic Models with Experimental Data Reveals Universally Essential Cofactors in Prokaryotes. en. *Metab. Eng.* **39,** 200 (Jan. 2017).

20. Pramanik, J. & Keasling, J. D. Effect of Escherichia coli biomass composition on central metabolic fluxes predicted by a stoichiometric model. en. *Biotechnol. Bioeng.* **60,** 230–238 (Oct. 1998).

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

22. Liu, J. K., O Brien, 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 (Sept. 2014).

23. Lerman, J. A., Hyduke, D. R., Latif, H., Portnoy, V. A., Lewis, N. E., Orth, J. D., Schrimpe-Rutledge, A. C., Smith, R. D., Adkins, J. N., Zengler, K. & Palsson, B. O. In silico method

for modelling metabolism and gene product expression at genome scale. *Nat. Commun.* **3,** 929 (July 2012).

24. LaCroix, R. A., Sandberg, T. E., O'Brien, 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. Environ. Microbiol.* **81,** 17–30 (Jan. 2015).

25. Monk, J. M., Lloyd, C. J., Brunk, E., Mih, N., Sastry, A., King, Z., Takeuchi, R., Nomura, W., Zhang, Z., Mori, H., Feist, A. M. & Palsson, B. O. iML1515, a knowledgebase that computes Escherichia coli traits. en. *Nat. Biotechnol.* **35,** 904–908 (Oct. 2017).

26. Heckmann, D., Lloyd, C. J., Mih, N., Ha, Y., Zielinski, D. C., Haiman, Z. B., Desouki, A. A., Lercher, M. J. & Palsson, B. O. Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. en. *Nat. Commun.* **9,** 5252 (Dec. 2018).

27. Kaval, K. G. & Garsin, D. A. Ethanolamine Utilization in Bacteria. en. *MBio* **9** (Feb. 2018).

28. Rowley, C. A., Anderson, C. J. & Kendall, M. M. Ethanolamine Influences Human Commensal Escherichia coli Growth, Gene Expression, and Competition with Enterohemorrhagic E. coli O157:H7. en. *MBio* **9** (Oct. 2018).

29. Holm, A. K., Blank, L. M., Oldiges, M., Schmid, A., Solem, C., Jensen, P. R. & Vemuri, G. N. Metabolic and transcriptional response to cofactor perturbations in Escherichia coli. en. *J. Biol. Chem.* **285,** 17498–17506 (June 2010).

30. Lloyd, C. J., King, Z. A., Sandberg, T. E., Hefner, Y., Olson, C. A., Phaneuf, P. V., O'Brien, E. J., Sanders, J. G., Salido, R. A., Sanders, K., Brennan, C., Humphrey, G., Knight, R. & Feist, A. M. The genetic basis for adaptation of model-designed syntrophic co-cultures. en. *PLoS Comput. Biol.* **15,** e1006213 (Mar. 2019).

31. Imlay, J. A. The mismetallation of enzymes during oxidative stress. en. *J. Biol. Chem.* **289,** 28121–28128 (Oct. 2014).

32. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L. & Mori, H. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. en. *Mol. Syst. Biol.* **2,** 2006.0008 (Feb. 2006).

33. Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst. Biol.* **7,** 74 (Aug. 2013).

34. Lloyd, C. J., Ebrahim, A., Yang, L., King, Z. A., Catoiu, E., O'Brien, E. J., Liu, J. K. & Palsson, B. O. COBRAme: A computational framework for genome-scale models of metabolism and gene expression. en. *PLoS Comput. Biol.* **14,** e1006302 (July 2018).

35. Yang, L., Ma, D., Ebrahim, A., Lloyd, C. J., Saunders, M. A. & Palsson, B. O. solveME: fast and reliable solution of nonlinear ME models. *BMC Bioinformatics* **17,** 391 (2016).

36. Ma, D., Yang, L., Fleming, R. M. T., Thiele, I., Palsson, B. O. & Saunders, M. A. Reliable and efficient solution of genome-scale models of Metabolism and macromolecular Expression. en. *Sci. Rep.* **7,** 40863 (Jan. 2017).

37. Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M. & Palsson, B. Ø. A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011. *Mol. Syst. Biol.* **7,** 535 (Oct. 2011).

38. Davidi, D., Noor, E., Liebermeister, W., Bar-Even, A., Flamholz, A., Tummler, K., Barenholz, U., Goldenfeld, M., Shlomi, T. & Milo, R. Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitrokcatmeasurements. *Proceedings of the National Academy of Sciences* **113,** 3401–3406 (2016).

39. Joyner, D., Čertík, O., Meurer, A. & Granger, B. E. Open source computer algebra systems: SymPy. *ACM Commun. Comput. Algebra* **45,** 225–234 (Jan. 2012).

Chapter 5

The effect of protein allocation on bacterial community characteristics

Understanding the fundamental characteristics of microbial communities has far reaching implications for human health and applied biotechnology. However, much is still unknown regarding the genetic basis and evolutionary strategies underlying the formation of viable synthetic communities. By pairing auxotrophic mutants in co-culture, it has been demonstrated that viable nascent *E. coli* communities can be established where the mutant strains are metabolically coupled. A novel algorithm, OptAux, was constructed to design 61 unique multi-knockout *E. coli* auxotrophic strains that require significant metabolite uptake to grow. These predicted knockouts included a diverse set of novel non-specific auxotrophs that result from inhibition of major biosynthetic subsystems. Three OptAux predicted non-specific auxotrophic strains–with diverse metabolic deficiencies–were co-cultured with an L-histidine auxotroph and optimized via adaptive laboratory evolution (ALE). Time-course sequencing revealed the genetic changes employed

by each strain to achieve higher community growth rates and provided insight into mechanisms for adapting to the syntrophic niche. A community model of metabolism and gene expression was utilized to predict the relative community composition and fundamental characteristics of the evolved communities. This work presents new insight into the genetic strategies underlying viable nascent community formation and a novel computational method to elucidate metabolic changes that empower the creation of cooperative communities.

5.1 Bacterial communities are ubiquitous in health and biotechnology

Microbial communities are capable of accomplishing many intricate biological feats due to their ability to partition metabolic functions among community members. Therefore, these microbial consortia have the attractive potential to accomplish complex tasks more efficiently than a single wild-type or engineered microbial strain . Past applications include applying communities to aid in waste decomposition, fuel cell development, and the creation of biosensors [1]. In the field of metabolic engineering, microbial communities have now been engineered capable of enhancing product yield or improving process stability by partitioning catalytic functions among community members [2–8]. Beyond biotechnology applications, studying microbial communities also has important health implications. This includes providing a better understanding of the gut microbiome and how it is affected by diet and other factors [9, 10]. For example, metabolic cross-feeding in communities has been shown to have a role in modulating the efficacy of antibiotics treatments [11]. Developing new computational and experimental approaches to better understand the creation of viable microbial communities and the inherent characteristics

of established communities could therefore have far reaching implications.

Experimental methods developed to study how simple communities form Synthetic communities have been constructed to study their interactions and new metabolic capabilities. One such study encouraged synthetic symbiosis between E. coli strains by co-culturing an L-isoleucine auxotroph with a L-leucine auxotroph [12, 13]. It was observed that the community was able to grow in glucose minimal media without amino acid supplementation due to amino acid cross-feeding between the mutant pairs. Mee et al. expanded upon this work by studying all possible binary pairs of 14 amino acid auxotrophs and developing methods to predict the results of combining the auxotrophic strains into 3-member, 13-member, and 14-member communities [14]. Similarly, Wintermute *et al.* observed community formation using a more diverse set of auxotrophs by co-culturing 46 conditionally lethal single gene knockouts from the E. coli Keio collection [15]. This work demonstrated that synthetic mutualism was possible in strains beyond amino acid auxotrophs [16]. These studies also demonstrated that new viable communities can be established in relatively short time frames (¡4 days) by pairing auxotrophic strains.

In addition to establishing syntrophic growth, nascent auxotrophic communities can be optimized by adaptive laboratory evolution (ALE) [17]. Expanding upon the experimental work in Mee *et al.* [14], Zhang *et al.* performed ALE on one of the co-culture pairs: a L-lysine auxotroph paired with a L-leucine auxotroph [17]. Separate co-cultures evolved to growth rates 3-fold greater than the parent, which was accomplished, in part, by forming different auxotroph strain abundances within the community. Similarly, Marchal *et al.* evolved co-cultures of two E. coli amino acid auxotrophs and sequenced the endpoint strains. This data was leveraged to identify mutations hinting at changes in the spatial structure that occurred during the evolution [18]. Studies of evolved co-culture pairs composed of different microbial species have also used

sequencing data and mutational analysis as a crucial component of interpreting adaptive strategies [19, 20]. The success of the above work demonstrated that ALE can be used to optimize auxotrophic communities and that mutational data provide valuable insight into mechanisms underlying the evolved improvements in community growth rates.

New computational methods are needed Computational methods have been established to study the characteristics of microbial communities. These methods often apply genome-scale metabolic models (M-models) [21–23]. Computational models have been created that use multi-compartmental flux balance analysis (FBA) [23–26], dynamic flux balance analysis (dFBA) [17, 27], dFBA integrated with spatial diffusion of extracellular metabolites (COMETS) [28], and FBA with game theory [29]. Novel algorithms have also been developed to describe general community characteristics (OptCom [30]) and dynamics (d-OptCom [31]). These algorithms employ a bilevel linear programming problem to find the metabolic state that maximizes community biomass while also maximizing the biomass objectives of each individual species [32]. Numerous ecological models have also been formulated to describe community dynamics [33–35].

Despite the significant advances made by the above modeling approaches, most methods were not intended to model suspension batch ALE experiments. For instance, ALE batch experiments in suspension assume growth in excess, well-mixed nutrients, thus negating the need for diffusion considerations (COMETS) or dynamic shifts in nutrient concentrations (dFBA). Also, in order for the strains to persist serial passage in an ALE experiment, it can be assumed that the cells in co-culture are growing, on average, at the same rate, thus negating the need for a bilevel growth objective that allows for varying growth rates of community members (OptCom). Additionally, given the growing appreciation for the role limited protein availability has on governing fundamental bacterial growth characteristics [36], it is likely that protein allocation plays

a role in defining fundamental community characteristics as well. Therefore, there is a need for an applicable approach to model this experimental condition in a way that accounts for the protein cost of metabolism.

Here, we elucidate the genetic mechanisms underlying the formation of syntrophy between co-cultures of auxotrophic mutants containing diverse biosynthetic deficiencies. We first introduce the OptAux algorithm for designing auxotrophic strains that require high amounts of supplemented metabolites to grow (Figure 5.1A). The OptAux solutions provided a catalog of auxotrophic mutants representing a diverse set of metabolic deficiencies. From the catalog, four auxotrophic mutants were selected to co-culture and optimized via adaptive laboratory evolution (ALE) (Figure 5.1B). To increase the growth rate of the nascent co-culture communities, significant metabolic rewiring had to occur to allow the strains to cross-feed the high levels of the necessary metabolites. Some strains additionally had to adapt to marked changes in their homeostatic metabolic state, resulting from the inhibition of a major biosynthetic subsystem. The genetic basis accompanying this rewiring was assessed by analyzing the genetic changes (mutations and observed genome region duplications) over the course of the ALE. This mutational analysis further enabled predictions of primary metabolite cross-feeding and community composition.

To study the characteristics of the ALE-optimized communities, a community model of metabolism and expression (ME-model) was constructed [37–39] (Figure 5.1C). Such a modeling approach was necessary since previous methods of genome-scale community modeling have focused on studying the metabolic flux throughout community members (using M-models) without consideration of the enzymatic cost of the proteins that drive these metabolic processes. As proteome optimization via niche partitioning and cell specialization is a driving factor of viable

community formation in ecological systems [40–43], it is essential to consider proteomic constraints when studying bacterial communities. To this end, community ME-models were utilized to interpret the nascent communities.

5.2 Results

5.2.1 OptAux Development and Simulation

The OptAux algorithm was designed to find metabolic reactions in E. coli that, when knocked out, will result in novel auxotrophies. This algorithm was implemented by selecting a metabolite of interest and applying OptAux to identify sets of reaction knockouts that will increase the required uptake of the metabolite in order for the cell to computationally grow (Figure 5.2A). OptAux was built by modifying an existing concept introduced for designing metabolite producing strains [44] which was later additionally implemented in a mixed-integer linear programming (MILP) algorithm (RobustKnock [45]). Three key modifications were made to derive OptAux from RobustKnock. First, the inner growth rate optimization was removed so that OptAux can be run at a predetermined growth rate (set_biomass constraint, Figure 5.2B). This ensures that OptAux designs computationally require the uptake of the target metabolite at all growth rates (Figure 5.2A, Figure A in S1 Appendix). Second, the objective coefficient was reversed in order to allow the algorithm to optimize for metabolite uptake as opposed to secretion. Third, a constraint was added to allow adjustments in the specificity of OptAux solutions (see Methods). This constraint allows the OptAux simulation to uptake any additional metabolite that can be consumed by the model (competing_metabolite_uptake_threshold constraint, Figure 5.2B). Without this constraint, many OptAux predicted designs have the

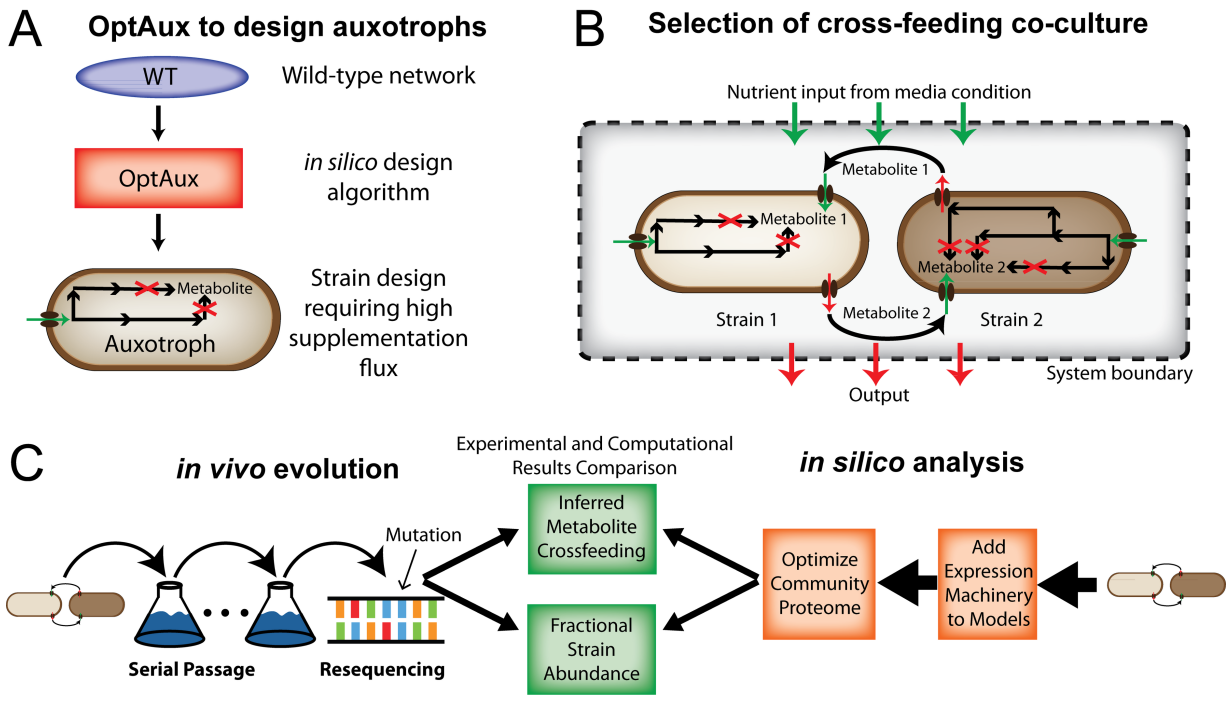

Figure 5.1: Study overview (A) An algorithm was developed to *de novo* predict reaction deletions that will produce *E. coli* strains auxotrophic for a target metabolite. (B) From the set of auxotrophic strain designs, pairs were selected to determine whether they were capable of forming a viable syntrophic community. (C) The chosen co-cultures were both evolved via adaptive laboratory evolution and modeled using a genome-scale model of *E. coli* metabolism and expression (ME-model) [37, 39]. The model predictions of fractional strain abundances and metabolite cross-feeding were compared to inferred results from the co-culture evolution experiments.

potential to additionally grow in the presence of other metabolites outside of the target metabolite. For instance, it is possible that OptAux-predicted L-glutamate auxotroph mutants could alternatively grow when supplemented with L-glutamine or other metabolites as well. Therefore, specificity, in this case, refers to whether the mutant strain will be auxotrophic for a given metabolite in the presence of other metabolites. High specificity solutions are auxotrophic for only one metabolite, regardless of whether other metabolites are present. The implementation described above allowed OptAux to identify strain designs requiring the targeted metabolite at all growth rates with varying degrees of metabolite specificity.

OptAux was utilized on the iJO1366 M-model of *E. coli* K-12 MG1655 [46, 47] to comprehensively examine auxotrophic strain designs. OptAux was run with 1, 2, and 3 reaction knockouts for 285 metabolite uptake reactions using 4 different competing_metabolite_uptake_threshold values (S1 Data). Of the given solutions, 233 knockout sets were found to be capable of producing 61 unique strain auxotrophies. This set of strain designs provides an expansive look into the auxotrophies possible in the *E. coli* K-12 MG1655 metabolic network, which could be used to understand the possible niches that *E. coli* could inhabit in natural or synthetic communities [48].

5.2.2 OptAux Solution Characteristics

The OptAux strain designs were broken into two major categories based on the number of individual metabolites that, when supplemented, can restore cell growth: 1) Essential Biomass Component Elimination Designs (EBC, Figure 5.3A) and 2) Major Subsystem Elimination Designs (MSE, Figure 5.3B). The EBC designs are characterized as auxotrophic strains with high metabolite specificity. They were broken into two subcategories: specific auxotrophs

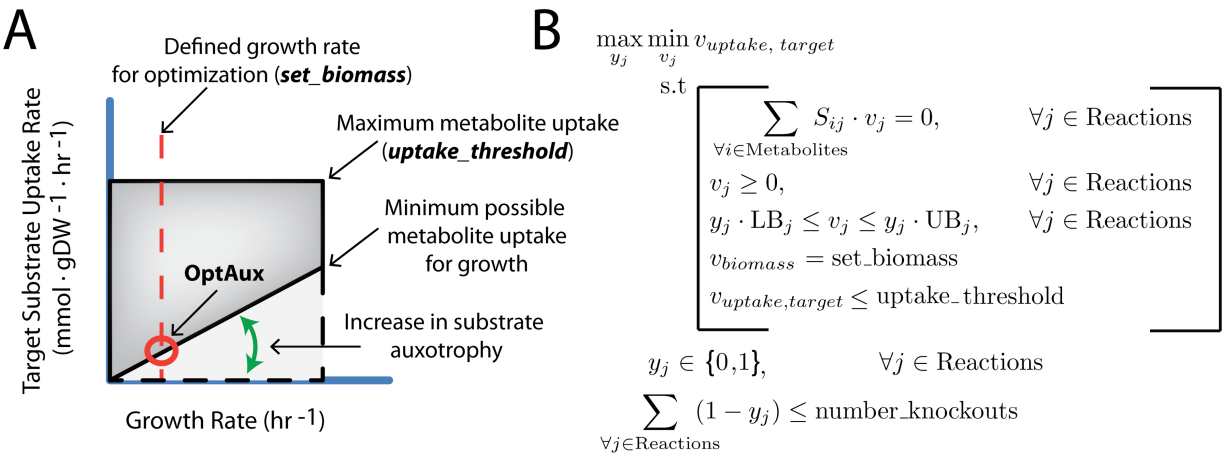

Figure 5.2: OptAux design A) OptAux was developed to maximize the minimum possible uptake of a target metabolite required for the model to grow. In other words, OptAux tries to increase the flux value at the intersection of the defined growth rate (set_biomass) and the minimum possible metabolite uptake flux (depicted with the red circle). Unlike algorithms such as OptKnock with tilting [44] and RobustKnock [45], the OptAux optimization occurs at a predetermined growth rate as opposed imposing an inner growth rate optimization. This change was made to ensure that all OptAux designs will computationally require the uptake of a target metabolite at all growth rates, particularly low growth rates. The dotted lines show the required uptake for the metabolite with no genetic interventions. In this case, uptake of the target metabolite is not required at any growth rate. The solid black lines depicts the maximum and minimum uptake required for a particular metabolite in an OptAux designed strain. (B) The OptAux optimization problem. See Methods for further description of the algorithm and underlying logic.

(only one metabolite can restore growth, Figure B in S1 Appendix) which consisted of 107 (20 unique) knockout sets and semi-specific auxotrophs (defined as strains in which less than 5 metabolites individually can restore growth, Figure B in S1 Appendix) which consisted of 67 (21 unique) knockout sets. The specific and semi-specific EBC designs were preferred at high competing_metabolite_uptake_threshold values.

There is notable overlap between OptAux predicted EBC designs (or those that are computationally identical), and known E. coli auxotrophic mutants [14, 49–60]. A summary of experimentally characterized OptAux designs is presented in Table A in S1 Appendix. Of note, there are 4 designs that were not found to be previously characterized in the scientific literature, and these present potential novel *E. coli* auxotrophs.

MSE designs were preferred at low competing_metabolite_uptake_threshold values and produce E. coli mutant strains with a diverse set of major metabolic deficiencies. These designs were defined as highly non-specific auxotrophic strains in which 5 or more metabolites could individually restore growth in the mutant strain. MSE designs consisted of the remaining 59 (20 unique) sets of knockouts. The MSE knockout strategy was often accomplished through knockouts that block metabolic entry points into key biosynthetic subsystems (Figure B in S1 Appendix). One such example of an MSE design is given in Figure 5.3B. Here a three reaction knockout design of the FUM, PPC, and MALS reactions can be rescued by one of the four compounds in the figure (i.e., citrate, L-malate, 2-oxoglutarate, or L-asparagine) at an average required uptake flux of 0.4 mmol gDW-1 hr-1 to grow at a rate of 0.1 hr-1. These rates are higher than the fluxes needed to rescue the EBC design in Figure 5.3A, which requires L-asparagine uptake of 0.024 mmol gDW-1 hr-1 on average to grow at a rate of 0.1 hr-1. Another example of a novel MSE design was a glutamate synthase (GLUSy) and glutamate dehydrogenase (GLUDy)

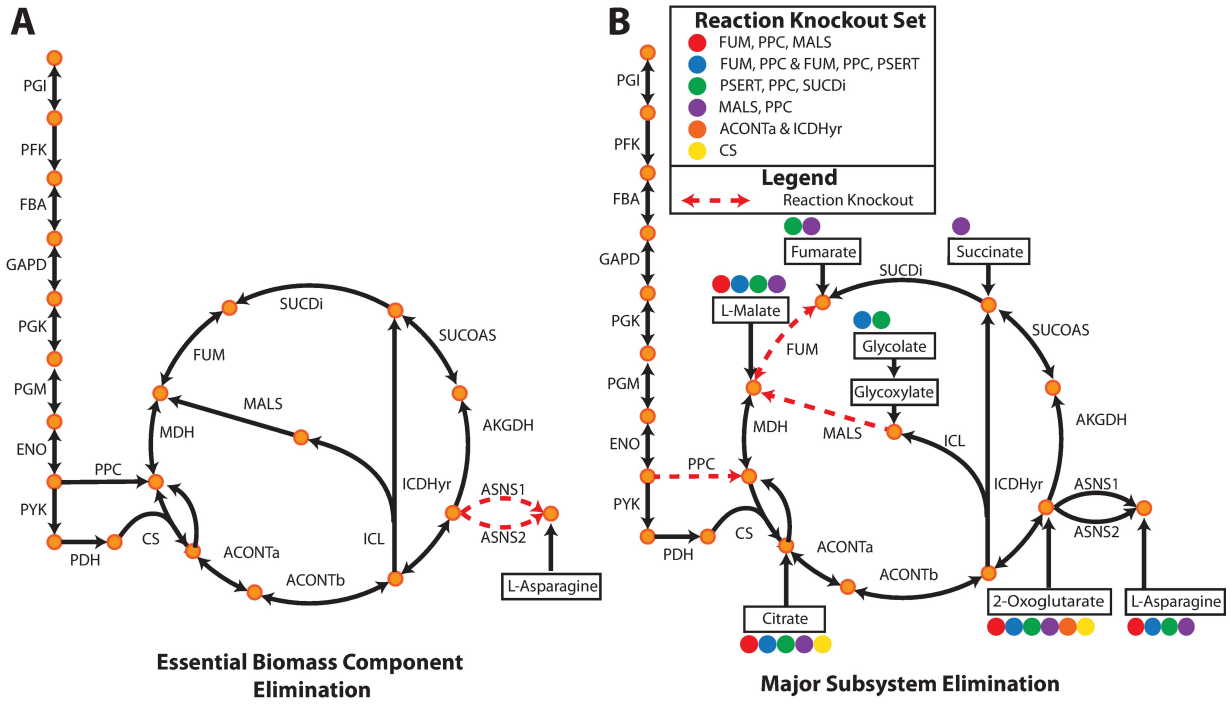

Figure 5.3: OptAux solutions Two major solution types are possible depending on the parameters used when running OptAux. (A) Essential Biomass Component Elimination designs, like the ASNS1 and ASNS2 knockout shown, can grow only when one specific metabolite is supplemented. For the case shown, this metabolite is L-asparagine. (B) Alternatively, Major Subsystem Elimination designs have a set of alternative metabolites that can individually restore growth in these strains. Examples of these designs are shown for citric acid cycle knockouts sets. One specific three reaction knockout design (FUM, PPC, MALS) is shown in red dashed lines where four metabolites in the figure can individually rescue this auxotroph (marked with solid red circles). The metabolites that can restore growth for each of the knockout strain designs listed in the legend are indicated by the colored circles.

double knockout which effectively blocks the entry of nitrogen into amino acid biosynthesis by preventing its incorporation into 2-oxoglutarate to produce L-glutamate. This renders the cell unable to produce all amino acids, nucleotides, and several cofactors. In order to grow at a rate of 0.1 hr-1, this strain is computationally predicted to require one of 19 individual metabolites at an average uptake of 0.62 mmol gDW-1 hr-1 (S2 Data).

MSE designs are of particular interest as they are often unique, non-trivial, and have not been studied in the context of E. coli auxotrophies. However, some of the MSE single knockouts have been used for a large-scale study of auxotrophic co-culture short term growth [16]. Since these predicted MSE knockouts disrupt major metabolic flows in the cells biochemical network, they produce auxotrophies that require much larger amounts of metabolite supplementation in order to grow, compared to EBC designs (e.g., Figure C in S1 Appendix). To grow in co-culture, MSE E. coli mutants would require a pronounced metabolic rewiring and likely additional adaptation to a new homeostatic metabolic state, making them attractive to study from a microbial community perspective. Additionally, any strain paired with an MSE strain in co-culture would be required to provide a relatively high amount of the MSE strains auxotrophic metabolites to enable community growth.

5.2.3 Adaptive Laboratory Evolution of Auxotrophic *E. coli* Co-cultures

To demonstrate how the OptAux algorithm can be leveraged to design strains and co-culture communities, *E. coli* auxotrophic mutants were validated in the wet lab and evolved in co-culture. Three communities were tested, each consisting of pairwise combinations of four OptAux predicted auxotrophs. This included one EBC design, $\Delta hisD$, which was validated as an L-histidine auxotroph, paired with each of three MSE designs, $\Delta pyrC$, $\Delta gltA\Delta prpC$, and

Table 5.1: Starting and final growth rates, along with fractional strain abundance of the Δ*hisD* strain (by characteristic mutation), for each ALE lineage. The cumulative number of cell division events that occurred throughout the experimental evolutions are also provided [61].

Combo	ALE #	Starting growth rate (hr^{-1})	Final growth rate (hr^{-1})	Relative Abunde -anc of Δ*hisD* (by Mutation)	Cumulative Cell Divisions ($x10^{11}$)
Δ*hisD* & Δ*pyrC*	2	0.03 ± 0.01	0.09 ± 0.02	0.29 ± 0.06	4.63
Δ*hisD* & Δ*pyrC*	3	0.03 ± 0.01	0.15 ± 0.01	0.25 ± 0.09	3.79
Δ*hisD* & Δ*pyrC*	4	0.03 ± 0.01	0.10 ± 0.02	0.21 ± 0.10	4.58
Δ*hisD* & Δ*gdhA*Δ*gltB*	5	0.04 ± 0.02	0.15 ± 0.01	0.57 ± 0.09	6.06
Δ*hisD* & Δ*gdhA*Δ*gltB*	6	0.04 ± 0.02	0.08 ± 0.01	0.55 ± 0.06	3.46
Δ*hisD* & Δ*gdhA*Δ*gltB*	8	0.04 ± 0.02	0.10 ± 0.02	0.57 ± 0.09	3.04
Δ*hisD* & Δ*gltA*Δ*prpC*	9	0.09 ± 0.02	0.19 ± 0.01	0.60 ± 0.10	7.50
Δ*hisD* & Δ*gltA*Δ*prpC*	10	0.09 ± 0.02	0.12 ± 0.02	0.50 ± 0.06	2.88
Δ*hisD* & Δ*gltA*Δ*prpC*	11	0.09 ± 0.02	0.13 ± 0.01	0.57 ± 0.09	4.77
Δ*hisD* & Δ*gltA*Δ*prpC*	12	0.09 ± 0.02	0.19 ± 0.01	0.56 ± 0.05	3.57

Δ*gdhA*Δ*gltB* . These three MSE strains had diverse metabolic deficiencies, including disruptions in pyrimidine synthesis, TCA cycle activity, and nitrogen assimilation into amino acids, respectively (Table B in S1 Appendix). The mutant was computationally predicted to be capable of growing when supplemented with one of 20 metabolites in iJO1366, and the Δ*gltA*Δ*prpC* and Δ*gdhA*Δ*gltB* mutants were predicted to grow in the presence of 14 and 19 metabolites, respectively (S2 Data, Table D in S1 Appendix).

Four replicates of each co-culture were inoculated and initially exhibited low growth rates ($\leq$ 0.1 hr^{-1}), suggesting the strains initially showed minimal cooperativity or metabolic cross-feeding (Figure D in S1 Appendix). Following approximately 40 days of ALE, all 3 co-culture combinations had evolved to establish a viable syntrophic community, indicated by an increase in the co-culture growth rate. There was diversity in the endpoint batch growth rates among the independently evolved triplicates for each of the Δ*hisD* & Δ*pyrC* and the Δ*hisD* & Δ*gdhA*Δ*gltB* co-cultures with endpoint growth rates ranging from 0.090.15 hr^{-1} and 0.080.15 hr^{-1}, respectively. The four successfully evolved independent replicates for the Δ*hisD* & Δ*gltA*Δ*prpC* co-cultures also showed endpoint growth rate diversity ranging from 0.120.19 hr^{-1} (Table 5.1,

Figure 5.4A). The relatively large range in endpoint growth rates for all co-cultures could suggest that a subset of replicates evolved to a less optimal state and thus could potentially be further improved if given more time to evolve. Alternatively, the slower growing co-cultures could have found a genetic state that resulted in a local maxima, rendering the co-culture less likely to increase its growth rate further.

To probe the adaptive strategies of the three co-culture pairs, the genomes of the populations were sequenced at several time points over the course of the 40 day evolution (Figure 5.4A). The sequencing data was used to identify genome region duplications and acquired mutations (Figure 5.4B), providing insight into the specific mechanisms employed by the co-cultures to establish cooperation.

The relative strain abundance of each mutant was tracked to observe the community composition throughout the course of the evolution. Each starting strain contained at least one unique characteristic mutation (Table C in S1 Appendix) that could act as a barcode to track the community composition (Figure 5.4B, Table 5.1). The breseq mutation identification software [62] was used to report the frequency of each of these characteristic mutations within a sequenced co-culture. The characteristic mutation frequency was then used to approximate the fraction of each strain within the co-culture population. This analysis showed that 2 of the 3 co-culture combinations maintained similar relative fractions of the two member strains, whereas one co-culture, $\Delta hisD$ & $\Delta pyrC$, consistently maintained a relative $\Delta pyrC$ abundance of around three quarters of the total population (71-79%, Table 5.1). The strains prevalence in the community could potentially be overestimated if the strains characteristic mutations fell within duplicated genome regions. To account for this possibility, the relative abundance of each strain in the populations was additionally computed by comparing the read coverage of the knocked out genes

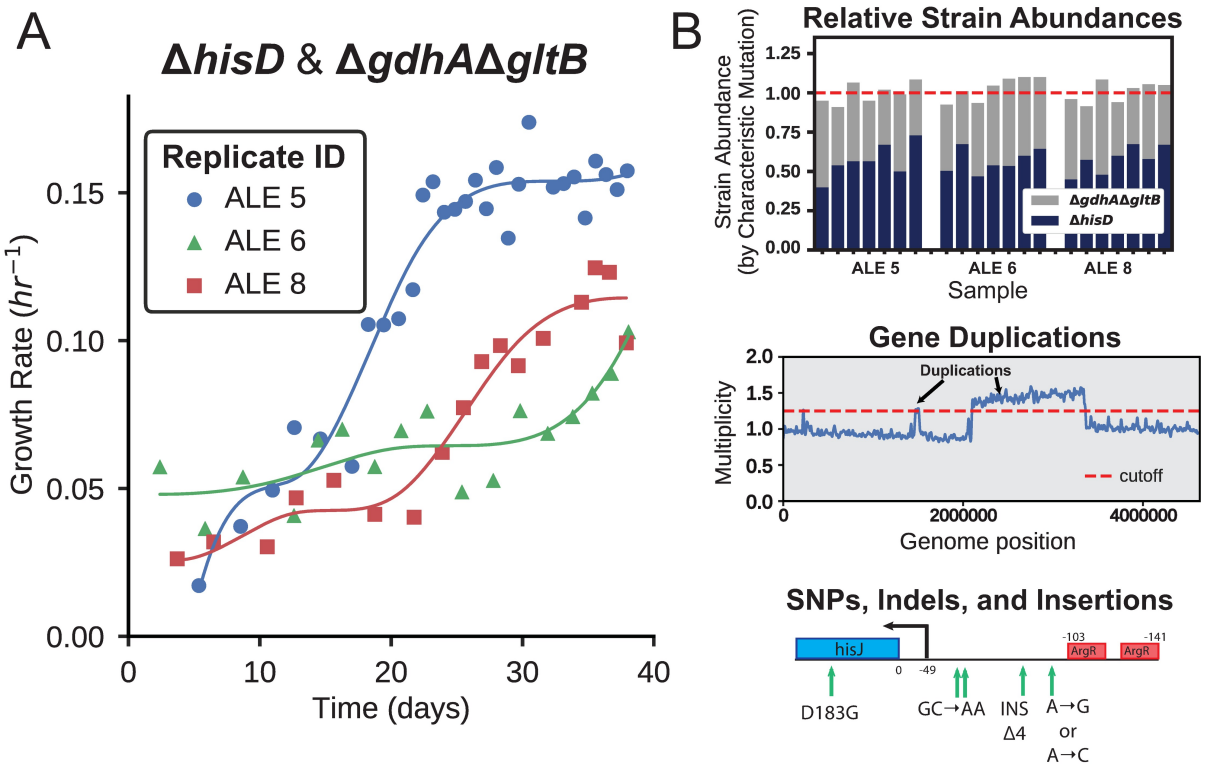

Figure 5.4: Representative example of an adaptive laboratory evolution and its downstream analysis A) E. coli co-cultures were evolved over a 40 day period and the growth rate was periodically measured. Over this time period the co-cultures evolved the capability to establish syntrophic growth, indicated by the improvement in community growth rate. (B) Each of the sampled co-cultures were resequenced at multiple points during the evolution. This information was used to predict the fractional strain abundances of each of the co-culture members (top panel, bars represent the computed fractional abundance of the strains in the legend). Sequencing data was also used to identify duplications in genome regions of the community members (middle panel) and infer causal mutations that improved community fitness (bottom panel). The complete set of ALE growth trajectories, inferred strain abundances, gene region duplications, and mutational analysis can be found in S1 Appendix, S3 Data, S4 Data, and Figs 5-7.

for each mutant relative to the average read depth. This orthogonal method gave predictions consistent with those obtained using the characteristic mutation-based method (Figures E-F in S1 Appendix).

Following the evolutions it was confirmed that all collected ALE endpoint clones remained auxotrophic and had not evolved the ability to grow in glucose M9 minimal media. Given that only the large subunit (*gltB*) of glutamate synthase (catalyzes both glutamate synthase and glutamate dehydrogenase reactions, Table B in S1 Appendix) was knocked out, it was important to verify that the cell could not adapt to restore glutamate synthase functionality using only the small subunit (*gltD*) [63].

5.2.4 Mutations Targeting Metabolite Uptake/Secretion

Several evolutionary strategies were observed in the mutations identified across the ten successfully evolved co-culture lineages (Tables E-G in S1 Appendix). One ubiquitous strategy across all three co-culture pairs, however, was to acquire mutations within or upstream of inner membrane transporter genes. For instance, numerous mutations were observed in every co-culture lineage in the *hisJ* ORF or upstream of the operon containing *hisJ*. This operon contains all four genes (*hisJ, hisM, hisP, hisQ*) composing the histidine ABC uptake complex, the primary mechanism for L-histidine uptake in *E. coli* K-12 MG1655 [64]. Seven mutations were found in the region directly upstream of the operons transcription start site (Figure 5.5). Two of the seven mutations were further observed in more than one co-culture pairing, with a SNP in one position (A $\rightarrow$ G, A $\rightarrow$ C, or A $\rightarrow$ T) at 86 base pairs upstream of *hisJ*) appearing to be particularly beneficial as it was identified in the endpoint clone of every lineage except one (ALE # 5). In three ALEs, a mutation was observed within the *hisJ* ORF that resulted in a substitution of

the L-aspartate residue at the 183 position by glycine. Based on the protein structure, this substitution could disrupt two hydrogen bond interactions with the bound L-histidine ligand in the periplasm [65]. Alternatively, this mutation could function to modulate translation of the *hisJ* operon by altering its mRNA secondary structure. Further mutations were observed that could affect the binding of the ArgR repressor upstream of the *hisJMPQ* operon (Table E in S1 Appendix) or affect the activity of the ArgR protein itself (Table F in S1 Appendix). This included a 121 base pair deletion and a SNP in the ArgR repressor binding site upstream of *hisJ* (Figure 5.5). The mutation in the *argR* ORF consisted of a frameshift insertion early in the coding sequence and persisted throughout ALE # 8, appearing in the Δ*hisD* endpoint clone (Table F in S1 Appendix). ArgR functions to repress L-arginine uptake and biosynthesis as well as repress the L-histidine ABC uptake complex [66] in response to elevated L-arginine concentrations. All of the above mutations could improve L-histidine uptake in the Δ*hisD* strains either by increasing the expression, improving the efficacy, or preventing ArgR mediated repression of the HisJMPQ ABC uptake system.

Beyond improving the uptake of L-histidine in the Δ*hisD* strain, mutations were observed that could improve metabolite uptake in the partnering strain. For instance, in the Δ*hisD* & Δ*gltA*Δ*prpC* co-culture, two of the evolutions acquired mutations in the *kgtP* ORF (a transporter of 2-oxoglutarate [67]) that were also present in the Δ*gltA*Δ*prpC* endpoint clones. These mutations include a substitution of an L-proline residue with an L-glutamine at the 124 position and a substitution of a glycine residue with an L-alanine at the 143 position (Table E in S1 Appendix). These two substitutions occurred in the fourth transmembrane helix in the protein and a cytoplasmic region [68], respectively. These mutations could act to augment the activity of the transporter or modulate its expression by changing the mRNA secondary structure. The

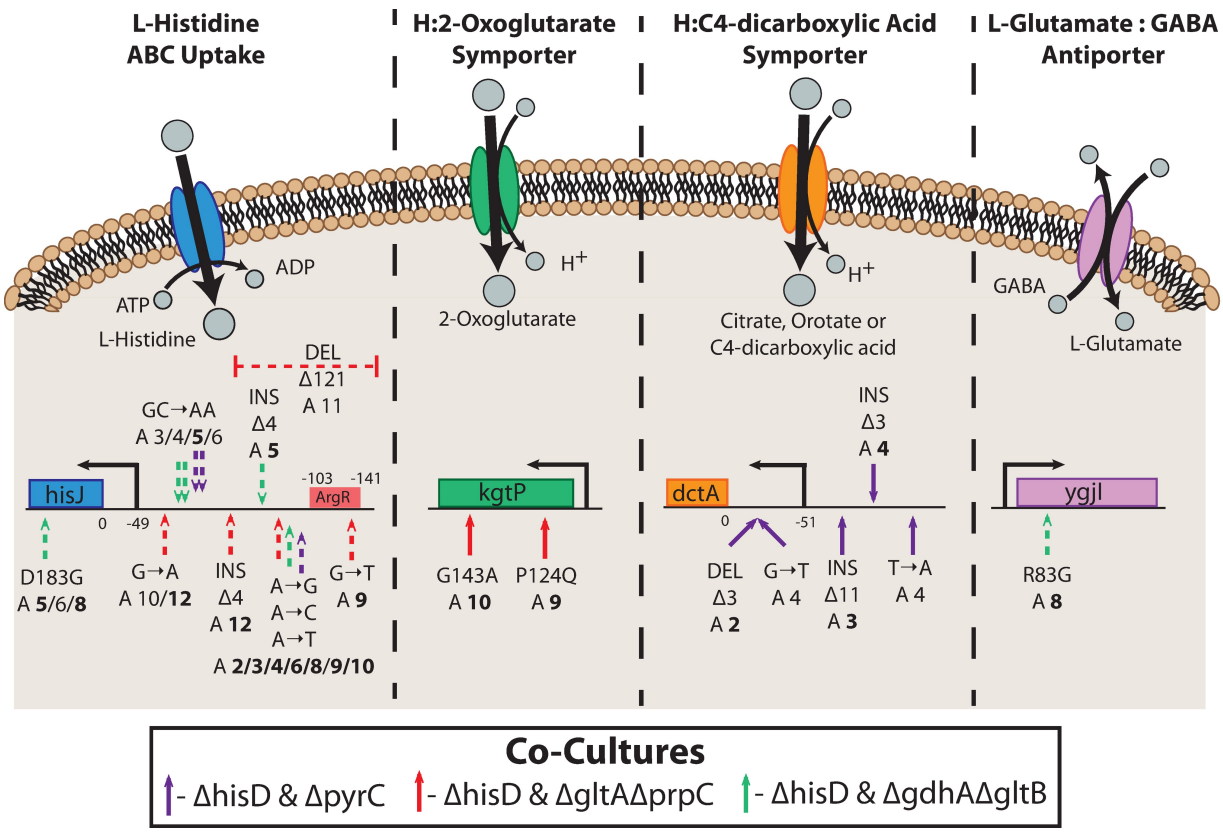

Figure 5.5: Mutations affecting inner membrane metabolite transport. Mutations were observed that possibly affect the activity of four inner membrane transporters. A schematic of the function or putative function of each transporter is shown. Depicted below the schematics are the locations of the observed mutations on the operon encoding each of the enzymatic complexes. For example, all ten evolved Δ*hisD* strain endpoints possessed at least one mutation in or upstream of *hisJ*. This operon includes genes coding for HisJMPQ, the four subunits of an L-histidine ABC uptake system. A depiction of the activity of this complex is shown, in which energy from ATP hydrolysis is used to transport L-histidine into the cytosol from the periplasm. Mutations are indicated on the operon schematics if mutations appear at >10% frequency in more than one flask in an ALE lineage, and ALE numbers are in bold if the mutation appears in the endpoint clone. The mutations indicated with a dashed arrow occured in the Δ*hisD* strain and a solid arrow indicates they occured in Δ*hisD* strains partner MSE strain

mutations further could complement the characteristic mutation upstream of the *kgtP* ORF observed in the starting clone of the Δ*gltA*Δ*prpC* mutant (Table C in S1 Appendix). Both the accumulation of mutations associated with this transporter and the fact that the citrate synthase knockout mutant is computationally predicted to grow in the presence of 2-oxoglutarate suggest

Table 5.2: Metabolite being cross-fed by the Δ*hisD* strain to its partner strain, as inferred from sequencing data.

Pair with Δ*hisD*	Inferred Metabolite	Mutation Evidence	Duplication Evidence
Δ*pyrC*	Orotate	Mutations upstream of *dctA* in Δ*pyrC* strain in all ALEs (Figure 5.5)	Broad duplication in portion of genome containing dctA coding region in all ALEs (Figure J in S1 Appendix, S4 Data)
Δ*gdhA*Δ*gltB*	L-Glutamate	Ale # 8 mutation in *ygjI* ORF in Δ*hisD* strain (Figure 5.5)	ALE # 5/6 targeted duplications in gltJ coding region (Figure 5.7, Figure I in S1 Appendix). ALE # 5 transient duplication in *abgT* coding region (Figure 5.7)
Δ*gltA*Δ*prpC*	2-Oxoglutarate	Starting mutation upstream of *kgtP* in Δ*gltA*Δ*prpC* strain (Table E in S1 Appendix). ALE # 9/10 mutations in *kgtP* ORF in Δ*gltA*Δ*prpC* strain (Figure 5.5)	

that Δ*gltA*Δ*prpC* could be cross-fed 2-oxoglutarate *in vivo* when in co-culture (Table 5.2.4).

For the Δ*hisD* & Δ*pyrC* co-culture, mutations were consistently observed upstream of dctA that could function to better facilitate the uptake of a metabolite being cross-fed from the Δ*hisD* strain to the Δ*pyrC* strain. The three independently evolved lineages each acquired at least one mutation upstream of *dctA*, which were confirmed to be in all Δ*pyrC* endpoint clones (Table G in S1 Appendix). The gene product of *dctA* functions as a proton symporter that can uptake orotate, malate, citrate, and C4-dicarboxylic acids [69] (Figure 5.5). Model simulations of a Δ*pyrC* strain predicted that growth is possible with orotate supplementation, but not with any of the other metabolites known to be transported by the *dctA* gene product. Thus, it is possible these mutations could act to increase the activity of this transporter to allow the Δ*pyrC* strain to more efficiently uptake orotate cross-fed by the Δ*hisD* strain (Table 5.2.4).

Lastly, one lineage of the Δ*hisD* & Δ*gdhA*Δ*gltB* co-culture acquired a SNP in the *ygjI* coding region and was present in the Δ*hisD* endpoint clone. This SNP resulted in a substitution of L-arginine for glycine at position 83, (Table F in S1 Appendix) within a periplasmic region

and one residue prior to a transmembrane helix of the protein [70]. The function of this protein has not been experimentally confirmed, but based on sequence similarity, it is predicted to be a GABA:L-glutamate antiporter [71]. Given that this mutation was seen in the $\Delta hisD$ clone, it is possible that this mutation had the effect of increasing the strains secretion of 4-aminobutyrate (GABA) or L-glutamate by increasing the expression or modulating the activity of YgjI. Such a mutation could improve the community growth rate by facilitating the cross-feeding of either these metabolites to the $\Delta gdhA\Delta gltB$ strain since this strain is predicted to grow when supplemented with either GABA or L-glutamate (Table D in S1 Appendix).

5.2.5 Mutations Targeting Nitrogen Regulation

Knocking out enzymatic reactions in major biosynthetic pathways likely disrupts the homeostatic concentrations of key sensor metabolites, thus activating non beneficial stress responses (e.g., nutrient limited stress responses). The sequencing data was used to elucidate some of the adaptive mechanisms employed by the co-cultures following these pathway disruptions. For example, three frameshift deletions and a SNP resulting in a premature stop codon were observed early in the *glnK* ORF. These mutations were present in three $\Delta gltA\Delta prpC$ endpoint clones and one $\Delta hisD$ endpoint clone from the $\Delta hisD$ & $\Delta gltA\Delta prpC$ co-cultures (Figure 5.6B). GlnK along with GlnB are two nitrogen metabolism regulators with many overlapping functions. Both regulators are uridylated depending on the relative concentrations of 2-oxoglutarate, ATP, and L-glutamate. In conditions of high 2-oxoglutarate and ATP concentrations relative to L-glutamate concentrations, GlnK and GlnB are uridylated causing an increase in glutamine synthetase activity [72]. However, unlike GlnB, when GlnK is not uridylated it binds to the AmtB nitrogen uptake complex, thus reducing AmtBs activity [73]. GlnK is also upregulated by

GlnG of the nitrogen two-component regulatory system in the absence of nitrogen, unlike GlnB [74]. The citrate synthase knockout strain ($\Delta gltA\Delta prpC$) in particular could see a disruption in the homeostatic concentrations of metabolites immediately downstream of the citrate synthase reaction, including 2-oxoglutarate and L-glutamate. This could impair the ability of the cell to respond to sensors of nitrogen excess or limitation and respond with the appropriate global regulatory changes. Removing the activity of this GlnK mediated response system would prevent any detrimental cellular responses (such as inhibition of the AmtB nitrogen uptake complex) due to atypical concentrations of the sensor metabolites within the co-culture strains. No mutations were observed in the alternative nitrogen regulator, GlnB, throughout any of the evolutions.

Mutations found in the $\Delta gdhA\Delta gltB$ strains imply a change in the activity of the two-component nitrogen regulatory system. The $\Delta gdhA\Delta gltB$ strain in all $\Delta hisD$ & $\Delta gdhA\Delta gltB$ lineages acquired mutations in the open reading frame of at least one gene in the two-component nitrogen regulator system, consisting of *glnG* (*ntrC*) and *glnL* (*ntrB*) (Figure 5.6A) [72]. Amino acid substitutions were observed in position 18, 86, and 105 of *glnG* corresponding to the response receiver domain of GlnG (based on protein families [75]), possibly augmenting its ability to interact with GlnL. The endpoint clone of ALE # 5 acquired an amino acid substitution of L-isoleucine to L-serine within a PAS domain of GlnL at position 12. This corresponds to the protein domain where regulatory ligands bind [76] suggesting this mutation could act to augment its activity in response to nitrogen availability. Like the citrate synthase knockout, the $\Delta gdhA\Delta gltB$ strain would likely experience a change in the homeostatic concentrations of metabolites used to sense nitrogen availability. Thus, it can be hypothesized that the mutations observed in the nitrogen two-component regulatory system act to augment the expression of nitrogen uptake and assimilation processes regulated by GlnGL.

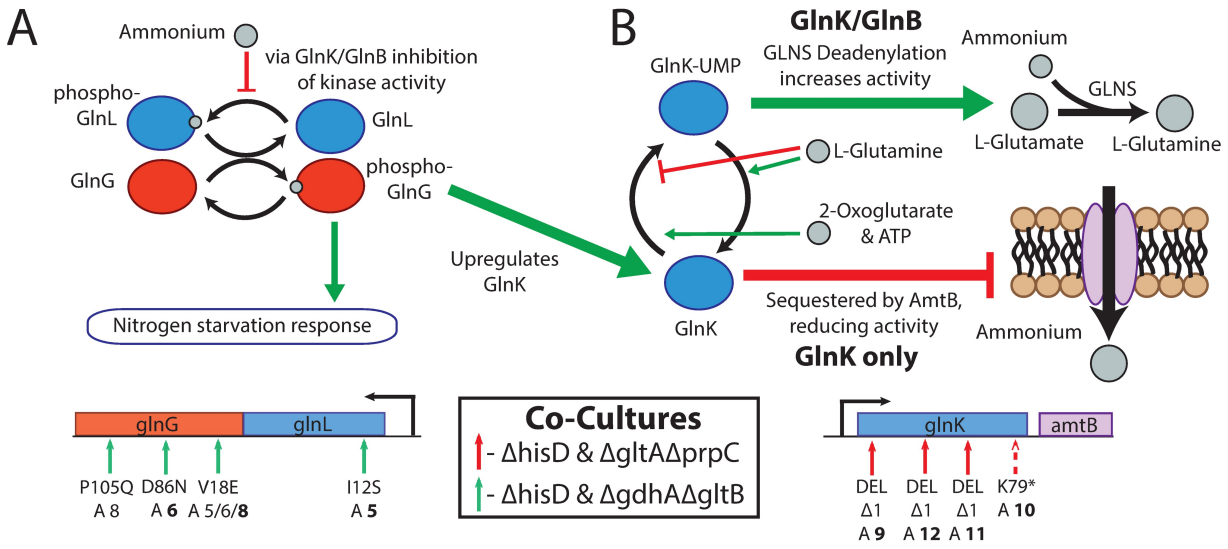

Figure 5.6: Mutations affecting nitrogen regulation. Functions of the mutated genes are summarized, and the location of all mutations are shown on the operon below the schematic. Mutations are shown if they appear at ¿10% frequency in more than one flask in an ALE lineage, and ALE numbers are in bold if the mutation appears in the endpoint clone. The mutations indicated with a dashed arrow occured in the Δ*hisD* strain and a solid arrow if they occured in Δ*hisD* strains partner MSE strain. (A) Mutations were acquired within the open reading frame of both genes comprising the nitrogen sensing two-component regulatory system. Shown in the schematic is the regulatory cascade in which nitrogen concentration is sensed (via GlnK or GlnB) by GlnL. In response to low nitrogen availability GlnL is autophosphorylated resulting in a subsequent transfer of the phosphorus group to GlnG. Phosphorylated GlnG upregulates general functions associated with nitrogen starvation, including increasing GlnK expression [74]. (B) Further, mutations were observed in the ORF of GlnK, one of two nitrogen metabolism regulators, sharing most functions with GlnB. Both genes become uridylylated in response to high concentrations of 2-oxoglutarate and ATP and low concentrations of glutamine, which is an indication of nitrogen limitation. GlnK-UMP can activate GLNS deadenylation, thus increasing its activity. Unlike GlnB, GlnK when in a deuridylylated state (indicative of high nitrogen availability) can be sequestered by the AmtB ammonium transporter reducing its activity [72]

Mutations were also observed targeting osmotic stress responses and nonspecific stress responses. These are summarized in the S1 Appendix.

5.2.6 Genome Duplications Complement Sequence Changes

A complementary adaptive strategy for improving co-culture community growth was to acquire duplications in particular regions of the genome (Figures H-J in S1 Appendix). This evo-

lutionary strategy possibly functioned in some cases to amplify expression of specific transporters to more efficiently uptake a metabolite that can rescue the strains auxotrophy (also observed in [77]). Alternatively, these duplications could function to provide genetic redundancy that increases the likelihood of acquiring mutations in the duplicated region [78, 79]. For example, one of the three $\Delta hisD$ & $\Delta gdhA\Delta gltB$ lineages displayed clear increases in sequencing depth near positions 674-683 kbp and 1,391-1,402 kbp with multiplicities exceeding 15. The former of these coverage peaks contains 9 genes, including the 4 genes composing the GltIJKL L-glutamate/L-aspartate ABC uptake system [80]. The latter peak consisted of 10 genes including the 4 genes in the abgRABT operon, which facilitates the uptake of p-aminobenzoyl-glutamate and its hydrolysis into glutamate and 4-aminobenzoate [81]. This suggests that either L-glutamate, L-aspartate, or p-aminobenzoyl-glutamate could be cross-fed to the $\Delta gdhA\Delta gltB$ strain in vivo. The *abgRABT* duplication, however, was depleted in favor of the *gltIJKL* duplication over the course of the evolution, suggesting L-glutamate or L-aspartate is the preferred cross-feeding metabolite over p-aminobenzoyl-glutamate (Figure 5.7, Table 5.2.4).

While the duplications mentioned above presented clear amplifications in targeted operons, some observed duplications consisted of 100,000s of basepairs and 100s of genes. Further, many of the duplications seen in the populations were not observed in the resequenced endpoint clones. Possible explanations for these observations can be found in the S1 Appendix.

5.2.7 Modeling Community Features of Auxotroph Communities

Community genome-scale models were applied to understand the basic characteristics of the co-culture communities generated in this study. Given the growing appreciation for the role of limited protein availability on governing many fundamental *E. coli* growth characteristics [36],

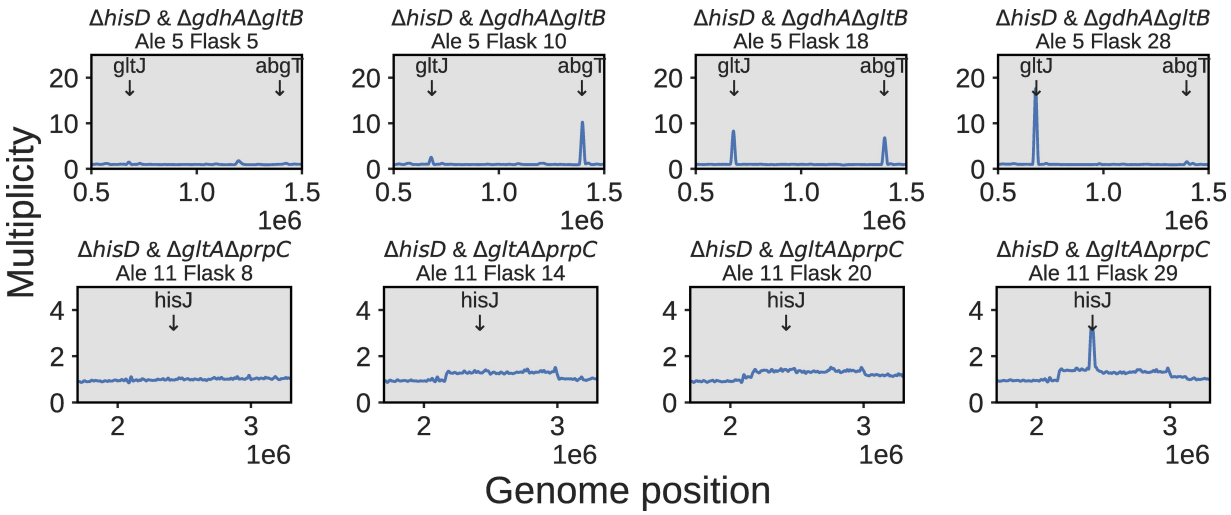

Figure 5.7: Duplication dynamics. The top panel depicts the dynamics of high multiplicity duplications in two transport complexes throughout the course of ALE # 5 of a Δ*hisD* & Δ*gdhA*Δ*gltB* co-culture. A small region containing the *abgT* symporter of p-aminobenzoyl glutamate was duplicated early in the evolution, but later duplications in a region containing gltJ, along with the rest of the genes comprising the GltIKJL L-glutamate/L-aspartate ABC uptake system, became more prevalent. The bottom panel depicts the course of ALE # 11, a Δ*hisD* & Δ*gltA*Δ*prpC* co-culture that initially showed a broad 1 Mbp duplication. By the end of the evolution either a nested duplication emerged in a small genome region containing *hisJ*, along with the rest of the HisJMPQ L-histidine ABC uptake system, or a significant subpopulation emerged containing this duplication

community genome-scale models of metabolism and gene expression (ME-models) were utilized. A new computational approach was also developed, as a community modeling method did not exist that was suitable for studying co-cultures growing in an ALE experiment while also being amenable to ME-models (see Methods).

Using community M- and ME-models, the role of substrate and proteome limitations on basic community characteristics was assessed. To that end, both types of community models were constrained to uptake no more than 5 $\frac{\text{mmol}}{\text{gDW·hr}}$ of glucose and simulated over a fractional Δ*hisD* strain abundance of 0 to 1 (Figure 5.8). The communities were allowed to cross-feed any metabolite that could restore growth in the partner strain (Table D in S1 Appendix). At this low glucose uptake rate the community ME-model was being simulated in the so-called

substrate-limited region [37], meaning that the community growth rate was determined solely by the amount of substrate available. In this region the protein allocation constraints inherent in the ME-model were mostly inactive. In the substrate-limited region, the ME-model and M-model behaved similarly and predicted little change in the community growth rate regardless of the fractional abundance of the strains in co-culture. Alternatively, the community ME-model was again simulated, but with an unlimited amount of glucose available to the in silico community. These simulations therefore occurred in the proteome-limited region of the community ME-model, meaning that the growth rate was determined by limitations in the protein available to carry out their enzymatic functions. When simulating the community ME-model in the proteome-limited region, notable composition-dependent variation in the community growth rate was observed across all fractional strain abundances (Figure 5.8). Metabolite exchange for substrate and proteome-limited ME-models was also observed (Figures M-N in S1 Appendix)

ME-model predictions are dependent on parameters that couple protein abundance to the flux values of the processes or reactions that they catalyze. These are called k_{eff}s and are analogous to the effective *in vivo* turnover rate of an enzyme. Obtaining these values on a genome-scale is a notoriously difficult problem [82], and no gold standard set of k_{eff}s currently exists. To account for uncertainty in these k_{eff}parameters, proteome limited community ME-model simulations were repeated using three different k_{eff}sets, including one set of naive values (all $k_{\text{eff}}\text{s} = 65$) and two sets derived using experimental data (default model [83] and *in vivo* estimated k_{eff}s [84, 85]). All fractional abundance values within 95% of the maximum community growth rate were compiled and represented as a kernel density plot. The computed optimal community compositions (i.e., strain ratios that enabled the fastest computed community growth) showed relatively good agreement with the experimentally inferred community compositions (Figure 5.8).

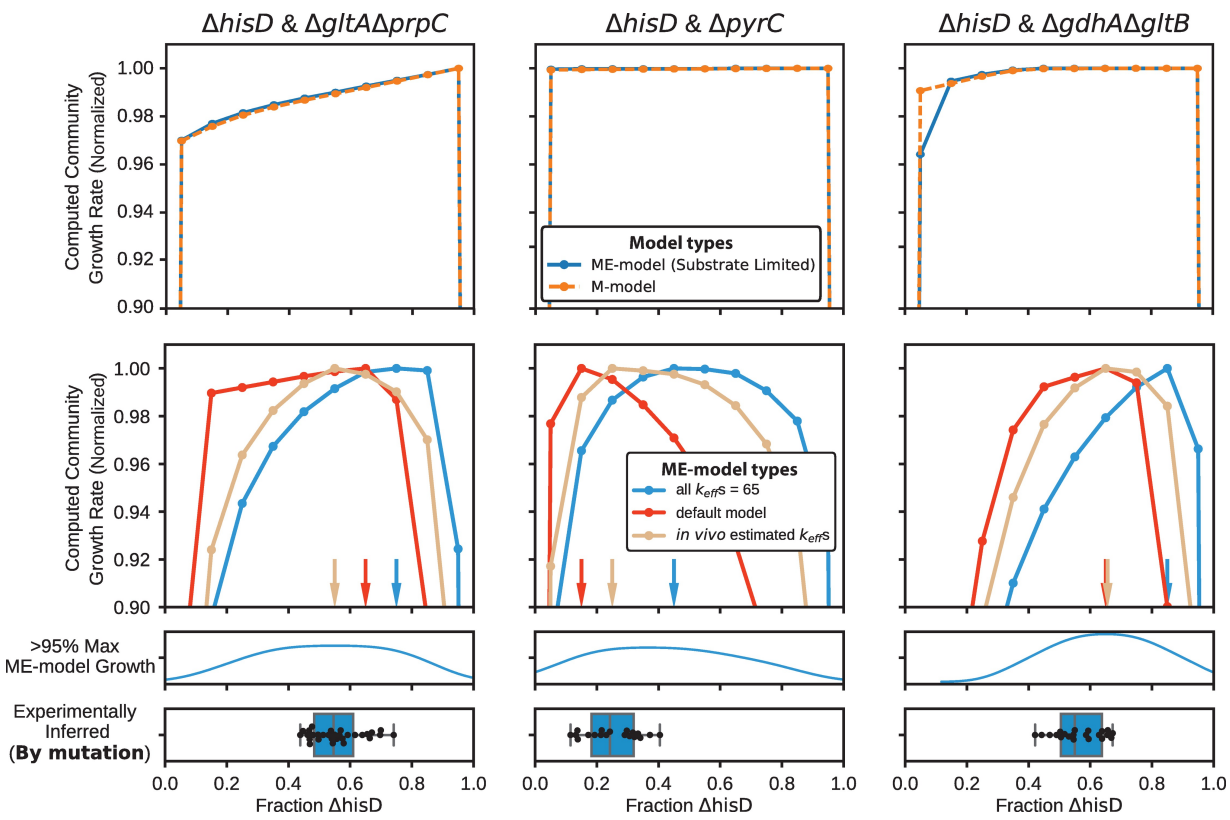

Figure 5.8: Comparison of community M- and ME-models. The simulated growth rates for fractional strain abundances of $\Delta hisD$ ranging from 0 to 1. The top panel shows the community growth rate predictions of the community M-model and the community ME-model simulated in glucose-limited *in silico* conditions. The bottom panel shows growth rate predictions for the community ME-model simulations in glucose excess conditions. The arrows correspond to the fractional abundance that provided the highest computed community growth rate. The fractional abundances with growth rates greater than 95% of the maximum computed value were represented as a kernel density plot. The high density regions of the kernel density plot aligned well with the experimentally inferred community compositions, shown in the box plot.

See the Methods for a description of the three k_{eff}sets.

The ME-modeling analysis suggested that it may be necessary to consider protein allocation when studying co-culture evolutions, therefore necessitating the use of resource allocation models, such as ME-models. The community ME-models thus were used to predict how the community composition could vary depending on basic characteristics of the co-cultures: 1) the identity of the metabolite that is cross-fed or 2) the enzyme efficiency of the community mem-

bers. These simulations predicted that the metabolite being cross-fed within the community could have a sizeable impact on both the community composition and growth rate. This is particularly true for the $\Delta hisD$ & $\Delta gdhA\Delta gltB$ and $\Delta hisD$ & $\Delta gltA\Delta prpC$ simulations which showed that metabolite cross-feeding affected the growth rate and community compositions by as much as 50% (Figure 5.9A).

The strains growing in co-culture *in vivo* each undoubtedly differed in the protein cost required to synthesize the metabolite required by its partner strain. Therefore a proteome efficiency analysis (see Methods) was performed which showed that the computed optimal community compositions (the fractional strain abundance that gave the maximum community growth rate) of all three co-cultures were moderately sensitive to the strains efficiency (Figure 5.9B). The computed optimal community composition was most sensitive when the $\Delta hisD$ strain's metabolite export was less proteome efficient than its partner MSE strain. This observation is not surprising given that the $\Delta hisD$ strain must secrete metabolite(s) to the MSE strain at a much higher flux than the MSE strain to the $\Delta hisD$ strain. Therefore, a decrease in protein efficiency will have a larger impact on the $\Delta hisD$ strain. The community models also unintuitively predicted that, if the $\Delta hisD$ strain required a greater protein investment to produce the metabolite required by the partner strain (i.e., if the $\Delta hisD$ strain was less efficient than its partner), the abundances of the $\Delta hisD$ strain would actually increase in the community.

The optimal predicted community composition for the two above computational analyses shown in Figure 5.9A and B are summarized in Figure 5.9C. The figure shows general agreement between the computed optimal community compositions and the experimentally inferred community composition, even after varying key features of the community simulation (metabolite cross-feeding and protein efficiency). This suggests that community ME-models have the poten-

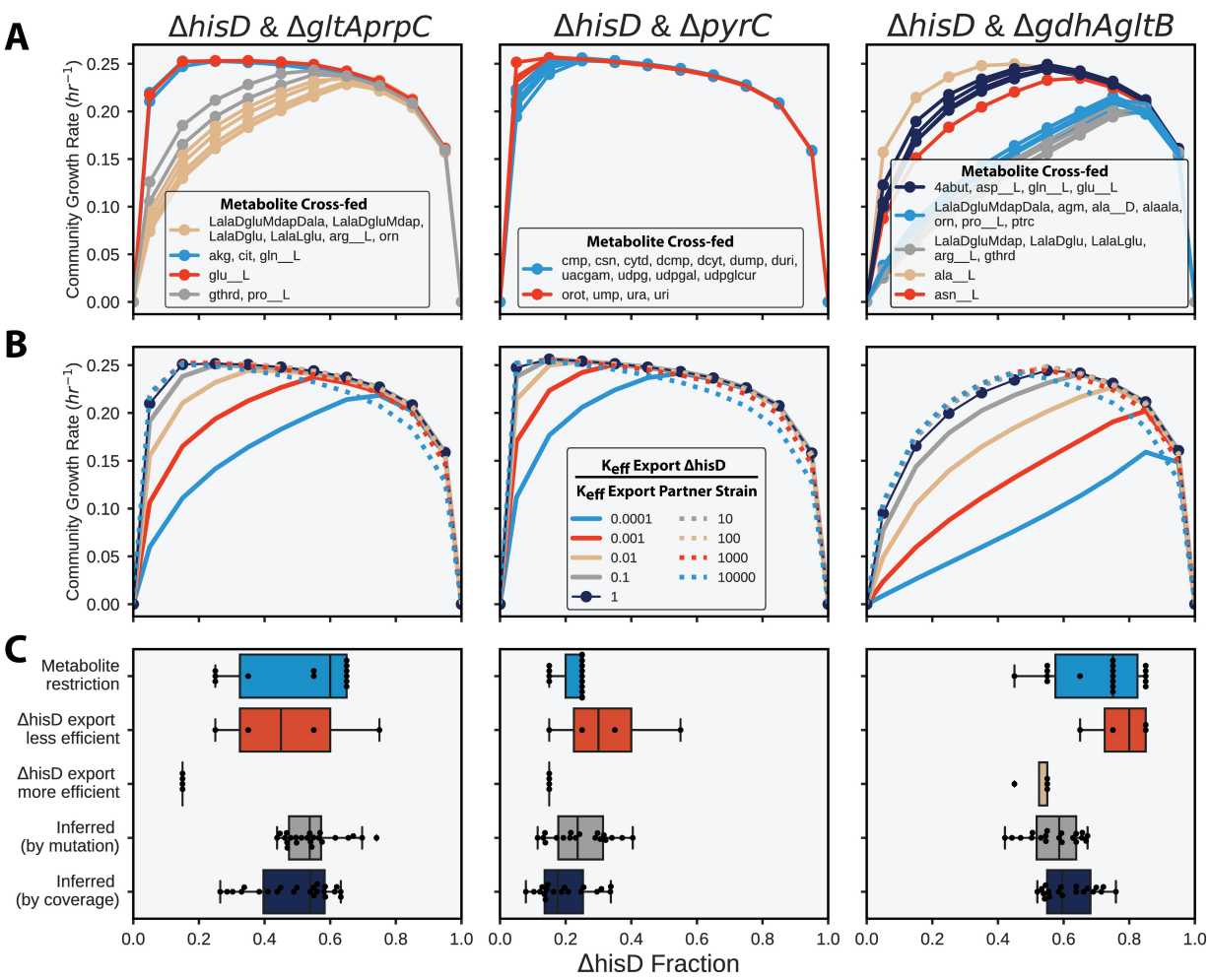

Figure 5.9: Community modeling. Community ME-model predicted growth rates computed with fractional strain abundances of Δ*hisD* ranging from 0 to 1. (A) The effect of metabolite cross-feeding on community structure. Each curve was computed after allowing each of the metabolites in the legend to exclusively be cross-fed to the MSE strain. Curves with identical computationally-predicted optimal strain abundances were grouped and given the same color. (B) The effect of varying the proteome efficiency of metabolite export on community structure (see Methods). The analysis was performed on models constrained to only cross-feed the metabolite that was considered most likely to be cross-fed to the Δ*gltA*Δ*prpC* , Δ*gdhA*Δ*gltB* , and Δ*pyrC* strains in vivo based on the sequencing data (2-oxoglutarate, orotate, and L-glutamate, respectively) (Table 5.2.4). (C) Box plots of experimentally inferred fractional strain abundances for each sample (bottom two rows, gray and dark blue) and the computationally-predicted optimal strain abundances following variation in the cross-feeding metabolite (top row, blue) and in strain proteome efficiency (second and third row, red and yellow).

tial to be useful tools for understanding the behavior of simple communities. The same analysis was performed with the *in vivo* estimated k_{eff}s set of k_{eff}s and showed similar behavior (Figure

O in S1 Appendix).

5.3 Discussion

This work provides genetic-level insight into the adaptation of model-designed nascent syntrophic communities growing cooperatively in suspension. This effort produced a novel algorithm, called OptAux, which was validated against historical auxotrophs and used to predict novel auxotrophic strain designs. OptAux-predicted designs with diverse metabolic deficiencies were co-cultured and community growth was optimized via adaptive laboratory evolution. Sequencing these co-cultures throughout the evolutions gave mutation and community composition information, thus providing insight into mechanisms of cellular cooperation. An additional modeling method was developed to interpret community features and demonstrated the importance of considering protein synthesis cost when studying cooperative communities in the utilized experimental conditions.

OptAux was demonstrated to be a useful tool for designing new types of cellular auxotrophies. Unlike many previously studied auxotrophies, OptAux enabled the prediction of auxotrophs stemming from a diverse set of major metabolic deficiencies. This included the prediction of 4 potential new essential biomass component elimination (EBC) designs and 20 unique major subsystem eliminations (MSE) designs. The OptAux-predicted MSE strains themselves could reveal further community insights if studied in co-culture. Such a combination would likely require a significant degree of metabolic rewiring in each strain to form a viable microbial community, thus probing the alternate evolutionary and cooperative paths such complex combinations could produce. OptAux is also suitable for predicting new auxotrophies in any organism outside of *E. coli* , provided the organism has an existing metabolic reconstruction [86].

Sequencing co-cultures throughout the course of the evolution experiments offered insight into the major adaptive mechanisms underlying the evolution of microbial cooperativity. The observed mutations indicated two major adaptive strategies employed by the strains in co-culture 1) mutating transporters, likely to improve uptake of auxotrophic metabolites (Figure 5.5) and 2) mutating to adapt to homeostatic changes as a result of metabolic disruptions upon imposing gene knockouts (Figure 5.6). The reported transporter mutations could prove useful for metabolic engineering applications, as optimizing the metabolite uptake characteristics of transporters can be an important component of improving the performance of engineering strains [87]. There, however, were no observed mutations, outside of mutations in a predicted GABA antiporter in a $\Delta hisD$ strain, hinting at how the strains were capable of rewiring their intracellular metabolism to supply their partner strain with the required metabolite (i.e., no observed mutations associated with biosynthetic pathways). A future direction of this work could be to further evolve these strains to observe if new mutations appear to enhance metabolite rewiring. Alternatively, it is possible that the co-cultures grew by clumping and employing nanotube-mediated cross-feeding [88], which may be explored using microscopy.

Community ME-models were applied to understand the factors that drive community composition. This was the first community modeling effort to demonstrate the necessity of considering protein allocation when computationally studying community features. Interestingly, some of the studied co-cultures evolved to consistent community compositions that skewed away from a 50:50 strain ratio, a feature the community ME-models were often capable of capturing (Figure 5.8). Additionally, the community ME-models predicted that, if the $\Delta hisD$ strain became less protein efficient at producing the necessary cross-feeding metabolite, the optimal abundance of the $\Delta hisD$ strain in the co-culture would actually increase (Figure 5.9). Though unintuitive,

this prediction is in agreement with a paradox highlighted in a previous computational study of community dynamics [89].

Despite the observed agreement between measured and computed optimal community compositions, this work highlighted the fact that there are a vast number of variables that could potentially influence basic features of simple communities. Experimentally assessing important features such as metabolite cross-feeding and community structure, as touched on here, on a large scale with many different cohorts and combinations is necessary in order to adequately understand the behavior of such bacterial communities. Model-driven design of communities and the use of community ME-models, however, present a more complete computational framework that can be leveraged as a tool to extract more knowledge from such experiments. Further, community ME-models offer a means to probe how factors outside of metabolism (e.g., translation efficiency and proteostasis) could affect community characteristics.

5.4 Methods

5.4.1 Computational Methods

All constraint-based modeling analyses were performed in Python using the COBRApy software package [90] and the iJO1366 metabolic model of *E. coli* K-12 MG1655 [46]. All optimizations were performed using the Gurobi (Gurobi Optimization, Inc., Houston, TX) mixed-integer linear programming (MILP) or linear programming (LP) solver. The community ME-models were solved using the qMINOS solver in quad precision [91, 92]. All scripts and data used to create the presented results can be found at www.github.com/coltonlloyd/optaux.

OptAux Algorithm Formulation For the presented work it was necessary to employ an algorithm capable of finding reaction knockouts that would ensure the target metabolite is computationally essential in the *in silico* growth media for all feasible growth rates. To this end, a new algorithm was written as opposed to implementing a reverse version of RobustKnock (i.e., RobustKnock where the target objective is metabolite uptake instead of secretion). A reverse RobustKnock implementation would optimize the minimum required uptake of a metabolite at the maximum growth rate, thus leading to strain designs that must uptake a high amount of the target metabolite only when approaching the maximum growth rate (Figure A in S1 Appendix). To prevent this computational phenotype with OptAux, the inner problem optimizing for growth rate, which was utilized in RobustKnock, was removed. The growth rate was instead constrained to the 'set_biomass' value, thus forcing the optimization to occur at a predefined growth rate. The constraint was implemented by setting the upper and lower bounds of the biomass objective function to 'set_biomass'. Using relatively low set_biomass values with OptAux ensured the target metabolite would be computationally required for all feasible growth rates. For the simulations ran in this study (S1 Data), the 'set_biomass' value was set to 0.1 hr^{-1}.

An additional constraint was included in OptAux to represent additional metabolites present in the *in silico* media that could alternatively be used for growth, called the 'competing_metabolite_uptake_threshold'. It was applied by finding all metabolites with exchange reactions and a default lower bound of 0 $\frac{\text{mmol}}{\text{gDW}\cdot\text{hr}}$ and increasing the bound to the 'competing_metabolite_uptake_threshold', thus allowing alternative metabolites in the *in silico* media to compete for uptake with the target metabolite. Increasing this threshold ultimately increases the specificity of the OptAux solution (i.e., whether other metabolites could potentially restore growth in addition to the target metabolite). In other words, if other metabolites were present

in the *in silico* media, would the model still be auxotrophic for the target metabolite? If the strain would still be auxotrophic, it can be said to have high specificity; if the strain would not be auxotrophic, it can be said to be non-specific or semi-specific.

The resulting OptAux algorithm is a bilevel MILP (Figure 5.2B) that can be found at www.github.com/coltonlloyd/optaux.

OptAux Simulations The OptAux algorithm was ran for all carbon containing metabolites with exchange reactions in iJO1366. The models default glucose M9 minimal *in silico* media was used for all optimizations with the maximum oxygen uptake set to 20 $\frac{\text{mmol}}{\text{gDW}\cdot\text{hr}}$. For each optimization the target metabolite was selected and the maximum uptake of the metabolite was set to 10 $\frac{\text{mmol}}{\text{gDW}\cdot\text{hr}}$. The model was then reduced by performing flux variability analysis (FVA) on every reaction in the model and setting the upper and lower bounds of each reaction to the FVA results. If FVA computed that no flux could be carried through the reaction, then it was removed from the model. Additionally, reactions were excluded from knockout consideration if they met one of the following criteria: 1) it was an iJO1366 false positive when glucose is the primary carbon substrate [93] 2) it was essential in LB rich media [15] 3) its annotated subsystem was one of the following: Cell Envelope Biosynthesis, Exchange, Inorganic Ion Transport and Metabolism, Lipopolysaccharide Biosynthesis / Recycling, Murein Biosynthesis, Murein Recycling, Transport, Inner Membrane, Transport, Outer Membrane, Transport, Outer Membrane Porin, or tRNA Charging 4) it involved a metabolite with more than 10 carbons 5) it was a spontaneous reaction.

Identifying Gene Mutations and Duplications The FASTQ data from the sequencing samples was filtered and trimmed using AfterQC version 0.9.6 [94]. The quality controlled reads were aligned to the genome sequence of E. coli K-12 BW25113 (CP009273.1) [95] using

Bowtie2 version 2.3.0 [96]. Mutations were identified based on the aligned reads using breseq version 0.32.0b [62]. If the sample was of a co-culture population and not a clone, the predict polymorphism option was used with a frequency cutoff of 0.025. The output of the breseq mutation analysis for all samples can be found in S3 Data and on www.aledb.org [97].

Duplications were found by analyzing the BAM sequence alignment files output from Bowtie using the pysam Python package [98]. Pysam was used to compute the sequencing read depth at each DNA position within the genome sequence. For population samples, a cutoff of 1.25 x coverage fit mean (a measure of average read alignment coverage over the genome) was used. This relatively low threshold was used to account for the varying fractional abundances of the strains in community. A gene was flagged as duplicated in the sample if over 80% of the base pairs in the genes ORF had alignment coverage above the duplication threshold. Duplications found in starting strains were excluded from the duplication analysis. Further, the set of duplicated genes were grouped together if they were located next to each other on the genome. A new group was created if there existed more than five genes separating a duplicated gene from the next duplicated gene in sequence (S4 Data).

Aligned read coverage across the *E. coli* genome is noisy and therefore was filtered before plotting in order to observe its dominant features. This was accomplished by first splitting the coverage vector into 50,000 segments, such that each segment represented 100 base pairs, and the average of the segments was found. Locally weighted scatterplot smoothing (LOWESS) was then applied to the array of concatenated segments using the statsmodel package in python [99]. For the smoothing, 0.5% of all of the segments was used when estimating each coverage value (y-value), and zero residual-based reweightings were performed. The remaining parameters were set to their default.

Calculating Strain Abundances from Sequencing Data The fractional abundances of the strains in co-culture were predicted using two features of the sequencing data obtained from each co-culture sample: 1) the frequency of characteristic mutations of each strain and 2) the read depth of the knocked out genes.

Each of the stains used in this study possessed a unique characteristic mutation (Table C in S1 Appendix), which could be used as a barcode to track the strain. The breseq mutation calling pipeline identified the characteristic mutations of each strain in co-culture and reported the frequency that the mutation was observed. This information was thus used to track the strains presence. For strains with two characteristic mutations (e.g., $\Delta hisD$ and $\Delta gdhA\Delta gltB$) the reported frequency of the genes was averaged and used as a prediction of the relative abundance of that strain. One mutation in particular, an IS element insertion in *yqiC*, which is characteristic of the $\Delta hisD$ strain, was not detected in several samples when $\Delta hisD$ was in co-culture with $\Delta pyrC$. This is likely due to the low abundance of the $\Delta hisD$ strain in that particular population. In those cases, the $\Delta hisD$ strains abundance was predicted using only the frequency of the *lrhA/alaA* intergenic SNP (Figure F in S1 Appendix). For one sample (A10 F23 I1 R1) the sequencing coverage was too low (14.5) and the $\Delta gltA\Delta prpC$ characteristic mutation was not detected. Therefore no relative abundance was computed for this sample.

The second method for computing fractional strain abundances used the sequencing read alignment to compare the coverage of the deleted genes in each strain to the average coverage of the sample. As an example, for a strain paired with the $\Delta hisD$ strain, the average coverage of the base pairs in the hisD ORF divided by the average coverage for that sample, would give an approximation of its relative abundance in the population. As with the characteristic mutation approach, if the two genes were knocked out in the strain, the average coverage of the two genes

was used to make the approximation (Figure E in S1 Appendix).

When reporting the relative abundance predictions (Figs 8 and 9), the computed abundances of each strain were normalized by the sum of the computed abundances of the two strains in co-culture. This ensured that the abundance predictions summed to one. Predictions made using the two described methods showed general agreement (Figure F in S1 Appendix).

Community Modeling A community modeling approach was formulated that was amenable to ME-models and consistent with the characteristics of the ALE experimental design. The ALE experimental design applies a constant growth rate selection pressure by ensuring the cells are maintained in exponential growth phase in nutrient excess media conditions. A consequence of this experimental design when applied to co-culture systems is that the strains in co-culture must be growing at the same growth rate, on average. If this was not the case, one strain would be diluted from the culture or there would be dramatic fluctuations in the community composition, which is not the case (Figure 5.9C). Further, ALE experiments ensure that the culture is well mixed and grown in an excess of nutrients. These experimental conditions are not amenable to most existing community modeling methods. One modeling framework exists to study communities growing in steady state, called SteadyCom [23] (Figure L in S1 Appendix), though this method is not compatible with ME-models. This is due to the ME-models use of non-linear macromolecular coupling constraint expressions that are formulated as a function of growth rate. Therefore, the conversion to aggregate biomass flux used in the SteadyCom formulation cannot be translated directly to ME-models.

Given the above considerations, a multicompartment FBA approach, similar to community FBA [26] was used where the growth rates of the co-culture strains were constrained to be equal. The community model included one compartment for each of the two mutant strains

in co-culture and a shared compartment where each of the strains could exchange metabolites. Further, the fluxes in and out of each strains compartment were scaled by the strains relative abundance to effectively mass balance the different model compartments (Figure K in S1 Appendix), thus allowing the relative abundance of each strain to be imposed as a parameter. For secretion, this was done by multiplying these exchange reactions as follows:

$$\text{metabolite}_{\text{Strain1}} \xrightarrow{\text{V}_{\text{secrete}}} \text{X}_{\text{Strain1}} \cdot \text{metabolite}_{\text{Shared}}$$

and for uptake:

$$\text{X}_{\text{Strain2}} \cdot \text{metabolite}_{\text{Shared}} \xrightarrow{\text{V}_{\text{uptake}}} \text{metabolite}_{\text{Strain2}}$$

where $\text{v}_{\text{secrete}}$ is the secretion flux from strain 1 and has units of $\frac{\text{mmol}}{\text{gDW}_{\text{Strain1}} \cdot \text{hr}}$ and $\text{X}_{\text{Strain1}}$ is the fractional abundance of strain 1 with units of $\frac{\text{gDW}_{\text{Strain1}}}{\text{gDW}_{\text{Community}}}$. Therefore applying this coefficient to $\text{metabolite}_{\text{Shared}}$ gives fluxes in the shared compartment units of $\frac{\text{mmol}}{\text{gDW}_{\text{Community}} \cdot \text{hr}}$. For the subsequent uptake of the shared metabolite by strain 2, the fractional abundance of strain 2 is applied giving units of $\frac{\text{mmol}}{\text{gDW}_{\text{Strain2}} \cdot \text{hr}}$ (Figure K in S1 Appendix).

Using this community modeling approach, the fractional abundance of each strain in the co-culture was implemented as a parameter that could be varied from 0 to 1, which in turn had an impact on the optimal growth state of the community. All presented simulations were ran by optimizing the community growth rate for 10 values of $\text{X}_{\text{Strain1}}$ (abundance of strain 1) ranging from 0.05 to 0.95. For $\text{X}_{\text{Strain1}}$ values of 0 or 1 the community growth rate was assumed to be 0 hr^{-1} given that the co-culture mutants are auxotrophic and require the presence of both mutants to grow. The metabolites that were allowed to be cross-fed in simulation were limited to the set of metabolites that can computationally restore the growth of each auxotroph mutant (Table D in S1 Appendix).

For the community simulations, the iJL1678b [39] ME-model and iJO1366 [46] M-model of *E. coli* K-12 MG1655 were used. For proteome-limited ME-models simulations, the uptake of metabolites in the *in silico* glucose minimal growth media into the shared compartment was left unconstrained, as the ME-model is self limiting [37]. For glucose-limited ME-model and M-model simulations, the maximum glucose uptake into the shared compartment was constrained to 5 $\frac{\text{mmol}}{\text{gDW}_{\text{Community}} \cdot \text{hr}}$. The non-growth associated ATP maintenance and the growth associated ATP maintenance were set to the default parameter values in the model. For ME-model simulations, the RNA degradation constraints were removed to prevent high ATP costs at the low community growth rates. Since the newly formed communities are unoptimized and growing slowly, the ME-models unmodeled/unused protein fraction parameter was set to a higher value, 0.75, for proteome limited simulations (an unmodeled/unused protein fraction of 0.65 was imposed when the *in vivo* estimated k_{eff}s parameter set was used, since these k_{eff}s give a lower maximum growth rate than the other two k_{eff}vectors used) and the default value, 0.36, for glucose-limited simulations. If a metabolite had a reaction to import the metabolite across the inner membrane but no export reaction, a reaction to transport the metabolite from the cytosol to the periplasm was added to the model. For more on the ME-model parameters, refer to [39] and [37].

Three different sets of enzyme turnover rates (k_{eff}s) were used for the community ME-model simulations (Figure 5.8). The first set of k_{eff}s (all k_{eff}s = 65) was imposed by setting all k_{eff}s in iJL1678b-ME equal to 65 s^{-1}. The next set of k_{eff}values (default model) used the default set of k_{eff}parameters included with iJL1678b-ME. Most of the metabolic k_{eff}s in this default set are determined by scaling a median k_{eff}value (65 s^{-1}) by an estimation of the solvent accessible surface area of the enzyme complex that catalyzes the reaction (reference [37] for further description). The default k_{eff}parameters further included a set of 284 metabolic k_{eff}s

derived using proteomics data and a computational method developed in Ebrahim *et al.* [83]. The last k_{eff}set (*in vivo* estimated k_{eff}s) included 234 k_{eff}s from Davidi *et al.* [84] that were estimated using model-computed fluxes and proteomics data. The k_{eff}s not estimated in Davidi *et al.* were imputed using the median estimated k_{eff}value from Davidi *et al.* (6.2 s^{-1}). For all three k_{eff}sets, all non-metabolic processes were assigned a $k_{\text{eff}}$of 65 s^{-1}.

Assessing the influence of metabolite cross-feeding on community composition was performed by restricting the simulation to cross-feed only one of the metabolites computationally predicted to restore growth in the MSE strain. In doing so, the identity of the metabolite being cross-fed could be related to the optimal community growth rate and structure.

To vary the proteome efficiency (k_{eff}) of secreting the cross-fed metabolites, first the exchange reactions into the shared compartment for all potential cross-feeding metabolites were constrained to zero, except the metabolite inferred from the experimental data (Table 5.2.4). Then the enzymatic efficiency of the outer membrane transport process of the inferred cross-feeding metabolite was altered in each strain. The outer membrane transport reactions for each inferred metabolite (i.e., HIStex, GLUtex, AKGtex, and OROTtex for L-histidine, L-glutamate, 2-oxoglutarate, and orotate, respectively) have multiple outer membrane porins capable of facilitating the transport process. To account for this, the k_{eff}kinetic parameter of each porin and reaction was changed by multiplying the default k_{eff}value by the appropriate multiplier. The COBRAme software was used for all ME-model computations [39].

Reproducability All code and data necessary to reproduce the presented results can be found on GitHub at https://github.com/coltonlloyd/OptAux.

5.4.2 Experimental Methods

***E. coli* Strain Construction** All single gene knockouts used in this work were obtained from the Keio collection, a collection of all single gene knockouts in *E. coli* K-12 BW25113 [15]. To generate double gene knockout strains, the second knockout genes were identified from the Keio collection as donor strains, and their P1 phage lysates were generated for the transduction into the receiving single knockout strains. For instance, the $\Delta gltA$ or $\Delta gltB$ knockout strain was a donor strain and the $\Delta prpC$ or $\Delta gdhA$ knockout strain was a receiving strain (Table B in S1 Appendix), respectively. These four knockout strains were used for the construction of the double knockout strains, $\Delta gltA\Delta prpC$ and $\Delta gdhA\Delta gltB$. Each mutant was confirmed not to grow in glucose M9 minimal media without supplementation of an auxotrophic metabolite predicted by the iJO1366 model.

Adaptive Laboratory Evolution Knockout mutants were each initially grown in lysogeny broth from a single colony, then washed 3 times and resuspended in M9-4g/L glucose medium. The washed cells from each knockout mutant preculture were then transferred to fresh M9-4g/L glucose medium and co-cultured with mutants from the partner strain. Cultures were initially inoculated with equal numbers of cells from the two relevant auxotrophs, then serially propagated (100 μL passage volume) in 15 mL (working volume) flasks of M9 minimal medium with 4 g/L glucose, kept at 37 °C and well-mixed for full aeration. An automated system passed the cultures to fresh flasks once they had reached an OD600 of 0.3 (Tecan Sunrise plate reader, equivalent to an OD600 of 1 on a traditional spectrophotometer with a 1 cm path length), a point at which nutrients were still in excess and exponential growth had not started to taper off. Four OD600 measurements were taken from each flask, and the slope of ln(OD600) vs. time determined the

culture growth rates. The timescale of the evolution was reported using the cumulative number of cell divisions, as opposed to generations or days, as mutations occur primarily during cell division events [61].

Resequencing Co-culture population samples were collected at multiple midpoints throughout the ALE and sequenced. Additionally, the starting mutant strains and clones of both mutants isolated from the ALE endpoints were sequenced. The $\Delta hisD$ endpoint clone was unable to be isolated via colony selection for ALE # 11. Genomic DNA of the co-culture populations and mutant clones was isolated using the Macherey-Nagel NucleoSpin tissue kit, following the manufacturers protocol for use with bacterial cells. The quality of isolated genomic DNA was assessed using Nanodrop UV absorbance ratios. DNA was quantified using the Qubit double-stranded DNA (dsDNA) high-sensitivity assay. Paired-end whole genome shotgun sequencing libraries were generated using KAPA HyperPlus kits and run on an Illumina MiSeq platform with a PE600v3 kit or an Illumina HiSeq 4000 with a PE-410-1001 kit for 150bp reads. DNA sequencing data from this study is available on the Sequence Read Archive database (accession no. SRP161177).

Acknowledgements

We thank Richard Szubin for help preparing samples for resequencing and thank Joshua Lerman and Justin Tan for informative discussions. The following authors contributed to this work: CJ Lloyd, ZA King and A.M Feist designed the study. CJ Lloyd. and ZA King. developed OptAux. CJ Lloyd, ZA King, and EJ O'Brien developed the community modeling methods. CJ Lloyd. performed all computation and analysis. Y Hefner and CA Olson constructed all *E. coli*

mutant strains and T Sandberg performed the adaptive laboratory evolution. JG Sanders, RA Salido, K Sanders, C Brennan, G Humphrey, and R Knight conducted DNA sequencing and PV Pheneuf aided with mutation analysis. CJ Lloyd and AM Feist wrote the manuscript and all authors reviewed the text and provided edits.

Funding for this work was provided by the Novo Nordisk Foundation through the Center for Biosustainability at the Technical University of Denmark under Grant NNF10CC1016517. Funding was also provided by the NIH National Institute of Allergy and Infectious Diseases under Grant no. U01AI124316 and the NIH National Institute of General Medical Sciences under Grant no. NIH R01 GM057089. CJL was supported by the National Science Foundation Graduate Research Fellowship under Grant no. DGE-1144086. 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. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Chapter 5 in part is a reprint of material published in: **CJ Lloyd**, ZA King, TE Sandberg, Y Hefner, CA Olson, EJ OBrien, JG Sanders, RA Salido, K Sanders, C Brennan, G Humphrey, R Knight, and AM Feist. 2019 "The genetic basis for adaptation of model-designed syntrophic co-cultures." *PLOS Computational Biology*, 15(3): e1006213. https://doi.org/10.1371/journal.pcbi.1006213. The dissertation author was the primary author.

5.5 References

1. Rittmann, B. E., Hausner, M., Löffler, F., Love, N. G., Muyzer, G., Okabe, S., Oerther, D. B., Peccia, J., Raskin, L. & Wagner, M. A vista for microbial ecology and environmental biotechnology. en. *Environ. Sci. Technol.* **40,** 1096–1103 (Feb. 2006).

2. Minty, J. J., Singer, M. E., Scholz, S. A., Bae, C.-H., Ahn, J.-H., Foster, C. E., Liao, J. C. & Lin, X. N. Design and characterization of synthetic fungal-bacterial consortia for direct production of isobutanol from cellulosic biomass. en. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 14592–14597 (Sept. 2013).

3. Bernstein, H. C. & Carlson, R. P. Microbial Consortia Engineering for Cellular Factories: in vitro to in silico systems. en. *Comput. Struct. Biotechnol. J.* **3,** e201210017 (Dec. 2012).

4. Zuroff, T. R., Xiques, S. B. & Curtis, W. R. Consortia-mediated bioprocessing of cellulose to ethanol with a symbiotic Clostridium phytofermentans/yeast co-culture. en. *Biotechnol. Biofuels* **6,** 59 (Apr. 2013).

5. Briones, A. & Raskin, L. Diversity and dynamics of microbial communities in engineered environments and their implications for process stability. en. *Curr. Opin. Biotechnol.* **14,** 270–276 (June 2003).

6. Zhang, H., Pereira, B., Li, Z. & Stephanopoulos, G. Engineering Escherichia coli coculture systems for the production of biochemical products. en. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 8266–8271 (July 2015).

7. Zhou, K., Qiao, K., Edgar, S. & Stephanopoulos, G. Distributing a metabolic pathway among a microbial consortium enhances production of natural products. en. *Nat. Biotechnol.* **33,** 377–383 (Apr. 2015).

8. Saini, M., Chen, M. H., Chiang, C.-J. & Chao, Y.-P. Potential production platform of n-butanol in Escherichia coli. *Metab. Eng.* **27,** 76–82 (2015).

9. Flint, H. J. The impact of nutrition on the human microbiome. *Nutr. Rev.* **70,** S10–S13 (2012).

10. Magnúsdóttir, S., Heinken, A., Kutt, L., Ravcheev, D. A., Bauer, E., Noronha, A., Greenhalgh, K., Jäger, C., Baginska, J., Wilmes, P., Fleming, R. M. T. & Thiele, I. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. en. *Nat. Biotechnol.* **35,** 81–89 (Jan. 2017).

11. Adamowicz, E. M., Flynn, J., Hunter, R. C. & Harcombe, W. R. Cross-feeding modulates antibiotic tolerance in bacterial communities. en. *ISME J.* (July 2018).

12. Hosoda, K., Suzuki, S., Yamauchi, Y., Shiroguchi, Y., Kashiwagi, A., Ono, N., Mori, K. & Yomo, T. Cooperative adaptation to establishment of a synthetic bacterial mutualism. en. *PLoS One* **6,** e17105 (Feb. 2011).

13. Hosoda, K. & Yomo, T. Designing symbiosis. *Bioeng. Bugs* **2,** 338–341 (2011).

14. Mee, M. T., Collins, J. J., Church, G. M. & Wang, H. H. Syntrophic exchange in synthetic microbial communities. en. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E2149–56 (May 2014).

15. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L. & Mori, H. Construction of Escherichia coli K-12 in-frame,

single-gene knockout mutants: the Keio collection. en. *Mol. Syst. Biol.* **2,** 2006.0008 (Feb. 2006).

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

17. Zhang, X. & Reed, J. L. Adaptive evolution of synthetic cooperating communities improves growth performance. en. *PLoS One* **9,** e108297 (Oct. 2014).

18. Marchal, M., Goldschmidt, F., Derksen-Müller, S. N., Panke, S., Ackermann, M. & Johnson, D. R. A passive mutualistic interaction promotes the evolution of spatial structure within microbial populations. en. *BMC Evol. Biol.* **17,** 106 (Apr. 2017).

19. Summers, Z. M., Fogarty, H. E., Leang, C., Franks, A. E., Malvankar, N. S. & Lovley, D. R. Direct exchange of electrons within aggregates of an evolved syntrophic coculture of anaerobic bacteria. en. *Science* **330,** 1413–1415 (Dec. 2010).

20. Hillesland, K. L., Lim, S., Flowers, J. J., Turkarslan, S., Pinel, N., Zane, G. M., Elliott, N., Qin, Y., Wu, L., Baliga, N. S., Zhou, J., Wall, J. D. & Stahl, D. A. Erosion of functional independence early in the evolution of a microbial mutualism. en. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 14822–14827 (Oct. 2014).

21. Zomorrodi, A. R. & Segrè, D. Synthetic Ecology of Microbes: Mathematical Models and Applications. en. *J. Mol. Biol.* **428,** 837–861 (Feb. 2016).

22. Perez-Garcia, O., Lear, G. & Singhal, N. Metabolic Network Modeling of Microbial Interactions in Natural and Engineered Environmental Systems. en. *Front. Microbiol.* **7,** 673 (May 2016).

23. Chan, S. H. J., Simons, M. N. & Maranas, C. D. SteadyCom: Predicting microbial abundances while ensuring community stability. en. *PLoS Comput. Biol.* **13,** e1005539 (May 2017).

24. Klitgord, N. & Segrè, D. Environments that induce synthetic microbial ecosystems. en. *PLoS Comput. Biol.* **6,** e1001002 (Nov. 2010).

25. Freilich, S., Zarecki, R., Eilam, O., Segal, E. S., Henry, C. S., Kupiec, M., Gophna, U., Sharan, R. & Ruppin, E. Competitive and cooperative metabolic interactions in bacterial communities. en. *Nat. Commun.* **2,** 589 (Dec. 2011).

26. Khandelwal, R. A., Olivier, B. G., Röling, W. F. M., Teusink, B. & Bruggeman, F. J. Community flux balance analysis for microbial consortia at balanced growth. en. *PLoS One* **8,** e64567 (May 2013).

27. Chiu, H.-C., Levy, R. & Borenstein, E. Emergent biosynthetic capacity in simple microbial communities. en. *PLoS Comput. Biol.* **10,** e1003695 (July 2014).

28. Harcombe, W. R., Riehl, W. J., Dukovski, I., Granger, B. R., Betts, A., Lang, A. H., Bonilla, G., Kar, A., Leiby, N., Mehta, P., Marx, C. J. & Segrè, D. Metabolic resource allocation in

individual microbes determines ecosystem interactions and spatial dynamics. en. *Cell Rep.* **7,** 1104–1115 (May 2014).

29. Zomorrodi, A. R. & Segrè, D. Genome-driven evolutionary game theory helps understand the rise of metabolic interdependencies in microbial communities. en. *Nat. Commun.* **8,** 1563 (Nov. 2017).

30. Zomorrodi, A. R. & Maranas, C. D. OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. en. *PLoS Comput. Biol.* **8,** e1002363 (Feb. 2012).

31. Zomorrodi, A. R., Islam, M. M. & Maranas, C. D. d-OptCom: Dynamic multi-level and multi-objective metabolic modeling of microbial communities. en. *ACS Synth. Biol.* **3,** 247–257 (Apr. 2014).

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

33. Biliouris, K., Babson, D., Schmidt-Dannert, C. & Kaznessis, Y. N. Stochastic simulations of a synthetic bacteria-yeast ecosystem. *BMC Syst. Biol.* **6,** 58 (2012).

34. Oliveira, N. M., Niehus, R. & Foster, K. R. Evolutionary limits to cooperation in microbial communities. en. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 17941–17946 (Dec. 2014).

35. Germerodt, S., Bohl, K., Lück, A., Pande, S., Schröter, A., Kaleta, C., Schuster, S. & Kost, C. Pervasive Selection for Cooperative Cross-Feeding in Bacterial Communities. en. *PLoS Comput. Biol.* **12,** e1004986 (June 2016).

36. Basan, M., Hui, S., Okano, H., Zhang, Z., Shen, Y., Williamson, J. R. & Hwa, T. Overflow metabolism in Escherichia coli results from efficient proteome allocation. en. *Nature* **528,** 99–104 (Dec. 2015).

37. Omiddle dotBrien, E. J., middle dot, 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,** 693–693 (2014).

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

39. Lloyd, C. J., Ebrahim, A., Yang, L., King, Z. A., Catoiu, E., O'Brien, E. J., Liu, J. K. & Palsson, B. O. COBRAme: A computational framework for genome-scale models of metabolism and gene expression. en. *PLoS Comput. Biol.* **14,** e1006302 (July 2018).

40. Wilson, M. & Lindow, S. E. Coexistence among Epiphytic Bacterial Populations Mediated through Nutritional Resource Partitioning. en. *Appl. Environ. Microbiol.* **60,** 4468–4477 (Dec. 1994).

41. Zhao, Q., Segre, D. & Paschalidisy, I. C. *Optimal allocation of metabolic functions among organisms in a microbial ecosystem* in *2016 IEEE 55th Conference on Decision and Control (CDC)* (2016).

42. Teague, B. P. & Weiss, R. SYNTHETIC BIOLOGY. Synthetic communities, the sum of parts. en. *Science* **349,** 924–925 (Aug. 2015).

43. Polz, M. F. & Cordero, O. X. Bacterial evolution: Genomics of metabolic trade-offs. en. *Nat Microbiol* **1,** 16181 (Oct. 2016).

44. Feist, A. M., Zielinski, D. C., Orth, J. D., Schellenberger, J., Herrgard, M. J. & Palsson, B. Ø. Model-driven evaluation of the production potential for growth-coupled products of Escherichia coli. *Metab. Eng.* **12,** 173–186 (2010).

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

46. Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M. & Palsson, B. Ø. A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011. en. *Mol. Syst. Biol.* **7,** 535 (Oct. 2011).

47. Monk, J. M., Lloyd, C. J., Brunk, E., Mih, N., Sastry, A., King, Z., Takeuchi, R., Nomura, W., Zhang, Z., Mori, H., Feist, A. M. & Palsson, B. O. iML1515, a knowledgebase that computes Escherichia coli traits. en. *Nat. Biotechnol.* **35,** 904–908 (Oct. 2017).

48. Zengler, K. & Zaramela, L. S. The social network of microorganisms — how auxotrophies shape complex communities. *Nat. Rev. Microbiol.* (2018).

49. Fotheringham, I. G., Dacey, S. A., Taylor, P. P., Smith, T. J., Hunter, M. G., Finlay, M. E., Primrose, S. B., Parker, D. M. & Edwards, R. M. The cloning and sequence analysis of the aspC and tyrB genes from Escherichia coli K12. Comparison of the primary structures of the aspartate aminotransferase and aromatic aminotransferase of E. coli with those of the pig aspartate aminotransferase isoenzymes. en. *Biochem. J* **234,** 593–604 (Mar. 1986).

50. Thèze, J., Margarita, D., Cohen, G. N., Borne, F. & Patte, J. C. Mapping of the structural genes of the three aspartokinases and of the two homoserine dehydrogenases of Escherichia coli K-12. en. *J. Bacteriol.* **117,** 133–143 (Jan. 1974).

51. Glansdorff, N. TOPOGRAPHY OF COTRANSDUCIBLE ARGININE MUTATIONS IN ESCHERICHIA COLI K-12. en. *Genetics* **51,** 167–179 (Feb. 1965).

52. Jones-Mortimer, M. C. Positive control of sulphate reduction in Escherichia coli. Isolation, characterization and mapping oc cysteineless mutants of E. coli K12. en. *Biochem. J* **110,** 589–595 (Dec. 1968).

53. Sirko, A. E., Zatyka, M. & Hulanicka, M. D. Identification of the Escherichia coli cysM gene encoding O-acetylserine sulphydrylase B by cloning with mini-Mu-lac containing a plasmid replicon. en. *J. Gen. Microbiol.* **133,** 2719–2725 (Oct. 1987).

54. Somers, J. M., Amzallag, A. & Middleton, R. B. Genetic fine structure of the leucine operon of Escherichia coli K-12. en. *J. Bacteriol.* **113,** 1268–1272 (Mar. 1973).

55. Wild, J., Hennig, J., Lobocka, M., Walczak, W. & Kłopotowski, T. Identification of the dadX gene coding for the predominant isozyme of alanine racemase in Escherichia coli K12. en. *Mol. Gen. Genet.* **198,** 315–322 (1985).

56. Lee, Y.-J. & Cho, J.-Y. Genetic manipulation of a primary metabolic pathway for L-ornithine production in Escherichia coli. en. *Biotechnol. Lett.* **28,** 1849–1856 (Nov. 2006).

57. Felton, J., Michaelis, S. & Wright, A. Mutations in two unlinked genes are required to produce asparagine auxotrophy in Escherichia coli. en. *J. Bacteriol.* **142,** 221–228 (Apr. 1980).

58. Vander Horn, P. B., Backstrom, A. D., Stewart, V. & Begley, T. P. Structural genes for thiamine biosynthetic enzymes (thiCEFGH) in Escherichia coli K-12. en. *J. Bacteriol.* **175,** 982–992 (Feb. 1993).

59. Cronan Jr, J. E., Littel, K. J. & Jackowski, S. Genetic and biochemical analyses of pantothenate biosynthesis in Escherichia coli and Salmonella typhimurium. en. *J. Bacteriol.* **149,** 916–922 (Mar. 1982).

60. Yang, Y., Tsui, H. C., Man, T. K. & Winkler, M. E. Identification and function of the pdxY gene, which encodes a novel pyridoxal kinase involved in the salvage pathway of pyridoxal 5'-phosphate biosynthesis in Escherichia coli K-12. en. *J. Bacteriol.* **180,** 1814–1821 (Apr. 1998).

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

62. Deatherage, D. E. & Barrick, J. E. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. en. *Methods Mol. Biol.* **1151,** 165–188 (2014).

63. Mantsala, P. & Zalkin, H. Active subunits of Escherichia coli glutamate synthase. en. *J. Bacteriol.* **126,** 539–541 (Apr. 1976).

64. Ardeshir, F. & Ames, G. F. Cloning of the histidine transport genes from Salmonella typhimurium and characterization of an analogous transport system in Escherichia coli. en. *J. Supramol. Struct.* **13,** 117–130 (1980).

65. Yao, N., Trakhanov, S. & Quiocho, F. A. Refined 1.89-A structure of the histidine-binding protein complexed with histidine and its relationship with many other active transport/chemosensory proteins. en. *Biochemistry* **33,** 4769–4779 (Apr. 1994).

66. Caldara, M., Charlier, D. & Cunin, R. The arginine regulon of Escherichia coli: whole-system transcriptome analysis discovers new genes and provides an integrated view of arginine regulation. en. *Microbiology* **152,** 3343–3354 (Nov. 2006).

67. Seol, W. & Shatkin, A. J. Escherichia coli alpha-ketoglutarate permease is a constitutively expressed proton symporter. en. *J. Biol. Chem.* **267,** 6409–6413 (Mar. 1992).

68. Seol, W. & Shatkin, A. J. Membrane topology model of Escherichia coli alpha-ketoglutarate permease by phoA fusion analysis. en. *J. Bacteriol.* **175,** 565–567 (Jan. 1993).

69. Baker, K. E., Ditullio, K. P., Neuhard, J. & Kelln, R. A. Utilization of orotate as a pyrimidine source by Salmonella typhimurium and Escherichia coli requires the dicarboxylate transport protein encoded by dctA. en. *J. Bacteriol.* **178,** 7099–7105 (Dec. 1996).

70. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* (2018).

71. Riley, M., Abe, T., Arnaud, M. B., Berlyn, M. K. B., Blattner, F. R., Chaudhuri, R. R., Glasner, J. D., Horiuchi, T., Keseler, I. M., Kosuge, T., Mori, H., Perna, N. T., Plunkett 3rd, G., Rudd, K. E., Serres, M. H., Thomas, G. H., Thomson, N. R., Wishart, D. & Wanner, B. L. Escherichia coli K-12: a cooperatively developed annotation snapshot–2005. en. *Nucleic Acids Res.* **34,** 1–9 (Jan. 2006).

72. Van Heeswijk, W. C., Westerhoff, H. V. & Boogerd, F. C. Nitrogen assimilation in Escherichia coli: putting molecular data into a systems perspective. en. *Microbiol. Mol. Biol. Rev.* **77,** 628–695 (Dec. 2013).

73. Javelle, A., Severi, E., Thornton, J. & Merrick, M. Ammonium sensing in Escherichia coli. Role of the ammonium transporter AmtB and AmtB-GlnK complex formation. en. *J. Biol. Chem.* **279,** 8530–8538 (Mar. 2004).

74. Van Heeswijk, W. C., Hoving, S., Molenaar, D., Stegeman, B., Kahn, D. & Westerhoff, H. V. An alternative PII protein in the regulation of glutamine synthetase in Escherichia coli. en. *Mol. Microbiol.* **21,** 133–146 (July 1996).

75. Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J. & Bateman, A. The Pfam protein families database: towards a more sustainable future. en. *Nucleic Acids Res.* **44,** D279–85 (Jan. 2016).

76. Song, Y., Peisach, D., Pioszak, A. A., Xu, Z. & Ninfa, A. J. Crystal structure of the C-terminal domain of the two-component system transmitter protein nitrogen regulator II (NRII; NtrB), regulator of nitrogen assimilation in Escherichia coli. en. *Biochemistry* **43,** 6670–6678 (June 2004).

77. Brown, C. J., Todd, K. M. & Rosenzweig, R. F. Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Mol. Biol. Evol.* **15,** 931–942 (1998).

78. Slack, A., Thornton, P. C., Magner, D. B., Rosenberg, S. M. & Hastings, P. J. On the mechanism of gene amplification induced under stress in Escherichia coli. en. *PLoS Genet.* **2,** e48 (Apr. 2006).

79. Serres, M. H., Kerr, A. R. W., McCormack, T. J. & Riley, M. Evolution by leaps: gene duplication in bacteria. en. *Biol. Direct* **4,** 46 (Nov. 2009).

80. Wallace, B., Yang, Y. J., Hong, J. S. & Lum, D. Cloning and sequencing of a gene encoding a glutamate and aspartate carrier of Escherichia coli K-12. en. *J. Bacteriol.* **172,** 3214–3220 (June 1990).

81. Carter, E. L., Jager, L., Gardner, L., Hall, C. C., Willis, S. & Green, J. M. Escherichia coli abg genes enable uptake and cleavage of the folate catabolite p-aminobenzoyl-glutamate. en. *J. Bacteriol.* **189,** 3329–3334 (May 2007).

82. Nilsson, A., Nielsen, J. & Palsson, B. O. Metabolic Models of Protein Allocation Call for the Kinetome. en. *Cell Syst* **5,** 538–541 (Dec. 2017).

83. Ebrahim, A., Brunk, E., Tan, J., O'Brien, E. J., Kim, D., Szubin, R., Lerman, J. A., Lechner, A., Sastry, A., Bordbar, A., Feist, A. M. & Palsson, B. O. Multi-omic data integration enables discovery of hidden biological regularities. en. *Nat. Commun.* **7,** 13091 (Oct. 2016).

84. Davidi, D., Noor, E., Liebermeister, W., Bar-Even, A., Flamholz, A., Tummler, K., Barenholz, U., Goldenfeld, M., Shlomi, T. & Milo, R. Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro kcat measurements. en. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 3401–3406 (Mar. 2016).

85. Heckmann, D., Lloyd, C. J., Mih, N., Ha, Y., Zielinski, D. C., Haiman, Z. B., Desouki, A. A., Lercher, M. J. & Palsson, B. O. Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. en. *Nat. Commun.* **9,** 5252 (Dec. 2018).

86. Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L. & Palsson, B. Ø. Reconstruction of biochemical networks in microorganisms. en. *Nat. Rev. Microbiol.* **7,** 129–143 (Feb. 2009).

87. Kell, D. B., Swainston, N., Pir, P. & Oliver, S. G. Membrane transporter engineering in industrial biotechnology and whole cell biocatalysis. en. *Trends Biotechnol.* **33,** 237–246 (Apr. 2015).

88. Shitut, S., Ahsendorf, T., Pande, S., Egbert, M. & Kost, C. *Nanotube-mediated cross-feeding couples the metabolism of interacting bacterial cells* 2017.

89. Kallus, Y., Miller, J. H. & Libby, E. Paradoxes in leaky microbial trade. en. *Nat. Commun.* **8,** 1361 (Nov. 2017).

90. Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. en. *BMC Syst. Biol.* **7,** 74 (Aug. 2013).

91. Yang, L., Ma, D., Ebrahim, A., Lloyd, C. J., Saunders, M. A. & Palsson, B. O. solveME: fast and reliable solution of nonlinear ME models. *BMC Bioinformatics* **17,** 391 (2016).

92. Ma, D., Yang, L., Fleming, R. M. T., Thiele, I., Palsson, B. O. & Saunders, M. A. Reliable and efficient solution of genome-scale models of Metabolism and macromolecular Expression. en. *Sci. Rep.* **7,** 40863 (Jan. 2017).

93. Orth, J. D. & Palsson, B. Gap-filling analysis of the iJO1366 Escherichia coli metabolic network reconstruction for discovery of metabolic functions. en. *BMC Syst. Biol.* **6,** 30 (May 2012).

94. Chen, S., Huang, T., Zhou, Y., Han, Y., Xu, M. & Gu, J. AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. en. *BMC Bioinformatics* **18,** 80 (Mar. 2017).

95. Grenier, F., Matteau, D., Baby, V. & Rodrigue, S. Complete Genome Sequence of Escherichia coli BW25113. en. *Genome Announc.* **2** (Oct. 2014).

96. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. en. *Nat. Methods* **9,** 357–359 (Mar. 2012).

97. Phaneuf, P. V., Gosting, D., Palsson, B. O. & Feist, A. M. ALEdb 1.0: a database of mutations from adaptive laboratory evolution experimentation. en. *Nucleic Acids Res.* (Oct. 2018).

98. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. en. *Bioinformatics* **25,** 2078–2079 (Aug. 2009).

99. Seabold, S. & Perktold, J. *Statsmodels: Econometric and statistical modeling with python* in *Proceedings of the 9th Python in Science Conference* **57** (2010), 61.

Chapter 6

Conclusions

The whole genome sequences that appeared in the mid to late 1990s ushered in the genome era for microbiology. The availability of all the genetic elements on a genome sequence, in turn, ushered in the era of microbial systems biology where the simultaneous function of all the gene products is considered. The basic paradigms of microbial systems biology that have solidified over the past 20 years are manifested in two fundamental steps: first in the reconstruction of all the biochemical, genetic, genomic, and structural information available about a target strain in a formalized fashion, called a reconstruction, and second, in the development of a mathematical framework that allows computational interrogation reconstruction properties, with special emphasis on phenotypic traits. These two steps together comprise *de facto* a multivariate mechanistic genotype-phenotype relationship.

The first chapter of this dissertation "The Promise of Systems Biology," describes the way metabolic and proteome synthetic networks have been reconstructed and characterized. Currently, protein structures, proteostasis, and stress responses are being reconstructed at the genome-scale. We are thus approaching the point in time where we can explicitly compute

over 90% of the proteome by mass and directly relate the outcome of such computations to phenotypic functions. These genome-scale models have invaluable utility to predict and guide hypothesis-driven experimental designs from a genome-scale standpoint and to compare properties of different strains. The work presented in this dissertation is the next step toward enabling the use of such models of protein expression to study the metabolic capabilities of new organisms and metabolic systems.

The second chapter of this dissertation "A Computational Framework to Empower ME-model Development," describes an important landmark toward the goal of empowering the systems biology community to reconstruct and apply genome-scale models of metabolism and gene expression (ME-models). These models represent vast knowledge-bases created by bringing together decades of biochemical research. As a result, these models are complex and often times difficult to understand. COBRAme was constructed to alleviate many of these difficulties and optimize the way ME-models are solved and reconstructed. With its release, ME-models for the first time have a documented computational framework to guide the use of such models. We anticipate that this software will mark an inflection point in the dissemination of ME-models throughout the scientific community.

At the heart of a ME-model is the underlying metabolic reconstruction (M-model). Therefore to bolster future ME-model reconstructions of *E. coli* K-12 MG1655, *i*ML1515 is presented in chapter three, "The next generation *E. coli* M-model". In addition to representing the most highly validated, comprehensive metabolic reconstruction of *E. coli* K-12 MG1655 to date, *i*ML1515 provides a rich knowledge-base that can be leveraged to study *E. coli* K-12 MG1655 in the context of the *E. coli* species. This was demonstrated by assessing protein sequence and structural variation among 1000 sequenced *E. coli* strains. *i*ML1515 was also used as a chassis

for generating strain specific models of *E. coli* clinical isolates and metagenome samples of infant microbiomes.

Chapter four, "Revealing the intricate relationships between proteome cofactor requirements and growth environments," applies the *E. coli* ME-model to study how the growth environment of *E. coli* can influence its enzyme cofactor use and its evolution of auxotrophies. This work extends some of the methods that have been used to assess the catabolic capabilities of the *E. coli* species to now assess the *growth capabilities* of the *E. coli* species. The growth capabilities are shown to be influenced by factors such as cofactor availability, etc.

Beyond influencing the growth capabilities of individual organisms, the evolutionary pressure for cells to optimize their proteome is thought to impact the evolution of microbial communities. Chapter five, "The effect of protein allocation on bacterial community composition," tested this hypothesis by pairing model-designed auxotrophic *E. coli* strains and evolving them via adaptive laboratory evolution in co-culture *in vivo*. In order to form a viable community, the strains in co-culture had to adapt the ability to cross-feed the metabolite required by their partner strain in order to grow. Sequencing data was collected to observe the adaptive strategies employed by the strains to form viable communities. Computations from a novel co-culture ME-model was validated against the experimental predictions of the relative strain abundance of each strain in co-culture. The community ME-model was used to hypothesize how changes in the protein efficiency of the strains in co-culture could impact overall community characteristics (e.g., relative strain abundances). This modeling technology could be applied to better design simple communities often used for industrial biotechnology and–increasingly–for human health applications. Future work should be performed to isolate individual strains from the co-culture communities and measure the metabolites each strain is cross-feeding in order for the community

to grow.

Constraint-based reconstruction and analysis (COBRA) methods have become widely used tools for biotechnology and basic science in both academic and industrial laboratories. They have been demonstrated as an effective approach to predict the catabolic capabilities of an organism from its genome sequence. A new generation of COBRA models and methods are now being developed–encompassing many biological processes and simulation strategies–and next-generation models enable new types of predictions. ME-models are one such method with enormous promise for enhancing the extent to which we can study an organisms. This dissertation work has resulted in computational tools to enhance the dissemination of the ME-model technologies. It has also demonstrated how the evolutionary pressure of efficient proteome use has had a profound impact on cellular metabolism, and how this can be applied to improve our understanding the factors underlying cell growth and how strains evolve. Lastly, ME-models were applied to study new co-culture systems, broadening the applications of ME-models to contribute to our understanding of community metabolism.

Acknowledgements

Chapter 6 in part is a reprint of material published in: **CJ Lloyd**, N Mih, L Yang and BO Palsson. 2019. "Fundamentals of Metabolic Systems Biology." *Encyclopedia of Microbiology*, The dissertation author was the primary author.