New Applications of Constraint-Based Modeling: Network Comparisons, Thermodynamic Feasibility, and Community Dynamics

by

Joshua James Hamilton

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Chemical Engineering)

at the

UNIVERSITY OF WISCONSIN-MADISON

2014

Date of final oral examination: August 5, 2014

This dissertation is approved by the following members of the Final Oral Committee:

James R. Luedtke, Assistant Professor, Industrial and Systems Engineering

Christos T. Maravelias, Associate Professor, Chemical and Biological Engineering

Sean P. Palecek, Professor, Chemical and Biological Engineering

Jennifer L. Reed, Associate Professor, Chemical and Biological Engineering

John Yin, Professor, Chemical and Biological Engineering

To my mentors, past and present

Acknowledgements

This dissertation would not have been possible without the support and encouragement of numerous colleagues, friends, and family. First and foremost, I wish to thank my advisor, Jennifer Reed, for her outstanding mentorship and inspiration. I greatly appreciate the freedom she gave me to pursue my own interests, and the guidance she provided while I charted my own path as a scientist. I am grateful for the many opportunities she has given me over the past five years.

I must also thank Brian Pfleger for his mentorship and guidance. It was my privilege to work with him in the classroom, and my great fortune to receive excellent advice on finding the perfect postdoctoral position. I also wish to acknowledge the members of my committee—James Luedtke, Christos Maravelias, Sean Palecek, and John Yin—for freely sharing their insights and expertise. Their contributions have been valuable both scientifically and professionally.

I must also thank the members of the Reed laboratory for their friendship and collegiality. Xiaolin Zhang has been a fantastic (and patient!) mentor in the wet lab, and her guidance was invaluable in translating my research from the desktop to the benchtop. I also wish to thank Trang Vu for her mentorship when I first joined the lab, and Dave Baumler, Camo Cotten, Joonhoon Kim, Wai Kit Ong, and Chris Tervo for valuable discussions over the years. I have also had the privilege to work with a number of undergraduate students during the course of my PhD, and I am grateful for the contributions of Vivek Dwivedi and Monse Calixto to the work presented here.

The completion of this dissertation would not have been possible without the unwavering support of my entire family. I am eternally indebted to my parents, Les and Lorrie, for the endless supply of love and encouragement they have given me while I pursue my dreams. I would also like to thank my brother, Jacob, for always reminding me there is life outside the laboratory.

Abstract

An organism's metabolism can be described via a genome-scale network reconstruction (GENRE), a structured collection of biochemical transformations and their associated genes. GENREs serve as platforms for the development of genome-scale metabolic models (GEMs), mathematical models which enable an organism's phenotype to be evaluated computationally via constraint-based methods (CBMs). Constraint-based modeling integrate optimization with physiochemical constraints to define and idenfity feasible cellular behaviors. This dissertation describes computational methods which advance the field of constraint-based modeling in three areas: network comparisons, thermodynamic constraints, and community CBMs.

Advances in genome sequencing and software development have enabled the rapid construction of GEMs, but methods for comparing GEMs remain in their infancy. We have developed an approach to identify functional differences between GEMs by comparing GENREs aligned at the gene level. Our approach (CONGA) seeks genes whose deletion disproportionately changes flux through a selected reaction (e.g., growth) in one model over another. Through a number of examples, we demonstrate this this approach can be used to identify differences in GENRE content which enable unique metabolic capabilities.

The predictive accuracy of CBMs depends on the degree to which constraints eliminate infeasible behaviors. Using thermodynamics-based metabolic flux analysis (TMFA), we implemented thermodynamic constraints on an *Escherichia coli* GENRE. We examined the effect of these constraints on the flux space, and assessed the predictive performance of TMFA against gene essentiality and quantitative metabolomics data. We propose that TMFA is a useful tool for validating phenotypes, and that additional types of data and constraints can improve predictions of metabolite concentrations.

In anaerobic syntrophic communities, electrons are transferred between species via reactions which are tightly constrained by thermodynamics. We developed and analyzed a thermodynamic coculture model of the syntrophic association between *Syntrophobacter fumaroxidans* and *Methanosprillum hungatei*. Our analysis revealed that thermodynamic constraints alone are insufficient to correctly predict the mechanism of H_2 production by *S. fumaroxidans*. Our model also describes the contributions of different H_2 production modes to electron transfer in the community, and predicts that *S. fumaroxidans* may alter its metabolic behavior in the presence of a high relative abundance of *M. hungatei*.

Some material in this manuscript has been previously published:

Hamilton JJ, Reed JL (2012) Identification of Functional Differences in Metabolic Networks Using Comparative Genomics and Constraint-Based Models. PLoS One 7(4): e34670.

Hamilton JJ, Dwivedi V, Reed JL (2013) Quantitative assessment of thermodynamic constraints on the solution space of genome-scale metabolic models. Biophys J 105(2): 512–522.

Hamilton JJ, Reed JL (2014) Software platforms to facilitate reconstructing genome-scale metabolic networks. Environ Microbiol 16(1): 49–59.

Table of Contents

List of Figures

List of Tables

Chapter 1: Genome-Scale Network Reconstructions, Genome-Scale Models and Constraint-Based Methods

Some material in this chapter has been adapted from:

> Hamilton JJ, Reed JL (2014) Software platforms to facilitate reconstructing genome-scale metabolic networks. Environ Microbiol 16(1): 49–59.

Systems-level analyses of microbial metabolism are facilitated by genome-scale network reconstructions (GENREs) of microbial biochemical networks, which collect and codify current knowledge about the metabolism of an organism. A GENRE is an organism-specific structured collection of biochemical transformations and associated genes obtained from the genome annotation and primary literature [56]. The past decade has seen enormous growth in the number of published GENREs for a wide range of organisms (e.g., *Escherichia coli* [139], *Saccharomyces cerevisiae* [168], *Shewanella oneidensis MR-1* [65], and *Geobacter* spp. [134]), and guidelines for developing a high-quality GENRE have recently been published [236].

A GENRE serves as a knowledge base for a particular organism, as well as a platform for the development of genome-scale metabolic models (GEMs) [183]. GEMs provide a concise mathematical representation of an organism's metabolism and enable its phenotype to be evaluated and manipulated computationally via constraint-based methods [123,265]. GEMs have also been used to drive and support experimental efforts in a variety of applications, including network characterization [158,265], metabolic engineering [217,265], evolution [170], drug discovery [38], contextualizing high-throughput data [27,192], and elucidating microbial community interactions [260].

Most notably, constraint-based methods were used to design and commercialize the first organism with direct biocatalytic routes to 1,4-butanediol [257]. Other examples

of GEM usage include: the identification of better antibiotics against *Vibrio vulnificus* [105]; a study of gene loss in the endosymbiont *Buchnera aphidicola* [258]; predictions of cooperative and competitive potential in bacterial communities [66]; and the design of a uranium bioremediation strategy for contaminated groundwater [263].

Chapters 1 and 2 provide background information on genome-scale models and constraint-based methods. Chapter 1 focuses on the construction of GENREs and GEMs, and describes the principles of constraint-based modeling. Chapter 2 describes current state-of-the-art methods for performing functional comparisons of GENREs (Section 2.1), enforcing thermodynamic constraints (Section 2.2) and constraint-based modeling of microbial communities (Section 2.3). The remaining chapters focus on the contributions of this dissertation to each of these areas.

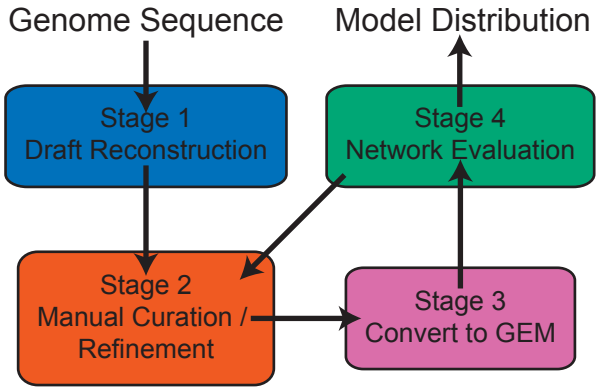
1.1: Genome-Scale Network Reconstructions

The network reconstruction process is divided into four stages [236], during which an annotated genome is converted to a high-quality metabolic network reconstruction and distributed to the scientific community (Figure 1.1). Reconstruction is an iterative process: stages may be repeated until the reconstruction's GEM predictions agree with experimental observations.

Reconstruction begins when an annotated genome gets converted into a draft reconstruction (Stage 1), during which biochemical databases are used to identify the metabolic functions associated with a genome's content. Once a draft reconstruction has been obtained, it must be refined in light of physiochemical considerations and expert knowledge about the organism (Stage 2). After a reconstruction is refined (or curated), it is then further evaluated in light of experimental evidence (Stage 4). These evaluations are performed using a GEM derived from the reconstruction (Stage 3), in the form of computational simulations. The results of these evalulations feed back into

Stage 2, and the reconstruction is refined until the GEM correctly predicts experimental observations. This results in an iterative reconstruction process, whose endpoint is determined by the desired scope and purpose of the reconstruction. One of the most comprehensive GENREs to date, for *E. coli*, has gone through four iterations since its initial publication in 2000 [50,54,166,190]. Other GENREs which have been subject to multiple rounds of iteration include the GENRE for *Homo sapiens,* with three reconstructions [49,82,238], and that for *S. cerevisiae,* with over a dozen reconstructions (reviewed in [168]).

Figure 1.1. Schematic of network reconstruction process.



Reconstruction is an iterative process, and Stages 2-4 are repeated until the model predictions agree with experimental observations. Stage numbers refer to the guidelines published by Thiele and Palsson [236].

Historically, network reconstruction has been a time- and labor-intensive process [236], and a number of tools have been developed to automate parts of the procedure. Most software tools have focused on developing draft reconstructions (such as [157,177], and many others) or performing simulations (such as [108,194,202], among others). Reviews of many tools for drafting GENREs or simulating GEMs have recently been published [40,116,187]. A number of software platforms have also been developed to provide support for all stages of the reconstruction process [1,89,102,119,227]; these have been recently reviewed as well [80]. These software platforms enable researchers who are

new to modeling to build a draft reconstruction and subsequently refine it to obtain a final, well-curated reconstruction.

1.1.1: Stage 1: Draft Reconstruction

In the first stage of network reconstruction, an annotated genome is used to assemble a collection of metabolic reactions. Annotated genomes can be obtained from a variety of sources, including the National Center for Biotechnology Information (NCBI) [247], the Kyoto Encyclopedia of Genes and Genomes (KEGG) [100,101], and the SEED [169]. A recent review describes a number of additional databases useful for network reconstruction [69].

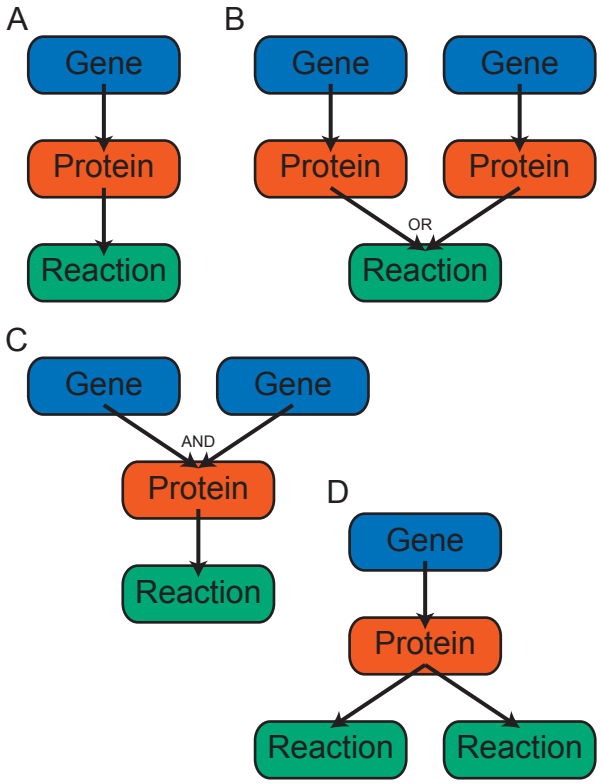
To obtain the appropriate metabolic reactions, metabolic genes are first identified (e.g., on the basis of enzyme names or enzyme commission (E.C.) numbers). Draft reconstructions also contain gene-protein-reaction (GPR) associations, which indicate which gene products carry out which biochemical transformations. Genes are connected to reactions via biochemical reaction databases, such as KEGG [100,101], MetaCyc [37], MetRxn [112], BIGG [200], or other existing GENREs. The draft reconstruction comprises all reactions and GPR associations retrieved in such a way.

1.1.2: Stage 2: Refinement and Curation

In the second stage, the draft reconstruction must be evaluated and refined to ensure consistency with physical principles and experimental evidence. The inclusion of each reaction is carefully scrutinized and GPR associations are validated.

The first task in Stage 2 is to verify the substrates and products of each reaction. Reaction databases such as KEGG may represent reactions in a generic way in order to capture a spectrum of catalytic activities. Manually curated or organism-specific databases are less likely to include such generic reactions. In all cases, generic reactions included in the draft reconstruction should be replaced with organism-specific ones.

Figure 1.2. Examples of detailed gene-protein-reaction (GPR) associations.



A: Simple association, in which a single gene encodes a single enzyme. B: Isozymes, in which multiple genes encode distinct proteins carrying out the same function. C: Multimeric protein complex, in where multiple genes encoding distinct protein subunits come together to form an active enzyme. D: One-to-many relationship, in which a single protein can carry out multiple reactions.

In addition, many databases represent metabolites in their uncharged state, when the actual charged state depends on the intracellular pH. Thus, all reactions should be checked to ensure all metabolites are in their proper charged state, and the overall reaction is mass- and charge-balanced. Unbalanced reactions may lead to the production of metabolites or energy (e.g., ATP) from nothing. Reactions should also be written in the proper direction, localized to the proper compartment (for eukaryotes), and associated with the proper metabolic pathway/subsystem.

Next, GPR associations for all reactions in the reconstruction need to be verified. All metabolic genes should be associated with the proper biochemical reactions (e.g., on the basis of annotations or experimental evidence) and each GPR should have the

proper form (Figure 1.2), as indicated by their gene annotations. In the simplest case, a single gene product carries out a single biochemical reaction (Figure 1.2A). Additionally, multiple enzymes may carry out the same reaction (called isozymes, Figure 1.2B); reactions may be carried out by the combined subunits of a multimeric protein complex (Figure 1.2C); or a single enzyme may carry out multiple reactions (Figure 1.2D). A detailed GPR captures these types of interactions between genes and reactions.

Next, a biomass reaction should be added to the reconstruction. The biomass reaction is a non-enzymatic reaction containing the macromolecules and other compounds which make up the dry weight of the cell [58]; the reaction is used to represent cellular growth when performing computational simulations. The construction of biomass equations has been recently reviewed [58]. Other reactions to be added in this stage of the reconstruction process include the ATP-maintenance reaction (representing cellular maintenance costs) and demand/sink reactions for metabolites whose biosynthesis or degradation pathways are unknown.

1.1.3: Stage 3: Conversion to a GEM

In this stage, the refined GENRE gets converted to a GEM, as described in Section 1.2. The resulting GEM serves as a basis for the simulations of Stage 4. Simulations can be performed using a variety of software platforms, including the popular COBRA Toolbox for Matlab [202], as well as many others (reviewed in [116,187]). More sophisticated researchers may prefer to perform simulations using a modeling language such as GAMS [41] or AMPL [64].

GENREs and other systems-biology models are distributed in one or more standard formats, such as Systems Biology Markup Language (SBML) [93] or BioPax [44]. Most software platforms for the reconstruction of GENREs support exporting the reconstruction in one or more standard formats.

1.1.4: Stage 4: Network Evaluation

The fourth and final stage of network reconstruction consists of network evaluation and validation against experimental data. During this stage, simulations are performed on a GEM derived from the reconstruction. The fundamental algorithm upon which most simulations are based is flux-balance analysis (FBA, [165]), a constraint-based method for predicting the flow of metabolites through a metabolic network. FBA can be applied to a variety of physiological analyses, including prediction of growth rates, byproduct secretion rates, and gene essentiality, as well as the calculation of theoretical yields [165].

The first evaluation step is to identify metabolic dead-ends, those metabolites which cannot be created or consumed. Such metabolites point to gaps, or missing reactions, in the network which may need to be filled. In particular, gaps associated with the production of biomass components or secretion products, or which may cause blocked reactions (i.e., reactions that can not carry any flux,), should be evaluated and filled.

The GEM predictions should also be validated against available experimental data. Common validation steps include prediction of experimental growth rates, gene deletion phenotypes, or other important physiological properties (such as P/O ratio, or flux splits in metabolic pathways).

Finally, we note that there are a number of algorithms to perform network evaluation and validation. These include: GapFind and GapFill for identifying and resolving dead-ends and gaps [113]; FVA for identifying blocked reactions [132]; and SMILEY [191], GrowMatch [198], GeneForce [13], and CROP [48] for resolving growth phenotype inconsistencies. Researchers should consult the details of each algorithm before selecting an approach.

1.2: Genome-Scale Models and Constraint-Based Methods

As described above, a GENRE is an organism-specific structured collection of the biochemical transformations and their associated genes. A GEM provides a mathematical representation of an organism's metabolism which enables its phenotype to be predicted and manipulated computationally. A model differs from a reconstruction in that it introduces relevant variables and equations which describe an organism's behavior. For metabolic models, these variables are most commonly fluxes through metabolic reactions, though others are possible (e.g., metabolite concentrations, Gibbs free energies).

1.2.1: Converting a GENRE to a GEM

During the conversion of a GENRE to a GEM, metabolic reactions are converted from a textual representation to a mathematical one, in the form of a stoichiometric matrix S. Each column of S corresponds to an individual reaction and each row of S corresponds to an individual metabolite. The stoichiometric coefficients of a reaction are represented

Figure 1.3. Difference in representation of glycolysis in a GENRE and a GEM.

Abbreviation	Glycolytic reactions
HEX1	[c]GLC + ATP → G6P + ADP + H
PGI	[c]G6P ↔ F6P
PFK	[c]ATP + F6P → ADP + FDP + H
FBA	[c]FDP ↔ DHAP + G3P
TPI	[c]DHAP ↔ G3P
GAPD	[c]G3P + NAD + PI ↔ 13DPG + H + NADH
PGK	[c]13DPG + ADP ↔ 3PG + ATP
PGM	[c]3PG ↔ 2PG
ENO	[c]2PG ↔ H_2O + PEP
PYK	[c]ADP + H + PEP → ATP + PYR

	HEX1	PGI	PFK	FBA	TPI	GAPD	PGK	PGM	ENO	PYK
ATP	−1	0	−1	0	0	0	1	0	0	1
GLC	−1	0	0	0	0	0	0	0	0	0
ADP	1	0	1	0	0	0	−1	0	0	−1
G6P	1	−1	0	0	0	0	0	0	0	0
H	1	0	1	0	0	1	0	0	0	−1
F6P	0	1	−1	0	0	0	0	0	0	0
FDP	0	0	1	−1	0	0	0	0	0	0
DHAP	0	0	0	1	−1	0	0	0	0	0
G3P	0	0	0	1	1	−1	0	0	0	0
NAD	0	0	0	0	0	−1	0	0	0	0
PI	0	0	0	0	0	−1	0	0	0	0
13DPG	0	0	0	0	0	1	−1	0	0	0
NADH	0	0	0	0	0	1	0	0	0	0
3PG	0	0	0	0	0	0	1	−1	0	0
2PG	0	0	0	0	0	0	0	1	−1	0
PEP	0	0	0	0	0	0	0	0	1	−1
H_2O	0	0	0	0	0	0	0	0	1	0
PYR	0	0	0	0	0	0	0	0	0	1

Left: GENRE representation of the first ten reactions in glycolysis. Right: The same reactions represented as a stoichiometric matrix. Columns of the matrix correspond to the indicated reactions, and rows correspond to the indicated metabolites.

as elements in the matrix, where each entry $S_{i,j}$ corresponds to the stoichiometric coefficient of the i^{th} metabolite in the j^{th} reaction. The process of constructing a stoichiometric matrix from a GENRE is illustrated for glycolysis in Figure 1.3.

Consider the reaction hexokinase, abbreviated HEX1. All metabolites have stoichiometry 1, with the reactants glucose (GLC) and ATP having negative coefficients, corresponding to their role as substrates in the reaction. The remaining metabolites, glucose-6-phosphate (G6P), ADP, and H^+, have positive coefficients, corresponding to their role as products.

The logical relationships describing GPR associations (Figure 1.2) must also be converted to mathematical ones. The technique for doing so is well-established [106] and described here. These mathematical relationships are formulated using a three dimensional array $GPR(j,p,g)$, with binary variables y_j, w_p, and z_g for each reaction j in the set of reactions J, enzyme p in the set of enzymes P, and gene g in the set of genes G, respectively. Each element of $GPR(j,p,g)$ has a value of 1 if there exists some association between reaction j, protein p, and gene g, and has a value of 0 otherwise. Additionally, each reaction with a known GPR ($j \in J_{GPR}$) can be carried out by one or more enzyme complexes ($p \in P(j)$), and each enzyme complex is associated with one or more genes ($g \in G(p)$). These three sets are defined as follows:

$$\begin{aligned} J_{GPR} &= \{j \in J \mid \exists (p,g) \text{ s.t. } GPR(j,p,g) = 1\} \\ P(j) &= \{p \in P \mid \exists g \text{ s.t. } GPR(j,p,g) = 1 \text{ for } j\} \\ G(p) &= \{g \in G \mid \exists j \text{ s.t. } GPR(j,p,g) = 1 \text{ for } p\} \end{aligned} \tag{1.1}$$

If any of the enzymes for reaction j are present (any $w_{p(n)} = 1$), the reaction can have a non-zero flux ($y_j = 1$), where y_j indicates whether a reaction is active or inactive. Conversely, if all the enzymes are absent (all $w_{p(n)} = 0$), then the reaction cannot occur ($y_j = 0$). This reaction-enzyme logical relationship can be formulated as:

$$y_j \geq w_p \qquad \forall j \in J_{GPR}, p \in P(j) \tag{1.2}$$

$$y_j \leq \sum_{p \in P(j)} w_p \qquad \forall j \in J_{GPR} \tag{1.3}$$

for the active and inactive cases, respectively.

If all of the genes associated with enzyme complex p are expressed (all $z_{g(p)} = 1$), then the enzyme is active ($w_p = 1$). If any of the subunits are not expressed (any $z_{g(p)} = 0$), then the enzyme is inactive ($w_p = 0$). This enzyme-gene logical relationship can be formulated as:

$$w_p - 1 \geq \sum_{g \in G(p)} \left(z_g - 1\right) \qquad \forall p \in P \tag{1.4}$$

$$w_p \leq z_g \qquad \forall p \in P, p \in G(p) \tag{1.5}$$

for the active and inactive cases, respectively. Reactions without GPR associations are not subject to these rules.

1.2.2: Constraint-Based Modeling and Flux Balance Analysis

GEMs provide a computational platform for the prediction of cellular phenotypes via constraint-based methods. These methods aim to predict the distribution of material through the cellular system (the flux distribution), as given by a flux vector v. Each reaction j has an entry in v, with the values of v corresponding to the rate at which each reaction occurs. In order to predict a flux distribution, the physiochemical constraints which govern cellular behavior must be identified written down mathematically. These constraints serve as the basis for constraint-based methods, which use mathematical programming and optimization [248] to identify a particular flux distribution.

Most constraint-based methods rely heavily on three constraints, the first of which is a steady-state mass balance constraint. Mass balances can be written for each metabolite i by taking the dot product of row $S_{i,\bullet}$ with the flux vector v. A system of mass balances for all metabolites in the network can be written:

$$\frac{dX}{dt} = S \bullet v \tag{1.6}$$

where X is a vector of metabolite concentrations. At steady-state, the time derivatives of the concentration vector are 0, and the constraint can be simplified:

$$S \bullet v = 0 \tag{1.7}$$

to give the steady-state mass balance constraint. The solution to the system of linear equations given by (1.7) is under-determined, due to the matrix S having fewer rows than columns (fewer metabolites than reactions). This necessitates the use of optimization to select a single flux distribution satisfying the constraint.

The second constraint imposes upper and lower bounds on the reaction rates v_j:

$$v_{min,j} \le v_j \le v_{max,j} \qquad \forall j \in J \tag{1.8}$$

The bounds v_{min} and v_{max} can be estimated from enzyme capacity (kinetic) information, or can be set to experimentally observed rates (e.g., substrate uptake rates). In the absence of specific information, v_{min} and v_{max} are typically set to -1000 and 1000 mmol/gDW/hr, respectively, where gDW refers to the gram dry weight of biomass.

Finally, some reactions are known to be proceed only in one direction (e.g., for thermodynamic reasons). Such reactions are assigned to the set J_{irrev} and assigned a v_{min} of 0:

$$0 \le v_j \qquad \forall j \in J_{irrev} \tag{1.9}$$

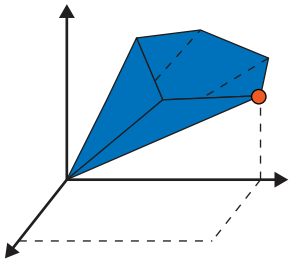
Virtually all constraint-based methods utilize these three constraints ((1.7) to (1.9)), but a wide variety of other constraints have been proposed [123,183].

These three constraints define a convex solution space of feasible steady-state flux distributions through the metabolic network (Figure 1.4). Optimization can then be used to select the flux distribution which best approximates observed cellular behavior, provided we can find an appropriate objective function (Figure 1.4).

A variety of objective functions have been proposed [207], with maximization of growth rate (v_{BM}) being most common. The growth rate is computed by maximizing

flux through the biomass equation, an artificial (sink) reaction representing the metabolic and energetic needs of the cell [58]. It is usually defined as a weighted linear combination of the metabolites found in one dry gram of cellular material. Other objective functions can be used to determine other phenotypes, such as maximal ATP production, or the yield of a desired by-product.

Figure 1.4. Visual representation of constraint-based analysis.



Physiochemical constraints define a convex solution space of feasible flux distributions (the blue cone). Optimization techniques such as flux-balance analysis identify a particular flux distribution (the red dot) maximizing or minimizing some function of interest.

The linear program which maximizes growth rate subject to the above constraints is flux-balance analysis, or FBA [164]:

$$\begin{array}{lll} \textbf{max} & v_{BM} & \\ \textbf{s.t.} & S \bullet v = 0 & \\ & v_{min,j} \le v_j \le v_{max,j} & \forall j \in J \\ & 0 \le v_j & \forall j \in J_{irrev} \end{array} \qquad \text{(FBA)}$$

FBA is an example of a linear program (LP), so-called because the objective and constraints are all linear functions of the variables. Mature, robust, and efficient solution algorithms exist for linear programs [59].

Other constraint-based methods contain integer variables, typically corresponding to discrete decisions such as the addition or removal of a gene. Such problems are called mixed-integer programs (MIPs), for which a variety of solution techniques also exist [250].

Chapter 2: Current State of Constraint-Based Modeling

Material in Section 2.1 was originally published in:

Hamilton JJ, Reed JL (2012) Identification of Functional Differences in Metabolic Networks Using Comparative Genomics and Constraint-Based Models. PLoS One 7(4): e34670.

Material in Section 2.1: was originally published in:

Hamilton JJ, Dwivedi V, Reed JL (2013) Quantitative assessment of thermodynamic constraints on the solution space of genome-scale metabolic models. Biophys J 105(2): 512–522.

Material in Section 2.3 is being prepared for publication as:

Hamilton JJ, Calixto Contreras, M, Reed JL (2014). Thermodynamics and H_2 Transfer in a Methanogenic, Syntrophic Community.

2.1: Comparative Analysis of Metabolic Networks

Advances in genome sequencing and computational modeling techniques have sparked the construction of genome-scale network reconstructions (GENREs) [56] for over 100 prokaryotic and eukaryotic organisms [158]. These reconstructions describe the functions of hundreds of metabolic genes, and enable a concise mathematical representation of an organism's biochemical capabilities via genome-scale models. Constraint-based methods [183] can then be applied to genome-scale models to understand and predict cellular behavior. Genome-scale models are becoming a common framework for representing genomic information, as evidenced by recent works simultaneously reporting genome sequences and metabolic models [8,105]. Efforts like the new Model SEED database will facilitate this process, by enabling the rapid construction and refinement of network reconstructions as genome annotations change [89].

The abundance of genome sequences has led to advances in comparative genomics, in which biological insight comes from interrogation of genome structure and function across species. The advent of tools such as the Model SEED paves the way for functional comparison of genome-scale reconstructions, but computational methods for comparing models at a functional level have not yet emerged. Existing network comparison approaches such as reconstruction jamborees [90,237] or metabolic network reconciliation [159] compare models of the same or closely-related organisms with the aim of identifying and reconciling differences between models.

These approaches rely on a manual mapping of metabolic compounds and reactions across the networks and then look at differences and similarities in reaction and gene content to identify *structural differences* (e.g., the presence or absence of particular genes or reactions). However, existing approaches do not identify *functional differences* (e.g., differences in organism behavior), or explain how structural differences impact the functional states of the network (e.g., achievable rates of growth or chemical production). Instead, models must be analyzed individually, and a number of simulations may be necessary before functional differences arising from structural differences are observed. Additionally, reaction alignment approaches can be time-consuming, since biochemical databases (such as BiGG, BioCyc, KEGG or SEED [37,100,169,200]) and model construction platforms (such as Pathway Tools [102] or the Model SEED [89]) may use different nomenclatures or abbreviations to describe metabolites and reactions.

2.2: Thermodynamic Constraints and Genome-Scale Metabolic Models

Genome-scale network reconstructions provide concise mathematical representations of an organism's biochemical capabilities, and serve as a platform for constraint-based techniques which can be used to understand and predict cellular behavior [123,158]. The

predictive accuracy of constraint-based methods depends on the degree to which constraints eliminate physiochemically and biologically infeasible behaviors.

Flux-balance analysis (FBA) [165] is commonly employed to predict the state of the network by identifying a steady-state flux distribution maximizing cellular growth, while also satisfying mass-balance and enzyme capacity constraints. Reaction directionality is typically assigned on the basis of enzyme assays or biological considerations (e.g., no free ATP synthesis), with no consideration given to thermodynamics. The second law of thermodynamics states that a negative Gibbs energy of reaction ($\Delta_r G$) is needed to drive a forward reaction flux, v, leading to the thermodynamic constraint $v \cdot \Delta_r G < 0$ for non-zero v.

The first attempt to enforce thermodynamic constraints on FBA was energy balance analysis (EBA) [14,185], which incorporated nonlinear constraints on chemical potentials into FBA in order to eliminate closed cycles. Closed cycles are sets of reactions for which the overall thermodynamic driving force is zero, and through which no net flux can occur (e.g., A→ B → C → A). These closed cycles have been variously referred to as Type III pathways [203], internal flux cycles [88], and loops [201]. Because its constraints are nonlinear, EBA is restricted to small models, though a scalable algorithm has recently been proposed [91]. The same scientists also developed a method to compute and eliminate closed cycles in flux-balance models [15,255] based solely on stoichiometry, although the technique remains computationally demanding for genome-scale networks. Recently, a scalable, mixed-integer approach to eliminating closed cycles has also been developed [201].

Due to the limited availability of free energy data for reactions and metabolites [2,74], many of the above approaches do not directly account for the relationships between $\Delta_r G$, metabolite concentrations, and free energies of formation ($\Delta_f G$). Fortunately, group contribution methods (GCMs) [96,135,136] have been developed to

estimate free energies of metabolites and reactions for which data are unavailable. A recent model of *E. coli* used one such GCM to assign reaction directionalities on the basis of thermodynamic constraints [54,96]. In other approaches, experimentally-measured thermodynamic data have been combined with heuristics and/or group contribution data to define feasible reaction directions in *E. coli* [62,114]. However, these approaches [54,62,114] neglect thermodynamic interactions between reactions in the network which arise due to shared metabolites. As a result, the directionality of a reaction is assigned independently of other reaction directions in the network. For example, two reactions may be feasible in both the forward and reverse directions, but due to a shared metabolite, the pair of reactions must proceed in the same direction. These approaches fail to capture this type of thermodynamic coupling between reactions.

GCMs have also been used in approaches which capture thermodynamic interactions by including metabolite concentrations as variables. EBA has been extended to predict intracellular metabolite concentrations in a small network [16], and two mixed-integer approaches have also been developed, in which thermodynamic constraints are imposed on top of pre-defined reaction directions. NET analysis [115] integrates quantitative metabolomics data with thermodynamic constraints to predict feasible free energy ranges for all reactions in the network. Another method, thermodynamics-based metabolic flux analysis (TMFA) [88], extends FBA with thermodynamic constraints, enabling the quantitative prediction of feasible ranges of metabolite concentrations and reaction free energies, without relying on metabolomic data. However, both of these methods have, to date, relied on prior knowledge of the reversibility or directionality of reactions [19,70,88,115], thereby restricting their predictive capabilities.

2.3: Metabolic Modeling of Microbial Consortia

Microorganisms in nature do not exist as pure cultures, but rather are engaged in a variety of interactions with other species in their environment. Syntrophy is one such type of inter-species interaction in which one species lives off the metabolic by-products of another. For example, two species might combine their metabolic capabilities to catabolize a substrate which neither could consume alone [142,205,222]. Syntrophic associations are an important step in the anaerobic degradation of organic matter to methane [143]. Syntrophic associations play an important role in sustaining a variety of natural and synthetic methanogenic communities [143], including aquatic sediments, landfills, and anaerobic digesters. These communities are typically tightly constrained by thermodynamics, as the oxidation reactions carried out by the first community member are thermodynamically unfavorable unless the degradation products are maintained at low levels by the second community member [204].

In anaerobic syntrophic communities, electrons can be transferred from one partner to the other, either through direct contact or small molecule diffusion [221]. Traditional biochemistry has been able to elucidate these electron transfer mechanisms [147,215,222,235], but it is difficult to evaluate these pathways in their metabolic and environmental context. This difficulty has led to the development of systems biology approaches that provide a link between an organism's genotype, proposed reaction mechanisms, and organismal and environmental phenotypes.

Genome-scale metabolic models (GEMs) and constraint-based methods have proven to be powerful computational tools for interrogating the link between genotype and phenotype [123,158,265]. GEMs provide a concise mathematical representation of an organism's metabolism, and constraint-based methods enable phenotypes to be predicted, evaluated, and manipulated computationally. These methods have provided significant insights into the genotype-phenotype relationship of isolated microbial species

[134,139,168], and recently have been deployed to investigate a variety of microbial interactions [31,66,86,109,121,149,212,223,249,266,267].

One of the earliest microbial community models used flux-balance analysis (FBA, [167]) to investigate the mutualism between the sulfate-reducing bacterium *Desulfovilbrio vulgaris* and the methanogenic archaeon *Methanococcus maripaludis* [223]. In this study, each organism was modeled as a compartment within a larger community-scale model. These compartmentalized approaches have been used to study the metabolic and environmental origins of cooperation and competition [66,109,249], as well as many specific microbial communities [31,86,121,149,212]. Community FBA (cFBA) extends these compartmentalized approaches to specifically account for individual species' biomass abundance [103]. These compartmentalized approaches have often used a single (joint) objective function to capture community behavior, typically the sum of individual species' growth rates. The method OptCom [266,267] instead uses a multi-level optimization framework, to capture the trade-offs between organismal and community fitness, with separate objective functions for the individual species and the community.

Thermodynamic constraints can be introduced into genome-scale models, in order to ensure that network predictions are qualitatively and quantitatively consistent with thermodynamic principles [14]. A variety of such methods have been proposed, all of which enforce the qualitative requirement that flux-carrying reactions have a thermodynamically favorable Gibbs free energy [14,185,201]. Some methods go further and use experimental [2] or statistical [96,135,156] methods to estimate the Gibbs free energy of reaction. These estimates can be combined with the appropriate mathematical relations to enable quantitative predictions of reaction free energies and metabolite concentrations [79,88,115]. We have previously used one such approach, thermodynamics-based metabolic flux analysis (TMFA, [79]) to construct a

thermodynamic model of *E. coli* where reaction directions are determined solely by thermodynamics.

Chapter 3: Comparative Analysis of Metabolic Networks

This material was originally published in:

Hamilton JJ, Reed JL (2012) Identification of Functional Differences in Metabolic Networks Using Comparative Genomics and Constraint-Based Models. PLoS One 7(4): e34670.

We have developed a bilevel mixed-integer linear programming (MILP) approach to identify functional differences between models by comparing network reconstructions aligned at the gene level, bypassing the need for a time-consuming reaction-level alignment. We call this new constraint-based method CONGA, or **C**omparison **o**f **N**etworks by **G**ene **A**lignment. We first use orthology prediction tools (e.g., bidirectional best-BLAST) to identify sets of orthologs in two organisms based on their genome sequences, and then we use CONGA to identify conditions under which differences in gene content (and thus reaction content) give rise to differences in metabolic capabilities. Because orthologs often encode proteins with the same function, we would expect their gene-protein reaction (GPR) associations, and thus their associated reactions, to be similar. Therefore, a gene-level alignment serves as a proxy for a reaction-level alignment. By identifying genetic perturbation strategies that disproportionately change flux through a selected reaction (e.g., growth or by-product secretion) in one model over another, we are able to identify functional differences (e.g., biomass yield) between the two organisms. Once these functional differences are found, they can be further evaluated to identify structural differences (e.g., gene and reaction differences) between the organisms' network reconstructions. By using an MILP approach, we are able to identify these differences directly and in an exhaustive fashion, without manually aligning all reactions in the two networks.

We demonstrate that this approach can be used to study both closely- and distantly-related organisms and to address a variety of biological questions, by applying it to three pairs of organisms with increasing phylogenetic distance. We first examine differences between two published metabolic reconstructions of *Escherichia coli* metabolism, *i*JR904 [189] and *i*AF1260 [54]. The *i*AF1260 model is an update to the *i*JR904 model, constructed to more accurately reflect experimental data, including gene essentiality data and growth phenotypes [42,191]. While both models have been used as tools to help design new chemical production strains [3,63,120,216], these two models have not been evaluated with respect to differences in their metabolic engineering predictions. By identifying knockout strategies where one model predicts a larger chemical production rate than the other, we are able to determine a small set of reactions responsible for predicted chemical production differences between the two models.

We have also used CONGA to aid in the development of a genome-scale network reconstruction of the photosynthetic cyanobacterium *Synechococcus sp.* PCC 7002, which we name *i*Syp611, by comparing it to the *i*Cce806 reconstruction of *Cyanothece sp.* ATCC 51142 [242]. Photoautotrophic microbes, such as cyanobacteria, possess the ability to fix carbon dioxide and transform light into chemical energy, making them strong candidates for biofuel production hosts [10,125,126,228]. Through our automated comparison, we also demonstrate the conserved aspects of cyanobacterial physiology, and gain insight into the unique properties of *Synechococcus* and *Cyanothece*.

Finally, we applied CONGA to compare the susceptibility of distantly-related human pathogens to loss of metabolic enzymes. We selected published networks of *Mycobacterium tuberculosis* H37Rv [95] and *Staphylococcus aureus* N315 [17] and sought gene knockout strategies that are predicted to be lethal in only one organism. We were then able to identify differences in their metabolic networks which point to unique

metabolic functions as possible targets for organism-specific antimicrobials. Such antibiotics are needed to expand the limited scope of existing broad-spectrum antibiotics [117] and to provide novel mechanisms of action which make the transfer of resistance across species less probable [33,81,153,243]. We show that many of the functions we identified have been experimentally verified as essential, demonstrating that our computational approach allows us to provide a list of candidate enzymes for more focused study. As a component of this comparison, we used three distinct orthology prediction tools to prepare a gene alignment between the pathogens. We then analyzed the number of false positive ortholog calls made by each method, and examined the effect these incorrect orthology assignments had on the results generated by CONGA.

Through these three case studies, we demonstrate that CONGA can be used to rapidly compare metabolic networks regardless of phylogenetic distance. We are also able to show that CONGA has applications in metabolic engineering, model development, and antibiotic discovery. We show that CONGA can facilitate jamboree and network reconciliation efforts by pinpointing those metabolic or genetic differences which give rise to differences in model predictions.

3.1: Results

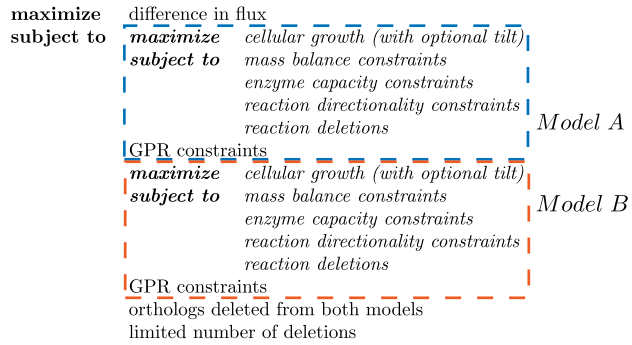
We have developed a bilevel mixed-integer linear programming (MILP) approach, called CONGA, to identify functional differences between two networks by comparing network reconstructions aligned at the gene level. We have constructed an illustrative example to demonstrate the types of functional differences CONGA can identify. We then present three case studies and demonstrate how CONGA results have implications in metabolic engineering (comparison of *E. coli* models), model development (comparison of cyanobacterial models), and drug discovery (comparison of human pathogen models).

3.1.1: Identification of Network Differences via CONGA

CONGA identifies functional differences between two networks by comparing network reconstructions aligned at the gene level. The constraint-based method identifies gene deletion strategies leading to different optimal flux distributions in the two networks. CONGA calculates the flux difference between two reactions in different models (e.g., Flux 1 in Species A minus Flux 2 in Species B) and identifies deletions such that the specified flux difference is maximized while both models are simultaneously maximizing biomass (Figure 3.1).

We refer to a solution identified by CONGA as a *gene deletion set*. CONGA can select any genes for deletion, with the restriction that orthologous genes present in both models be deleted simultaneously from both models. We note that while CONGA can calculate the flux difference between any two reactions, we believe that selecting

Figure 3.1. Conceptual structure of the CONGA formulation.



CONGA employs a bilevel optimization problem to identify genetic perturbations with nonidentical effects in each of two networks. The outer problem is an MILP which finds gene deletions maximizing the difference in flux value between two reactions in two different models. The inner problems (in italics) are flux-balance analysis (FBA) problems which ensure the flux difference is maximized while both models are maximizing biomass. An optional tilt can be added to the inner problem which forces the flux in the outer problem to the lowest value that still support maximum biomass production. FBA imposes constraints based on reaction stoichiometry, reaction directionality, and enzyme capacities. GPR constraints associate genes to reactions and are used to enforce the reaction deletions associated with the gene deletions in the outer problem. CONGA can select any genes for deletion, with the restriction that orthologous genes present in both models be deleted simultaneously from both models. Finally, a limit may be imposed on the total number of gene deletions

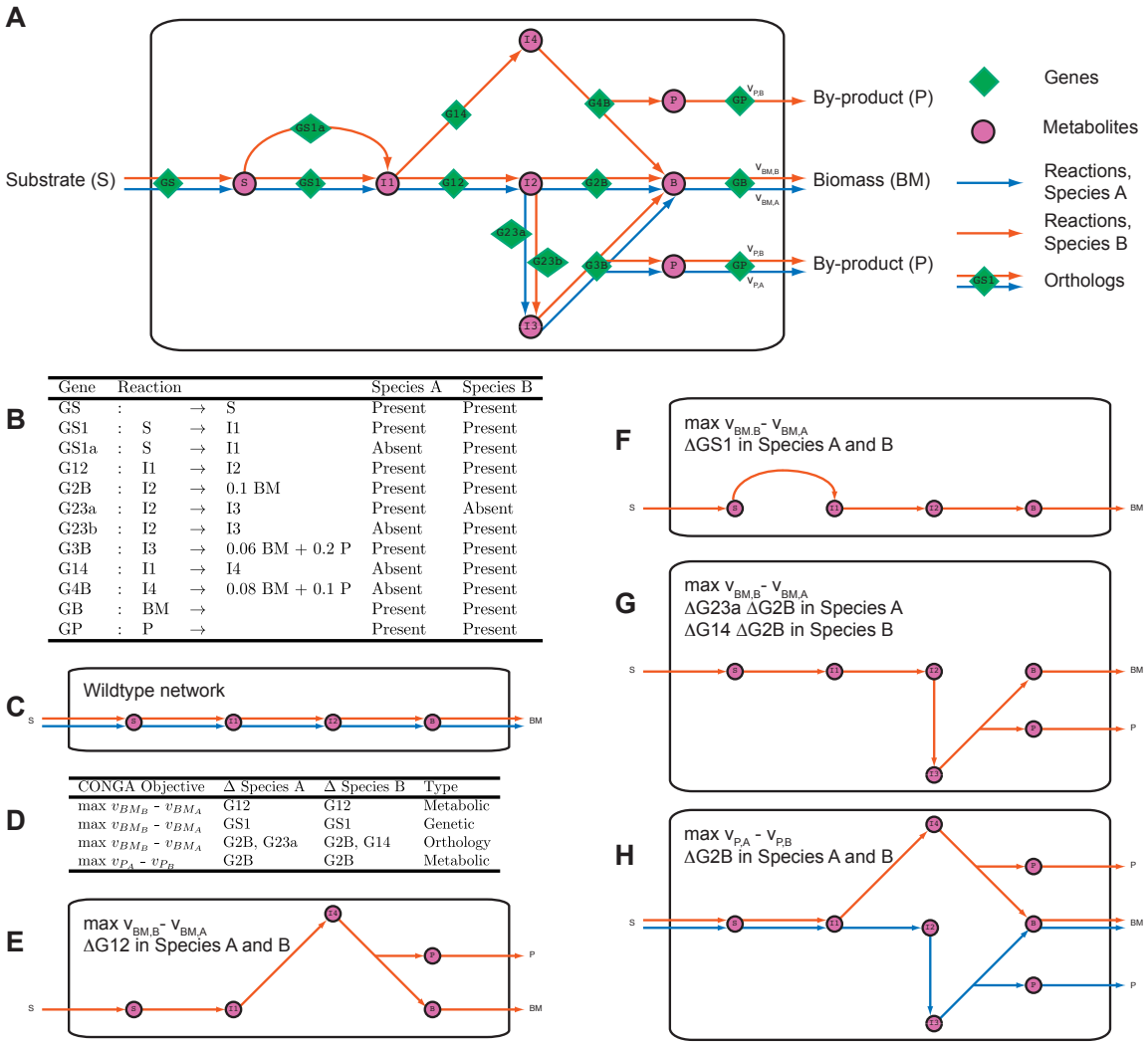
equivalent reactions (e.g., biomass) provides the most useful objective for comparing models. Via manual investigation of the results, we are able to classify gene deletion sets identified by CONGA as arising due to one of four types of functional network differences:

1. *genetic differences*, in which gene-protein-reaction (GPR) relationships differ between models;
2. *orthology differences*, in which genes encoding enzymes with identical functions cannot be assigned as orthologs (e.g., due to sequence dissimilarity);
3. *metabolic differences*, where one organism has additional reactions which enable it to carry out unique biochemical transformations; and
4. *mixed differences*, which arise due to some combination of types 1-3.

Using two example networks, we demonstrate the types of functional differences CONGA can identify (Figure 3.2). Each reaction network catalyzes the conversion of substrate (S) to biomass (BM) and some by-product (P) (Figure 3.2A). We refer to the two species as A and B, and the biomass- and by-product-producing reactions as v_{BM} and v_P, respectively. Each pathway producing biomass gives different yields for BM and P (Figure 3.2B), though the optimal flux distributions maximizing biomass without any gene deletions are identical in the two organisms (Figure 3.2C). By applying CONGA with different objective functions, we can identify gene deletion conditions under which network differences become apparent (Figure 3.2D).

We first used CONGA to compute gene deletion sets maximizing v_{BM} in Species B over Species A ($v_{BM_B} - v_{BM_A}$). This objective will be greatest when a gene deletion set is predicted to be lethal in Species A and not in Species B. One such deletion set contains the ortholog G12, which is present in both models (Figure 3.2E). Under this deletion, growth becomes impossible in Species A, whereas Species B has additional

Figure 3.2. Application of CONGA to an example pair of metabolic networks.



Gene	Reaction			Species A	Species B
GS	:	→	S	Present	Present
GS1	: S	→	I1	Present	Present
GS1a	: S	→	I1	Absent	Present
G12	: I1	→	I2	Present	Present
G2B	: I2	→	0.1 BM	Present	Present
G23a	: I2	→	I3	Present	Absent
G23b	: I2	→	I3	Absent	Present
G3B	: I3	→	0.06 BM + 0.2 P	Present	Present
G14	: I1	→	I4	Absent	Present
G4B	: I4	→	0.08 BM + 0.1 P	Absent	Present
GB	: BM	→		Present	Present
GP	: P	→		Present	Present

CONGA Objective	Δ Species A	Δ Species B	Type
max $v_{BM_B} - v_{BM_A}$	G12	G12	Metabolic
max $v_{BM_B} - v_{BM_A}$	GS1	GS1	Genetic
max $v_{BM_B} - v_{BM_A}$	G2B, G23a	G2B, G14	Orthology
max $v_{P_A} - v_{P_B}$	G2B	G2B	Metabolic

(**A**) In these two example networks, substrate (S) is utilized to produce biomass (BM) and some by-product (P). We refer to the two species as A and B, and the biomass- and product-producing reactions as v_{BM} and v_P, respectively. (**B**) List of genes and reactions present or absent in each network. All shared reactions have orthologs present in both networks, except for the reaction associated with genes G23a and G23b, which are not orthologs. (**C**) A schematic view of the wildtype network behaviors in which flux through v_{BM} is maximized. (**D**) Gene deletion sets identified by CONGA for the stated CONGA objectives The first three objectives maximize v_{BM} in Species B over Species A. The last objective maximizes v_P in Species A over Species B. The type of model difference (genetic, orthology, or metabolic) associated with each deletion set is also given. (**E** through **H**) Schematic views of the flux distributions associated with each gene deletion set in **D**. The optimal flux distributions in the example networks change as a result of the gene deletion sets in **D**. Differences in the optimal flux distributions are due to differences in the two networks.

reactions which allow it to convert I1 to B via metabolite I4. Thus, this gene deletion set points to a metabolic difference between the two models. CONGA can also be used

to identify genetic differences (Figure 3.2F). For instance, the deletion of GS1 is lethal only in Species A, because Species B has an additional isozyme (GS1a) which carries out the same transformation. Thus, this deletion set points to a genetic difference. Other deletion sets point to orthology differences (Figure 3.2G). For example, genes G23a and G23b are not orthologs even though they carry out the same reaction. Thus, the deletion of G2B and G23a is lethal in Species A, but Species B can still carry flux through the reaction associated with G23b.

CONGA can also identify how metabolic differences affect cellular phenotypes other than growth rate (Figure 3.2H). In this example, the objective is to maximize the difference in flux through v_P in Species A over Species B ($v_{P_B} - v_{P_A}$). (The resulting phenotypes for each model are analogous to production phenotypes predicted by OptORF [106].) Deleting G2B forces Species A to utilize the lower reaction pathway, producing 0.06 BM and 0.2 P per S. However, the optimal flux distribution for Species B uses the upper reaction pathway, as this route produces more biomass (0.08 BM per S vs 0.06 BM per S via the lower pathway). As a consequence, Species A produces more by-product: 0.2 P per S in Species A vs. 0.1 P per S in Species B.

Because production values may not be unique at the maximum growth rate, CONGA can artificially inflate flux differences between models. This can only occur when the fluxes whose difference is being maximzed (e.g., chemical production rates) differ from the fluxes maximized by each model (e..g, biomass). In this case, we impose a tilt on the objective of the inner problem. This tilt forces CONGA to identify deletions such that the specified flux difference is maximized when the individual fluxes through each reaction are at their lowest values that still support maximum biomass production.

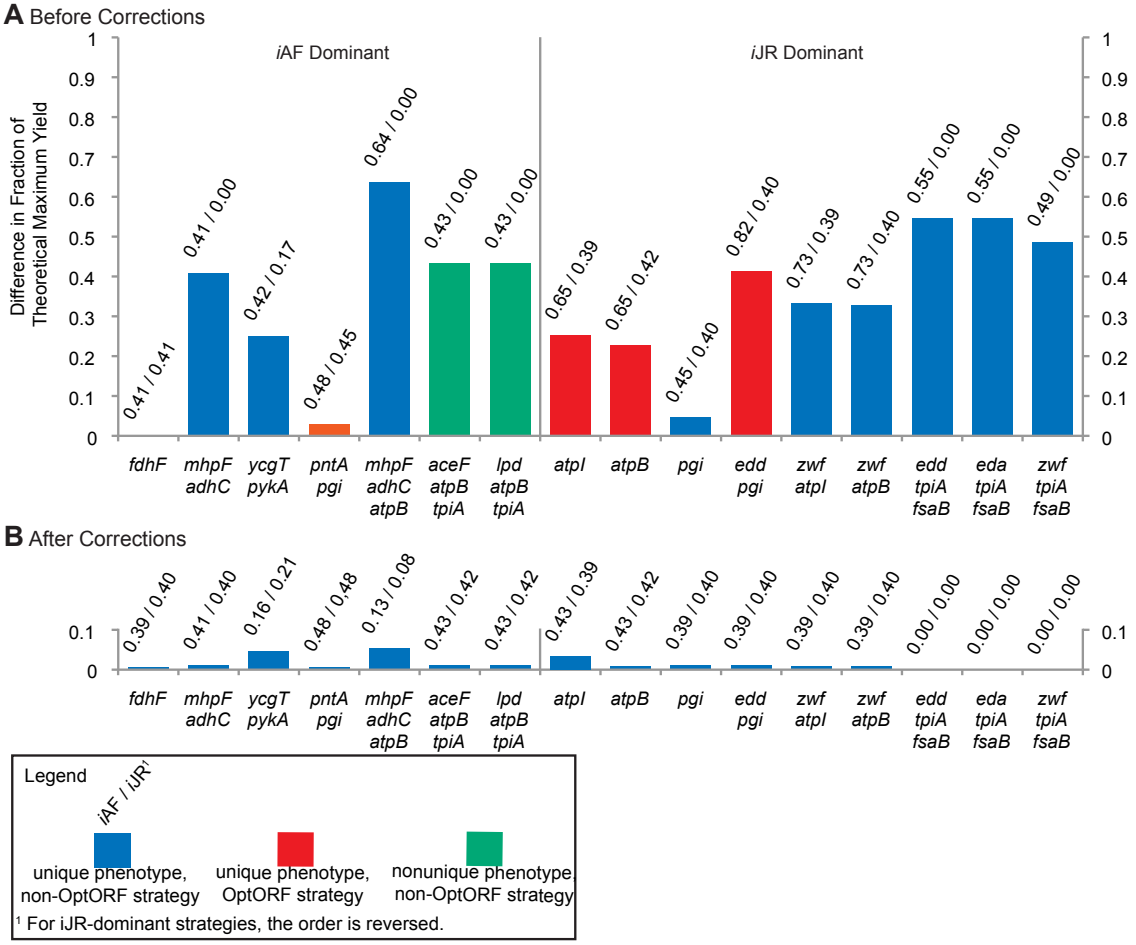
3.1.2: Comparison of *E. coli* Metabolic Models

We first used CONGA to compare two genome-scale metabolic models of *E. coli*, the *i*JR904 model [189] and the *i*AF1260 model [54]. The *i*AF1260 model extends the *i*JR904 model by compartmentalizing the network (separating the cytoplasm and periplasm), improving the biomass composition, and adding new metabolic reactions. The *i*JR904 model has been used frequently for metabolic engineering studies [55], but to our knowledge no studies have examined the extent to which the *i*AF1260 model's additional metabolic content affects computationally derived strain designs.

To explore the effect of the *i*AF1260 model's larger network, we used CONGA to identify gene deletion strategies for three commonly studied fermentation products—ethanol, lactate, and succinate—seeking identical knockout conditions where the *i*AF1260 model predicted higher production rates than the *i*JR904 model, and vice versa. We refer to such strategies as *model-dominant strategies*. For example, an *i*AF1260-dominant strategy is one in which the same gene deletion set predicts higher chemical production in the *i*AF1260 model than in the *i*JR904 model. Because some of these knockout strategies result in nonunique chemical production rates, model-dominant strategies were identified with respect to the lowest possible production rate consistent with the maximum growth rate.

Our initial CONGA results revealed a need to reconcile the fermentation pathways between the two models, due to changes in representation made in the *i*AF1260 model. We thus modified the *i*JR904 model to reflect these changes and repeated the simulations using the reconciled models. (See Dataset S2 in the original publication for details.) For ethanol, succinate, and lactate, we identified the top three model-dominant strategies for each model for up to three, four, and five knockouts, respectively. We observed that multiple deletions are necessary to detect differences in production of these latter metabolites, and the difference in yield does not improve

Figure 3.3. Model-dominant production strategies for ethanol.



(**A**) Deletion strategies for ethanol production. Each bar represents the absolute difference in predicted ethanol yields between the *i*JR904 and *i*AF1260 models as a fraction of the maximum theoretical yield (2 ethanol/glucose). Left side: Strategies for which the *i*AF1260 model predicts higher production. Right side: Strategies for which the *i*JR904 model predicts higher production. Corresponding gene deletion strategies involving 1, 2, or 3 genes are given below the figure. Numbers above each bar indicate the fraction of the theoretical maximum yield obtained by each model, with the dominant model listed first. Some strategies have a nonunique ethanol production phenotype, in which multiple ethanol production values can occur at the maximum growth rate. For these scenarios, the production difference calculated by CONGA is from the lowest expected level of ethanol production in each model, and such strategies are indicated in green. Strategies for which the yield of the dominant model meets or exceeds the yield for the third-best OptORF strategy for that model are known as OptORF strategies, and such strategies are indicated in red. (**B**) The same gene deletion strategies after reconciliation of the *i*JR904 and *i*AF1260 networks with respect to metabolic differences.

significantly beyond four or five knockouts, depending on the model and product. We also employed OptORF [106], without transcriptional regulation, to identify the top

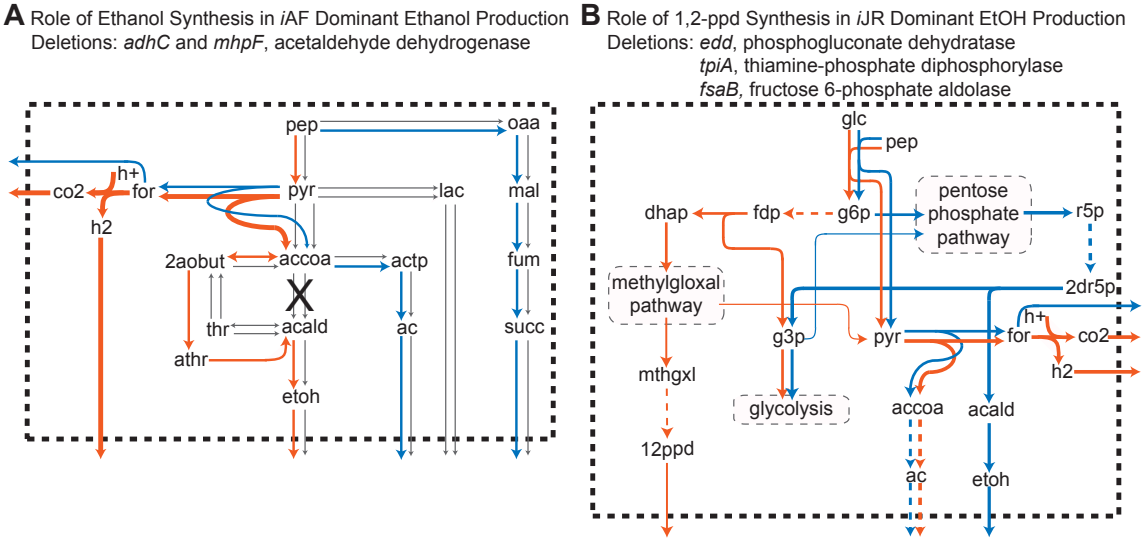
three deletion strategies for each model and product, for each number of gene deletions. We refer to these strategies as *OptORF strategies*. These strategies were then compared to the model-dominant strategies identified by CONGA, to determine if optimal OptORF strategies are likely to give similar or different predictions between the two models.

The CONGA results for the model-dominant strategies for ethanol production are presented in Figure 3.3A. We observed that only 4 of the 16 (25%) model-dominant strategies were also OptORF strategies (red bars), and none of the triple-deletion model-dominant strategies were OptORF strategies. This suggests that, when examining optimal OptORF strategies for higher numbers of gene knockouts, either model's predictions are likely to be similar at the maximum growth rate. However, the models may predict different ethanol production rates using the same gene deletion set for strategies which do not result in the maximum level of chemical production.

The CONGA results for model-dominant strategies for the production of lactate and succinate were quite different (data not shown). Here, 15 of the 30 model-dominant strategies are also OptORF strategies. Of these 15 strategies, 13 are *i*JR904-dominant strategies, with 11 involving the deletion of *mhpF* and *adhC* (thereby removing acetaldehyde dehydrogenase). When these two genes are deleted, ethanol synthesis is no longer possible in the *i*JR904 model, while the *i*AF1260 model can synthesize ethanol via a second pathway (Figure 3.4A). The double deletion of *mhpF* and *adhC* enables *i*JR904-dominant strategies for lactate and succinate production, with additional deletions determining whether lactate or succinate is the dominant product. We also observed that the *i*AF1260-dominant strategies for succinate production are all of low-yield (less than 10% the theoretical maximum). In fact, the *i*AF1260 model requires five gene deletions to obtain yields greater than 10% of the theoretical maximum, while the *i*JR904 model requires only two gene deletions. These results demonstrate that CONGA

can also be used to identify differences in the ease of coupling growth to chemical production in different models or organisms.

Figure 3.4. Flux maps illustrating differences in metabolic pathways in *E. coli* GENREs.



The text above each map indicates the pathway responsible for the phenotypic difference, the phenotype with which the strategy is associated, and the gene deletion for which the phenotype occurs. (**A**-**B**) Schematic views of the flux distributions associated with the indicated gene deletion set. Metabolites are represented in plain text. Metabolic transformations are indicated via arrows, with thicker arrows indicating higher flux. In some instances, multiple transformations are combined into a single dashed arrow or lumped into a subsystem. Subsystems are indicated by plain text enclosed in a grey rectangle. Fluxes active in the *i*AF1260 network are in red, fluxes active in the *i*JR904 network are in blue, and inactive fluxes are in grey. If gene (reaction) deletions occur in the fermentation pathway, they are indicated by black 'X's. Fluxes crossing the dashed boundary indicate transport to the extracellular environment. Metabolite abbreviations: 12ppd, 1,2-propanediol; 2aobut, L-2-Amino-3-oxobutanoate; actp, acetyl phosphate; athr, allo-threonine. All other abbreviations match those used in the *i*Syp611 metabolic model (see Dataset S2 in the original publication).

We then set out to investigate which network differences between *i*JR904 and *i*AF1260 account for the production differences associated with each gene deletion set found by CONGA. Of the 46 total model-dominant strategies, 34 (74%) could be attributed to at least one of six metabolic differences between the two models (Table 3.1). The remaining 12 model-dominant strategies predicted production differences of less than 10% the theoretical maximum yield, and in many cases much less. Two of the network differences (1,2-propanediol synthesis and hexokinase) were associated only

with *i*JR904-dominant ethanol production strategies, while others were responsible for more than one set of model-dominant strategies. For example, differences in the succinate transport mechanism were implicated in strategies associated with *i*AF1260-dominant production of ethanol and lactate, and with *i*JR904-dominant production of succinate.

Table 3.1. Explanation of metabolic differences between the *i*JR904 and *i*AF1260 models of *E. coli*.

Metabolic Difference	**Description of Metabolic Difference**	**Functional Effect**
1,2-Propanediol Synthesis	The *i*AF1260 model has the ability to secrete 1,2-propanediol; the *i*JR904 model does not.	The ability to convert glucose to 1,2-propanediol gives the *i*AF1260 model greater flexibility in choosing fermentation products under some conditions.
Aldehyde Dehydrogenase	The *i*AF1260 model has a unique aldehyde dehydrogenase which the *i*JR904 model lacks.	This reaction grants the *i*AF1260 model the ability to convert acetaldehyde to acetate using NADP. This reaction was selected for deletion by CONGA in *i*JR dominant strategies, but was never directly implicated in a solution.
Ethanol Synthesis	The *i*AF1260 model has unique reactions to convert acetyl-CoA to acetaldehyde which the *i*JR904 model lacks.	Deletions are possible in which the *i*JR904 model produces no ethanol while the *i*AF1260 model produces ethanol at high levels.
Hexokinase	The *i*AF1260 model has a unique hexokinase that it can use as an alternative to phosphoglucose isomerase (PGI).	The *i*AF1260 model has the ability to recover from multiple-reaction deletions containing PGI, while the *i*JR904 model does not.
Hydrogen Transport	The *i*AF1260 model has the ability to secrete hydrogen gas; the *i*JR904 model does not.	The ability to secrete hydrogen gas allows the *i*AF1260 model to convert formate to CO_2 and H_2, consuming a proton in the process. This provides the *i*AF1260 model an additional way to consume cytoplasmic H^+, and changes the preferred fermentation products under some conditions.
Succinate Transport	The *i*AF1260 model employs a hydrogen antiporter for succinate; the *i*JR904 model employs a hydrogen symporter.	Production of succinate becomes less energetically favorable in the *i*AF1260 model, as the synthesis route consumes fewer cytoplasmic protons.

Six metabolic differences accounted for the majority of the model-dominant strategies identified by CONGA.

Many of these network differences affect the balance of possible fermentation products (Figure 3.4). For example, the *i*AF1260 network contains an additional pathway to convert acetyl-CoA to ethanol via L-2-amino-3-oxobutanoate and allo-threonine (Figure 3.4A). As noted above, this extra pathway for ethanol synthesis in the *i*AF1260 model carries flux in many of the *i*JR904-dominant lactate and succinate production strategies, demonstrating that a single network difference can be found under multiple simulation conditions. In other instances, network differences affect flux balances outside the central fermentation pathways (Figure 3.4B). For example, when the genes *edd* (or *eda*), *tpiA*, and *fsaB* are deleted, disrupting glycolysis and the Entner-Doudoroff pathway, the *i*JR904 and *i*AF1260 models produce different products. The *i*JR904 model converts glucose into ribose-5-phosphate (r5p) via the oxidative and non-oxidative branches of the pentose phosphate pathway. The r5p is then converted to deoxyribose-5-phosphate and broken down into glyceraldehyde-3-phosphate (g3p), which enters glycolysis, and acetaldehyde (acald), which gets converted to ethanol. In contrast, the *i*AF1260 model converts glucose to g3p and dihydroxyacetone phosphate (dhap). As in the *i*JR904 model, g3p enters glycolysis, while dhap enters the methylglyoxal (mthgxl) pathway. Some of the mthgxl is converted to 1,2-propanediol (12ppd) via a unique 12ppd synthesis pathway, while the remaining mthgxl continues through the pathway to make pyruvate.

After identifying the metabolic differences that lead to model-dominant strategies, we modified the *i*JR904 and *i*AF1260 networks to contain identical representations of each pathway (Dataset S2 in the original publication) and re-evaluated the phenotype predictions of each knockout strategy. After the network reconciliation, we found that all but one of the knockout mutants are now predicted to have similar production rates (Figure 3.3B).

While other studies have identified functional differences between the *i*JR904 and *i*AF1260 models with respect to growth phenotypes (e.g., gene essentiality predictions [54]) using an enumerative approach, here we compared the two reconstructions with respect to their metabolic engineering predictions using an algorithmic approach that identifies just those conditions resulting in different model predictions. We hypothesized that coupling of metabolites to biomass would be more difficult in the larger *i*AF1260 model, and that the model might have higher production levels (or larger production ranges if multiple products are possible), due to the larger network containing more ways to balance internal fluxes. These hypotheses were not borne out (with the notable exception of coupling succinate production to biomass), as we were able to predict similar production levels using both models. In fact, the production differences we did observe were due to only 21 reactions that represent just 3.5% of the 594 unique metabolic reactions in the *i*AF1260 model (described previously in [54]).

3.1.3: Cyanobacterial Metabolic Differences

Having analyzed two models of the same organism, we then sought to analyze two models of closely related but distinct organisms, and to examine organisms less well-studied than *E. coli*, to see if CONGA can be used to generate new physiological insights. For this application, we selected two cyanobacteria, *Synechococcus sp.* PCC 7002 and *Cyanothece sp.* ATCC 51142. Very few genome-scale metabolic reconstructions of cyanobacteria have been published to date [68,110,144], and our group has recently developed two more, the *i*Syp611 model of *Synechococcus* (this paper) and the *i*Cce806 model of *Cyanothece* [242]. In order to gain insight into the metabolic similarities and differences between these two cyanobacterial strains, we used CONGA to identify gene deletion sets that were predicted to be lethal in only one cyanobacterial metabolic model, as well as to improve our draft *Synechococcus* reconstruction.

Table 3.2. Number of lethal gene deletion sets for the cyanobacterial models *i*Syp611 and *i*Cce806.

	***i*Syp611**	***i*Cce804**	**Interpretation**	**Example**
Genetic	20 (12)	22 (9)	A gene-protein-reaction (GPR) relationship differs between models.	The *i*Syp611 model has a unique isozyme for phosphoglucomutase.
Orthology	4 (4)	4 (4)	Genes encoding enzymes with identical functions cannot be assigned as orthologs.	Both organisms have annotations for dihydroorotase, but the genes are not matched as orthologs due to sequence dissimilarity.
Metabolic	4 (2)	10 (5)	One organism has an additional reaction which enables it to carry out a unique biochemical transformation.	The double deletion of glutamate dehydrogenase and glutamate synthase is lethal only in the *i*Cce806 model.
Mixed	2 (2)	0 (0)	More than one of the above types is implicated in the predicted phenotype difference.	The *Synechococcus* gene for malic enzyme (NADP-catalyzed) is predicted to be an ortholog to the *Cyanothece* gene for malic enzyme (NAD-catalyzed) (orthology difference). The *i*Cce806 has both NAD- and NADP-catalyzed versions of malic enzyme (metabolic difference).
Total	30 (20)	36 (18)		

Functional network differences were classified into one of four types based on their biological interpretation. In many cases, different gene deletion sets led to the same reaction deletion set. The number of unique reaction deletion sets is given in parentheses.

We first applied CONGA to the draft *i*Syp611 model. Some of the gene deletion sets identified by CONGA arose due to missing genes in the draft *i*Syp611 model. For example, CONGA identified gene deletion sets containing protein synthesis enzymes present only in the *i*Cce806 network. *Synechococcus* also has these proteins, but they had not been included in the model. Other network differences arose due to incomplete GPR associations in the draft *i*Syp611 model. For example, the *i*Cce806 model associated HisB with both histidinol-phosphatase and imidazoleglycerol-phosphate dehydratase, while the draft *i*Syp611 network only associated the protein with histidinol-phosphatase. The original annotation indicated the gene was bifunctional, and

the draft *i*Syp611 model was updated accordingly. This approach increased the size of the *i*Syp611 model from 542 to 611 genes, an increase in gene content of 13%. This increase in gene content is comparable to that seen in metabolic network reconciliation [159], which was used to expand the gene content of genome-scale models of *Pseudomonas aeruginosa* and *Pseudomonas putida* by 3% and 18%, respectively.

Table 3.3. Explanation of metabolic differences between the cyanobacterial models iSyp611 (Synechococcus) and iCce806 (Cyanothece).

Reaction Deletion Set	**Lethal In**	**Explanation of Metabolic Difference**
PDH	*i*Syp611	Acetyl-CoA synthase (ACS), pyruvate dehydrogenase (PDH), and phosphotransacetylase (PTA) are responsible for acetyl-CoA synthesis. The *i*Syp611 model requires PDH to supplement the activity of ACS, while the *i*Cce806 model requires PTA. Thus, the deletion of PDH is lethal only in the *i*Syp611 model.
MDH and ME2	*i*Syp611	Fumarate, produced as a byproduct of arginine biosynthesis, is converted to malate and then to oxaloacetate (by malate dehydrogenase, MDH). In the absence of MDH, malic enzyme (ME) can instead convert malate to pyruvate. The *i*Syp611 model contains NADP-catalyzed malic enzyme (ME2), while the *i*Cce806 model contains both NADP- (ME2) and NAD-catalyzed (ME1) malic enzyme. Thus, the deletion of MDH and ME2 is lethal only in the *i*Syp611 model.

We identified two unique reaction deletion sets lethal only in the *i*Syp611 model. We identified a total of seven metabolic differences between the two models.

After refining the draft model based on these results, the resulting model (*i*Syp611) was compared again to *i*Cce806 using CONGA. We identified 30 gene deletion sets that are lethal only in the *i*Syp611 model and 36 gene deletion sets that are lethal only in the *i*Cce806 model (Table 3.2). We found that in many instances different gene deletion sets mapped to the same set of reaction deletions (or *reaction deletion set*). For example, we identified six gene deletion sets lethal in the *i*Syp611 model that all mapped to photosystem II. As a result of these and other redundancies, the 30 gene deletion sets for the *i*Syp611 model reduced to 20 unique reaction deletion sets, and the 36 gene deletion sets for the *i*Cce806 model reduced to 18 unique reaction deletion sets.

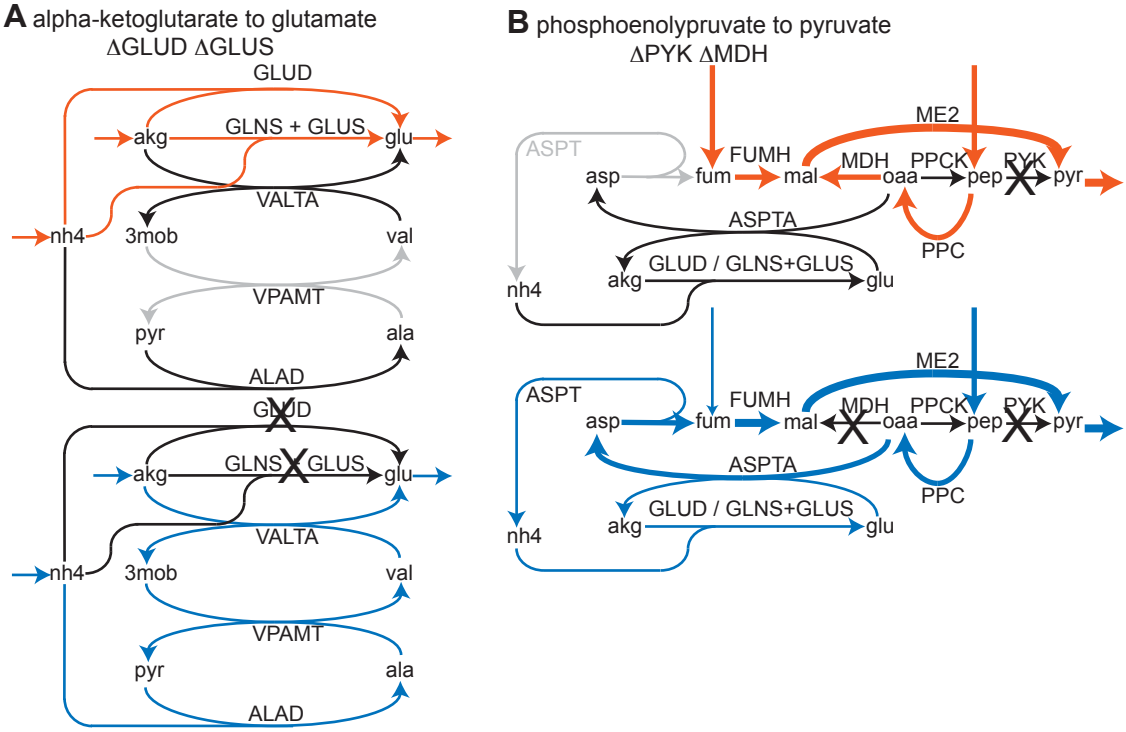
Table 3.4. Explanation of metabolic differences between the cyanobacterial models iSyp611 (Synechococcus) and iCce806 (Cyanothece).

Reaction Deletion Set	**Lethal In**	**Explanation of Metabolic Difference**
ASNS1	*i*Cce806	This reaction synthesizes asparagine. The *i*Syp611 model does not contain this reaction, because *Synechococcus* instead aminates aspartyl-tRNA to asparaginyl-tRNA prior to protein synthesis.
PQPCOR	*i*Cce806	*Cyanothece* is unique among the two cyanobacteria in using plastocyanin during photosynthesis. Hence, the *i*Cce806 model contains the reaction PQPCOR, while the *i*Syp611 model does not.
PTA	*i*Cce806	Acetyl-CoA synthase (ACS), pyruvate dehydrogenase (PDH), and phosphotransacetylase (PTA) are responsible for acetyl-CoA synthesis. The *i*Cce806 model requires PTA to supplement the activity of ACS, while the *i*Syp611 model requires PDH. Thus, the deletion of PTA is lethal only in the *i*Cce806 model.
GLUD and GLUS	*i*Cce806	GLUD (glutamate dehydrogenase) and GLUS (glutamate synthase) synthesize glutamate from alpha-ketoglutarate. This step incorporates ammonia into the metabolism and begins amino acid synthesis. The *i*Syp611 model has an extra reaction, valine-pyruvate aminotransferase (VPAMT), which allows it to recover from this deletion. Under the deletion scenario, ammonia gets combined with pyruvate to make alanine. Alanine is converted to valine which in turn is convered to glutamate.
MDH and PYK	*i*Cce806	Pyruvate synthesis is necessary to meet biomass demands. Pyruvate is normally synthesized from phosphoenolpyruvate via pyruvate kinase (PYK). In the absence of PYK, pyruvate can be synthesized from malate. Malate is produced as a result of biomass demands for arginine and tetrahydrofolate, but in insufficient levels to meet demand. Malate dehydrogenase (MDH) can make up for the demand by converting oxaloacetate to malate. As a consequence, deletion of both genes is lethal. The *i*Syp611 model has the unique reaction aspartase (ASPT), which it can use instead of MDH to convert oxaloacetate to malate, by way of aspartate. As a consequence, MDH function is no longer required in the absence of PYK, and the double deletion is nonlethal.

We identified five unique reaction deletion sets lethal only in the *i*Cce806 model. We identified a total of seven metabolic differences between the two models.

Of the four types of functional network differences, we were most interested in metabolic differences, although the other types are also important. For example, genetic differences may occur because the genes encoding an essential protein have not yet been identified in one organism. In total, the metabolic differences accounted for 4 of 30 gene

Figure 3.5. Identified metabolic differences in cyanobacteria..



(**A**) Top: Pathways for synthesis of glutamate (glu) from alpha-ketoglutarate (akg) used in *i*Cce806. Bottom: Pathway predicted by the *i*Syp611 model when glutamate dehydrogenase (GLUD) and glutamate synthase (GLUS) are deleted. Valine aminotransferase (VPAMT) enables the synthesis of glutamate from pyruvate (pyr). (**B**) Top: Pathway for conversion of phosphoenolpyruvate (pep) to pyruvate when pyruvate kinase (PYK) is deleted from *i*Cce806. Bottom: Pathway predicted by the *i*Syp611 model when malate dehydrogenase (MDH) is also deleted. Aspartase (ASPT) allows malate (mal) to be synthesized entirely from fumarate (fum), rather than from fumarate and oxaloacetate (oaa). (**A** and **B**) Red arrows indicate flux in the *i*Cce806 model. Blue arrows represent flux in the *i*Syp611 model under the indicated knockout condition. Black arrows indicate inactive reactions and reaction deletions are indicated by black 'X's. Gray arrows (top panels) indicate reactions not present in the *i*Cce806 model. Arrow thickness corresponds to relative flux levels. Reaction and metabolite abbreviations are identical in the *i*Syp611 and *i*Cce806 models and are given in the Dataset S2 in the original publication.

deletion sets (or 2 of 20 reaction deletion sets) for the *i*Syp611 model (Table 3.3) and 10 of 36 gene deletion sets (or 5 of 18 reaction deletion sets) for the *i*Cce806 model (Table 3.4).

Two of the reaction deletion sets which are lethal only in the *i*Cce806 model require deletion of two reactions from both models (Figure 3.5). In the first deletion set (Figure 3.5A), deletion of glutamate dehydrogenase and glutamate synthase prevents

the *i*Cce806 model from synthesizing glutamate. The *i*Syp611 model has a unique reaction, valine amino-transferase (VPAMT), which allows it to recover from this double deletion (blue arrows). In the second deletion set (Figure 3.5B), deletion of pyruvate kinase and malate dehydrogenase prevents the *i*Cce806 model from making pyruvate. The *i*Syp611 model has another unique reaction, aspartase (ASPT), which enables it to produce pyruvate and recover from the double deletion. A search of the *Cyanothece* genome failed to reveal candidate genes for ASPT and VPAMT, lending support to the hypothesis that they may be true metabolic differences between the two cyanobacteria.

CONGA reveals differences that can be used to reconcile and improve genome-scale metabolic models of closely-related species. We intend to use the remaining genetic and orthology differences found by CONGA as a starting point in further updating our reconstruction, as they may indicate missing or incorrectly annotated genes. CONGA can also identify differences in metabolic capabilities between models: our analysis here indicates *Synechococcus* and *Cyanothece* share a significant number of pathways, with important differences in central and amino acid metabolism.

3.1.4: Drug Targeting in Human Pathogens

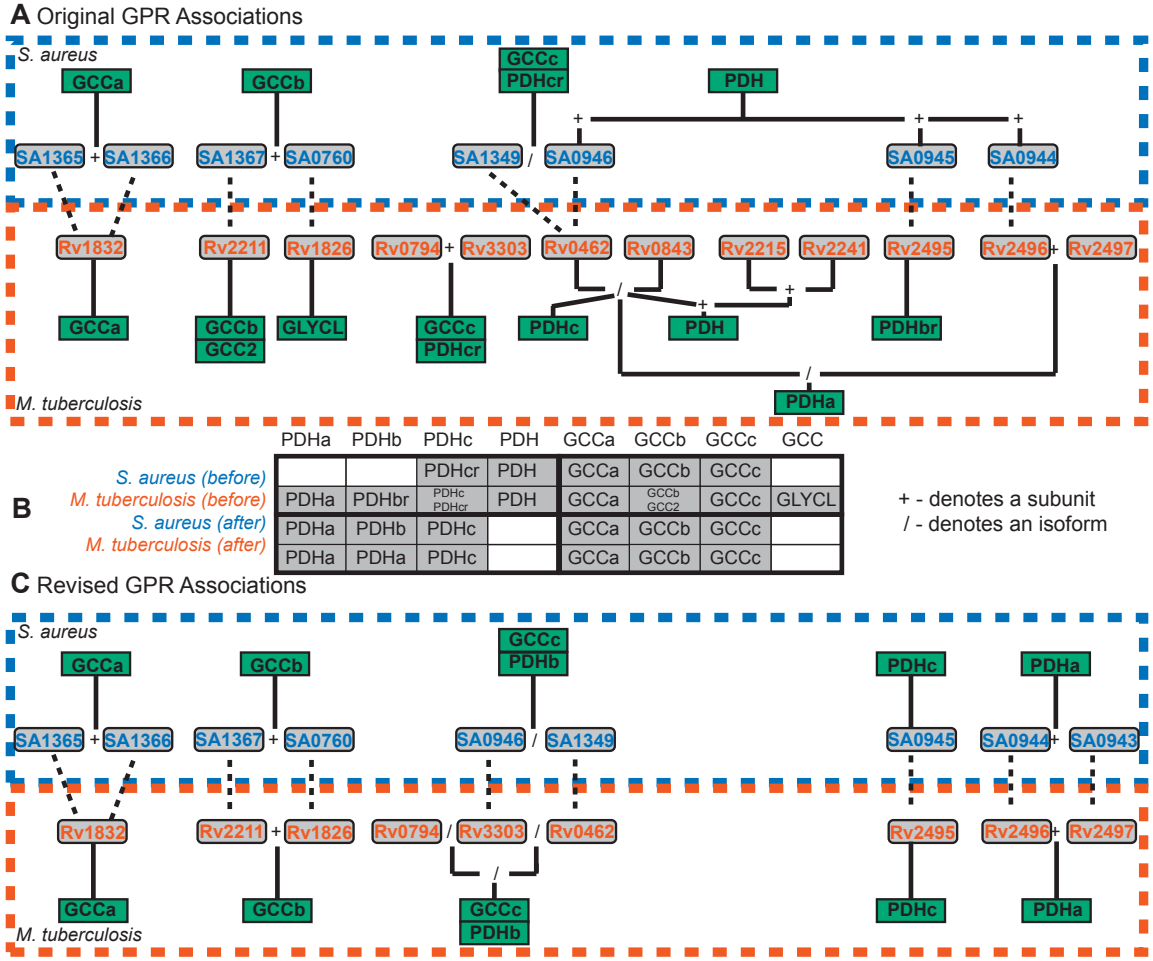
While were able to idenitfy metabolic differences between the two cyanobacteria, many of the differences identified by CONGA were not due to reaction-level differences. We thus sought to use CONGA to explore differences in metabolic capabilities between two dissimilar oganisms, and to exploit those differences to identify organism-specific drug targets. For this application, we applied CONGA to existing models of two phylogenetically distant human pathogens, the *i*NJ661 model of *M. tuberculosis* [95] and the *i*SB619 model of *S. aureus* [17], in order to explore differences in pathogenicity and drug resistance based on differences in reaction and gene content. As with our analysis

of the cyanobacterial models, we sought genetic perturbation strategies that were predicted to be lethal in only one organism.

Our preliminary analysis identified a total of 168 unique gene deletion sets, of which 139 (83%) could be traced in whole or in part to genetic or orthology differences. As these differences made up the majority of identified differences, we manually evaluated the quality of the orthology assignments and the original GPR associations. This analysis resulted in the modification of the GPR associations for 19 reactions in the *i*SB619 model and 36 reactions in the *i*NJ661 model (Dataset S2 in the original publication). As a result of these changes, 7 genes were eliminated from and 3 added to the *i*SB619 model, with 10 genes eliminated from and 4 added to the *i*NJ661 model.

A number of these initial genetic- and orthology-related gene deletion sets arose due to different representations of the glycine cleavage complex (GCC) and pyruvate dehydrogenase system (PDH) in the two models (Figure 3.6A). Both GCC and PDH are composed of three separate enzymes (a, b, and c), each of which carries out a distinct catalytic activity. Deletion of GCC is predicted to be lethal in both organisms, and because one subunit is shared by GCC and PDH, deletions to one complex may affect the other. In its original form, the *i*SB619 reconstruction modeled PDH as an overall reaction, and GCC via its three individual reactions (Figure 3.6B). In contrast, the *i*NJ661 model represented both PDH and GCC as individual and overall reactions (Figure 3.6B). Due to these differences, a number of ortholog deletions are lethal in only one model. For example, deletion of the ortholog pair (SA0945, Rv2495) deletes PDH from the *i*SB619 network, but only deletes PDHb from the *i*NJ661 network. The deletion is lethal only in the *i*SB619 model. We thus revised the GPR associations for these complexes to give a consistent representation between the two models (Figure 3.6C). These changes also required changes to the stoichiometric matrices in each model. (See Dataset S2 in the original publication for details.)

Figure 3.6. Example adjustment of pathogen models following preliminary analysis.



(A) Original model annotations for the glycine cleavage (GCC) and pyruvate dehydrogenase (PDH) complexes. Green boxes represent reactions and gray boxes represent genes. *S. aureus* loci are in blue text and *M. tuberculosis* loci are in red text. Dashed lines indicate orthologs and solid lines connect genes to reactions. SA1365 and SA1366 are orthologous to the N-terminus and C-terminus of Rv1832, respectively, and together are orthologous to the entire Rv1832 sequence. A '+' sign between genes indicates a complex; a '/' sign indicates isozymes. **(B)** The two models were inconsistent in their representation of these two enzyme complexes. This table indicates the presence or absence of individual (a, b, c) and lumped (GCC, PDH) reactions before and after model adjustments. The shaded gray boxes indicate the presence of a particular function, and the small black text indicates that model's specific reaction. **(C)** Revised model annotations for the GCC and PDH complexes. The color scheme is the same as in **A**.

We applied CONGA again after this initial reconciliation, and identified 71 gene deletion sets lethal only in the *i*SB619 model and 84 gene deletion sets lethal only in the *i*NJ661 model (Table 3.5). Of these, a total of 99 gene deletion sets (64%) were still due

to genetic or orthology differences. Nevertheless, CONGA identified 18 gene deletion sets arising from metabolic differences which were lethal only in the *i*SB619 model, and 38 such gene deletion sets lethal only in the *i*NJ661 model. As with the cyanobacteria, in some instances multiple gene deletion sets mapped to the same reaction deletion set (Table 3.5). Of these, we examined only those gene deletion sets arising from metabolic differences, and identified 17 unique reaction deletion sets lethal only in the *i*SB619 model and 28 unique reaction deletion sets lethal only in the *i*NJ661 model.

Table 3.5. Number of lethal gene deletion sets for the human pathogen models iSB619 and iNJ661.

	***i*SB619**	***i*NJ661**	**Interpretation**	**Example**
Genetic	3 (3)	13 (8)	A gene-protein-reaction (GPR) relationship differs between models.	Only the *i*SB619 model has a gene associated with sulfur reductase.
Orthology	17 (17)	14 (14)	Genes encoding enzymes with identical functions can- not be assigned as orthologs.	Both organisms have putative annotations for chorismate mutase, which are not matched as orthologs due to sequence dissimilarity.
Metabolic	18 (17)	38 (28)	One organism has an additional reaction which enables it to carry out a unique biochemical transformation.	The deletion of homoserine kinase is lethal only in the *i*SB619 model.
Mixed	33 (26)	19 (11)	More than one of the above types is implicated in the predicted phenotype difference.	Only the *i*NJ661 model has a gene associated with phosphoserine transaminase (genetic difference). This reaction deletion is nonlethal in the *i*SB619 model because it can utilize alternative pathways to perform this function (metabolic difference).
Total	71 (63)	84 (61)		

Functional network differences were classified into one of four types based on their biological interpretation. In many cases, different gene deletion sets led to the same reaction deletion set. The number of unique reaction deletion sets is given in parentheses.

These 45 unique reaction deletion sets served as the starting set of potential drug targets. We employed a multi-step process to reduce these reaction deletion sets to a set of candidate antibiotic targets. First, because genes may be associated with more than

Table 3.6. Potential drug targets in the human pathogens *S. aureus* and *M. tuberculosis*.

Organism	**Reaction Deletion Set**	**Subsystem**	**Known Drugs**
S. aureus	ALATA_D	Cell wall synthesis	Vancomycin [252]
S. aureus	DHFS	Cofactor synthesis	Trimethoprim and Sulfonamides [131]
S. aureus	KAS11 or KAS12 or KAS13	Cell membrane synthesis	Small molecules [85,154]
S. aureus	NNAM	Cofactor synthesis	Small molecules [67]
S. aureus	TECA1S or TECA2S or TECA3S or TECA4S	Cell wall synthesis	Vancomycin [252]
M. tuberculosis	CHRPL	Cell membrane synthesis	None
M. tuberculosis	FACOAL80 or FACOAL160 or FACOAL200 or FACOALPHDCA	Cell wall synthesis	Small molecules [9]
M. tuberculosis	FAS80_L or FAS100 or FAS120 or FAS140 or FAS160 or FAS180 or FAS200 or FAS240_L or FAS260 or FASPHDCA	Cell wall synthesis	Pyrazinamide [26]
M. tuberculosis	FASm220 or FASm240 or FASm260 or FASm280 or FASm300 or FASm320 or FASm340 or FASm2201 or FASm2202 or FASm2401 or FASm2402 or FASm2601 or FASm2602 or FASm2801 or FASm2802	Cell wall synthesis	Isoniazid [261,264]
M. tuberculosis	MCBTS	Siderophore synthesis	Small molecules [152]
M. tuberculosis	PREPPACPH	Cell membrane synthesis	None
M. tuberculosis	PPTGS or PPTGS_TB1 or PPTGS_TB1 or UDCPDP	Cell wall synthesis	Ethambutol [261,264]

We identified five unique reaction deletion sets lethal only in the *i*SB619 model, and seven unique reaction deletion sets lethal only in the *i*NJ661 model. From these, we identified 10 candidate antibiotic targets in *S. aureus* and 37 candidate antibiotic targets in *M. tuberculosis*. Antibiotics targeting some of these reactions have already been developed.

one reaction, we eliminated from each unique reaction deletion set any reactions that were nonessential to the set. For example, CONGA identified the deletion of SA1487 as lethal in *S. aureus*, leading to the reaction deletion set DHFS and THFGLUS. However, the deletion of THFGLUS is not lethal, so THFGLUS was removed from the reaction

deletion set, giving the reduced reaction deletion set DHFS. We then examined the reduced reaction deletion sets and eliminated those sets where more than one reaction deletion was required to give a lethal prediction. Such reaction deletion sets are likely to be poor candidates for potential drug targets, because they may require development of a multiple-drug treatment strategy. For example, CONGA identified the reaction deletion set RNDR1, RNDR4 as being lethal in *M. tuberculosis*, with both reaction deletions necessary to give a lethal prediction. This set was subsequently eliminated from the set of candidate antibiotic targets. Finally, we eliminated those reactions included in the Recon 1 genome-scale metabolic model of human metabolism [49], as drugs targeting these reactions may cause adverse side-effects in humans. This procedure yielded 10 reactions as candidate antibiotic targets in *S. aureus* and 37 reactions as candidate antibiotic targets in *M. tuberculosis* (Table 3.6).

Many of the candidate antibiotic targets are already targeted by existing antibiotics (Table 3.6), demonstrating that our approach can correctly identify candidate metabolic functions for drug targeting. Most of the reactions for which antimicrobials exist are involved in cell wall and cell membrane synthesis. While both organisms require these biosynthetic capabilities, their cell walls and membranes are structurally different, and so different proteins and reactions are required. These differences are reflected in the standard antimicrobial treatments for these two pathogens. For example, vancomycin binds to the D-alanine terminus of peptidoglycan and prevents the incorporation of teichoic acids into the matrix [252]. Mycobacteria, such as *M. tuberculosis*, have structurally distinct cell walls, for which isoniazid, ethambutol, and pyrazinamide are required treatments [26,261,264]. We were also able to find reports of small molecule inhibitors of fatty acid synthesis in both *S. aureus* [85,154] and *M. tuberculosis* [9].

We also identified a variety of other metabolic functions which antibiotics do not yet target. For example, the *i*SB619 model requires tetrahydrofuran (THF) and NAD to produce biomass. Unfortunately, many staphylococci are already resistant to inhibitors of THF synthesis [131], while inhibitors of the nicotinamidases *S. aureus* uses for NAD synthesis have only recently been identified [67]. However, *M. tuberculosis* can grow in media lacking THF and NAD [95], suggesting the lack of THF and NAD in the *i*NJ661 biomass equation may reflect a model development choice, rather than a biological difference. We identified *M. tuberculosis*' unique use of siderophores for iron transport, for which biosynthesis inhibitors have been identified [152]. We also identified mycobacteria's use of unique glycolipids, but we were unable to identify inhibitors that have been reported in the literature, making glycolipid synthesis a potential new target for new *M. tuberculosis*-specific antibiotics. Of the remaining organism-specific metabolic functions, two candidate antibiotic targets (nicotinamidase in *S. aureus* and siderophore synthesis in *M. tuberculosis*) had not been identified by previous computational studies of these models [17,95].

By comparing pathogens against each other, we are able to identify essential functions unique to a particular pathogen. This enables the identification of narrow-spectrum antibiotics tailored to individual pathogens. It is believed that the use of such antibiotics can overcome multi-drug resistance through novel mechanisms of action [81,243] and slow the rate of resistance transfer across species [33,153]. We believe our framework provides a rapid means of identifying unique metabolic functions as possible targets for new antimicrobials, and will provide a useful tool for combating the rapid rise of multi-drug resistant bacteria.

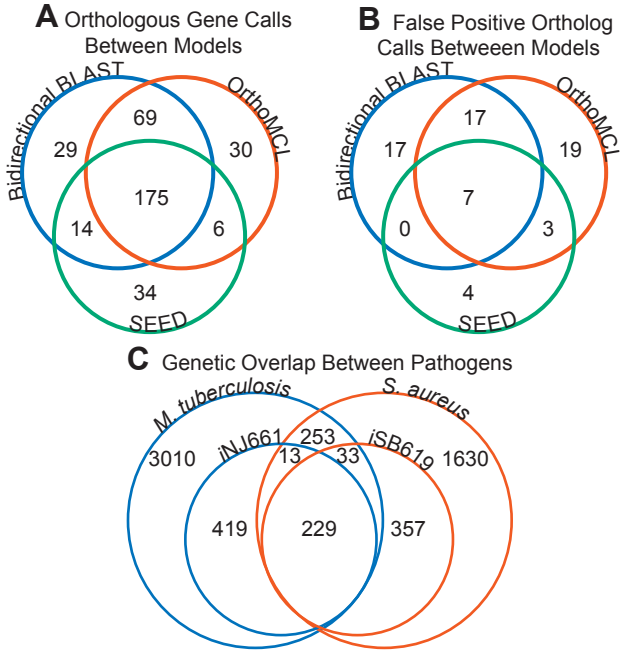
3.1.5: Assessment of Ortholog Calling Methods

Before CONGA can be applied to a pair of metabolic models, a gene-level alignment must be performed. We perform this alignment by identifying the orthologous genes between the two models, and we force CONGA to select ortholog pairs as a single unit. Prior to applying CONGA to the pathogen models, we examined three methods for identifying orthologous genes (Figure 3.7). The first method utilized a BLAST search [4] to identify those pairs of *M. tuberculosis* and *S. aureus* genes which were mutual best-BLAST hits of each other, called *bidirectional best-BLAST hits*. An E-value of 10^{-5} was employed as a cutoff. The second method used OrthoMCL [124] to identify pairs of genes belonging to the same ortholog group (a cross-taxa group of genes in which all genes are bidirectional best-BLAST hits of one another). The last method utilized the SEED [169] to identify genes belonging to the same FIGfam (sets of proteins homologous along their entire length).

We first identified ortholog pairs where both the *M. tuberculosis* and *S. aureus* genes were included in the *i*NJ661 and *i*SB619 models, respectively. We found that the number and content of ortholog calls depended on the method used (Figure 3.7A). The bidirectional best-BLAST search identified a total of 287 of a possible 619 genes. SEED identified the fewest, with only 229 orthologs. Of these, 175 orthologs were common to all methods, with smaller numbers of orthologs being shared by pairs of methods.

We also analyzed the three methods for false positive ortholog calls (Figure 3.7B). A false positive ortholog call is one in which two orthologs are associated with different reactions in their respective models. We found that all three methods identified 7 ortholog pairs for which model annotations were distinctly different (Table 3.7). SEED identified the fewest additional false positives, giving 14 total. Full details of orthologs assigned by each method can be found in Dataset S3 in the original publication. We then analyzed the effect of each ortholog calling method on the gene deletion sets

Figure 3.7. Comparison of ortholog identification methods for *S. aureus* and *M. tuberculosis*.



(A) Number of model genes identified as orthologs by each of the three methods. Only the orthologs present in both models are included in the diagram. Overlapping areas indicate orthologs identified by one or more methods. **(B)** Number of false positive orthology assignments made by each of the three methods. A false positive orthology assignment indicates the two genes are associated with different reactions in their respective models. Overlapping areas indicate false positives identified by one or more methods. **(C)** Overlap of gene content between the two pathogens, based on SEED FIGfams. Smaller circles represent genetic content of the two models, with the larger circles representing the entire genome. Numbers within overlapping areas indicate numbers of orthologs.

identified by CONGA. We found that using orthologs identified by bidirectional best-BLAST and OrthoMCL yielded numerous gene deletion sets containing false positive ortholog pairs. In contrast, the number of gene deletion sets containing true ortholog pairs was relatively insensitive to the method used to call orthologs. We thus chose to perform all simulations using SEED orthologs.

Using the orthologs identified by SEED, we then assessed the metabolic overlap between the two models (Figure 3.7C). In addition to the 224 orthologs present in both models, the *i*SB619 model contains 33 genes with orthologs that are not included in the *i*NJ661 model, and the *i*NJ661 model contains 13 genes with orthologs that are not in

Table 3.7. False positive ortholog calls in the *i*SB619 (*S. aureus*) and *i*NJ661 (*M. tuberculosis*) human pathogen models.

Ta		***M. tuberculosis***	
SA0486	Glutamyl-tRNA synthetase	Rv2992	Alanyl-tRNA synthetase
SA0760	Glycine cleavage complex, subunit B	Rv1826	Glycine cleavage complex, entire complex
SA1059	Methionyl-tRNA synthetase	Rv1406	Methionyl-tRNA formyltransferase
SA1131	2-oxoglutarate synthase	Rv2455	Ferredoxin oxidoreductase
SA1132	2-oxoglutarate synthase	Rv2454	Ferredoxin oxidoreductase
SA1519	L-alanine, glycine, and L-serine transport via ABC system	Rv1704	D-alanine, D-serine, glycine, and L-serine transport via proton symport
SA2467	Imidazole-glycerol-3-phosphate synthase	Rv1602	Glutamine phosphoribosyldiphosphate amidotransferase

All three methods for assigning ortholog pairs identified seven pairs of orthologs which carried out different functions in the *i*SB619 and *i*NJ661 models.

the *i*SB619 model. These 46 genes can likely be used to expand the scope of each model. Additionally, we identified 253 orthologs included in neither model. Using SEED, we were able to classify these 253 orthologs into subsystems and found that 45% were involved in protein, DNA, or RNA metabolism, while 15% were involved in non-metabolic functions such as cell division, regulation, and the stress response. An additional 22% were of unknown or uncertain function. The remaining 18% were spread across a variety of metabolic subsystems, with 35 of the 253 (8%) orthologs being involved in vitamin and cofactor synthesis. Many of these 35 genes are involved in the assembly of metal clusters and would not generally be included in a metabolic model. Finally, we observed that metabolic genes are enriched for members of an ortholog pair: 37% (229 of 619) of genes in the *i*SB619 model had orthologs in the *i*NJ661 model, while only 21% (528 of 2515) of genes in the *S. aureus* genome had orthologs in the *M. tuberculosis* genome (χ^2 P-value < 0.001).

3.2: Discussion

In this work, we developed a bilevel mixed-integer programming approach to identify the functional differences between networks by comparing network reconstructions aligned at the gene level. The constraint-based method first identifies a set of orthologous genes based on genome sequence, and then identifies conditions under which differences in gene content give rise to differences in metabolic capabilities. Our gene-centric approach allows for the rapid identification of *functional differences* between networks which can be traced back to the presence or absence of particular genes or reactions (*structural differences*) in one network or the other. We demonstrate that our algorithm can be used to identify genetic, orthology, and metabolic differences between reaction networks with applications in metabolic engineering, model development, and antibiotic discovery.

Increasingly, new genome-scale reconstructions are being created by identifying bidirectional best-BLAST hits against genomes for which models have already been constructed. GPR and reaction annotation information can then be copied into the new model (see for example [176,186,199,224,225]). Our results point to two possible challenges with this approach. First, a bidirectional best-BLAST search might not identify all orthologs: the *i*Syp611 model was constructed from a draft *i*Cce806 model containing 591 genes. Orthologs for 537 of these genes were copied to the *i*Syp611 model, representing a 9% gene loss. Of the 54 *Cyanothece* genes for which a bidirectional best-BLAST search did not identify orthologs in *Synechococcus*, manual curation identified orthologs for 26 of them. While these orthologs were not bidirectional best-BLAST hits, we decided the genes had sufficiently high sequence similarity and sufficiently similar annotations to be considered orthologs. (Annotations were collected from NCBI, the Kyoto Encyclopedia of Genes and Genomes (KEGG) [100], and SEED

[169].) This suggests that construction of new models using only bidirectional best-BLAST hits may exclude significant numbers of genes from new reconstructions. Second, using bidirectional best-BLAST hits to identify orthologs may also generate large numbers of false positive ortholog pairs. Our bidirectional best-BLAST comparison of the manually curated *S. aureus* and *M. tuberculosis* models yielded 41 false positives (14% of the 287 orthologs, where a false positive indicates orthologs were associated with different metabolic reactions). If one model had been created from the other, these genes would have incorrect reactions associated with them. Manual assessment of the cyanobacterial bidirectional best-BLAST hits yielded 35 (of 537, or 7%) false positive orthologs in the draft *i*Syp611 model, which were subsequently removed from the final reconstruction. Thus, false positive ortholog calls represent a significant problem even for closely-related organisms.

Our approach represents a significant advance in comparing genome-scale network reconstructions. CONGA is a single instance of a broader approach, in which two different networks are compared and analyzed for functional differences. This represents a significant advance over existing model-comparison approaches [90,159,237], which typically do not identify the effect of network differences on achievable functional states. However, CONGA is not a replacement for more exhaustive approaches such as jamborees or network reconciliation: CONGA will not lead to the identification of all structural differences between models, just those causing different functional states. For example, a reaction-level alignment of the *i*Syp611 and *i*Cce806 models identified 172 reactions unique to the *i*Cce806 model and 57 reactions unique to the *i*Syp611 model. Of these 229 reaction differences, 126 cannot be utilized under the photoautotrophic conditions studied here. Of the remaning 113 unique reactions, only 15 were identified by CONGA as leading to differences in gene essentiality in the two cyanobacterial models under carbon-limited photoautotrophic conditions (when all genes are considered

for deletion). Additional reaction differences could be picked up by CONGA if other environments (e.g., dark fermentation), growth conditions (e.g., suboptimal instead of lethal gene deletions), and objective functions (e.g, chemical production rates) were considered, and if orphan reactions (those without a GPR association) could be deleted as well (since 20 of the 229 unique reactions did not have GPR associations). Despite the inability to identify all structural differences, CONGA can identify those gene (and thus reaction) differences which give rise to differences in predicted growth and production rates, as well as other phenotypes. As a result, we believe that it will be a useful tool to complement existing model reconciliation and comparison efforts, such as jamborees.

3.3: Methods

3.3.1: Formulation of Optimization Problem for Identification of Gene Deletion Sets

The CONGA framework employs a bilevel optimization problem to identify genetic perturbations which disproportionately change flux through a selected reaction (e.g., growth or by-product secretion) in one organism over another (Figure 3.1). The outer problem is a mixed-integer linear program (MILP) which finds gene deletions maximizing the flux difference between two reactions in different models. The two inner problems (one for each model) are flux-balance analysis (FBA) problems [165], linear programs (LPs) which maximize growth subject to reaction stoichiometry, thermodynamics, and enzyme capacities. We alter the FBA problems using deletions given by the outer problem. Gene-protein-reaction (GPR) constraints associate genes with reactions and are used to enforce the gene deletions given by the outer problem. These constraints are formulated using the logical relationships developed previously

[106]. CONGA can select any genes for deletion, with the restriction that orthologous genes present in both models be deleted simultaneously from both models.

The FBA formulation for each model's inner problem is shown below:

$$\textbf{max} \quad \sum_j c_j v_j \tag{3.1}$$

$$\textbf{s.t.} \quad \sum_i S_{i,j} v_j = 0 \qquad \forall i \in I \tag{3.2}$$

$$\alpha_j \le v_j \le \beta_j \qquad \forall j \in J \tag{3.3}$$

$$v_j = 0 \qquad \forall j \in J \mid y_j = 0 \tag{3.4}$$

Each reaction j in the set of reactions J has a flux given by v_j. The FBA objective is a linear combination of fluxes $\sum_j c_j v_j$, where c is a vector of weights. We choose to maximize for biomass alone, in which case c_j is a standard basis vector along biomass, and the objective is written as v_{BM}. Each reaction j consumes and produces some metabolites i in the set of metabolites I, with stoichiometry given by $S_{i,j}$. By conservation of mass, net production of each metabolite across the entire network must be zero at steady-state (3.2). Each reaction is constrained to have flux within an appropriate range as given by enzyme capacities and thermodynamics (3.3). For reactions deleted by the outer problem, a binary variable (y_j) takes a zero value ($y_j = 0$), and the corresponding flux v_j is constrained to zero (3.4).

On-off reaction states are given by the binary variable y and determined by GPR constraints embedded in the outer problem:

$$y_j = f(z_{\hat{g}}, w_{\hat{p}}) \;\; \forall GPR(j, \hat{p}, \hat{g}) \in J, P, G \tag{3.5}$$

Each gene g in the set of genes G, protein p in the set of proteins P, and reaction j in the set of reactions J has a corresponding binary variable z, w, and y, respectively, which determines the gene, protein, or reaction's on-off state. (See [106] for details.) Each reaction j with a known GPR association can be carried out by a subset of enzymes $\hat{p}$, and each enzyme is specified by the subset of gene products $\hat{g}$. The outer

problem selects one or more genes for deletion ($z_g = 0$), and the GPR constraints $GPR(j, \hat{p}, \hat{g})$ implement the necessary logical relationships to determine the set of deleted reactions ($y_j = 0$).

To identify lethal gene deletion sets, the outer problem identifies deletions such that the growth rate of one species (A) is maximized with respect to the other (B). So long as growth is unconstrained, an objective of the form

$$\textbf{max} \quad v_{BM_A} - v_{BM_B} \tag{3.6}$$

will first identify gene deletions lethal only in species B. Finally, additional constraints are added which impose a limit K on the total number of gene deletions,

$$\sum_g 1 - z_g \leq K \tag{3.7}$$

and which ensure that all pairs of orthologous genes are deleted in common:

$$z_{g_A} = z_{g_B} \; \forall (z_{g_A}, z_{g_B}) \in O \tag{3.8}$$

The set of orthologs O contains all pairs of genes z_{gA} and z_{gB} found to be orthologous between Species A and Species B.

The final formulation results from using (3.6) as the outer objective, and accumulating (3.1)-(3.5), (3.7), and (3.8) as constraints. Constraints (3.1)-(3.5) and (3.7) must be imposed for each species:

max	$v_{BM_A} - v_{BM_B}$		
s.t.	constraints (3.1)-(3.4)	$\forall$ Species A and B	
	constraint (3.5)	$\forall$ Species A and B	(CONGA)
	constraint (3.7)	$\forall$ Species A and B	
	constraint (3.8)		

3.3.2: Reformulation to Single-Level Optimization Problem

To facilitate the solution process, we reformulated the bilevel program as single-level MILP by replacing the inner maximization problems with their optimality conditions, in accordance with strong duality [59]. The strong duality theorem for a linear program states that, at optimality, the values of the primal and dual objectives are equal, and

the primal and dual variables satisfy the primal and dual constraints, respectively [59]. Thus, each inner problem, (3.1)-(3.4), can be replaced by formulating its dual, equating the primal and dual objectives, and accumulating the primal and dual constraints. This reformulation was first proposed for the bilevel strain design problem OptKnock [36] and has since been described for other bilevel problems [106,174,230].

This reformulation requires a new variable for each constraint of the inner problem. Each metabolite i must satisfy the mass balance, for which we introduce the unconstrained dual variable $u_{1,j}$. Active reactions are further constrained to be within the range $\alpha_j \le v_j \le \beta_j$, for which we introduce the positive dual variables $u_{2,j}$ and $u_{3,j}$, respectively. In many cases, α and β are assigned large, arbitrary values. To reduce the size of the reformulation, we eliminated the upper bound constraint ($v \le \beta$) and imposed the lower bound constraint ($\alpha \le v$) only on uptake fluxes and irreversible reactions, collectively the set J_{LL}. Finally, reactions removed by gene knockouts are constrained to zero flux, for which we introduce a free dual variable $u_{4,j}$. This allows the dual of each inner problem to be formulated as:

$$\textbf{min} \quad -\sum_j \alpha_j u_{3,j} \tag{3.9}$$

$$\textbf{s.t.} \quad \sum_i S_{i,j} u_{1,j} - u_{3,j} + u_{4,j} = c_j \qquad \forall j \in J \tag{3.10}$$

$$u_{3,j} = 0 \qquad \forall j \notin J_{LL} \tag{3.11}$$

$$u_{3,j} = 0 \qquad \forall j \in J \mid y_j = 0 \tag{3.12}$$

$$u_{4,j} = 0 \qquad \forall j \in J \mid y_j = 1 \tag{3.13}$$

$$u_{3,j} \ge 0 \qquad \forall j \in J \tag{3.14}$$

Constraints (3.11) to (3.13) can be implemented using big-M constraints [248] or using the GAMS/CPLEX indicator constraint facility (the latter was used in this work).

The single-level formulation can then be constructed by using (3.6) as the outer objective, equating the primal and dual objectives (3.1) and (3.9) for each network, including constraints (3.2) to (3.5), (3.7), and (3.10) to (3.14) for each network, and

adding constraint (3.8). Equating the primal and dual objectives of the inner problem gives

$$\sum_j c_j v_j = -\sum_j \alpha_j u_{3,j} \tag{3.15}$$

so that the final, single-level formulation can be expressed as:

max $v_{BM_A} - v_{BM_B}$

s.t. constraints (3.2)-(3.5), (3.7), (3.10)-(3.15) ∀ Species A and B
constraint (3.8)

We also implemented integer cut constraints [34] to allow the generation of multiple solutions.

3.3.3: Modifications for Identification of Model-Dominant Strategies

To identify model-dominant chemical production strategies in the *E. coli* models, we sought gene deletions maximizing chemical production in one model with respect to the other. For these simulations, a few modifications from the previous formulation are required. First, the outer objective, (3.6), was altered to reflect chemical production flux. The vector c was changed to a standard basis vector along the production flux of interest. We denote this objective as v_P.

$$\textbf{max} \quad v_{P_A} - v_{P_B} \tag{3.16}$$

Some knockout conditions result in a *nonunique phenotype* for a particular chemical, in which multiple chemical production values can occur at the maximum growth rate. Under such conditions, CONGA can artificially inflate flux differences between models, by choosing a large production rate in one model and a small production rate in the second. We thus imposed a tilted objective function on each inner problem, which maximizes biomass while imposing a small penalty (γ) on chemical production; this causes the inner problem to return the value of v_P representing the lowest expected flux through the reaction [57].

$$\textbf{max} \quad v_{BM} - \gamma v_P \tag{3.17}$$

Because flux values in general are not necessarily unique, this tilted objective is necessary whenever the fluxes whose difference is being maximzed (e.g., chemical production rates) differ from the fluxes maximized by each model (e..g, biomass). We found this formulation is sensitive to the value chosen for γ. If γ is too small, the tilt does not correctly return the value of v_P representing the lower bound through the reaction, and if γ is too large, the tilt returns solutions with a slightly suboptimal growth rate. For our comparison of the *E. coli* models, we found that setting $\gamma=10^{-4}$ avoided both of these problems. However, the tradeoff between growth rate and chemical production flux varies from model to model and product to product, suggesting our value of $\gamma = 10^{-4}$ may not be generally applicable. Modifying the inner objective in this way requires modifying the weight vector c in (3.15) to include the value $c_j = -\gamma$ where $j = P$. We also imposed this tilted objective function when using OptORF to identify the top deletion strategies for each model and product.

We also constrained the dual variables associated with the reaction deletion constraint to be between [-1, 1] to improve solver performance [106,107].

$$-1 \le u_{4,j} \le 1 \tag{3.18}$$

Finally, we constrained both models to have nonzero biomass. The final, single-level formulation can be expressed as:

max	$v_{P_A} - v_{P_B}$	
s.t.	constraint (3.2)-(3.5), (3.7), (3.10)-(3.15), (3.18)	∀ Species A and B
	$v_{BM} > 0$	∀ Species A and B
	constraint (3.8)	

Finally, we note that CONGA can be rewritten to consider a reaction alignment and reaction deletions, by redefining the set O and using reaction instead of gene deletions.

3.3.4: Reducing the Number of Variables: General Procedure

In order to reduce simulation times, we eliminated essential and blocked genes from consideration as possible deletions by CONGA. For each model, we performed single-gene deletion simulations with no constraints on uptake fluxes to identify essential genes (those required for cellular growth). Genes whose orthologs were essential in both models as well as essential genes without orthologs (collectively the set G_E) were then excluded from consideration by CONGA. Eliminating single-deletion essential genes from consideration enables CONGA to focus only on conditionally essential genes, as well as those deletions for which exhaustive searches of all combinations are computationally prohibitive.

Table 3.8. Variable Reduction Procedures.

Model	Size[1]	Essential	Blocked	Redundant	Selectable	% Original Size
*i*JR904	905	171	171	168	395	44
*i*AF1260	1260	215	279	253	523	42
*i*Cce806	806	214	178	---	414	51
*i*Syp611	611	254	131	---	226	37
*i*SB619	615	80	142	---	393	64
*i*NJ661	655	128	107	---	420	64

Numbers of genes eliminated via identification of essential, blocked, and redundant genes in each of the six models studied. The final number of genes left for CONGA to select from is also given. A solid black line indicates the procedure was not carried out on the indicated model.

Flux-variability analysis (FVA) [132] was also used to identify blocked reactions (those incapable of carrying flux), and genes encoding only blocked reactions were also identified (e.g., if a gene encodes both a blocked and a nonblocked reaction, the gene is not considered a blocked gene). As above, genes whose orthologs were blocked in both models as well as blocked genes without orthologs (collectively the set G_B) were then excluded from consideration by CONGA. Because blocked reactions cannot carry flux under any circumstances, we know a priori that blocked genes cannot contribute to functional network differences.

For each pair of models, we fixed the essential genes of each organism to be on ($z_g = 1$), thereby excluding them from consideration by CONGA. We also fixed the blocked genes of each organism to be off ($z_g = 0$), and excluded them from the gene-deletion constraint, $\sum_g 1 - z_g \leq K$. The remaining genes which do not have a fixed on-off state make up the selectable gene set, G_S (Table 3.8). For our *E. coli* comparisons, the sets G_E and G_B were identified on glucose media under anaerobic conditions (i.e., the simulation conditions). Because these simulations enforced a nonzero biomass requirement, essential genes cannot play a role in model-dominant gene deletion strategies. We also added all genes encoding transport reactions were also added G_E, allowing us to focus on enzymatic, rather than transport, reaction differences between the networks.

3.3.5: Reducing the Number of Variables: Additional Operations for *E. coli* Models

For our comparison of the *E. coli* models, we performed additional steps in order to improve the run-time performance of CONGA. First, we identified essential and blocked genes (as described above) on glucose media under anaerobic conditions (i.e., the simulation conditions). We also removed from consideration all genes associated with membrane transporters. These two steps forced CONGA to consider only metabolic genes that can be active under the simulation conditions. Finally, we developed a procedure to reduce the number of genes needed to determine the on-off state of each reaction, by identifying conserved sets of subunits and isozymes across models.

We first constructed and solved an optimization problem to determine the reactions which can be activated by each gene $\bar{g}$ in the model: we turn off all genes but $\bar{g}$, and identify those reactions which can be turned on ($y_j = 1$) by activating $\bar{g}$ alone,

subject to GPR constraints. We refer to the set of such reactions as the *activated reaction set*; genes with the same activated reaction set correspond to isozymes.

$$\textbf{max} \quad \sum_j y_j \tag{3.19}$$

$$\textbf{s.t.} \quad z_{\bar{g}} = 1 \tag{3.20}$$

$$z_g = 0 \qquad \forall g \in G \setminus \bar{g} \tag{3.21}$$

$$y_j = f(z_{\hat{g}}, w_{\hat{p}}) \quad \forall GPR(j, \hat{p}, \hat{g}) \in J, P, G \tag{3.22}$$

We then constructed and solved an optimization problem to determine the reactions which can be deactivated by each gene $\bar{g}$ in the model: we turn on all genes but $\bar{g}$, and find those reactions which can be turned off ($y_j = 0$) by deleting $\bar{g}$ alone, subject to GPR constraints. We refer to the set of such reactions as the *deactivated reaction set*; genes with the same deactivated reaction set correspond to subunits (or members of the same protein complex).

$$\textbf{max} \quad \sum_j y_j \tag{3.23}$$

$$\textbf{s.t.} \quad z_{\bar{g}} = 0 \tag{3.24}$$

$$z_g = 1 \qquad \forall g \in G \setminus \bar{g} \tag{3.25}$$

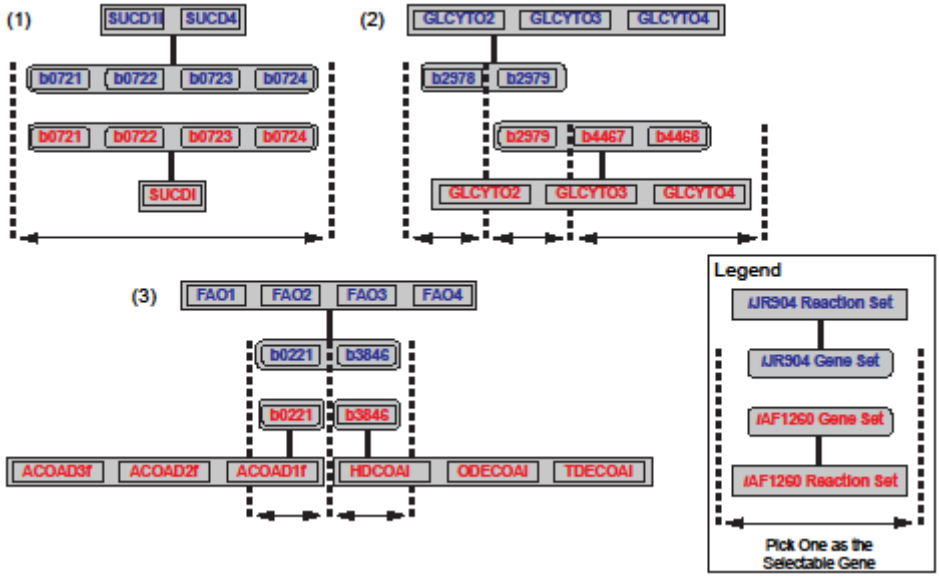
$$y_j = f(z_{\hat{g}}, w_{\hat{p}}) \quad \forall GPR(j, \hat{p}, \hat{g}) \in J, P, G \tag{3.26}$$

We then identified all isozymes and subunits (called *gene sets*) by identifying the sets of genes which all have the same activated or deactivated reaction sets. Some isozymes or subunits within a gene set may be multi-functional (e.g., be associated with multiple reactions); these are included in the gene set only if all isozymes or subunits within the set share the same multi-functionality. We then manually aligned the sets of isozymes and subunits between the models. We find that subunits and isozymes can align in different ways, as illustrated in Figure 3.8.

In scenario 1, the gene sets are identical in both models. (Alternatively, the genes in a gene set are only found in one model.) In this case, we select a single gene to represent the state of the entire gene set. We call this gene the *selectable gene*. In scenario 2, gene sets in the two models partially overlap, with each set containing both

conserved and unique genes. In this case, we select a single gene to represent the state of the conserved genes, and a single gene to represent the unique portion of each gene set, for a total of three selectable genes. And finally in scenario 3, a gene set in one model overlaps with multiple gene sets in the second model. In this case, each individual gene remains selectable. Many additional clarifying examples can be found can be found in Dataset S3 in the original publication.

Figure 3.8. Alignment of isozymes and subunits.



We find that subunits and isozymes can align in different ways, as illustrated below. See the main text for a description of each pattern.

For each gene set corresponding to a group of isozymes, we fix all but the selectable gene from each set to the off state, $z_g = 0$ (collectively the "off set", G_{off}). For a reaction with isozymes, this procedure ensures that all but one common isozyme is fixed to the off state, so that deleting selectable isozymes forces the reactions to the off state. For each gene set corresponding to a group of subunits, we fix all but the selectable gene from each set to the on state, $z_g = 1$ (collectively the "on set", G_{on}). For a protein with subunits, this procedure ensures that all but one common subunit is fixed to the on state, so that deleting selectable subunits forces reactions to the off state. This

procedure also prevents equivalent solutions from being found by CONGA (e.g., multiple gene deletion sets corresponding to the same reaction deletion set). The full list of gene sets in the *i*JR904 and *i*AF1260 models can be found in Dataset S3 in the original publication.

The genes fixed on or off by the above two procedures (referred to as redundant genes in Table 3.8) were removed from the selectable gene set, G_S. The new, smaller selectable gene set represents a subset set of genes which can determine the on-off state of all non-essential, non-blocked reactions in a pair of models (Table 3.8). With these new constraints, the bilevel form of CONGA can be written as:

$$
\begin{array}{lll}
\textbf{max} & \text{difference in flux} & \\
\textbf{s.t.} & \text{constraints (3.1) to (3.4)} & \forall \text{ Species A and B} \\
 & \text{constraint (3.5)} & \forall \text{ Species A and B} \\
 & \sum_{g \in G_S} 1 - z_g \le K & \forall \text{ Species A and B} \\
 & z_g = 1 \qquad \forall g \in G_E \text{ and } G_{on} & \forall \text{ Species A and B} \\
 & z_g = 0 \qquad \forall g \in G_B \text{ and } G_{off} & \forall \text{ Species A and B} \\
 & \text{constraint (3.8)} &
\end{array}
\qquad \text{(CONGA)}
$$

3.3.6: Identification of Orthologs

A gene-based alignment of two networks requires a method for identifying orthologous genes between two genomes. Since the *E. coli* simulations studied two models of the same organism, we were able to immediately match gene loci. For the cyanobacterial simulations, we used the set of bidirectional best-BLAST hits identified during the first step of the *i*Syp611 reconstruction process. Genes added during the manual reconstruction process were checked against the final *i*Cce806 model, and additional orthologs were identified. For the pathogen studies, we used SEED to identify orthologs, as this method identified the smallest number of false positive ortholog pairs (Table 3.7

Table 3.9. Comparison of *i*Syp611 (*Synechococcus*) and *i*Cce806 (*Cyanothece*) cyanobacterial models.

	Draft *i*Syp611 Model	***i*Syp611 Model**	***i*Cce806 Model**
Genes (Orthologs)	542 (497)	611 (529)	806 (529)
Proteins	461	533	690
Reactions	491	552	667
Reactions w/ GPRs	491	517	625
Metabolites	529	542	587

The draft and final reconstructions of the *i*Syp611 model differ considerably in size. The size of the *i*Cce806 model is given as a point of reference.

and 3.2: Discussion). A full summary of ortholog pairs used in each simulation can be found in Dataset S3 in the original publication.

3.3.7: Construction of the *i*Syp611 Metabolic Network

We have formulated a genome-scale network reconstruction of the photosynthetic cyanobacterium *Synechococcus sp.* PCC 7002 consisting of 611 genes, 533 proteins, 552 reactions, and 542 metabolites (Table 3.9). A total of 517 reactions (94%) are associated with genes and proteins, represented by gene-protein-reaction (GPR) associations.

The model was constructed from a draft version of the *i*Cce806 reconstruction of *Cyanothece sp.* ATCC 51142 via a gene-level comparison. The *Synechococcus sp.* PCC 7002 genome sequence was downloaded from the GenBank database at the National Center for Biotechnology Information (NCBI) website [21]. A bidirectional best-BLAST search was used to identify potential orthologs between the two genomes. The validity of the associations was manually assessed using annotation information available from NCBI, KEGG [100], and SEED [169]. For those genes deemed highly probable orthologs, protein and reaction associations were copied from the *i*Cce806 model to create a draft reconstruction using SimPheny (Genomatica Inc., San Diego, CA).

After assembling the draft network, missing functions were added to ensure production of all biomass components. Candidate reactions were selected based on pathways in other cyanobacterial strains. Potential genes for these reactions were

located by best-hit BLAST analysis against other cyanobacterial genomes as well as annotation information obtained from NCBI, KEGG, and SEED. In cases where genomic information was unavailable, reactions were selected based on their frequency of occurrence in related strains. This draft model contained 542 genes, of which 497 were orthologous to genes in *Cyanothece* (Table 3.9).

We also applied CONGA to our draft reconstruction, and use the results to add new subunits and isozymes to existing reactions. In all, nearly 70 genes were added to the reconstruction. A complete list of reactions and GPR associations in the *i*Syp611 model is included in Datasets S1 and S2 in the original publication.

Wherever possible, the reactions used to represent RNA, DNA, protein, fatty acid, and lipid synthesis were updated to reflect the particulars of *Synechococcus sp.* PCC 7002. DNA and RNA composition was based on genomic GC content, and protein composition was obtained from amino acid counts of the proteome. Fatty acid composition was taken from *Synechococcus sp.* PCC 7002 [148,196], and lipid composition was taken from *Synechococcus sp.* PCC 7942 [20]. The biomass equation was formulated using weight fractions of macromolecules (DNA, RNA, protein, lipid, fatty acids, glycogen) measured from *Synechosystis sp.* PCC 6803 in batch culture [210], and the composition of the soluble pool was copied from the *i*JR904 model [189].

The final metabolic reconstruction was used to formulate a constraint-based model of *Synechococcus* metabolism. Experimentally, *Synechococcus* is able to grow phototrophically using light, carbon dioxide, and ammonium. Our constraint-based model was capable of predicting growth under photoautotrophic and glycerol heterotrophic conditions via FBA.

3.3.8: Models and Simulation Conditions

The CONGA analysis of the two published models of *E. coli*, *i*JR904 [189] and *i*AF1260 [54]f were performed under anaerobic, glucose-limited conditions (uptake rate 18.5 mmol/gDW/hr). All reported chemical production levels were normalized to the theoretical maximum (2 mol ethanol/mol glucose, 2 mol lactate/mol glucose, and 1.71 mol succinate/mol glucose).

For the *i*Syp611 and *i*Cce806 comparisons, several reactions in the *i*Syp611 model were replaced with their *i*Cce806 equivalents, including biomass, ATP synthase, DNA, RNA, lipid, and protein synthesis, and cytochrome oxidases unique to *Cyanothece* (Dataset S2 in the original publication). Simulations were performed under carbon-limited photoautotrophic conditions, with maximum uptake fluxes for photons for both photosystems, carbon dioxide, and ammonium constrained to 100 mmol photons/gDW/hr, 20 mmol/gDW/hr, and 10 mmol/gDW/hr, respectively. Unconstrained uptake of inorganic phosphate, oxygen, magnesium(II), protons, sulfate, and water was also allowed. Growth-associated and non-growth associated ATP maintenance requirements were set to zero.

For the *i*NJ661 and *i*SB619 models, simulations were performed on nine distinct minimal media with different carbon and nitrogen sources (Dataset S2 in the original publication). The set of gene deletion sets common to all conditions was then analyzed for potential drug targets.

All simulations were performed using CPLEX 12 (IBM, Armonk, NY) accessed via the General Algebraic Modeling System (GAMS, GAMS Development Corporation, Washington, DC). Simulations were performed on a Red Hat Enterprise Linux server with 2.66 GHz Intel Xeon processors and 8 GB of RAM. CONGA can identify a lethal gene deletion set containing just one gene in less than a second. Lethal gene deletion sets containing two or three genes took on average 3 minutes to identify. Identifying

model-dominant chemical production strategies is more time-consuming, with a model-dominant strategy containing two genes requiring on average 2 minutes. Model-dominant strategies containing three, four, or five genes took an average of 15 minutes, 75 minutes, and 5 hours to identify, respectively. A full summary of differences identified by each simulation can be found in the Dataset S3 in the original publication.

Chapter 4: Thermodynamic Constraints and Genome-Scale Metabolic Models

This material was originally published in:

> Hamilton JJ, Dwivedi V, Reed JL (2013) Quantitative assessment of thermodynamic constraints on the solution space of genome-scale metabolic models. Biophys J 105(2): 512–522.

In this work, we examine the extent to which thermodynamics-based flux-balance methods can make genome-scale, quantitative predictions, in the absence of outside information on flux directions, considering both the presence and absence of uncertainty in thermodynamic estimates. To this end, we applied TMFA to the *i*JR904 model of *Escherichia coli* [189]. This model was used since group contribution estimates are available for a higher fraction of the metabolites in the *i*JR904 model than in newer models [96]. We assessed the predictive performance of TMFA against a number of large-scale datasets [11,19,94,99], encompassing metabolite concentrations and gene deletion phenotypes, under both aerobic and anaerobic, and optimal and suboptimal growth conditions. Through this analysis, we highlight the importance of quantitative concentration measurements and thermodynamic coupling in achieving physiologically realistic predictions of growth rates and flux distributions. We were also able to generate hypotheses regarding metabolite concentrations and thermodynamic bottlenecks, and we discuss additional types of data and constraints which can improve predictions of metabolite concentrations.

4.1: Materials and Methods

4.1.1: Overview and Relationship to Previous Thermodynamic Models

Given a stoichiometric matrix S and set of reactions J, flux-balance analysis (FBA) seeks a steady-state flux distribution (v) maximizing the flux through the biomass reaction (v_{BM}), while also satisfying mass-balance and enzyme capacity constraints for individual reactions, j:

$$\textbf{max} \quad v_{BM} \tag{4.1}$$
$$\textbf{s.t.} \quad S \cdot v = 0 \tag{4.2}$$
$$0 \le v_j \le v_{max} \quad \forall j \in J_{irrev} \tag{4.3}$$
$$v_{min} \le v_j \le v_{max} \ \forall j \notin J_{irrev} \tag{4.4}$$

Because the network is at steady-state, net production of all metabolites is zero. Each reaction is further constrained to have flux within an appropriate range as given by enzyme capacities, with some reactions assumed to be irreversible ($j \in J_{irrev}$). The limits through most reactions are set to $v_{max} = 1000$ mmol/gDW/hr and $v_{min} = -1000$ mmol/gDW/hr, except for measured fluxes (e.g., carbon uptake rates).

Previous implementations of TMFA [70,88] have imposed thermodynamic constraints on top of FBA, thereby allowing these constrains to further restrict reaction directions; however, a reaction deemed irreversible in FBA cannot become reversible even if indicated by the thermodynamic constraints. In contrast, our implementation makes no assumptions about reaction (ir)reversibility, allowing thermodynamic constraints to decide the directions of all reactions. We replaced constraints (4.3) and (4.4) with:

$$v_{min} \le v_j \le v_{max} \ \forall j \in J \tag{4.5}$$

and rely on thermodynamic constraints alone to determine reaction directions.

4.1.2: Calculating Free Energies of Reaction

Enforcing thermodynamic constraints requires knowledge of the standard transformed Gibbs energy of reaction ($\Delta_r G'^0$) for the reactions in the model. Due to a paucity of experimental data, group contribution methods [135,136] are used to provide estimates and uncertainties of the standard transformed Gibbs energy of formation ($\Delta_f G'^0$) for metabolites and of reaction ($\Delta_r G'^0$) for reactions.

Group contribution methods (GCMs) assume that the $\Delta_f G'^0$ of a metabolite i is a linear combination of the formation energies of its constituent molecular substructures (or groups), k:

$$\Delta_f G_i'^0 = \sum_k n_{i,k} \Delta_{gr} G_k'^0 \tag{4.6}$$

where $\Delta_{gr} G_k'^0$ is the estimated contribution of group k to the overall $\Delta_f G'^0$, and $n_{i,k}$ is the number of groups k in the molecular structure of compound i. We used a software implementation of the GCM of Jankowski, et al. [61,87,96], to obtain estimates and uncertainties of $\Delta_f G'^0$ and $\Delta_r G'^0$ for metabolites and reactions in the *i*JR904 network (see Tables S1 and S2 in the Supporting Material of the original publication for values).

The GCM method returns estimates of $\Delta_f G'^0$ for the predominant ionic species (pseudoisomer) at biochemical standard state: pH 7, zero ionic strength, and temperature 298K. The major pseudoisomeric form of each molecule in the *i*JR904 model was determined using pKa estimation software (Marvin pKa plug-in, version 5.11.4, ChemAxon, Budapest, Hungary).

We found that the metabolite charges predicted by the GCM differed in some instances from the charges used in the *i*JR904 model stoichiometric matrix. Because mass and charge balancing plays an important role in calculating $\Delta_r G'^0$, we adjusted the $\Delta_f G'^0$ estimates for these metabolites using the pKa of the relevant

protonation/deprotonation reaction [2]. pKa values were determined using pKa estimation software. We also adjusted the $\Delta_f G'^0$ estimates for all extracellular metabolites to a pH of 7.4, in accordance with the pH used in our simulations [2].

We define the change in Gibbs energy due to pKa adjustments as $\Delta_{pKa} G'^0$, and compute it as follows, following a standard derivation [2,83]. Consider the deprotonation reaction, in which an acid HA dissociates by forming its conjugate base A^- and a proton, H^+.

$$HA \leftrightarrow H^+ + A^- \tag{4.7}$$

Given $\Delta_f G'^0_{HA}$ from the GCM, $\Delta_f G'^0_{A^-}$ is given by $\Delta_f G'^0_{A^-} = \Delta_f G'^0_{HA} + RT \log_{10} pK_a$. Therefore,

$$\Delta_{pKa} G'^0_{A^-} = \Delta_f G'^0_{HA} - \Delta_f G'^0_{A^-} = -RT \log_{10} pK_a \tag{4.8}$$

We further define the change in Gibbs energy due to pH adjustments as $\Delta_{pH} G'^0$, and compute it as follows, again following a standard derivation [2]:

$$\Delta_{pH} G'^0_i = N_i RT \ln(10) \Delta pH \tag{4.9}$$

where N_i is the number of hydrogen atoms in the molecule, and ΔpH is the pH difference across the membrane (i.e., 7.4 – 7.0).

These computations resulted in the standard transformed Gibbs energy of formation, $\Delta_f G'^0$, computed as follows:

$$\Delta_f G'^0_i = \Delta_f G'^0_{i,GCM} + \Delta_{pH} G'^0_i + \Delta_{pKa} G'^0_i \tag{4.10}$$

where, $\Delta_{pH} G'^0_i$ and $\Delta_{pKa} G'^0_i$ are the changes in Gibbs energy due to pH and pKa adjustments, respectively, and $\Delta_f G'^0_{i,GCM}$ denotes the estimate computed by the GCM as given in (4.6). Simulations were performed under conditions of zero ionic strength and temperature of 298K, to avoid the need for additional adjustments to $\Delta_f G'^0$.

Using our new estimates of $\Delta_f G'^0$, we calculated $\Delta_r G'^0$ from the stoichiometry of each reaction:

$$\Delta_r G_j'^0 = \sum_i S_{i,j} \Delta_f G_i'^0 \tag{4.11}$$

where $S_{i,j}$ is the stoichiometric coefficient of metabolite i in reaction j and $\Delta_f G'^0$ is computed as in (4.10). This method of calculating $\Delta_r G'^0$ is equivalent to

$$\Delta_r G_j'^0 = \sum_i S_{i,j} \left(\sum_k n_{i,k} \Delta_{gr} G_k'^0 + \Delta_{pH} G_i'^0 + \Delta_{pKa} G_i'^0 \right) \tag{4.12}$$

and is consistent with the previous implementation of TMFA [88]. In this approach, any structural groups unchanged during the reaction cancel out, thereby reducing the uncertainty in $\Delta_r G'^0$. We note, however, that this approach treats the GCM assumption of linearly additive free energy contributions as fact, rather than assumption.

We then calculated the transformed Gibbs energy of reaction ($\Delta_r G'$) as a function of metabolite concentration, x_i:

$$\Delta_r G_j' = \Delta_r G_j'^0 + RT \sum_i S_{i,j} \ln(x_i) + \Delta_t G_j'^0 \tag{4.13}$$

where R is the gas constant, T is the temperature (298K), and $\Delta_t G'^0$ reflects the contribution to $\Delta_r G'$ from the transport of metabolites across the membrane. Following an established derivation [88], we calculated $\Delta_t G'^0$ as a function of the electrochemical potential ($\Delta\Psi$) and pH gradient (ΔpH) across the membrane:

$$\Delta_t G_j'^0 = c_j F \Delta\Psi - 2.3 h_j RT \Delta pH \tag{4.14}$$

where F is Faraday's constant, c_j is the net charge transported from outside to inside the cell, and h_j is the number of protons transported across the membrane (see Table S3 in the Supporting Material of the original publication for values). Our constraints on intra- and extracellular proton concentrations resulted in values of -130 mV for $\Delta\Psi$, and 0.4 for ΔpH [88,151].

The GCM software provides estimates ($\Delta_{gr}G'^{0}_{k,est}$) and uncertainties ($SE_{\Delta_{gr}G'^{0}_{k,est}}$) of each constituent group, which can be used to estimate uncertainties for $\Delta_f G'^{0}$ ($SE_{\Delta_f G'^{0}_{est}}$) and $\Delta_r G'^{0}$ ($SE_{\Delta_r G'^{0}_{est}}$) as follows:

$$SE_{\Delta_f G'^{0}_{est}} = \sqrt{\left(\sum_k n_{i,k} SE_{\Delta_{gr} G'^{0}_{k,est}}\right)^2} \tag{4.15}$$

$$SE_{\Delta_r G'^{0}_{est}} = \sqrt{\left(\sum_i S_{i,j}\left(\sum_k n_{i,k} SE_{\Delta_{gr} G'^{0}_{k,est}}\right)\right)^2} \tag{4.16}$$

A full list of adjusted $\Delta_f G'^{0}$ and $\Delta_r G'^{0}$ values and uncertainties can be found in Tables S1 and S2 in the Supporting Material of the original publication. In TMFA, we fixed values of $\Delta_{gr}G'^{0}$ to their estimated values ($\Delta_{gr}G'^{0}_{est}$) as given by the group contribution method,

$$\Delta_{gr}G'^{0}_{k} = \Delta_{gr}G'^{0}_{k,est} \tag{4.17}$$

while in relaxed TMFA (RTMFA), we allowed $\Delta_{gr}G'^{0}_{k}$ of each group, metabolite, and reaction to vary within its 95% confidence interval, as determined by the standard error (SE) reported by the GCM software (i.e., $SE_{\Delta_{gr}G'^{0}_{k,est}}$):

$$\Delta_{gr}G'^{0}_{k,est} - 2SE_{\Delta_{gr}G'^{0}_{k,est}} \le \Delta_{gr}G'^{0}_{k} \le \Delta_{gr}G'^{0}_{k,est} + 2SE_{\Delta_{gr}G'^{0}_{k,est}} \tag{4.18}$$

Such a constraint indirectly also ensures that $\Delta_f G'^{0}_{i}$ and $\Delta_r G'^{0}_{j}$ of each metabolite and reaction also varies within its 95% confidence interval.

4.1.3: Calculating Free Energies of Reaction: Special Cases

The *i*JR904 model contains 13 metabolites for which the molecular formula is based on the average experimentally measured fatty acid composition of membrane phospholipids. The formula utilized in the stoichiometric matrix represents 50 copies of the compound,

while the formula utilized to calculate $\Delta_f G'^0$ represents a single copy. This necessitated a scaling of $\Delta_f G'^0$ in (4.10) by a factor of 50. A list of these metabolites can be found in Table S1 in the Supporting Material of the original publication.

Due to limitations in the GCM, there were some metabolites for which $\Delta_f G'^0$ was unknown, because one or more of their constituent groups had unknown $\Delta_{gr} G'^0_{est}$. In these cases, we allowed $\Delta_{gr} G'^0$ for these groups to vary freely, and linearly combined reactions involving these metabolites into "lumped" representations so that $\Delta_r G'$ could be calculated. (We call the set of lumped reactions J_L, and introduce it into the S matrix as a subset of the set J.) We then developed constraints that ensured both the lumped reactions and their constituent reactions remained thermodynamically consistent, irrespective of the value of $\Delta_f G'^0$ of the unknown metabolites.

For lumped reactions, we first calculated $\Delta_r G'$ as described in (4.12) and (4.13), while simultaneously calculating it from $\Delta_r G'$ of their constituent reactions (for which the $\Delta_{gr} G'^0$ of unknown groups varies freely). To do this, we defined parameters $\alpha_{j,l} = 1$ if reaction j (one of the reactions with unknown $\Delta_r G'$) combined in the forward direction to make up lumped reaction l, and defined $\alpha_{j,l} = -1$ if reaction j combined in the reverse direction. We then computed $\Delta_r G'$ of the lumped reaction:

$$\Delta_r G'_l = \sum_j \alpha_{j,l} \Delta_r G'_j \qquad \forall l \in J_L \tag{4.19}$$

This constraint ensures that the thermodynamics of the lumped reaction and its constituents are internally consistent, irrespective of the value of $\Delta_{gr} G'^0$ of the unknown groups. The complete list of lumped reactions J_L, and parameters α can be found in Table S5 in the Supporting Material of the original publication.

We also identified 39 reactions for which this lumping was not possible, for which different approaches to calculating $\Delta_r G'^0$ were employed. We first identified 18 reactions

which were blocked on the basis of stoichiometry. We allowed $\Delta_{gr}G'^{0}$ of the unknown groups associated with these 18 reactions to vary freely, knowing that the value of $\Delta_{gr}G'^{0}$ would have no effect on thermodynamics elsewhere in the network.

An additional 11 reactions were transport reactions, involving metabolites which could only be transported back and forth across the membrane. For these reactions, we required the $\Delta_f G'^{0}$ of the intracellular and extracellular metabolites to be equal, thereby allowing $\Delta_r G'^{0}$ to be calculated (the net contribution to (4.11) is zero). Because the metabolites under consideration were not involved in other reactions, the value of $\Delta_f G'^{0}$ could have no effect on thermodynamics elsewhere in the network.

We found another set of 6 reactions in which the reactions could only carry flux if involved in a closed cycle (e.g. A→B→A) with another member of the set. Because enforcement of thermodynamics would render the reactions inactive, they cannot carry flux in any feasible solution. Therefore, these reactions were removed from the *i*JR904 model.

Finally, there were 2 reactions involving lipopolysaccharide (LPS), another compound for which no $\Delta_f G'^{0}$ was available. Both reactions are coupled to the biomass reaction, and must operate in the forward direction for growth to occur. We allowed $\Delta_f G'^{0}$ of LPS to vary freely, knowing that the coupling requirement would ensure proper enforcement of thermodynamics.

The full list of reactions and metabolites requiring each treatment can be found in Table S4 in the Supporting Material of the original publication.

4.1.4: Enforcing Thermodynamic Consistency

Thermodynamic consistency requires that reaction fluxes are consistent with predicted values of $\Delta_r G'$ (i.e., $v \cdot \Delta_r G' < 0$). We employed a mixed-integer approach to enforce this

requirement, in which a binary variable δ indicates if a reaction is operating in the forward ($\delta = 1$) or reverse ($\delta = 0$) direction. We then added the following equations to our model:

$$(1-\delta)v_{min} \le v \le \delta v_{max} \quad (4.20)$$

$$-M\delta + \varepsilon \le \Delta_r G' \le M(1-\delta) - \varepsilon \quad (4.21)$$

where (4.21) is a big-M constraint [248] in which M is an upper limit on $\Delta_r G'$ (300kcal/mol) and ε is a small non-zero number (10^{-6}).

4.1.5: Final Formulation

In its final formulation, TMFA combines FBA, thermodynamic constraints, and constraints on metabolite concentrations.

max v_{BM}
s.t. FBA constraints, (4.2) and (4.5)
$\Delta_f G'$ constraints (4.6), (4.8) to (4.10) $i \in I$
$\Delta_r G'$ constraints (4.11), (4.13), (4.14), and (4.19) $j \in J$ (TMFA)
consistency constraints, (4.20) and (4.21) $j \in J$
$\Delta_{gr} G'^{0}_{k,est}$ definition, (4.17) or (4.18) $k \in K$
concentration constraints (as needed)

Note that because the set of lumped reactions J_L is a subset of set J, (4.2) was modified to include only those $j \in J \setminus J_L$, as the reactions in J_L are artificial and not a part of the organism's metabolic network.

4.1.6: Flux and Thermodynamic Variability Analysis

We performed flux and thermodynamic variability analysis under a variety of conditions to identify thermodynamically feasible flux and metabolite concentration ranges. In flux variability analysis (FVA) (16), the flux v through each reaction is minimized and maximized subject to the constraints of TMFA. That is, for each reaction j, we solve the following optimization problem:

max (min) v_j
s.t. FBA constraints, (4.2) and (4.5) (FVA)

$\Delta_f G^{'}$ constraints (4.6), (4.8) to (4.10) $i \in I$

$\Delta_r G^{'}$ constraints (4.11), (4.13), (4.14), and (4.19) $j \in J$

consistency constraints, (4.20) and (4.21) $j \in J$

$\Delta_{gr} G^{'0}_{k,est}$ definition, (4.17) or (4.18) $k \in K$

concentration constraints (as needed)

In thermodynamic variability analysis (TVA) [88], the concentration x of each metabolite (or $\Delta_r G^{'}$ of each reaction) is minimized and maximized subject to the constraints of TMFA. For metabolites, we solve the following optimization problem

max (min) x_i

s.t. FBA constraints, (4.2) and (4.5)

$\Delta_f G^{'}$ constraints (4.6), (4.8) to (4.10) $i \in I$

$\Delta_r G^{'}$ constraints (4.11), (4.13), (4.14), and (4.19) $j \in J$

consistency constraints, (4.20) and (4.21) $j \in J$

consistency constraints, (4.20) and (4.21) $j \in J$

$\Delta_{gr} G^{'0}_{k,est}$ definition, (4.17) or (4.18) $k \in K$

concentration constraints (as needed)

(TVA)

and for reactions, we solve the following optimization problem:

max (min) $\Delta_r G^{'}_j$

s.t. FBA constraints, (4.2) and (4.5)

$\Delta_f G^{'}$ constraints (4.6), (4.8) to (4.10) $i \in I$

$\Delta_r G^{'}$ constraints (4.11), (4.13), (4.14), and (4.19) $j \in J$

consistency constraints, (4.20) and (4.21) $j \in J$

consistency constraints, (4.20) and (4.21) $j \in J$

$\Delta_{gr} G^{'0}_{k,est}$ definition, (4.17) or (4.18) $k \in K$

concentration constraints (as needed)

(TVA)

Using our FVA results, we defined sets of reactions which can operate only in the forward (J_{for}) or reverse (J_{rev}) directions, or which are blocked (J_{bl}) entirely. By imposing the following constraints on our model, we were able to reduce simulation times by over an order of magnitude.

$$\delta_j = 1 \qquad \forall j \in J_{for} \tag{4.22}$$

$$v_j \geq 0 \qquad \forall j \in J_{for} \tag{4.23}$$

$$\delta_j = 0 \qquad \forall j \in J_{rev} \tag{4.24}$$

$$v_j \le 0 \qquad \forall j \in J_{rev} \tag{4.25}$$

$$v_j = 0 \qquad \forall j \in J_{bl} \tag{4.26}$$

For some reactions, flux maximization using RTMFA was computationally intensive, so all FVA simulations were performed with a time limit of 5 minutes. As a result, there is a possibility that true flux ranges are larger than reported.

4.1.7: Differences in Phenotype: CONGA

We then used an algorithm we previously developed, CONGA [78], to identify single, double, and triple gene deletions predicted to be lethal in FBA but not TMFA (or RTMFA). CONGA employs a bilevel optimization problem to identify gene deletion strategies maximizing the difference in biomass flux between two different models, FBA and TMFA (or RTMFA). The outer problem is a mixed-integer linear program which finds gene deletions maximizing the difference in biomass flux between two reactions in two different models. The inner problems are flux-balance analysis (FBA) problems which ensure the flux difference is maximized while both models are simultaneously maximizing biomass. Gene-protein-reaction (GPR) constraints [106] associate genes to reactions and enforce the reaction deletions in the inner problems which are associated with the gene deletions in the outer problem. CONGA can select any genes for deletion, with the restriction that orthologous genes present in both models be deleted simultaneously from both models. Because the formulations were of the same model, genes were easily aligned by matching gene loci.

CONGA relies on duality theory to reformulate the bilevel problem as a single-level problem, so we were unable to utilize TMFA directly in the inner problem. Instead, we used a linear programming relaxation of TMFA (TMFA-LP), in which we combined FBAr with reaction directions assigned to be consistent with TMFA predictions, as defined by the sets J_{for}, J_{rev}, and J_{bl} described in Methods. Finally, because TMFA-LP solutions are a superspace of TMFA solutions, it was necessary to

ensure that the gene-deletion strategies identified by CONGA produced the same phenotype using both TMFA and TMFA-LP. Of all gene-deletion strategies identified by CONGA, the vast majority produced the same phenotype using both formulations.

The formulation is given below, and additional details on implementation and reformulation as a single-level problem can be found in [78].

$$
\begin{aligned}
&\textbf{max} && v_{BM,TMFA\text{-}LP} - v_{BM,FBA} \\
&\textbf{s.t.} && \textbf{max} \quad v_{BM,TMFA\text{-}LP} \\
&&& \textbf{s.t.} \quad \text{FBA constraints, (4.2) and (4.5)} \\
&&&\qquad \text{reaction direction constraints, (4.23), (4.25), and (4.26)} \\
&&&\qquad \text{reaction deletions from the outer problem} \\
&&& \text{GPR constraints} \\
&&& \textbf{max} \quad v_{BM,FBA} \\
&&& \textbf{s.t.} \quad \text{FBA constraints, (4.2) to (4.4)} \\
&&&\qquad \text{reaction deletions from the outer problem} \\
&&& \text{GPR constraints} \\
&&& \text{limited number of gene deletions} \\
&&& \text{orthologs deleted from both models}
\end{aligned}
\qquad \text{(CONGA)}
$$

4.1.8: Synthetic Lethals and Phenotype Correction

In order for FBA and TMFA to make different predictions for the same gene deletion, TMFA must enable a reaction to proceed in a direction not allowed by FBA. By comparing thermodynamically feasible reaction directions from FVA analysis to those assigned by FBA, we can identify the set of such reactions, called J_{SL}.

Once we identified gene deletions lethal only in the FBA model, we employed the Synthetic Lethals (SL) Finder [226] to identify the reaction(s) responsible for rescuing the phenotype in TMFA (the SL reaction). We then performed the gene deletion in TMFA to determine the direction of the SL reaction by examining the predicted flux distribution.

SL Finder employs a bilevel optimization problem to identify reactions whose deletion causes a lethal phenotype. The inner problem is an FBA problem which ensures that biomass is maximized, while the outer problem is a mixed-integer linear program

which seeks reaction deletions which lower the maximum biomass. As with CONGA, SL Finder reformulates the bilevel problem to a single-level problem, and once again we employed TMFA-LP in the inner problem. The formulation is given below and additional details on implementation and reformulation as a single-level problem can be found in [226].

$$
\begin{array}{lll}
\textbf{min} & v_{BM} & \\
\textbf{s.t.} & \textbf{\textit{max}} & v_{BM} \\
 & \textbf{\textit{s.t.}} & \text{FBA constraints, (4.2) and (4.5)} \\
 & & \text{reaction direction constraints, (4.23), (4.25), and (4.26)} \\
 & & \text{reaction deletions from the outer problem} \\
 & \text{GPR constraints} & \\
 & \text{delete gene of interest} & \\
 & \text{select reaction from set } J_{SL} \text{ for deletion} &
\end{array}
\qquad \text{(SL)}
$$

For some gene deletions predicted to be lethal in FBA, the TMFA prediction disagreed with experimental observation. This suggested that the SL reaction did not operate *in vivo* in the direction predicted by FBA. In these instances, we developed a constraint on metabolite concentrations that prevented the SL reaction from operating in the TMFA-predicted direction, thereby correcting the phenotype. We refer to such constraints as phenotype-correction constraints. For example, if TMFA predicted that a reaction operated in the reverse direction, we sought a constraint that forced the reaction to operate in the forward direction. For such a reaction, $\Delta_r G' < 0$. That is, for a non-transport reaction,

$$\Delta_r G_j'^0 + RT \sum_i S_{i,j} \ln(x_i) < 0 \tag{4.27}$$

Then, as long as

$$\min RT \sum_i S_{i,j} \ln(x_i) < -\max \Delta_r G_j'^0 \tag{4.28}$$

adding a new constraint of the form

$$RT \sum_i S_{i,j} \ln(x_i) < -\max \Delta_r G_j'^0 \tag{4.29}$$

to TMFA ensures the synthetic lethal reaction will operate in the forward direction. The value of $\max \Delta_r G_j'^0$ is found by solving TVA (see above), subject to the additional constraint given by (4.27).

4.1.9: Simulation Conditions

All simulations were performed using CPLEX 12 (IBM, Armonk, NY) accessed via the General Algebraic Modeling System, Version 23.3.3 (GAMS, GAMS Development Corporation, Washington, DC). Simulations were performed on a Red Hat Enterprise Linux server with 2.66 GHz Intel Xeon processors and 8 GB of RAM. The TMFA formulation solves in a few seconds, while the RTMFA formulation takes approximately an hour to prove global optimality.

4.1.10: Sources of Experimental Data

This study used experimental measurements of concentration data from two distinct sources [19,94]. The first study, from Ishii, et al. [94], examined *E. coli* grown in continuous culture at a specific growth rate of 0.2 hr^{-1}, while the second, from Bennett, et al. [19], examined exponential growth of *E. coli* in batch culture. We use these data in simulations of suboptimal and optimal growth, respectively. The dataset of Bennett, et al., reported mean values and 95% confidence intervals (CIs) for 107 metabolites in the *i*JR904 model, while the dataset of Ishii, et al., reported between two and five distinct measurements for 88 metabolites in the *i*JR904 model, from which we computed the mean values and 95% CIs. In some instances, calculated CIs were larger than the mean, in which case the metabolite was excluded from our analysis. Values of lower and upper CIs for all metabolites in both datasets are presented in Tables S14 and S15 in the Supporting Material of the original publication.

We also relied on two large-scale gene deletion studies for single-gene knockout phenotypes. The first study, from Baba, et al. [11], examined *E. coli* grown on glucose

minimal media under aerobic conditions, while the second, from Joyce, et al. [99], examined *E. coli* grown on glycerol minimal media under aerobic conditions. The dataset of Joyce, et al. described a computational cutoff to identify lethal deletions based on OD_{600} measurements (the lowest $1/9^{th}$, based on OD values), which we applied to the dataset of Baba, et al. (at 24 hrs) to identity a list of lethal deletions. We then excluded from this list any genes found to be rescued by the overexpression of at least one noncognate *E. coli* gene [172], as those single-gene deletions may still exhibit low levels of growth.

4.1.11: Experimental Methods

The Keio collection of in-frame single-gene deletion strains [11,253] and wild-type *E. coli K-12 BW25113* were used in all experiments. Double mutants were generated using P1 transduction [239].

Strains were screened in triplicate for aerobic growth at 37°C in a Tecan Infinite® 200 PRO microplate reader (Tecan Group Ltd., Zurich, Switzerland). Optical density measurements at 600 nm (OD_{600}) were taken by the microplate reader every 15 minutes for 20 hours. Strains were pre-cultured overnight in LB medium. Pre-cultured cells were washed three times and resuspended in M9 minimal media so that the starting OD_{600} was around 0.05, as measured in a spectrophotometer with a 1 cm path length. M9 minimal medium was supplemented with 2 g/L glucose and 50 mg/mL amino acid (aspartate, arginine, or tyrosine), as appropriate. Growth curves represent the average OD_{600} at each time point across three replicates.

Strains were screened for anaerobic growth by streaking onto M9 agar plates with 0.2% glucose. Plates were placed into AnaeroPack® rectangular jars and incubated at 37°C in the presence of AnaeroPack®-Anaero anaerobic gas generators and Resazurin anaerobic indicator strips (all Thermo Fisher Scientific, Waltham, MA). Wild-type *E.*

coli K-12 BW25113 colonies appeared after 48 hours, and plates of mutant strains were removed and analyzed after 96 hours.

4.2: Results

4.2.1: Optimization of Aerobic Growth on Glucose using TMFA

We first used thermodynamics-based metabolic flux analysis (TMFA) to determine the maximum growth rate of *E. coli* under glucose aerobic conditions, using the *i*JR904 genome-scale metabolic model [189]. Building on previous work [88], we constrained the concentration of intracellular metabolites to a range of 0.001 to 20 mM. We constrained the concentrations of extracellular nutrients to that of 0.4% glucose MOPS medium [150]. The concentration of intracellular H^+ was fixed to a pH of 7, and extracellular H^+ to a pH of 7.4. The concentration of extracellular gases (O_2 and CO_2) were based on experimental measurement [254], with a requirement that the intracellular concentration be less than the extracellular concentration. All simulations were performed with glucose as the limiting substrate, with a maximum uptake rate of 10 mmol/gDW/hr. The full set of metabolite concentration and exchange flux constraints can be found in Table S6 in the Supporting Material of the original publication.

Under these conditions, TMFA predicted a maximum growth rate of zero, suggesting that growth was not possible due to the thermodynamic infeasibility of one or more essential reactions. We expanded the concentration range of seven metabolites to enable thermodynamic feasibility of eight essential reactions. Among these were four reactions involved in histidine biosynthesis, including histidinol-phosphatase,

$$\text{water} + \text{L-histidinol phosphate} \leftrightarrow \text{L-histidinol} + \text{phosphate} \tag{4.30}$$

which must overcome a $\Delta_r G'^0$ of 14.96 kcal/mol to proceed in the forward direction. Histidinol itself is predicted by the GCM to have a $\Delta_f G'^0$ of -2.21 kcal/mol and a charge of +1. However, the *i*JR904 S-matrix utilizes a charge of +2, requiring a $\Delta_{pKa} G'^0$

offset of 9.71 kcal/mol. The switch from positive to negative $\Delta_f G'^0$ contributes to the thermodynamic infeasibility of the histidine biosynthesis pathway. We note that a similar problem of thermodynamic infeasibility was observed in the original implementation of TMFA, albeit with different reactions being infeasible. This suggests that the predictions made by thermodynamic flux-balance models will be sensitive to both the GCM method and any subsequent adjustments made to GCM predictions. Table S6 in the Supporting Material of the original publication lists the metabolites requiring expanded bounds, as well as the concentration ranges necessary for thermodynamic feasibility.

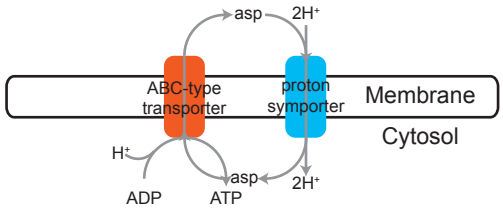
We first performed simulations neglecting the uncertainty in the standard Gibbs energy of formation of each group ($\Delta_{gr} G'^0$) (4.17). The maximum growth rate predicted by TMFA exceeded that of FBA by only 2%, despite the increased network flexibility made possible by the lack of predefined reaction directions in TMFA.

We then introduced uncertainty in $\Delta_{gr} G'^0$ (4.18) and observed that the maximum growth rate exceeded that of FBA by approximately 12%. This elevated growth rate highlights the additional network flexibility made possible by the uncertainty in free energy estimates arising from the GCM. In particular, the growth rate difference was due to mechanisms in the relaxed TMFA model (RTMFA) which enable ATP to be synthesized at lower energetic cost than occurs physiologically (and is reflected in FBA). For example, RTMFA should predict ATP synthesis via ATP synthase, using energy released from the transport of four protons across the plasma membrane. Instead, the RTMFA model identifies numerous cycles which synthesize ATP using energy released from the transport of fewer numbers of protons, by shuttling metabolites back and forth across the plasma membrane (Figure 4.1A). One such metabolite is aspartate (asp), which enables ATP to be synthesized using energy

released from the transport of only two protons across the plasma membrane. However, this shuttling relies on the aspartate ABC-type transporter pumping out aspartate, which requires ADP to be present at a concentration higher than ATP (Figure 4.1B). By additionally constraining the concentration of ATP to its experimentally measured range [19], in which ATP is present at a higher concentration than ADP, we can force the RTMFA model to use ATP synthase to synthesize ATP. When a constraint on ATP concentration is added to the model, the predicted growth rate exceeds FBA by only 3%. This constraint on ATP concentration was used in all subsequent RTMFA simulations discussed in this work.

Figure 4.1. Examples of thermodynamically feasible but physiologically implausible behavior.

A)



B)

RTFMA	$[x]_{min}$ (mM)	$[x]_{max}$ (mM)	$[x]_{meas}$ (mM)
ADP	0.89	20	0.437 to 0.704
ATP	0.01	0.22	**8.13 to 11.4**

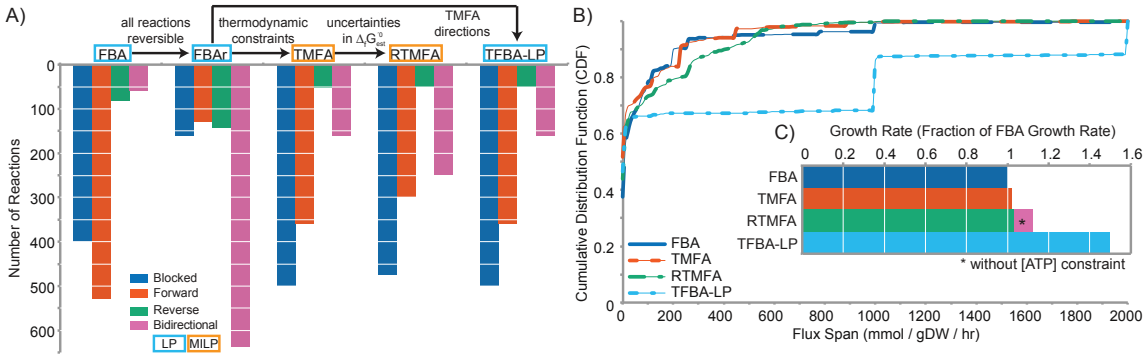
(A) RTMFA predicts ATP can be synthesized by cycling small molecules across the membrane. One such cycle involves asp. **(B)** Metabolite concentration ranges predicted by RTMFA for which the cycle shown in **(A)** is thermodynamically feasible (columns 1 and 2), and experimentally measured metabolite concentrations (column 3). Bold type: constraining this concentration in the model renders the cycle thermodynamically infeasible.

4.2.2: Flux Variability Analysis: Thermodynamically Feasible Reaction Directions

We then used flux variability analysis (FVA) to determine thermodynamically feasible directions for all reactions in the network under glucose aerobic conditions (Figure 4.2A, Table 4.1, and Table S8 in the Supporting Material of the original publication). We classified reactions as fully bidirectional, or constrained in one of three ways: blocked entirely, or capable of operating in the forward or reverse direction only. We first

performed FVA on a fully-reversible version of the *i*JR904 model without any thermodynamic constraints (FBAr), in which all reactions (except biomass) were allowed to be reversible. This allowed us to identify directionally-constrained reactions on the basis of stoichiometry and the external environment.

Figure 4.2. Comparison of thermodynamic formulations under glucose aerobic conditions.



(A) Reactions are classified as blocked, forward, reverse, or bidirectional based on their thermodynamic feasibility. Differences between each formulation are given above the chart. **(B)** Cumulative distribution function (CDF) for the flux span for each formulation. **(C)** Plot of growth rate for each formulation, normalized to the FBA growth rate. **(Formulations)** FBAr: FBA with all reaction directions fully reversible. TMFA: FBAr with thermodynamic constraints, global metabolite bounds, and media constraints. RTMFA: cTMFA with uncertainties in $\Delta_f G'^{0}_{est}$ and a constraint on ATP concentration. TMFA-LP: LP FBAr with reaction directions consistent

Table 4.1. Summary of thermodynamically feasible reaction directions in different models under glucose aerobic growth conditions.

	Growth Rate (hr^{-1})	**Forward**	**Reverse**	**Bidirectional**	**Blocked**
FBA	0.92	529	81	58	402
FBAr	1.45	129	143	637	161
TMFA	0.94	359	52	160	499
RTMFA	0.95	298	50	248	474
TMFA-LP	1.38	359	51	162	498

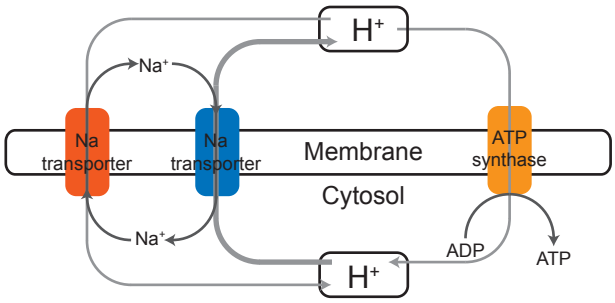
We observed that in FBAr, the majority of reactions are bidirectional (60%), with the remainder being constrained in some way. On the other hand, the FBA model has a large set of irreversible reactions, causing the number of bidirectional reactions to decrease significantly, to a mere 5% of the network. The majority of reactions in the FBA model operate in the forward direction only, or not at all. When we neglect *a*

priori reaction directionality assignments and instead allow reaction directions to be determined solely by thermodynamic constraints (TMFA), the fraction of directionally constrained reactions becomes 85%, compared to 95% in FBA, and 40% in FBAr. This reveals that thermodynamic constraints play a major role in eliminating some of the network flexibility resulting from eliminating predefined reaction directions (FBAr). We also see that, relative to FBA (with predefined reaction directions), TMFA enabled previously forward- or reverse-only reactions to become bidirectional, and previously blocked reactions to become feasible in the reverse direction.

Just as moving from FBA to TMFA led to a decrease in the number of constrained reactions, so too does moving from TMFA to RTMFA. The number of constrained reactions decreases from 85% in TMFA to 77% in RTMFA. The number of bidirectional reactions increases by 50%, as many previously forward-only reactions become feasible in the reverse direction.

When the reaction directionalities from FVA analysis of the TMFA model were used to further constrain fluxes in FBAr (TMFA-LP), the predicted growth rate was 150% of that predicted by FBA (Figure 4.2C). This increase in growth rate indicates ATP- or other energy-generating cycles were present in the network, that are not present when thermodynamic constraints are imposed directly. One such cycle involves the shuttling of sodium ions back and forth across the membrane, resulting in a proton gradient used to synthesize ATP (Figure 4.3). This cycle is infeasible under TMFA, despite each reaction operating in a thermodynamically feasible direction. This emphasizes the need to actively impose thermodynamic constraints that account for thermodynamic interactions between reactions [88,115,201], as opposed to methods which impose thermodynamically feasible reaction directions without accounting for thermodynamic coupling between reactions [54,62,114].

Figure 4.3. Example of a thermodynamically infeasible cycle in TMFA-LP.



Imposition of TMFA-feasible reaction directions on top of FBA is insufficient to eliminate cycles. This closed loop is feasible in TMFA-LP and infeasible in TMFA. The cycle generates ATP at no energetic cost by shuttling sodium ions back and forth across the membrane. The cycle relies on the sodium transport reactions NAt3_1 and NAt3_2 and the ATP synthase reaction ATPS4r.

Finally, we used our FVA results from each formulation to find a cumulative distribution function (CDF) for the flux span, the difference between the maximum and minimum flux through a given reaction (Figure 4.2B and Table S9 in the Supporting Material of the original publication). The CDF for FBA is the sharpest, with over 90% of the reactions having a span less than 300 mmol/gDW/hr. The CDF for TMFA is similar to that of FBA, despite the increase in network flexibility. The CDF becomes more shallow for RTMFA (more reactions with larger spans), as a reaction's span can increase if it becomes directionally unconstrained. The CDF for TMFA-LP was the most shallow: jumps in the CDF value at flux spans of 1000 and 2000 mmol/gDW/hr point to the existence of many thermodynamically infeasible closed cycles (e.g., A→ B → C → A) in which participating reactions operate at their maximum flux. These results highlight the important role thermodynamic interactions play in shaping a feasible flux space. We also observed that the sets of bidirectional and constrained reactions vary slightly across media conditions, though the overall CDFs remain qualitatively similar (data not shown).

4.2.3: Gene Deletion Studies: Comparison of FBA to TMFA

Under glucose aerobic conditions, CONGA identified 22 single gene deletions for which TMFA and FBA made different growth phenotype predictions (Table 4.2). In 19 cases, TMFA predicted a nonlethal phenotype and FBA predicted a lethal one, while in the remaining 3 cases TMFA predicted a lethal phenotype and FBA predicted a nonlethal one. Relative to experimental data [11], this corresponded to a better prediction by TMFA in 7 cases, and a worse prediction in the remaining 15 cases. Using RTMFA instead of TMFA introduces another 2 worse and 1 better predictions (Table 4.2).

In order for TMFA to predict growth when FBA predicts no growth, TMFA must enable a reaction to proceed in a direction not allowed by FBA. We used a variant of SL Finder [226] to identify these reactions for each knockout mutant. In cases where TMFA makes a better prediction, we hypothesized that the synthetic lethal (SL) reaction is active under the mutant phenotype. This hypothesis can be tested by knocking out the SL reaction from the single mutant: the resulting mutant should be incapable of growth. Conversely, when TMFA makes a worse prediction, the SL reaction may not operate in the predicted direction *in vivo*.

Of the seven deletions for which TMFA made a better prediction than FBA, we selected two for experimental validation: Δ*aspC* and Δ*argD*. Both mutants exhibit robust growth, and removing their SL reactions only requires a single gene deletion. For the Δ*aspC* mutant, SL Finder identified aspartase (encoded by *aspA*) as the SL reaction. After an *aspA::kan aspC* double mutant proved viable (Figure 4.4A), we identified two studies reporting *tyrB* as an isozyme for *aspC* [22,72]. A *tyrB::kan* Δ*aspC* double mutant proved nonviable (Figure 4.4B), confirming that *tyrB* and not *aspA* rescues the Δ*aspC* mutant. We can thus correct the phenotype by adding *tyrB* as an isozyme for *aspC* in the model, and by imposing a constraint that prevents aspartase from operating in the reverse direction. SL Finder also predicted that Δ*argD* was rescued by ornithine

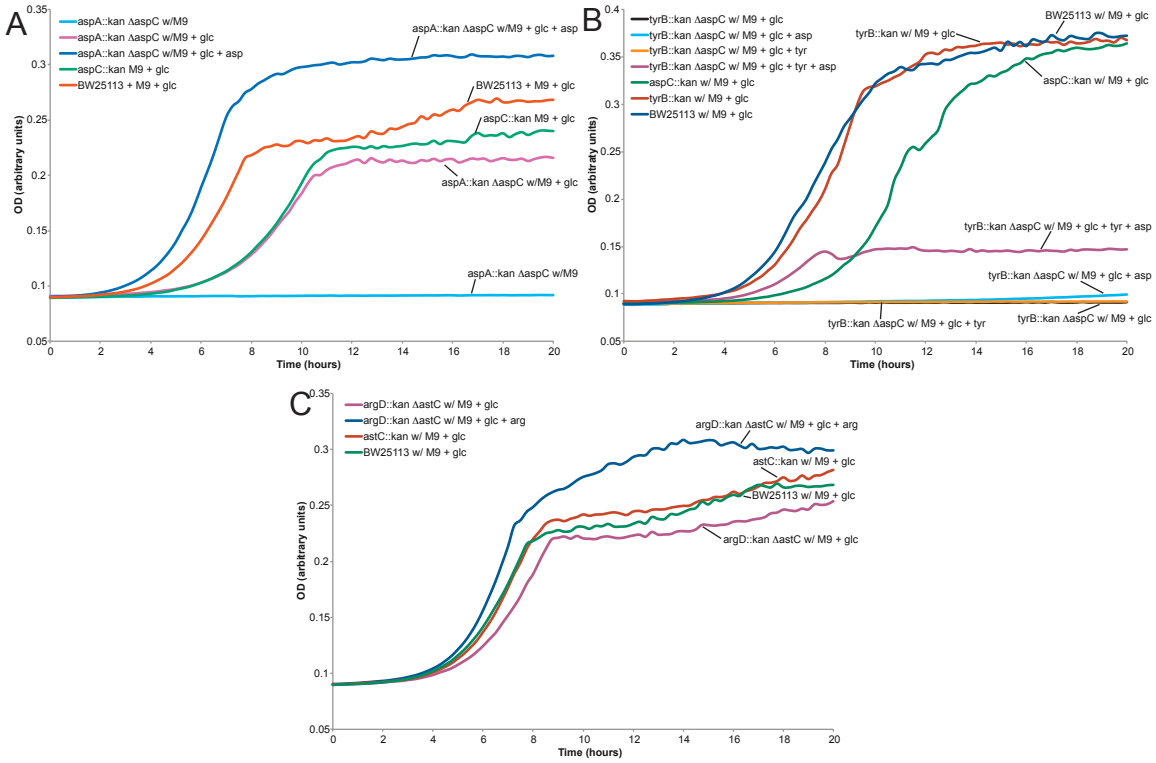
Table 4.2. Single-gene deletions for which FBA and/or (R)TMFA predict different growth phenotypes under glucose aerobic conditions.

	Gene Locus	**Gene Name**	**Phenotype**	**FBA**	**TMFA**	**TMFA (corrected)**	**RTMFA**	**RTMFA (corrected)**
Better in TMFA	b0907	*serC*	Nonlethal	Lethal	Nonlethal	Nonlethal	Nonlethal	Nonlethal
	b0928	*aspC*	Nonlethal	Lethal	Nonlethal	Nonlethal	Nonlethal	Nonlethal
	b2913	*serA*	Nonlethal	Lethal	Nonlethal	Nonlethal	Nonlethal	Nonlethal
	b3359	*argD*	Nonlethal	Lethal	Nonlethal	Lethal	Nonlethal	Lethal
	b3429	*glgA*	Nonlethal	Lethal	Nonlethal	Nonlethal	Nonlethal	Nonlethal
	b3430	*glgC*	Nonlethal	Lethal	Nonlethal	Nonlethal	Nonlethal	Nonlethal
	b4388	*serB*	Nonlethal	Lethal	Nonlethal	Nonlethal	Nonlethal	Nonlethal
Worse in Both	b0474	*adk*	Lethal	Lethal	Nonlethal	Lethal	Nonlethal	Lethal
	b0720	*gltA*	Lethal	Lethal	Nonlethal	Lethal	Nonlethal	Lethal
	b1207	*dnaR*	Lethal	Lethal	Nonlethal	Lethal	Nonlethal	Lethal
	b2615	*nadK*	Lethal	Lethal	Nonlethal	Lethal	Nonlethal	Lethal
	b2818	*argA*	Lethal	Lethal	Nonlethal	Lethal	Nonlethal	Lethal
	b3607	*cysE*	Lethal	Lethal	Nonlethal	Lethal	Nonlethal	Lethal
	b3608	*gpsA*	Lethal	Lethal	Nonlethal	Lethal	Nonlethal	Nonlethal
	b3729	*glmS*	Lethal	Lethal	Nonlethal	Lethal	Nonlethal	Nonlethal
	b3957	*argE*	Lethal	Lethal	Nonlethal	Lethal	Nonlethal	Lethal
	b3958	*argC*	Lethal	Lethal	Nonlethal	Lethal	Nonlethal	Lethal
	b3959	*argB*	Lethal	Lethal	Nonlethal	Lethal	Nonlethal	Lethal
	b4226	*ppa*	Lethal	Lethal	Nonlethal	Lethal	Nonlethal	Lethal
	b3919	*tpiA*	Nonlethal	Nonlethal	Lethal	Lethal	Lethal	Lethal
Worse in TMFA	b1849	*purT*	Nonlethal	Nonlethal	Lethal	Lethal	Nonlethal	Nonlethal
	b2500	*purN*	Nonlethal	Nonlethal	Lethal	Lethal	Nonlethal	Nonlethal
Better in RTMFA	b0888	*trxB*	Lethal	Lethal	Lethal	Lethal	Nonlethal	Nonlethal
Worse in RTMFA	b1136	*icd*	Lethal	Lethal	Lethal	Lethal	Nonlethal	Lethal
	b2780	*pyrG*	Lethal	Lethal	Lethal	Lethal	Nonlethal	Lethal

Columns labeled 'corrected' indicate (R)TMFA predictions after the additional phenotype-correction constraints (described in Table S10 in the Supporting Material of the original publication) are included.

transaminase, a reaction for which *argD* and *astC* are reported to have activity [23,24]. However, allowing ornithine transaminase to be reversible results in TMFA making four worse predictions while correcting the single Δ*argD* prediction. This suggests that ornithine transaminase might not rescue Δ*argD*. Indeed, an *argD::kan* Δ*astC* double mutant proved viable (Figure 4.4C), suggesting some other reaction or isozyme rescues the Δ*argD* mutant.

Figure 4.4. Growth curves for selected E. coli mutant strains.



(A) Growth curves for Δ*aspA::kan* Δ*aspC* mutant. **(B)** Growth curves for Δ*tyrB::kan* Δ*aspC* mutant. **(C)** Growth curves for Δ*argD::kan* Δ*astC* mutant.

In cases where TMFA made a worse prediction than FBA, we developed a phenotype-correction constraint that prevented the SL reaction from operating in the rescuing direction (Table S10 in the Supporting Material of the original publication). In some instances, SL Finder predicted multiple reactions acted together to rescue the phenotype, or that the same reaction rescued multiple phenotypes. For example, SL Finder predicted the reactions ornithine transaminase and N-acetylornithine deacetylase acted together to rescue the phenotypes of Δ*argA*, Δ*argB*, and Δ*argC*. All told, the 12 genes for which TMFA made an incorrect nonlethal prediction were associated with 10 SL reactions. We were able to identify metabolite concentration constraints for all 10 of these reactions (Table S10 in the Supporting Material of the original publication), which when implemented in TMFA resulted in correct predictions for these 12 genes (Table

4.2). For RTMFA, there were 19 SL reactions associated with the 14 genes for which RTMFA made an incorrect nonlethal prediction. We were able to identify concentration constraints for 17 of these reactions (Table S10 in the Supporting Material of the original publication), which when implemented in RTMFA resulted in correct predictions for 12 of these 14 genes (Table 4.2). With the exception of Δ*argD* noted above, correct TMFA and RTMFA predictions were unaffected by these additional phenotype-correction constraints.

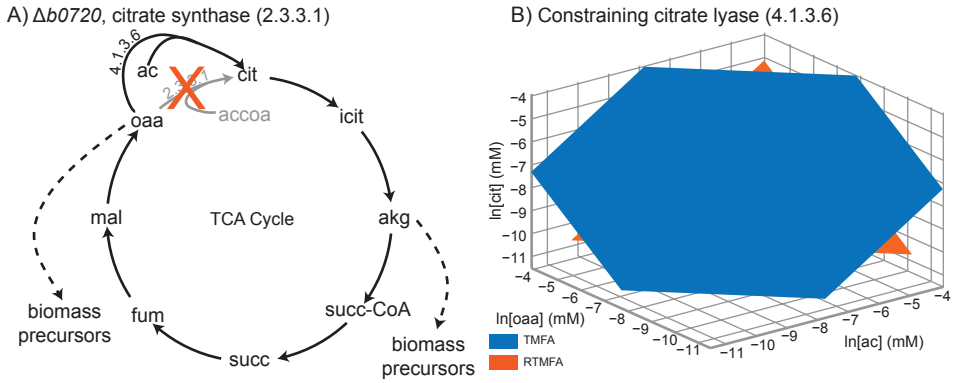
An example phenotype-correction constraint is illustrated in Figure 4.5. TMFA predicts that the deletion of citrate (cit) synthase (*gltA*, EC 2.3.3.1) is nonlethal, with citrate lyase (EC 4.1.3.6) rescuing growth. In this case citrate is synthesized from oxaloacetate (oaa) and acetate (ac), instead of oaa and acetyl-CoA (accoa) (Figure 4.5A). We found the phenotype-correction constraint for Δ*gltA* to be $\ln(ac)+\ln(oaa)-\ln(cit)<-2.76$ for TMFA, and $\ln(ac)+\ln(oaa)-\ln(cit)<-7.77$ for RTMFA. In Figure 4.5B, these constraints are indicated by the shaded planes, with concentrations in the half-space below the plane satisfying the constraint.

CONGA also identified a total of 20 double- and triple-deletions for which TMFA predicted a nonlethal phenotype and FBA a lethal one. (Table S11 in the Supporting Material of the original publication). We were able to find experimental phenotypes for 14 of these multi-gene deletions, and TMFA made a worse prediction in all cases. Taken together with the single-gene deletion data, this suggests that additional concentration measurements are needed to more accurately define the metabolic flux space and predict growth phenotypes.

We also identified 3 gene deletions for which TMFA falsely predicted a lethal phenotype (Δ*b1949*, Δ*b2500*, Δ*b3919*) when FBA predicted a nonlethal one (Table 4.2). In these cases, we hypothesized that the SL reaction active in FBA was thermodynamically infeasible in TMFA. Indeed, in 2 of the 3 cases, when uncertainty in

free energies was included, RTMFA made the correct prediction (the exception being Δ*b3919*). RTMFA also picked up an additional gene deletion (Δ*b0888*) for which TMFA and FBA both incorrectly predicted a lethal phenotype (Table 4.2). Thus, while there were cases in which thermodynamic assignment of some reaction directions led to inaccurate growth predictions, there were other cases where thermodynamic constraints were needed to explain observed growth phenotypes.

Figure 4.5. Example of reduced concentration spaces imposed by phenotype-correction constraints.



(A) TMFA incorrectly predicts that the deletion of *gltA* (citrate synthase, 2.3.3.1) is nonlethal. When *gltA* is deleted (large 'X'), TMFA predicts that citrate lyase (4.1.3.6) synthesizes citrate (cit) by operating in reverse. **(B)** To correct the phenotype, citrate lyase must be constrained to operate only in the forward direction. This requires the ac, oaa, and cit concentrations to lie beneath the shaded surface.

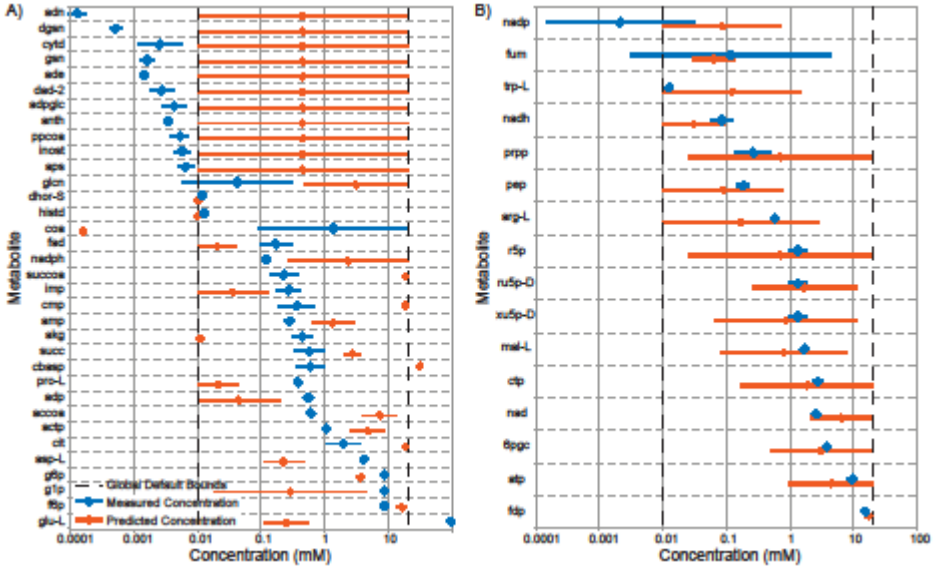
We also evaluated TMFA for aerobic growth on glycerol and anaerobic growth on glucose. We found that to enable aerobic growth in glycerol M9 medium TMFA required the same concentration constraints as the glucose case; however, for anaerobic growth on glucose TMFA required expanded concentration ranges on a slightly different set of metabolites. We also performed CONGA on single-gene deletions under both conditions. Under glycerol aerobic conditions, we found 20 gene deletions for which FBA and TMFA made different predictions, with 6 better and 14 worse predictions (Table S12 in the Supporting Material of the original publication). Under glucose anaerobic conditions, we found 26 gene deletions for which FBA and TMFA made different

predictions. We then performed growth phenotype screens for the 20 deletion strains available in the Keio collection. Assuming the 6 deletion strains that are unavailable in the Keio collection involve essential genes [11], these screens reveal that TMFA makes a better prediction than FBA in 8 cases and a worse prediction in 18 cases (Table S13 in the Supporting Material of the original publication).

4.2.4: Thermodynamic Variability Analysis: Ranges of Metabolite Concentration

Thermodynamic variability analysis (TVA) was used to study the ranges of metabolite concentrations that allow maximal growth on glucose minimal media in the absence of uncertainty in the standard Gibbs energy of formation of each group ($\Delta_{gr}G'^{0}$). Using TMFA, we identified a total of 124 (out of 618) intracellular metabolites whose feasible

Figure 4.6. Comparison of model-predicted and experimentally observed metabolite concentrations for maximal growth.



The dataset of Bennett, et al contains metabolite concentrations for 107 metabolites in *i*JR904. **(A)** Plot of the 38 metabolites for which experimental and theoretical concentration measurements do not overlap. **(B)** Plot of the 12 metabolites for which experimental and theoretical concentration measurements overlap, and for which metabolite concentration ranges are constrained by thermodynamics. (**A** and **B**) Circles (diamonds) indicate the mean predicted (measured) metabolite concentration, with horizontal bars denoting the full concentration range. Metabolite abbreviations can be found in the Supporting Material of the original publication.

concentration range was less than the default global bounds (0.001mM-20mM, see Table 4.3, and Table S14 and Figure S5 in the Supporting Material of the original publication), indicating that the thermodynamic constraints impose restrictions on metabolite concentrations.

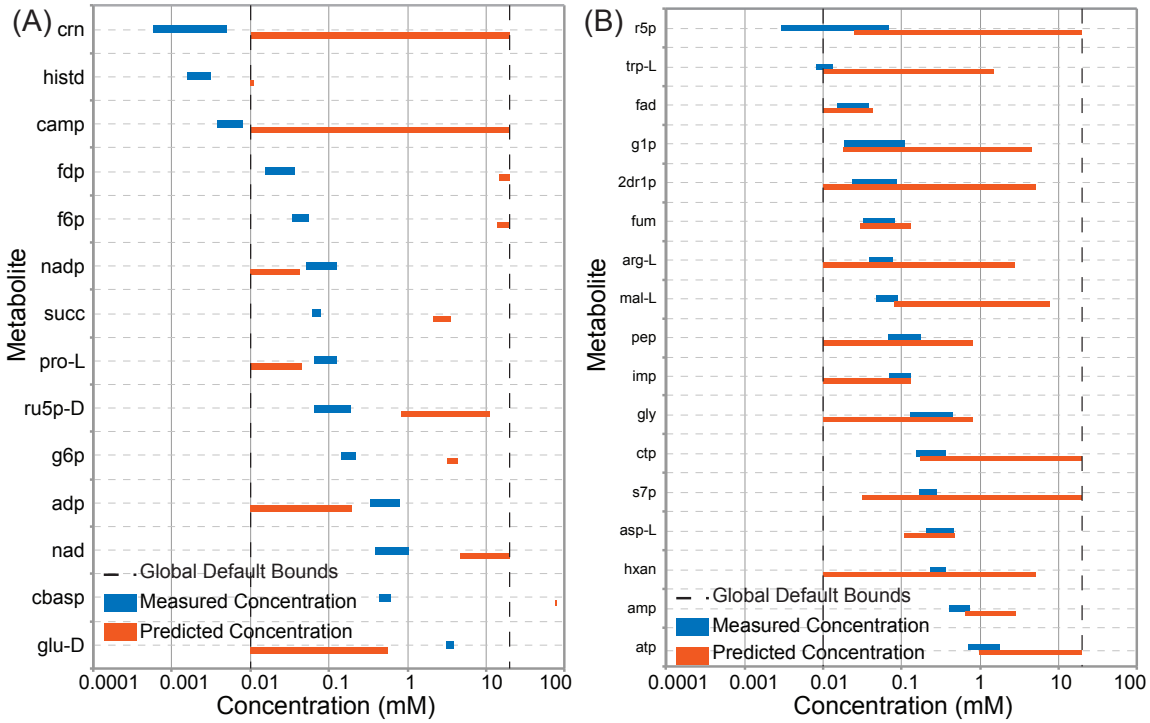
Table 4.3. Comparison of model-predicted and experimentally measured metabolite concentration ranges for glucose aerobic conditions at maximal growth, simulating growth in a batch reactor.

		Predicted by TMFA		
		Unconstrained by Thermodynamics	**Constrained by Thermodynamics**	**Total**
Experimentally Measured	**Overlap**	57	16	73
	No Overlap	11	23	34
	No Data	426	85	511
Total		494	124	618

'Constrained' indicates the concentration ranges are tighter than the global bounds (0.001mM to 20mM).

The study of Bennett, et al. [19] examined exponential growth of *E. coli*, and reported measurements for 107 metabolites in the *i*JR904 model. Of these, predicted concentration ranges overlapped with their measured values in 73 instances, and failed to overlap in 34 (Table 4.3, Figure 4.6, and Figure S6 in the Supporting Material of the original publication). Of these 34 metabolites (Figure 4.6A), 12 measurements did not overlap with their predicted values because the measurements fell outside the global bounds defined by our model. In addition, for 11 of these 12 metabolites, the concentration range predicted by TMFA spanned the full range allowed by our global bounds, indicating that changing the global bounds would likely resolve these conflicts. However, doing so is unlikely to result in tighter ranges on predicted metabolite concentrations. For the remaining 22 (out of 34) conflicting metabolites whose measurements did fall within the global bounds, thermodynamic consistency (i.e., predicted concentrations consistent with experimental measurement) could be achieved for all 22 metabolites by allowing for uncertainty using RTMFA. However, the predicted

Figure 4.7. Comparison of model-predicted (using suboptimal growth) and experimentally observed metabolite concentrations in continuous culture under glucose aerobic conditions.



The dataset of Ishii, et al. contains metabolite concentrations for 88 metabolites in the *i*JR904 model. **(A)** Plot of the 14 metabolites for which experimental and theoretical concentration measurements do not overlap. **(B)** Plot of the 17 metabolites for which experimental and theoretical concentration measurements overlap, and for which metabolite concentration ranges are predicted to be constrained by TMFA. Metabolite abbreviations can be found in Table S1 of the Supporting Material of the original publication.

concentration range for all of these metabolites in RTMFA spanned the global range, indicating that RTMFA is unable to predict these metabolite concentrations with great precision. Of the 73 instances of overlap between measured and TMFA predicted values, 16 metabolites had a feasible concentration range that was restricted by thermodynamic constraints (Figure 4.6B), while 57 did not (i.e., the ranges were the same as the global bounds). These results suggest TMFA may be more suitable as a framework for incorporating measured concentration data into constraint-based models, rather than a tool to predict experimental concentration measurements.

Table 4.4. Comparison of model-predicted and experimentally measured metabolite concentration ranges for glucose aerobic conditions, simulating growth in a CSTR.

	Predicted by TMFA			
		Unconstrained by Thermodynamics	**Constrained by Thermodynamics**	**Total**
Experimental Data	**Overlap**	57	17	74
	No Overlap	2	12	14
	No Data	433	97	530
Total				618

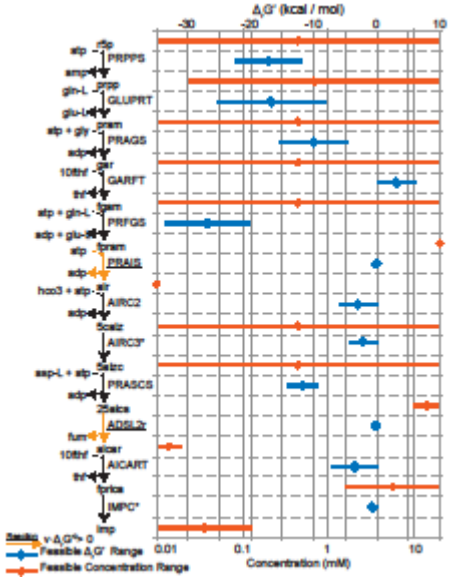
'Constrained' indicates the concentration ranges are tighter than the global bounds (0.001mM to 20mM).

A second study, from Ishii, et al. [94], examined *E. coli* grown in continuous culture at 0.2 hr^{-1}, and reported measurements for 88 metabolites in the *i*JR904 model. We performed TVA at this growth rate to examine the effect of suboptimal growth on predicted metabolite concentrations. In general, we find that fewer metabolites are predicted to have constrained ranges during suboptimal growth (97 compared to 124, Table S15 and Fig S7 in the Supporting Material of the original publication), most likely due to the increase in network flexibility associated with suboptimal growth. We also compared the model predictions to measurements taken from the dataset of Ishii, et al.[94] and found that more predictions agree with experimental measurements (74 of 88 measurements, Table 4.4, Figure 4.7 and Figure S10 in the Supporting Material of the original publication). Of the 68 metabolites measured in both studies, 42 predictions agreed with experimental measurements in both studies.

4.2.5: Examination of Thermodynamic Bottlenecks

Previous studies have utilized thermodynamic constraints to identify candidate reactions for regulation (those with transformed Gibbs energy of reaction ($\Delta_r G'$) far from zero) [70,115], and to identify thermodynamic bottlenecks in cellular metabolism [70,88]. Thermodynamic bottlenecks were first defined as reactions which render metabolic pathways infeasible for a given system with known concentrations [137,138]. The term was later used [70,88] to describe reactions for which $\Delta_r G'$ is close to equilibrium. Such

Figure 4.8. Model-predicted concentrations and $\Delta_r G'$ ranges for metabolites and reactions in the purine biosynthesis pathway.



Circles (diamonds) indicate the mean predicted concentration ($\Delta_r G'$), with horizontal bars denoting the full concentration ($\Delta_r G'$) range. Underlined reaction abbreviations signify that $\Delta_r G'$ and $\Delta_r G'^0$ have opposite signs. Starred reaction abbreviations indicate that the reaction direction in the *i*JR904 model is opposite that shown in the figure, and $\Delta_r G'$ values have had their sign reversed to agree with the direction shown. Metabolite and reaction abbreviations can be found in the Supporting Material of the original publication.

reactions are feasible for only a narrow concentration range. Thus the term thermodynamic bottleneck refers to a bottleneck in the space of potential metabolite concentrations. Thermodynamic models can provide quantitative values for the feasible concentration range of metabolites associated with a bottleneck. TVA was used to study the ranges of reaction Gibbs energies ($\Delta_r G'$) that allow maximal growth on glucose minimal media in the absence of uncertainty in $\Delta_{gr} G'^0$. Using TVA, we identified a total of 168 reactions which must operate in a single direction and whose free energy range includes equilibrium (Table S17 in the Supporting Material of the original publication).

One such reaction operating very close to equilibrium is phosphoribosylaminoimidazole synthase (PRAIS), an intermediate step in purine biosynthesis. We used TVA to identify the range of $\Delta_r G'$ and metabolite concentrations consistent with cellular growth for each reaction and metabolite in the pathway (Figure 4.8). TVA shows that the driving force for PRAIS is a large concentration gradient between 2-formamido-N(1)-(5-phospho-D-ribosyl)acetamidine (fpram) and 5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxamide (air), and PRAIS is only feasible for a narrow range of fpram and air concentrations. Conversely, reactions with a large positive or negative $\Delta_r G'$ should be relatively insensitive to metabolite concentration, and TMFA confirms this prediction. Reactions such as GLUPRT ($\Delta_r G' < 0$) remain feasible for concentration ranges spanning several orders of magnitude. Finally, PRPP synthetase (PRPPS), the primary regulatory control point for purine biosynthesis [151], also has a free energy range far from zero.

4.3: Discussion

Thermodynamics-based metabolic flux analysis (TMFA) modifies flux balance analysis with thermodynamic constraints, allowing for expanded predictive capability of constraint-based methods. TMFA ensures that all reactions operate in thermodynamically feasible directions, eliminates thermodynamically infeasible closed cycles, and accounts for thermodynamic coupling between reactions in the network. In this work, we demonstrate that thermodynamic constraints can provide a guide for predicting reaction directions in the absence of prior knowledge. We also systematically evaluated the impacts of these thermodynamic constraints on metabolic flux distributions and cellular growth rates, and we highlight the importance of explicitly accounting for thermodynamic coupling between reactions. We used TMFA to make both qualitative and quantitative predictions, and have validated these predictions

against a variety of genome-scale datasets. We show how these predictions can generate hypotheses regarding reaction directions and thermodynamic bottlenecks. We found many instances in which predictions of metabolite concentrations lack precision and/or accuracy, in which case additional types of data or constraints can be incorporated into TMFA. Identifying what additional data are most useful is an important question that should be addressed in the years ahead.

In this work, we found that TMFA was able to achieve physiologically realistic predictions of growth rates and flux distributions in the absence of uncertainty in the estimated contributions of groups ($\Delta_{gr}G'^{0}_{k,est}$) to formation energies ($\Delta_f G'^{0}$), but concentration measurements for ATP were required in the presence of uncertainty. While it is encouraging that TMFA can reproduce wild-type growth rates with a minimum of experimental data, we found that additional concentration measurements may be necessary to refine growth rate predictions for other conditions (e.g., for knockout mutants). We also observed that slightly different global concentration bounds were necessary to support growth in aerobic versus anaerobic conditions, suggesting that concentration predictions are media-dependent.

TMFA can also be used to generate hypotheses regarding reaction directions and thermodynamic bottlenecks. For example, cofactor pairs such as ATP/ADP appear together in numerous reactions, resulting in constraints on the concentration ratio of the two metabolites [88]. However, we found that constraining the ATP concentration was necessary to achieve physiologically realistic behavior in the presence of uncertainty in $\Delta_{gr}G'^{0}_{k,est}$ estimates. Thus, it may be that constraints on metabolite ratios serve to drive reactions in their physiological directions. Physiological reaction directions are often assigned based on *in vitro* characterization of enzymes under conditions which may vary significantly from those found *in vivo*. Thus, we envision TMFA complementing other

approaches [54,62,83,114] which are used to calculate thermodynamically feasible reaction directions for new genome-scale models. TMFA also promises to be a useful tool for metabolic engineering applications, by identifying thermodynamic bottlenecks in engineered pathways [211] or by pinpointing those reactions whose reversible operation enables new routes for chemical synthesis [43].

We also observed that the TMFA model is limited in its ability to predict metabolite concentrations. This may be because the majority of reactions in the *i*JR904 network are thermodynamically favorable, and thus relatively insensitive to metabolite concentrations. Obtaining tighter bounds on predicted metabolite concentrations may require the use of a penalty function [92], a thermodynamic objective [16], or the use of kinetic constraints. Recent studies have identified correlations between metabolite concentrations and physiochemical properties [12,262], including the K_M of metabolic enzymes [19]. Alternatively, incorporation of known metabolite concentrations may enable TMFA to predict concentrations of metabolites for which data are unavailable. In light of these results, we suggest that, without additional constraints, TMFA is better suited for validating phenotypes and generating hypotheses than for quantitative prediction of metabolite concentrations.

Furthermore, our results suggest that additional types of data and constraints will be needed to improve TMFA's predictions of growth phenotypes and metabolite concentrations due to uncertainties in $\Delta_f G'^0_{est}$. A recently published GCM provides tighter estimates for $\Delta_f G'^0_{est}$ [155], while other approaches have successfully combined group contribution estimates with experimentally measured $\Delta_f G'^0$ values or equilibrium constants [62,115]. We also observed that the inclusion of additional constraints on cofactor concentrations and formation energies (ATP, NAD, NADP, etc) further constrains the flux space (data not shown). Thus, experimental measurements of $\Delta_f G'^0$

for cofactors may be a promising way to improve the accuracy of thermodynamic models. Finally, we note that our approach underestimates uncertainty in $\Delta_r G'^0$ by neglecting the error associated with structural groups unchanged by a reaction, an approach which is valid only if the contributions of $\Delta_{gr} G'^0$ to $\Delta_f G'^0$ are in fact linearly additive. Additional types of data will likely be necessary when considering the error associated with these unchanged groups.

Associated with a need to reduce uncertainty in $\Delta_f G'^0_{est}$ is a need to improve model run-time performance, as large mixed-integer programs such as TMFA can be quite cumbersome. A recent Master's thesis examines a number of thermodynamic approaches (EBA, TMFA, etc.) from a theoretical and practical perspective, and provides insights into how to improve solver performance of different formulations [146]. We also observed that *a priori* thermodynamic constraints on reaction directions ((4.22) to (4.26)) reduced solution times by over an order of magnitude. Alternatively, rather than enforcing strict thermodynamic requirements, one could use thermodynamic and metabolomic data as a guide, and seek a solution which maximizes the consistency with the available data (e.g., by allowing thermodynamic and concentration constraints to be violated, and employing a penalty in the objective). One study suggests that metabolite concentrations remain stable in response to perturbations, implying a single set of metabolomics data [94] could be used to model a variety of conditions.

Chapter 5: Metabolic Modeling of Microbial Consortia

This material is being prepared for publication as:

> Hamilton JJ, Calixto Contreras, M, Reed JL (2014). Thermodynamics and H_2 Transfer in a Methanogenic, Syntrophic Community.

We developed a thermodynamic, multi-species model of the syntrophic association between the anaerobic bacterium *Syntrophobacter fumaroxidans* and the methanogenic archaeon *Methanosprillum hungatei*. In pure culture, *S. fumaroxidans* ferments fumarate to succinate and CO_2, while in the presence of *M. hungatei* it converts propionate to acetate, CO_2, and H_2 [180,219]. H_2 serves as the electron carrier between the two species, and its production is only observed during syntrophic growth. Using a thermodynamic, constraint-based model, we set out to test the proposed hypothesis that this behavior is governed by thermodynamics [147,204,215,221,222].

We first developed genome-scale network reconstructions of both microorganisms, and stoichiometrically and thermodynamically verified proposed mechanisms of electron transfer for each species. We then identified additional constraints and the cellular objective function required to predict the proper flux through experimentally characterized carbon and electron transfer pathways during monoculture and syntrophic (i.e., coculture) growth. Our analysis revealed that thermodynamic constraints alone are insufficient to correctly predict the dynamics of H_2 production by *S. fumaroxidans*.

We also extended TMFA to model the syntrophic association between the two micro-organisms. The association is modeled as a continuous coculture system with constraint-based models for each microbe and a mass balance around the reactor. The resulting coculture model accounted for the biomass concentrations of each species. We predicted the behavior of this syntrophic association under a variety of conditions, and

identified two distinct regimes of behavior, depending on H_2 availability. The coculture model describes the contributions of different H_2 production modes to electron transfer in the community, and predicts that *S. fumaroxidans*' may alter its metabolic behavior in the presence of a high relative abundance of *M. hungatei*.

5.1: Results

Draft reconstructions of *Syntrophobacter fumaroxidans* and *Methanosprillum hungatei* were tested and parameterized against experimental data, leading to the *i*Sfu648 and *i*Mhu273 reconstructions, respectively. The *i*Sfu648 and *i*Mhu273 models were used to stoichiometrically and thermodynamically verify proposed mechanisms of electron transfer for each species. Analysis of H_2 production in the *i*Sfu658 model revealed that thermodynamic constraints alone are insufficient to correctly predict the mechanism of H_2 production by *S. fumaroxidans*. Analysis of the community using a thermodynamic coculture model revealed the contributions of different H_2 production modes to electron transfer in the community. The coculture model also predicts that *S. fumaroxidans*' may alter its metabolism in the presence of a high relative abundance of *M. hungatei*.

5.1.1: Testing and Parameterizing the *i*Mhu273 Metabolic Model

Methanogens can be classified into two groups based on the presence or absence of cytochromes in their energy transfer mechanisms [235]. To date, genome-scale models (GEMs) have been constructed for two methanogens, *Methanosarcina barkeri* [53,75] and *Methanosarcina acetivorans* [18,199], both of which use cytochromes. To the best of our knowledge, there are no published GEMs for methanogens lacking cytochromes, such as *M. hungatei*.

The reconstruction of *M. hungatei* was built from the *i*MB745 reconstruction of *M. acetivorans* [18], the newest methanogen reconstruction available at the time this work began. A preliminary draft reconstruction was built based on sequence homology,

but the reconstruction contained less than 200 genes (results not shown). Instead of performing extensive gapfilling, all reactions from the *i*MB745 *M. acetivorans* reconstruction were copied into the *M. hungatei* reconstruction, with modifications to reflect key metabolic features of *M. hungatei*. A comparison of the *i*Mhu273 reconstruction to other recent reconstructions of methanogens is given in Table 5.1.

Table 5.1. A comparison of the *i*Mhu273 *M. hungatei* reconstruction to other recent methanogen reconstructions.

	***i*Mhu273**	***i*MB745**	***i*MG746**
Organism	*M. hungatei*	*M. acetivorans*	*M. barkeri*
Genes	273	745	746
Reactions	737	756	741
GPRs	285	629	615
Metabolites	638	715	642

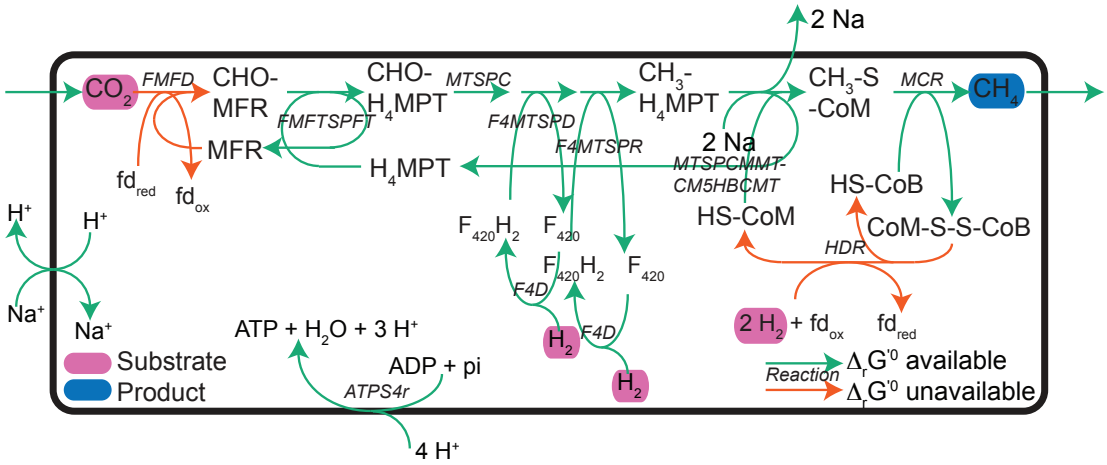
'Reaction counts do not include exchange reactions or biomass. Metabolite counts do not include extracellular metabolites. GPR stands for gene-protein-reaction association.

The reconstruction was then converted to a thermodynamic model, as described in Methods. Molecular structure files (molfiles) were obtained for 94% of the metabolites in the reconstruction, enabling the calculation of the standard transformed Gibbs free energy of reaction ($\Delta_r G'^0$) for 83% of the reactions in the network. Thermodynamics-based metabolic flux analysis (TMFA) was used to predict growth and ATP generation mechanisms in the *i*Mhu273 model and compared those against known mechanisms. Simulations were performed in monoculture under a defined minimal medium.

TMFA predicted a no-growth phenotype, due to the inability of the *i*Mhu273 model to oxidize H_2S to SO_3^{2-} via sulfite reductase, a necessary reaction for biomass production. The estimated $\Delta_r G'$ for this reaction was between 55.6 kJ/mol and 284 kJ/mol, indicating that the reaction could not proceed in the direction required for growth. However, replacing the coenzyme F_{420} with a generic ferredoxin enabled sulfite reductase to proceed in the required direction.Because sulfur metabolism in

methanogens remains poorly understood [128], this difference in sulfur metabolism between *M. acetivorans* and *M. hungatei* seems plausible.

Figure 5.1. Carbon and electron transfer mechanisms in *M. hungatei*.



The methanogenesis pathway has a net stoichiometry of CO_2 + 4 H_2 → CH_4. The substrates and products of this transformation are indicated with purple and blue ovals, respectively. Reactions for which $\Delta_r G'^0$ estimates are available are indicated with green arrows; reactions for which $\Delta_r G'^0$ estimates are not available are indicated with orange arrows. Metabolite abbreviations: CH_3-H_4MPT: 5-Methyl-5,6,7,8-tetrahydromethanopterin; CH_3-S-CoM: methylcoenzyme m; CHO-H_4MPT: 5-Formyl-5,6,7,8-tetrahydromethanopterin; CHO-MFR: formylmethanofuran b; CoM-S-S-CoB: heterodisulfide; F_{420}: coenzyme ferredoxin 420-2 (oxidized); $F_{420}H_2$: coenzyme ferredoxin 420-2 (reduced); Fd_{ox}: ferredoxin (oxidized) 2[4Fe-4S]; Fd_{red}: ferredoxin (reduced) 2[4Fe-4S]; H_4MPT: 5,6,7,8-tetrahydromethanopterin; HS-CoB: coenzyme b; HS-CoM: coenzyme m; MFR: Methanofuran b. Reaction abbreviations: ATPS4R: ATP synthase (3.6.3.14); F4D: coenzyme F420 hydrogenase/dehydrogenase (1.12.98.1); F4MTSPD: methylenetetrahydromethanopterin dehydrogenase (1.5.98.1); F4MTSPR: 5,10-methylenetetrahydromethanopterin reductase (1.5.98.2); FMFD: formylmethanofuran dehydrogenase (1.2.99.5); FMFTSPFT: formylmethanofuran--tetrahydromethanopterin formyltransferase (2.3.1.101); HDR: heterodisulfide reductase (1.8.98.1); MCR: methyl-coenzyme M reductase (2.8.4.1); MTSPC: N(5),N(10)-methenyltetrahydromethanopterin cyclohydrolase (3.5.4.27); MTSPCMMT-CM5HBCMT: tetrahydromethanopterin S-methyltransferase (2.1.1.86).

Experimental evidence suggests that *M. hungatei* is able to generate 0.5 mol ATP per mol of CO_2 converted to CH_4 [235]. A methanogenesis pathway which produces this ATP yield could be identified (Figure 5.1), but the $\Delta_r G'^0$ of one reaction (FMFTSPFT, formylmethanofuran-tetrahydromethanopterin formyltransferase) had to be allowed to vary within a 99% confidence interval (rather than 95% used for all other reactions) of

its estimated standard transformed Gibbs energy of reaction ($\Delta_r G'^{0}_{est}$). The mechanism for carbon transfer is well-characterized [60,127,209,234,235], but uncertainty remains about the stoichiometry of electron transfer and small ion transport [235]. Na^+ transport stoichiometries associated with tetrahydromethanopterin S-methyltransferase (MTSPCMMT_CM5HBCMT, E.C. 2.1.1.86) and an Na^+/H^+ antiporter were selected to give an ATP gain matching the experimental estimates.

Additionally, the *i*Mhu273 model predicted other, higher-yielding, ATP-generating mechanisms outside the methanogenesis pathway. The enumeration of all such ATP-generating cycles [176] proved computationally intractable, so a previously proposed probability-based approach [62] was used to qualitatively constrain reaction directions in the *i*Mhu273 model. This approach calculates the probability that a reaction's $\Delta_r G'^{0}$ is negative. If the probability is greater (less) than 70% (30%), the reaction is constrained to the forward (reverse) direction. Constraining all 497 such reactions eliminated these higher ATP yielding mechanisms, but also resulted in a no-growth phenotype.

To overcome this problem, a new optimization approach, Minimal Probabilistic Sets (MPS) was developed. MPS minimizes the number of qualitative reaction direction constraints needed to reduce the maximum ATP gain to 0.5 mol ATP/mol CO_2, while still allowing for growth. A set of 41 reactions had to be constrained to a single direction. Many of the reactions identified by MPS may be thermodynamically unidirectional *in vivo* due to intracellular metabolite concentrations related to the cellular energy charge [5], but these constraints are not captured in the *i*Mhu273 model. These additional reaction direction constraints were used in all subsequent analyses with the *i*Mhu273 model.

It has been observed that biomass and energy generation are independent processes in *M. hungatei*, with CO_2 being the sole source of ATP (using methanogenesis), and acetate being the sole source of carbon for biomass [51]. In

contrast, the *i*Mhu273 model predicts that *M. hungatei* can produce some biomass using CO_2 alone, and MPS could not identify any qualitative reaction direction constraints which would prevent CO_2 from being used to generate biomass while still allowing biomass production from acetate. Nonetheless, during growth on CO_2 and acetate, over 90% of the CO_2 consumed is used for methanogenesis.

Experimental data were subsequently used to identify substrate uptake rates (SUR), and the growth- (GAM) and non-growth-associated (NGAM) ATP maintenance requirements for *M. hungatei*. NGAM represents the amount of energy spent to maintain the cell (i.e., maintenance energy), while GAM represents energy spent on growth-related functions (e.g., protein synthesis). For the *i*Mhu273 model, the NGAM was estimated to be 0.6 mmol ATP/gDW/day, GAM was estimated to be 47 mmol ATP/gDW, SUR_{CO2} was estimated to be 75.7 mmol/gDW/day, and $SUR_{formate}$ was estimated to be 955 mmol/gDW/day.

5.1.2: Testing and Parameterizing the *i*Sfu648 Metabolic Model

The *i*Sfu648 reconstruction of *S. fumaroxidans* was built using the RAVEN Toolbox [1], which uses protein orthology to construct draft reconstructions from the proteins and reactions in the KEGG database. The resulting draft reconstruction was manually refined following recommendations given in a recent review [80].

5.1.2.1: Curating Carbon and Electron Transfer Modes

Experimental studies have elucidated five growth modes for *S. fumaroxidans*: four in monoculture and one in coculture with *M. hungatei* (Table 5.2) [180,181,219]. In monoculture, *S. fumaroxidans* ferments fumarate to succinate and CO_2 (Figure 5.2A) and the bacterium can also reduce fumarate to succinate alone, using formate or H_2 as an electron acceptor. *S. fumaroxidans* can also reduce propionate to succinate using fumarate as an electron acceptor (Figure 5.2B). In coculture with *M. hungatei*, *S.*

fumaroxidans grows on propionate, producing acetate and CO_2 (Figure 5.2C); however, *S. fumaroxidans* can not grow on propionate alone in monoculture.

Table 5.2. Experimentally observed and computationally predicted extracellular flux distributions for *S. fumaroxidans*

Growth Mode	**Ideal Stoichiometry (No Growth)**	**Predicted Stoichiometry (Max Growth)**	**Observed Growth (1/days)**	**Predicted Growth (1/days)**
Fumarate, Monoculture	7 fumarate → 6 succinate + 4 CO_2	7 fumarate → 5.60 succinate + 4.37 CO_2	0.3261	0.3261
Fumarate + Formate, Monoculture	fumarate + formate → succinate + CO_2	fumarate + 0.93 formate → 0.94 succinate + 1.01 CO_2	0.5455	0.2748
Fumarate + H_2, Monoculture	fumarate + H_2 → succinate	fumarate + 0.93 H_2 → 0.94 succinate + 0.09 CO_2	0.4615	0.2748
Propionate + Fumarate, Monoculture	propionate + 3 fumarate → acetate + CO_2 + succinate	0.37 propionate + 3 fumarate → 0.53 acetate + 1.77 CO_2 + 2.34 succinate	0.7317	0.574
Propionate, Coculture	propionate → acetate + CO2 + 3 H_2	propionate → 0.93 acetate + 0.99 CO2 + 2.93 H_2	0.1351	0.1351

The *i*Sfu648 model predicts the following carbon transfer pathways for five major growth modes in *S. fumaroxidans* in the absence of biomass growth [180,219].

Electron transfer mechanisms in *S. fumaroxidans* are considerably less clear than carbon transfer mechanisms. A number of studies have identified gene clusters encoding a variety of hydrogenases, dehydrogenases, and other electron transfer enzymes [29,30,147,181,215,251]. Among the more notable of these are:

- a confurcating hydrogenase, which couples the favorable oxidation of reduced ferredoxin with the unfavorable production of H_2 from NADH
- a confurcating formate dehydrogenase, similar to the confurcating hydrogenase, but producing formate instead of H_2 from NADH and CO_2

- a proton-pumping, RNF-type ferredoxin:NAD+ oxidoreductase, which uses an ion gradient to drive the unfavorable formation of reduced ferredoxin from NADH

All told, 7 enzymes which catalyze 11 different electron transfer reactions were identified. In many cases, the proposed reactions catalyzed by these enzymes differ between publications, and the most consistent annotation in the literature was used when assigning reactions to enzymes.

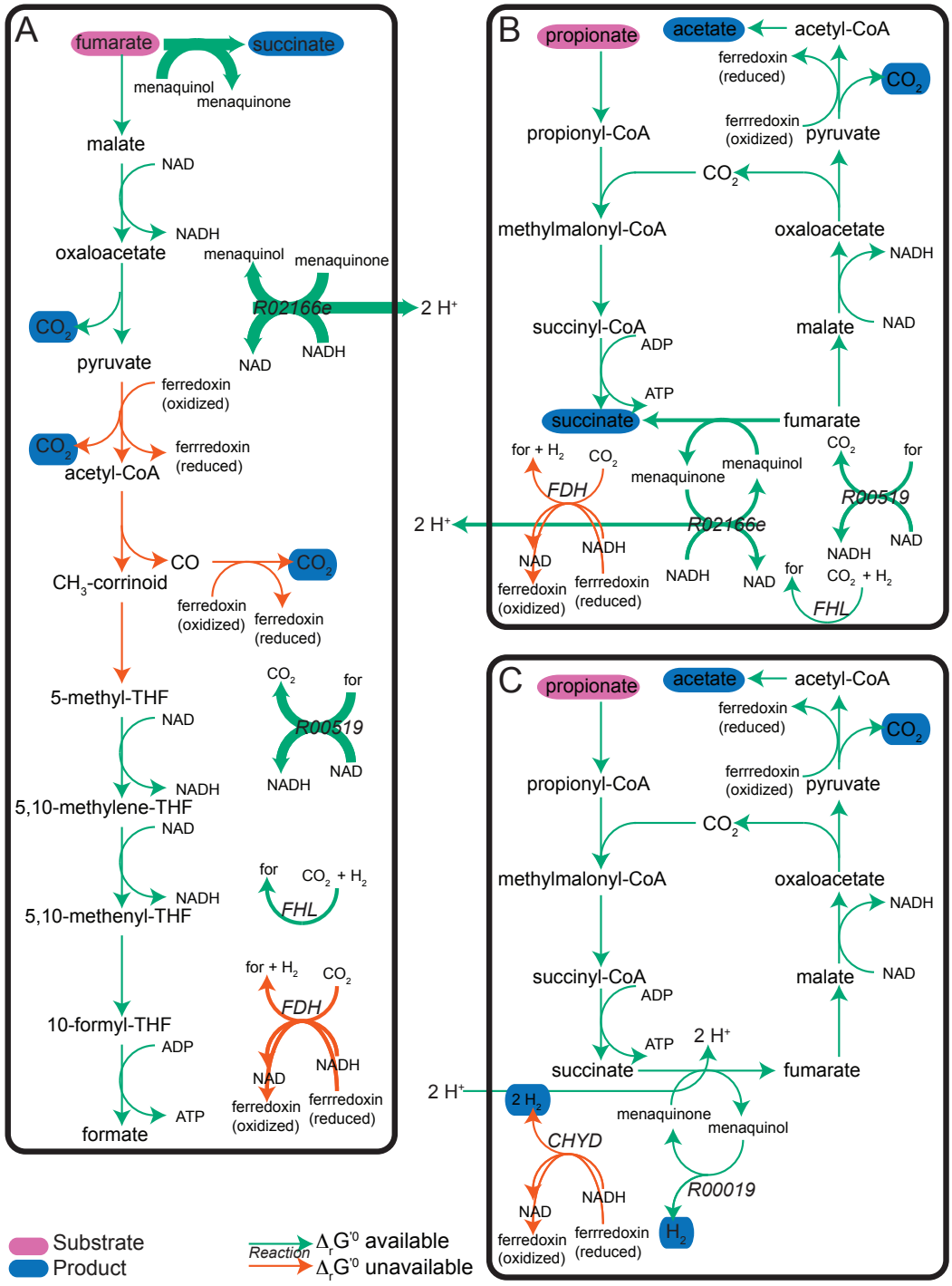
The draft reconstruction was updated with these carbon and electron transfer reactions, and the resulting stoichiometric model was converted to a thermodynamic model. Estimates for $\Delta_r G'^0$ were obtained for approximately 84% of the reactions in the *i*Sfu648 metabolic reconstruction. For reactions involved in electron transfer for which $\Delta_r G'^0$ could not be estimated, a "lumped" reaction approach was used to constrain the transformed Gibbs energy of reaction ($\Delta_r G'$).

5.1.2.2: Validation of ATP Synthesis Modes

A thermodynamically consistent metabolic network should contain no thermodynamically infeasible closed cycles [14]. Such cycles may become problematic if they enable undesirable behaviors, such as free ATP production. The thermodynamic model for *i*Sfu648 allowed ATP producing cycles when no nutrients were provided. As with the *i*Mhu273 reconstruction, MPS was used to minimize the number of qualitative reaction direction constraints needed to eliminate these cycles, while maintaining biomass growth. Due to different carbon and electron transfer mechanisms in mono- and coculture, different sets of reaction direction constraint were identified for the two conditions.

Experimental evidence suggests that the carbon and electron transfer mechanisms described above provide the sole source for ATP production in *S. fumaroxidans*, either

Figure 5.2. Carbon and electron transfer mechanisms in *S. fumaroxidans* for each metabolic mode.



Simulations were carried out by maximizing ATP production with no biomass production requirement. The overall stoichiometry of each pathway is given in Table 5.1. Plot legend information can be found in Figure 5.1. (A) Monoculture growth on fumarate alone. (B) Monoculture growth on fumarate and propionate. (C) Syntrophic growth on propionate.

by substrate-level phosphorylation or through establishment of an proton gradient used

by ATP synthase [204]. However, the *i*Sfu648 model includes additional pathways for ATP synthesis with higher ATP yields. To identify these pathways, the *i*Sfu648 model was first used to identify the highest amount of ATP that could be produced using the experimentally known mechanisms. The *i*Sfu648 model was then analyzed to find and eliminate other higher ATP yielding pathways.

To first determine the maximum ATP yield of these known pathways, FBA was used to maximize ATP production under each of the five growth modes, only allowing flux through known carbon and electron transfer pathways. During this stage, additional constraints necessary to give the observed extracellular flux distributions were identified. For example, during fumarate fermentation (Figure 5.2A), fumarate gets directly converted to succinate and malate at a ratio 6:1. This ratio does not fall naturally out of the pathway stoichiometry, and was introduced as an additional constraint.

Figure 5.2 shows the results for three of the five conditions. During fumarate fermentation (Figure 5.2A), seven moles of fumarate produce at most four ATP molecules: one via substrate-level phosphorylation and three via ATP synthase. NADH: menaquinone oxidoreductase (R02166, 1.6.5.3) transfers electrons from NADH to menaquinone and pumps protons across the membrane. Fumarate fermentation provides some of the necessary NADH, with formate dehydrogenase (R00519, 1.2.1.2) supplying the remainder. Fumarate fermentation also requires oxidized ferredoxin, which is generated via the confurcating formate dehydrogenase (CFDH). Additional flux through formate dehydrogenase (R00519, 1.2.1.2) and formate hydrogen-lyase (FHL) balances the NAD and formate supply. The simulations also revealed that when *S. fumaroxidans* reduces propionate to succinate using fumarate as an electron acceptor (Figure 5.2B), 2 mols of ATP can be produced per mol of propionate: one via substrate-level phosphorylation and one via ATP synthase. Electrons are transferred via the same

reactions as in fumarate fermentation, though the reactions carry different fluxes. Finally, during growth in coculture on propionate alone, 1 mol ATP gets produced per mol propionate (Figure 5.2C). Here, ATP is generated via substrate-level phosphorylation, which is then used to generate a proton gradient (via reverse action of ATP synthase). The proton gradient is necessary to drive the endergonic oxidation of succinate to fumarate, giving a net yield of 0.5 ATP. This is in sharp contrast to monoculture growth on fumarate, in which fumarate reduction to succinate is used to establish a proton gradient for ATP synthesis. The *i*Sfu648 model also predicts a different electron transfer mechanism than is seen in monoculture, in which hydrogen:ferredoxin oxidoreductase (R00019, 1.12.7.2) provides the menaquinone for succinate oxidation, and the confurcating hydrogenase supplies NAD for fumarate oxidation.

Finally, these known ATP-generating mechanisms were compared to other ATP-generating mechanisms in the network. As with the *i*Mhu273 reconstruction, mechanisms with higher ATP yields were identified, and MPS was used to eliminate them. For each growth condition, MPS identified the smallest number of qualitative reaction direction constraints needed to maintain biomass growth while ensuring the appropriate maximal ATP gain. MPS could not identify a minimal set of qualitative reaction direction constraints which was valid under all conditions. Many of the reactions identified by MPS may be thermodynamically unidirectional *in vivo* due to intracellular metabolite concentrations related to the cellular energy charge [5], but these constraints are not captured in the *i*Sfu648 model. These additional reaction direction constraints were used in all subsequent analyses with the *i*Sfu648 model.

5.1.2.3: Parameterization of the *i*Sfu648 Metabolic Model

Experimental data was used to identify the substrate uptake rates (SUR), and the growth- (GAM) and non-growth-associated (NGAM) maintenance requirements for *S. fumaroxidans*. These parameters were estimated using data from growth on fumarate alone and propionate alone. For the *i*Sfu648 model, the following parameters resulted in the best fit of the model to the experimental data: NGAM = 5.04 mmol ATP/gDW/day, GAM = 31 mmol ATP/gDW, $SUR_{propionate}$ = 23.84 mmol/gDW/day, and $SUR_{fumarate}$ = 50.97 mmol/gDW/day.

Using these parameter values, the *in-silico* growth rates under each growth condition were predicted (Table 5.2). Not surprisingly, the predicted growth rates for fumarate alone and propionate alone conditions agree with experimental observations (since these were used to estimate the parameter values). However, the *i*Sfu648 model significantly under-predicts growth rates in three monoculture modes (fumarate with formate; fumarate with H_2; and propionate with fumarate).

5.1.2.4: Validation of Extracellular Flux Distributions

When maximizing biomass production on the entire network, the *i*Sfu648 model predicted a wide range of extracellular flux distributions [180,181,219]. When the enzyme cost (i.e. total flux) was minimized at maximum biomass growth (pTMFA [122]) the *i*Sfu648 model correctly predicted the extracellular flux distribution for three of five growth modes (Table 5.2): monoculture growth on fumarate alone, monoculture growth on fumarate with H_2, and coculture growth. These results indicate that the majority of carbon is diverted to fermentation products, consistent with the expectation that high fluxes through the low-energy fermentation pathways are needed to meet cellular energy demands.

However, for monoculture growth on fumarate with formate, and monoculture growth on propionate with fumarate, pTMFA incorrectly predicted H_2 as a byproduct (data not shown). Hydrogen production is not seen experimentally under these conditions, as it is widely thought that H_2 production is only thermodynamically favorable at low partial pressures [147,215,221,222]. In particular, methanogens in syntrophic communities enable sustained H_2 production by consuming H_2 and keeping its partial pressure low [147,204,215,221,222]. Indeed, when H_2 production was observed in monoculture, H_2 production ceased at a partial pressure of approximately 10 Pa [206].

Since the *i*Sfu648 model included thermodynamic constraints it was unclear why H_2 was being predicted as a by-product. Further investigation found that the directions of three reactions (confurcating hydrogenase, confurcating dehydrogenase, and formate-hydrogen lyase) needed to be qualitatively constrained to prevent H_2 production under these two conditions. Previously, metabolite concentration ratios have been calculated such that, if imposed as constraints, would make a thermodynamic model's reaction directionality consistent with the qualitative reaction direction constraints [79]. However, metabolite concentration ratios that would prevent these three reactions from enabling H_2 production could not be identified in this study.

We speculate that we could not obtain metabolite concentration ratio constraints because too much uncertainty regarding estimates for Gibbs free energies exists in the *i*Sfu648 model. Specifically, by using $\Delta_r G'^0$ as a basis for thermodynamic calculations, thermodynamic interconnectivity arising from shared metabolites cannot be accounted for. As a result, a constraint blocking H_2 production under these two conditions was used. This constraint allowed H_2-producing reactions to proceed as long as the network produced no net H_2. Upon doing so, pTMFA correctly predicted the extracellular flux distributions in four out of the five conditions tested, as shown in Table 5.2.

The remaining disagreement between the *i*Sfu648 model predictions and experimental observations occurs during growth on propionate with fumarate. Under this condition, fumarate was predicted to supply all the carbon for biomass, with propionate providing additional ATP via substrate-level phosphorylation. Fumarate fermentation is predicted to provide most of the cell's energy needs under this condition, with just enough propionate consumed to satisfy the energy balance.

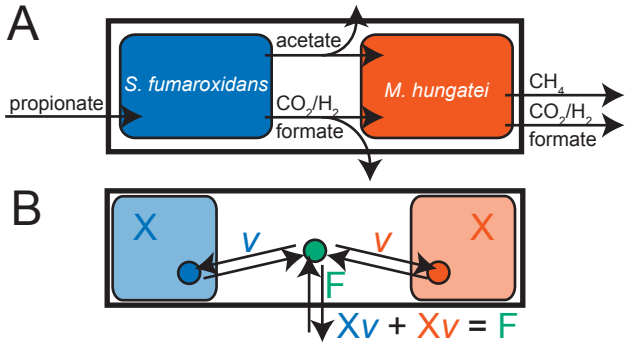
5.1.3: Behavior of *M. hungatei* and *S. fumaroxidans* in Coculture

5.1.3.1: Predicting Extracellular Flux Distributions

The two individual models (*i*Sfu468 and *i*Mhu273) were combined to build a coculture model that was used to examine the syntrophic association between *M. hungatei* and *S. fumaroxidans* (Figure 5.3). During growth in coculture, *S. fumaroxidans* takes up propionate from the environment, converting it to acetate, CO_2 and H_2, or formate (Figure 5.3A) [180,219]. *M. hungatei* consumes these by-products, converting acetate to biomass and CO_2/H_2 and formate to CH_4 [51]. *M. hungatei* can also optionally interconvert excess CO_2 and formate via a formate dehydrogenase [28].

To simulate this association, a continuous coculture system was modeled using constraint-based models for each microbe and equations for the reactor. The resulting coculture model accounted for the biomass concentrations of each species (Figure 5.3B). In this system, both species grow at the same rate (equal to the dilution rate) and exchange metabolites (colored circles) with the medium. The medium components (cells and metabolites) are subject to mass balance constraints, which relate each species' biomass concentration (X), individual species uptake/secretion fluxes (v), and the net flux into or out of the reactor (F). Propionate was the only substrate in the reactor feed (i.e., it had a net flux into the reactor), ensuring that all carbon and electrons used by *M. hungatei* must be produced by *S. fumaroxidans*.

Figure 5.3. Diagram of coculture simulations.



(A) Allowed carbon exchange fluxes. Propionate is fed into the system, which *S. fumaroxidans* converts to acetate, and CO_2/H_2 or formate. These metabolites are secreted into the media, where they can be consumed by *M. hungatei*. *M. hungatei* converts acetate to biomass, and CO_2/H_2 and formate to CH_4. *M. hungatei* can also optionally interconvert excess CO_2 and formate via a formate-hydrogen lyase. (B) Mass balance around the coculture system. *S. fumaroxidans* and *M. hungatei* are present in the reactor at biomass concentration X. They exchange metabolites (circles) into a shared environment with flux v. The flux F represents the net flux into / out of the system, and is related to v and X for each species via the given equation.

While the overall conversion of propionate by this community in continuous culture has not been characterized experimentally, an overall reaction of

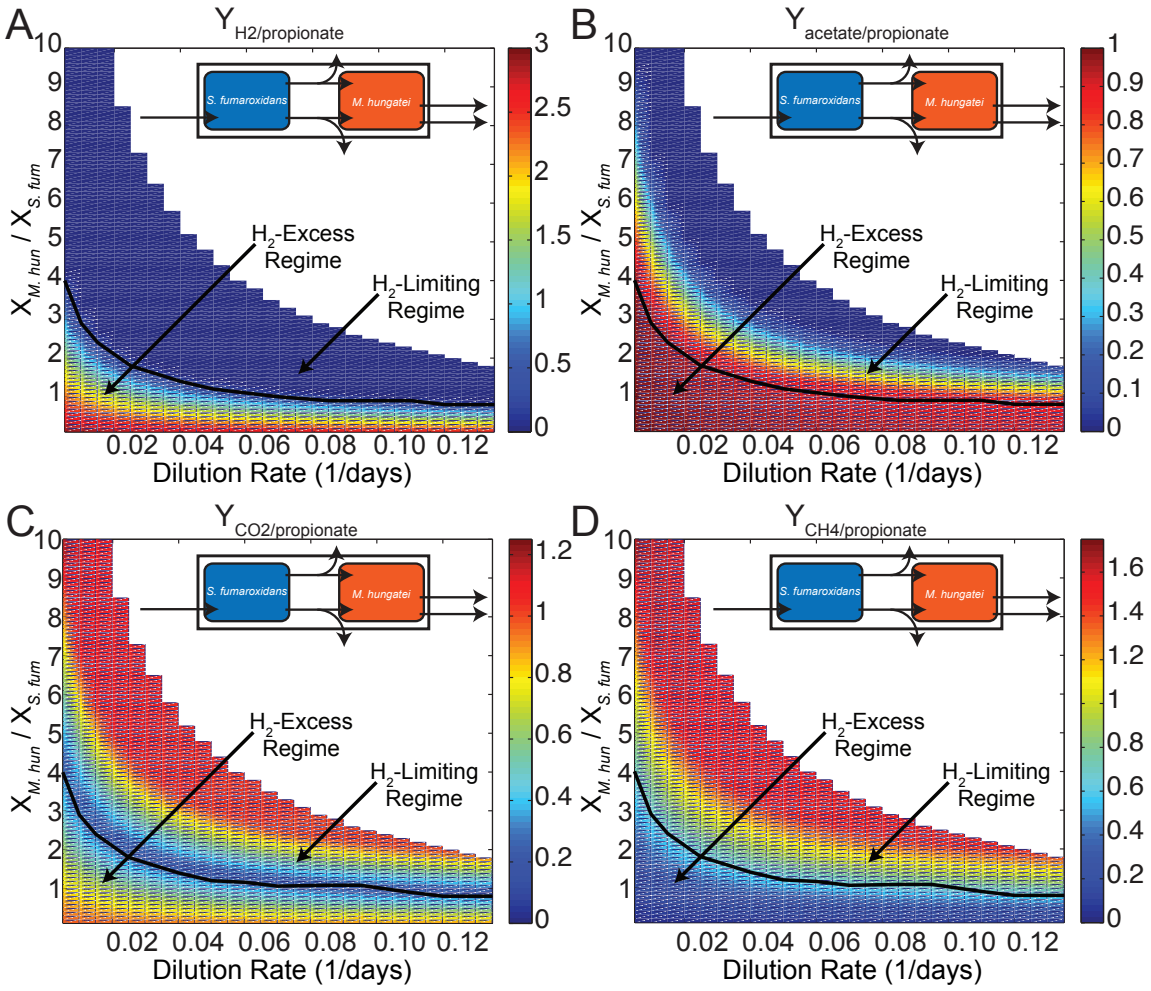
$$\text{propionate} \rightarrow \text{acetate} + 0.25\ CO_2 + 0.75\ CH_4 \qquad (5.1)$$

has been proposed previously for a ratio of 3 *M. hungatei* to 4 *S. fumaroxidans* ($X_{M.hun}/X_{S.fum} = 0.75$) [204,206]. When the dilution rate approached 0.135 $days^{-1}$ (the maximum growth rate for *S. fumaroxidans*), the coculture model predicted an overall propionate conversion reaction by *M. hungatei* and *S. fumaroxidans* (when $X_{M.hun}/X_{S.fum} = 0.75$) which was similar to the proposed reaction:

$$\text{propionate} \rightarrow 0.78\ \text{acetate} + 0.29\ CO_2 + 0.77\ CH_4 \qquad (5.2)$$

The coculture model was then used to explore the behavior of the community under a variety of operating conditions, by systematically changing the reactor dilution rate and the relative ratio of *M. hungatei* to *S. fumaroxidans*, while allowing unlimited propionate uptake by the reactor (Figure 5.4).

Figure 5.4. Diagram of community by-product yields in the coculture system.



The plots show the yield of (A) H_2 , (B) acetate, (C) CO_2, and (D) CH_4 per mole of propionate as a function of the dilution rate of the reactor (X-axis) and the ratio of species biomass concentrations (*M. hungatei* to *S. fumaroxidans,* Y-axis). Plot colors correspond to values as indicated to the right of the chart. The black curve indicates the approximate onset of the H_2-limited regime, and the white region indicates conditions under which the reactor balance is infeasible. The maximal dilution rate corresponds to the maximal growth rate of the slower-growing *S. fumaroxidans*. (Insets) Diagram of coculture simulations as in (3A), illustrating that the yields are calculated around the entire community.

At a fixed dilution rate, the coculture model predicted that changing the ratio of *M. hungatei* to *S. fumaroxidans* affected the community's product yields. At low $X_{M.hun}/X_{S.fum}$ ratios, *S. fumaroxidans* produced H_2 in excess of *M. hungatei*'s energy needs (labeled as the H_2-excess regime), while at higher $X_{M.hun}/X_{S.fum}$ ratios H_2 became

the limiting nutrient (labeled as the H_2-limiting regime) (Figure 5.4A). Within this H_2-limiting regime, the $X_{M.hun}/X_{S.fum}$ ratio could increase until *M. hungatei*'s demand for H_2 exceeded *S. fumaroxidans*' ability to supply it, and reactor balance became infeasible (Figure 5.4A).

The coculture model also predicted that at low dilution (growth) rates, the onset of H_2 limitation and reactor infeasibility occurred at higher $X_{M.hun}/X_{S.fum}$ ratios than it did at high dilution rates. Methanogenesis in *M. hungatei* provides the energy needed for growth- (GAM) and non-growth-associated (NGAM) maintenance. At low growth rates, *M. hungatei*'s maintenance energy needs are lower; thus, requiring less CO_2 and H_2, and fewer *S. fumaroxidans* cells, to support growth.

The coculture model predicted other interesting changes to the community's product yields as well. First, the coculture model predicted that the community's proposed product yields occur along the line demarcating the H_2-excess and H_2-limitng regimes (the black curves in Figure 5.4). This result is unsurprising, as the proposed community yield assumed H_2 limitation. However, as the $X_{M.hun}/X_{S.fum}$ ratio increased within the H_2-limiting regime, the community's product yields shifted away from the proposed product yields. Specifically, the coculture model predicted a decrease in community acetate yield (Figure 5.4B) and increases in community CO_2 and CH_4 yield (Figure 5.4C and D). The coculture model predicted that this change in the community's product yields was driven by a shift in *S. fumaroxidans*' metabolic behavior, whereby acetate was converted to CO_2/H_2 or formate. These molecules were then metabolized by *M. hungatei* to produce CH_4. Finally, the coculture model predicted that both formate and H_2 could serve as electron shuttles in this community.

5.1.3.2: Metabolic Flexibility of *S. fumaroxidans*

The coculture model simulations indicated that formate and H_2 could both serve as electron shuttles. In the H_2-excess regime, electrons could be transferred via H_2, or electrons could be transferred via both H_2 and formate.

When all electrons were transferred via H_2, *S. fumaroxidans*' confurcating hydrogenase (CHYD) coupled the favorable oxidation of reduced ferredoxin with the unfavorable production of H_2 from NADH, as in Figure 5.2C. The overall reaction catalyzed by *S. fumaroxidans* was approximately

$$\text{propionate} \rightarrow \text{acetate} + CO_2 + 3\ H_2 \tag{5.3}$$

(Figure 5.5A). When formate also served as an electron shuttle, *S. fumaroxidans*' confurcating formate dehydrogenase (CFDH) coupled the oxidation of reduced ferredoxin with the production of formate from NADH and CO_2, with pyruvate oxidation supplying the CO_2. Flux through CHYD decreased to compensate for increasing CFDH flux. The net effect was an increase in formate production and an equimolar decrease in CO_2 and H_2 production. If formate production was maximized, the overall reaction catalyzed by *S. fumaroxidans* was approximately

$$\text{propionate} \rightarrow \text{acetate} + \text{formate} + 2\ H_2 \tag{5.4}$$

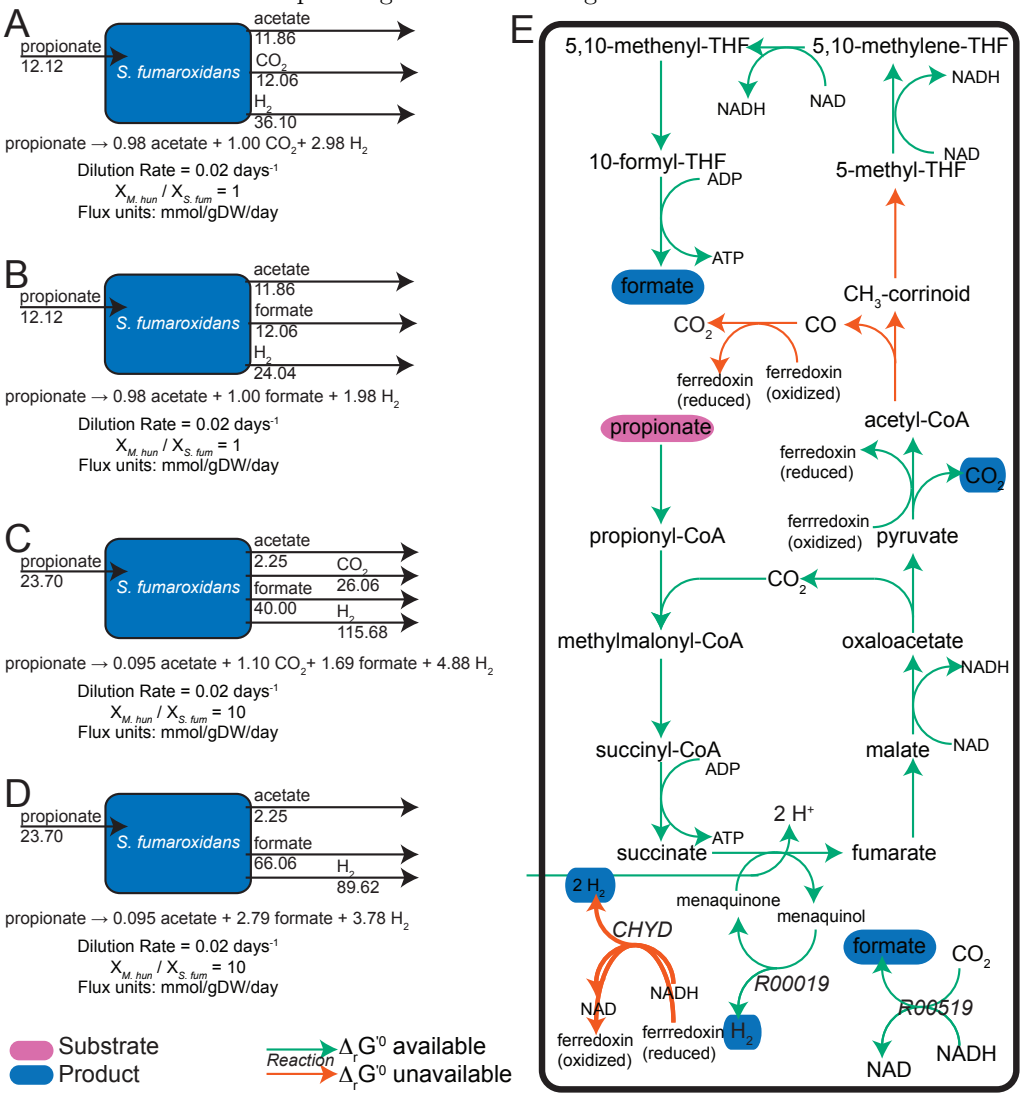
(Figure 5.5B).

Within the H_2-limiting regime, the coculture model predicted a shift in the extracellular flux distribution of *S. fumaroxidans* as the $X_{M.hun}/X_{S.fum}$ ratio increased: acetyl-CoA was further oxidized to CO_2/H_2 and formate. Under this condition, the overall reaction catalyzed by *S. fumaroxidans* was approximately:

$$\text{propionate} \rightarrow CO_2 + 2\ \text{formate} + 5\ H_2 \tag{5.5}$$

(Figure 5.5C). Propionate was oxidized to acetyl-CoA as expected, but acetate was further oxidized to formate and CO_2 (as in fumarate fermentation in monoculture) (Figure 5.5E). Hydrogen:ferredoxin oxidoreductase (R00019, 1.12.7.2) still provided the menaquinone for succinate oxidation, and the confurcating hydrogenase (CHYD)

Figure 5.5. The coculture model predicts distinct extracellular flux distributions around *S. fumaroxidans* as the reactor operating conditions change.



(A) A representative flux distribution around *S. fumaroxidans* in the H_2-excess regime when H_2 serves as the sole electron shuttle. (B) A representative flux distribution around *S. fumaroxidans* in the H_2-excess regime when both formate and H_2 serve as electron shuttles. (C) A representative flux distribution around *S. fumaroxidans* in the H_2-limiting regime when both formate and H_2 serve as electron shuttles. Both CO_2 and formate provide carbon for methanogenesis in *M. hungatei*. (D) A representative flux distribution around *S. fumaroxidans* in the H_2-limiting regime when both formate and H_2 serve as electron shuttles. Only formate provides carbon for methanogenesis in *M. hungatei*. (A) to (D) Simulations were carried out at a dilution rate of 0.02 $days^{-1}$, with *S. fumaroxidans* and *M. hungatei* present at the indicated ratios. Flux units are in mmol/gDW/day. (E) Carbon and electron transfer mechanisms in *S. fumaroxidans* at high levels of *M. hungatei*. Plot legend information can be found in Figure 5.1.

supplied NAD for fumarate oxidation. Fumarate was oxidized to 2 mol CO_2 and 1 mol

formate. The coculture model also predicted that fumarate oxidation drove the formate dehydrogenase (R00519, 1.2.1.2), which converted one molecule of CO_2 to formate. The resulting additional supply of formate/H_2 (7 mol of electron pairs, instead of 3 mol) enabled *S. fumaroxidans* to supply the energy needed to support a high relative abundance of *M. hungatei*.

As in the H_2-excess regime, multiple extracellular flux distributions around *S. fumaroxidans* were possible in the H_2-limiting regime. The confurcating formate dehydrogenase (CFDH) could also supply NAD for fumarate oxidation, in which case the overall reaction catalyzed by *S. fumaroxidans* was approximately:

$$\text{propionate} \rightarrow 3 \text{ formate} + 4 \text{ } H_2 \tag{5.6}$$

(Figure 5.5D).

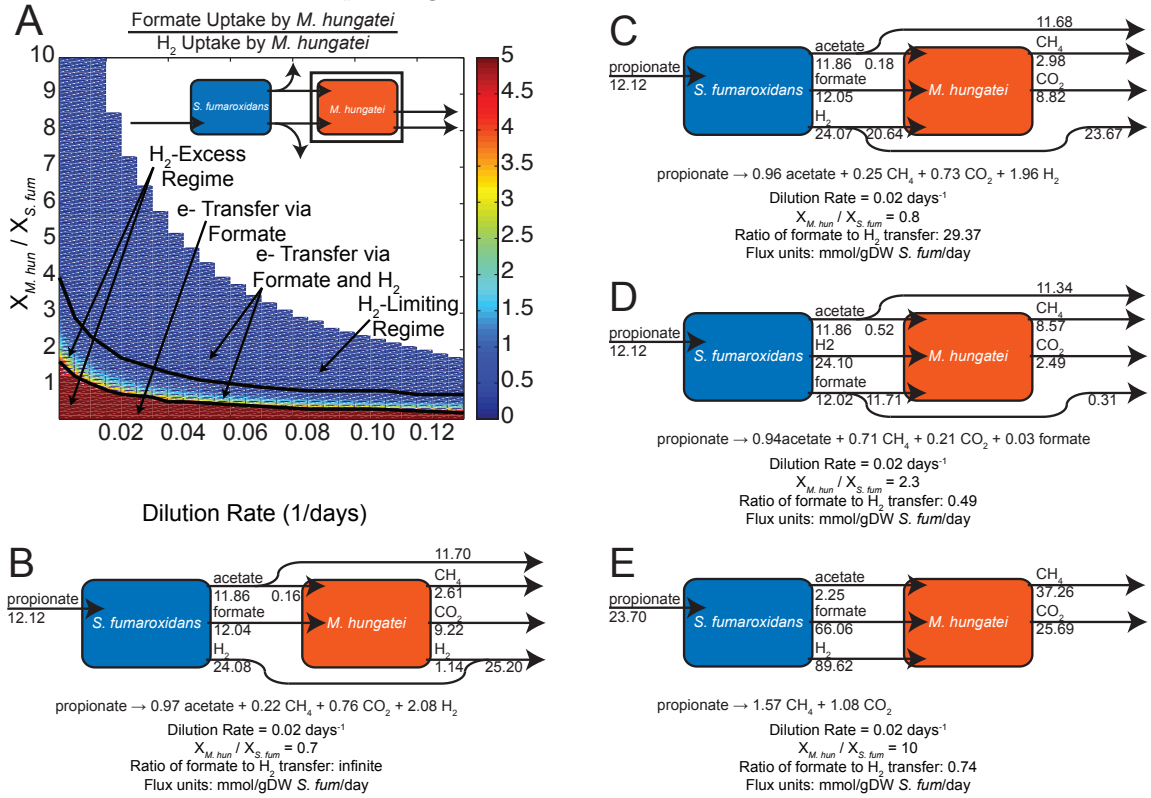
5.1.3.3: Interspecies Formate and Hydrogen Transfer

Studies have shown that in coculture, *S. fumaroxidans* passes electrons to *M. hungatei* via formate and H_2 [28,47,220]. Formate is believed to be the dominant electron shuttle between the two microbes based on estimated diffusion rates [47,220] and the high formate dehydrogenase activities in syntrophically grown *S. fumaroxidans* and *M. hungatei* [28]. Additionally, other studies have shown that *S. fumaroxidans* cannot grow in coculture with methanogens that consume only H_2 [46,220].

Having established that the coculture model predicted electron transfer via formate and H_2, the coculture model then was used to explore how the maximum ratio of formate to H_2 consumed by *M. hungatei* (the transfer ratio) changed with the reactor operating conditions (Figure 5.6A).

In the H_2-excess regime, the simulations identified dilution rate and $X_{M.hun}/X_{S.fum}$ ratio conditions where formate could be the sole source of electrons for *M. hungatei* (labeled as the "electron transfer via formate" regime, Figure 5.6A). In this regime, formate could provide all of the electrons needed to satisfy *M. hungatei*'s energy needs.

Figure 5.6. Maximum ratio of formate to H_2 transfer between *S. fumaroxidans* and *M. hungatei*, as a function of the reactor operating conditions.



(A) The plot shows the maximum ratio of formate to H_2 transfer between *S. fumaroxidans* and *M. hungatei*, as a function of the dilution rate of the reactor (X-axis) and the ratio of species biomass concentrations (*M. hungatei* to *S. fumaroxidans,* Y-axis). Plot colors correspond to values as indicated to the right of the chart. For clarity of presentation, transfer ratios above 5 have been rounded down to 5. The black curve indicates the approximate onset of the indicated regimes, and the white region indicates conditions under which the reactor balance is infeasible. The maximal dilution rate corresponds to the maximal growth rate of the slower-growing *S. fumaroxidans*. (Insets) Diagram of coculture simulations as in (3A), illustrating that the ratio is calculated using uptake rates of *M. hungatei*. (B) Species-weighted representative flux distributions around *S. fumaroxidans* and *M. hungatei* in the H_2-excess regime when formate alone serves as the electron shuttle. (C) Species-weighted representative flux distributions around *S. fumaroxidans* and *M. hungatei* in the H_2-excess regime when both formate and H_2 serve as electron shuttles. (D) and (E) Species-weighted representative flux distributions around *S. fumaroxidans* and *M. hungatei* in the H_2-limiting regime when both formate and H_2 serve as electron shuttles. (B) to (E) Simulations were carried out at a dilution rate of 0.02 $days^{-1}$, with *S. fumaroxidans* and *M. hungatei* present at the indicated ratios.. Flux units are in mmol/gDW *S.fumaroxidans*/day.

M. hungatei converted formate to CO_2 and H_2 for methanogenesis, and secreted the excess CO_2 and H_2 into the reactor (Figure 5.6B). None of the H_2 produced by *S. fumaroxidans* was taken up by *M. hungatei*, so the formate to H_2 transfer ratio was

effectively infinite. Thus, it seems plausible that formate is the dominant electron shuttle between the two microbes, provided that the community operates in the H_2-excess regime.

At higher $X_{M.hun}/X_{S.fum}$ ratios, formate alone could no longer provide all of the electrons needed to satisfy *M. hungatei*'s energy needs. As a result, *M. hungatei* needed to uptake H_2 as well as formate (Figure 5.6C), and the formate to H_2 transfer ratio decreased (labeled as the "electron transfer via formate and H_2" regime, Figure 5.6A). The transfer ratio decreased until the onset of H_2 limitation, at which point the formate to H_2 transfer ratio was approximately one to two (Figure 5.6D).

In the H_2-excess regime, *M. hungatei* still needed to uptake H_2 as well as formate (again labeled as the "electron transfer via formate and H_2" regime). The shift in *S. fumaroxidans*' metabolic behavior resulted in an increase in the transfer of both formate and H_2, and the formate to H_2 transfer ratio increased somewhat with the $X_{M.hun}/X_{S.fum}$ ratio, becoming approximately three to four when the reactor balance became infeasible (Figure 5.6E).

5.2: Discussion

The *i*Mhu273 and *i*Sfu648 thermodynamic models were successfully used to stoichiometrically and thermodynamically verify proposed carbon and electron transfer pathways in *M. hungatei* and *S. fumaroxidans*. Nonetheless, interesting questions about these pathways remain unanswered. A thermodynamic coculture model of the syntrophic association between these species confirmed the importance of formate and H_2 to electron transfer in the community. The use of a single-level optimization to describe the coculture resulted in novel predictions which bear further scrutiny.

5.2.1: Validation and Parameterization of *i*Mhu273 Metabolic Model

The final reconstruction of *M. hungatei* contained only 273 genes, compared to 745

genes in the *M. acetivorans* reconstruction from which it was built. Even after altering the sequence homology parameters to their most generous, the RAVEN Toolbox did not generate a noticeably larger reconstruction. This suggests that manual curation of RAVEN reconstructions remains necessary, unless high-quality reconstructions of one (or more) closely related organisms are available. (*M. acetivorans* and *M. hungatei* diverge taxonomically below the class level.)

Despite containing a complete methanogenesis pathway, the *i*Mhu273 model was unable to say anything about the stoichiometry of the energy-converting (Eha- or Ehb-type) hydrogenase (EHA, 1.12.7.2), which is thought to pump H^+/Na^+ while reducing ferredoxin [235]. The heterodisulfide reductase (HDR, 1.8.98.1) can also reduce ferredoxin, and the *i*Mhu273 model predicted it to be the only ferredoxin-reducing reaction required for methanogenesis. This observation is consistent with the observation that the expression of Eha/Ehb is considerably lower than that of HDR [231].

Additionally, different stoichiometries for small molecule transport during methanogenesis remain thermodynamically possible. For example, the group contribution method predicted that tetrahydromethanopterin S-methyltransferase (MTSPCMMT_CM5HBCMT, E.C. 2.1.1.86) could drive transport of up to 4 Na^+ ions at standard conditions, instead of the 2 Na+ ions in the iMhu273 reconstruction. Furthermore, some studies suggest the archaeal A_1A_0 ATP synthase is coupled to Na^+ instead of H^+ translocation [178,235]. In this case, the *i*Mhu273 model predicted the Na+/H+ antiporter was not involved in energy generation from methanogenesis, as Na^+ ions from tetrahydromethanopterin S-methyltransferase were solely responsible for ATP synthesis. Thus, while the proposed methanogenesis pathway is consistent with available data, it is not the only possibility.

5.2.2: Validation and Parameterization of *i*Sfu648 Metabolic Model

The *i*Sfu648 model also has some important limitations. In particular, the *i*Sfu648 reconstruction does not distinguish between cytosolic, periplasmic, and membrane-bound versions of electron transfer complexes (such as H_2ase). For example, the fumarate reductase and succinate dehydrogenase complexes are membrane-bound, and menaquinol should transfer electrons to a membrane-bound or periplasmic enzyme complex. However, the *i*Sfu648 reconstruction does not differentiate between differently-localized versions of the same complex. As a consequence, the *i*Sfu648 model's predicted electron transfer mechanisms may be simpler than occur *in vivo*.

In addition, the *i*Sfu648 model was unable to confirm the hypothesized directions of important electron transfer reactions, including the confurcating hydrogenase, formate dehydrogenase, and the RNF-type oxidoreductase. This problem arose due to the inability of the group contribution method to estimate the standard transformed Gibbs energy of formation ($\Delta_f G'^0$) of ferredoxin, which resulted in no estimate for $\Delta_r G'^0$ for these reactions.

This work also raises interesting questions about the appropriate mathematical basis for deriving thermodynamic constraints. Previously, we used the $\Delta G'^0$ of groups directly when modeling thermodynamics [79], and found that introducing uncertainty into the *i*Sfu648 thermodynamic model made the problem computationally difficult. In this work, we used $\Delta_r G'^0$ as the basis for thermodynamic calculations, and could handle uncertainty without any computational difficulties. However, using $\Delta_r G'^0$ as a basis for thermodynamic calculations results in too much network flexibility, as it does not account for thermodynamic interconnectivity arising from shared metabolites. As a result, the model-predicted feasible $\Delta_r G'^0$ range through a linear combination of reactions considerably exceeds the group-contribution predicted $\Delta_r G'^0$ range of the

overall reaction. For example, pTMFA erroneously predicted H_2 production from fumarate alone with the following stoichiometry:

$$\text{fumarate} + 2\ H^+ + 4\ H_2O \rightarrow 4\ CO_2 + 6\ H_2 \quad (5.7)$$

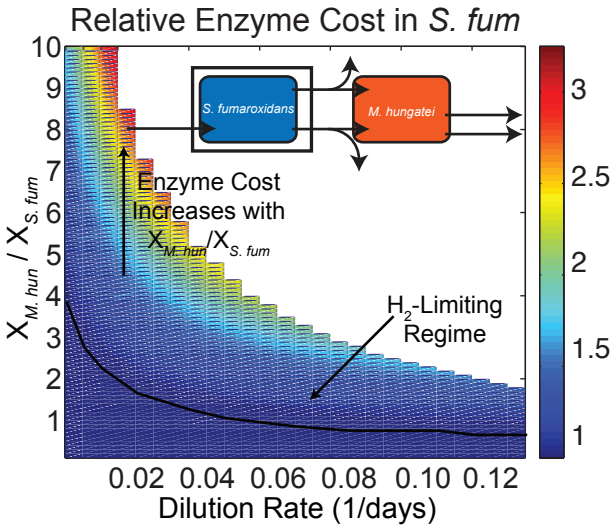
The group contribution method predicted a 95% confidence interval for $\Delta_r G'^0$ of 147 to 279 kJ/mol: thermodynamically unfavorable, as expected. Taking a linear combination of $\Delta_r G'^0$ for the reactions which make up this H_2-producing pathway, the *i*Sfu648 model predicted a $\Delta_r G'^0$ range of -315 to 5052 kJ/mol. The $\Delta_r G'^0$ range is approximately 40 times larger than that given by group contribution, and indicates that H_2 production is thermodynamically possible. This explains why pTMFA predicted H_2 production under conditions where it is not seen experimentally.

Thermodynamic interconnectivity can be captured using the $\Delta G'^0$ of molecules or groups directly when modeling thermodynamics, as well as through lumping constraints. Enumerating all such constraints which could be applied to a network is computationally infeasible, but it may be possible to identify important linear combinations of reactions which would benefit from a lumping constraint.

5.2.3: Behavior of *M. hungatei* and *S. fumaroxidans* in Coculture

Some constraint-based approaches to the study of microbial communities have been formulated as bilevel optimization problems (e.g., OptCom [266,267]). To facilitate the solution process, these optimization problems are typically reformulated as single-level optimization problems, by replacing the inner maximization problems with optimality [35,59] or complementary slackness [107] conditions. However, no such conditions exist for mixed-integer programs [77] such as TMFA, so we developed a single-level formulation with a community-level objective and thermodynamic constraints.

Figure 5.7. As the ratio of *M. hungatei* to *S. fumaroxidans* increases, so does the enzyme investment required by *S. fumaroxidans*.



The plot shows the enzyme cost in the presence of the *M. hungatei* to *S. fumaroxidans* ratio constraint, relative to the enzyme cost in the absence of this constraint. Plot legend information can be found in Figure 5.4.

This formulation predicted a shift in the metabolic behavior of *S. fumaroxidans* at high relative abundances of *M. hungatei*. Such a shift implicitly assumes that *S. fumaroxidans* is aware of the levels of *M. hungatei* in the environment and can alter its metabolism accordingly. At present, mechanisms for communication between *S. fumaroxidans* and *M. hungatei* are unknown, and known genes for quorum sensing are missing from related syntrophic bacteria [141,214]. As a result, the coculture model may overpredict the feasible ratios of *M. hungatei* to *S. fumaroxidans*. Additionally, as a consequence of this metabolic shift, the enzyme utilization of *S. fumaroxidans* was predicted to increase (Figure 5.7) at high ratios of *M. hungatei* to *S. fumaroxidans*. At the highest feasible ratios, the increase was approximately three-fold, an increase which may be unachievable in practice if the metabolism of *S. fumaroxidans* is insufficiently flexible.

The coculture model correctly predicted that both H_2 and formate are involved in electron transfer in this community. The coculture model also predicted formate could

be the dominant electron shuttle in the H_2-excess regime, while H_2 was always predicted to be dominant in the H_2-limiting regime. However, experimental measurements showing formate to be the dominant electron shuttle appear to have been taken from the H_2-limiting regime, with H_2 having a measured soluble concentration of approximately 20 nM [28]. We speculated that formate may be preferred over H_2 for electron transfer for thermodynamic reasons, as the $\Delta_r G'^0$ of (5.4) is more favorable (negative) than that of equation (5.3) in which formate is not exchanged. Thus by exchanging formate in place of CO_2/H_2, the metabolism of *S. fumaxoridans* becomes slightly less sensitive to thermodynamics.

5.3: Methods

5.3.1: Reconstruction of the *i*Mhu273 Metabolic Model

The *i*Mhu273 reconstruction of *M. hungatei* was built from the *i*MB745 reconstruction of *M. acetivorans* [18], the newest methanogen reconstruction available at the time this work began. A preliminary draft reconstruction was built using the RAVEN Toolbox [1], which uses sequence homology to construct draft reconstructions from the proteins and reactions in the KEGG database. Briefly, the RAVEN Toolbox uses protein homology to identify the KEGG Orthology (KO) ID, which best matches each gene. The reactions and genes corresponding to that KO ID are then imported into the reconstruction. Sequence homology was computed using default parameters: e-value $< 10^{-30}$, alignment length > 200 nucleotides, sequence similarity $> 40\%$.

Unfortunately, the reconstruction contained less than 200 genes. Instead of performing extensive gapfilling, all reactions from the *i*MB745 *M. acetivorans* reconstruction were copied into the *i*Mhu273 reconstruction, with modifications to reflect key metabolic features of *M. hungatei*.

Because *M. hungatei* can only utilize acetate and CO_2 [51], reactions which enabled growth on CO and methylated carbon sources were removed. In addition, the oxidative arm of the TCA cycle was replaced with the reductive arm [6]. The methanogensis pathway was also replaced [235].

The *i*Mhu273 reconstruction was then converted to a thermodynamic model. On defined minimal medium, TMFA predicted a no-growth phenotype, due to the inability of the *i*Mhu273 model to oxidize H_2S to SO_3^{2-} via sulfite reductase, a necessary reaction for biomass production. The estimated $\Delta_r G'$ for this reaction was between 55.6 kJ/mol and 284 kJ/mol, indicating that the reaction could not proceed in the forward direction required for growth. However, replacing the coenzyme F_{420} with a generic ferredoxin enabled sulfite reductase to proceed in the required direction. Because sulfur metabolism in methanogens remains poorly understood [128], this difference in sulfur metabolism between *M. acetivorans* and *M. hungatei* seems possible.

Substrate uptake rates (SURs), and growth- (GAM) and non-growth-associated (NGAM) maintenance requirements for the *i*Mhu273 model of *M. hungatei* were estimated using experimental data. NGAM represents the amount of energy spent to maintain the cell (i.e., maintenance energy), while GAM represents energy spent on growth-related functions (e.g., protein synthesis). In GEMs, these ATP requirements are usually expressed using ATP hydrolysis reactions: in the case of NGAM, the ATP hydrolysis reaction is constrained to some lower bound, and in the case of GAM, an ATP hydrolysis term is added to the biomass equation.

The value of the NGAM parameter was found by maximizing ATP hydrolysis at the reported maintenance cost as measured in terms of CO_2 uptake [206]. The value of the GAM parameter is typically obtained by plotting a series of substrate uptake and growth rates taken from chemostat experiments [241], but no such data were available for *M. hungatei*. However, studies have reported the growth rate of *M. hungatei* on CO_2

[251], and its yield per mol methane ($Y_{X/P}$) [241]. From these data, the following iterative procedure was used to simultaneously estimate SUR_{CO2} and GAM:

1. Estimate a value for the GAM.
2. Identify the SUR_{CO2} necessary to achieve the observed growth rate.
3. Calculate the *in-silico* CH_4 production rate at the observed growth rate and compute $Y_{X/P}$.
4. Adjust the value of the GAM upwards or downwards as appropriate.
5. Repeat steps 2 to 4 until the computed $Y_{X/P}$ agrees with the experimental measurement.

An identical procedure was followed to compute the SUR for growth on formate, assuming the same NGAM and GAM as for growth on CO_2. For the *i*Mhu273 model, the NGAM was estimated to be 0.6 mmol ATP/gDW/day, GAM was estimated to be 47 mmol ATP/gDW, SUR_{CO2} was estimated to be 75.7 mmol/gDW/day, and $SUR_{formate}$ was estimated to be 955 mmol/gDW/day.

The final *i*Mhu273 reconstruction contains 806 reactions, 273 genes (associated with 285 reactions), 706 metabolites. Of the 273 genes, 196 were added based on sequence homology, and 77 were added manually.

5.3.2: Reconstruction of the *i*Sfu648 Metabolic Model

The *i*Sfu648 reconstruction of *S. fumaroxidans* was built using the RAVEN Toolbox [1], which uses sequence orthology to construct draft reconstructions from the proteins and reactions in the KEGG database. Briefly, the RAVEN Toolbox uses protein homology to identify the KEGG Orthology (KO) ID, which best matches each gene. The reactions and genes corresponding to that KO ID are then imported into the *i*Sfu648 reconstruction. Sequence homology was computed using default parameters: e-value $< 10^{-30}$, alignment length > 200 nucleotides, sequence similarity $> 40\%$. The resulting

draft reconstruction was manually refined following recommendations given in a recent review [80]. While the *i*Sfu648 reconstruction was built using the RAVEN Toolbox, curation and refinement occurred in the GAMS modeling environment.

5.3.2.1: Curation of Draft *i*Sfu648 Model

First, non-metabolic reactions were removed from the draft reconstruction (primarily those involved in tRNA charging). Next, all generic metabolites (e.g., acceptor) were identified and replaced with specific metabolites (e.g., NAD). Then, all reactions in the *i*Sfu648 reconstruction were mass- and charge-balanced, with a generic ferredoxin molecule used for charge balancing if atomic balancing was insufficient to do so.

In addition, all gene-protein-reaction (GPR) relationships were evaluated. Instead of providing detailed GPRs, the RAVEN Toolbox generates lists of genes associated with each reaction, and requires users to define the detailed GPR structure themselves.

First, gene annotations were examined to ensure that they matched the function of the associated reactions. This was particularly problematic in the case of one-to-many relationships, in which a single protein can carry out multiple reactions. For each gene which matches a particular Kegg Orthology (KO) group, the RAVEN Toolbox associates all reactions in that group with the gene. Many such reactions were removed, due to a low likelihood of physiological relevance. In addition, 80 of the reactions contained in the final *i*Sfu648 reconstruction had their GPRs adjusted through the removal of one or more genes, and 16 had their GPRs replaced entirely (approximately 11% of all reactions). The draft reconstruction contained a number of hypothetical genes, and both these genes and their associated reactions were removed from the *i*Sfu648 reconstruction.

The logical structure of the GPRs provided by the RAVEN Toolbox were also determined. In addition to simple associations, in which a single gene encodes a single

enzyme, GPRs can take the form of isozymes, in which multiple genes encode distinct proteins carrying out the same function, and multimeric protein complexes, in which multiple genes encoding distinct protein subunits come together to form an active enzyme. Genes annotated as separate subunits of a complex were given an 'AND' relationship, while genes with no such annotation were given an 'OR' relationships. Of the reactions retained in the final *i*Sfu648 reconstruction, approximately 32% of them required manual curation of the logical relationships among the subunits and isozymes.

All told, the final *i*Sfu648 reconstruction retained 717 of 859 reactions (83%) and 556 of 720 genes (77%) from the draft reconstruction, and around 50% of the GPRs required some level of curation.

Finally, the RAVEN Toolbox does not provide reaction direction information, instead assuming reactions are bidirectional in the absence of any specifying information. Reaction directions for the *i*Sfu648 reconstruction were computed using group contribution.

5.3.2.2: Manual Curation of Known Growth and Electron Transfer Mechanisms

S. fumaroxidans metabolizes carbon in five well-defined ways, as described in Table 5.1. The steps of each pathway have been determined by ^{13}C-NMR experiments [180,219], and a recent genomic survey identified the genes associated with each enzyme [181]. These pathways were incorporated into the final *i*Sfu648 reconstruction, replacing draft reconstruction content where it disagreed with experimental evidence.

S. fumaroxidans also contains a wide variety of hydrogenases, dehydrogenases, and other enzyme complexes involved in electron transfer. A number of genomic surveys have proposed genes and functional roles associated with each complex [147,181,215], which were used to guide manual curation (Table 5.2). As described in Results, the

*i*Sfu648 reconstruction was able to stoichiometrically and thermodynamically verify proposed stoichiometries for these reactions.

5.3.2.3: Development and Gapfilling of Biomass Equation

The RAVEN Toolbox does not provide users with a biomass equation; they must instead construct one manually. Unfortunately, the dry cell weight biomass composition of *S. fumaroxidans* has not yet been described. In this work, a template biomass equation (proposed by the developers of the Model SEED [89]) was used as a scaffold to construct the biomass equation based on data from related organisms.

The weight fractions of major macromolecule classes (proteins, DNA, etc) were taken from *Geobacter sulfurreducens*, a deltaproteobacterium which has been extensively studied and modeled [133,134]. (*S. fumaroxidans* is also a deltaproteobacterium). Mole fractions of individual macromolecules were obtained as follows:

- amino acids, from the *i*AF1260 model of *E. coli* [54];
- DNA and RNA, from the *S. fumaroxidans* genome sequence [181];
- lipids and carbohydrates, from *G. sulfurreducens* [129];
- cell wall, cofactors, and small molecules, from the Model SEED template.

The *i*Sfu648 model was unable to sustain flux through this biomass equation under any known growth condition. Individual biomass precursors which the *i*Sfu648 model could not synthesize were identified, and SMILEY was used to identify those reaction additions which would enable biomass growth [191]. Manual curation was used to identify reactions necessary for lipid, cell wall, and carbohydrate biosynthesis. Where possible, those solutions for which genomic evidence could be found were selected. In the case of many cofactors, no genomic evidence was found for any solution. These cofactors were eliminated from the biomass equation.

SMILEY is a mixed-integer programming approach for model refinement, which calculates the minimum number of reactions from a universal reaction database which must be added to a reconstruction to enable cellular growth. For this study, the universal reaction database was a manually-curated subset of version 57 of the KEGG database, available through the BioWebDB Consortium (http://www.biowebdb.org). During curation, all reactions which met any of the following criteria were removed:

- contained an elongation;
- contained the same metabolite on both sides;
- contained compounds in the KEGG glycan database;
- contained a compound with an R group;
- contained a compound without a formula.

For reactions containing a generic acceptor molecule, two separate versions of the reaction were created, one using NAD as an acceptor, and one using NADP. Additionally, any reaction which could not be balanced through the addition of protons or ferredoxins was removed. Reaction directions for the gapfilling database were computed using group contribution, as described in the TMFA-LP section of Methods. Metabolite formulas and molfiles were obtained directly from KEGG. The curated database contains over 6000 reactions and represents approximately 70% of the original KEGG version 57 database.

5.3.2.4: Gapfilling of Blocked Reactions

Flux Variability Analysis (FVA) [132] was used to identify reactions which were blocked under one or more growth conditions. SMILEY was used to identify gapfilling solutions for these blocked reactions, and those reactions with genomic evidence were added to the *i*Sfu648 model.

5.3.2.5: Conversion to a Thermodynamic Model

The *i*Sfu648 reconstruction was then converted to a thermodynamic model. On defined minimal medium, TMFA predicted a no-growth phenotype, most likely because biosynthetic routes which were feasible under TMFA-LP may not be feasible under TMFA [79]. Once again, SMILEY identified those biomass precursors which the *i*Sfu648 model could not synthesize, and solutions which would enable biomass growth under TMFA were added to the *i*Sfu648 reconstruction.

The final *i*Sfu648 reconstruction contains 874 reactions, 648 genes (associated with 770 reactions), 893 metabolites. Considering the entire content of the final *i*Sfu648 reconstruction, reactions and genes contained in the draft reconstruction make up 82% and 86% of the total reaction and gene content, respectively.

5.3.2.6: *i*Sfu648 Model Parameterization

Substrate uptake rates (SURs), and growth- (GAM) and non-growth-associated (NGAM) maintenance requirements for the *i*Sfu648 model of *S. fumaroxidans* were estimated using experimental data. NGAM represents the amount of energy spent to maintain the cell (i.e., maintenance energy), while GAM represents energy spent on growth-related functions (e.g., protein synthesis). In GEMs, these ATP requirements are usually expressed using ATP hydrolysis reactions: in the case of NGAM, the ATP hydrolysis reaction is constrained to some lower bound, and in the case of GAM, an ATP hydrolysis term is added to the biomass equation.

The value of the NGAM parameter was found by maximizing ATP hydrolysis at the reported maintenance cost as measured in terms of propionate uptake [206]. The value of the GAM parameter is typically obtained by plotting a series of substrate uptake and growth rates taken from chemostat experiments [241], but no such data were available for *S. fumaxoridans*. However, studies have reported the growth rate of *S.*

fumaxoridans on a variety of substrates [251], and its yield per mol propionate ($Y_{X/S}$) [241]. From these data, the following iterative procedure was used to simultaneously estimate $SUR_{propionate}$ and GAM:

1. Estimate a value for the GAM.
2. Identify the $SUR_{propionate}$ necessary to achieve the observed growth rate and compute $Y_{X/S}$.
3. Adjust the value of the GAM upwards or downwards as appropriate.
4. Repeat steps 2 to 3 until the computed $Y_{X/S}$ agrees with the experimental measurement.

An identical procedure was followed to compute the SUR for growth on fumarate and formate, assuming the same NGAM and GAM as for growth on propionate. For the *i*Sfu648 model, the NGAM was estimated to be 5.04 mmol ATP/gDW/day, GAM was estimated to be 31 mmol ATP/gDW, $SUR_{propionate}$ was estimated to be 23.84 mmol/gDW/day, and $SUR_{fumarate}$ was estimated to be 50.97 mmol/gDW/day.

5.3.3: Preparation for Thermodynamic Modeling

In preparation for thermodynamic modeling, molfile structure files for all metabolites in the reconstructions were obtained. Molfiles for the *S. fumaroxidans* reconstruction were downloaded from KEGG, while molfiles for the *M. hungatei* reconstruction were downloaded from KEGG or manually reconstructed.

All metabolites were then converted to their predominant ionic species (pseudoisomer) at biochemical standard state: pH 7, zero ionic strength, and temperature 298K. The major pseudoisomeric form of each molecule was determined using pKa estimation software (Marvin pKa plug-in, version 5.11.4, ChemAxon, Budapest, Hungary). Finally, all reactions in the reconstructions were mass- and charge-

balanced using the new metabolite formulas. Supplemental Files 1 and 2 contain molfiles for all compounds in the reconstructions.

5.3.4: Thermodynamics-Based Metabolic Flux Analysis (TMFA)

Flux-balance analysis (FBA) [165] is a constraint-based technique for predicting the state of a metabolic network consistent with physiochemical principles. FBA identifies a flux distribution which maximizes cellular growth, subject to steady-state mass-balance and enzyme capacity constraints.

Specifically, given a stoichiometric matrix S and set of reactions J, FBA seeks a steady-state flux distribution (v) maximizing the flux through the biomass reaction (v_{BM}), while also satisfying mass-balance and enzyme capacity constraints for individual reactions, j:

$$\textbf{max} \quad v_{BM} \tag{5.8}$$

$$\textbf{s.t.} \quad S \cdot v = 0 \tag{5.9}$$

$$v_{\min} \le v_j \le v_{\max} \qquad \forall j \in J \tag{5.10}$$

Enzyme capacities, v_{min} and v_{max}, should be set on the basis of available evidence, such as thermodynamic irreversibility. In the absence of evidence, fluxes are typically constrained to $v_{max} = 1000$ mmol/gDW/day and $v_{min} = -1000$ mmol/gDW/day, except for measured fluxes (e.g., carbon uptake rates).

Thermodynamics-based metabolic flux analysis (TMFA, [79,88]) extends FBA via the introduction of thermodynamic constraints. TMFA makes no *a priori* assumptions about reaction directions, instead relying on the second law of thermodynamics, which states that the transformed Gibbs energy of reaction ($\Delta_r G'$) and its flux (v) have opposite signs.

$$v_j \cdot \Delta_r G_j^{'} < 0 \tag{5.11}$$

This nonlinear constraint is converted to a mixed-integer constraint as described previously [79].

The $\Delta_r G^{'}$ of a reaction is in turn a function of the standard transformed Gibbs energy of reaction ($\Delta_r G^{'0}$) and the concentrations (x_i) of those metabolites participating in the reaction:

$$\Delta_r G_j^{'} = \Delta_r G_j^{'0} + RT \sum_i S_{i,j} \ln(x_i) \tag{5.12}$$

where R is the gas constant, and T is the temperature. In the absence of specific information, metabolite concentrations are constrained to global bounds of 0.01 mM to 20 mM.

Due to a paucity of experimental data, group contribution methods (GCMs) [135,136] are used to provide estimates ($\Delta_r G_{j,est}^{'0}$) and uncertainties ($SE_{\Delta_r G_{j,est}^{'0}}$) of $\Delta_r G^{'0}$. These estimates and uncertainties were obtained using a software implementation of the latest GCM for biological systems, which the authors refer to as component contribution (CC) [156]. The implementation is available via the von Bertalanffy 2.0 add-on to the COBRA Toolbox [202].

The GCM method returns estimates of $\Delta_r G^{'0}$ for each reaction, given the pKa values for all compounds in the reconstruction, and information on temperature (T), pH, ionic strength (I) and electrical potential (Φ) in each cellular compartment. pKa values were calculated using pKa estimation software (Marvin pKa plug-in, version 5.11.4, ChemAxon, Budapest, Hungary) from molfile structure files.

The standard transformed Gibbs free energy ($\Delta_r G^{'0}$) of each reaction was allowed to vary within its 95% confidence interval, as determined by the standard error (SE) reported by the GCM software:

$$\Delta_r G_{j,est}^{'0} - 2SE_{\Delta_r G_{j,est}^{'0}} \leq \Delta_r G_j^{'0} \leq \Delta_r G_{j,est}^{'0} + 2SE_{\Delta_r G_{j,est}^{'0}} \tag{5.13}$$

If the GCM method is unable to obtain an estimate of $\Delta_r G^{'0}$ for a particular reaction, $\Delta_r G^{'0}$ was allowed to vary freely.

The GCM method estimated $\Delta_r G'^0$ for approximately 82% of the reactions in the *i*Mhu273 reconstruction, and approximately 84% of the reactions in the *i*Sfu648 reconstruction. For some reactions for which $\Delta_r G'^0$ could not be estimated, a lumping approach was used to constrain $\Delta_r G'$, as described in "Thermodynamic Lumping."

Aggregating the above constraints gives the final formulation for TMFA:

$$\begin{array}{lll} \textbf{max} & v_{BM} & \\ \textbf{s.t.} & \text{FBA constraints, (5.9) and (5.10)} & \forall j \in J \\ & \Delta_r G' \text{ constraints, (5.12) and (5.13)} & \forall j \in J \\ & \text{consistency constraints, (5.11)} & \forall j \in J \end{array} \qquad \text{(TMFA)}$$

5.3.5: Thermodynamic Lumping

Due to limitations in the GCM [156], there were some reactions for which $\Delta_r G'^0$ could not be estimated. Among these reactions were the formate dehydrogenase and confuracting hydrogenase involved in H_2 production during syntophic growth. To understand why these reactions were predicted to be active under monoculture conditions, a previously described lumping approach [79] was used to constrain the free energies of these reactions.

In this approach, reactions with unknown $\Delta_r G'^0$ are linearly combined into lumped representations for which $\Delta_r G'^0$ can be calculated. The set of lumped reactions is called J_L, and introduced into the S matrix as a subset of the set J. Constraints are derived which ensure both the lumped reactions and their constituent reactions remained thermodynamically consistent, irrespective of the value of $\Delta_r G'^0$ of the unknown reactions.

For lumped reactions, $\Delta_r G'$ is calculated as described in (5.12), with (5.13) modified to allow $\Delta_r G'^0$ to vary freely. The transformed Gibbs free energy ($\Delta_r G'$) of the lumped reactions is calculated from $\Delta_r G'$ of their constituent reactions. To do this, parameters are defined such that $\alpha_{j,l} = 1$ if reaction j (one of the reactions with

unknown $\Delta_r G'^0$) combines in the forward direction to make up lumped reaction l, and defined $\alpha_{j,l} = -1$ if reaction j combines in the reverse direction. The transformed Gibbs free energy ($\Delta_r G'$) of the lumped reaction is then computed:

$$\Delta_r G'_l = \sum_j \alpha_{j,l} \Delta_r G'_j \qquad \forall l \in J_L \tag{5.14}$$

This constraint ensures that the thermodynamics of the lumped reaction and its constituents are internally consistent, irrespective of the value of $\Delta_r G'^0$ of the unknown reactions.

5.3.6: Linear Programming Approximation of TMFA (TMFA-LP)

Because TMFA is a mixed-integer program, its use may be impractical under certain conditions. For example, TMFA cannot be used as the inner problem in a bilevel optimization, and extending TMFA to large networks (as in gapfilling) is computationally intensive. As a consequence, during model refinement TMFA-LP, rather than TMFA is used, as described previously [79]. In this approach, extreme values for $\Delta_r G'$ are computed as given by (5.12), assuming $\Delta_r G'^0$ falls within the 95% confidence interval given by (5.13). If the predicted $\Delta_r G'$ range for a reaction was entirely negative, the reaction is assumed to proceed only in the forward direction (i.e., v_{min} = 0 mmol/gDW/day); if the predicted $\Delta_r G'$ range is entirely positive, the reaction is assumed to proceed only in the reverse direction (i.e., v_{max} = 0 mmol/gDW/day). Otherwise, the reaction is allowed proceed in both directions (i.e., v_{min} = -1000 mmol/gDW/day, v_{min} = +1000 mmol/gDW/day).

5.3.7: Parsimonious TMFA (pTMFA)

While FBA assumes selective pressure for the fastest growing strains, other selective forces may shape an organism's phenotype [208]. pFBA [122] is a constraint-based approach which assumes selective pressure not only for the fastest growing strains, but

also for those that require the lowest overall flux through the network (a proxy for minimizing the total enzyme mass required to sustain optimal growth through the network). pTMFA uses the same assumptions as pFBA while implementing the thermodynamic constraints of TMFA.

pTMFA was implemented as a two-stage optimization process. In Stage 1, TMFA is solved as described previously. In the second stage, the overall flux through the network is minimized, by decomposing each flux v into its forward and reverse components, v^+ and v^-:, and summing over the fluxes of all reactions in the network. To ensure that the pTMFA solution is also optimal with respect to growth, the growth rate is fixed to the Stage 1 solution. Aggregating the new constraints with those from TMFA gives pTMFA:

$$
\begin{aligned}
&\textbf{min} \quad \sum_j v_j^+ + v_j^- \\
&\textbf{s.t.} \quad \text{FBA constraints, (S2) and (S3)} && \forall j \in J \\
&\qquad \text{thermodynamic constraints (S4) to (S6)} && \forall j \in J \\
&\qquad v = v^+ - v^- && \forall j \in J \\
&\qquad v^+ \geq 0 && \forall j \in J \\
&\qquad v^- \geq 0 && \forall j \in J \\
&\qquad v_{BM} = \textbf{max} \quad v_{BM} \\
&\qquad\qquad \textbf{s.t.} \quad \text{FBA constraints, (S2) and (S3)} && \forall j \in J \\
&\qquad\qquad \text{thermodynamic constraints (S4) to (S6)} && \forall j \in J
\end{aligned}
\qquad \text{(pTMFA)}
$$

While FBA assumes selective pressure for the fastest growing strains, other selective forces may shape an organism's phenotype [208]. pFBA [122] is a constraint-based approach which assumes selective pressure not only for the fastest growing strains, but also for those that require the lowest overall flux through the network (a proxy for minimizing the total enzyme mass). pTMFA uses the same assumptions as pFBA while implementing the thermodynamic constraints of TMFA.

pTMFA was implemented as a two-stage optimization process. In Stage 1, TMFA is solved as described previously. In the second stage, the overall flux through

the network is minimized, by decomposing each flux v into its forward and reverse components, v^+ and v^-:, and summing over the fluxes of all reactions in the network:

$$\textbf{min} \quad \sum_j v_j^+ + v_j^- \tag{5.15}$$

$$\textbf{s.t.} \quad v = v^+ - v^- \tag{5.16}$$

$$v^+ \geq 0$$
$$v^- \geq 0 \tag{5.17}$$

To ensure that the pTMFA solution is also optimal with respect to growth, the growth rate is fixed to the Stage 1 solution:

$$\begin{array}{lll} v_{BM} = & \textbf{max} & v_{BM} \\ & \textbf{s.t.} & \text{FBA constraints, (5.9) and (5.10)} \quad \forall j \in J \\ & & \Delta_r G' \text{ constraints, (5.12) and (5.13)} \quad \forall j \in J \\ & & \text{consistency constraints, (5.11)} \quad \forall j \in J \end{array} \tag{5.18}$$

Aggregating the new constraints with those from TMFA gives pTMFA:

$$\begin{array}{ll} \textbf{min} & \sum_j v_j^+ + v_j^- \\ \textbf{s.t.} & \text{FBA constraints, (5.9) and (5.10)} \quad \forall j \in J \\ & \Delta_r G' \text{ constraints, (5.12) and (5.13)} \quad \forall j \in J \\ & \text{consistency constraints, (5.11)} \quad \forall j \in J \\ & \text{flux decomposition constraints, (5.16) and (5.17)} \quad \forall j \in J \\ & \text{optimal biomass given by TMFA, (5.18)} \end{array} \tag{pTMFA}$$

5.3.8: Community Formulation

For the coculture simulations, a formulation was developed to simulate growth in a continuous stirred-tank reactor (CSTR), taking into account the biomass concentrations X_n of each species n. Conceptually, the formulation attempts to minimize the species-weighted overall flux through the network, subject to TMFA constraints for each species, and a mass balance around the entire reactor:

$$\begin{array}{lll} \textbf{min} & \sum_n \left(X_n \sum_{j,n} \left(v_{j,n}^+ + v_{j,n}^- \right) \right) & \\ \textbf{s.t.} & \text{FBA constraints, (5.9) and (5.10)} \quad \forall j \in J & \forall n \in N \\ & \Delta_r G' \text{ constraints, (5.12) and (5.13)} \quad \forall j \in J & \forall n \in N \end{array} \tag{CSTR}$$

consistency constraints, (5.11) $\forall j \in J \qquad \forall n \in N$

reactor balance, (5.19)

$v_{BM,n} = D \qquad \forall n \in N$

As described in Results and illustrated in Figure 5.3B, both species exchange metabolites to a shared pool of metabolites i. Each metabolite i_n is exported to (or imported from) the shared pool via an exchange flux j_n at rate $v_{j,n}$. Likewise, each metabolite i has a net in (or out) flow from the reactor, denoted by rate F_j. Thus, each metabolite can flow in/out of the shared pool via $n+1$ reactions. Such tuples of reactions are indicated by the set $j_{n=1},..., j_{n=n}, j$, called J_{Shared}. The mass balance around the reactor can be then written:

$$\sum_n X_n v_{j,n} = F_j \qquad \forall (j_{n=1},...,j_{n=n},j) \in J_{Shared} \tag{5.19}$$

Realizing that the reactor operates at a fixed dilution rate (D) [213], the final constraint can be derived from a mass-balance around each species:

$$v_{BM,n} = D \qquad \forall n \in N \tag{5.20}$$

As presented, this formulation is general and can be applied to a community with any number of species.

The use of thermodynamic constraints necessitates the use of concentrations as variables in the coculture formulation. Additional constraints enforce this consistency of metabolite concentrations. A set of tuples ($i_{n=1},...,i_{n=n}, i$, called I_{Shared}) indicates those tuples of *in-silico* metabolites corresponding to the same physical metabolite (e.g., CO_2 has an extracellular version for each strain, and a version in the shared pool, for a total of three versions). A constraint forces each version of a metabolite to have the same concentration, thus ensuring that a concentration constraint on metabolite i in the shared pool constrains the metabolite concentration in each model as well:

$$\ln(x_{i,n=1}) = ... = \ln(x_{i,n=n}) = \ln(x_i) \qquad \forall (i_{n=1},...,i_{n=n},i) \in I_{Shared} \tag{5.21}$$

Failing to do so may result in inconsistent thermodynamic predictions arising from a single metabolite having multiple concentrations.

Finally, metabolites known to be exported to (or imported from) the shared space by each species via diffusion are identified. For each such metabolite, a pair (i_n i_n) containing the extra- and intracellular versions of that metabolite gets defined, as are sets $I_{n,Out}$ and $I_{n,In}$ containing those pairs which diffuse out and in, respectively. Constraints on these metabolite concentrations ensure consistency with diffusion:

$$\begin{aligned} \ln(x_{i,n}) &< \ln(x_{i,n}) \qquad \forall (i_n, i_n) \in I_{n,Out} \\ \ln(x_{i,n}) &> \ln(x_{i,n}) \qquad \forall (i_n, i_n) \in I_{n,In} \end{aligned} \tag{5.22}$$

The expanded coculture model formulation includes those constraints necessary for thermodynamic consistency across models.

$$\textbf{min} \quad \sum_n \left(X_n \sum_{j,n} \left(v^+_{j,n} + v^-_{j,n} \right) \right)$$

$$\begin{aligned} \textbf{s.t.} \quad & \text{FBA constraints, (5.9) and (5.10)} && \forall j \in J \ \forall n \in N \\ & \Delta_r G' \text{ constraints, (5.12) and (5.13)} && \forall j \in J \ \forall n \in N \\ & \text{consistency constraints, (5.11)} && \forall j \in J \ \forall n \in N \\ & \text{concentration consistency constraints, (5.21)} && \forall n \in N \\ & \text{diffusion constraints, (5.22)} && \forall n \in N \\ & \text{reactor balance, (5.19)} \\ & v_{BM,n} = D && \forall n \in N \end{aligned} \tag{CSTR}$$

This formulation is a mixed-integer non-linear program (MINLP): the consistency constraints (5.11) are integer constraints, and the reactor balance constraint (5.19) is nonlinear. To avoid solving the MINLP, the dilution rate D and biomass concentrations of X_n for all species were fixed, rendering the MINLP an easily-solved MIP.

To explore the behavior of the community under a variety of operating conditions, the reactor dilution rate and the relative ratio of *M. hungatei* to *S. fumaroxidans* were systematically changed, while allowing unlimited propionate uptake by the reactor. Simulations were performed at dilution rates between 0 to 0.135 days^{-1} at an interval of 0.005 days^{-1}. The ratio of *M. hungatei* to *S. fumaroxidans* was varied from 0 to 10, with simulations performed at intervals of 0.1. The data were plotted in Matlab (The MathWorks, Inc, Natick, Massachusetts, USA) with the Matlab function

interp being used to interpolate between simulated points. The maximal dilution rate of 0.135 $days^{-1}$ corresponds to the maximal growth rate of the slower-growing *S. fumaroxidans*.

5.3.9: Minimal Probabilistic Sets (MPS)

Under some conditions, thermodynamic constraints predict that a solution space contains undesired flexibility (e.g., the maximal predicted rate of ATP hydrolysis exceeds the experimental observation). A previous study of thermodynamic constraints used probability to qualitatively constrain reaction directions and reduce network flexibility [62]. This approach calculated the probability that a reaction's $\Delta_r G'^0$ was negative. If the probability was greater (less) than 70% (30%), the reaction was constrained to the forward (reverse) direction.

As described in Results, this probabilistic approach was applied to reduce network flexibility in *M. hungatei*, but was found to reduce network flexibility too much, and phenotypes which were previously correct become incorrect (e.g., the model no longer predicted growth). To obtain a balance between too much and not enough network flexibility, an optimization procedure was developed to identify the smallest number of probabilistic (qualitative reaction direction) constraints needed to correct one (or more) phenotypes, while preserving one (or more) other phenotypes. The required formulation, Minimal Probabilistic Sets (MPS), is defined as follows.

The sets $J_{probFwd}$ and $J_{probRev}$ consist of those reactions j in J which probability suggests should be constrained to the forward or reverse directions, respectively. The binary variables y and z take a value of 0 if a reaction is to be constrained to the forward or reverse direction, respectively, and take a value of 1 otherwise. The appropriate constraints are:

$$y_j = 0 \text{ or } 1 \qquad \forall j \in J_{probFwd} \tag{5.23}$$

$$z_j = 0 \text{ or } 1 \qquad \forall j \in J_{probRev} \tag{5.24}$$

$$y_j = z_j = 1 \text{ otherwise} \qquad \forall j \in J \tag{5.25}$$

$$y_j = z_j \qquad \forall j \in J \tag{5.26}$$

The objective is to maximize the network flexibility (as measured by the number of unconstrained reactions), while maintaining consistency with each n of N phenotypes:

$$\textbf{max} \quad \sum_{j,n} y_{j,n} + \sum_{j,n} z_{j,n} \tag{5.27}$$

$$\textbf{s.t.} \quad \text{phenotype constraints} \qquad \forall n \in N \tag{5.28}$$

$$\text{computed cellular phenotype} \qquad \forall n \in N \tag{5.29}$$

$$y_{j,n} = y_{j,\hat{n}} \qquad \forall (n,\hat{n}) \in N \tag{5.30}$$

$$z_{j,n} = z_{j,\hat{n}} \qquad \forall (n,\hat{n}) \in N$$

Any reactions which the optimization identifies as requiring a qualitative reaction direction constraint can be easily identified, as the y or z variable associated with that reaction will have a value of 0. The desired phenotypes (5.28) can be imposed as a variety of constraints, e.g.,

$$\begin{aligned} v_{BM} &\geq \varepsilon \\ v_{ATP} &= \gamma \end{aligned} \tag{5.31}$$

and may enforce phenotypes to be corrected (e.g., ATP gain) or maintained (e.g., cellular growth). Individual cellular phenotypes (5.29) are computed by an inner problem, which maximizes or minimizes some cellular objective, $v_{obj,}$ subject to mass-balance and qualitative reaction direction constraints under a particular media condition:

$$\textbf{max} \quad v_{obj} \tag{5.32}$$

$$\textbf{s.t.} \quad S \cdot v = 0 \tag{5.33}$$

$$v_{\min} \leq v_j \leq v_{\max} \qquad \forall j \in J \tag{5.34}$$

$$\text{media constraints} \tag{5.35}$$

$$v_j \geq 0 \qquad \forall j \in y_j = 0 \tag{5.36}$$

$$v_j \leq 0 \qquad \forall j \in z_j = 0 \tag{5.37}$$

The inner problem enforces the qualitative reaction direction (probabilistic) constraints (5.36) and (5.37) identified by the outer problem. Because thermodynamic constraints

cannot be enforced directly in the inner problem, reactions are constrained using their TMFA-LP directions instead.

The final outer problem constraint (5.30) ensures that the same qualitative reaction direction constraints are enforced on each inner problem. Aggregating all of the above constraints gives the final formulation for MPS:

$$\begin{aligned}
\textbf{max} \quad & \sum_{j,n} y_{j,n} + \sum_{j,n} z_{j,n} \\
\textbf{s.t.} \quad & \text{computed cellular phenotype(s), (5.32) to (5.37)} \quad \forall n \in N \\
& \text{phenotype constraint(s), (5.31)} \quad \forall n \in N \\
& \text{allowed probabilistic constraints, (5.23) to (5.26)} \quad \forall n \in N \\
& \text{consistency of probabilistic constraints, (5.29)} \quad \forall n \in N
\end{aligned} \tag{MPS}$$

For purposes of implementation, this bilevel problem is converted to a single-level problem via well-established techniques which rely on duality theory [59,78].

The formulation enables any number of phenotypes to be predicted by the inner problems, provided that a constraint for each is given in the outer problem. MPS was able to correct a number of phenotypes in the *i*Mhu273 and *i*Sfu648 models.

5.3.9.1: MPS for Validating *i*Mhu273: ATP Gain

Experimental evidence suggests that *M. hungatei* is able to generate 0.5 mol ATP per mol of CO_2 converted to CH_4 [235]. A methanogenesis pathway which produces this ATP yield could be identified, but the *i*Mhu273 model predicted other, higher-yielding, ATP-generating mechanisms outside the methanogenesis pathway. MPS was used to identify qualitative reaction direction constraints which would eliminate these extra mechanisms, while simultaneously ensuring biomass growth:

$$\begin{aligned}
\textbf{max} \quad & \sum_{j} y_{j} + \sum_{j} z_{j} \\
\textbf{s.t.} \quad & \textbf{\textit{max}} \quad \text{ATP gain} \\
& \textbf{\textit{s.t.}} \quad \text{FBA constraints, (5.9) and (5.10)} \quad \forall j \in J \\
& \qquad \text{carbon uptake via } CO_2 \text{ and acetate} \\
& \qquad \text{enforce probabilistic constraints, (5.36) and (5.37)} \quad \forall j \in J
\end{aligned} \tag{MPS}$$

ATP gain = 0.5
$v_{BM} \geq 0.01$
allowed probabilistic constraints, (5.23) to (5.26) $\forall n \in N$
consistency of probabilistic constraints, (5.29) $\forall n \in N$

5.3.9.2: MPS for Validating *i*Mhu273: Biomass Growth

It has been observed that biomass and energy generation are independent and non-interacting in *M. hungatei*, with methanogenesis via CO_2 being the sole source of ATP, and acetate being the sole source of biomass components [51]. In contrast to experimental evidence, the *i*Mhu273 model predicts that *M. hungatei* can produce some biomass from CO_2 alone. MPS was used to resolve the discrepancy with two inner problems: the first inner problem tries to enforce a no growth phenotype on CO_2 alone, while the second inner problem tries to maintain growth on acetate alone:

max $\sum_{j,n} y_{j,n} + \sum_{j,n} z_{j,n}$
s.t. ***max*** v_{BM}
s.t. FBA constraints, (5.9) and (5.10) $\forall j \in J$
carbon uptake via CO_2
enforce probabilistic constraints, (5.36) and (5.37) $\forall j \in J$
$v_{BM} = 0$
max v_{BM}
s.t. FBA constraints, (5.9) and (5.10) $\forall j \in J$
carbon uptake via acetate
enforce probabilistic constraints, (5.36) and (5.37) $\forall j \in J$
$v_{BM} \geq 0.01$
allowed probabilistic constraints, (5.23) to (5.26) $\forall n \in N$
consistency of probabilistic constraints, (5.29) $\forall n \in N$ (MPS)

The problem proved infeasible, meaning that with the current *i*Mhu273 model topology ATP generation and biomass growth cannot be fully uncoupled.

5.3.9.3: MPS for Validating *i*Sfu648: Closed-Network ATP Cycles

Thermodynamic constraints dictate that solutions containing closed cycles (e.g., A → B → C → A) should be infeasible. In the absence of complete thermodynamic information,

such cycles remain possible. MPS was used to identify qualitative reaction direction constraints which would eliminate closed-cycles which synthesize ATP, while simultaneously ensuring biomass growth:

$$
\begin{aligned}
&\textbf{max} \quad \sum_{j,n} y_{j,n} + \sum_{j,n} z_{j,n} \\
&\textbf{s.t.} \quad \textbf{\textit{max}} \quad v_{ATPM} \\
&\qquad \textbf{\textit{s.t.}} \quad \text{FBA constraints, (5.9) and (5.10)} \qquad \forall j \in J \\
&\qquad\qquad \text{no uptake fluxes} \\
&\qquad\qquad \text{enforce probabilistic constraints, (5.36) and (5.37)} \; \forall j \in J \\
&\qquad v_{ATPM} = 0 \\
&\qquad \textbf{\textit{max}} \quad v_{BM} \\
&\qquad \textbf{\textit{s.t.}} \quad \text{FBA constraints, (5.9) and (5.10)} \qquad \forall j \in J \\
&\qquad\qquad \text{growth in monoculture (or coculture)} \\
&\qquad\qquad \text{enforce probabilistic constraints, (5.36) and (5.37)} \; \forall j \in J \\
&\qquad v_{BM} \geq 0.01 \\
&\qquad \text{allowed probabilistic constraints, (5.23) to (5.26)} \qquad \forall n \in N \\
&\qquad \text{consistency of probabilistic constraints, (5.29)} \qquad \forall n \in N
\end{aligned}
\tag{MPS}
$$

MPS was used also to identify a single set qualitative reaction direction constraints which would ensure growth in both monoculture and coculture simultaneously (formulation not shown). That problem was infeasible, indicating that distinct sets of qualitative reaction direction (probabilistic) constraints are required under the two conditions.

5.3.9.4: MPS for Validating *i*Sfu648: ATP Gain

In-silico simulations identified the theoretical maximum ATP yield for each growth mode of *S. fumaroxidans*. However, the *i*Sfu648 model enabled other mechanisms of ATP generation with a higher yield. MPS was used to identify qualitative reaction direction constraints which would eliminate these extra mechanisms, while simultaneously ensuring biomass growth:

$$
\begin{aligned}
&\textbf{max} \quad \sum_{j} y_{j} + \sum_{j} z_{j} \\
&\textbf{s.t.} \quad \textbf{\textit{max}} \quad \text{ATP gain}
\end{aligned}
\tag{MPS}
$$

s.t. FBA constraints, (5.9) and (5.10) $\forall j \in J$
appropriate growth mode uptake constraints
enforce probabilistic constraints, (5.36) and (5.37) $\forall j \in J$
maximum ATP gain for that mode
$v_{BM} \geq 0.01$
allowed probabilistic constraints, (5.23) to (5.26) $\forall n \in N$
consistency of probabilistic constraints, (5.29) $\forall n \in N$

5.3.10: Simulation Conditions

Simulations were performed for both individual models (*i*Sfu648 and *i*Mhu273) and the coculture model. All simulations were performed using CPLEX 12 (IBM, Armonk, NY) accessed via the General Algebraic Modeling System, Version 23.9.5 (GAMS, GAMS Development Corporation, Washington, DC).

Chapter 6: Conclusions

Some material in this chapter has been adapted from:

> Hamilton JJ, Reed JL (2014) Software platforms to facilitate reconstructing genome-scale metabolic networks. Environ Microbiol 16(1): 49–59.

This dissertation is focused on developing novel methods to advance the field of constraint-based modeling. We first developed CONGA, an algorithm to reveal the impact of structural network differences on predicted functional states (Chapter 3 and [78]). Using CONGA, we can not only identify but also analyze the phenotypic impact of genetic differences between organisms. With TMFA, we introduced thermodynamic constraints into flux-balance approaches, and evaluated the impact of these constraints on metabolic flux distributions and cellular growth rates (Chapter 4 and [79]). Finally, we extended TMFA to model the syntrophic association between two micro-organisms (Chapter 5). In that work, we developed a community-level optimization problem which we used to predict community stoichiometry at different dilution rates and biomass abundance ratios. Despite these advances, there remains considerable room for improvement in the areas of thermodynamic and community models. This chapter discusses remaining challenges in these areas (Sections 6.2 and 6.3), describes the use of CONGA by other investigators (Section 6.1), and concludes with some final thoughts on the future of constraint-based modeling (Sections 6.4 and 6.5).

6.1: CONGA: A New Tool for Metabolic Comparisons

We developed a mathematical programming approach, CONGA, to identify the functional differences between networks by comparing network reconstructions aligned at the gene level. Our gene-centric approach allows for the rapid identification of

functional differences between networks which can be traced back to the presence or absence of particular genes or reactions (*structural differences*) in one network or the other.

Working with collaborators, we have also constructed and analyzed GENREs for two strains of *Lactobacillus casei*, *L. casei* 12A and *L. casei* ATCC 334 (manuscript in preparation). Our reconstructions revealed that *L. casei* 12A can utilize a wider variety of carbon sources than *L. casei* ATCC 334. However, when analyzing carbon sources common to both strains, CONGA found that both strains were equally robust. We also performed simulations with different nitrogen sources (amino acids), and we were able to identify isozymes unique to each strain which influenced robustness on different nitrogen sources.

CONGA has also been used by other scientists to compare GENREs of four *Shewanella* strains, as well as a core genome GENRE containing the genes shared by all 21 sequenced *Shewanella* strains [161]. The authors used CONGA to identify functional differences between all pairs of strains, as well as between each strain and the core genome. They found the *Shewanella* core genome to be least robust (e.g., CONGA identified the largest number of uniquely lethal gene deletions), due the presence of unique isozymes in individual *Shewanella* strains. When examining just the GENREs of the four strains, the authors found that relative strain robustness varied with environmental condition. The authors also identified a new type of metabolic difference, a *biomass difference*, in which differences in biomass composition give rise to differences in gene essentiality.

Thus far, CONGA has been used to identify gene deletions pointing to functional metabolic differences. However, other network perturbations may be equally effective indicators of network differences. Robust algorithms for identifying other types of perturbations have been developed [106,130,171,173,174,188,256] and can be easily

incorporated into CONGA. Furthermore, gene and reaction differences may not be the only source of differences between models, For example, CONGA could be used to examine differences at the level of constraints, by comparing identical models with otherwise different constraints based on gene expression, regulation, or thermodynamics. Finally, we envision our approach being used to examine cellular behavior under different environmental conditions, or to compare evolved and un-evolved cellular phenotypes. Ultimately, a comparative approach such as CONGA will enable rapid evaluation of the influence of network and model differences on predicted functional states.

6.2: Remaining Challenges in Thermodynamic Models

Thermodynamics-based metabolic flux analysis (TMFA) extends flux balance analysis with thermodynamic constraints, to ensure that all reactions operate in thermodynamically feasible directions. Group contribution methods (GCMs) are used to provide estimates and uncertainties of Gibbs free energies. Our studies incorporating thermodynamic constraints into genome-scale models revealed three important areas for further research.

First, group contribution methods (GCMs) enable computation of Gibbs free energies and uncertainties at three levels: the reaction ($\Delta_r G'^0$), the metabolites which participate in the reactions ($\Delta_f G'^0$), or the groups which make up the metabolites ($\Delta_{gr} G'^0$). There are difficulties associated with using both groups and reactions as a basis for thermodynamic constraints, and it remains unclear which thermodynamic basis best balances those difficulties. For example, in our study of syntrophic association, we found that performing calculations with reactions as the basis for thermodynamic constraints resulted in too much network flexibility. In particular, the model-predicted feasible $\Delta_r G'^0$ range through a pathway considerably exceeded the GCM-predicted

$\Delta_r G'^0$ range of the pathway's overall stoichiometry. This discrepancy arises because using $\Delta_r G'^0$ directly fails to account for the interdependency between $\Delta_r G'^0$s for all reactions sharing a given metabolite. Using groups or metabolites as a basis for thermodynamic calculations enables the model to capture these interdependencies by explicitly account for the relationship between $\Delta_r G'^0$ and $\Delta_f G'^0$ (i.e., $\Delta_r G'^0_j = \sum_i S_{i,j} \Delta_f G'^0_i$). However, when analyzing our thermodynamic model of *E. coli*, we found that using metabolites or groups as a basis for thermodynamic constraints resulted in very long run-times, especially when accounting for uncertainty in the estimates of $\Delta_f G'^0$ and $\Delta_{gr} G'^0$

Second, part of the computational difficulty associated with thermodynamic constraints may arise from the weakness of the LP-relaxation of TMFA. A recent Master's thesis [146] studied a special case of TMFA, in which $\Delta_r G'$ is not constrained by group contribution estimates:

$$\begin{array}{lll} \textbf{max} & v_{BM} & \\ \textbf{s.t.} & S \cdot v = 0 & \\ & v_{min,j} \le v_j \le v_{max,j} & \forall j \in J \\ & (1-\delta_j) v_{min,j} \le v_j \le \delta_j v_{max,j} & \forall j \in J \\ & -M\delta_j + \varepsilon \le \Delta_r G'_j \le M(1-\delta_j) - \varepsilon & \forall j \in J \end{array} \qquad \text{(TMFA)}$$

where the last two constraints enforce thermodynamic feasibility ($v \cdot \Delta_r G' < 0$). The author showed that the relaxation corresponds to the thermodynamically unconstrained problem [146]. As a result, branch-and-bound solvers cannot use the relaxation to say anything useful about the original MIP. The author also developed and implemented cutting plane algorithms for this special case [146]. However, we were unable to replicate the performance enhancements when imposing additional constraints on $\Delta_r G'$ and $\Delta_r G'^0$ (i.e., $\Delta_r G' = \Delta_r G'^0 + RT \sum_i S_{i,j} \ln(x_i)$ and $\Delta_r G'^0$ constrained by its confidence interval).

Third, the utility of thermodynamic constraints can also be improved through additional thermodynamic measurements. The latest GCM [156] relies heavily on a collection of measured thermodynamic parameters published by the National Institute of Standards and Technology (NIST) [74]. However, the NIST database lacks information on important biological compounds (notably cofactors containing metal ions), so $\Delta_f G'^0$ of these metabolites cannot be estimated. This limitation proved to be a major obstacle in our studies of syntrophic association, as we were unable to say anything about the energetics of many important electron transfer reactions. The addition of thermodynamic measurements for metal-containing cofactors (e.g., in the form of reduction potentials [184]) would expand the coverage of GCMs to many biologically important reactions.

Nevertheless, TMFA remains a useful tool with many applications. TMFA ensures that all reactions operate in thermodynamically feasible directions, and can predict reaction directions in the absence of prior knowledge. Thus, we envision TMFA complementing other approaches [54,62,83,114] which are used to calculate thermodynamically feasible reaction directions for new genome-scale models. TMFA also promises to be a useful tool for metabolic engineering applications, by identifying thermodynamic bottlenecks in engineered pathways [211] or by pinpointing those reactions whose reversible operation enables new routes for chemical synthesis [43]. Finally, thermodynamic constraints could serve as a filter to test the feasibility of alternative routes for biochemical production.

6.3: Towards a Theory of Community Systems Biology

Constraint-based analysis has been applied to biological reaction networks for over 20 years [241]. In the past few years, the field of constraint-based modeling has matured from a descriptive to a predictive one, and predictions from genome-scale models are

being used in a prospective manner to drive translational applications [32]. In light of these successes, researchers have begun investigating new applications of constraint-based modeling. In particular, constraint-based methods have been deployed to investigate simple microbial communities, as described in Section 2.3. A recent road map for the development of community systems biology [260] suggests that constraint-based, bottom-up approaches can complement traditional top down ('meta-omics') approaches [259] by providing a mechanistic understanding of microbial interactions. However, a number of hurdles must be overcome before this vision becomes a reality.

The foremost problem is accurate determination of a community's metabolic content. Traditionally, community composition has been determined through identification of phylogenetic marker genes (e.g., 16S rRNA, [140]). However, these genes cannot be used to infer metabolic capabilities, as the genome content of organisms with identical 16S rRNA sequences can be highly divergent [97]. In addition, the most abundant organisms might not be the most important, as numerically under-represented microbes can still carry out important roles [218]. Fortunately, new single-cell technologies [25] will enable whole-genome sequencing of individual community members, for which GENREs can then be constructed.

However, high-quality GENREs require high-quality gene annotations. Newly sequenced genomes are typically annotated automatically [193], and gene functions are predicted on the basis of sequence similarity to reference genes. However, these annotations become less reliable at larger phylogenetic distances from the reference organism (often *E. coli*) [195,240]. Fortunately, a number of both experimental [45,182] and computational [179,232] approaches have been recently developed to identify and validate new annotations. However, these techniques can only be applied to organisms which can be cultured and genetically manipulated, and focusing only on these organisms may result in a biased picture of the community.

In order for community systems biology models to make meaningful predictions, individual species models must be integrated into a community model in a biologically relevant manner. This integration will require an understanding of the objectives and constraints governing the behavior of each community member, and potentially the entire community. As we saw with our model of a syntrophic association, identifying the proper constraints and objectives may require extensive experimental characterization of the community, and the necessary constraints may not arise as a result of stoichiometry alone.

Unfortunately, the mechanisms of metabolite exchange in communities remain poorly understood [175]. In cases where mechanisms are known, integrating meta-omics data with constraint-based community models may help elucidate the biological principles underpinning the association [149]. In addition, tools such as microbial imaging mass spectrometry [145] can identify metabolites which are being exchanged, and constraint-based models of individual microbes could be use to propose mechanisms for metabolite production and consumption. Lastly, we note that microbes within a community may communicate indirectly (e.g., indirect nutrient exchange via diffusion) or non-metabolically (e.g., quorum sensing). Modeling of these types of communities may require novel extensions to traditional constraint-based modeling [84].

Despite these significant hurdles, community systems biology has a promising future. The success of constraint-based modeling as it applies to individual microbes has been driven in large part by advances in both molecular and computational biology, and continuing advances in these fields will benefit the development of community systems biology as well.

6.4: Software for Genome-Scale Network Reconstruction

This dissertation required the development and refinement of genome-scale network reconstructions for three distinct organisms using a variety of software platforms. Our experience with these software platforms inspired us to write a review which introduces these tools to non-specialists [80].

While we found that these software platforms greatly facilitate the reconstruction process, we also found that users must be aware of each platform's potential pitfalls and should actively evaluate software inputs and outputs to ensure a high quality reconstruction. The utility of these software platforms would be enhanced through features which actively guide users through the reconstruction process.

For example, these platforms could provide better support for mass- and charge-balancing reactions. Many reactions involve generic metabolites which must be replaced before a reaction can be balanced. Reaction balancing could be facilitated by automatic flagging of generic metabolites and reactions and providing users the opportunity to replace them. It is also important that metabolites be represented in their properly charged state. For example, the cheminformatics software MarvinBeans (Marvin pKa plug-in, ChemAxon, Budapest, Hungary) can be used to identify the proper charge for each metabolite. Users could then be notified of unbalanced reactions and given the opportunity to balance them. It is also important that reactions be assigned their proper direction, to prevent stoichiometrically balanced cycles and other unrealistic behaviors (such as free ATP production). Reaction directions can be predicted on the basis of thermodynamics, and one such method has been implemented in the COBRA Toolbox [156].

Automatic reconstruction platforms could also be improved by allowing users to evaluate both the inputs and outputs of reconstruction steps (such as growth media and gap-filled reactions), and warning users when automated reconstruction steps encounter

difficulties (such as when balancing reactions). These warnings would alert users to potential inaccuracies in their model, which they could later examine and correct by hand.

Finally, the scientific community would benefit from making network reconstruction tools more accessible to the non-specialist. A powerful interactive visual interface for developing reconstructions would provide a strong incentive for the broader community to become engaged with the reconstruction process.

6.5: The Future of Constraint-Based Modeling

Genome-scale models have been used to support experimental efforts in a variety of areas, and constraint-based methods are beginning to drive translational research [32]. However, while constraint-based methods can rapidly generate novel phenotypes through sophisticated *in-silico* manipulations (e.g., controlled up- and down-regulations), performing the manipulations in the laboratory remains time-consuming. Fortunately, a number of experimental advances promise the rapid generation of complex genetic inventions, including:

- a variety of mutagenesis techniques, for the construction of large populations of knockout mutants [71,76,118,160,162,163];
- genome engineering techniques which enable targeted manipulations and the generation of genetic diversity [39,52,98,104,244,246];
- large-scale assembly methods, for the *de novo* assembly of novel genetic constructs [7,73,111];
- and pathway optimization techniques, for the rapid tuning of gene expression levels [197,229,245].

When combined with ongoing efforts to develop algorithms for experimental design and hypothesis testing [232,233], these experimental advances should

substantially accelerate the model refinement and strain design processes, thereby facilitating the rapid construction of novel bacterial phenotypes.

Further adoption of constraint-based modeling techniques can also be facilitated by the development of user-friendly environments for model reconstruction and analysis. Recently, a number of such software tools have been published [1,80,89,102,119,227], and while not perfect (see Section 6.4), they do enable non-specialists to develop GENREs and perform some constraint-based analyses. In particular, the Pathway Tools environment [102,119] stands out for providing a comprehensive software platform with support for all stages of the reconstruction process. The continued development of Pathway Tools and other software platforms will be crucial in ensuring the widespread adoption of constraint-based modeling by specialists and non-specialists alike.

As a consequence of these developments, the most exciting applications of constraint-based modeling lie ahead. New technologies for making and testing novel predictions will surely expand the scope of biological discoveries achievable by constraint-based modeling.

References

1. Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, et al. (2013) The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for Penicillium chrysogenum. PLoS Comput Biol 9(3): e1002980.

2. Alberty RA (2003) Thermodynamics of Biochemical Reactions. Hoboken, NJ: John Wiley & Sons.

3. Alper HS, Jin Y-S, Moxley JF, Stephanopoulos G (2005) Identifying gene targets for the metabolic engineering of lycopene biosynthesis in Escherichia coli. Metab Eng 7(3): 155–164.

4. Altschul SF, Gish W, Miller W, Myers E, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3): 403–410.

5. Andersen KB, von Meyenburg K (1977) Charges of nicotinamide adenine nucleotides and adenylate energy charge as regulatory parameters of the metabolism in Escherichia coli. J Biol Chem 252(12): 4151–4156.

6. Anderson IJ, Ulrich LE, Lupa B, Susanti D, Porat I, et al. (2009) Genomic characterization of methanomicrobiales reveals three classes of methanogens. PLoS One 4(6): e5797.

7. Annaluru N, Muller H, Mitchell LA, Ramalingam S, Stracquadanio G, et al. (2014) Total Synthesis of a Functional Designer Eukaryotic Chromosome. Science 344(6179): 55–58.

8. Archer CT, Kim JF, Jeong H, Park JH, Vickers CE, et al. (2011) The genome sequence of E. coli W (ATCC 9637): comparative genome analysis and an improved genome-scale reconstruction of E. coli. BMC Genomics 12: 9.

9. Arora P, Goyal A, Natarajan VT, Rajakumara E, Verma P, et al. (2009) Mechanistic and functional insights into fatty acid activation in Mycobacterium tuberculosis. Nat Chem Biol 5(3): 166–173.

10. Atsumi S, Higashide W, Liao JC (2009) Direct photosynthetic recycling of carbon dioxide to isobutyraldehyde. Nat Biotechnol 27(12): 1177–1180.

11. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, et al. (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol 2: 2006.0008.

12. Bar-Even A, Noor E, Flamholz A, Buescher JM, Milo R (2011) Hydrophobicity and Charge Shape Cellular Metabolite Concentrations. PLoS Comput Biol 7(10): e1002166.

13. Barua D, Kim J, Reed JL (2010) An automated phenotype-driven approach (GeneForce) for refining metabolic and regulatory models. PLoS Comput Biol 6(10): e1000970.

14. Beard DA, Liang S, Qian H (2002) Energy balance for analysis of complex metabolic networks. Biophys J 83(1): 79–86.

15. Beard DA, Babson E, Curtis E, Qian H (2004) Thermodynamic constraints for biochemical networks. J Theor Biol 228(3): 327–333.

16. Beard DA, Qian H (2005) Thermodynamic-based computational profiling of cellular regulatory control in hepatocyte metabolism. Am J Physiol Endocrinol Metab 288(3): E633–E644.

17. Becker SA, Palsson BØ (2005) Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation. BMC Microbiol 5: 8.

18. Benedict MN, Gonnerman MC, Metcalf WW, Price ND (2012) Genome-scale metabolic reconstruction and hypothesis testing in the methanogenic archaeon Methanosarcina acetivorans C2A. J Bacteriol 194(4): 855–865.

19. Bennett BD, Kimball EH, Gao M, Osterhout RE, Van Dien SJ, et al. (2009) Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. Nat Chem Biol 5(8): 593–599.

20. Benning C (1998) Biosynthesis and Function of the Sulfolipid Sulfoquinovosyl Diacylglycerol. Annu Rev Plant Physiol Plant Mol Biol 49: 53–75.

21. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2010) GenBank. Nucleic Acids Res 38(S1): D46–D51.

22. Berg CM, Wang M, Vartak NB, Liu L (1988) Acquisition of new metabolic capabilities: multicopy suppression by cloned transaminase genes in Escherichia coli K-12. Gene 65(2): 195–202.

23. Billheimer JT, Carnevale HN, Leisinger T, Eckhardt T, Jones EE (1976) Ornithine delta-transaminase activity in Escherichia coli: its identity with acetylornithine delta-transaminase. J Bacteriol 127(3): 1315–1323.

24. Billheimer JT, Shen MY, Carnevale HN, Horton HR, Jones EE (1979) Isolation and characterization of acetylornithine delta-transaminase of wild-type Escherichia coli W. Comparison with arginine-inducible acetylornithine delta-transaminase. Arch Biochem Biophys 195(2): 401–413.

25. Blainey PC (2013) The future is now: single-cell genomics of bacteria and archaea. FEMS Microbiol Rev 37(3): 407–427.

26. Blanchard JS (1996) Molecular mechanisms of drug resistance in Mycobacterium tuberculosis. Annu Rev Biochem 65: 215–239.

27. Blazier AS, Papin JA (2012) Integration of expression data in genome-scale metabolic network reconstructions. Front Physiol 3: 299.

28. De Bok FAM, Luijten MLGC, Stams AJM (2002) Biochemical Evidence for Formate Transfer in Syntrophic Propionate-Oxidizing Cocultures of Syntrophobacter fumaroxidans and Methanospirillum hungatei. Appl Environ Microbiol 68(9): 4247–4252.

29. De Bok FAM, Roze EHA, Stams AJM (2002) Hydrogenases and formate dehydrogenases of Syntrophobacter fumaroxidans. Antonie Van Leeuwenhoek 81(1-4): 283–291.

30. De Bok FAM, Hagedoorn P-L, Silva PJ, Hagen WR, Schiltz E, et al. (2003) Two W-containing formate dehydrogenases (CO2-reductases) involved in syntrophic propionate oxidation by Syntrophobacter fumaroxidans. Eur J Biochem 270(11): 2934–2942.

31. Bordbar A, Lewis NE, Schellenberger J, Palsson BØ, Jamshidi N (2010) Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions. Mol Syst Biol 6: 422.

32. Bordbar A, Monk JM, King ZA, Palsson BØ (2014) Constraint-based models predict metabolic and associated cellular functions. Nat Rev Genet 15(2): 107–120.

33. Bumann D (2008) Has nature already identified all useful antibacterial targets? Curr Opin Microbiol 11(5): 387–392.

34. Burgard AP, Vaidyaraman S, Maranas CD (2001) Minimal reaction sets for Escherichia coli metabolism under different growth requirements and uptake environments. Biotechnol Prog 17(5): 791–797.

35. Burgard AP, Maranas CD (2003) Optimization-based framework for inferring and testing hypothesized metabolic objective functions. Biotechnol Bioeng 82(6): 670–677.

36. Burgard AP, Pharkya P, Maranas CD (2003) OptKnock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol Bioeng 84(6): 647–657.

37. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, et al. (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 38(S1): D473–D479.

38. Chavali AK, D'Auria KM, Hewlett EL, Pearson RD, Papin JA (2012) A metabolic network approach for the identification and prioritization of antimicrobial drug targets. Trends Microbiol 20(3): 113–123.

39. Cong L, Ran FA, Cox D, Lin S, Barretto R, et al. (2013) Multiplex genome engineering using CRISPR/Cas systems. Science 339(6121): 819–823.

40. Copeland WB, Bartley BA, Chandran D, Galdzicki M, Kim KH, et al. (2012) Computational tools for metabolic engineering. Metab Eng 14(3): 270–280.

41. Corporation GD (2014) GAMS: A User's Guide. 2014th ed. Washington, DC: GAMS Development Corporation.

42. Covert MW, Knight EM, Reed JL, Herrgård MJ, Palsson BØ (2004) Integrating high-throughput and computational data elucidates bacterial networks. Nature 429(6987): 92–96.

43. Dellomonaco C, Clomburg JM, Miller EN, Gonzalez R (2011) Engineered reversal of the β-oxidation cycle for the synthesis of fuels and chemicals. Nature 476(7360): 355–359.

44. Demir E, Cary MP, Paley SM, Fukuda K, Lemer C, et al. (2010) The BioPAX community standard for pathway data sharing. Nat Biotechnol 28(9): 935–942.

45. Deutschbauer AM, Price MN, Wetmore KM, Shao W, Baumohl JK, et al. (2011) Evidence-based annotation of gene function in Shewanella oneidensis MR-1 using genome-wide fitness profiling across 121 conditions. PLoS Genet 7(11): e1002385.

46. Dong X, Plugge CM, Stams AJM (1994) Anaerobic degradation of propionate by a mesophilic acetogenic bacterium in coculture and triculture with different methanogens. Appl Environ Microbiol 60(8): 2834–2838.

47. Dong X, Stams AJM (1995) Evidence for H2 and formate formation during syntrophic butyrate and propionate degradation. Anaerobe 1(1): 35–39.

48. Dreyfuss JM, Zucker JD, Hood HM, Ocasio LR, Sachs MS, et al. (2013) Reconstruction and Validation of a Genome-Scale Metabolic Model for the Filamentous Fungus Neurospora crassa Using FARM. PLoS Comput Biol 9(7): e1003126.

49. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, et al. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci 104(6): 1777–1782.

50. Edwards JS, Palsson BØ (2000) The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci 97(10): 5528–5533.

51. Ekiel I, Smith ICP, Sprott GD (1983) Biosynthetic pathways in Methanospirillum hungatei as determined by 13C nuclear magnetic resonance. J Bacteriol 156(1): 316–326.

52. Esvelt KM, Wang HH (2013) Genome-scale engineering for systems and synthetic biology. Mol Syst Biol 9: 641.

53. Feist AM, Scholten JCM, Palsson BØ, Brockman FJ, Ideker T (2006) Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri. Mol Syst Biol 2: 2006.0004.

54. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, et al. (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 3: 121.

55. Feist AM, Palsson BØ (2008) The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. Nat Biotechnol 26(6): 659–667.

56. Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ (2009) Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol 7(2): 129–143.

57. Feist AM, Zielinski DC, Orth JD, Schellenberger J, Herrgård MJ, et al. (2010) Model-driven evaluation of the production potential for growth-coupled products of Escherichia coli. Metab Eng 12(3): 173–186.

58. Feist AM, Palsson BØ (2010) The biomass objective function. Curr Opin Microbiol 13(3): 344–349.

59. Ferris MC, Mangasarian OL, Wright SJ (2008) Linear Programming with Matlab. Phildelphia, PA: The Society for Industrial and Applied Mathematics and the Mathematical Programming Society.

60. Ferry JG (1999) Enzymology of one-carbon metabolism in methanogenic pathways. FEMS Microbiol Rev 23(1): 13–38.

61. Finley SD, Broadbelt LJ, Hatzimanikatis V (2009) Thermodynamic analysis of biodegradation pathways. Biotechnol Bioeng 103(3): 532–541.

62. Fleming RMT, Thiele I, Nasheuer HP (2009) Quantitative assignment of reaction directionality in constraint-based models of metabolism: application to Escherichia coli. Biophys Chem 145(2-3): 47–56.

63. Fong SS, Burgard AP, Herring CD, Knight EM, Blattner FR, et al. (2005) In silico design and adaptive evolution of Escherichia coli for production of lactic acid. Biotechnol Bioeng 91(5): 643–648.

64. Fourer R, Gay DM, Kernighan BW (2002) AMPL: A Modeling Language for Mathematical Programming. 2nd ed. Boston, MA: Cengage Learning.

65. Fredrickson JK, Romine MF, Beliaev AS, Auchtung JM, Driscoll ME, et al. (2008) Towards environmental systems biology of Shewanella. Nat Rev Microbiol 6(8): 592–603.

66. Freilich S, Zarecki R, Eilam O, Segal ES, Henry CS, et al. (2011) Competitive and cooperative metabolic interactions in bacterial communities. Nat Commun 2: 589.

67. French JB, Cen Y, Vrablik TL, Xu P, Allen E, et al. (2010) Characterization of nicotinamidases: steady state kinetic parameters, classwide inhibition by nicotinaldehydes, and catalytic mechanism. Biochemistry 49(49): 10421–10439.

68. Fu P (2009) Genome-scale modeling of Synechocystis sp. PCC 6803 and prediction of pathway insertion. J Chem Technol Biotechnol 84(4): 473–483.

69. Garcia-Albornoz MA, Nielsen J (2013) Application of Genome-Scale Metabolic Models in Metabolic Engineering. Ind Biotechnol 9(4): 203–214.

70. Garg S, Yang L, Mahadevan R (2010) Thermodynamic analysis of regulation in metabolic networks using constraint-based modeling. BMC Res Notes 3: 125.

71. Gawronski JD, Wong SMS, Giannoukos G, Ward D V, Akerley BJ (2009) Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for Haemophilus genes required in the lung. Proc Natl Acad Sci 106(38): 16422–16427.

72. Gelfand DH, Steinberg RA (1977) Escherichia coli mutants deficient in the aspartate and aromatic amino acid aminotransferases. J Bacteriol 130(1): 429–440.

73. Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang R-Y, et al. (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. Science 329(5987): 52–56.

74. Goldberg RN, Tewari YB, Bhat TN (2004) Thermodynamics of enzyme-catalyzed reactions--a database for quantitative biochemistry. Bioinformatics 20(16): 2874–2877.

75. Gonnerman MC, Benedict MN, Feist AM, Metcalf WW, Price ND (2013) Genomically and biochemically accurate metabolic reconstruction of Methanosarcina barkeri Fusaro, iMG746. Biotechnol J 8(9): 1070–1079.

76. Goodman AL, Mcnulty NP, Zhao Y, Leip D, Mitra RD, et al. (2009) Identifying Genetic Determinants Needed to Establish a Human Gut Symbiont in Its Habitat. Cell Host Microbe 6(3): 279–289.

77. Guzelsoy M, Ralphs TK (2010) Integer Programming Duality. In: Cochran JJ, editor. Encyclopedia of Operations Research and Management Science. Hoboken, NJ: John Wiley & Sons.

78. Hamilton JJ, Reed JL (2012) Identification of Functional Differences in Metabolic Networks Using Comparative Genomics and Constraint-Based Models. PLoS One 7(4): e34670.

79. Hamilton JJ, Dwivedi V, Reed JL (2013) Quantitative assessment of thermodynamic constraints on the solution space of genome-scale metabolic models. Biophys J 105(2): 512–522.

80. Hamilton JJ, Reed JL (2014) Software platforms to facilitate reconstructing genome-scale metabolic networks. Environ Microbiol 16(1): 49–59.

81. Hancock REW (2005) Mechanisms of action of newer antibiotics for Gram-positive pathogens Major targets. Lancet Infect Dis 5(4): 209–218.

82. Hao T, Ma H-W, Zhao X-M, Goryanin I (2010) Compartmentalization of the Edinburgh Human Metabolic Network. BMC Bioinformatics 11: 393.

83. Haraldsdóttir HS, Thiele I, Fleming RMT (2012) Quantitative Assignment of Reaction Directionality in a Multicompartmental Human Metabolic Reconstruction. Biophys J 102(8): 1703–1711.

84. Harcombe WR, Riehl WJ, Dukovski I, Granger BR, Betts A, et al. (2014) Metabolic Resource Allocation in Individual Microbes Determines Ecosystem Interactions and Spatial Dynamics. Cell Rep 7(4): 1104–1115.

85. He X, Reynolds KA (2002) Purification, characterization, and identification of novel inhibitors of the beta-ketoacyl-acyl carrier protein synthase III (FabH) from Staphylococcus aureus. Antimicrob Agents Chemother 46(5): 1310–1318.

86. Heinken A, Sahoo S, Fleming RMT, Thiele I (2013) Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. Gut Microbes 4(1): 1–13.

87. Henry CS, Jankowski MD, Broadbelt LJ, Hatzimanikatis V (2006) Genome-scale thermodynamic analysis of Escherichia coli metabolism. Biophys J 90(4): 1453–1461.

88. Henry CS, Broadbelt LJ, Hatzimanikatis V (2007) Thermodynamics-based metabolic flux analysis. Biophys J 92(5): 1792–1805.

89. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, et al. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol 28(9): 977–982.

90. Herrgård MJ, Swainston N, Dobson PD, Dunn WB, Arga KY, et al. (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nat Biotechnol 26(10): 1155–1160.

91. Heuett WJ, Qian H (2006) Combining flux and energy balance analysis to model large-scale biochemical networks. J Bioinform Comput Biol 4(6): 1227–1243.

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

93. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, et al. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19(4): 524–531.

94. Ishii N, Nakahigashi K, Baba T, Robert M, Soga T, et al. (2007) Multiple high-throughput analyses monitor the response of E. coli to perturbations. Science 316(5824): 593–597.

95. Jamshidi N, Palsson BØ (2007) Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. BMC Syst Biol 1: 26.

96. Jankowski MD, Henry CS, Broadbelt LJ, Hatzimanikatis V (2008) Group contribution method for thermodynamic analysis of complex metabolic networks. Biophys J 95(3): 1487–1499.

97. Jaspers E, Overmann J (2004) Ecological significance of microdiversity: identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. Appl Environ Microbiol 70(8): 4831–4839.

98. Jiang W, Bikard D, Cox D, Zhang F, Marraffini LA (2013) RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat Biotechnol 31(3): 233–239.

99. Joyce AR, Reed JL, White A, Edwards RA, Osterman AL, et al. (2006) Experimental and computational assessment of conditionally essential genes in Escherichia coli. J Bacteriol 188(23): 8259–8271.

100. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28(1): 27–30.

101. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40(D1): D109–D114.

102. Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, et al. (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Brief Bioinform 11(1): 40–79.

103. Khandelwal RA, Olivier BG, Röling WFM, Teusink B, Bruggeman FJ (2013) Community flux balance analysis for microbial consortia at balanced growth. PLoS One 8(5): e64567.

104. Kim H, Kim J-S (2014) A guide to genome engineering with programmable nucleases. Nat Rev Genet 15(5): 321–334.

105. Kim HU, Kim SY, Jeong H, Kim TY, Kim JJ, et al. (2011) Integrative genome-scale metabolic analysis of Vibrio vulnificus for drug targeting and discovery. Mol Syst Biol 7: 460.

106. Kim J, Reed JL (2010) OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains. BMC Syst Biol 4: 53.

107. Kim J, Reed JL, Maravelias CT (2011) Large-scale bi-level strain design approaches and mixed-integer programming solution techniques. PLoS One 6(9): e24162.

108. Klamt S, Saez-Rodriguez J, Gilles ED (2007) Structural and functional analysis of cellular networks with CellNetAnalyzer. BMC Syst Biol 1: 2.

109. Klitgord N, Segrè D (2010) Environments that induce synthetic microbial ecosystems. PLoS Comput Biol 6(11): e1001002.

110. Knoop H, Zilliges Y, Lockau W, Steuer R (2010) The metabolic network of Synechocystis sp. PCC 6803: systemic properties of autotrophic growth. Plant Physiol 154(1): 410–422.

111. Kosuri S, Church GM (2014) Large-scale de novo DNA synthesis: technologies and applications. Nat Methods 11(5): 499–507.

112. Kumar A, Suthers PF, Maranas CD (2012) MetRxn: A Knowledgebase of Metabolites and Reactions Spanning Metabolic Models and Databases. BMC Bioinformatics 13: 6.

113. Kumar VS, Dasika MS, Maranas CD, Satish Kumar V (2007) Optimization based automated curation of metabolic reconstructions. BMC Bioinformatics 8: 212.

114. Kummel A, Panke S, Heinemann M (2006) Systematic assignment of thermodynamic constraints in metabolic network models. BMC Bioinformatics 7: 512.

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

116. Lakshmanan M, Koh G, Chung BKS, Lee D-Y (2014) Software applications for flux balance analysis. Brief Bioinform 15(1): 108–122.

117. Lange RP, Locher HH, Wyss PC, Then RL (2007) The targets of currently used antibacterial agents: lessons for drug discovery. Curr Pharm Des 13(30): 3140–3154.

118. Langridge GC, Phan M, Turner DJ, Perkins TT, Parts L, et al. (2009) Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. Genome Res 19(12): 2308–2316.

119. Latendresse M, Krummenacker M, Trupp M, Karp PD (2012) Construction and completion of flux balance models from pathway databases. Bioinformatics 28(3): 388–396.

120. Lee SJ, Lee D, Kim TY, Kim BH, Lee J, et al. (2005) Metabolic Engineering of Escherichia coli for Enhanced Production of Succinic Acid , Based on Genome Comparison and In Silico Gene Knockout Simulation. Appl Environ Microbiol 71(12): 7880–7887.

121. Lewis NE, Schramm G, Bordbar A, Schellenberger J, Andersen MP, et al. (2010) Large-scale in silico modeling of metabolic interactions between cell types in the human brain. Nat Biotechnol 28(12): 1279–1285.

122. Lewis NE, Hixson KK, Conrad TM, Lerman J a, Charusanti P, et al. (2010) Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Mol Syst Biol 6: 390.

123. Lewis NE, Nagarajan H, Palsson BØ (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. Nat Rev Microbiol 10(4): 291–305.

124. Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13(9): 2178–2189.

125. Lindberg P, Park S, Melis A (2010) Engineering a platform for photosynthetic isoprene production in cyanobacteria, using Synechocystis as the model organism. Metab Eng 12(1): 70–79.

126. Liu X, Sheng J, Curtiss R (2011) Fatty acid production in genetically modified cyanobacteria. Proc Natl Acad Sci 108(17): 6899–6904.

127. Liu Y, Whitman WB (2008) Metabolic, phylogenetic, and ecological diversity of the methanogenic archaea. Ann N Y Acad Sci 1125: 171–189.

128. Liu Y, Beer LL, Whitman WB (2012) Methanogens: a window into ancient sulfur metabolism. Trends Microbiol 20(5): 251–258.

129. Lovley DR, Giovannoni SJ, White DC, Champine JE, Phillips EJP, et al. (1993) Geobacter metallireducens gen. nov. sp. nov., a microorganism capable of coupling the complete oxidation of organic compounds to the reduction of iron and other metals. Arch Microbiol 159(4): 336–344.

130. Lun DS, Rockwell G, Guido NJ, Baym M, Kelner JA, et al. (2009) Large-scale identification of genetic design strategies using local search. Mol Syst Biol 5: 296.

131. Lyon BR, Skurray R (1987) Antimicrobial resistance of Staphylococcus aureus: genetic basis. Microbiol Rev 51(1): 88–134.

132. Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab Eng 5(4): 264–276.

133. Mahadevan R, Bond DR, Butler JE, Coppi M V (2006) Characterization of Metabolism in the Fe(III)-Reducing Organism Geobacter sulfurreducens by Constraint-Based Modeling. Appl Environ Microbiol 72(2): 1558–1568.

134. Mahadevan R, Palsson BØ, Lovley DR (2011) In situ to in silico and back: elucidating the physiology and ecology of Geobacter spp. using genome-scale modelling. Nat Rev Microbiol 9(1): 39–50.

135. Mavrovouniotis ML (1990) Group contributions for estimating standard Gibbs energies of formation of biochemical compounds in aqueous solution. Biotechnol Bioeng 36(10): 1070–1082.

136. Mavrovouniotis ML (1991) Estimation of standard Gibbs energy changes of biotransformations. J Biol Chem 266(22): 14440–14445.

137. Mavrovouniotis ML (1993) Identification of localized and distributed bottlenecks in metabolic pathways. Proc Int Conf Intell Syst Mol Biol 1: 275–283.

138. Mavrovouniotis ML (1996) Duality theory for thermodynamic bottlenecks in bioreaction pathways. Chem Eng Sci 51(9): 1495–1507.

139. McCloskey D, Palsson BØ, Feist AM (2013) Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. Mol Syst Biol 9: 661.

140. McDonald DT, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, et al. (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 6(3): 610–618.

141. McInerney MJ, Rohlin L, Mouttaki H, Kim U, Krupp RS, et al. (2007) The genome of Syntrophus aciditrophicus: life at the thermodynamic limit of microbial growth. Proc Natl Acad Sci 104(18): 7600–7605.

142. McInerney MJ, Struchtemeyer CG, Sieber JR, Mouttaki H, Stams AJM, et al. (2008) Physiology, ecology, phylogeny, and genomics of microorganisms capable of syntrophic metabolism. Ann N Y Acad Sci 1125: 58–72.

143. McInerney MJ, Sieber JR, Gunsalus RP (2009) Syntrophy in anaerobic global carbon cycles. Curr Opin Biotechnol 20(6): 623–632.

144. Montagud A, Navarro E, Fernández de Córdoba P, Urchueguía JF, Patil KR (2010) Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium. BMC Syst Biol 4: 156.

145. Moree WJ, Phelan V V, Wu C-H, Bandeira N, Cornett DS, et al. (2012) Interkingdom metabolic transformations captured by microbial imaging mass spectrometry. Proc Natl Acad Sci 109(34): 13811–13816.

146. Muller AC (2012) Thermodynamic Constraints for Metabolic Networks. Freie Universität Berlin.

147. Müller N, Worm P, Schink B, Stams AJM, Plugge CM (2010) Syntrophic butyrate and propionate oxidation processes: from genomes to reaction mechanisms. Environ Microbiol Rep 2(4): 489–499.

148. Murata N, Wada H (1995) Acyl-lipid desaturases and their importance in the tolerance and acclimatization to cold of cyanobacteria. Biochem J 308(1): 1–8.

149. Nagarajan H, Embree M, Rotaru A-E, Shrestha PM, Feist AM, et al. (2013) Characterization and modelling of interspecies electron transfer mechanisms and microbial community dynamics of a syntrophic association. Nat Commun 4: 2809.

150. Neidhardt FC, Bloch PL, Smith DF (1974) Culture medium for enterobacteria. J Bacteriol 119(3): 736–747.

151. Neidhardt FC (1996) Escherichia coli and Salmonella: Cellular and Molecular Biology. 2nd ed. Neidhardt FC, editor Washington, DC: ASM Press.

152. Neres J, Labello NP, Somu R V, Boshoff HI, Wilson DJ, et al. (2008) Inhibition of siderophore biosynthesis in Mycobacterium tuberculosis with nucleoside bisubstrate analogues: structure-activity relationships of the nucleobase domain of 5'-O-[N-(salicyl)sulfamoyl]adenosine. J Med Chem 51(17): 5349–5370.

153. Neu HC (1992) The crisis in antibiotic resistance. Science 257(5073): 1064–1073.

154. Nie Z, Perretta C, Lu J, Su Y, Margosiak S, et al. (2005) Structure-based design, synthesis, and study of potent inhibitors of beta-ketoacyl-acyl carrier protein synthase III as potential antimicrobial agents. J Med Chem 48(5): 1596–1609.

155. Noor E, Bar-Even A, Flamholz A, Lubling Y, Davidi D, et al. (2012) An integrated open framework for thermodynamics of reactions that combines accuracy and coverage. Bioinformatics 28(15): 2037–2044.

156. Noor E, Haraldsdóttir HS, Milo R, Fleming RMT (2013) Consistent Estimation of Gibbs Energy Using Component Contributions. PLoS Comput Biol 9(7): e1003098.

157. Notebaart RA, van Enckevort FHJ, Francke C, Siezen RJ, Teusink B (2006) Accelerating the reconstruction of genome-scale metabolic networks. BMC Bioinformatics 7: 296.

158. Oberhardt MA, Palsson BØ, Papin JA (2009) Applications of genome-scale metabolic reconstructions. Mol Syst Biol 5: 320.

159. Oberhardt MA, Puchałka J, Martins dos Santos VAP, Papin JA (2011) Reconciliation of genome-scale metabolic reconstructions for comparative systems analysis. PLoS Comput Biol 7(3): e1001116.

160. Oh J, Fung E, Price MN, Dehal PS, Davis RW, et al. (2010) A universal TagModule collection for parallel genetic analysis of microorganisms. Nucleic Acids Res 38(14): e146.

161. Ong WK, Vu TT, Lovendahl KN, Llull JM, Serres MH, et al. (2014) Comparisons of Shewanella strains based on genome annotations, modeling, and experiments. BMC Syst Biol 8: 31.

162. Van Opijnen T, Bodi KL, Camilli A (2009) Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat Methods 6(10): 767–772.

163. Van Opijnen T, Camilli A (2013) Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. Nat Rev Microbiol 11(7): 435–442.

164. Orth JD, Fleming RMT, Palsson BØ (2010) Reconstruction and Use of Microbial Metabolic Networks: the Core Escherichia coli Metabolic Model as an Educational Guide. In: Böck A, III RC, Kaper JB, Karp PD, Neidhardt FC, et al., editors. Escherichia coli and Salmonella: Cellular and Molecular Biology. Washington, DC: ASM Press.

165. Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? Nat Biotechnol 28(3): 245–248.

166. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, et al. (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011. Mol Syst Biol 7(535): 535.

167. Orth JD, Palsson BØ (2012) Gap-filling analysis of the iJO1366 Escherichia coli metabolic network reconstruction for discovery of metabolic functions. BMC Syst Biol 6: 30.

168. Osterlund T, Nookaew I, Nielsen J (2012) Fifteen years of large scale metabolic modeling of yeast: Developments and impacts. Biotechnol Adv 30(5): 979–988.

169. Overbeek RA, Begley T, Butler RM, Choudhuri J V, Chuang H-Y, et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 33(17): 5691–5702.

170. Papp B, Notebaart RA, Pál C (2011) Systems-biology approaches for predicting genomic evolution. Nat Rev Genet 12(9): 591–602.

171. Patil KR, Rocha I, Förster J, Nielsen J (2005) Evolutionary programming as a platform for in silico metabolic engineering. BMC Bioinformatics 6: 308.

172. Patrick WM, Quandt EM, Swartzlander DB, Matsumura I, Information S (2007) Multicopy suppression underpins metabolic evolvability. Mol Biol Evol 24(12): 2716–2722.

173. Pharkya P, Burgard AP, Maranas CD (2004) OptStrain: a computational framework for redesign of microbial production systems. Genome Res 14(11): 2367–2376.

174. Pharkya P, Maranas CD (2006) An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. Metab Eng 8(1): 1–13.

175. Phelan V V, Liu W-T, Pogliano K, Dorrestein PC (2012) Microbial metabolic exchange--the chemotype-to-phenotype link. Nat Chem Biol 8(1): 26–35.

176. Pinchuk GE, Hill EA, Geydebrekht O V, De Ingeniis J, Zhang X, et al. (2010) Constraint-based model of Shewanella oneidensis MR-1 metabolism: a tool for data analysis and hypothesis generation. PLoS Comput Biol 6(6): e1000822.

177. Pinney JW, Shirley MW, McConkey GA, Westhead DR (2005) metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of Plasmodium falciparum and Eimeria tenella. Nucleic Acids Res 33(4): 1399–1409.

178. Pisa KY, Huber H, Thomm M, Müller V (2007) A sodium ion-dependent A1AO ATP synthase from the hyperthermophilic archaeon Pyrococcus furiosus. FEBS J 274(15): 3928–3938.

179. Plata G, Fuhrer T, Hsiao T-L, Sauer U, Vitkup D (2012) Global probabilistic annotation of metabolic networks enables enzyme discovery. Nat Chem Biol 8(10): 848–854.

180. Plugge CM, Dijkema C, Stams AJM (1993) Acetyl-CoA cleavage pathway in a syntrophic propionate oxidizing bacterium growing on fumarate in the absence of methanogens. FEMS Microbiol Lett 110(1): 71–76.

181. Plugge CM, Henstra AM, Worm P, Swarts DC, Paulitsch-Fuchs AH, et al. (2012) Complete genome sequence of Syntrophobacter fumaroxidans strain (MPOB(T)). Stand Genomic Sci 7(1): 91–106.

182. Price MN, Deutschbauer AM, Kuehl J V, Liu H, Witkowska HE, et al. (2011) Evidence-based annotation of transcripts and proteins in the sulfate-reducing bacterium Desulfovibrio vulgaris Hildenborough. J Bacteriol 193(20): 5716–5727.

183. Price ND, Reed JL, Palsson BØ (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2(11): 886–897.

184. Prince RC, Adams MWW (1987) Oxidation-reduction properties of the two Fe4S4 clusters in Clostridium pasteurianum ferredoxin. J Biol Chem 262(11): 5125–5128.

185. Qian H, Beard DA, Liang S (2003) Stoichiometric network theory for nonequilibrium biochemical systems. Eur J Biochem 270(3): 415–421.

186. Raghunathan A, Reed JL, Shin S-I, Palsson BØ, Daefler S (2009) Constraint-based analysis of metabolic capacity of Salmonella typhimurium during host-pathogen interaction. BMC Syst Biol 3: 38.

187. Raman K, Chandra N (2009) Flux balance analysis of biological systems: applications and challenges. Brief Bioinform 10(4): 435–449.

188. Ranganathan S, Suthers PF, Maranas CD (2010) OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. PLoS Comput Biol 6(4): e1000744.

189. Reed JL, Vo TD, Schilling CH, Palsson BØ (2003) An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome Biol 4(9): R54.

190. Reed JL, Palsson BØ (2003) Thirteen Years of Building Constraint-Based In Silico Models of Escherichia coli. J Bacteriol 185(9): 2692–2699.

191. Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, et al. (2006) Systems approach to refining genome annotation. Proc Natl Acad Sci 103(46): 17480–17484.

192. Reed JL (2012) Shrinking the Metabolic Solution Space Using Experimental Datasets. PLoS Comput Biol 8(8): e1002662.

193. Richardson EJ, Watson M (2013) The automatic annotation of bacterial genomes. Brief Bioinform 14(1): 1–12.

194. Rocha I, Maia P, Evangelista P, Vilaça P, Soares S, et al. (2010) OptFlux: an open-source software platform for in silico metabolic engineering. BMC Syst Biol 4: 45.

195. Rost B (2002) Enzyme function less conserved than anticipated. J Mol Biol 318(2): 595–608.

196. Sakamoto T, Shen G, Higashi S, Murata N, Bryant DA (1998) Alteration of low-temperature susceptibility of the cyanobacterium Synechococcus sp. PCC 7002 by genetic manipulation of membrane lipid unsaturation. Arch Microbiol 169(1): 20–28.

197. Sandoval NR, Kim JYH, Glebes TY, Reeder PJ, Aucoin HR, et al. (2012) Strategy for directing combinatorial genome engineering in Escherichia coli. Proc Natl Acad Sci 109(26): 10540–10545.

198. Satish Kumar V, Maranas CD, Kumar VS (2009) GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. PLoS Comput Biol 5(3): e1000308.

199. Satish Kumar V, Ferry JG, Maranas CD (2011) Metabolic reconstruction of the archaeon methanogen Methanosarcina Acetivorans. BMC Syst Biol 5: 28.

200. Schellenberger J, Park JO, Conrad TM, Palsson BØ (2010) BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinformatics 11: 213.

201. Schellenberger J, Lewis NE, Palsson BØ (2011) Elimination of thermodynamically infeasible loops in steady-state metabolic models. Biophys J 100(3): 544–553.

202. Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, et al. (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat Protoc 6(9): 1290–1307.

203. Schilling CH, Letscher D, Palsson BØ (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. J Theor Biol 203(3): 229–248.

204. Schink B (1997) Energetics of syntrophic cooperation in methanogenic degradation. Microbiol Mol Biol Rev 61(2): 262–280.

205. Schink B (2002) Synergistic interactions in the microbial world. Antonie Van Leeuwenhoek 81(1-4): 257–261.

206. Scholten JCM, Conrad R (2000) Energetics of syntrophic propionate oxidation in defined batch and chemostat cocultures. Appl Environ Microbiol 66(7): 2934–2942.

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

208. Schuetz R, Zamboni N, Zampieri M, Heinemann M, Sauer U (2012) Multidimensional Optimality of Microbial Metabolism. Science 336(6081): 601–604.

209. Schwörer B, Thauer RK (1991) Activities of formylmethanofuran dehydrogenase, methylenetetrahydromethanopterin dehydrogenase, methylenetetrahydromethanopterin reductase, and heterodisulfide reductase in methanogenic bacteria. Arch Microbiol 155(5): 459–465.

210. Shastri AA, Morgan JA (2005) Flux balance analysis of photoautotrophic metabolism. Biotechnol Prog 21(6): 1617–1626.

211. Shen CR, Lan EI, Dekishima Y, Baez A, Cho KM, et al. (2011) Driving forces enable high-titer anaerobic 1-butanol synthesis in Escherichia coli. Appl Environ Microbiol 77(9): 2905–2915.

212. Shoaie S, Karlsson F, Mardinoglu A, Nookaew I, Bordel S, et al. (2013) Understanding the interactions between bacteria in the human gut through metabolic modeling. Sci Rep 3: 2532.

213. Shuler ML, Kargi F (2002) Bioprocess Engineering. 2nd ed. Upper Saddle River, NJ: Prentice Hall.

214. Sieber JR, Sims DR, Han C, Kim E, Lykidis A, et al. (2010) The genome of Syntrophomonas wolfei: new insights into syntrophic metabolism and biohydrogen production. Environ Microbiol 12(8): 2289–2301.

215. Sieber JR, McInerney MJ, Gunsalus RP (2012) Genomic insights into syntrophy: the paradigm for anaerobic metabolic cooperation. Annu Rev Microbiol 66: 429–452.

216. Singh A, Cher Soh K, Hatzimanikatis V, Gill RT (2011) Manipulating redox and ATP balancing for improved production of succinate in E. coli. Metab Eng 13(1): 76–81.

217. Smolke CD (2009) The Metabolic Pathway Engineering Handbook: Tools and Applications. 1st ed. Smolke CD, editor Boca Raton, FL: CRC Press.

218. Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci 103(32): 12115–12120.

219. Stams AJM, Van Dijk JB, Dijkema C, Plugge CM (1993) Growth of syntrophic propionate-oxidizing bacteria with fumarate in the absence of methanogenic bacteria. Appl Environ Microbiol 59(4): 1114–1119.

220. Stams AJM, Dong X (1995) Role of formate and hydrogen in the degradation of propionate and butyrate by defined suspended cocultures of acetogenic and methanogenic bacteria. Antonie Van Leeuwenhoek 68(4): 281–284.

221. Stams AJM, de Bok FAM, Plugge CM, van Eekert MHA, Dolfing J, et al. (2006) Exocellular electron transfer in anaerobic microbial communities. Environ Microbiol 8(3): 371–382.

222. Stams AJM, Plugge CM (2009) Electron transfer in syntrophic communities of anaerobic bacteria and archaea. Nat Rev Microbiol 7(8): 568–577.

223. Stolyar SM, Van Dien SJ, Hillesland KL, Pinel N, Lie TJ, et al. (2007) Metabolic modeling of a mutualistic microbial community. Mol Syst Biol 3: 92.

224. Sun J, Sayyar B, Butler JE, Pharkya P, Fahland TR, et al. (2009) Genome-scale constraint-based modeling of Geobacter metallireducens. BMC Syst Biol 3: 15.

225. Sun J, Haveman SA, Bui OT, Fahland TR, Lovley DR (2010) Constraint-based modeling analysis of the metabolism of two Pelobacter species. BMC Syst Biol 4: 174.

226. Suthers PF, Zomorrodi AR, Maranas CD (2009) Genome-scale gene/reaction essentiality and synthetic lethality analysis. Mol Syst Biol 5: 301.

227. Swainston N, Smallbone K, Mendes P, Kell DB, Paton N (2011) The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks. J Integr Bioinform 8(2): 186.

228. Tan X, Yao L, Gao Q, Wang W, Qi F, et al. (2011) Photosynthesis driven conversion of carbon dioxide to fatty alcohols and hydrocarbons in cyanobacteria. Metab Eng 13(2): 169–176.

229. Temme K, Zhao D, Voigt CA (2012) Refactoring the nitrogen fixation gene cluster from Klebsiella oxytoca. Proc Natl Acad Sci 109(18): 7085–7090.

230. Tepper N, Shlomi T (2010) Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways. Bioinformatics 26(4): 536–543.

231. Tersteegen A, Hedderich R (1999) Methanobacterium thermoautotrophicum encodes two multisubunit membrane-bound [NiFe] hydrogenases . Transcription of the operons and sequence analysis of the deduced proteins. Eur J Biochem 264(3): 930–943.

232. Tervo CJ, Reed JL (2012) FOCAL: an experimental design tool for systematizing metabolic discoveries and model development. Genome Biol 13(12): R116.

233. Tervo CJ, Reed JL (2013) BioMog: A Computational Framework for the De Novo Generation or Modification of Essential Biomass Components. PLoS One 8(12): e81322.

234. Thauer RK (1998) Biochemistry of methanogenesis : a tribute to Marjory Stephenson. Microbiology 144(9): 2377–2406.

235. Thauer RK, Kaster A-K, Seedorf H, Buckel W, Hedderich R (2008) Methanogenic archaea: ecologically relevant differences in energy conservation. Nat Rev Microbiol 6(8): 579–591.

236. Thiele I, Palsson BØ (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc 5(1): 93–121.

237. Thiele I, Hyduke DR, Steeb B, Fankam G, Allen DK, et al. (2011) A community effort towards a knowledge-base and mathematical model of the human pathogen Salmonella Typhimurium LT2. BMC Syst Biol 5: 8.

238. Thiele I, Swainston N, Fleming RMT, Hoppe A, Sahoo S, et al. (2013) A community-driven global reconstruction of human metabolism. Nat Biotechnol 31(5): 419–425.

239. Thomason LC, Costantino N, Court DL (2007) E. coli genome manipulation by P1 transduction. Current Protocols in Molecular Biology. Hoboken, NJ: John Wiley & Sons. pp. 1.17.1–1.17.8.

240. Tian W, Skolnick J (2003) How Well is Enzyme Function Conserved as a Function of Pairwise Sequence Identity? J Mol Biol 333(4): 863–882.

241. Varma A, Palsson BØ (1994) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. Appl Environ Microbiol 60(10): 3724–3731.

242. Vu TT, Stolyar SM, Pinchuk GE, Hill EA, Kucek LA, et al. (2012) Genome-scale modeling of light-driven reductant partitioning and carbon fluxes in diazotrophic unicellular cyanobacterium Cyanothece sp. ATCC 51142. PLoS Comput Biol 8(4): e1002460.

243. Walsh CT (2003) Where will new antibiotics come from? Nat Rev Microbiol 1(1): 65–70.

244. Wang HH, Isaacs FJ, Carr PA, Sun ZZ, Xu G, et al. (2009) Programming cells by multiplex genome engineering and accelerated evolution. Nature 460(7257): 894–898.

245. Wang HH, Kim H, Cong L, Jeong J, Bang D, et al. (2012) Genome-scale promoter engineering by coselection MAGE. Nat Methods 9(6): 591–593.

246. Warner JR, Reeder PJ, Karimpour-Fard A, Woodruff LBA, Gill RT (2010) Rapid profiling of a microbial genome using mixtures of barcoded oligonucleotides. Nat Biotechnol 28(8): 856–862.

247. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2007) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 35(S1): D5–D12.

248. Williams HP (1999) Model Building in Mathematical Programming. 4th ed. Hoboken, NJ: John Wiley & Sons.

249. Wintermute EH, Silver PA (2010) Emergent cooperation in microbial metabolism. Mol Syst Biol 6: 407.

250. Wolsey LA (1998) Integer Programming. 1st ed. Hoboken, NJ: Wiley-Interscience.

251. Worm P, Stams AJM, Cheng X, Plugge CM (2011) Growth- and substrate-dependent transcription of formate dehydrogenase and hydrogenase coding genes in Syntrophobacter fumaroxidans and Methanospirillum hungatei. Microbiology 157(1): 280–289.

252. Wright GD, Walsh CT (1992) D-Alanyl-D-alanine ligases and the molecular mechanism of vancomycin resistance. Acc Chem Res 25(10): 468–473.

253. Yamamoto N, Nakahigashi K, Nakamichi T, Yoshino M, Takai Y, et al. (2009) Update on the Keio collection of Escherichia coli single-gene deletion mutants. Mol Syst Biol 5: 335.

254. Yang C, Hua Q, Baba T, Mori H, Shimizu K (2003) Analysis of Escherichia coli anaplerotic metabolism and its regulation mechanisms from the metabolic responses to altered dilution rates and phosphoenolpyruvate carboxykinase knockout. Biotechnol Bioeng 84(2): 129–144.

255. Yang F, Qian H, Beard DA (2005) Ab initio prediction of thermodynamically feasible reaction directions from biochemical network stoichiometry. Metab Eng 7(4): 251–259.

256. Yang L, Cluett WR, Mahadevan R (2011) EMILiO: a fast algorithm for genome-scale strain design. Metab Eng 13(3): 272–281.

257. Yim H, Haselbeck R, Niu W, Pujol-Baxley C, Burgard AP, et al. (2011) Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. Nat Chem Biol 7(7): 444–445.

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

259. Zarraonaindia I, Smith DP, Gilbert JA (2013) Beyond the genome: community-level analysis of the microbial world. Biol Philos 28(2): 261–282.

260. Zengler K, Palsson BØ (2012) A road map for the development of community systems (CoSy) biology. Nat Rev Microbiol 10(5): 366–372.

261. Zhang Y, Wade MM, Scorpio A, Zhang H, Sun Z (2003) Mode of action of pyrazinamide: disruption of Mycobacterium tuberculosis membrane transport and energetics by pyrazinoic acid. J Antimicrob Chemother 52(5): 790–795.

262. Zhu Q, Qin T, Jiang Y-Y, Ji C, Kong D-X, et al. (2011) Chemical Basis of Metabolic Network Organization. PLoS Comput Biol 7(10): e1002214.

263. Zhuang K, Ma E, Lovley DR, Mahadevan R (2012) The design of long-term effective uranium bioremediation strategy using a community metabolic model. Biotechnol Bioeng 109(10): 2475–2483.

264. Zimhony O, Cox JS, Welch JT, Vilchèze C, Jacobs WR (2000) Pyrazinamide inhibits the eukaryotic-like fatty acid synthetase I (FASI) of Mycobacterium tuberculosis. Nat Med 6(9): 1043–1047.

265. Zomorrodi AR, Suthers PF, Ranganathan S, Maranas CD (2012) Mathematical optimization applications in metabolic networks. Metab Eng 14(6): 672–686.

266. Zomorrodi AR, Maranas CD (2012) OptCom: A Multi-Level Optimization Framework for the Metabolic Modeling and Analysis of Microbial Communities. PLoS Comput Biol 8(2): e1002363.

267. Zomorrodi AR, Islam MM, Maranas CD (2014) d-OptCom: Dynamic Multi-level and Multi-objective Metabolic Modeling of Microbial Communities. ACS Synth Biol 3(4): 247–257.