UNIVERSITY OF CALIFORNIA, SAN DIEGO

# Predicting growth optimization strategies with metabolic/expression models

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

in

Bioinformatics and Systems Biology

by

Joanne K. Liu

Committee in charge:

Professor Karsten Zengler, Chair
Professor Nathan Lewis, Co-Chair
Professor Michael Burkart
Professor Terry Gaasterland
Professor Bernhard Palsson
Professor Milton Saier

2017

The dissertation of Joanne K. Liu is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2017

DEDICATION

To my mom and dad, who I cannot thank enough for supporting me throughout my education, and to The One.

# EPIGRAPH

*Essentially, all models are wrong, but some are useful.*

—George E. P. Box

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

# LIST OF SUPPLEMENTARY FILES

- Liu_chapter4_Additional_file.xlsx

ACKNOWLEDGEMENTS

I wish to acknowledge Professor Karsten Zengler for his support as my mentor and chair of my committee. I am beyond grateful for his encouragement throughout my Ph.D. career. I also wish to acknowledge Professor Nathan Lewis for his steady support and insight as co-chair of my committee. I also thank the past and present members of the Zengler, Lewis, and Palsson labs, as well as my fellow students from Bioinformatics and Systems Biology.

Chapter 2, in part, is currently being prepared for submission for publication of the material. Joanne Liu, Ali Ebrahim, Mahmoud Al Bassam, Colton Lloyd, Ji-Nu Kim, Connor Olson, and Karsten Zengler. A systems biology approach to investigate proteome control of acetate and ethanol production in *Clostridium ljungdahlii* (working title). The dissertation author was the primary investigator and author of this material.

Chapter 3, in full, is currently being prepared for submission for publication of the material. Joanne Liu, Nathan Lewis, and Karsten Zengler. Exploring the evolutionary significance of tRNA operon structure using metabolic and gene expression models (working title). The dissertation author was primary investigator and author of this paper.

Chapter 4, in full, is a modified reprint of the material as it appears in BMC Systems Biology 2014. Joanne K Liu, Edward J O'Brien, Joshua A Lerman, Karsten Zengler, Bernhard O Palsson and Adam M Feist. Reconstruction and

modeling protein translocation and compartmentalization in *Escherichia coli* at the genome-scale. BMC Systems Biology 2014 8:110. The dissertation author was a primary investigator and author of this paper.

Chapter 5, in part, is currently being prepared for submission for publication of the material. Joanne Liu, Ali Ebrahim, Mahmoud Al Bassam, Colton Lloyd, Ji-Nu Kim, Connor Olson, and Karsten Zengler. A systems biology approach to investigate proteome control of acetate and ethanol production in *Clostridium ljungdahlii* (working title). The dissertation author was the primary investigator and author of this material.

## VITA

| | |
|---|---|
| 2017 | Doctor of Philosophy, University of California, San Diego<br>Bioinformatics and Systems Biology |
| 2011 | Bachelors of Science with honors and *Summa Cum Laude*,<br>College of Biological Sciences, University of Minnesota, Twin Cities<br>Genetics, Cell Biology & Development |

## PUBLICATIONS

Colton J Lloyd, Ali Ebrahim, Laurence Yang, Zachary A King, Edward Catoiu, Edward J O'Brien, **Joanne K Liu**, Bernhard O Palsson. *bioRXiv*, 2017. DOI: 10.1101/106559

Neil Swainston, Kieran Smallbone, Hooman Hefzi, Paul D Dobson, Judy Brewer, Michael Hanscho, Daniel C Zielinski, Kok Siong Ang, Natalie J Gardiner, Jahir M Gutierrez, Sarantos Kyriakopoulos, Meiyappan Lakshmanan, Shangzhong Li, **Joanne K Liu**, Veronica S Martnez, Camila A Orellana, Lake-Ee Quek, Alex Thomas, Juergen Zanghellini, Nicole Borth, Dong-Yup Lee, Lars K. Nielsen, Douglas B Kell, Nathan E Lewis, Pedro Mendes. Recon 2.2: from reconstruction to model. *Metabolomics*, 2016, 12:109. DOI: 10.1007/s11306-016-1051-4

Mallory Embree, **Joanne K Liu**, Mahmoud M Al-Bassam, Karsten Zengler. Networks of energetic and metabolic interactions define dynamics in microbial communities. *Proceedings of the National Academy of Sciences*, 2015, 112: 15450-15455. DOI: 10.1073/pnas.1506034112

**Joanne K Liu**, Edward J O'Brien, Joshua A Lerman, Karsten Zengler, Bernhard O Palsson, Adam M Feist. Reconstruction and modeling protein translocation and compartmentalization in *Escherichia coli* at the genome-scale. *BMC Systems Biology*, 2014, 8:110. DOI: 10.1186/s12918-014-0110-6

ABSTRACT OF THE DISSERTATION

**Predicting growth optimization strategies with metabolic/expression models**

by

Joanne K. Liu

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2017

Professor Karsten Zengler, Chair
Professor Nathan Lewis, Co-Chair

Systems biology strives to understand complex multi-component biological processes and capture knowledge of their function through models. With metabolic and gene expression models (ME-models), we can mathematically and simultaneously represent the majority of these processes, including transcription, translation, and metabolism. This enables us to compute the molecular constituents of a cell as a function of genetic and environmental parameters. ME-models represent an improvement in current capabilities to predict phenotypes, as demonstrated

by the reconstruction and validation of a ME-model for the acetogen *Clostridium ljungdahlii*. *C. ljungdahlii* can grow autotrophically on carbon monoxide (CO), and/or carbon dioxide + hydrogen ($CO_2$+$H_2$) and fix these gases into multicarbon organics, an ability that can be redirected to produce biocommodities. The *C. ljungdahlii* ME-model was able to improve growth rate predictions, identify previously unknown secretion products, and compute the transcriptome of *C. ljungdahlii* accurately.

ME-models offer the opportunity to systematically explore the interface between protein and function. First, perturbations of tRNA co-expression in ME-models revealed unique organization solutions to two different selective pressures: Optimization of growth through minimal co-expression of tRNAs, and efficiency of resources through optimal grouping of tRNAs. Second, because of the incorporation of protein translocation and membrane function, a ME-model was able to recapitulate acetate production during glucose consumption due to membrane overcrowding. Third, a ME-model highlighted how variations in nickel availability impacts metalloproteins, thereby controlling growth and secretion rates of fermentation products. Thus, three features that could constrain the proteome of an organism - genome architecture, cell structure, and media composition - were successfully interrogated using ME-models.

# Chapter 1

# Systems biology, metabolism, and non-metabolic constraints

## 1.1 Introduction to metabolic and gene expression models

Although a prokaryote is a single cell, the many intricate components that enable the organism to function are complex and interconnected. Systems biology strives to understand the network of these biological processes and capture the knowledge through models. Constraint-based models of metabolism (M-models) are one of the few models that reach genome-scale while maintaining molecular detail and have proven to be effective for predictions of metabolic flux and strain design [1].

An organism is more than the sum of it's biochemical reactions. For these reactions to occur, genes and proteins are necessary to catalyze them. Thus, the

predictive capability of M-models was expanded to include gene expression pathways (i.e., ME-model) [2, 3]. With the development of the ME-model, we can mathematically represent major cellular processes such as macromolecular synthesis, basic transcriptional regulation, and protein translocation. Also included are accounts for production of transcriptional units (TU), functional RNAs (i.e., tRNAs, rRNAs, etc.), proteins, prosthetic groups, and cofactors, as well as the formation and translocation of multimeric complexes. The energetic costs associated with all aspects of these processes are represented too. With these ME-models, we can now compute the molecular constitution of cells as a function of genetic and environmental parameters instead of utilizing experimental data to constrain the models. The ME-model explicitly predicts transcript and protein abundances, which allows direct evaluation with quantitative experimental transcriptomics and proteomics. To date, it has only been possible to perform indirect comparative analysis between omics data and M-models or to neglect the complexity of the genotype-phenotype relationship and use omics data as *ad hoc* constraints for enzyme activities [4, 5, 6, 7, 8]. The ME-model approach overcomes this previous lack of a mechanistic systems-level framework for analyzing a myriad of molecular components in the context of cellular physiology.

For example, a comparative *in silico* transcriptomics analysis with the ME-model of *Thermatoga maritima* enabled prediction of regulons. With the ME-model's unprecedented capacity to investigate the interdependence of cellular processes, we were able to predict and validate genes that were differentially regulated for growth on L-arabinose versus growth on cellobiose in minimal medium and were able to discover the regulons of the transcription factors (AraR and CelR)

governing this shift [3]. Furthermore, the *Escherichia coli* ME-model demonstrated that accounting for macromolecular costs caused intrinsic limitations that led to accurate *in silico* predictions of maximum growth rates, substrate uptake rates, and secretion rates [9, 10]. Thus, by expanding the numbers of represented cellular processes, ME-models have significantly broadened the scope and predictability of microbial systems biology.

### 1.1.1 *Clostridium ljungdahlii*

The acetogen *Clostridium ljungdahlii* has emerged as a potential chassis for strain designed chemical production for not only can it grow heterotrophically on a diverse set of sugars, but it can also grow autotrophically on carbon monoxide (CO), carbon dioxide ($CO_2$) and hydrogen ($H_2$), or a mixture of all three gases (*i.e.*,syngas). When grown autotrophically, *C. ljungdahlii* metabolizes the gases into multicarbon organics, an ability that can be redirected and engineered to produce biocommodities from low cost substrates.

To advance towards this goal, a *C. ljungdahlii* ME-model was reconstructed and validated. This ME-model, named iJL965-ME, accounts for 965 ORFs that are responsible for the production of transcriptional units, functional RNAs (*e.g.*, tRNAs, rRNAs), prosthetic groups, cofactors, and protein complexes that are necessary for all of the major central metabolic, amino acid, nucleotide, and lipid biosynthesis pathways. iJL965-ME was able to compute the molecular constitution (*i.e.*, transcriptome, proteome, and fluxome) of *C. ljungdahlii* as a function of genetic and environmental parameters, and was able to do so accurately. The ME-model recapitulated results from autotrophic and heterotrophic, including growth

rate, product secretion, and transcriptomics.

## 1.2 Demonstration of ME-model capabilities

ME-models offer the opportunity to systematically explore the interface between protein and function. Attempts to creatively constrain the proteome revealed and corroborated evolutionary strategies to optimize resources. In particular, three different features of an organism - genome architecture, cell structure, and media composition - were successfully interrogated using ME-models.

### 1.2.1 Genome architecture: tRNA operon structure

Translation must be carefully controlled because it requires the highest energy and resource expenditure of any process in fast-growing cells. The formation and maintenance of operons, which is a co-regulated cluster of genes that are expressed on the same RNA transcript, may promote gene expression efficiency in organisms. Interrogation of tRNA gene location and co-expression revealed that selective pressures to minimize resource costs extend to the molecular level, affecting even genomic tRNA organization. However, the exact solution for surviving selection is different between organisms, as highlighted by a comparison between the *E. coli* and *C. ljungdahlii* ME-models.

### 1.2.2 Cell structure: Protein translocation and compartmentalization

Membranes play a crucial role in cellular functions. They provide a physical barrier, control trafficking of substances entering and leaving the cell, are a major determinant of cell ultra-structure, and more. However, membrane- and location-based processes were not yet reconstructed and integrated into genome-scale models. The *E. coli* ME-model, iOL1650-ME, was expanded to include protein translocation and compartmentalization. The results of the updated *E. coli* ME-model, iJL1678-ME, recapitulated *in vivo* data. Furthermore, iJL1678-ME was used to support the hypothesis that limited membrane space reduces the respiratory capabilities of *E coli*.

### 1.2.3 Media composition: Nickel availability

Trace metals are essential for all living organisms, for they are required in transcription, translation, and metabolism. However, when faced with suboptimal amounts of metal, not all enzymes will be functional due to a lack of adequate cofactors. In such a situation, certain functions will be prioritized over others. This results in a non-linear relationship between nickel availability and growth rate for *C. ljungdahlii* grown on CO, and a decrease in acetate production when nickel-limited *C. ljungdahlii* was grown on fructose.

## 1.3 References

[1] Aarash Bordbar, Jonathan M. Monk, Zachary A. King, and Bernhard O. Palsson. Constraint-based models predict metabolic and associated cellular functions. *Nature Reviews Genetics*, 15(2):107–120, 2014.

[2] Ines Thiele, Neema Jamshidi, Ronan M. T. Fleming, Bernhard Ø. Palsson, and P Stothard. Genome-Scale Reconstruction of *Escherichia coli's* Transcriptional and Translational Machinery: A Knowledge Base, Its Mathematical Formulation, and Its Functional Characterization. *PLoS Computational Biology*, 5(3):e1000312, 2009.

[3] Joshua A. Lerman, Daniel R. Hyduke, Haythem Latif, Vasiliy A. Portnoy, Nathan E. Lewis, Jeffrey D. Orth, Alexandra C. Schrimpe-Rutledge, Richard D. Smith, Joshua N. Adkins, Karsten Zengler, and Bernhard O. Palsson. In silico method for modelling metabolism and gene product expression at genome scale. *Nature Communications*, 3:929, 2012.

[4] Mats Åkesson, Jochen Förster, and Jens Nielsen. Integration of gene expression data into genome-scale metabolic models. *Metabolic Engineering*, 6(4): 285–293, 2004.

[5] Scott A. Becker and Bernhard O. Palsson. Context-Specific Metabolic Networks Are Consistent with Experiments. *PLoS Computational Biology*, 4(5): e1000082, 2008.

[6] Tomer Shlomi, Moran N Cabili, Markus J Herrgård, Bernhard Ø Palsson, and Eytan Ruppin. Network-based prediction of human tissue-specific metabolism. *Nature biotechnology*, 26(9):1003–10, 2008.

[7] Caroline Colijn, Aaron Brandes, Jeremy Zucker, Desmond S. Lun, Brian Weiner, Maha R. Farhat, Tan-Yun Cheng, D. Branch Moody, Megan Murray, and James E. Galagan. Interpreting Expression Data with Metabolic Flux Models: Predicting *Mycobacterium tuberculosis* Mycolic Acid Production. *PLoS Computational Biology*, 5(8):e1000489, 2009.

[8] Nathan E Lewis, Kim K Hixson, Tom M Conrad, Joshua A Lerman, Pep Charusanti, Ashoka D Polpitiya, Joshua N Adkins, Gunnar Schramm,

Samuel O Purvine, Daniel Lopez-Ferrer, Karl K Weitz, Roland Eils, Rainer König, Richard D Smith, and Bernhard Ø Palsson. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Molecular Systems Biology*, 6:390, 2010.

[9] Laurence Yang, Ding Ma, Ali Ebrahim, Colton J. Lloyd, Michael A. Saunders, and Bernhard O. Palsson. solveME: fast and reliable solution of nonlinear ME models. *BMC Bioinformatics*, 17(391), 2016.

[10] E. J. O'Brien, J. A. Lerman, R. L. Chang, D. R. Hyduke, and B. O. Palsson. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Molecular Systems Biology*, 9(1):693–693, 2014.

# Chapter 2

# Accurate prediction of bacterial phenotypes using the ME-model framework in *Clostridium ljungdahlii*

## 2.1 Introduction

Acetogens have been investigated as a promising sustainable alternative to convert waste gases containing $CO_2$, $H_2$, and CO (i.e., syngas) into multi-carbon commodities like biofuels [1, 2]. The Wood-Ljungdahl pathway (WLP) in *Clostridium ljungdahlii* enables the use of either $H_2$ or CO as electron donors with accompanied reduction of $CO_2$, thereby making WLP the only known $CO_2$-fixing pathway coupled to energy conservation [3]. The feasibility of autotrophic growth

was poorly understood for a long time as no ATP was gained at the substrate level. Knowledge of how a bacterium completely lacking cytochrome-encoding genes could maintain the proton motive force was lacking. It was then discovered that the RNF complex couples ferredoxin oxidation, $NAD^+$ reduction and proton exportation by a novel mechanism called electron bi furcation [4]. To explore how growth strategies occur, models like constraint-based genome scale models of metabolism (i.e., M-models) have been useful for gaining insight to possible energy flux routes [5, 6, 7, 8]. While M-models have enabled much progress in elucidating cofactor fluxes, critical components of the cell, such as the production of macromolecules and the mechanistic utilization of metals, vitamins, and cofactors, are usually absent in these models thereby limiting in-depth understanding of cellular life.

So-called metabolic and gene expression models (ME-models) include not only metabolic reactions, but they also include explicit representations of major cellular processes such as macromolecular synthesis and basic transcriptional regulation, which significantly broadens the scope and predictability of microbial systems biology [9, 10]. Specifically, the ME-model will: 1) Account for the transcriptional and translational cost of proteins and complex formation; 2) Incorporate the energetics associated with cofactor dependencies and prosthetic group usage; 3) Quantitatively predict transcript and protein levels; 4) Predict optimal codon usage for heterologous pathways. With these ME-models, the optimal molecular constitution of cells can be computed as a function of genetic and environmental parameters. Since both RNA and protein abundances are explicitly predicted, cofactor requirements can now be explored.

Here, we chose the model acetogen *C. ljungdahlii* to reconstruct the first ME-model of a gram-positive bacterium. There are several attractive features that endorse *C. ljungdahlii* for as a platform for gaining in-depth knowledge necessary to better understand acetogens. It is readily cultured in the laboratory in simple medium, either on a diverse set of five and six carbon sugars, or with CO or $H_2$ as electron donor. Furthermore, genetic manipulation tools have also been developed for this organism, so that genes may be knocked out, knocked in, and over-expressed [11, 12, 13, 14]. Thus, a foundation for potential large-scale production of chemicals from CO and $CO_2$ using systems-guided, rational strain design is currently feasible for *C. ljungdahlii*.

The completed *C. ljungdahlii* ME-model, named iJL965-ME, captures all major central metabolic, amino acid, nucleotide, lipid, major cofactors, and vitamin synthesis pathways as well as pathways to synthesis RNA and protein molecules necessary to catalyze these reactions. Furthermore, the reconstruction includes the WLP, with updated cofactors, and its associated mechanisms for energy conservation. iJL965-ME was used to reveal how protein allocation and media composition influence metabolic pathways and energy conservation in *C. ljungdahlii*, and to accurately predict secretion of acetate, ethanol, and glycerol during changing carbon.

## 2.2 Results

### 2.2.1 Updating the M-model

An existing genome-scale M-model (iHN637) was updated [5]. By using recent literature and genome annotations as reference [15, 16, 17, 18, 19], 28 reactions were added and four reactions removed from iHN637. The updated M-model (iJL680) consisted of 43 additional genes and contained updated cofactor stoichiometry and directionality of redox reactions based on experimental data (Fig 2.1, Table 2.1).

### 2.2.2 Reconstructing a *C. ljungdahlii* ME-model named iJL965-ME

Following established methods, the gene expression network (i.e., E-matrix) for *C. ljungdahlii* was reconstructed [20, 21, 22, 23]. This reconstruction included additional 196 protein-coding open reading frames (ORFs), 89 RNA genes, 576 transcription units (TUs) (415 of which were rho-dependent and 29 RNA-stable), 19 types of rRNA modifications, 17 types of tRNA modifications, 735 protein complexes with updated stoichiometry, 219 modified protein complexes, and 134 translocated proteins (Supplemental materials). Lastly, the turnover rate for metabolic enzymes (approximated by the keff constant and a required parameter for ME-models) was set to the average turnover rate of all enzymes found in acetogens in the enzyme database Brenda, 25 $s^{-1}$ [24]. Coupling constraints linking macromolecular synthesis costs with reactions were calculated using the formulation in COBRAme [10, 23]. Relative ratios between and within simula-

tions reflect biology, as shown within this study. Using the COBRAme framework, the *C. ljungdahlii* E-matrix was integrated with iJL680 to create the ME-model, iJL965-ME. iJL965-ME accounts for all of the major central metabolic pathways and biomass synthesis pathways as well as transcription, translation, RNA modifications, protein modifications, and translocation reactions necessary to catalyze these reactions (Fig 2.2). Therefore, iJL965-ME enables the prediction of fermentation profiles, including overflow metabolism, gene expression and usage of co-factors and metals, which are described in detail below.

### 2.2.3 Accuracy of predicted growth and yield phenotypes improves with iJL965-ME

Unlike the M-model, iJL965-ME enabled batch simulations (i.e., maximum nutrient uptake) and nutrient-limited growth conditions for *C. ljungdahlii*. Due to internal constraints on protein production and catalysis, referred to as proteomic limitations, iJL965-ME growth rate was a non-linear function of the substrate uptake rate. Thus, optimal carbon uptake rate and maximum growth rate could be simultaneously predicted, whereas M-models require information of one rate to predict the other [10]. As a result, we predicted unique growth rate and yield functions for growth with CO, $CO_2$+$H_2$, or fructose (Fig 2.3). In general, M-models for acetogens could not predict alternative fermentation products other than acetate without additional constraints on redox fluxes, oxygen uptake, or the objective function [5, 6, 7]. However, iJL965-ME was able to intrinsically predict changes in the primary fermentation product over substrate availability for CO and fructose growth. When protein production approached proteome limitations (exemplified

by maximum growth rate *in silico* and stationary phase *in vivo*), iJL965-ME correctly predicted a switch from acetate secretion to ethanol secretion (Fig 2.3A, C; Fig 2.4). Thus, iJL965-ME was able to recapitulate overflow metabolism due to a combination of excess electrons and proteome limitations.

When the model was forced to produce ethanol (by removing acetate secretion) under growth on $CO_2$+$H_2$, a 0.3% drop in maximum growth rate was predicted (hence why iJL965-ME never predicts ethanol production in $CO_2$+$H_2$ conditions), and since ethanol was a less oxidized product than acetate, more $H_2$ was required, which increases the ratio of consumed $H_2$:$CO_2$ from 2.1 to 2.5.

The ME-model also predicted substrate-specific growth rates with high accuracy. Due to distinct resource requirements (*e.g.*, proteome) when metabolizing different substrates, iJL965-ME predicted unique maximum growth rates for individual substrates. Unlike the M-model (iJL680), which predicted that glucose and fructose would have identical growth rates, iJL965-ME correctly predicted slower growth on glucose than for fructose. iJL965-MEs growth rate predictions were more accurate (Pearsons r: $0.68 > 0.29$; Spearman $\rho$: $0.60 > 0.091$; Fig 2.5A). Furthermore, iJL965-ME was better at predicting the ratio of maximum acetate secretion rate to substrate uptake rate than the M-models iHN637 and iJL680 (r: $0.97 > 0.22$; Fig 2.5B).

Interestingly, iJL965-ME predicted previously unknown secretion of glycerol (<2.5e-3 mmol*gDW$^{-1}$*h$^{-1}$) following acetate and ethanol production during growth on xylose or glucose, but not on arginine or pyruvate. According to the model, glycerol secretion occurred due to proteomic-limitations and overflow metabolism, as the cell no longer invested the resources necessary to recycle glyc-

erol, a byproduct of cardiolipin production (Fig 2.5C). In order to verify glycerol production, we carried out HPLC analysis and measured 0.024±0.012 mM and 0.083±0.018 mM of glycerol of cultures grown on either xylose or glucose, respectively (Fig 2.6).

### 2.2.4 Predicted gene expression recapitulates *in vivo* data

One advantage of iJL965-ME is the integration of the optimal RNA and protein content as part of the biomass composition of *C. ljungdahlii*, which enables *in silico* predictions of transcription and translation through their flux reactions (mmol*gDW$^{-1}$*h$^{-1}$ [10, 23]). To test the accuracy of our model, genes were categorized by RAST subsystems and summed as per predicted transcription flux reactions [15]. The model predictions were strongly correlated to RNA-seq data for *C. ljungdahlii* grown on CO, $CO_2$+$H_2$, or fructose ($r \geq 0.82$, Fig 2.7). At the highest correlation, all categories fell within the prediction interval of the linear regression (Fig 2.8A-C).

At the gene level, many genes could be strongly linked to biomass production and growth rate (r>0.9, p<0.05*Bonferonni, Fig 2.9). Such genes were not the same between carbon substrate conditions, with commonly shared growth rate correlated proteins reduced to rRNA genes and some, but not all, tRNAs (Fig 2.9 callout). Under autotrophic conditions, WLP proteins were correlated more with substrate availability than growth rate ($r_{CO}$: 0.983>0.955, $r_{CO_2+H_2}$: 0.996>0.884; Fig 2.8D, E). In addition, both carbon monoxide dehydrogenase (CODH4) and 5,10-methylenetetrahydrofolate reductase (MTHFR5), essential reactions in the WLP, were linearly related to CO uptake during growth on CO, while other non-

WLP redox reactive enzyme activities were correlated with growth rate (Fig 2.10). Similarly, WLP reactions were linearly linked to $CO_2$ uptake in $CO_2$+$H_2$ conditions, in addition to the linear response of ferredoxin:NADPH hydrogenase to $H_2$, while non-WLP redox reactions were correlated with growth rate (Fig 2.11).

In heterotrophic conditions, the WLP was more active under nutrient-limitations than proteomic-limitations, as its activity level was related to acetate secretion ($r = 0.993$, $p<0.01$, Fig 2.8F). The WLP was recapturing $CO_2$ for biomass production using the reducing power gained by metabolizing fructose. At greater than 57% of the optimal fructose uptake (Fig 2.8F), the primary provider of oxidized ferredoxin switched from WLP to ferredoxin:NADP reductase (FRNDPR2r) and acetaldehyde:ferredoxin oxidoreductase (AOR_CL) (Fig 2.12). Extraneous reducing power captured by $NAD^+$ from glyceraldehyde-3-phosphate dehydrogenase (GADP) was removed by producing ethanol (alcohol dehydrogenase; ALCD2x) (Fig 2.12). These findings are corroborated by a previous report that *C. ljungdahlii* grows mixotrophically, instead of heterotrophically, when presented with sugar as a carbon source [25].

## 2.3 Discussion

We showed that the incorporation of the E-matrix into constraint-based genome-scale models significantly widens the scope of their application, including prediction of overflow metabolism and optimal expression levels, as well as media optimization strategies. Such capabilities proved useful for exploring and understanding system responses of *C. ljungdahlii*. The reconstructed *C. ljung-*

*dahlii* ME-model (iJL965-ME) was not only more accurate than the M-model at predicting growth rates and acetate secretion rates, but was also capable of predicting secretion of ethanol ($H^2$, as a less effective oxidizing agent than CO, was an exception) and the novel secretion of glycerol (Fig 2.3, 2.5). Furthermore, *in silico* predictions of gene/subsystem expression were highly comparable to *in vivo* transcriptomics for three separate conditions, bolstering confidence in predicting macromolecular responses to environmental changes (Fig 2.8A-C).

As demonstrated in this study, ME-models like iJL965-ME provide a comprehensive, genome-scale, systems biology approach that links the environment and macronutrient metabolism.

## 2.4 Methods

### 2.4.1 Bacterial growth conditions

*Clostridium ljungdahlii* (ATCC 55383) was grown under anaerobic conditions containing PETC medium (ATCC medium 1754) at 37°C. Fructose cultures were grown in 125 mL serum bottles containing 100 mL of medium plus 28 mM fructose, CO in 125 mL serum bottles containing 25 mL of media and bottles were pressurized once with CO to 18 PSI. Pyruvate, xylose, glucose, and arginine experiments were performed in test tubes containing 10 mL of medium and 30 mM of carbon source. Medium contained 0.10 mM of $NiCl_2$*6 $H_2O$ (*i.e.*, 1x). Growth was routinely determined by measurement of $OD_{600}$. Concentrations of fructose, acetate, ethanol, and glycerol were determined by high-performance liquid chromatography (Waters) as previously described [26]. Detection was performed by

UV absorption at 410 nm.

### 2.4.2 RNA isolation, removal of rRNA and library preparation of CO-grown cells

All experiments were performed using two biological replicates. Cell pellets were collected by centrifugation at room temperature for 5 mins at 5000 rcf. Growth medium was removed and cell pellets were snap frozen immediately in liquid nitrogen, then kept at -80°C. Cell lysates were prepared by grinding the pellets in liquid nitrogen. The lysates were cleared by maximum speed centrifugation at 4°C. To stabilize RNA, 500 $\mu$l of Trizole reagent (Thermo Fisher Scientific) was added to 50-100 $\mu$l of cleared cell lysates, vortex mixed and stored at -80°C. The samples were brought to room temperature and 140 $\mu$l of chloroform was added to each tube, vortex mixed and centrifuged at maximum speed at 4°C for 10 mins. The aqueous fraction was isolated and total RNA was extracted using the RNeasy mini kit (Qiagen), the volume was brought to 900 $\mu$l using RLT, 600 $\mu$l of 95% ethanol was added and mixed in order to bind the RNA. The RNeasy protocol was then followed as recommended by the manufacturer to isolate pure RNA. The ribosomal RNA (rRNA) was depleted using the Ribo-Zero rRNA Removal kit (Epicentre). Strand-specific RNA-seq libraries were prepared using the Stranded RNA-seq Kit (Kapa Biosystems). The libraries were paired-end sequenced with Illumina HiSeq 4000. The sequencing reads were mapped to the *C. ljungdahlii* genome NC_014328 with Bowtie2. FeatureCounts was used to estimate reads per gene. DESeq2 was used to determine differentially expressed genes. RNA-seq values were FPKM-normalized. Reads were deposited to BioSample as

SAMN07391098.

### 2.4.3 Revision of M-model

A previously published M-model, iHN637, was updated to remove obsolete metabolic reactions and to add new reactions to reflect current literature [18, 19, 21]. The *C. ljungdahlii* genome was reannotated using RAST and PROKKA to account for the most recent information and methods in functional annotations [17, 15]. If both start and end sites of ORFs matched that of the original annotation but the functions did not, the new function was also considered during reconstruction of both M- and ME-models. Flux Balance Analysis simulations [27] were carried out as described previously using COBRApy [28]. All M-model simulations maximized growth through the biomass objective function [29].

### 2.4.4 Reconstructing the ME-model

Bidirectional hits and functional overlaps (using RAST annotations) between *Escherichia coli*, *Bacilllus subtilis*, and *C. ljungdahlii*, as well as manual curation of the published annotation, and genome annotations obtained by RAST and PROKKA were used to identify potential E-matrix proteins [17, 16, 15]. Using *E. coli* [9, 10, 20] and *B. subtilis* [15] as reference and the method established by Thiele et al. [20] to fill in missing knowledge, template reactions for the following functions were reconstructed: essential rRNA and tRNA modifications, transcription, translation, translocation, a single bilayer membrane constraint, and Fe-S cluster formation. Transcription units (TU) were predicted using the method established by Lerman et al. [22] and rho independent TUs were predicted using

ARNold [30]. The gene-protein-reactions in iHN673 were converted into protein complexes and updated using Uniprot and PDB annotations as well as functional similarity to *E. coli* and *B. subtilis* proteins [31, 32]. The modeled protein complexes not only contained updated stoichiometry, but also included protein modifications. COBRAme was used to comprise this information into a cohesive model [23]. A complete list of proteins, protein complexes, template reactions, and parameters can be found in supplemental materials.

### 2.4.5 Analyzing the ME-model

Using SoPlex and cobrapy, growth rate was optimized using the binary_search function, as described in COBRAme [23, 28, 33]. All analysis was carried out using python in Jupyter Notebooks, and visualization was provided by matplotlib [34, 35]. Scipy and statsmodels were used for statistical analysis [36, 37]. The code for confidence intervals was taken from [38]. All error bars were 1 standard deviation. In comparing *in vivo* data to *in silico* data, RNA-seq reads from *C. ljungdahlii* grown on fructose [5], $CO_2$+$H_2$ [5], and CO (SAMN07391098) that correspond to 965 modeled ORFs were summed and logged.

## 2.5 Figures and tables

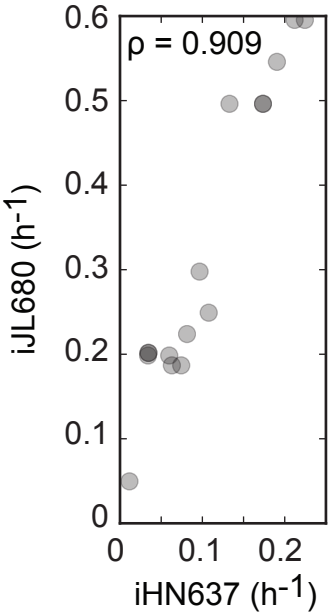



**Figure 2.1**: **Comparing predicted growth rates of iHN637 to iJL680, the updated *C. ljungdahlii* M-model.** Queried substrates and uptake rates are the same from table 1 in [5]

.

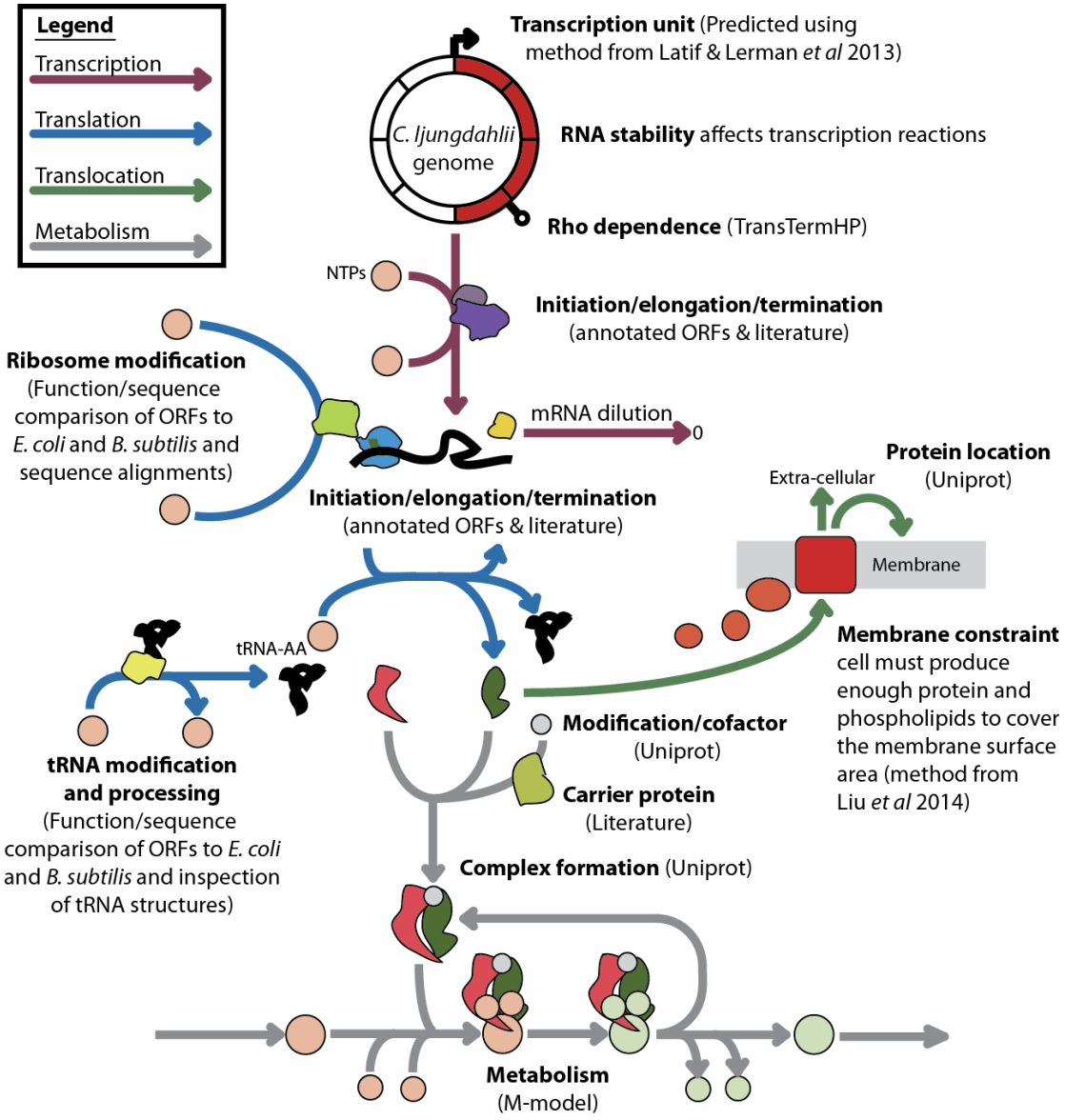



**Figure 2.2**: **Visual representation of the ME-model reconstruction workflow.** The E-matrix reconstruction accounted for transcription, translation, and translocation as well as associated reactions to produce functional enzymes. Integration of the E-matrix (colored arrows) with the M-model (grey arrows) resulted in the ME-model.

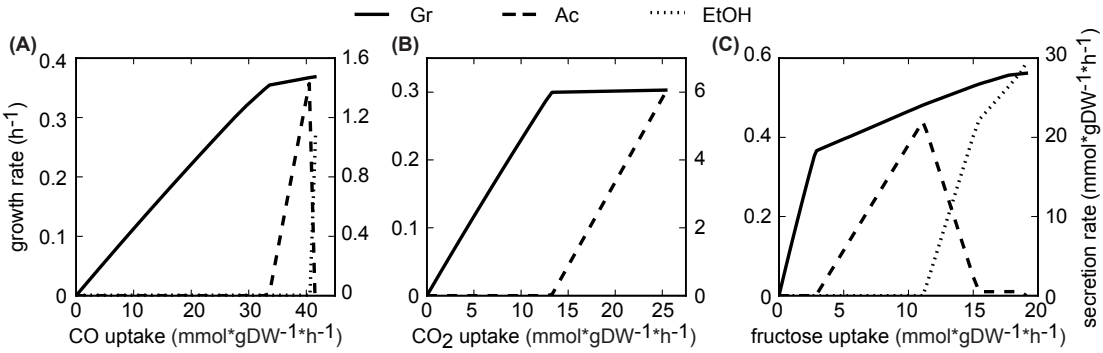



**Figure 2.3**: **Predicted growth rate and yield.** Maximum growth rate, acetate secretion rate, and ethanol secretion rate changed as a function of (A) CO, (B) $CO_2$, and (C) fructose uptake rate on minimal media.

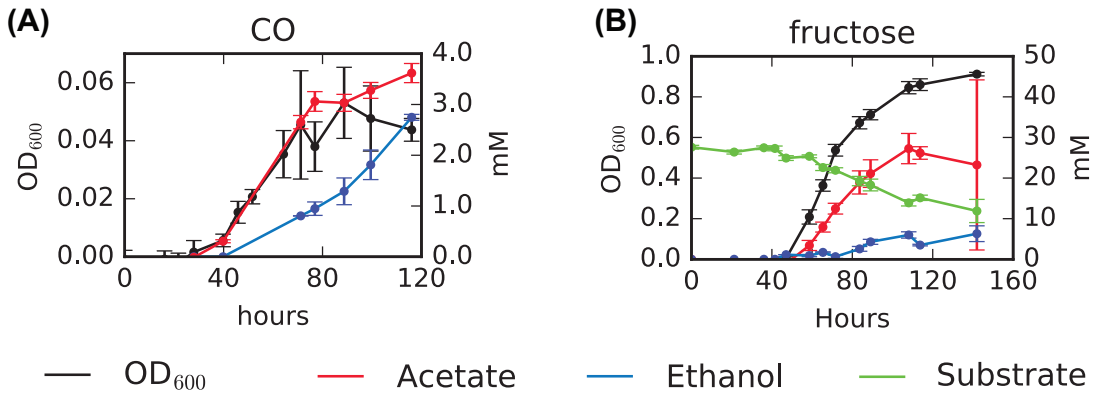



**Figure 2.4**: **Measured growth curves, substrate uptake, and products for CO and fructose conditions.** Growth curves ($OD_{600}$, black line) and HPLC measured molecules were plotted against hours for (A) CO and (B) fructose as carbon source.

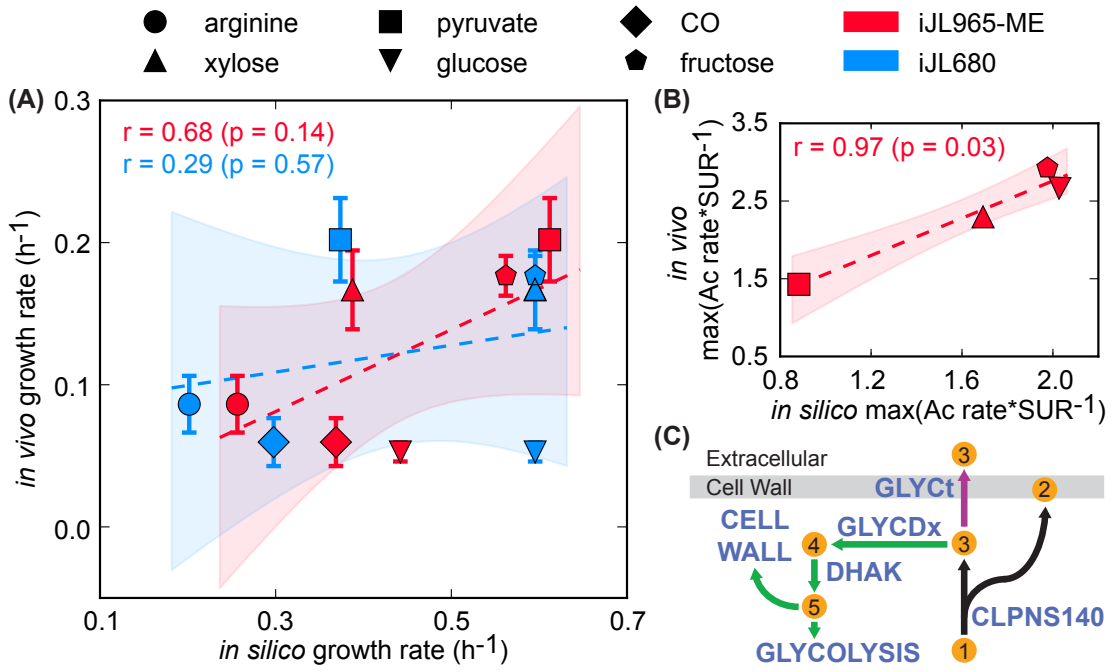



**Figure 2.5**: **Accuracy of growth rate and product predictions.** (A) Two sets of predicted growth rates, from iJL680 and iJL965-ME, were plotted against *in vivo* measured growth rates for six substrates (+- std, n=3). Linear regressions and 95% confidence intervals were represented by dashed lines and shaded areas, respectively. In iJL680, carbon atom uptake was constrained to 30 mmol*gDW$^{-1}$*h$^{-1}$, while in iJL965-ME, the optimal carbon uptake was constrained by inherent proteome limitations. r and p represent Pearson's correlation and p-value. (B) Predicted maximum acetate secretion rate (Ac; mmol*gDW$^{-1}$*h$^{-1}$) to substrate uptake rate (SUR; mmol*gDW$^{-1}$*h$^{-1}$) was plotted against measured averaged values. (C) Predicted pathway mechanism for observed glycerol in spent media. Glycerol was a byproduct of cell membrane formation during cardiolipin production. While the cell was carbon-limited, glycerol was recycled into biomass using the green pathway. When the cell was proteome-limited, *C. ljungdahlii* secreted glycerol (purple arrow). Abbreviations: 1 = phosphatidylglycerol (n-C14:0), 2 = cardiolipin (n-C14:0), 3 = glycerol, 4 = dihydroxyacetone, 5 = dihydroxyacetone phosphate, CLPNS140 = cardiolipin synthase (n-C14:0), GLYCt = glycerol transport, GLYCDx = glycerol dehydrogenase, DHAK = dihydroxyacetone kinase.

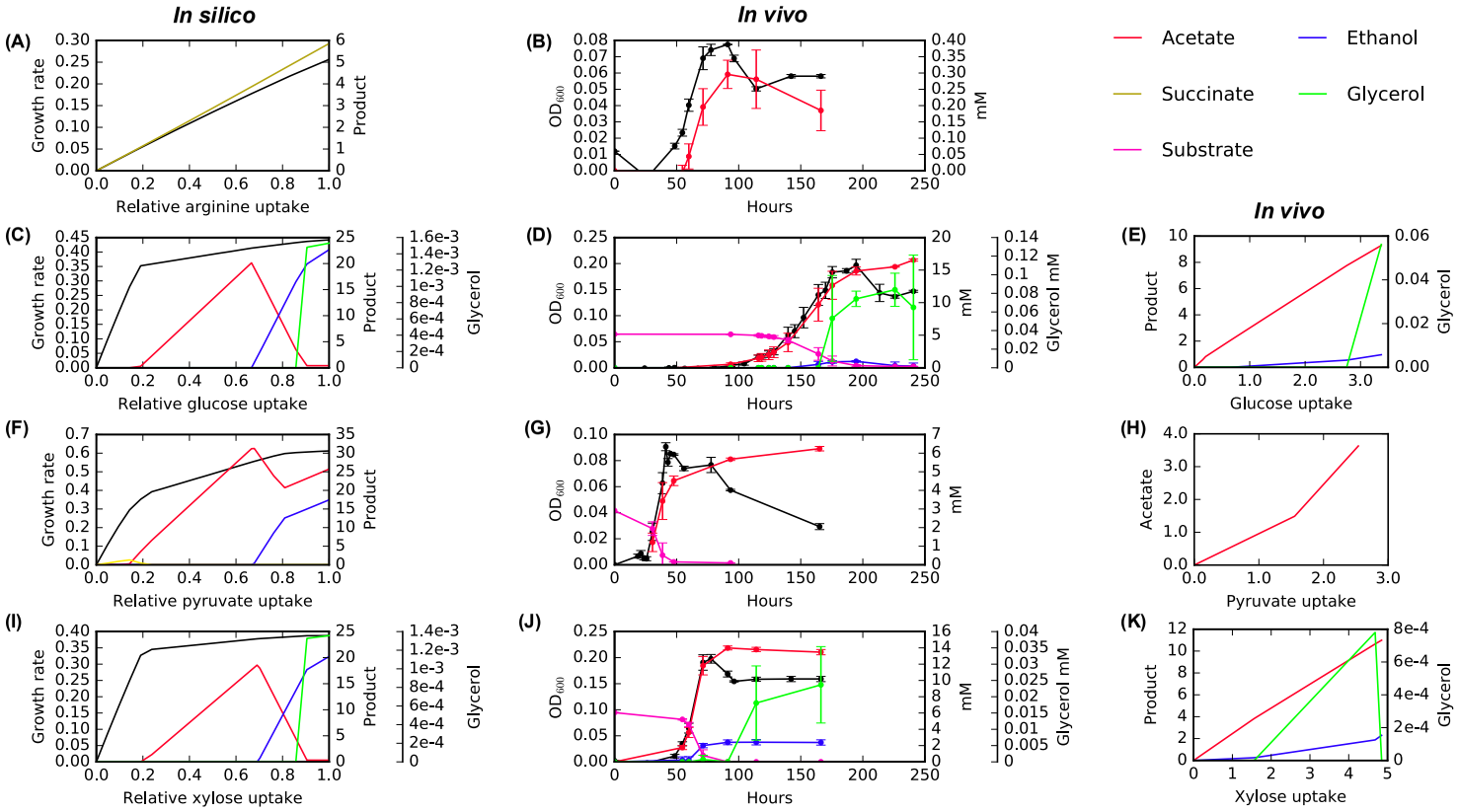



**Figure 2.6**: **Predicted and measured growth curves, substrate uptake rates, and yield.** Predicted growth rate (black line; $h^{-1}$) and secreted products (acetate, ethanol, succinate, glycerol; mmol*$gDW^{-1}$*$h^{-1}$) were plotted against substrate uptake rate (mmol*$gDW^{-1}$*$h^{-1}$) in the left column. Growth curves ($OD_{600}$, black line) and HPLC measured molecules were plotted against hours in the middle column. Rates (mmol*$gDW^{-1}$*$h^{-1}$) were calculated from the middle column using the first derivative of a Savitzky-Golay filter, and products were plotted against carbon uptake rate. Carbon source per row was as follows: (A & B) arginine, (C-E) glucose, (F-H) pyruvate, and (I-K) xylose.

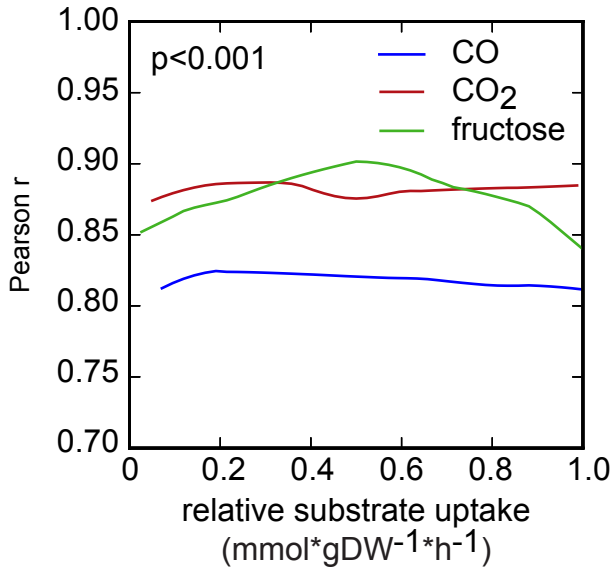



**Figure 2.7**: **Pearson r correlation between categorized and summed predicted transcriptomics.** The Pearson r correlation between categorized and summed predicted transcription flux reactions and RNA-seq data was calculated for discrete substrate uptake rates that ranged from maximum to low uptake rates (n > 30, 0 unfeasible). Relative substrate uptake rate for CO, $CO_2$+$H_2$, and fructose was plotted against the Pearson r.

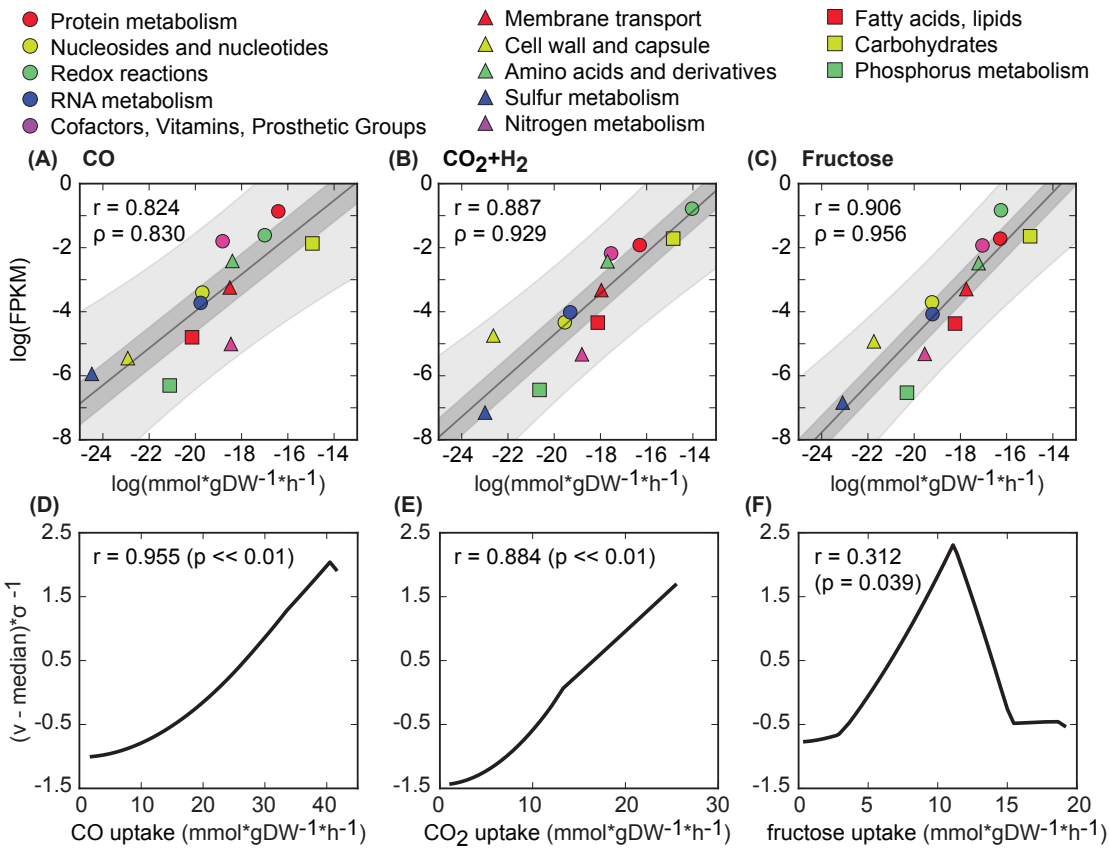



**Figure 2.8**: **Predicted and experimental gene expression.** Categorized by RAST subsystem and summed, predicted gene expression (transcription flux reactions) was compared to RNA-seq data for *C. ljungdahlii* grown on (A) CO, (B) $CO_2$+$H_2$, and (C) fructose. Linear regressions, 95% confidence intervals of the regression, and 95% prediction intervals are represented by lines, dark shaded areas, and light shaded areas respectively. Scatter plots shown are for the highest Pearson r between predicted and experimental data. Normalized total transcription flux (mmol*gDW$^{-1}$*h$^{-1}$) of the Wood-Ljungdahl pathway was plotted against carbon substrate uptake rate for (D) CO, (E) $CO_2$+$H_2$, and (F) fructose, Pearson r reflects correlation with growth rate.

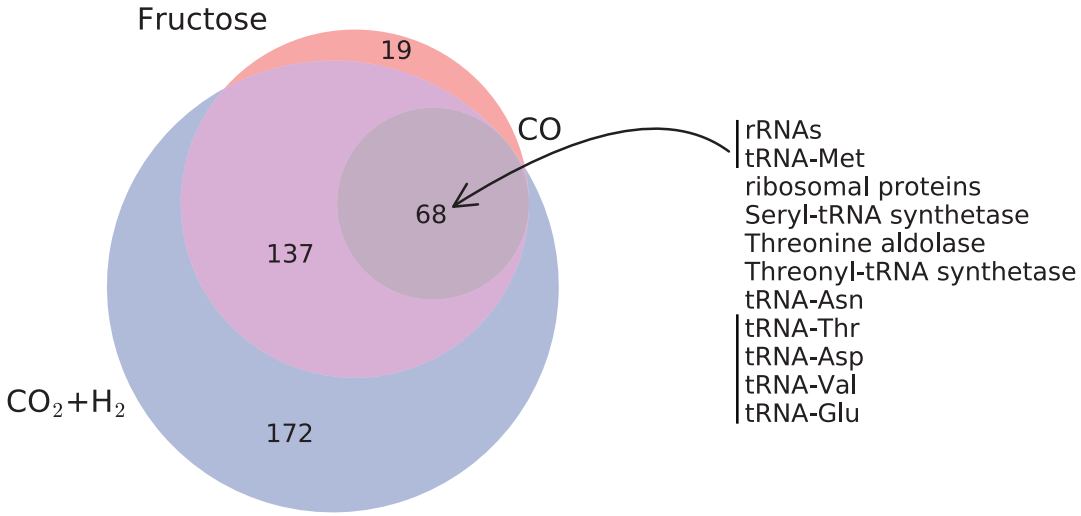



**Figure 2.9**: **Genes highly correlated with growth rate.** Genes that were highly correlated with growth rate more so than substrate uptake rate ($r > 0.9$, p-val $< 0.05$*Bonferonni) were identified, and overlap of genes between the three substrate conditions were plotted in a venn diagram. In the call-out of gene functions shared in all three conditions, the black line indicates that these genes were in the same operon.

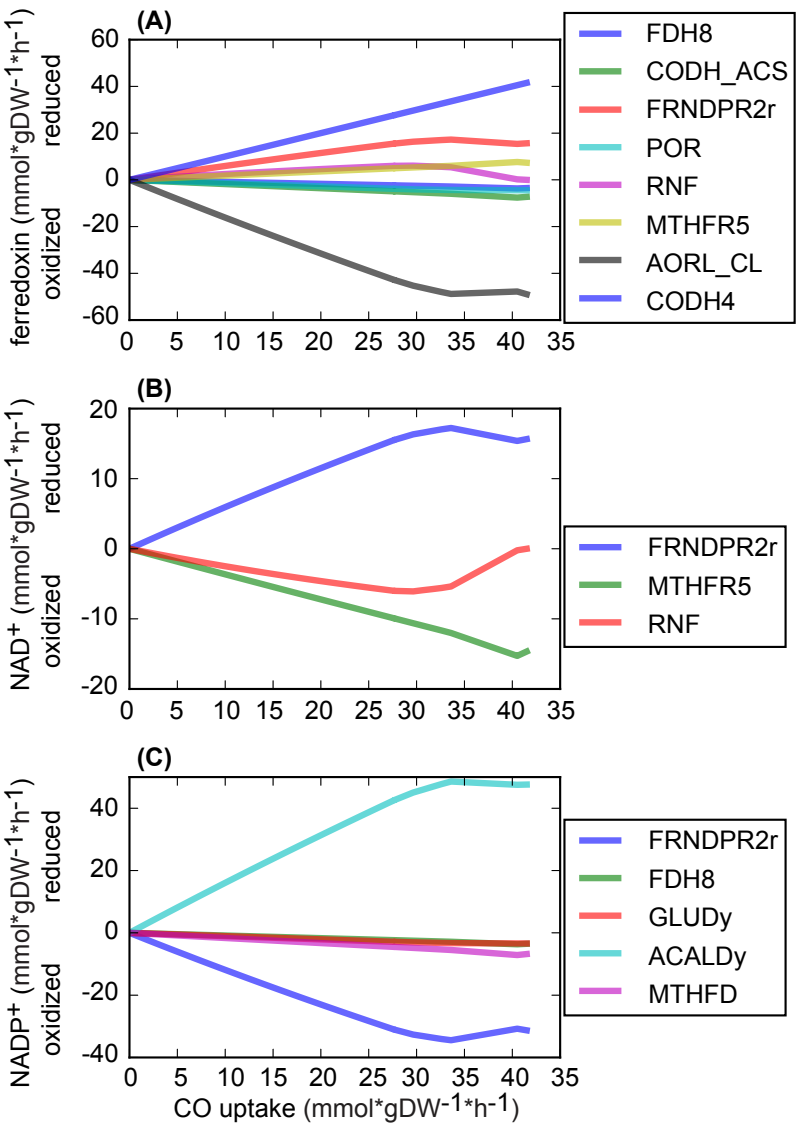



**Figure 2.10**: **Predicted high flux-carrying redox reactions on CO-growth.** Fluxes were plotted against CO uptake rate for reactions involving (A) ferredoxin, (B) $NAD^+$, and (C) $NADP^+$. High flux was defined as the absolute sum of flux across the nutrient spectrum greater than 40 mmol*$gDW^{-1}$*$h^{-1}$. Abbreviations: FDH8 = ferredoxin dehydrogenase, CODH_ACS = carbon monoxide dehydrogenase, FRNDPR2r = ferredoxin:NADP reductase, POR = pyruvate synthase, RNF = ferredoxin:NAD oxidoreductase, = 5,10-methylenetetrahydrofolate reductase (ferredoxin), AOR_CL = acetaldehyde:ferredoxin oxidoreductase, CODH4 = carbon monoxide dehydrogenase, GLUDy = Glutamate dehydrogenase (NADP), ACALDy = acetaldehyde dehydrogenase, MTHFD = methylenetetrahydrofolate dehydrogenase (NADP).

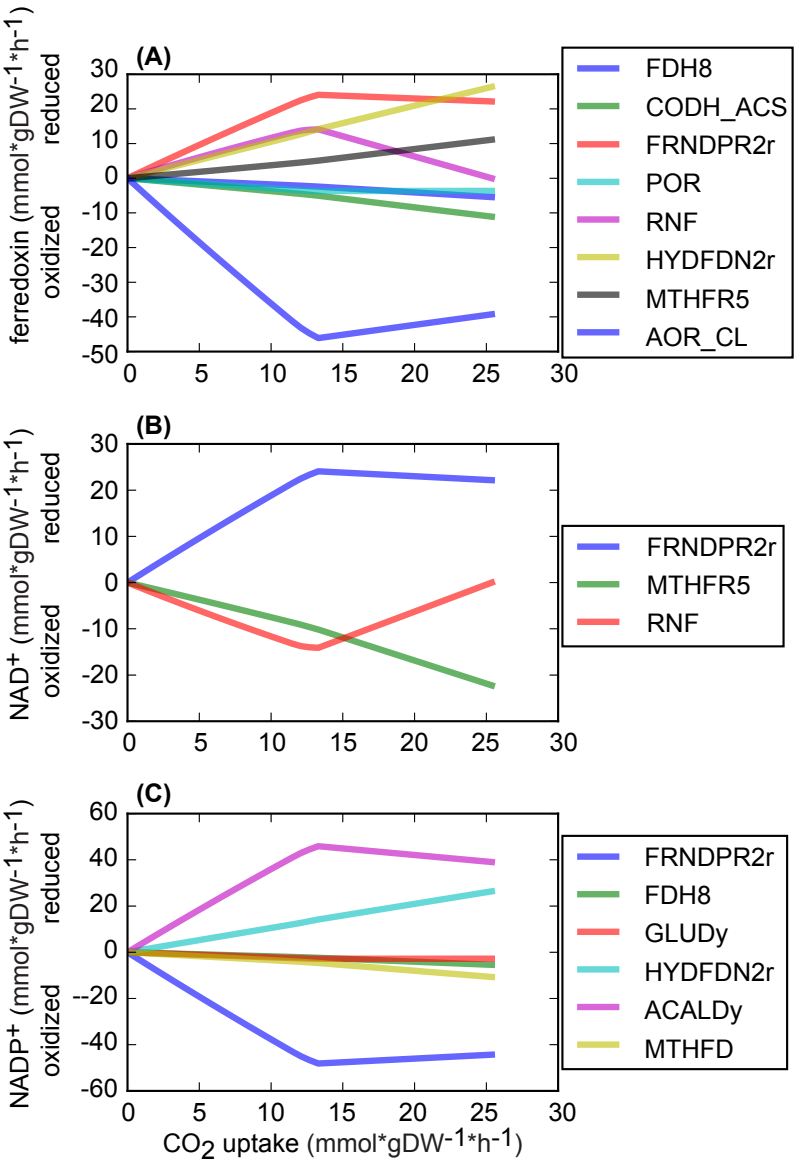



**Figure 2.11**: **Predicted high flux-carrying redox reactions on $CO_2$+$H_2$-growth.** Fluxes were plotted against $CO_2$ uptake rate for reactions involving (A) ferredoxin, (B) $NAD^+$, and (C) $NADP^+$. High flux was defined as the absolute sum of flux across the nutrient spectrum greater than 40 mmol*$gDW^{-1}$*$h^{-1}$. Abbreviations: FDH8 = ferredoxin dehydrogenase, CODH_ACS = carbon monoxide dehydrogenase, FRNDPR2r = ferredoxin:NADP reductase, POR = pyruvate synthase, RNF = ferredoxin:NAD oxidoreductase, = 5,10-methylenetetrahydrofolate reductase (ferredoxin), AOR_CL = acetaldehyde:ferredoxin oxidoreductase, GLUDy = Glutamate dehydrogenase (NADP), ACALDy = acetaldehyde dehydrogenase, MTHFD = methylenetetrahydrofolate dehydrogenase (NADP), HYDFDN2r = ferredoxin NADPH linked hydrogenase.

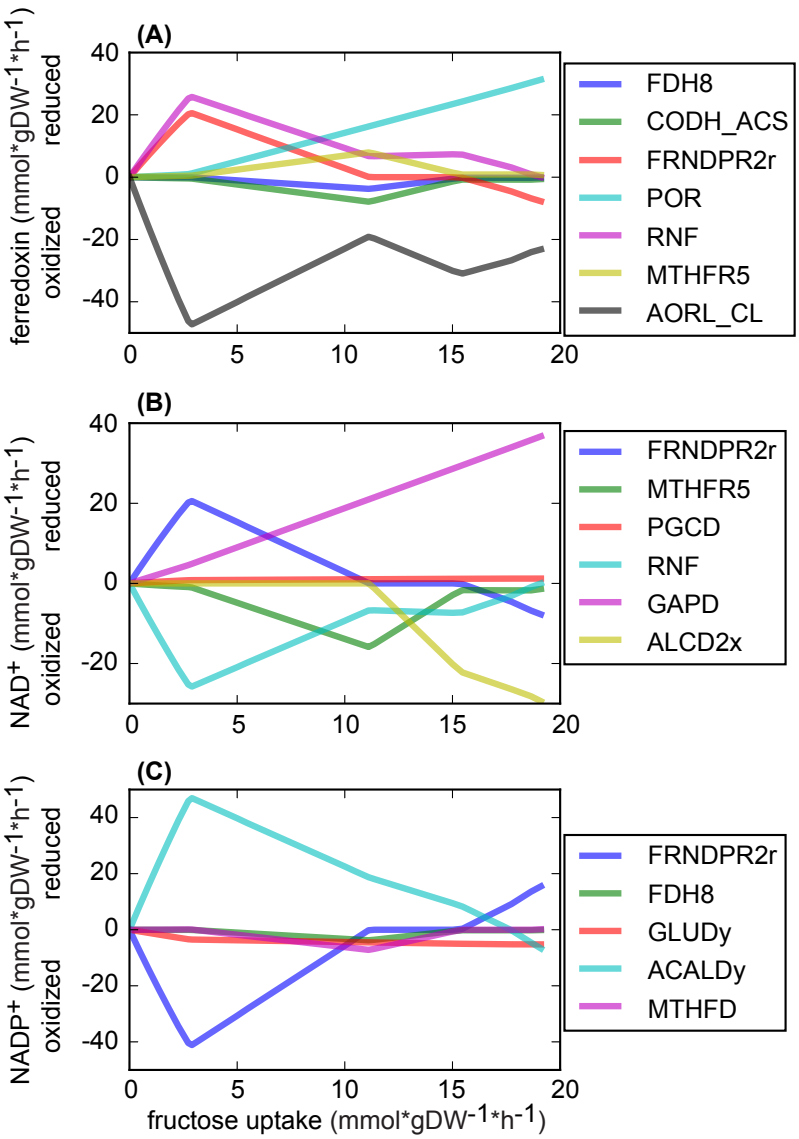



**Figure 2.12**: **Predicted high flux-carrying redox reactions on fructose-growth.** Fluxes were plotted against fructose uptake rate for reactions involving (A) ferredoxin, (B) $NAD^+$, and (C) $NADP^+$. High flux was defined as the absolute sum of flux across the nutrient spectrum greater than 40 mmol*$gDW^{-1}$*$h^{-1}$. Abbreviations: FDH8 = ferredoxin dehydrogenase, CODH_ACS = carbon monoxide dehydrogenase, FRNDPR2r = ferredoxin:NADP reductase, POR = pyruvate synthase, RNF = ferredoxin:NAD oxidoreductase, = 5,10-methylenetetrahydrofolate reductase (ferredoxin), AOR_CL = acetaldehyde:ferredoxin oxidoreductase, GLUDy = Glutamate dehydrogenase (NADP), ACALDy = acetaldehyde dehydrogenase, MTHFD = methylenetetrahydrofolate dehydrogenase (NADP), GAPD = glyceraldehyde-3-phosphate dehydrogenase, PGCD = phosphoglycerate dehydrogenase.

**Table 2.1**: **Names of reactions added or removed from iHN657 to produce iJL680.** See BiGG [39] for more details on individual reactions.

| **New** | **Removed** |
|---|---|
| ACALDy | ACALD |
| BMOCOS | BTDD-RRx |
| BMOGDS | FDH7 |
| BTDD-RRy | HYDFDi |
| CCGt | HYDFDNr |
| CPMPS | |
| DADt | |
| DAPAL | |
| DHORD-NAD | |
| FDH8 | |
| FHL | |
| FMNRx2 | |
| GLCt3 | |
| GLYCLTDx | |
| GLYCTt | |
| LIPATPT | |
| LIPOt2 | |
| LTHRK | |
| METabc | |
| MOCOS | |
| MOGDS | |
| MPTAT | |
| MPTS | |
| PGLYCP | |
| PNTOt | |
| THRt2 | |
| TMDPK | |
| ZN2t | |

## 2.6 Acknowledgements

Chapter 2, in part, is currently being prepared for submission for publication of the material. Joanne Liu, Ali Ebrahim, Mahmoud Al Bassam, Colton Lloyd, Ji-Nu Kim, Connor Olson, and Karsten Zengler. A systems biology approach to investigate proteome control of acetate and ethanol production in *Clostridium ljungdahlii* (working title). The dissertation author was the primary investigator and author of this material. We are also thankful to Cameron Martino, Kristine Ly, and Kevin Tang for assisting with growth experiments, and to Nathan Lewis, Cristal Zuñiga and Lívia Zaramela for fruitful discussions and input. Without them, this study would not have been possible.

## 2.7 References

[1] Bettina Schiel-Bengelsdorf and Peter Dürre. Pathway engineering and synthetic biology using acetogens. *FEBS Letters*, 586(15):2191–2198, 2012.

[2] Haythem Latif, Ahmad A Zeidan, Alex T Nielsen, and Karsten Zengler. Trash to treasure: production of biofuels and commodity chemicals via syngas fermenting microorganisms. *Current Opinion in Biotechnology*, 27:79–87, 2014.

[3] Kirsten Küsel and Harold L. Drake. Acetogens. pages 1–5. Springer Netherlands, 2011.

[4] Kai Schuchmann and Volker Müller. Autotrophy at the thermodynamic limit of life: a model for energy conservation in acetogenic bacteria. *Nature Reviews Microbiology*, 12(12):809–821, 2014.

[5] Harish Nagarajan, Merve Sahin, Juan Nogales, Haythem Latif, Derek R Lovley, Ali Ebrahim, and Karsten Zengler. Characterizing acetogenic metabolism

using a genome-scale metabolic reconstruction of *Clostridium ljungdahlii*. *Microbial cell factories*, 12(1):118, 2013.

[6] Kaspar Valgepea, Renato de Souza Pinto Lemgruber, Kieran Meaghan, Robin William Palfreyman, Tanus Abdalla, Björn Daniel Heijstra, James Bruce Behrendorff, Ryan Tappel, Michael Köpke, Séan Dennis Simpson, Lars Keld Nielsen, and Esteban Marcellin. Maintenance of ATP Homeostasis Triggers Metabolic Shifts in Gas-Fermenting Acetogens. *Cell Systems*, 4(5):505–515.e5, 2017.

[7] Kaspar Valgepea, Kim Q. Loi, James B. Behrendorff, Renato de S.P. Lemgruber, Manuel Plan, Mark P. Hodson, Michael Köpke, Lars K. Nielsen, and Esteban Marcellin. Arginine deiminase pathway provides ATP and boosts growth of the gas-fermenting acetogen *Clostridium autoethanogenum*. *Metabolic Engineering*, 41:202–211, 2017.

[8] M. Ahsanul Islam, Karsten Zengler, Elizabeth A. Edwards, Radhakrishnan Mahadevan, Gregory Stephanopoulos, D. Horsman, S. J. Jones, M. A. Marra, N. Lewis, S. Rahmanian, J. Kang, D. Hyduke, B. Palsson, N. Ivanova, N. Kyrpides, B. Department Of, R. U. H. T. Cell Biology, Y. I. Wolf, R. L. Tatusov, F. Sabathe, L. Doucette-Stamm, P. Soucaille, M. J. Daly, G. N. Bennett, E. V. Koonin, and D. R. Smith. Investigating *Moorella thermoacetica* metabolism with a genome-scale constraint-based metabolic model. *Integr. Biol.*, 7(8): 869–882, 2015.

[9] Joanne K Liu, Edward J O'Brien, Joshua A Lerman, Karsten Zengler, Bernhard O Palsson, and Adam M Feist. Reconstruction and modeling protein translocation and compartmentalization in Escherichia coli at the genome-scale. *BMC Systems Biology*, 8(1):110, 2014.

[10] E. J. O'Brien, J. A. Lerman, R. L. Chang, D. R. Hyduke, and B. O. Palsson. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Molecular Systems Biology*, 9(1):693–693, 2014.

[11] John T. Heap, Sarah A. Kuehne, Muhammad Ehsaan, Stephen T. Cartman, Clare M. Cooksley, Jamie C. Scott, and Nigel P. Minton. The ClosTron: Mutagenesis in Clostridium refined and streamlined. *Journal of Microbiological Methods*, 80(1):49–55, 2010.

[12] Ching Leang, Toshiyuki Ueki, Kelly P Nevin, and Derek R Lovley. A genetic system for Clostridium ljungdahlii: a chassis for autotrophic production of biocommodities and a model homoacetogen. *Applied and environmental microbiology*, 79(4):1102–9, 2013.

[13] Areen Banerjee, Ching Leang, Toshiyuki Ueki, Kelly P Nevin, and Derek R Lovley. Lactose-inducible system for metabolic engineering of Clostridium ljungdahlii. *Applied and environmental microbiology*, 80(8):2410–6, 2014.

[14] Nigel P. Minton, Muhammad Ehsaan, Christopher M. Humphreys, Gareth T. Little, Jonathan Baker, Anne M. Henstra, Fungmin Liew, Michelle L. Kelly, Lili Sheng, Katrin Schwarz, and Ying Zhang. A roadmap for gene system development in Clostridium. *Anaerobe*, 41:104–112, 2016.

[15] Scott A Becker, Bernhard Ø Palsson, Aaron A Best, Matthew DeJongh, Terrence Disz, Robert A Edwards, Kevin Formsma, Svetlana Gerdes, Elizabeth M Glass, Michael Kubal, Folker Meyer, Gary J Olsen, Robert Olson, Andrei L Osterman, Ross A Overbeek, Leslie K McNeil, Daniel Paarmann, Tobias Paczian, Bruce Parrello, Gordon D Pusch, Claudia Reich, Rick Stevens, Olga Vassieva, Veronika Vonstein, Andreas Wilke, and Olga Zagnitko. The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Microbiology*, 5(1):8, 2005.

[16] Michael Köpke, Claudia Held, Sandra Hujer, Heiko Liesegang, Arnim Wiezer, Antje Wollherr, Armin Ehrenreich, Wolfgang Liebl, Gerhard Gottschalk, and Peter Dürre. *Clostridium ljungdahlii* represents a microbial production platform based on syngas. *Proceedings of the National Academy of Sciences of the United States of America*, 107(29):13087–92, 2010.

[17] T. Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, 2014.

[18] Johanna Mock, Yanning Zheng, Alexander P. Mueller, San Ly, Loan Tran, Simon Segovia, Shilpa Nagaraju, Michael Köpke, Peter Dürre, and Rudolf K. Thauer. Energy Conservation Associated with Ethanol Formation from H2 and CO2 in *Clostridium autoethanogenum* Involving Electron Bifurcation. *Journal of Bacteriology*, 197(18):2965–2980, 2015.

[19] Yang Tan, Zi-Yong Liu, Zhen Liu, and Fu-Li Li. Characterization of an acetoin reductase/2,3-butanediol dehydrogenase from *Clostridium ljungdahlii* DSM 13528. *Enzyme and Microbial Technology*, 79-80:1–7, 2015.

[20] Ines Thiele, Neema Jamshidi, Ronan M. T. Fleming, Bernhard Ø. Palsson, and P Stothard. Genome-Scale Reconstruction of *Escherichia coli's* Transcriptional and Translational Machinery: A Knowledge Base, Its Mathematical Formulation, and Its Functional Characterization. *PLoS Computational Biology*, 5(3):e1000312, 2009.

[21] Ines Thiele and Bernhard Ø Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1):93–121, 2010.

[22] Joshua A. Lerman, Daniel R. Hyduke, Haythem Latif, Vasiliy A. Portnoy, Nathan E. Lewis, Jeffrey D. Orth, Alexandra C. Schrimpe-Rutledge, Richard D. Smith, Joshua N. Adkins, Karsten Zengler, and Bernhard O. Palsson. In silico method for modelling metabolism and gene product expression at genome scale. *Nature Communications*, 3:929, 2012.

[23] Colton J Lloyd, Ali Ebrahim, Laurence Yang, Zachary Andrew King, Edward Catoiu, Edward J O'Brien, Joanne K Liu, and Bernhard O Palsson. COBRAme: A Computational Framework for Building and Manipulating Models of Metabolism and Gene Expression. *bioRxiv*, 2017.

[24] Sandra Placzek, Ida Schomburg, Antje Chang, Lisa Jeske, Marcus Ulbrich, Jana Tillack, and Dietmar Schomburg. BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Research*, 45(D1):D380–D388, 2017.

[25] Shawn W Jones, Alan G Fast, Ellinor D Carlson, Carrissa A Wiedel, Jennifer Au, Maciek R Antoniewicz, Eleftherios T Papoutsakis, and Bryan P Tracy. CO2 fixation by anaerobic non-photosynthetic mixotrophy for improved carbon conversion. *Nature communications*, 7:12800, 2016.

[26] Vasiliy A Portnoy, Markus J Herrgård, and Bernhard Ø Palsson. Aerobic fermentation of D-glucose by an evolved cytochrome oxidase-deficient *Escherichia coli* strain. *Applied and environmental microbiology*, 74(24):7561–9, 2008.

[27] Adam M Feist and Bernhard O Palsson. The biomass objective function. *Current Opinion in Microbiology*, 13(3):344–349, 2010.

[28] Ali Ebrahim, Joshua A Lerman, Bernhard O Palsson, and Daniel R Hyduke. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Systems Biology*, 7(1):74, 2013.

[29] Adam M Feist and Bernhard O Palsson. The biomass objective function. *Current Opinion in Microbiology*, 13(3):344–349, 2010.

[30] Magali Naville, Adrien Ghuillot-Gaudeffroy, Antonin Marchais, and Daniel Gautheret. ARNold: A web tool for the prediction of Rho-independent transcription terminators. *RNA Biology*, 8(1):11–13, 2011.

[31] UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, 2017.

[32] H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The Protein Data Bank. *Nucleic acids research*, 28(1):235–42, 2000.

[33] Roland Wunderling. Paralleler und Objektorientierter Simplex. 1996.

[34] Fernando Pérez and Brian E. Granger. IPython: A System for Interactive Scientific Computing Python. *Computing in Science and Engineering*, 9(3): 21–29, 2007.

[35] John D. Hunter. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[36] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and Statistical Modeling with Python. *PROC. OF THE 9th PYTHON IN SCIENCE CONF*, 2010.

[37] Eric Jones, Travis Oliphant, and Pearu Peterson. SciPy: Open source scientific tools for Python. 2001.

[38] http://markthegraph.blogspot.com/2015/05/using-python-statsmodels-for-ols-linear.html.

[39] Jan Schellenberger, Junyoung O Park, Tom M Conrad, and Bernhard Ø Palsson. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC bioinformatics*, 11(1):213, 2010.

# Chapter 3

# Exploring the evolutionary significance of tRNA operon structure using metabolic and gene expression models

## 3.1 Introduction

An operon is a co-regulated cluster of genes that are expressed on the same RNA transcript. These genomic features arise through a variety of means, including horizontal gene transfer that places a gene under another gene's promoter, horizontal gene transfer of whole operons, deletion of intervening sequences, and genome rearrangement [1]. Though the presence of an operon may be a random event, selection pressures can drive the maintenance of operons. For example, po-

tential benefits bestowed by an operon onto the host organism include a reduction in regulation costs [2], diminished stochastic gene expression through synchronicity of protein ratios [3, 4], and insurance that all functional steps in a pathway are produced [5]. Such theories hint at an evolutionary optimization problem to promote efficiency in gene expression.

In order to optimize cellular efficiency, translation must be carefully controlled because it requires the highest energy and resource expenditure of any process in fast-growing cells. Since the available tRNA pool could be rate-limiting during protein translation [6], close correspondence between codon usage and the available tRNA pool, often quantified through the tRNA adaptation index (tAI) [7], must be maintained efficiently. Even though tRNA co-expression explained *E. coli's* tRNA profile better than tRNA gene copy number (widely recognized as a correlated estimate for tRNA profile [8, 9, 10]), relatively few papers have investigated the influence of operons on tRNA expression levels [11, 12]. Yet rRNA and tRNA genes can often be found on the same operon, and 23.8% of all tRNA genes from prokaryotic genomes sequenced by 2014 were found to be located in an operon with another tRNA gene [12]. Such evidence implies that evolutionary pressures may also shape genomic tRNA structure.

Constraint-based modeling offers a biophysically-based approach to estimate tRNA concentrations and usage. In particular, constraint-based metabolic and gene expression models (*i.e.*, ME-models) are well-suited for examining potential insights into operon structure. The scope of predictions that ME-models cover are extensive; these models account for transcription, tRNA charging, translation, and metabolic reactions. Additionally, ME-models incorporate the underlying

genome architecture through transcriptional units that account for co-expression of genes. ME-models have been used to successfully recapitulate several levels of phenotypes, from growth rates to pathway expression levels, and even undiscovered operons [13, 14, 15]. As of writing, only *E. coli* and *C. ljungdahlii* have completed ME-models that use the COBRAme framework, which allowed comparisons of model perturbations with the knowledge that the constraints within the models (e.g., coupling constraints) were similarly formulated [16, 17].

Using the two available COBRAme-based ME-models, one for *Escherichia coli* and one for *Clostridium ljungdhalii* [16, 17], we examined the systematic importance of tRNA co-expression. We validated the two models for the purposes of this study and examined the tRNA operon structures, thereby identifying unique tRNA operon solutions to two different selective pressures. One solution led to optimization of phenotype through fragmenting operons and the other solution to optimized efficiency through optimal grouping of tRNAs.

## 3.2 Results and Discussion

### 3.2.1 tRNA operon structure: Fragmentation versus modularity

Examination of tRNA-containing operons organization in two bacteria, the fast-growing generalist *E. coli* and the slower-growing homoacetogen *C. ljungdahlii*, revealed two different strategies (Fig 3.1 & 3.2) [18]. These two strategies will be referred to as fragmentation, where tRNA organization leads to both a high number of singly-transcribed tRNA genes and a minimization of co-transcribed

tRNA species, and modularization, which is the tendency towards polycistronic tRNA genes.

In *E. coli*, 23% of tRNA genes could be transcribed monocistronically, and 37% could be expressed as polycistronic transcripts that lack other tRNA genes. When considering unique tRNA species by anticodon, the number of single transcripts that can be uniquely expressed increased to 54%, and for tRNA species by amino acid (AA), 56%. Furthermore, *E. coli* appeared to favor less tRNA genes per transcript and did not have an operon containing more than seven tRNAs, while the highest number of unique tRNA species per operon was four. Thus, *E. coli* displays a fragmentation strategy for its tRNA operon structure (blue bars, Fig 3.3).

In case of *C. ljungdahlii*, the analysis revealed that only 8.4% of tRNA genes could be expressed monocistronically and 26% could be expressed as polycistronic transcripts lacking other tRNA genes. Looking at tRNA species, only 32% of tRNAs by anticodon and 34% of tRNAs by AA were capable of being uniquely expressed on a single transcript. Thus, *C. ljungdahlii* had the majority of its tRNA species co-transcribed with another tRNA type, and *C. ljungdahlii* could express fifteen tRNAs, including the only tRNA-his gene, on a single transcript. The bias towards polycistronic tRNA genes means that *C. ljungdahlii* prefers modularization in comparison to *E. coli* (green bars, Fig 3.3).

### 3.2.2 Predicted tRNA charging amino acid usage is consistent with amino acid requirements

AA compositions predicted by the *E. coli* ME-model (iLE1678-ME) and the *C. ljungdahlii* ME-model (iJL965-ME) were compared against *in vivo* data. AA composition was calculated from transcriptomic data using RNA-seq (FPKM) data from *E. coli* batch-grown on glucose, glycerol, xylose, and acetate and *C. ljungdahlii* batch-grown on fructose, CO and $CO_2$+$H_2$ as a proxy for protein count. Only proteins reconstructed in the ME-models were considered. For each substrate condition, the ME-models were simulated at maximum growth rate (which was calculated when substrate availability was greater than what can be consumed, and considered to be equivalent to *in vivo* batch growth), half of the maximum substrate uptake rate, and minimal substrate availability (*i.e.*, tenth of maximum substrate uptake rate). Predicted AA compositions were calculated from tRNA charging reactions (mmol*gDW$^{-1}$*h$^{-1}$) which reflects the exact AA requirements of the *in silico* cell.

The predicted and measured AA compositions were highly comparable ($R^2 \geq 0.964$ for all batch-growth conditions in both models; Fig 3.4, 3.5). The high correlation between *in silico* and *in vivo* values continued to hold true for tRNA molecule concentrations (uM) and calculated AA composition from protein expression (ribosome profiling, RPKM) in *E. coli*, both of which were more appropriate comparisons for *in silico* tRNA expression and tRNA charging reactions (Fig 3.6, 3.7). With these validations for AA composition and our knowledge of the genome architecture, we have confidence in the output of translation and the

underlying structure of transcription in the ME-models for batch conditions.

The goodness of fit decreased when *in vivo* batch-grown cells were compared to *in silico* growth on half of the maximum substrate uptake rate and minimal substrate availability. Thereby iLE1678-ME and iJL965-ME demonstrated their capability to predict variable AA compositions dependent on substrate availability. Furthermore, expression values from *in silico* minimal and half substrate availability were able to explain tRNA molecule concentrations in low growth rate ($0.4\ h^{-1}$) better than *in silico* maximum growth rate could (Fig 3.6). Although the higher correlations imply that ME-models continue to be accurate at lower growth rates, the actual influence of growth rate on tRNA pools is currently inconclusive and requires more investigation [8, 11].

Despite the lack of evidence to support conclusions from non-optimal growth rates, ME-models still provide an opportunity to specifically examine the effects of varying tRNA operon structure.

### 3.2.3 Optimized tRNA operon structure meets tRNA abundance requirements

To examine whether tRNA gene location and co-transcription influences the cell, 1000 models with all tRNAs randomly shuffled into another tRNA's location, henceforth referred to as Monte-Carlo (MC) tRNA location models, were built for *E. coli* and *C. ljungdahlii* each. The MC tRNA location models were then simulated with validated substrates (glucose, glycerol, xylose, and acetate for *E. coli* and fructose, CO, and $CO_2$+$H_2$ for *C. ljungdahlii*) at maximum growth rate, half of the maximum substrate uptake rate, and minimal substrate uptake (Fig 3.8).

With this setup, we can examine whether the two organisms' different tRNA organization strategies, fragmentation and modularization, promote optimization for translational purposes under particular growth conditions.

Shuffling tRNA order and location has a dramatic effect on tRNA expression, as the range of AA-categorized tRNA (tRNA-AA) expression can vary drastically in relation to other tRNA-AA molecules (Fig 3.9). When tRNA-AA expressions of the MC tRNA location models were compared against the original models' (iLE1678-ME and iJL965-ME which contain published genome architectures), tRNA expression was revealed to be minimized. Both iLE1678-ME and iJL965-ME performed better than the median MC tRNA location model because they expressed less total tRNA for a significant number of tRNA-AA molecules ($p < 0.02$ for all maximum growth rate conditions; Fig 3.9, Table 3.1, 3.2). Thus, the original tRNA operon structures led to reduced cost of tRNA expression.

In contrast to the flux ranges of tRNA expression, the AA composition of the cell, as represented by tRNA charging reactions, remains relatively constant. Regardless, iLE1678-ME and iJL965-ME revealed that the published tRNA operon structures also promoted utilization of tRNA usage (*i.e.*, tRNA charging reactions) at maximum growth rate. For a significant number of tRNA-AA molecules, iLE1678-ME and iJL965-ME used more tRNA in tRNA charging reactions than the median MC tRNA location model ($p < 0.05$ for all conditions but two; Fig 3.9, Table 3.1, 3.2). *E. coli* on acetate and *C. ljungdahlii* on fructose were the exceptions, as tRNA expression was minimized, but tRNA usage was not maximized (Table 3.1, 3.2). Thus, the original tRNA operon structures generally led to increased tRNA usage.

If tRNA expression could be likened to capital costs and tRNA usage to operating costs, then *E. coli* and *C. ljungdahlii* have minimized capital costs by optimizing expression of necessary tRNAs. The operating costs have likewise been maximized, even though tRNA operon structure does not influence operating costs as strongly as it does capital costs, as seen through the lack of fluctuation in tRNA usage and the non-optimal tRNA usage in acetate-grown iLE1678-ME. Together, these observations suggest that the cells partly control their capital expenses at maximum growth rate though tRNA operon structure.

At least half of the tRNA-AA molecules in the original models have both lower expression and higher usage than the median MC tRNA location model (*i.e.*, tRNA-AA optimization) at maximum growth in multiple substrate conditions (Fig 3.9, Table 3.3, 3.4). However, *E. coli* and *C. ljungdahlii* did not optimize the same tRNA-AA molecules, with only F, G, K, M, and Y being shared between the two models, thereby showing that optimized tRNA-AA molecules may differ by organism.

Both iLE1678-ME and iJL965-ME displayed less efficient tRNA expression and tRNA charging usage as growth rate dropped from maximum, and they were no longer efficient at minimum growth rate (Table 3.1, 3.2) with the exception of $CO_2$+$H_2$, implying that tRNA operon structures have been optimized for growth when nutrients were abundant. The number of optimized tRNA-AA molecules also decreased with growth rate (Table 3.3, 3.4). *E. coli* on xylose and *C. ljungdahlii* stood out as retaining the most optimized number of tRNA-AA molecules with 9 AAs and 7 AAs respectively. Perhaps this optimization of tRNA-AA molecules for lower growth rate inducing substrates ($gr_{glucose} = 0.92$ vs $gr_{xylose} = 0.87$; $gr_{CO}$

= 0.38 vs $gr_{CO_2+H_2}$ = 0.31) hints at an evolutionary process that ensured continued resource efficiency in less desirable conditions once preferred substrates are depleted.

### 3.2.4 Positive selection for high tRNA efficiency

Despite a trend towards minimization in capital expenses, iLE1678-ME (*E. coli*) performed at an average in total tRNA efficiency, as measured by the total tRNA usage to total tRNA expression ratio, compared to the MC tRNA location models (Fig 3.10e). Its maximum growth rate was also average (Fig 3.10g). However, when the range of tRNA efficiency values and growth rates of the MC tRNA location models were compared against *C. ljungdahlii's* ranges, *E. coli* has evolved to minimize the potential error around tRNA efficiency, rRNA expression, and growth rate (Fig 3.10). Fragmentation of the operon structure ensured that regardless of tRNA order or location, potential phenotypes cannot deviate too far from the original value (Fig 3.10a, b), which may reflect a history of tRNA genes being regularly added and subtracted from the genome to reach its current, optimal state [19, 12].

The only non-random gene locations in tRNA-containing operons were occupied by rRNA genes, which refers to the set of 16S, 5S, and 23S rRNAs. In iLE1678-ME, all seven rRNA gene sets were co-expressed with tRNA genes, and rRNA expression was driven, in part, by the need for the associated tRNA genes. All three of the tRNAs with anticodon UGC, which codes for tRNA-ala, were on a polycistronic transcript with an rRNA gene set (Fig 3.1). Since alanine was the most required AA, iLE1678-ME subsequently expressed a significant amount of

rRNA genes at maximum growth rate (Fig 3.10f, 3.11). The selective maximization of rRNA expression points at growth rate optimization in *E. coli*, as ribosome amount is linearly correlated to growth rate [20].

While *E. coli* has been optimized for output, particularly rRNA production, *C. ljungdahlii* seemed to be focused on minimizing capital expenditures, as demonstrated by the significantly high tRNA efficiency in iJL965-ME which remained high even as growth rate dropped (Fig 3.12), while both growth rate and rRNA expression were average compared to the MC tRNA location models (Fig 3.10). However, average rRNA expression may also point to efficient resource usage. Unlike rRNA arrangement in *E. coli*, seven of iJL965-ME's nine rRNA gene sets were co-expressed with tRNAs. Furthermore, *C. ljungdahlii* does not associate a specific tRNA species with rRNA, which allowed *C. ljungdahlii* the ability to fine tune its rRNA need by expressing operons with the necessary amount of tRNAs per species, thereby minimizing resources spent on producing more rRNA, while *E. coli* has evolved so that an abundant amount of rRNA is available for maximum growth rate. Finally, unlike *E. coli's* tight range of values, shuffling of tRNA locations would lead to drastic changes in tRNA efficiency, rRNA expression, and growth rate. Thus, in contrast to *E. coli's* fragmentation, modularization in *C. ljungdahlii* sacrificed growth rate for tRNA efficiency and resource frugality.

Although tRNA efficiency, rRNA expression, and growth rate were not correlated ($R^2 \leq 0.176$, Fig 3.10a-d), there were operon structures that resulted in higher growth rates. This may not be so important for *E. coli*, since its range of potential growth rates was limited, and the payoff between tRNA efficiency and growth rate was low (slope of the upper soft boundary, $m_{E.coli}$=0.010;

Fig 3.10b). *C. ljungdahlii* could potentially improve growth rate by 1.5% solely through tRNA operon rearrangement, but the majority of organizations were less efficient (Fig 3.10d), which means that the organism will grow faster at a cost to translational efficiency ($m_{C.ljungdahlii} = 0.016$; Fig 3.10d).

### 3.2.5 Does tRNA operon structure reflect K/r strategists?

Fragmentation and modularization may hint at a deeper understanding of the differences between K- and r-strategists, where K-strategists are typically associated with slow growth due to limitations by density-dependent controls, and r-strategists with fast growth (Note: K and r strategists can be differentiated by their maximum specific growth rate under conditions with excess substrate (*i.e.*, batch growth) [21]). *C. ljungdahlii*, as a K-strategist (max *in silico* growth rate on fructose is 0.57 $h^{*-1}$), evolved to maximize efficiency of resources at the tRNA operon structure level. Thus, *C. ljungdahlii* matches cost to need, which may provide *C. ljungdahlii* with a slight edge over competitors when nutrients are limiting for the ecological community. *E. coli*, an r-strategist (max *in silico* growth rate on glucose is 0.92 $h^{*-1}$), has evolved to always perform near optimum in regards to its tRNA operon structure, and rRNA expression, which is tied to tRNA expression, is maximized. Furthermore, *E. coli's* fractured tRNA-containing operon structure may allow *E. coli* to quickly match tRNA-demands specific to available substrates, as *E. coli* is a generalist that consumes multiple carbon sources. Thus, *E. coli* has optimized its output, which may allow it to persist in an ecological community through rapid growth.

## 3.3 Conclusions

Although ME-models currently lack other factors that affect tRNA amounts (*e.g.*, regulation, proximity to the origin of replication, leading verses lagging strand, individualized aminoacyl-synthase turnovers [11]), ME-models account for genome architecture (gathered from publicly available databases), transcription, tRNA charging, and translation (validated through a combination of 'omics and Northern blot data), which allowed us the ability to interrogate the importance of tRNA operon structure for two organisms, *E. coli* and *C. ljungdahlii*. Examination of these two organisms' operon structures revealed two different strategies: Fragmentation in *E. coli* and modularization in *C. ljungdahlii*. Using iLE1678-ME (*E. coli*) and iJL965-ME (*C. ljungdahlii*) as a basis, 1000 models with randomly shuffled tRNA locations for each organism were built. Predictions from these MC tRNA location models compared to those from iLE1678-ME or iJL965-ME showed that tRNA operon structure was optimized for tRNA abundance requirements. In iLE1678-ME, the tRNA operon structure also lead to high rRNA expression, while in iJL965-ME, tRNA efficiency was optimized. these conclusions regarding optimization primarily hold strong for batch growth conditions, which implies that tRNA operon structure is a nonrandom result of selective pressures for maximizing growth rate.

## 3.4 Methods

### 3.4.1 *In silico* modeling

iLE1678-ME was obtained from [16], while iJL965-ME was available in lab and described in chapter 2. While no changes were made to iLE1678-ME, iJL965-ME was updated so that tRNAs were associated to specific codons based on their anticodons using figure 1 from [7]. Codons that were not covered from the initial changes described above were covered by assigning all tRNAs associated to the coded AA.

COBRAme formulated ME-models do not have a tRNA to rRNA constraint, unlike previous ME-models [16]. However, an additional rRNA-to-tRNA constraint had to be added to iLE1678-ME. This is because iLE1678-ME rRNA production was internally constrained by the tRNAs associated to their operons, otherwise all of the MC tRNA location models obtained growth rates over 0.97. The severe difference in phenotypes made iLE1678-ME uncomparable to the MC tRNA location models, so an upper bound constraint was placed on the three operons that initally contained rRNA+tRNA-ala that limited their expression to the highest expressed tRNA-AA.

### 3.4.2 Model building

Transcription unit structures were downloaded from [16, 18]. Transcription units that contained tRNA genes were identified in *E. coli* and *C. ljungdahlii*. The exact stoichiometries for the complete transcription of each individual tRNA molecule were calculated using the COBRAme function

*add_transcription_reaction* [16]. These stoichiometries by tRNA were subtracted from each respective operon transcription reaction, so that the left-over reaction accounted for non-tRNA genes and intergenic regions [22]. All tRNA locations in the "left-over reactions" were assigned a number. To incorporate a new tRNA, an ordered list of tRNA genes were randomly shuffled using Python's "random" package (seed set to 86519), and the stoichiometries of the new tRNA gene were then added into the "left-over reaction". In total, 2000 such models were built, 1000 for each species.

### 3.4.3 Analysis

All analysis was performed using Python 2.7 in Jupyter notebooks, and visualization provided by Matplotlib [23, 24]. Both ME-models were solved using qminos with precision set to 1e-15 [15]. To obtain maximum growth rate, carbon substrate lower bound was set to -1000 per model. To get lower growth rates, the substrate uptake rate from the maximum growth rate solution was then multiplied by either 0.5 or 0.1. RNA-seq data were obtained from [25, 17, 26], ribosome profiling data were obtained from [27], and tRNA molecule concentrations were obtained from [8]. All statistical analysis were performed using [28].

Flux variable analysis on transcriptional units revealed that the exact expression and usage per tRNA could vary while other reactions remained the same [22]. However, when tRNA expression and usage were summed by AA, all of the different results from flux variable analysis led to the same results, hence why tRNA-AA is the lowest level of expression considered.

The soft upper boundaries in Fig 3.10b, c were calculated by splitting the

range of tRNA efficiency values into intervals of 30, finding the highest growth rate value for each value, then fitting the best fit line through these points.

## 3.5 Figures and tables

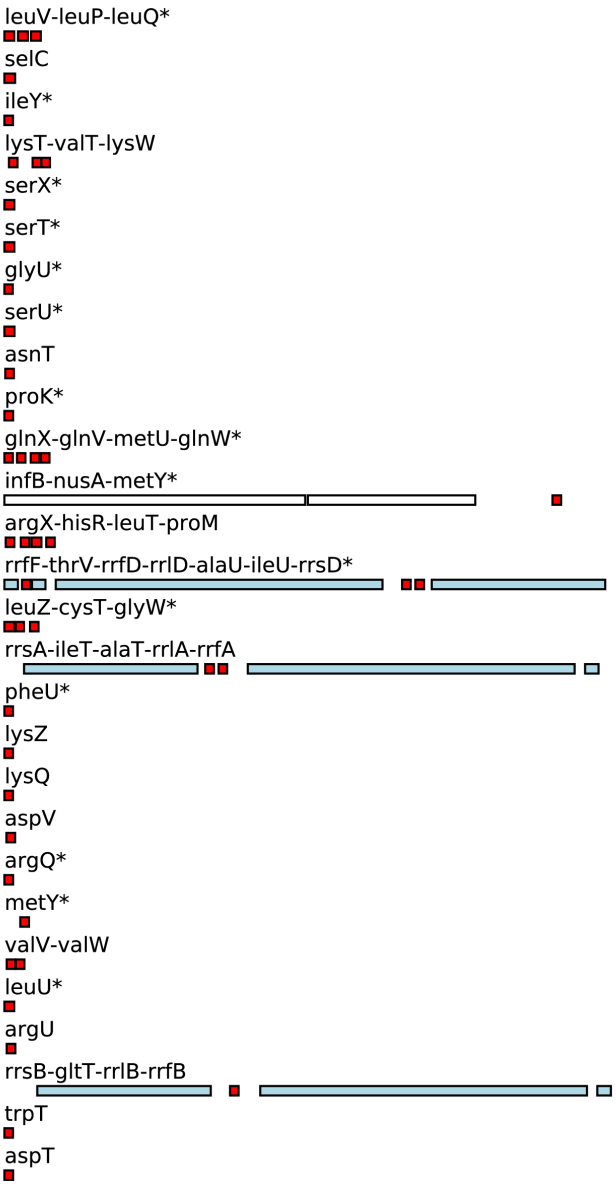



**Figure 3.1**: **Organization of *E. coli's* tRNA-containing operons.** Transcription units are those as published in COBRAme [16]. Red genes are tRNAs, blue genes are rRNAs, and white genes are CDS. Box sizes and spacing between boxes represent relative bp distances. Green lines indicate alternate transcription unit start and stop sites.


**Figure 3.1** ***(continued)*****: Organization of** ***E. coli's*** **tRNA-containing operons.**

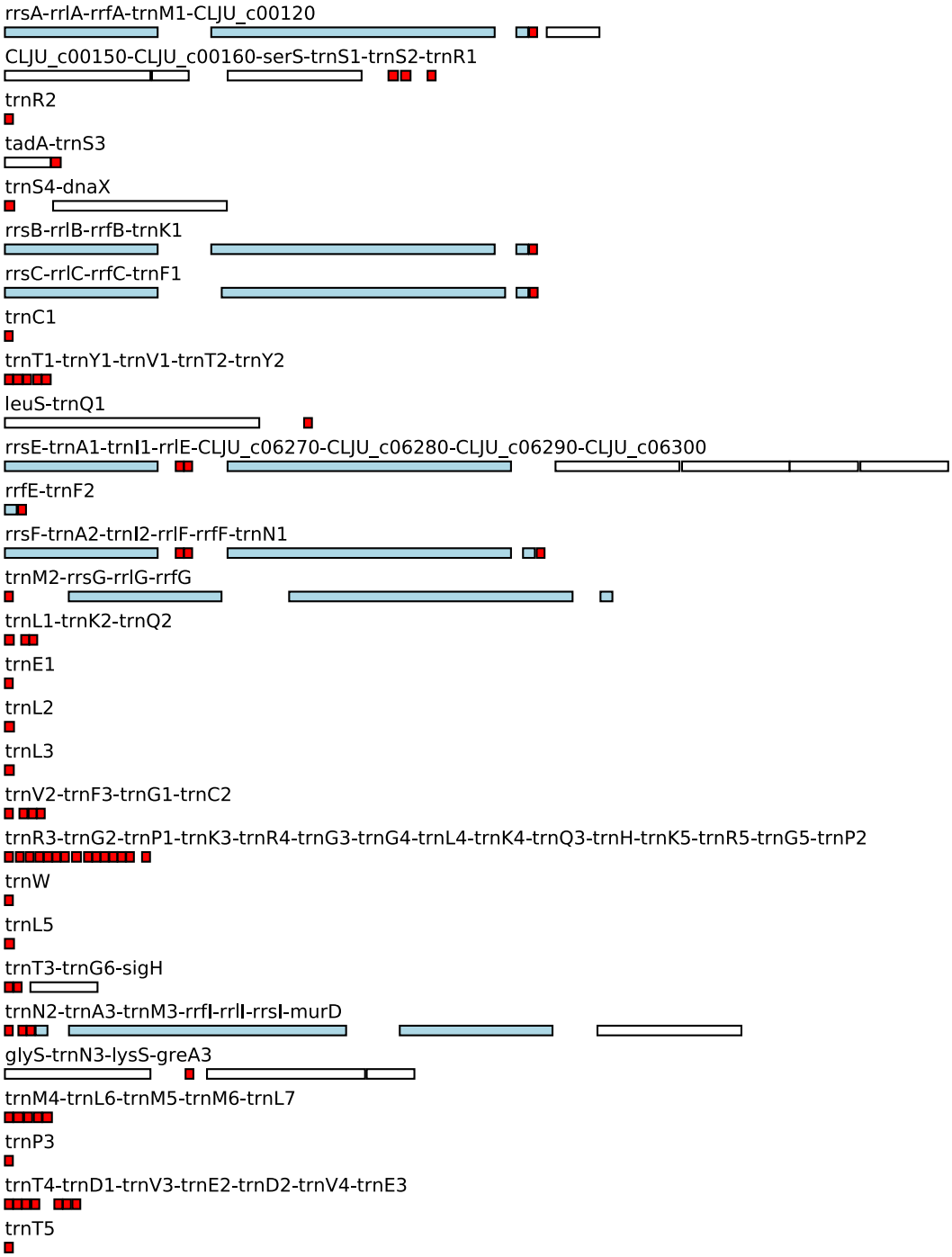



**Figure 3.2**: **Organization of *C. ljungdahlii's* tRNA-containing operons.** Transcription units were obtained from BioCyc on March 22, 2017 [18]. Red genes are tRNAs, blue genes are rRNAs, and white genes are CDS. Box sizes and spacing between boxes represent relative bp distances.

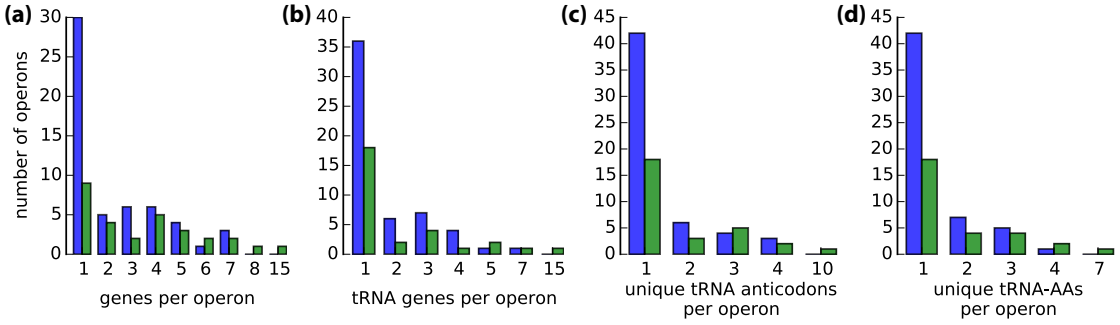



**Figure 3.3**: **Distribution of tRNAs by operon in *E. coli* and *C. ljungdahlii*.** Bar graphs show operon count by (a) the number of genes per operon, (b) the number of tRNAs per operon, (c) the number of unique anticodons as represented by tRNAs per operon, and (d) the number of unique amino acids as represented by tRNAs per operon for *E. coli* (blue) and *C. ljungdahlii* (green). All potential operons, including alternative start and end sites, are included.

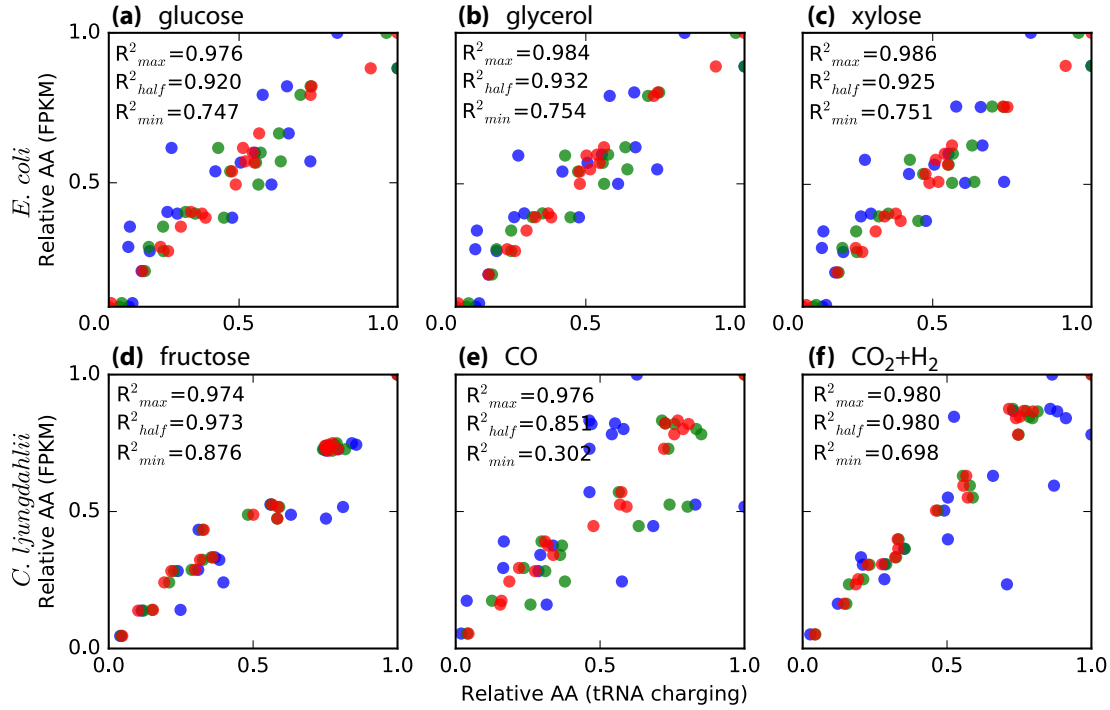



**Figure 3.4**: **Comparing *in silico* and *in vivo* AA composition for *E. coli* and *C. ljungdahlii*.** *In vivo* AA compositions were calculated using RNA-seq, harvested mid-log phase from batch-grown cells, as a proxy for protein count. *In silico* AA compositions were the sum of AA-categorized tRNA charging reactions (mmol*gDW*$^{-1}$*h$^{-1}$) at maximum growth rate (red), half of the maximum substrate uptake (green), and minimum (*i.e.*, tenth) of the maximum substrate uptake rate (blue) on glucose, glycerol, or xylose for *E. coli* (top row) and fructose, CO, or $CO_2$+$H_2$ for *C. ljungdahlii* (bottom row). Values are relative to the most AA required, which is alanine for *E. coli* and lysine for *C. ljungdahlii*.

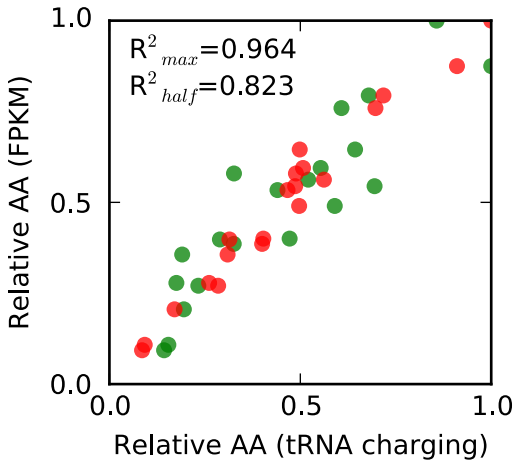



**Figure 3.5**: **Comparing *in vivo* and *in silico* AA composition for *E. coli* grown on acetate.** *In vivo* AA composition was calculated using RNA-seq [25], harvested mid-log phase from batch-grown cells, as a proxy for count of proteins that were also modeled in iLE1678-ME. *In silico* AA composition was calculated from summed tRNA charging reactions at maximum growth rate (red) and half of the maximum glucose uptake (green). Values are relative to the most used AA, alanine.

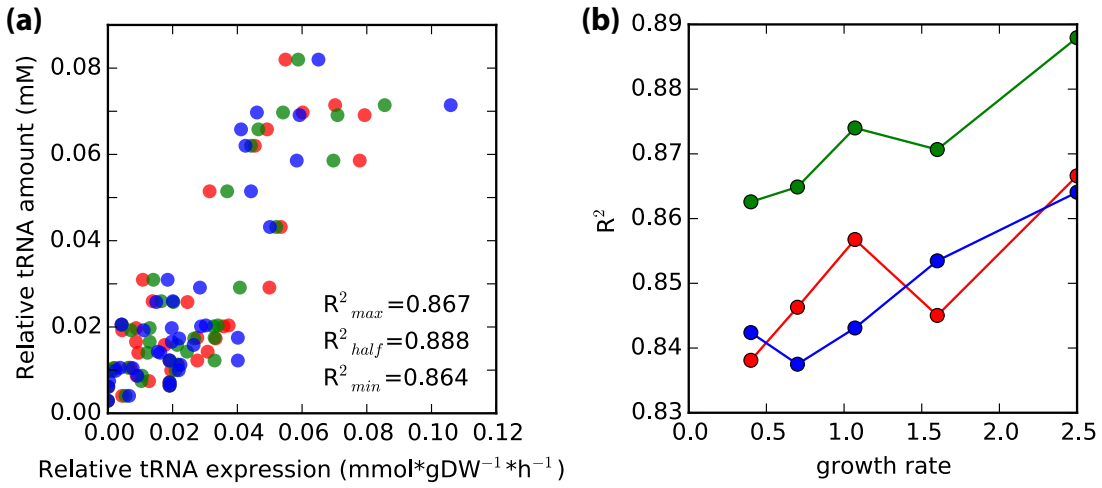



**Figure 3.6**: **Comparing *in vivo* and *in silico* tRNA expression for *E. coli*.** *In vivo* tRNA amounts was taken from maximum growth rate from [8] (Table 5), while *in silico* results were calculated from tRNA charging and tRNA expression reactions at maximum growth rate (red), half of the maximum glucose uptake (green), and one tenth of the maximum glucose uptake rate (blue). $R^2$ values from linear regressions between *in vivo* measurements from growth rates of 0.4, 0.7, 1.07, 1.6, and 2.5 and *in silico* predictions were plotted in (b).

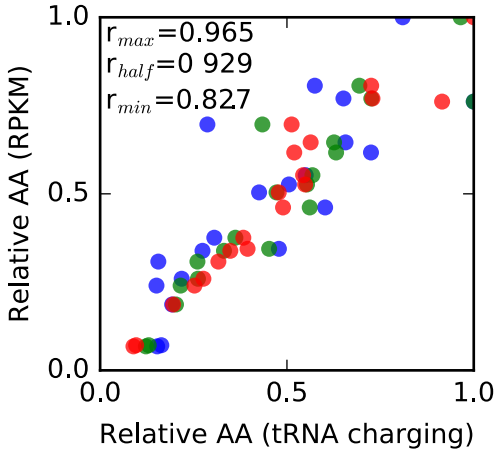



**Figure 3.7**: **Comparing *in vivo* and *in silico* AA composition for *E. coli*.** *In vivo* AA composition was calculated from Ribo-seq using batch-grown mapped reads for protein-coding genes that are represented in the corresponding ME-model [27]. *In silico* aa composition was calculated from summed tRNA charging reactions at maximum growth rate (red), half of the maximum glucose uptake (green), and one tenth of the maximum glucose uptake rate (blue). Values are relative to the most used AA, which is alanine for *E. coli*.

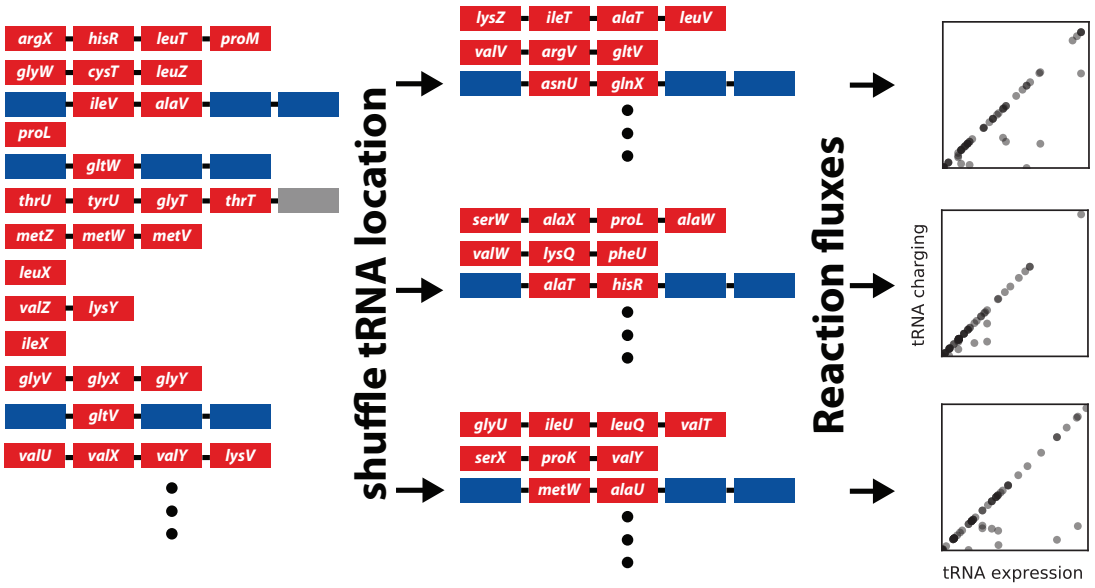



**Figure 3.8**: **Diagram of the Monte-Carlo method for tRNA location shuffling.** Red boxes represent tRNA genes, blue boxes represent rRNA genes, and grey boxes represent open reading frames. Operon diagram is not to scale for gene size and distance.

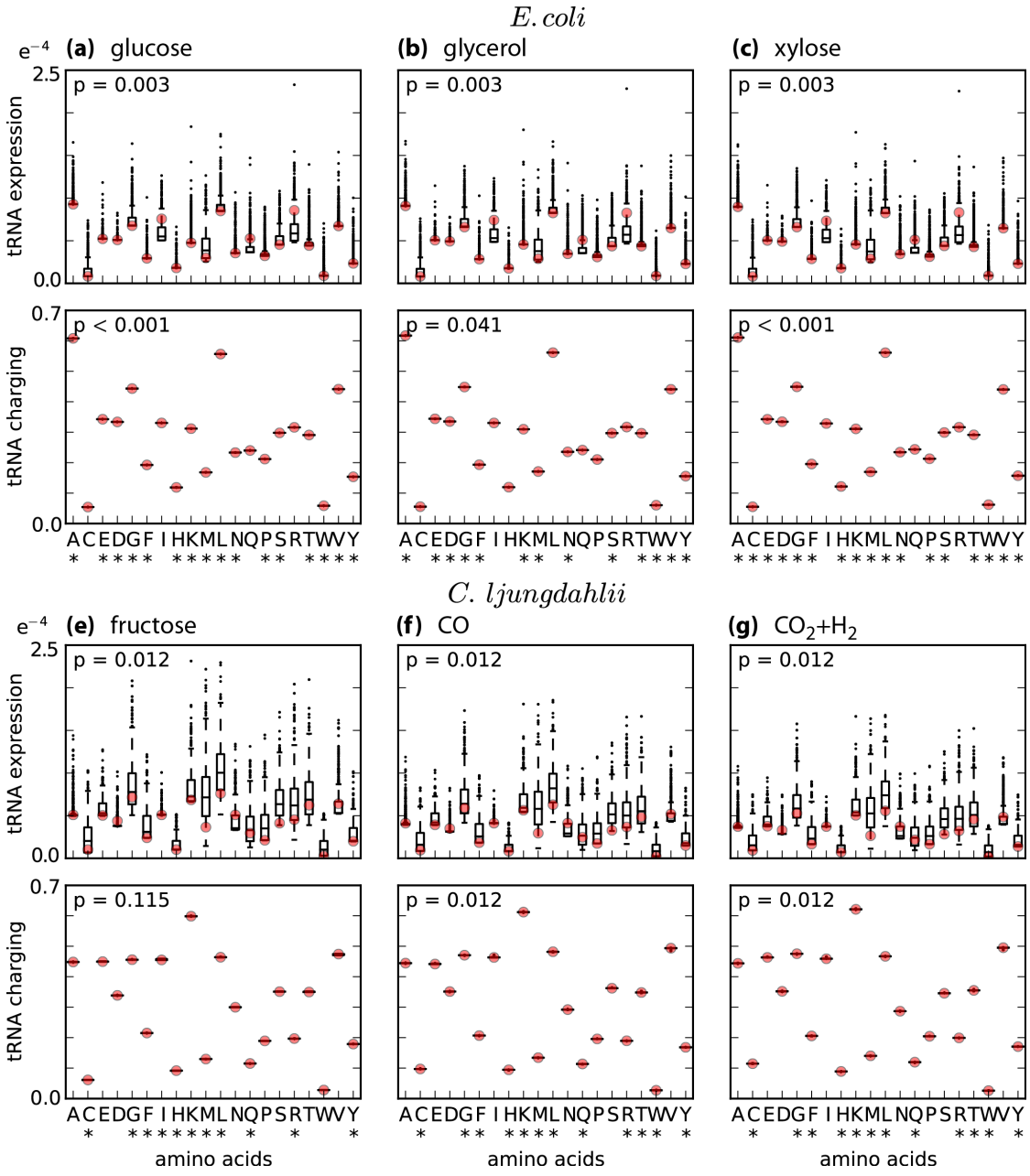
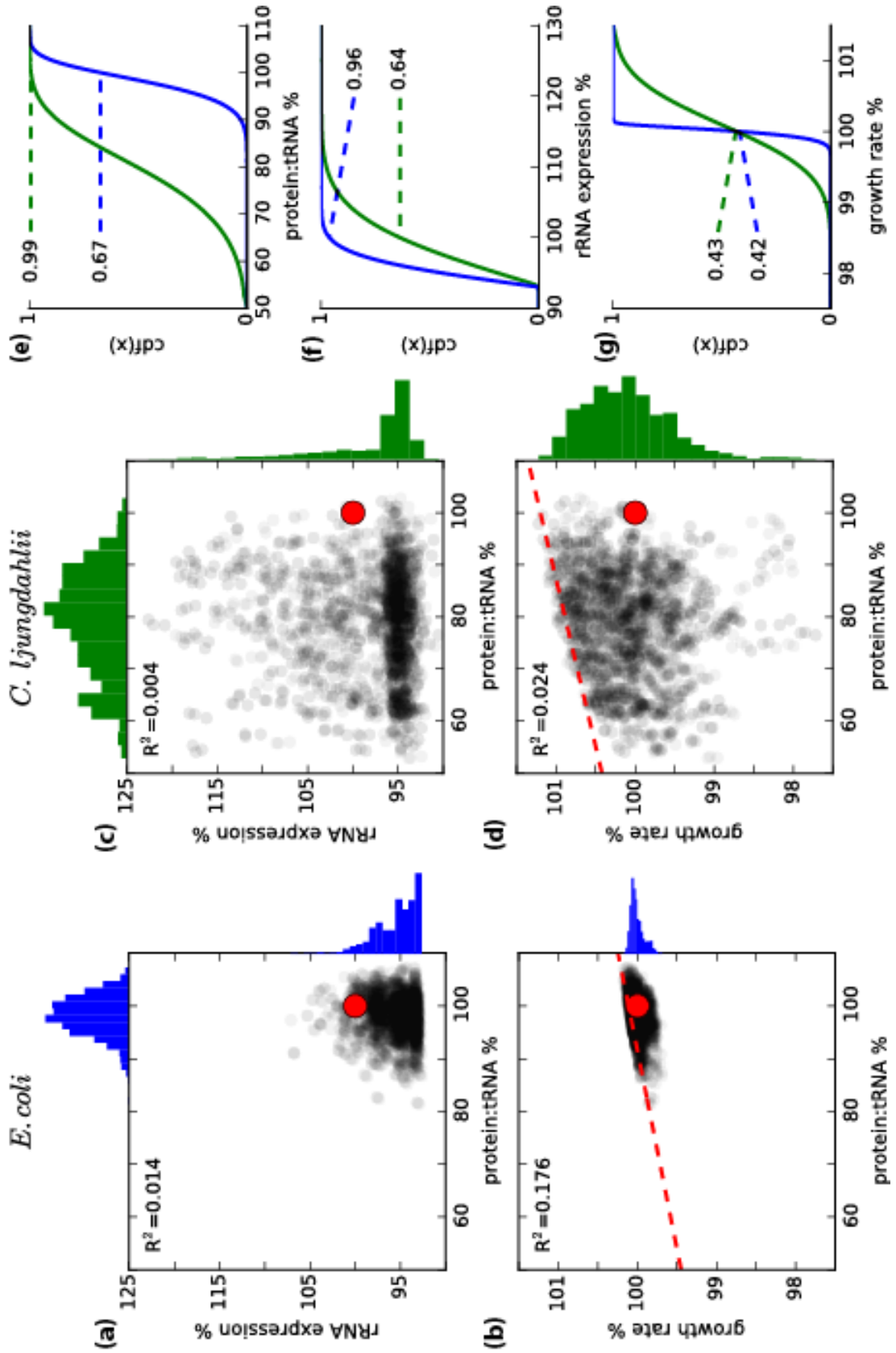



**Figure 3.9**: **Comparing tRNA expression and tRNA charging fluxes against the original models'.** AA-categorized *in silico* tRNA expression (mmol*gDW$^{-1}$) and tRNA charging fluxes (mmol*gDW$^{-1}$) from the MC tRNA location models were plotted as box-plots, and red dots indicate the original models' predictions. *E. coli* was batch simulated on (a) glucose, (b) glycerol, and (c) xylose, and *C. ljungdahlii* on (d) fructose, (e) CO, and (f) $CO_2$+$H_2$. P values are from binomal tests of whether the original models give rise to lower expression levels or higher tRNA usage than the median values from the MC tRNA location models. Asterisks indicate tRNAs by AA that had both less than average tRNA expression and greater than average tRNA usage.

**Figure 3.10*(next page)*: Comparing efficiencies and growth rates from the MC tRNA location models as a percentage of the original models'.** All results are from batch simulations on different substrates, with *E. coli* data coming from glucose, glycerol, or xylose conditions (n=3000), and *C. ljungdahlii* from fructose, CO, or $CO_2$+$H_2$ conditions (n=3000). (a & c) rRNA expression ($\sum$rRNA mmol*gDW$^{-1}$) was plotted against protein:tRNA ($\sum$tRNA charging mmol*gDW$^{-1}$:$\sum$tRNA mmol*gDW$^{-1}$). (b & d) Growth rate (h$^{-1}$) was plotted against protein:tRNA, and soft upper boundaries (red dashed line) were found. Red dots represent the original models' averaged results. $R^2$ values are from linear regressions. A histograph of each dataset is displayed opposite to its axis in (a-d). Cumulative density functions (cdf) calculated from the histographs in (a-d) were plotted against (e) protein:tRNA, (f) rRNA expression, and (g) growth rate. Dotted lines indicate the probability of obtaining a value less than the original models' prediction for *E. coli* (blue) and *C. ljungdahlii* (green).


E. coli
C. ljungdahlii
(a)
(b)
(c)
(d)
(e)
(f)
(g)
R² =0.014
R² =0.176
R² =0.004
R² =0.024
rRNA expression %
growth rate %
protein:tRNA %
cdf(x)
0.99
0.67
0.96
0.64
0.43
0.42

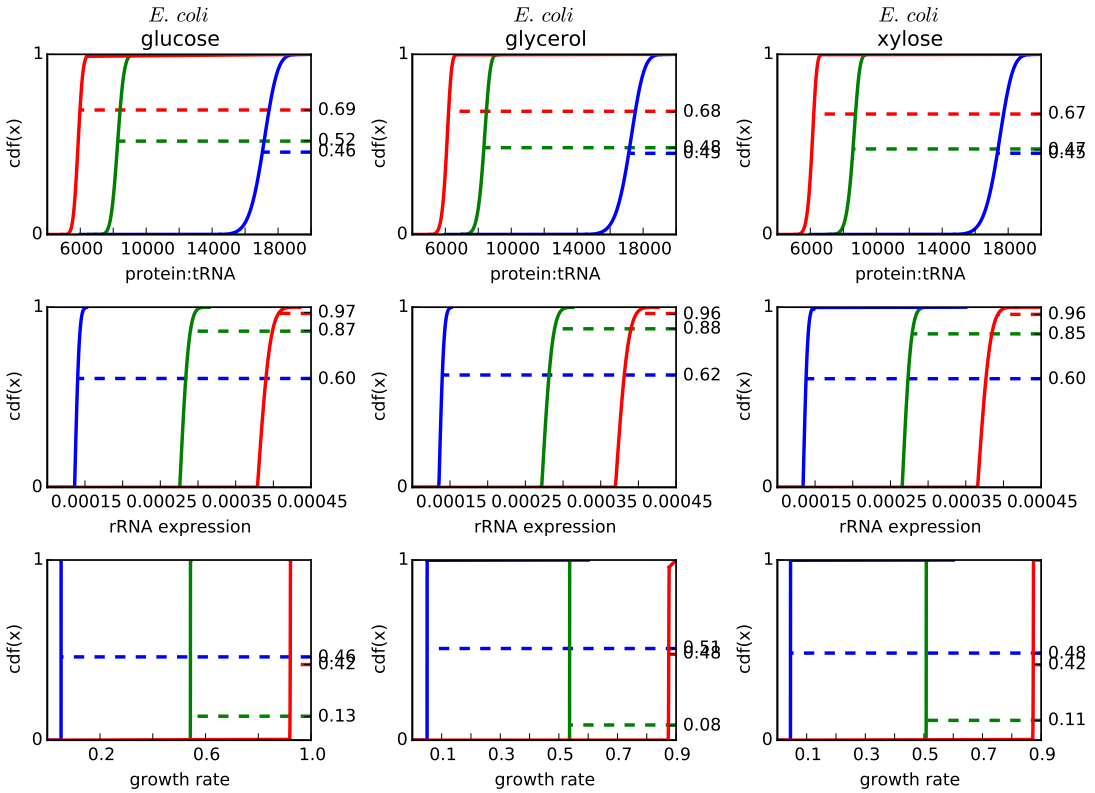



**Figure 3.11**: **Cumulative density functions of tRNA efficiency, rRNA expression, and growth rate for *E. coli* grown on glucose, glycerol, and xylose from the MC tRNA location models.** Cumulative density functions (cdf) of tRNA efficiency ($\sum$tRNA charging reaction fluxes:$\sum$tRNA expression fluxes), rRNA expression (mmol*gDW$^{-1}$), and growth rate (h$^{-1}$) for *E. coli* grown on glucose (left column), glycerol (middle column), and xylose (right column) from the MC tRNA location models. Dotted lines indicate the probability of obtaining a value less than the original model's when grown at maximum growth rate (red), half of the maximum substrate uptake (green), and one tenth of the maximum substrate uptake rate (blue).

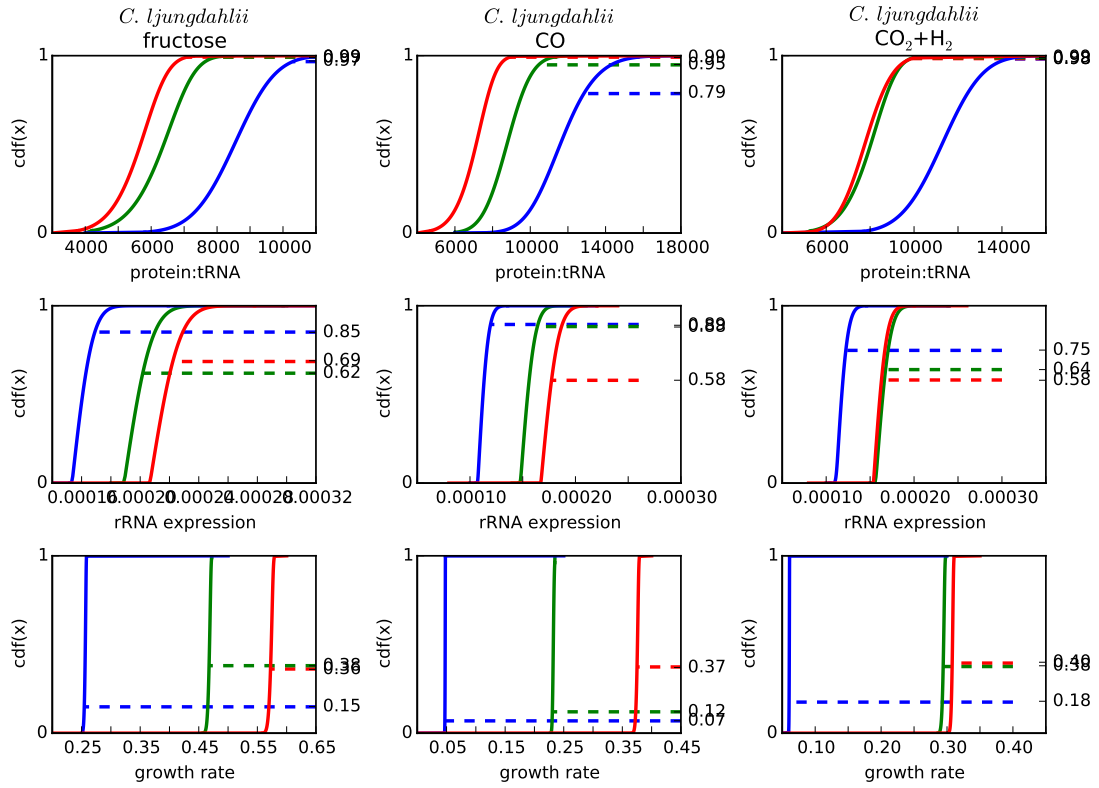



**Figure 3.12**: **Cumulative density functions of tRNA efficiency, rRNA expression, and growth rate for *C. ljungdahlii* grown on fructose, CO, and $CO_2$+$H_2$ from the MC tRNA location models.** Cumulative density functions (cdf) of tRNA efficiency ($\sum$tRNA charging reaction fluxes:$\sum$tRNA expression fluxes), rRNA expression (mmol*gDW$^{-1}$), and growth rate (h$^{-1}$) for *C. ljungdahlii* grown on fructose (left column), CO (middle column), and $CO_2$+$H_2$ (right column) from the MC tRNA location models. Dotted lines indicate the probability of obtaining a value less than the original model's when grown at maximum growth rate (red), half of the maximum substrate uptake (green), and one tenth of the maximum substrate uptake rate (blue).

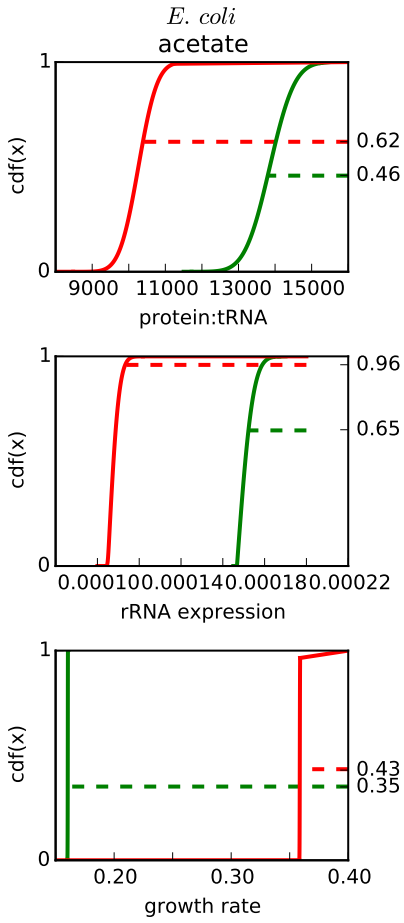



**Figure 3.13**: **Cumulative density functions of tRNA efficiency, rRNA expression, and growth rate for *E. coli* grown on acetate from the MC tRNA location models.** Dotted lines indicate the probability of obtaining a value less than the original model's when grown at maximum growth rate (red) and half of the maximum substrate uptake (green).

**Table 3.1**: **Efficient tRNA expression and tRNA usage in *E. coli*.** P values of binomal tests on whether the original models gives rise to lower tRNA expression levels or higher tRNA charging reactions by AA compared to the median values from the MC tRNA location models.

| Substrate | Growth rate | tRNA expression | tRNA charging |
|---|---|---|---|
| Glucose | Max | 0.0026 | <0.0001 |
| Glucose | Half | 0.0118 | 0.0026 |
| Glucose | Min | 0.2632 | 0.8238 |
| Glycerol | Max | 0.0026 | 0.0414 |
| Glycerol | Half | 0.0118 | 0.2632 |
| Glycerol | Min | 0.2632 | 0.8238 |
| Xylose | Max | 0.0026 | <0.0001 |
| Xylose | Half | 0.0118 | 0.0414 |
| Xylose | Min | 0.1153 | 0.8238 |
| Acetate | Max | 0.0026 | <0.0001 |
| Acetate | Half | 0.0414 | 0.0118 |

**Table 3.2**: **Efficient tRNA expression and tRNA usage in *C. ljungdahlii*.** P values of binomal tests on whether the original models gives rise to lower tRNA expression levels or higher tRNA charging reactions by AA compared to the median values from the MC tRNA location models.

| **Substrate** | **Growth rate** | **tRNA expression** | **tRNA charging** |
|---|---|---|---|
| Fructose | Max | 0.0118 | 0.1153 |
| Fructose | Half | 0.0118 | 0.5034 |
| Fructose | Min | 0.8238 | 0.0414 |
| CO | Max | 0.0118 | 0.0118 |
| CO | Half | 0.5034 | 0.0414 |
| CO | Min | 0.5034 | 0.0414 |
| $CO_2$+$H_2$ | Max | 0.0118 | 0.0118 |
| $CO_2$+$H_2$ | Half | 0.0118 | 0.0414 |
| $CO_2$+$H_2$ | Min | 0.0414 | 0.0414 |

**Table 3.3**: **Optimized tRNA-AA molecules in *E. coli*.** AAs that the original models both express lower tRNA amounts and use more than the median MC tRNA location model for three carbon substrates and availability. Common amino acids by substrate, growth rate, or all conditions were identified. When similar analysis was done on acetate for *E. coli*, maximum growth rate has no optimized tRNA-AA molecules, while at half growth rate, optimized tRNA-AA molecules are related to A, E, D, G, F, H, K, L, N, P, S, T, V, and Y.

| Growth rate | Glucose | Glycerol | Xylose | Intersection |
|---|---|---|---|---|
| **Max** | A, E, D, G, F, H, K, M, L, N, P, S, T, W, V, Y | A, E, D, G, F, K, M, N, S, T, W, V, Y | A, C, E, D, G, F, H, K, M, L, N, P, S, T, W, V, Y | A, E, D, G, F, K, M, N, S, T, W, V, Y |
| **Half** | A, E, D, G, F, H, K, M, L, N, P, S, T, W, V, Y | E, D, G, H, L, P, S, W, V | A, E, D, G, F, H, K, M, N, P, S, T, V, Y | E, D, G, H, P, S, V |
| **Min** | Y, P, K, T, V | A, G, F, K, N, P, T, V, Y | A, G, F, K, N, P, T, V, Y | Y, P, K, T, V |
| **Intersection** | Y, P, K, T, V | G, V | A, G, F, K, N, P, T, V, Y | V |

**Table 3.4**: **Optimized tRNA-AA molecules in *C. ljungdahlii*.** AAs that the original models both express lower tRNA amounts and use more than the median MC tRNA location model for three carbon substrates and availability. Common amino acids by substrate, growth rate, or all conditions were identified.

| Growth rate | Fructose | CO | $CO_2$+$H_2$ | Intersection |
|---|---|---|---|---|
| **Max** | C, G, F, I, H, K, M, L, Q, R, Y | C, G, F, H, K, M, L, Q, R, T, W, Y | C, G, F, H, K, M, L, Q, R, T, W, Y | C, G, F, H, K, M, L, Q, R, Y |
| **Half** | G, F, H, K, M, L, Q, R, Y | C, E, M, L, Q, P, W | E, G, F, H, K, M, L, Q, P, R, Y | Q, M, L |
| **Min** | C, E, M, L, Q, P, W | C, E, M, L, Q, W, V | C, E, F, K, M, L, Q, P, R, T, Y | Q, M, C, E, L |
| **Intersection** | Q, M, L | Q, C, M, L, W | F, K, M, L, Q, R, Y | Q, M, L |

## 3.6 Acknowledgments

Chapter 3, in full, is currently being prepared for submission for publication of the material. Joanne Liu, Nathan Lewis, and Karsten Zengler. Exploring the evolutionary significance of tRNA operon structure using metabolic and gene expression models (working title). The dissertation author was primary investigator and author of this paper. We also thank Joseph Liu, Colton Lloyd, and Laurence Yang for their contributions to this work.

## 3.7 References

[1] Anne E. Osbourn and Ben Field. Operons. *Cellular and Molecular Life Sciences*, 66(23):3755–3775, 2009.

[2] Morgan N Price, Katherine H Huang, Adam P Arkin, and Eric J Alm. Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome research*, 15(6):809–19, 2005.

[3] J. Christian J Ray and Oleg A. Igoshin. Interplay of Gene Expression Noise and Ultrasensitive Dynamics Affects Bacterial Operon Organization. *PLoS Computational Biology*, 8(8):e1002672, 2012.

[4] Pablo A. Nuñez, Héctor Romero, Marisa D. Farber, and Eduardo P.C. Rocha. Natural Selection for Operons Depends on Genome Size. *Genome Biology and Evolution*, 5(11):2242–2254, 2013.

[5] Alon Zaslaver, Avi Mayo, Michal Ronen, and Uri Alon. Optimal gene partition into operons correlates with gene functional order. *Physical Biology*, 3(3):183–189, 2006.

[6] Charles G Kurland. Major codon preference: theme and variations. *Genetics*

*Evol. Biol. J. Mol. Evol. Gene Nature (London) Mol. Microbiol. J. Mol. Evol. Yeast Lloyd, A. T. and Sharp, P. M*, 167140(230):378–39797, 1993.

[7] Mario dos Reis, Renos Savva, and Lorenz Wernisch. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic acids research*, 32(17):5036–44, 2004.

[8] Hengjiang Dong, Lars Nilsson, and Charles G. Kurland. Co-variation of tRNA Abundance and Codon Usage in *Escherichia coli* at Different Growth Rates. *Journal of Molecular Biology*, 260(5):649–663, 1996.

[9] S Kanaya, Y Yamada, Y Kudo, and T Ikemura. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, 238(1):143–55, 1999.

[10] Michael J McDonald, Chih-Hung Chou, Krishna B S Swamy, Hsien-Da Huang, and Jun-Yi Leu. The evolutionary dynamics of tRNA-gene copy number and codon-use in *E. coli*. *BMC evolutionary biology*, 15:163, 2015.

[11] David H Ardell and Leif A Kirsebom. The Genomic Pattern of tDNA Operon Expression in *E. coli*. *PLoS Computational Biology*, 1(1):e12, 2005.

[12] Naama Wald and Hanah Margalit. Auxiliary tRNAs: large-scale analysis of tRNA genes reveals patterns of tRNA repertoire dynamics. *Nucleic acids research*, 42(10):6552–66, 2014.

[13] Joshua A. Lerman, Daniel R. Hyduke, Haythem Latif, Vasiliy A. Portnoy, Nathan E. Lewis, Jeffrey D. Orth, Alexandra C. Schrimpe-Rutledge, Richard D. Smith, Joshua N. Adkins, Karsten Zengler, and Bernhard O. Palsson. In silico method for modelling metabolism and gene product expression at genome scale. *Nature Communications*, 3:929, 2012.

[14] Edward J O'Brien, Joshua A Lerman, Roger L Chang, Daniel R Hyduke, and Bernhard Ø Palsson. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Molecular systems biology*, 9: 693, 2013.

[15] Laurence Yang, Ding Ma, Ali Ebrahim, Colton J. Lloyd, Michael A. Saunders, and Bernhard O. Palsson. solveME: fast and reliable solution of nonlinear ME models. *BMC Bioinformatics*, 17(391), 2016.

[16] Colton J Lloyd, Ali Ebrahim, Laurence Yang, Zachary Andrew King, Edward Catoiu, Edward J O'Brien, Joanne K Liu, and Bernhard O Palsson. COBRAme: A Computational Framework for Building and Manipulating Models of Metabolism and Gene Expression. *bioRxiv*, 2017.

[17] Joanne Liu, Colton Lloyd, Mahmoud Al Bassam, Ali Ebrahim, Jinu Kim, Connor Olsen, and Karsten Zengler. Predicting proteome allocation, overflow metabolism, and metal requirements in *Clostridium ljungdahlii*. *To be submitted*, 2017.

[18] Ron Caspi, Richard Billington, Luciana Ferrer, Hartmut Foerster, Carol A. Fulcher, Ingrid M. Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A. Mueller, Quang Ong, Suzanne Paley, Pallavi Subhraveti, Daniel S. Weaver, and Peter D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, 44(D1):D471–D480, 2016.

[19] Mike Withers, Lorenz Wernisch, and Mario dos Reis. Archaeology and evolution of transfer RNA genes in the Escherichia coli genome. *RNA (New York, N.Y.)*, 12(6):933–42, 2006.

[20] Matthew Scott, Carl W. Gunderson, Eduard M. Mateescu, Zhongge Zhang, and Terence Hwa. Interdependence of Cell Growth and Gene Expression: Origins and Consequences. *Science*, 330(6007):1099–1102, 2010.

[21] John H. Andrews and Robin F. Harris. r- and K-Selection and Microbial Ecology. pages 99–147. Springer, Boston, MA, 1986.

[22] Ali Ebrahim, Joshua A Lerman, Bernhard O Palsson, and Daniel R Hyduke. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Systems Biology*, 7(1):74, 2013.

[23] John D. Hunter. Matplotlib: A 2D Graphics Environment. *Computing in*

*Science & Engineering*, 9(3):90–95, 2007.

[24] Fernando Pérez and Brian E. Granger. IPython: A System for Interactive Scientific Computing Python. *Computing in Science and Engineering*, 9(3): 21–29, 2007.

[25] Troy E Sandberg, Colton J Lloyd, Bernhard O Palsson, and Adam M Feist. Laboratory Evolution to Alternating Substrate Environments Yields Distinct Phenotypic and Genetic Adaptive Strategies. *Applied and environmental microbiology*, 83(13):e00410–17, 2017.

[26] Harish Nagarajan, Merve Sahin, Juan Nogales, Haythem Latif, Derek R Lovley, Ali Ebrahim, and Karsten Zengler. Characterizing acetogenic metabolism using a genome-scale metabolic reconstruction of *Clostridium ljungdahlii*. *Microbial cell factories*, 12(1):118, 2013.

[27] Fuad Mohammad, Christopher J. Woolstenhulme, Rachel Green, and Allen R. Buskirk. Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling. *Cell Reports*, 14(4):686–694, 2016.

[28] Eric Jones, Travis Oliphant, and Pearu Peterson. SciPy: Open source scientific tools for Python. 2001.

# Chapter 4

# Reconstructing and modeling protein translocation and compartmentalization in *Escherichia coli*

## 4.1 Introduction

Compartmentalization provided by membranes is essential for life. Compartmentalization allows unique internal microenvironments, permits harvestable energy gradients, provides organizational structure, protects the cell, and more. Membranes also represent significant physical barriers. Thus, cells have evolved pathways that allow molecule transport between compartments. As a gram-negative bacterium, *Escherichia coli* has two membranes: An inner, tightly regulated mem-

brane and an outer, more porous membrane (see [1, 2] for review). In order to achieve desired membrane functions, *E. coli* has evolved a system to translocate protein into their appropriate locations.

There is a wealth of scientific information on protein translocation processes, but holistic studies on their system-wide effects are lacking. Such genome-wide studies are important as protein translocation enables key cellular functions. These functions need to be put into context of all other cellular functions to understand their energetic requirements, general interactions and balance with the rest of the cell. To do so, one must take a systems approach, where comprehensive molecular processes and interactions are reconstructed into a self-consistent and computable format. A couple of recently published studies have taken steps in this direction. In a comprehensive approach to cellular processes, the recent whole-cell model of *Mycoplasma genitalium* incorporates a SecA+Sec translocase pathway into one of its protein formation modules [3]. In this model, translation is uncoupled from translocation, even though the two processes can happen concurrently [4]. Furthermore, protein translocation rates are not calculated *de novo* but are instead based on user-inputted gene expression levels and energy-carrier metabolite concentrations (calculated prior from a separate module). Thus, set expression levels of protein translocases operate as a constraint on other processes; for example, metabolism uptake is dependent on the number of transporters. Additionally, membrane lipid formation is driven by a biomass objective function [3], whereas a computation based on a cell's surface area might be more appropriate. In another study, a larger effort was focused on the genome-scale reconstruction of the protein secretion pathway in *Saccharomyces cerevisiae* [5]. This model of protein

secretion is stand-alone' and is not integrated with additional cellular processes. It can be used as a scaffold on which omics data (e.g., RNA-seq) can be overlaid to estimate effects of protein abundance and metabolic costs of translocation on the cell. Although these models contain some detail about protein translocation, both are reliant on expression data input and are not dependent on the demands of cellular events. Finally, another notable model incorporated membrane space into a genome-scale model of *E. coli* to demonstrate that while the membrane may cap certain fluxes, leading to simultaneous respiration and fermentation at high growth rates, metabolic demands drive the membrane proteome. Although this model lacks the process of protein translocation and has only four integral proteins, it demonstrated that the consequence of protein translocation, namely compartment formation, truly constrains cellular events [6].

A recent genome-scale model of metabolism and gene-expression of *E. coli*, called a ME-model [7] or specifically, the retroactively named iOL1650-ME model (following a previous naming convention [8]), affords us the opportunity to integrate protein translocation seamlessly with cellular processes. Although iOL1650-ME describes the synthesis of all the proteins in the proteome, the proteins are not compartmentalized. In this work, we significantly expanded the validated iOL1650-ME model [7] to include a comprehensive reconstruction of protein translocation pathways. The expanded iOL1650-ME includes a reconstruction of lipoprotein biogenesis, the incorporation of four distinct protein compartments (cytoplasm, periplasm, the inner and the outer membrane), published enzymatic rates of the translocases and diffusion rates of outer membrane porins, and a membrane constraint based on cell morphology all integrated into one reconstruction. The ex-

panded model, hereafter referred to as iJL1678-ME, allows for *de novo* prediction of enzyme abundances and their cellular location as well as the constraining effects of membrane production. We apply iJL1678-ME to show how it is predictive of compartmentalized cellular content for validation, describe its utility and limitations, and show how it can be applied to examine a broadened scope of applications including targeted inhibition of proteins.

## 4.2 Results and Discussion

All proteins in *E. coli* are synthesized in the cytoplasm, but over 20% of *E. coli*'s protein-coding open reading frame (pORF) are annotated to encode protein with non-cytoplasmic functions, and an estimated 15% of cellular protein mass is in the cell envelope [9, 10]. These proteins are assisted by translocase complexes to get to their cellular destinations. Depending on their final location and biochemical properties, the translocation route taken for a particular protein involves one of three integral inner membrane translocases (Sec, Tat, and YidC) and perhaps an outer membrane translocase (LolB and Bam) (see [1, 11] for review). The most-studied and ubiquitous translocase is the Sec complex [12]. The channel-forming Sec protein has two chaperone pathways that converge on it. One, the SRP/Sec pathway, brings nascent peptides to the Sec complex and primarily uses the kinetic energy of translation to drive protein integration into the inner membrane [4, 13, 14]. Sometimes, the mediator YidC binds to Sec complex to enhance proper membrane integration, but on its own, YidC is an insertase that translocates a couple of essential proteins [15, 16, 17]. Alternatively, proteins moving to

the periplasm and beyond, generally follow the SecB/Sec pathway which uses an ATPase, SecA, to thread chaperoned, unfolded proteins through the Sec complex and into the periplasm [18, 19, 20, 21]. Furthermore, non-cytoplasmic, folded proteins which often containing cofactors, take the Tat translocase, a dynamic protein complex that recruits TatA subunits to adjust its channel size appropriately and is driven by an electrochemical gradient [22, 23, 24]. To get to the outer membrane, proteins must first cross the inner membrane, then take one of the two pathways: Lol and Bam. The Lol pathway excises lipoproteins from the inner membrane and incorporate them into the outer membrane [25, 26]. In the Bam pathway, unfolded $\beta$-barrels are chaperoned in the periplasm, typically by SurA [27, 28], to the Bam complex, which facilitates their proper insertion into the outer membrane [29]. Alterations to these pathways exist, but these five translocation pathways are thought of as canonical pathways [25, 30]. All this information enables a bottom-up reconstruction of the protein translocation network in *E. coli*.

### 4.2.1 Reconstruction of protein translocation processes and their incorporation into iOL1650-ME

A bottom-up procedure to reconstruct the network of protein translocation and lipoprotein biogenesis within a genome-scale model of metabolism and gene-expression in *E. coli* [7] was developed (Fig 4.1A). The result of implementing this procedure was a biochemically, genetically, and genomically structured network [31] that enabled the analysis of the molecular effects of protein translocation in context of other networks using constraint-based analysis methods. The network reconstruction procedure involved five major phases.

**Reconstruction of protein translocation pathways**

Through an extensive literature search, the SecB/Sec, SRP/Sec, Tat, Lol, Bam, and YidC insertion translocation pathways were identified for inclusion into the reconstruction (Fig 4.1B) (see [1, 11] for review). Three additional pathways were also included, based on case studies demonstrating that the SRP/Sec pathway occasionally requires assistance from YidC and/or SecA to have properly formed integral proteins [25, 30, 32]. In addition to protein translocation, lipoprotein biogenesis pathways were reconstructed, as lipoproteins are located in membranes and are essential through their structural and functional uses (Methods) [33, 34, 35]. In the end, 27 pORFs and one RNA gene, which together form 16 protein complexes, were added to the model to enable protein translocation (Additional file 1 Tables 1 & 2). Furthermore, based on the sequence of events in each of these pathways, a set of mechanistic reactions (*i.e.*, template reactions [36]) were developed that could be applied to and individualized for every pORF.

**Compartmentalization**

The incorporation of protein translocation pathways requires proteins to have defined compartmentalization. First, two new compartments, inner and outer membranes, were added to the three existing compartments in iOL1650-ME (cytoplasm, periplasm, and extra-cellular) [7]. Using the protein databases EchoLocation [37], Uniprot [38], and Ecocyc [39] as well as the bioinformatic programs PSORTb [40] and TMHMM [41], the 1,568 pORFs included in the reconstruction were assigned to compartments (Fig 4.1C). pORFs with a transmembrane component or a lipid membrane anchor were assigned to either the inner or outer

membrane; otherwise, pORFs were either cytoplasmic or periplasmic. Proteins composed of multiple pORFs were assigned to the compartment of its components (Additional file 1 Table 2), but if any of its pORFs was in a membrane then the entire complex was assigned to that membrane, with the outer membrane taking precedent over the inner(*e.g.*, AcrAB-TolC multidrug efflux system is assigned to the outer membrane). For example, ATP synthase has pORFs located in the inner membrane (AtpB, AtpC, AtpE, AtpF) and cytoplasm (AtpA, AtpD, AtpG, AtpH), but the synthase itself is assigned to the inner membrane so that it may interact with metabolites in both the cytoplasm and periplasm.

The compartment assignment resulted in 71% of pORFs being assigned to the cytoplasm, 21% to the inner membrane, 6% to the periplasm, and 2% to the outer membrane.

**Assigning translocated proteins to pathways**

Protein translocation reactions were formulated for each pORF. Using a set of rules based on experimental data, protein location, and physical properties (Additional file 1 Table 3), non-cytoplasmic annotated pORFs were assigned to translocation pathways (Fig 4.1D). The developed template reactions allowed for the methodological creation of each pORF's translocation reactions and their subsequent incorporation into the reconstruction. Additional pathway development steps included determining the amount of ATP hydrolyzed by SecA for each pORF (*i.e.*, 1 ATP per $\backsim$25 amino acids) [42], assigning 23 pORFs to lipoprotein biogenesis [37], and calculating the number of TatA's needed for each Tat-translocated pORF [23] (Additional file 1 Table 1, Fig 4.2). Published translocase $k_{cat}$ values

were associated with appropriate proteins in the translocation pathways. These values [43, 44, 45, 46, 47] were incorporated into the model through coupling constraints [36, 48], which account for turnover rates by linking gene expression to metabolism through the dependence of reaction fluxes on enzyme concentration (Fig 4.1D) [35]. Additionally, outer membrane porins were represented to behave as passive-diffusion channels [2] in the reconstruction. Instead of identical turnover rates for all outer membrane porins in the cell, incorporation of porin-specific coupling constraints allowed the model to account for individualized solute diffusion rates based on effective porin radius, hydrodynamic solute radius, membrane thickness, and growth rate (see Additional file 1 Table 4 for list of solutes, which are also exchange metabolites). This formulation represents the cross-sectional area a solute can pass through and distance a solute had to travel to reach the periplasm [49]. Without these coupling constraint updates, the model was unable to predict accurate translocase levels (Fig 4.3).

**Incorporating cell-size and membrane constraints**

Cell envelope production was fundamentally changed to reflect the cell's shape and composition more accurately. The previously-developed iOL1650-ME accounts for production of kdo 2 lipid IV, phospholipids, and murein through growth rate dependent demands scaled to cell size [7]. These demands were identified as key areas for improvement to a more mechanistic description in iJL1678-ME. Changes to the model included adding murein recycling, a lipoprotein demand, and a membrane spatial constraint. The peptidoglycan layer protects the cell from lysis by providing a physical structure, and it also dynamically renews its components

by using enzymes located in all compartments of the cell (see [50] for review). To reflect this renewal process, AmpG permease transports anhydro-muropeptides to equal 45% of the murein demand, which causes a murein recycling loop [51]. Lipoproteins are also important for structural integrity, and the number of lipoproteins that have been estimated in a cell, $7x10^5$, is a significant amount of mass [10], so a growth-rate scalable lipoprotein demand, using Braun's lipoprotein [52], was added. Finally, because there are inner and outer membrane compartments, membrane demands and composition can be more explicitly described with the genome-scale model. Membrane surface area, which is a function of growth rate, is required to be occupied completely by proteins and lipids (see Methods). The surface area of integral proteins was calculated from their mass, except for lipoproteins which were set to the approximate cross-sectional area of their lipid moieties (Additional file 1 Table 5) [10, 23, 53]. The rest of the outer membrane outer leaflet is filled in with kdo 2 lipid IV while the other three membrane leaflets are occupied by a mixed composition of phospholipids (see Methods for mathematical formulation of the membrane constraint) [54, 55]. This novel membrane constraint not only allows a variable membrane proteome, but it also ensures that the cell is completely covered by two membranes.

**Updating model parameters**

Two model parameters were updated to reflect the new reconstruction content. The growth-associated maintenance (GAM) was updated from 35 to 34.98 ATP mmol *$gDW^{-1}$ to account for the ATP spent translocating proteins out of the cytoplasm, which is small compared to the cell's total energy production but

expensive per non-cytoplasmic protein (0.02 for translocating 2.3x$10^{-3}$ protein mmol*$gDW^{-1}$, or 85.7 ATP for each non-cytoplasmic protein) . Also, the out-of-scope protein proportion of proteome, a parameter introduced in iOL1650-ME to account for proteins expressed *in vivo* but actively utilized by the network reconstruction [7, 56], was changed. As iJL1678-ME includes more pORFs, this parameter's value had to be reduced by the expressed mass of new protein content. Thus, the out-of-scope protein proportion was changed from 0.45 to 0.36 to reflect iJL1678-ME's increased comprehensiveness. Taken in whole, the improved network reconstruction demonstrated that there is enough scientific literature to accurately reconstruct protein translocation in a genome-scale model. As a result of having this reconstruction, it was possible to compute physiological aspects of the cell envelope, which converges to a fully comprehensive *in silico* model of *E. coli*.

### 4.2.2 Proteomic shifts highlight the significance of new content in iJL1678-ME

iOL1650-ME and iJL1678-ME enable quantitative predictions of genome-scale proteome abundances. Instead of requiring input expression data, these models calculate the proteins necessary to maximize growth rate through a metabolism-centered network. However, not only does iJL1678-ME contain more reconstructed content, but it also has a reformulated cell envelope representation that requires more membrane production, phospholipid variety, and murein recycling. To demonstrate the difference between the two ME-models, the computed protein expression fluxes in glucose M9 minimal media were compared (Fig 4.4, *in silico* media com-

position given in Additional file 1 Table 6). Although the majority of pORFs (1475) were approximately the same in both model simulations, 32 of the genes were differentially expressed, and a number of proteins were uniquely expressed (Fig 4.4A). Clearly, accommodating protein translocation has a systemic effect on the computed proteome.

Looking first at pORFs expressed in both models, the largest outlying subgroup is the cell membrane and envelope related proteins. This differential expression was due to the addition of murein recycling, which increases overall murein production (145%) and associated ATP expenditure (140%, which is 2.3% of all ATP production in iJL1678-ME). It has been previously reported that murein recycling can come to a significant cost to the cell [51]. As for carbohydrate metabolism, the porin coupling constraint forced iJL1678-ME to consider the slower diffusion rate of acetate verses gaseous molecules; thus, iJL1678-ME utilized acetate overflow (*i.e.*, fermentation) pathways less than iOL1650-ME. Not only was its acetate secretion less (1.5 verses 8.1 mmol*gDW$^{-1}$*h$^{-1}$), but it also downregulated two genes involved in small carbon molecule metabolism (eutD and purT). Instead, iJL1678-ME adjusted its energy production pathways so that more of its ATP was generated through oxidative phosphorylation. As a consequence, expression of TCA cycle proteins and succinate dehydrogenase was greater. Finally, the collective increase in protein expression due to the expanded scope of iJL1678-ME led to greater expression of transcription, vitamin B12 transporters, and nucleotide metabolism proteins.

When examining the uniquely expressed genes, 65 genes were unique to iJL1678-ME (Fig 4.4B), and 6 to iOL1650-ME. Of the uniquely expressed pORFs

in iJL1678-ME, 42% were reconstructed in this paper and thus not contained in iOL1650-ME. The rest were due to murein recycling, more phospholipid variety (as part of the membrane constraint), and an increase in oxidative phosphorylation, which in turn required heme metabolism. As for the uniquely expressed proteins in iOL1650-ME, these proteins were due to isozymes employed (*e.g.*, AcnA verses AcnB in iJL1678-ME).

In summary, the increased scope of modeled genes in iJL1678-ME caused a notable change in protein expression levels, and these shifts can be directly attributed to model updates and constraints derived from biochemical knowledge available in literature. The resulting proteomic content was examined further.

### 4.2.3 *In silico* computations recapitulate *in vivo* data

To estimate the accuracy of the iJL1678-ME *in silico* proteome, glucose M9 minimal media simulation results were compared to experimental data (Additional file 1 Table 6). Unlike iOL1650-ME, iJL1678-ME calculates a compartment-specific proteome with absolute protein levels. Although this ability may be especially useful in studying the membrane proteome, an area plagued by hardship due to its hydrophobic and amphiphilic nature, it has also created difficulty in comprehensively evaluating iJL1678-ME's results. Even though the correlation between the transcriptome and proteome is poor on a protein-to-transcript level [57, 58], RNA-seq is a robust currently-available omic data-source which covers genome-scale expression in all compartments. Assuming that discrepancies in transcript-to-protein ratios are reduced through averaging, RNA-seq data (GEO accessions: GSE48324 [59] and GSE55367 [60]) was assumed as a one-to-one proxy for protein levels. Pro-

tein masses were calculated from amino acid sequences and normalized by relative fractional proteome mass. Once a comprehensive quantitative proteomics dataset is available, it will be important to validate that the same functional groups are under-predicted.

Since the network reconstruction expanded the scope of iOL1650-ME, we sought to validate the new features of the genome-scale model. The computed mass of all proteins associated with a translocation pathway (color labeled in Fig 4.1B) as a fraction of total cellular protein mass is largely similar to *in vivo* data (Fig 4.5A, Fig 4.6). The most notable outlier is the Tat pathway. The difference between *in silico* and *in vivo* expression may be due to the fact that a TatBC complex forms multiple channels to simultaneously translocate substrates [61, 62], but in iJL1678-ME model, each TatBC complex translocates a single substrate at any point in time. To explore the possibility of a different representation for TatBC, the mass of TatBC was adjusted by four-fold (the maximum demonstrated number of bound precursor proteins) and this improved the *in vivo* to *in silico* correlation ($R^2$=0.897 to 0.925, p-value=0.014 to 0.009), which hints at the possibility TatBC commonly forms multiple channels per complex *in vivo*. These results demonstrate that bottom-up reconstruction approaches and constraint-based modeling can estimate relative protein levels when incorporated with turnover rates and metabolic demands and serves as validation of the reconstructed content (see Fig 4.3 for translocation without $k_{cat}$).

iJL1678-ME's ability to accurately compute protein amounts extends to compartmentalization, which is enabled due to protein translocation (Fig 4.5B). Simulation results predict that the mass of cytoplasmic proteins constitute approx-

imately 79% of the proteome, while the inner membrane protein masses are 10%, periplasmic 1.0%, and outer membrane 10%. Calculating these same values for *in vivo* measurements gave 76.6%, 10.6%, 4.9%, and 7.9%, respectively. In a complementary analysis, iJL1678-ME estimated outer membrane protein values closer to published numbers than *in vivo* (RNA-seq) data's approximation of the outer membrane proteome. The *in silico* protein numbers reflect experimental published amounts at $7.2 \times 10^5$ lipoproteins verses $7 \times 10^5$ and $1.5 \times 10^5$ porins verses $2 \times 10^5$ [10], which implies that the RNA-to-protein ratio is not one-to-one for outer membrane proteins. As there are less proteins in the non-cytosolic compartments, the averaging effect of large groups is less effective, which may explain the discrepancy.

Where do the similarities and differences between the computed and measured compartment-specific protein mass arise? To answer this question, the protein masses were broken down into smaller subgroups, as labeled in iJO1366 which used EcoCyc and GO annotations [39, 63]. All 1,568 pORFs were categorized by functional annotation as opposed to a gene-by-gene comparison, with the assumption that a larger sample size would reduce the discrepancies between protein and RNA abundances. A comparison between computational predictions and experimental data was performed using linear regression of log-log values with zero values being removed from further calculations (Fig 4.7). A normal probability plot of the standardized residuals of the initial model (Fig 4.8) revealed that while most points could be described by a normal distribution, five points describing lowly-expressed functions in iJL1678-ME were out of range (Fig 4.7A). These five points were separated for further analysis while the reduced set of points was refitted, resulting in a more accurate linear model (Fig 4.7B).

Due to their departure from normalcy, the five outliers in Fig 4.7A were examined to identify reasons for modeled discrepancies. The five points covered genes involved with inorganic ions, cofactor and prosthetic groups, protein maturation, and metabolite transportation. Not only is the available knowledge of metal ion and cofactor requirements sparse [64], but the model demands the incorporation of only the most necessary groups into proteins. As result, expression of inorganic ion, cofactor, and prosthetic related pORFs are low. Similarly, protein maturation pORFs are required for proper inclusion of ions and groups; they also assist misfolded proteins, whose possibility are not computed in optimal situations. Lastly, iJL1678-ME predicts a lower periplasmic mass for small metabolite transportation as compared to *in vivo* data. Closer examination of this functional group revealed that the model has severely decreased the diversity of ABC transporters to five protein species. However, *E. coli* produces multiple species of ABC transporters in preparation for environmental changes [65]. This readiness to consume a variety of substrates improves the cell's overall fitness, but when confronted with glucose as the sole carbon substrate, the varied over-expression limited the predicted optimal growth rate, according to iJL1678-ME.

### 4.2.4 Applications predict the effect of molecular perturbations

Genome-scale models of metabolism have enjoyed many successes in elucidating interactions, metabolic engineering, drug targeting, and more. Up to this point in time, perturbations in genome-scale models are often focused on gene knockouts and constraining a particular reaction to a bound [66]. iJL1678-ME

can be used to provide new insights which cannot be currently be achieved with existing models; that is, iJL1678-ME can be used to estimate the detailed effects of molecular processes and physical parameters and on a much broader scale. This ability of iJL1678-ME will be demonstrated through two examples: Membrane crowding and Sec pathway inhibition.

**Assessing the consequences of membrane crowding**

Molecular crowding in the finite space of cells limits metabolic activity [6, 67]. Such crowding constraints are found both in the volume of the cell (also called packing' constraints) as well as the surface area of its membranes. iOL1650-ME, and consequently iJL1678-ME, implicitly considers volume crowding effects because density is constrained based on the overall growth rate [7]. Limited surface area in the membranes are thought to constrain major aspects of metabolism and physiology; for example, it may force *E. coli* to employ a mixture of respiration and fermentation to maximize growth rate [6, 68]. Thus, as part of the reconstruction process, a constraint on the fraction of protein in the membranes was incorporated into iJL1678-ME. This membrane constraint is mechanistic and imposed on a genome-scale, thereby representing a unique opportunity for a detailed assessment of the consequences of limited membrane space. The results of restricting the total surface area of integral membrane proteins in the model are described.

Computations of growth optimization were performed with constraints on the protein-to-lipid surface area ratios in both the inner and outer membranes. These computations revealed that the maximum growth rate was achieved when the fraction of membrane surface area occupied by protein was 42% and 25% for the

inner membrane and outer membrane, respectively. Furthermore, over- and under-production of membrane proteins did not affect the maximum growth rate with the same severity. The uneven slopes from the apex at 42% and 25% indicates that over-expression of membrane proteins may be less taxing on growth rate than under-expression, suggesting that it may in the cell's favor to over-produce membrane proteins than under-produce (Fig 4.9A).

As the inner membrane contains a diverse set of proteins that are important for metabolism, iJL1678-ME was used to examine the effects of spatial limitations on the inner membrane proteome. Although oxidative phosphorylation is much more efficient than alternate energy producing pathways, *E. coli* at high growth-rates and in excess glucose also employs fermentation pathways [69]. The electron transport system (ETS) is embedded in the membrane, and limited membrane space for the ETS may be why *E. coli* resorts to the mixed energy-production strategy [6]. iOL1650-ME, on the other hand, predicted that such a phenomenon occurs based on the trade-off between ATP generation and protein production costs [7].

In iJL1678-ME, acetate secretion has been almost eliminated compared to iOL1650-ME (8.1 to 1.5 mmol*gDW$^{-1}$*h$^{-1}$), due to the porin constraint. Differences in diffusion rates for each metabolite allowed the model to recognize that gases diffuse faster than solubilized carbon molecules, and complete metabolism of a carbon source becomes a better investment. However, fermentation returned when the inner membrane protein surface area decreased below 50%, as demonstrated by the increased secretion of acetate (Fig 4.9B). Within these regions of constraining protein-occupied surface area, the cell model produced less oxidative

phosphorylation products, which includes the ETS, instead of glucose PTS permeases and transporters for continued and increased glucose uptake, as previously hypothesized (Fig 4.9B & C) [6]. At extremely low surface areas allocated to proteins ($\leq$10%), there was not enough room to accommodate NADH dehydrogenase in the membrane. Instead, alternate dehydrogenases were expressed. Thus, to maximize growth rate, iJL1678-ME choses to increase fermentation rates with decreased membrane space.

Once membrane space permits complete metabolism of glucose influx at 50% protein-occupied surface area, fermentation pathways are no longer heavily employed which improves metabolic efficiency, hence the drop in in glucose uptake and increase oxygen uptake (Fig 4.9B). However, beyond 50%, iJL1678-ME makes a trade-off between producing more ETS, an expensive investment, to alternative proteins (Fig 4.9C). This shift in protein expression to accommodate the trade-off of ETS may play out similarly for proteins not required for metabolism, protein translocation, or metabolite transport but are essential for other processes (*e.g.*, expression of flagella for locomotion).

Where do *in vivo* cells fall along this scan across inner membrane occupancy? The calculated *in vivo* surface area of 28.5%, based on RNA-seq data, puts a cell below optimal membrane occupancy. Within this range of *in vivo* surface area, the increased acetate secretion hints that membrane space constraints may indeed be why cells employ combinatorial energy production pathways at maximum growth rates, as Zhuang et al. had hypothesized [6]. Furthermore, oxygen uptake drops severely when the protein surface area approaches the *in vivo* value of 28.5% (17 mmol*gDW$^{-1}$*h$^{-1}$ which is close to the measured values of

15 mmol*gDW$^{-1}$*h$^{-1}$ [70] and 18 mmol*gDW$^{-1}$*h$^{-1}$ [71]). This finding implies that a finite inner membrane protein surface area can limit the oxygen uptake and usage rate, thereby lowering the growth rate to less than the maximum potential.

**Perturbations in network performance by changing enzymatic efficiency**

The Sec pathway is a key pharmaceutical target due to its ubiquity and essentiality. For example, SecA is particularly attractive since it does not have a human homologue, and a recent non-cellular assay for SecA activity was developed specifically for drug discovery [72]. However, effects of decreased Sec translocase activity on a cell are largely unknown. While reactions in metabolic models can be capped to mimic protein inhibition, iJL1678-ME takes this ability further by targeting enzymatic efficiencies, similar to the effects of drugs. Thus, the impact of inhibiting Sec translocation on overall cellular phenotype was analyzed with iJL1678-ME by targeting key enzymes. SecA is the energy driver for the SecB/Sec pathway, and the ribosome is the energy driver for the SRP pathway. Together, these two pathways meet at SecYEGDF (Fig 4.1B). Due to their importance, these three proteins were inhibited.

When the $k_{cat}$ values of SecA, SecYEGDF, and the ribosome were reduced in a step-wise manner, growth rate was affected differently in each situation (Fig 4.10A). The relationship between ribosome inhibition and growth rate is nearly linear. SecA and SecYEGDF, on the other hand, behave in a hyperbolic manner. Thus, unlike ribosome, the activity of SecA or SecYEGDF must be nearly eliminated (*i.e.*, SecA<2.5%, SecYEGDF<5%) to reduce the growth rate by half. A closer look at these extremely low enzymatic rates reveals that the *in*

*silico* membrane proteome was dominated by SecYEGDF. Therefore, membrane occupancy was capped at 50%, as done by Zhuang et al [6], to determine whether spatial limitations may change the overall behavior to Sec pathway perturbations. The inhibition simulations were repeated, showing that ribosome was not affected by membrane limitations, while effects were observed when SecA and SecYEGDF's turnover rates dropped below two amino acids per second (Fig 4.11). However, regardless of membrane space, both SecA and SecYEGDF must be severely inhibited to significantly decrease growth rate. This example of targeting Sec translocation shows that iJL1678-ME can be used to discover cellular effects of selected perturbations. Other molecular behaviors, like combinatorial drug effects, may find similar answers through iJL1678-ME. For example, simultaneously targeting the two chaperone pathways for SecYEGDF, namely SecA and ribosome, is not a synergistic approach, and SecA must still be targeted for complete inhibition to significantly lower the growth rate (Fig 4.10B).

## 4.3 Conclusions

Taken in whole, iJL1678-ME stoichiometrically represents the wealth of knowledge known for protein translocation of *E. coli* in an integrated and computable format. For the first time, a bottom-up stoichiometric reconstruction (with turnover rates) predicted protein levels without expression data as inputs and imposed constraints. Furthermore, the ability to explicitly model protein translocation and compartmentalization of proteins is a significant advancement for genome-scale models, as it alleviates the need for fixed demands for the newly

reconstructed content. In combination with the membrane constraint, proteomic predictions represent a milestone for constraint-based modeling. As an example, iJL1678-ME could be utilized for designing fine-tuned engineered strains by identifying how the membrane proteome may react to overexpression of non-cytoplasmic proteins and for determining ways to counteract undesired effects through selective gene manipulation. Through exploration of modeled membrane formation contextualized within protein translocation and metabolism, iJL1678-ME demonstrated that bottom-up systems-biology can be used to predict and analyze cellular physiology, thereby providing an opportunity to assist and supplement research on fundamentally challenging areas which may otherwise be difficult to study.

Improvements in iJL1678-ME are likely to come through further experimental evidence. For example, more elucidation is required on the exact stoichiometry of TatA proteins per substrate and complex before such information can be incorporated into iJL1678-ME. Other ME-model based reconstructions may include a module to simulate plasmid induction and subsequent protein secretion. Finally, iJL1678-ME's predictive capabilities could be improved by incorporating data types such as ribosome profiling, quantitative proteomics, and additional $k_{cat}$ values. In conclusion, ME-models with compartmentalization and membrane constraints open exciting new avenues for the use of genome-scale models to interpret biological functions, to form the basis for strain designs, and understand infectious disease.

## 4.4 Methods

### 4.4.1 Reconstruction

A metabolism and gene expression model of *E. coli*, retroactively named here iOL1650-ME following an established convention [8], was used as the starting basis on which protein translocation reconstruction was built upon [7].

Literature review led to identification of five main translocation pathways plus three alternate assisting proteins. These pathways were developed into template reactions to which each of iJL1678-ME's pORFs could be applied to (see "Simplifed templates for translocation pathways" below).

Based on subcellular location annotations in Echolocation, EcoCyc and Uniprot (discrepancies and unknowns settled through PSORTb and TMHMM), all pORFs and protein complexes were assigned to one of four compartments: Cytosol, inner membrane, periplasm, and outer membrane [9, 37, 38, 39, 40]. The inner and outer membrane compartments are new additions to iOL1650-ME. New genes were also added to allow protein translocation and lipoprotein biogenesis. Reactions in iOL1650-ME were modified so that all proteins are compartmentalized. Furthermore, reactions were curated to ensure that reactions account for physical barrier membranes present. For example, if a reaction involves metabolites located in the cytoplasm and the periplasm, an inner membrane protein must be present for the reaction to occur.

Proteins with known experimental evidence were assigned to their respective translocase pathways. Based on these known peptides and current hypotheses, a set of rules was developed so that proteins without an experimentally-validated

pathway could be assigned to one. These rules were established primarily by annotated subcellular location and secondarily by the type of protein (Additional file 1 Table 1). However, each pathway operates at its own speed. iOL1650-ME's coupling constraints offer a solution for this problem, as the coupling constraints put limits on fluxes by linking reactions to enzyme degradation and the catalytic rate $k_{cat}$ [7]. Using this established constraint, turnover rates were applied to the translocase pathways to improve the model's ability to predict the membrane proteome (see Fig 4.3 for translocation without $k_{cat}$). Key proteins of each pathway had calculated turnover rates, and these $k_{cat}$ values were applied to all other enzymes in the pathway that have an interaction with that enzyme. The turnover rates of SecA, LolCDE, Bam, and Tat were all known from literature while the turnover rate for the SRP pathway was assumed to be equal to ribosome translation because of co-translational translocation [43, 44, 45, 46, 47]. For Tat-translocated proteins, a best fit polynomial equation for the number of TatA's verses average channel diameter was used to calculate the number of TatA's required for each [23]. Protein diameter was calculated by multiplying molecular weight by 1.21 to get volume and assuming a sphere shape [53]. Values were rounded up to the nearest integer.

Lipoprotein biogenesis was also determined to be relevant, and thus was included in the reconstruction process. The model has the flexibility to choose fatty acids from any available phospholipid. The proteins are modified by Lgt, Lsp, and Lnt to become lipoproteins.

Murein demand was adapted from the original iOL1650-ME model. However, since it is known that 45% of murein is recycled, the model is forced to utilize

the muropeptide transporter (AmpG), which has been implicated in the process of murein recycling [51], so that the flux of transported murein peptides is 45% of the murein demand (0.01389 mmol*gDW$^{-1}$).

**Simplified templates for translocation pathways**

*SRP/Sec pathway*

Ribosome-nascent_chain + SRP + GTP + FtsY-GTP + Sec_complex →
Ribosome-nascent_chain-Sec_complex (translation) + SRP + FtsY + 2GDP + 2Pi
→
Protein (inner membrane) + Ribosome + Sec_complex

*SecB/Sec pathway*

Peptide (cytosol) + SecB + SecA +Sec_complex + 1 ATP/25 aa → Peptide (periplasm) + SecB+ SecA + Sec_complex + (1ADP + 1Pi)/25 aa

*YidC insertion*

Ribosome-nascent_chain + SRP + YidC → Protein (inner membrane) + Ribosome + SRP + YidC

*Tat pathway*

Peptide (cytosol) + Tat_chaperone + TatBC + (1+)TatA → Protein (cytosol) + TatBC + (1+)TatA

*Lol pathway*

Lipoprotein (inner membrane) + LolCDE + LolA + ATP →

Lipoprotein-LolA (periplasm) + LolCDE + ADP + Pi + LolB →

Lipoprotein (outer membrane) + LolB + LolA

*Bam pathway*

Peptide (periplasm) + SurA + Bam_complex →

Peptide (outer membrane) + SurA + Bam_complex

*Simplified template reactions for lipoprotein biogenesis*

Peptide (inner membrane) + Lgt + pg →

Prolipoprotein + g3p + Lgt + LspA →

Apolipoprotein + LspA+ amide_linked_fatty_acid + Lnt → Lipoprotein + apg + Lnt

### 4.4.2 Outer membrane porins

As many as $2x10^5$ porins have been determined to be in the outer membrane [10]. Thus, to accurately account for these pathways, the outer membrane porins were coupled with diffusion rates [49, 73, 74]. In iJL1678-ME, the $k_{cat}$ values of the outer membrane porins are individualized for every combination of solute and porin, producing unique reactions reflecting effective diffusion rates based on diameters of solute and porin (Additional file 1 Table 4). To calculate the concentration difference between the extra-cellular environment and the periplasm, only porins with calculated effective diameters remained in the model (Table 4.1). The diameters for all possible solutes were calculated using MarvinSketch assum-

ing (1) the solutes were suspended in water (solvent radius: 1.4 Å) and (2) the solvent accessible surface area was a sphere, MarvinSketch 6.1.0, 2013, ChemAxon (http://www.chemaxon.com). With all the values known and inputted, this leaves the concentration difference between the extracellular ($C_e$) and periplasm ($C_p$), $C_e$-$C_p$, as the sole variable. Using an initial batch culture simulation in glucose M9 minimal media with the assumption $C_p << C_e$, the total flux of metabolite passage through outer membrane porins was calculated. Using iJL1678-ME's flux results of outer membrane trafficking, the known number of porins ($2x10^5$ per cell) [10], the solute diffusion rate through porins, and the porin constrain equations, a series of simulations with varying total solute concentration differences were run to estimate the approximate difference to such that number of porins produced equals the experimental value [75]. This concentration difference, $6.5x10^{-4}$ was incorporated into the porin diffusion rates as the default value, which may be adjusted by the user.

**Formulation for diffusion rates through outer membrane porins**

Variable definitions

$P$ Permeability coefficient

$D$ Free diffusion coefficient

$d$ Thickness of the membrane

$a_o$ total cross-sectional areas of all pores

$A$ Total area of the outer membrane

$V$ Rate of diffusion of solutes across the outer membrane

$C_e$ Concentration of the extra-cellular solutes

$C_p$ Concentration of periplasmic solutes

$R$ Gas constant

$T$ Temperature

$N_A$ Avogadro's number

$n$ Dynamic viscosity

$r$ Radius of Solute

$R$ Radius of pore

$g$ Growth rate

Theoretical permeability coefficient

$$P_{theory} = \frac{D}{d} * \frac{a_o}{A} * \frac{a}{a_o}$$

Ficks first law

$$V = P * A * (C_e - C_p)$$

Renkin equation

$$\frac{a}{a_o} = \left(1 - \frac{r}{R}\right)^2 * \left(1 - 2.104 * \frac{r}{R} + 2.09 * \left(\frac{r}{R}\right)^3 - 0.95 * \left(\frac{r}{R}\right)^5\right)$$

Stokes-Einstein

$$D = \frac{R*T}{N_A*6*n*\pi*r}$$

Flux of solute through porins of type i

$$V_{solute} = \frac{V_i*T*R^3*\left(1-2.104*\frac{r}{R}+2.09*\left(\frac{r}{R}\right)^3-0.95*\left(\frac{r}{R}^5\right)\right)*(C_e-C_p)}{6*n*r*d*g*1200}$$

### 4.4.3 Updating parameters

In order to determine how much more cellular mass iJL1678-ME explicitly accounts for, RNA-seq was first assumed to be a one-to-one proxy for protein expression levels, and in this dataset, the new pORFs and outer membrane proteins

summed to 9.5% of all proteomic mass. As a comparison point, the outer membrane protein mass (*i.e.*, lipoproteins and porins) was experimentally derived to be 7.4% of total proteomic mass [10]. Supplementing 7.4% with the estimated mass of protein translocases and lipoprotein biogenesis proteins from RNA-seq (as there were no experimental protein estimates available in literature) summed to 9.2% of total proteomic mass, which is similar to 9.5%.

The GAM (growth associated maintenance) was updated to account for the amount of ATP used in protein translocation. The ATP flux used in protein translocation by SecA and LolCDE was calculated and subtracted from the GAM value established in iOL1650-ME, reducing it from 35 to 34.98.

### 4.4.4 Membrane constraints

The combined surface area (SA) of membrane proteins, phospholipids (PE is phosphatidylethanolamine, PG is phosphatidylglycerol, and CLPN is cardiolipin), and lipopolysaccharides (LPS) must equal the total surface area of a cell (equation 1) times four membrane leaflets (equation 2) [55, 76]. The surface area of each membrane molecule was determined by its classification (Additional file 1 Table 5). If the molecule was a protein, the protein was assumed to extend through the lipid bilayer and occupy twice the amount of calculated surface area. An additional constraint was imposed so that phospholipid composition would better reflect the diversity of known membranes (equation 3).

(1) $SA(\mu) = 0.456\pi * 2^{\frac{\mu * ln(2)}{3}} * \left(3.9 * 2^{\frac{\mu * ln(2)}{3}} - 0.456 * 2^{\frac{\mu * ln(2)}{3}}\right) + \left(0.912\pi * 2^{\frac{\mu * ln(2)}{3}}\right)^2$

(2) $4 * SA(\mu) = \Sigma_{i \in proteins} SA\, of\, membrane\, protein_i + \Sigma SA\, of\, LPS +$

$\Sigma SA\,of\,phospholipids$

(3) $\Sigma SA\,of\,phospholipids = 77\% * \Sigma SA_{PE} + 18\% * \Sigma SA_{PG} + 5\% * \Sigma SA_{CLPN}$

Additional constraints in the iJL1678-ME include a variable maximum cap on protein surface area and the option to force the model to produce nonfunctional membrane protein.

This cell envelope demand for LPS and lipids originally appearing in iOL1650-ME was removed, which makes the production of these two types of molecules a function of growth rate, protein production, and membrane size. Membrane size was taken to growth-rate dependent as formulated by O'Brien et al (see [7] supplemental materials).

### 4.4.5 Analyzing the model

The model was run using batch simulations, as described by O'Brien et al using resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC0205CH11231 [7]. For all analyses performed, the *in silico* media composition was M9 and an excess of glucose (4 g*$L^{-1}$) (Additional file 1 Table 6).

Since membrane proteomics is difficult to study; it is even more difficult to obtain absolute numbers comparing relative ratios of protein amounts. Therefore, RNA-seq was used as an *in vivo* proxy for comparison (GEO accessions: GSE48324 [59] and GSE55367 [60]). A 1:1 ratio of protein expression levels to RNA-seq levels (FPKM normalized to overall expression) was assumed. Mass was calculated based on the atomic mass of the primary protein structure multiplied by the flux of

protein being produced. In comparing *in vivo* data to *in silico* data, mass was summed up by compartment location, functional annotation, or both (Additional file 1 Table 2). Error bars are 1 standard deviation from two RNAseq runs.

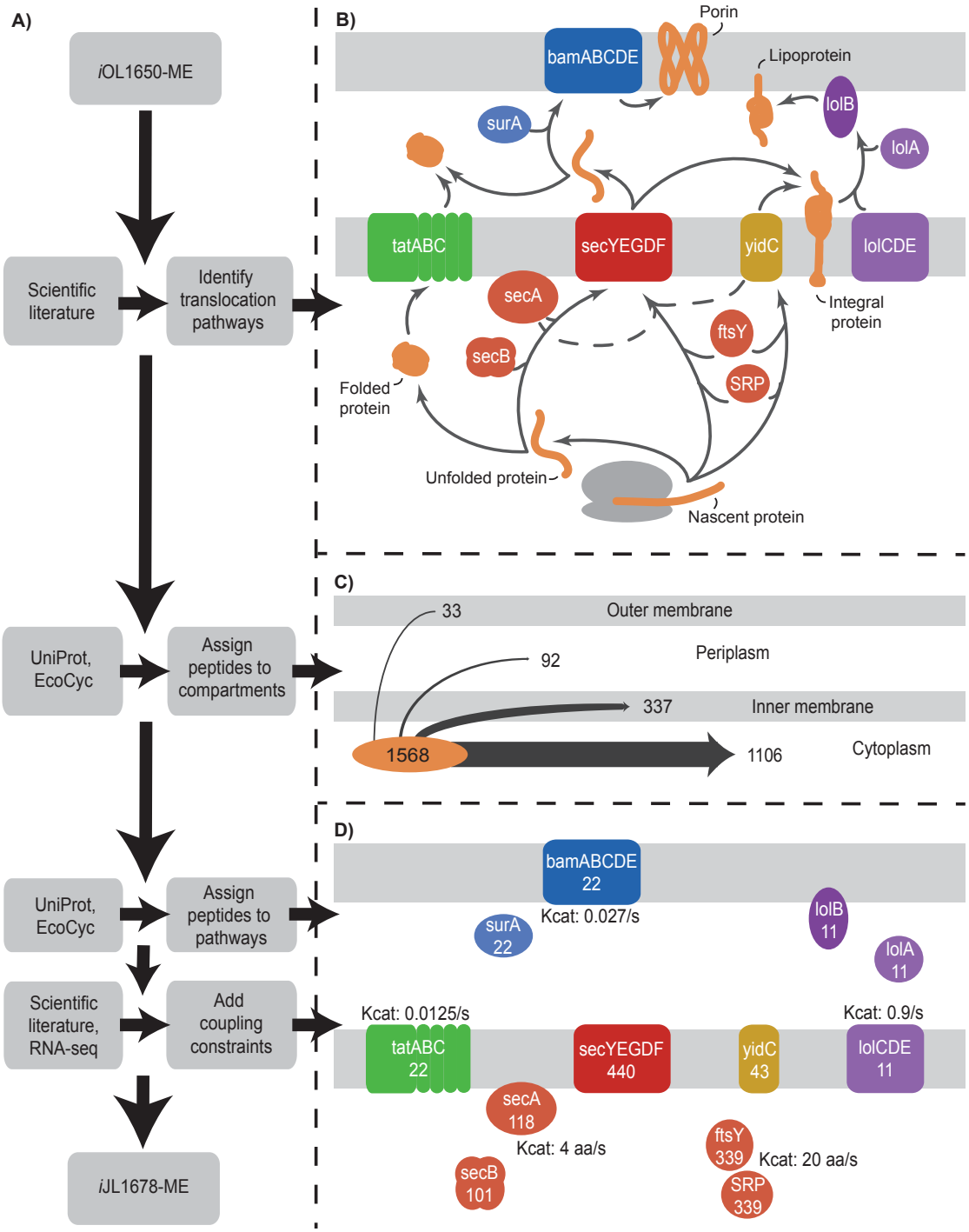
The mass of compartmentalized functional annotations between *in vivo* and *in silico* data was compared on a log-log basis. A simple linear regression model was calculated between the two datasets. The standardized residuals (residual i / standard deviation of residual i) of the *in silico* data was plotted against a rankit score (expected values of the order statistics if the sample is normally distributed), creating a normal probability plot. A line passing through the first and third quartiles revealed points that deviated from a normal distribution (*i.e.*, deviated from the quartile line). These points were removed from the dataset for further analysis and the simple linear regression model was recalculated for the reduced dataset.

### 4.4.6 Protein inhibition

To adjust the turnover rate of SecA, the coupling constraint was modified so that it would reflect numbers lower than the published value of 4.0 $s^{-1}$ [44]. Similarly, all coupling constraints involved with SecYEGDF or ribosome were multiple by fractions to lower enzyme efficiencies. To limit membrane inner membrane protein surface area, the variable maximum cap (included as part of the membrane constraint formulation) was set to 0.5.

## 4.5 Figures and tables

**Figure 4.1*(next page)***: **Workflow utilized and resulting network for reconstructing protein translocation in *E. coli*.** (A) An outline of the workflow used to reconstruct the protein translocation network in *E. coli*. At each step, various sources of data were used as inputs to the workflow. The resulting general network, compartmentalized content, and pathway breakdown are shown in greater detail to the right. (B) A diagram of the translocation pathways included in the reconstruction: SRP/Sec, SecB/Sec, Tat, YidC, Lol, Bam pathways, and three alternatives (dashed lines). Proteins that allow translocation are labeled in white while translocated protein types are labeled in black. Lipoprotein biogenesis is not depicted. (C) Model-simulated pORFs were assigned to one of four compartments. The numbers denote how many of the 1,568 proteins will end up in each compartment. (D) Each non-cytosolic pORF was assigned to a translocation pathway. Numbers in white are how many pORFs require that translocation-associated protein. The model also underwent several other updates, including the addition of known turnover rates that are denoted by black numbers.


A)
iOL1650-ME
Scientific literature
Identify translocation pathways
UniProt, EcoCyc
Assign peptides to compartments
UniProt, EcoCyc
Assign peptides to pathways
Scientific literature, RNA-seq
Add coupling constraints
iJL1678-ME
B)
Porin
bamABCDE
Lipoprotein
lolB
lolA
surA
tatABC
secYEGDF
yidC
lolCDE
Integral protein
secA
ftsY
secB
SRP
Folded protein
Unfolded protein
Nascent protein
C)
33
Outer membrane
92
Periplasm
337
Inner membrane
1568
1106
Cytoplasm
D)
bamABCDE 22
Kcat: 0.027/s
surA 22
lolB 11
lolA 11
Kcat: 0.0125/s
Kcat: 0.9/s
tatABC 22
secYEGDF 440
yidC 43
lolCDE 11
secA 118
Kcat: 4 aa/s
ftsY 339
Kcat: 20 aa/s
secB 101
SRP 339

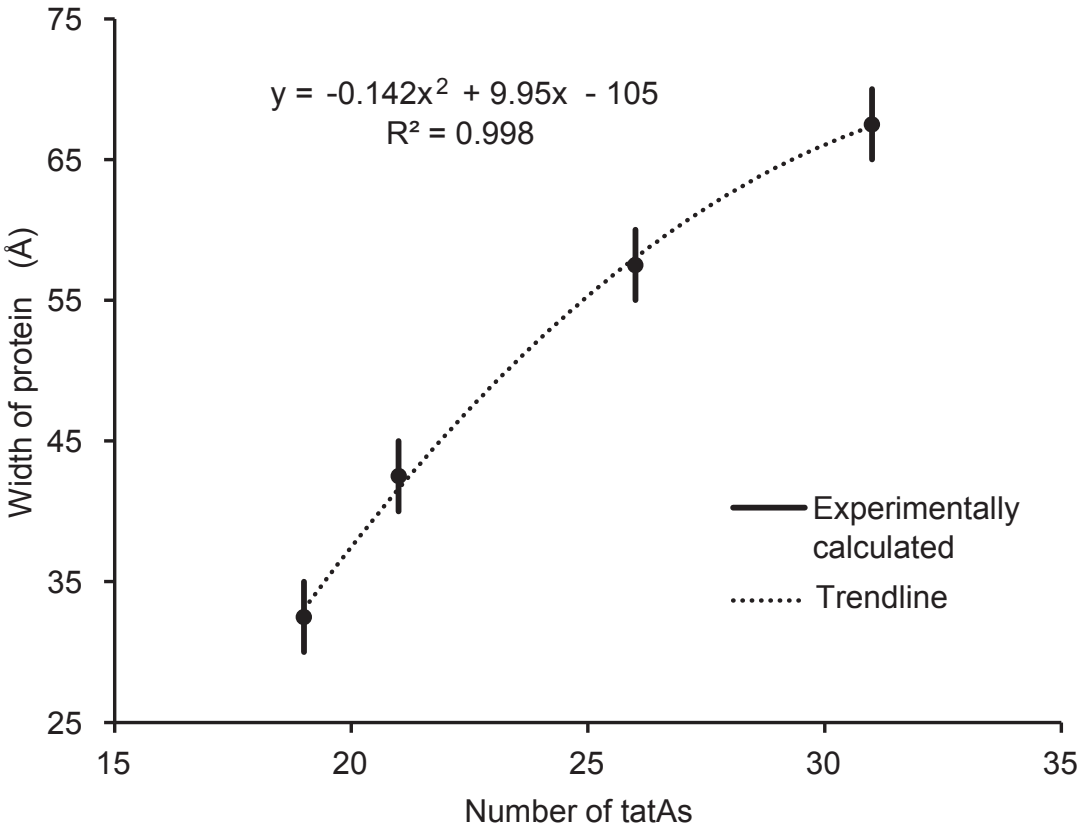



**Figure 4.2**: **Calculation of the number of TatA proteins required for each translocated protein.** Using data from [23], the diameter of the channel formed by TatA proteins was plotted to determine how many TatAs are required to transport a Tat-translocated protein. The estimated width of each Tat-translocated protein was calculated from their molecular weight, assuming a spherical shape [53]. The resulting value was plugged into the trendline equation. The number of required TatA proteins was rounded up to the nearest integer and inserted into the Tat-pathway template reactions.

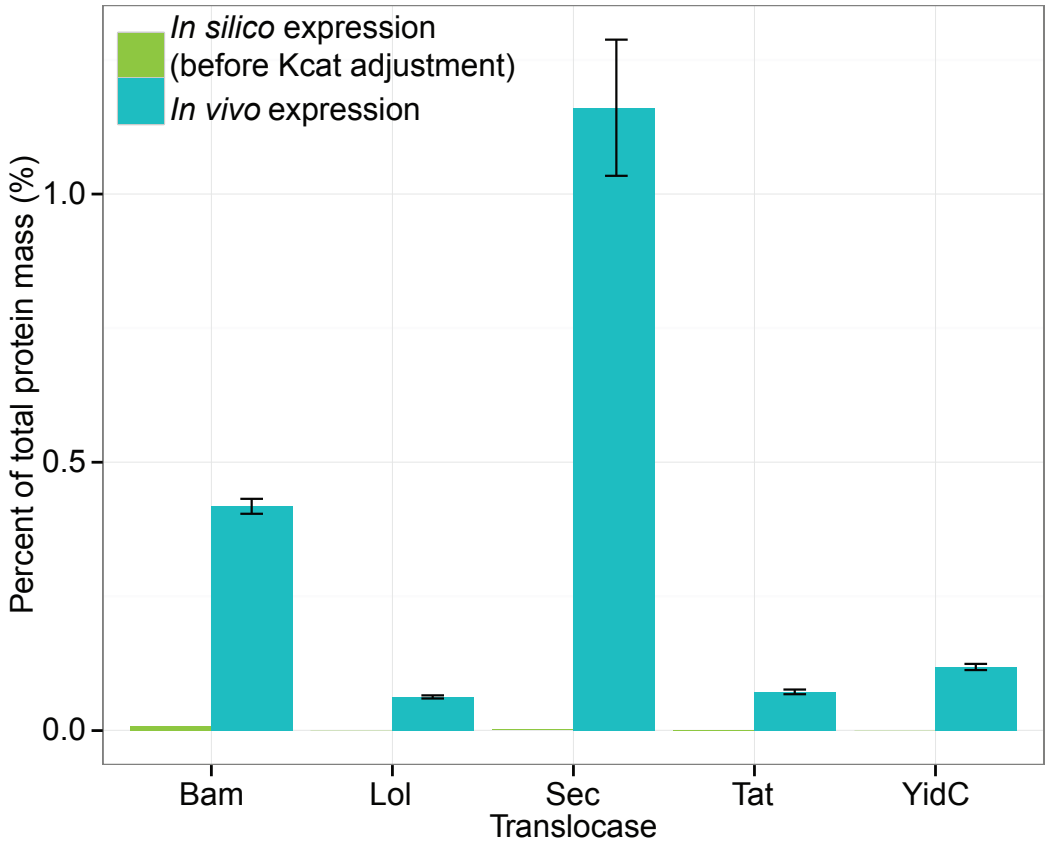



**Figure 4.3**: ***In silico* protein expression of translocase pathways before the addition of enzyme turnover rates.** A bar graph showing simulation results (green) of translocase pathway protein levels from iJL1678-ME without translocase turnover rates and measured *in vivo* expression levels (blue) using RNA-seq as a proxy for protein production ($R^2$=0.047, p-val=0.73). Results were taken from glucose M9 minimal media conditions.

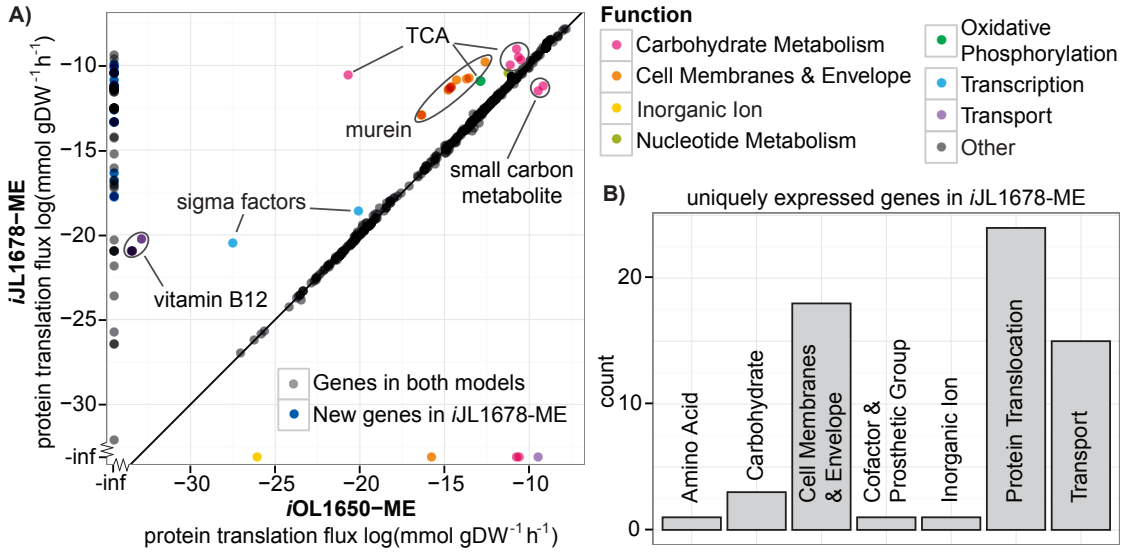



**Figure 4.4**: **Proteome expression comparison between iOL1650-ME and iJL1678-ME.** The difference that the protein translocation reconstruction brings to iOL1650-ME is compared through computed protein expression in glucose M9 minimal media conditions. (A) Protein translation flux between iJL1678-ME and iOL1650-ME. The majority of pORF expression (94%) are approximately the same in both model simulations, but 4% are uniquely expressed in iJL1678-ME, and 0.8% is uniquely expressed in iOL1650-ME (points along the -inf line). 1.5% of the proteins are differentially expressed, the majority of which are expressed to a greater extent in iJL1678-ME than in iOL1650-ME, but two proteins involved in small carbon metabolism (EutD and PurT) are expressed lower. (B) Histograms detailing the functional annotations of the uniquely expressed genes within the two models.

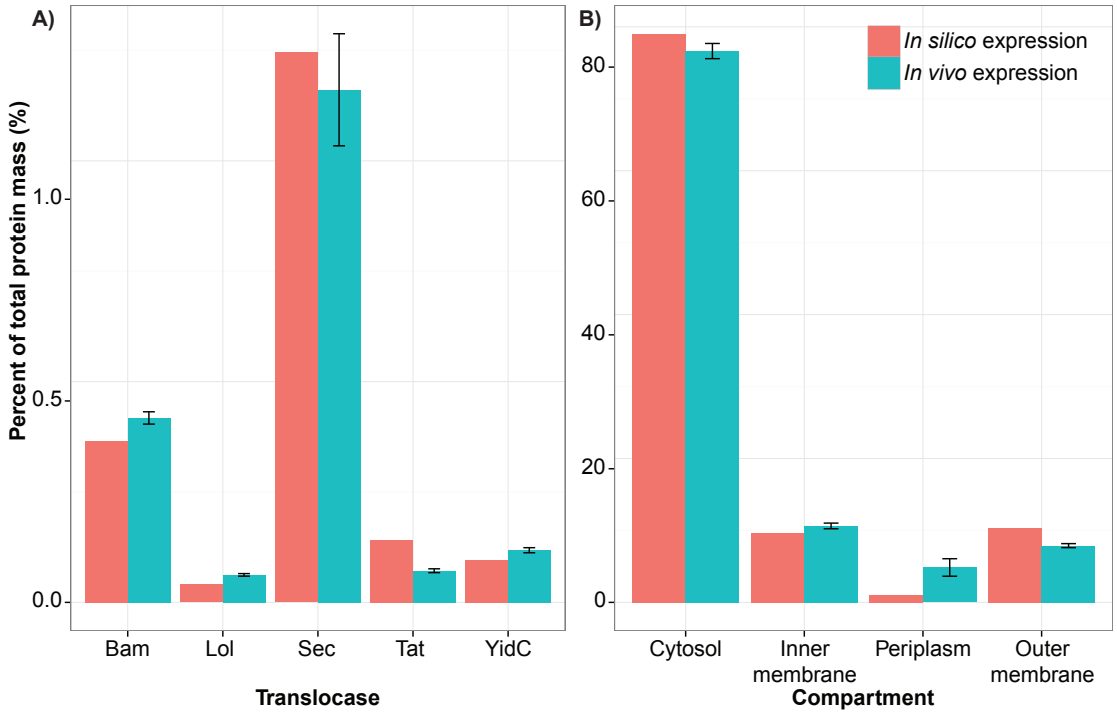



**Figure 4.5**: **Comparison of *in silico* predicted protein masses verses *in vivo* measurements for reconstructed content specific to iJL1678-ME.** Bar graphs showing simulation results (pink) of protein levels from the reconstructed iJL1678-ME verses measured *in vivo* expression levels (blue) using averaged RNA-seq as a proxy for protein production. Results were taken from glucose M9 minimal media conditions. (A) Translocase protein levels. (B) Percentage of protein mass in each of the four compartments.

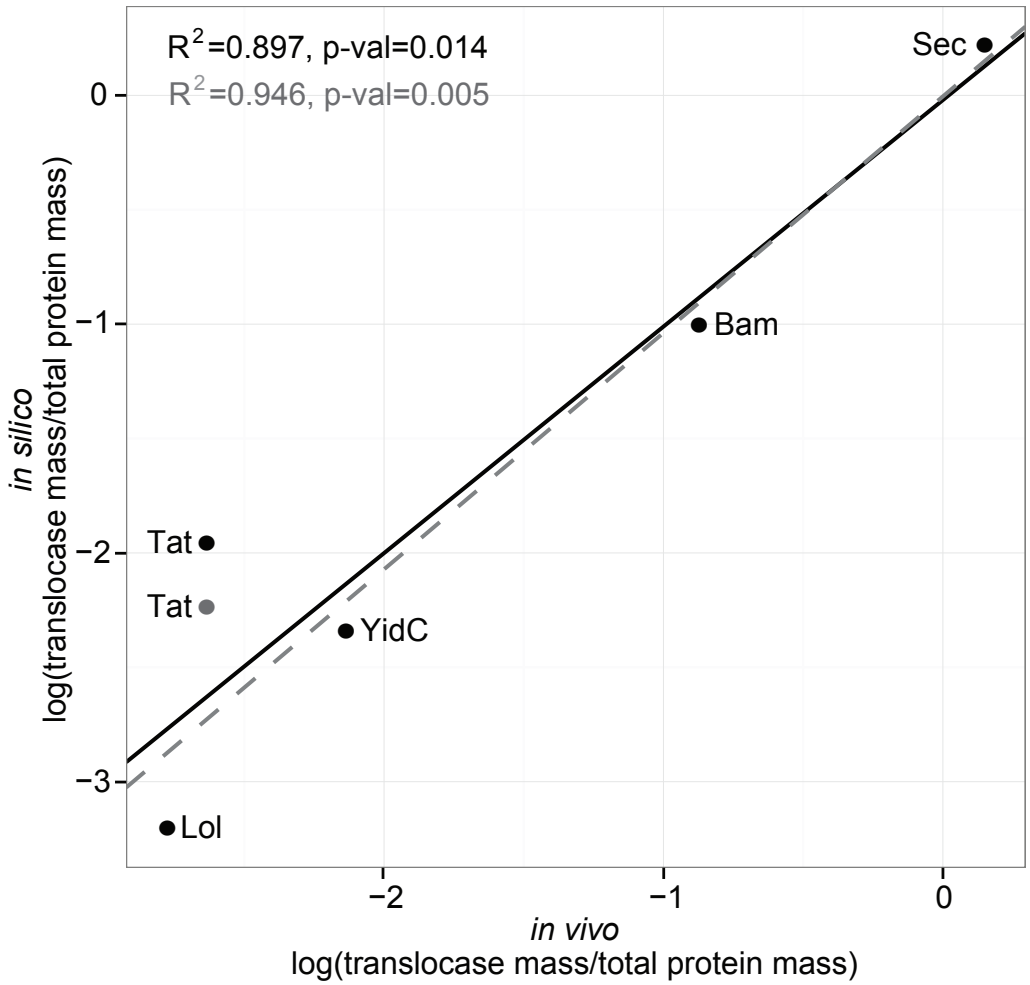



**Figure 4.6**: **Comparison of *in silico* verses *in vivo* protein expression of translocase pathways.** Shown is a scatterplot comparing *in silico* and *in vivo* translocase protein levels. Gray represents new calculations when the mass of TatBC is lowered four-fold to account for TatBC's ability to simultaneously translocate multiple substrates.

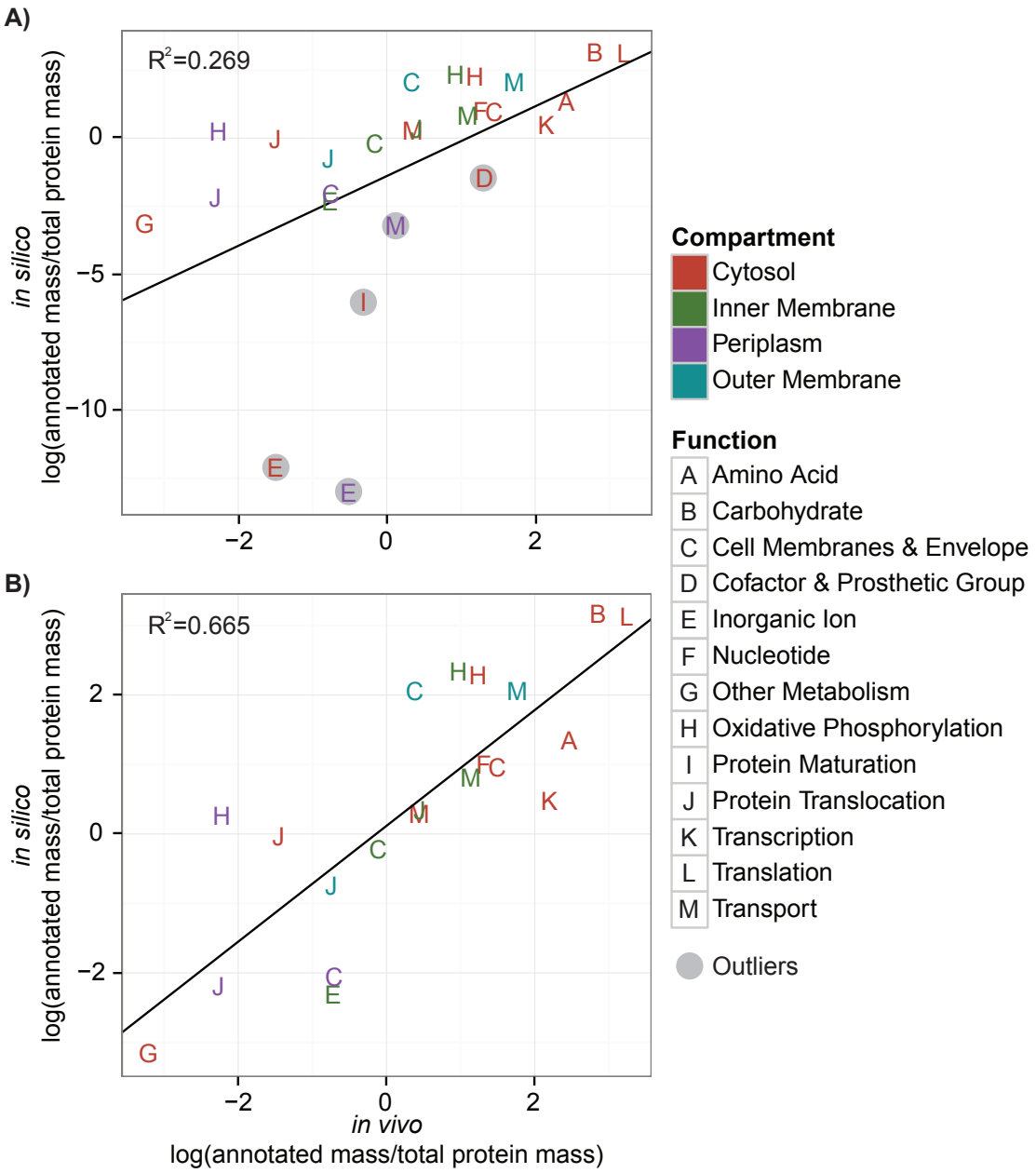



**Figure 4.7**: **Analysis of *in silico* predicted protein masses verses *in vivo* measurements.** Predicted (*in silico* ) versus measured (*in vivo*) protein masses that were reconstructed in iJL1678-ME were categorized by function and compartment. Subgroups with zero values were removed from further calculations. (A) The linear model between *in silico* and *in vivo* protein mass predictions (p-value=$6.6\text{x}10^{-3}$). The outliers had standardized residues that fell outside of the normal distribution curve as formed by the other points (Fig 4.8). (B) The outliers were removed, and the linear model between *in silico* and *in vivo* protein mass predictions was recalculated (p-value=$6.6\text{x}10^{-6}$).

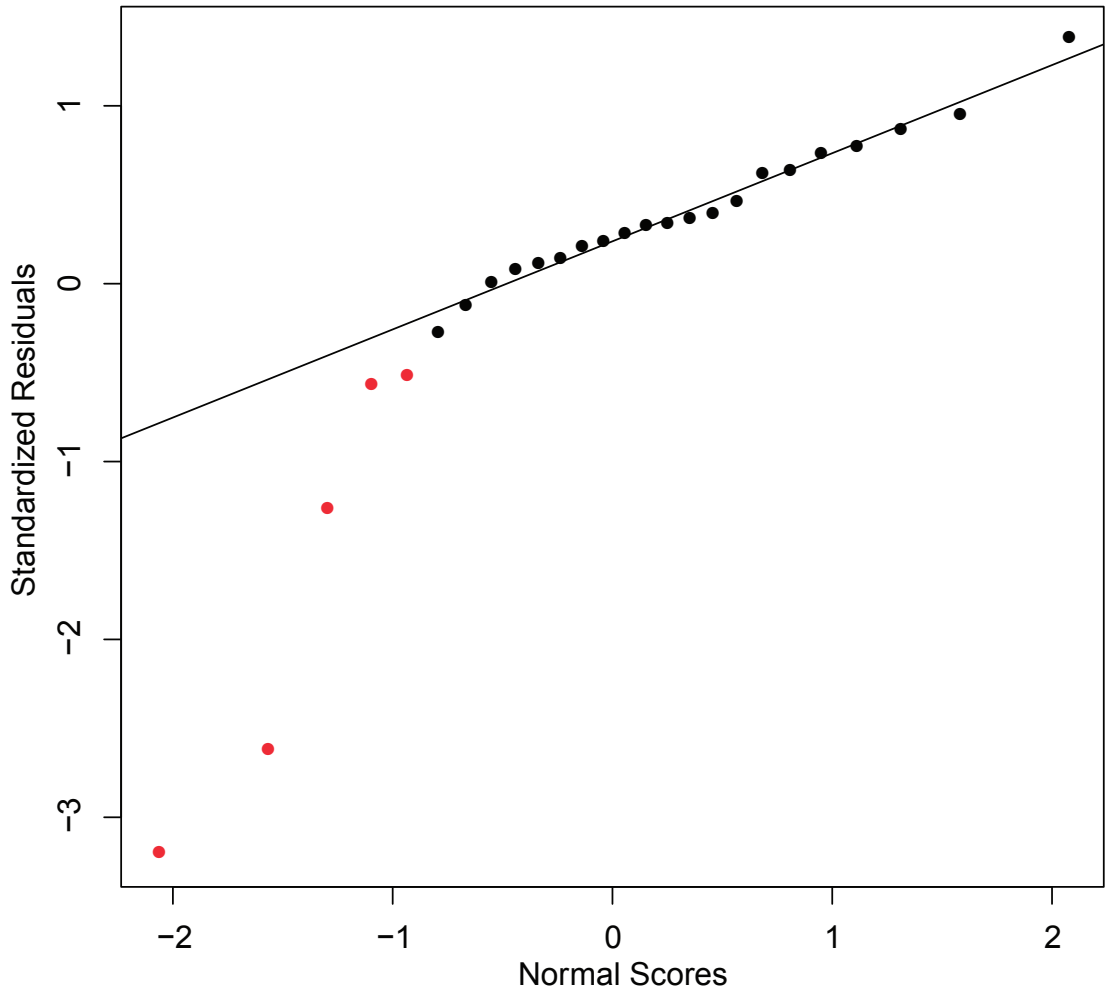
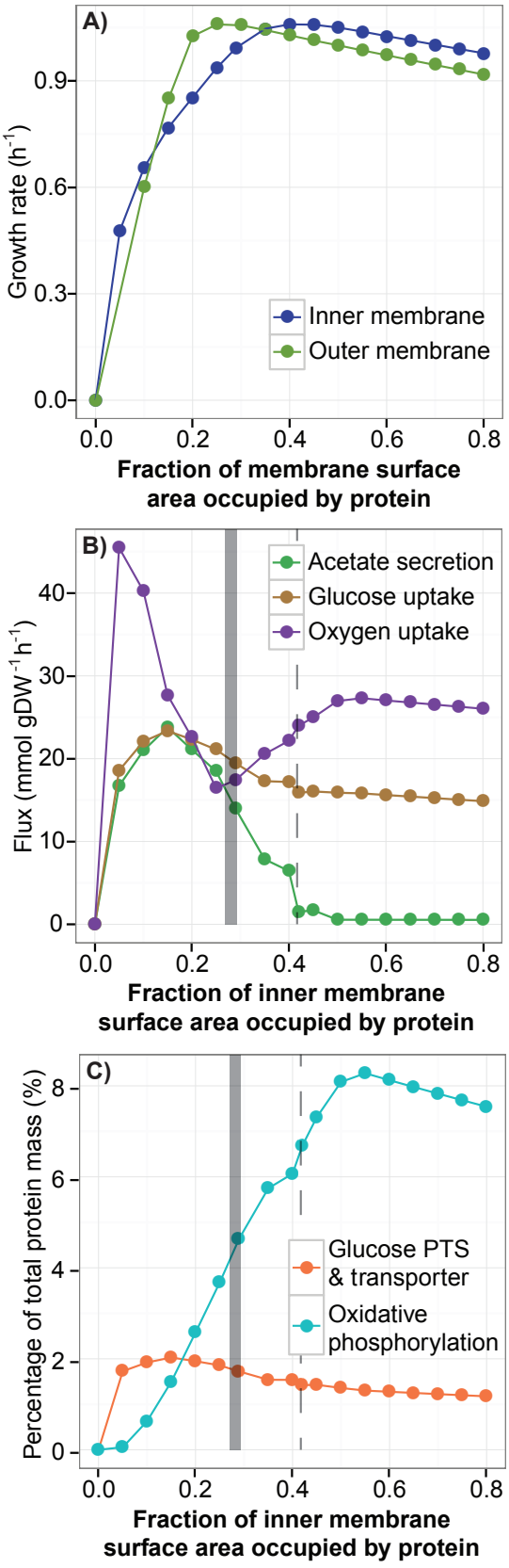



**Figure 4.8**: **Additional data for the linear model analysis.** Points represent predicted (*in silico*) versus measured (*in vivo*) protein masses categorized by function and compartment for the proteins which were reconstructed in iJL1678-ME. The normal probability plot of rankit scores against standardized residuals of a linear regression over all data points demonstrates that several points (red) do not fall within a normal distribution.

**Figure 4.9*(next page)***: **Effects of constraining the amount of membrane surface area that may be occupied by protein.** Shown here is a scatterplot comparing the effects of controlled protein occupancy in the membranes. (A) The effects of constraining the protein surface area in the inner and outer membranes. The apex of growth rate occurs at 0.42 fractional area for protein occupancy for the inner membrane and 0.25 for the outer membrane. The growth rate decreases more rapidly if membranes protein were under-produced verses over-produced. (B) Acetate secretion, glucose uptake, and oxygen uptake fluxes when constraining inner membrane protein surface area. The gray solid bar represents the calculated *in vivo* surface area (+/- one standard deviation), and the dashed line represents the optimal inner membrane surface area occupancy. (C) Mass of the electron transport system complexes and glucose transporters when constraining inner membrane protein surface area.


A)
Growth rate (h-1)
0.0
0.3
0.6
0.9
Inner membrane
Outer membrane
0.0
0.2
0.4
0.6
0.8
Fraction of membrane surface area occupied by protein
B)
Flux (mmol gDW-1 h-1)
0
10
20
30
40
Acetate secretion
Glucose uptake
Oxygen uptake
0.0
0.2
0.4
0.6
0.8
Fraction of inner membrane surface area occupied by protein
C)
Percentage of total protein mass (%)
0
2
4
6
8
Glucose PTS & transporter
Oxidative phosphorylation
0.0
0.2
0.4
0.6
0.8
Fraction of inner membrane surface area occupied by protein

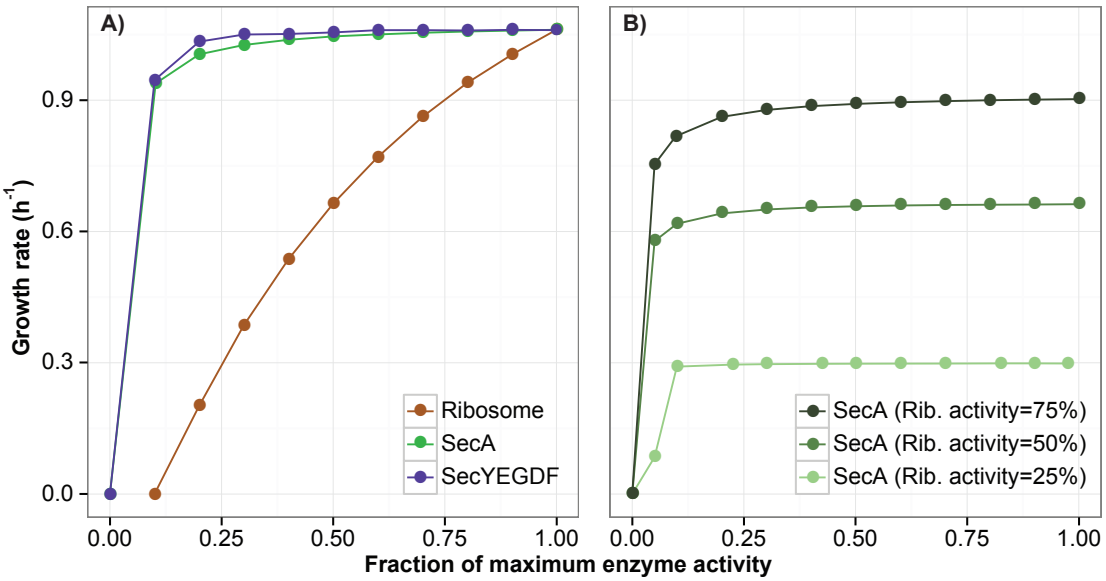



**Figure 4.10**: **Effects of inhibiting SecA on growth rate.** (A) A scatterplot showing the effects of decreasing enzyme efficiency of several key enzymes involved in Sec translocation (ATPase SecA, the channel SecYEGDF, and ribosome) have on growth rate. The growth rate was predicted by decreasing turnover rate (*i.e.*, $k_{cat}$) of SecA, SecYEGDF, and ribosome and optimizing for growth rate. Simulations were performed with an upper limit of 0.5 of the membrane protein surface area occupancy. (B) The effects of simultaneously inhibiting SecA and ribosome.

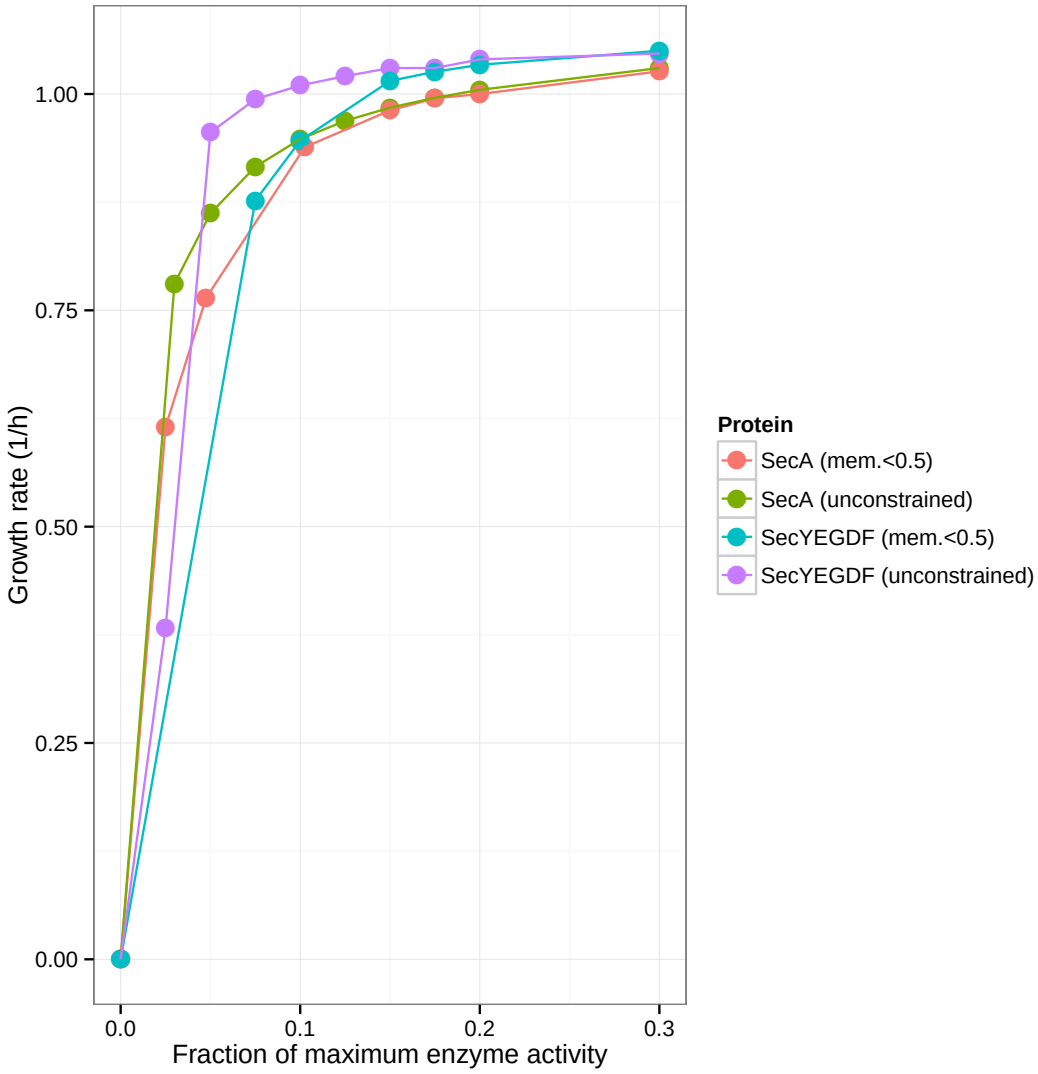



**Figure 4.11**: **Effects of limiting the Sec pathway with membrane limitations.** Shown is a plot comparing SecA and SecYEGDF inhibition while the membrane is and is not constrained to 0.5. The membrane constraint affects overall growth rate at very low enzymatic levels (<0.2). Ribosome inhibition is not shown since the membrane constraint does not affect simulation results.

**Table 4.1**: Outer membrane porin effective diameters

| **Porin** | **Effective diameter** | **Reference** |
|---|---|---|
| ompA | 1.0 nm | Sugawara and Nikaido 1991 |
| ompC | 0.54 nm | Nikaido and Rosenberg 1982 |
| ompF | 0.58 nm | Nikaido and Rosenberg 1982 |

## 4.6 Acknowledgments

Chapter 4, in full, is a modified reprint of the material as it appears in *BMC Systems Biology* 2014. Joanne K Liu, Edward J O'Brien, Joshua A Lerman, Karsten Zengler, Bernhard O Palsson and Adam M Feist. Reconstruction and modeling protein translocation and compartmentalization in *Escherichia coli* at the genome-scale. *BMC Systems Biology* 2014 8:110. The dissertation author was a primary investigator and author of this paper. Many thanks to Ali Ebrahim for his invaluable assistance during the startup of the experiment and his continuous support. We also thank Haythem Latif and Gabriela Guzman for producing the RNA-seq data.

## 4.7 References

[1] Ross E. Dalbey, Peng Wang, and Andreas Kuhn. Assembly of Bacterial Inner Membrane Proteins. *Annual Review of Biochemistry*, 80(1):161–187, 2011.

[2] Hiroshi Nikaido. Molecular basis of bacterial outer membrane permeability revisited. *Microbiology and molecular biology reviews : MMBR*, 67(4):593–656, 2003.

[3] Jonathan R Karr, Jayodita C Sanghvi, Derek N Macklin, Miriam V Gutschow, Jared M Jacobs, Benjamin Bolival, Nacyra Assad-Garcia, John I Glass, and Markus W Covert. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401, 2012.

[4] Joen Luirink and Irmgard Sinning. SRP-mediated protein targeting: structure and function revisited. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 2004.

[5] Amir Feizi, Tobias Österlund, Dina Petranovic, Sergio Bordel, and Jens

Nielsen. Genome-Scale Modeling of the Protein Secretory Machinery in Yeast. *PLoS ONE*, 8(5):e63284, 2013.

[6] Kai Zhuang, Goutham N Vemuri, and Radhakrishnan Mahadevan. Economics of membrane occupancy and respiro-fermentation. *Molecular Systems Biology*, 7:500–500, 2011.

[7] Edward J O'Brien, Joshua A Lerman, Roger L Chang, Daniel R Hyduke, and Bernhard Ø Palsson. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Molecular systems biology*, 9: 693, 2013.

[8] Jennifer L Reed, Thuy D Vo, Christophe H Schilling, and Bernhard O Palsson. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome biology*, 4(9):R54, 2003.

[9] Anders Krogh, Björn Larsson, Gunnar von Heijne, and Erik L.L Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. *Journal of Molecular Biology*, 305(3):567–580, 2001.

[10] Frederick C. Neidhardt and Roy Curtiss. *Escherichia coli and Salmonella : cellular and molecular biology*. ASM Press, 1996. ISBN 1555810845.

[11] Martine P. Bos, Viviane Robert, and Jan Tommassen. Biogenesis of the Gram-Negative Bacterial Outer Membrane. *Annual Review of Microbiology*, 61(1): 191–214, 2007.

[12] Mechthild Pohlschröder, Will A. Prinz, Enno Hartmann, and Jon Beckwith. Protein Translocation in the Three Domains of Life: Variations on a Theme. *Cell*, 91(5):563–566, 1997.

[13] Sandra Angelini, Sandra Deitermann, and Hans-Georg Koch. FtsY, the bacterial signal-recognition particle receptor, interacts functionally and physically with the SecYEG translocon. *EMBO reports*, 6(5):476–481, 2005.

[14] David J. F. du Plessis, Greetje Berrelkamp, Nico Nouwen, and Arnold J. M.

Driessen. The Lateral Gate of SecYEG Opens during Protein Translocation. *Journal of Biological Chemistry*, 284(23):15805–15814, 2009.

[15] Edith N.G. Houben, Pier A. Scotti, Quido A. Valent, Josef Brunner, Jan-Willem L. de Gier, Bauke Oudega, and Joen Luirink. Nascent Lep inserts into the *Escherichia coli* inner membrane in the vicinity of YidC, SecY and SecA. *FEBS Letters*, 476(3):229–233, 2000.

[16] Nico Nouwen and Arnold J M Driessen. SecDFyajC forms a heterotetrameric complex with YidC. *Molecular microbiology*, 44(5):1397–405, 2002.

[17] Martin van der Laan, Philipp Bechtluft, Stef Kol, Nico Nouwen, and Arnold J.M. Driessen. F <sub>1</sub> F <sub>0</sub> ATP synthase subunit c is a substrate of the novel YidC pathway for membrane protein biogenesis. *The Journal of Cell Biology*, 165(2):213–222, 2004.

[18] Louise Baars, A. Jimmy Ytterberg, David Drew, Samuel Wagner, Claudia Thilo, Klaas Jan van Wijk, and Jan-Willem de Gier. Defining the Role of the *Escherichia coli* Chaperone SecB Using Comparative Proteomics. *Journal of Biological Chemistry*, 281(15):10024–10034, 2006.

[19] Anastassios Economou and William Wickner. SecA promotes preprotein translocation by undergoing ATP-driven cycles of membrane insertion and deinsertion. *Cell*, 78(5):835–843, 1994.

[20] Eva M Murén, Dominic Suciu, Traci B Topping, Carol A Kumamoto, and Linda L Randall. Mutational alterations in the homotetrameric chaperone SecB that implicate the structure as dimer of dimers. *The Journal of biological chemistry*, 274(27):19397–402, 1999.

[21] Franz-Ulrich Hartl, Stewart Lecker, Elmar Schiebel, Joseph P. Hendrick, and William Wickner. The binding cascade of SecB to SecA to SecYE mediates preprotein targeting to the E. coli plasma membrane. *Cell*, 63(2):269–279, 1990.

[22] Claire-Lise Santini, Bérengère Ize, Angélique Chanal, Matthias Müller, Gérard Giordano, and Long-Fei Wu. A novel Sec-independent periplasmic protein

translocation pathway in Escherichia coli. *The EMBO Journal*, 17(1):101–112, 1998.

[23] Ulrich Gohlke, Lee Pullan, Christopher A. McDevitt, Ida Porcelli, Erik de Leeuw, Tracy Palmer, Helen R. Saibil, and Ben C. Berks. The TatA component of the twin-arginine protein transport system forms channel complexes of variable diameter. *Proceedings of the National Academy of Sciences*, 102(30):10482–10486, 2005.

[24] Umesh K. Bageshwar and Siegfried M. Musser. Two electrical potentialdependent steps are required for transport by the *Escherichia coli* Tat machinery. *The Journal of Cell Biology*, 179(1):87–99, 2007.

[25] Linda Fröderberg, Edith N. G. Houben, Louise Baars, Joen Luirink, and Jan-Willem de Gier. Targeting and Translocation of Two Lipoproteins in *Escherichia coli* via the SRP/Sec/YidC Pathway. *Journal of Biological Chemistry*, 279(30):31026–31032, 2004.

[26] Krishnan Sankaran and Henry C Wu. Lipid modification of bacterial prolipoprotein. Transfer of diacylglyceryl moiety from phosphatidylglycerol. *The Journal of biological chemistry*, 269(31):19701–6, 1994.

[27] Eduard Bitto and David B. McKay. The Periplasmic Molecular Chaperone Protein SurA Binds a Peptide Motif That Is Characteristic of Integral Outer Membrane Proteins. *Journal of Biological Chemistry*, 278(49):49316–49322, 2003.

[28] J. G. Sklar, T. Wu, D. Kahne, and T. J. Silhavy. Defining the roles of the periplasmic chaperones SurA, Skp, and DegP in Escherichia coli. *Genes & Development*, 21(19):2473–2484, 2007.

[29] Juliana C Malinverni and Thomas J Silhavy. Assembly of Outer Membrane $\beta$-Barrel Proteins: the Bam Complex. *EcoSal Plus*, 4(2), 2011.

[30] Sandra Deitermann, Grit Sophie Sprie, and Hans-Georg Koch. A Dual Function for SecA in the Assembly of Single Spanning Membrane Proteins in *Escherichia coli*. *Journal of Biological Chemistry*, 280(47):39077–39085, 2005.

[31] Jan Schellenberger, Junyoung O Park, Tom M Conrad, and Bernhard Ø Palsson. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC bioinformatics*, 11(1):213, 2010.

[32] James C. Samuelson, Minyong Chen, Fenglei Jiang, Ines Möller, Martin Wiedmann, Andreas Kuhn, Gregory J. Phillips, and Ross E. Dalbey. YidC mediates membrane protein insertion in bacteria. *Nature*, 406(6796):637–641, 2000.

[33] Alain Bernadac, Marthe Gavioli, Jean-Claude Lazzaroni, Satish Raina, and Roland Lloubès. Escherichia coli tol-pal mutants form outer membrane vesicles. *Journal of bacteriology*, 180(18):4872–8, 1998.

[34] J. G. Sklar, T. Wu, L. S. Gronenberg, J. C. Malinverni, D. Kahne, and T. J. Silhavy. Lipoprotein SmpA is a component of the YaeT complex that assembles outer membrane proteins in Escherichia coli. *Proceedings of the National Academy of Sciences*, 104(15):6400–6405, 2007.

[35] Hajime Tokuda and Shin-Ichiro Narita. Biogenesis and Membrane Targeting of Lipoproteins. *EcoSal Plus*, 4(1), 2010.

[36] Ines Thiele, Neema Jamshidi, Ronan M. T. Fleming, Bernhard Ø. Palsson, and P Stothard. Genome-Scale Reconstruction of *Escherichia coli's* Transcriptional and Translational Machinery: A Knowledge Base, Its Mathematical Formulation, and Its Functional Characterization. *PLoS Computational Biology*, 5(3):e1000312, 2009.

[37] Richard S. P. Horler, Andrew Butcher, Nikitas Papangelopoulos, Peter D. Ashton, and Gavin H. Thomas. EchoLOCATION: an in silico analysis of the subcellular locations of Escherichia coli proteins and comparison with experimentally derived locations. *Bioinformatics*, 25(2):163–166, 2009.

[38] The UniProt Consortium. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Research*, 41(D1):D43–D47, 2013.

[39] Ingrid M Keseler, Julio Collado-Vides, Socorro Gama-Castro, John Ingraham, Suzanne Paley, Ian T Paulsen, Martín Peralta-Gil, and Peter D Karp. EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic acids research*,

33(Database issue):D334–7, 2005.

[40] Nancy Y. Yu, James R. Wagner, Matthew R. Laird, Gabor Melli, Sébastien Rey, Raymond Lo, Phuong Dao, S. Cenk Sahinalp, Martin Ester, Leonard J. Foster, and Fiona S. L. Brinkman. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, 26(13):1608–1615, 2010.

[41] Anders Krogh, Björn Larsson, Gunnar von Heijne, and Erik L.L Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology*, 305(3):567–580, 2001.

[42] Elmar Schiebel, Arnold J.M. Driessen, Franz-Ulrich Hartl, and William Wickner. $\Delta\mu$H+ and ATP function at different steps of the catalytic cycle of preprotein translocase. *Cell*, 64(5):927–939, 1991.

[43] Patrick P. Dennis and Hans Bremer. Modulation of Chemical Composition and Other Parameters of the Cell at Different Exponential Growth Rates. *EcoSal Plus*, 3(1), 2008.

[44] Danuta Tomkiewicz, Nico Nouwen, Ruud van Leeuwen, Sander Tans, and Arnold J. M. Driessen. SecA Supports a Constant Rate of Preprotein Translocation. *Journal of Biological Chemistry*, 281(23):15709–15713, 2006.

[45] Neal Whitaker, Umesh K. Bageshwar, and Siegfried M. Musser. Kinetics of Precursor Interactions with the Bacterial Tat Translocase Detected by Real-time FRET. *Journal of Biological Chemistry*, 287(14):11252–11260, 2012.

[46] Kyoko Kanamaru, Naohiro Taniguchi, Shigehiko Miyamoto, Shin-ichiro Narita, and Hajime Tokuda. Complete reconstitution of an ATP-binding cassette transporter LolCDE complex from separately isolated subunits. *FEBS Journal*, 274(12):3034–3043, 2007.

[47] Christine L. Hagan and Daniel Kahne. The Reconstituted *Escherichia coli* Bam Complex Catalyzes Multiple Rounds of $\beta$-Barrel Assembly. *Biochemistry*, 50(35):7444–7446, 2011.

[48] Joshua A. Lerman, Daniel R. Hyduke, Haythem Latif, Vasiliy A. Portnoy, Nathan E. Lewis, Jeffrey D. Orth, Alexandra C. Schrimpe-Rutledge, Richard D. Smith, Joshua N. Adkins, Karsten Zengler, and Bernhard O. Palsson. In silico method for modelling metabolism and gene product expression at genome scale. *Nature Communications*, 3:929, 2012.

[49] Hiroshi Nikaido and Emiko Y Rosenberg. Effect on solute size on diffusion rates through the transmembrane pores of the outer membrane of Escherichia coli. *The Journal of general physiology*, 77(2):121–35, 1981.

[50] Waldemar Vollmer, Didier Blanot, and Miguel A. De Pedro. Peptidoglycan structure and architecture. *FEMS Microbiology Reviews*, 32(2):149–167, 2008.

[51] T. Uehara, K. Suefuji, N. Valbuena, B. Meehan, M. Donegan, and J. T. Park. Recycling of the Anhydro-N-Acetylmuramic Acid Derived from Cell Wall Murein Involves a Two-Step Conversion to N-Acetylglucosamine-Phosphate. *Journal of Bacteriology*, 187(11):3643–3649, 2005.

[52] Klaus Hantke and Volkmar Braun. Covalent Binding of Lipid to Protein. *European Journal of Biochemistry*, 34(2):284–296, 1973.

[53] Yehouda Harpaz, Mark Gerstein, and Cyrus Chothia. Volume changes on protein folding. *Structure*, 2(7):641–649, 1994.

[54] Peter F. Mühlradt and Jochen R. Golecki. Asymmetrical Distribution and Artifactual Reorientation of Lipopolysaccharide in the Outer Membrane Bilayer of Salmonella typhimurium. *European Journal of Biochemistry*, 51(2): 343–352, 1975.

[55] J. Pramanik and J. D. Keasling. Stoichiometric model ofEscherichia coli metabolism: Incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. *Biotechnology and Bioengineering*, 56 (4):398–421, 1997.

[56] Matthew Scott, Carl W. Gunderson, Eduard M. Mateescu, Zhongge Zhang, and Terence Hwa. Interdependence of Cell Growth and Gene Expression: Origins and Consequences. *Science*, 330(6007):1099–1102, 2010.

[57] Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie. Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. *Science*, 329(5991):533–538, 2010.

[58] J. W. Walley, Z. Shen, R. Sartor, K. J. Wu, J. Osborn, L. G. Smith, and S. P. Briggs. Reconstruction of protein networks from an atlas of maize seed proteotypes. *Proceedings of the National Academy of Sciences*, 110(49):E4808–E4817, 2013.

[59] Aarash Bordbar, Harish Nagarajan, Nathan E Lewis, Haythem Latif, Ali Ebrahim, Stephen Federowicz, Jan Schellenberger, and Bernhard O Palsson. Minimal metabolic pathway structure is consistent with associated biomolecular interactions. *Molecular systems biology*, 10:737, 2014.

[60] Troy E. Sandberg, Margit Pedersen, Ryan A. LaCroix, Ali Ebrahim, Mads Bonde, Markus J. Herrgard, Bernhard O. Palsson, Morten Sommer, and Adam M. Feist. Evolution of Escherichia coli to 42 C and Subsequent Genetic Engineering Reveals Adaptive Mechanisms and Novel Mutations. *Molecular Biology and Evolution*, 31(10):2647–2662, 2014.

[61] Xianyue Ma and Kenneth Cline. Multiple precursor proteins bind individual Tat receptor complexes and are collectively transported. *The EMBO Journal*, 29(9):1477–1488, 2010.

[62] M. J. Tarry, E. Schafer, S. Chen, G. Buchanan, N. P. Greene, S. M. Lea, T. Palmer, H. R. Saibil, and B. C. Berks. Structural analysis of substrate binding by the TatBC component of the twin-arginine protein transport system. *Proceedings of the National Academy of Sciences*, 106(32):13284–13289, 2009.

[63] Jeffrey D Orth, Tom M Conrad, Jessica Na, Joshua A Lerman, Hojung Nam, Adam M Feist, and Bernhard Ø Palsson. A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011. *Molecular systems biology*, 7:535, 2011.

[64] Aleksandar Cvetkovic, Angeli Lal Menon, Michael P. Thorgersen, Joseph W. Scott, Farris L. Poole II, Francis E. Jenney Jr, W. Andrew Lancaster, Jeremy L. Praissman, Saratchandra Shanmukh, Brian J. Vaccaro, Sunia A.

Trauger, Ewa Kalisiak, Junefredo V. Apon, Gary Siuzdak, Steven M. Yannone, John A. Tainer, and Michael W. W. Adams. Microbial metalloproteomes are largely uncharacterized. *Nature*, 466(7307):779–782, 2010.

[65] Mingzhu Liu, Tim Durfee, Julio E. Cabrera, Kai Zhao, Ding J. Jin, and Frederick R. Blattner. Global Transcriptional Programs Reveal a Carbon Source Foraging Strategy by *Escherichia coli*. *Journal of Biological Chemistry*, 280(16):15921–15927, 2005.

[66] Aarash Bordbar, Jonathan M. Monk, Zachary A. King, and Bernhard O. Palsson. Constraint-based models predict metabolic and associated cellular functions. *Nature Reviews Genetics*, 15(2):107–120, 2014.

[67] Q. K. Beg, A. Vazquez, J. Ernst, M. A. de Menezes, Z. Bar-Joseph, A.-L. Barabasi, and Z. N. Oltvai. Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity. *Proceedings of the National Academy of Sciences*, 104(31):12663–12668, 2007.

[68] Samuel Wagner, Louise Baars, A. Jimmy Ytterberg, Anja Klussmeier, Claudia S. Wagner, Olof Nord, Per-Åke Nygren, Klaas J. van Wijk, and Jan-Willem de Gier. Consequences of Membrane Protein Overexpression in *Escherichia coli*. *Molecular & Cellular Proteomics*, 6(9):1527–1550, 2007.

[69] G. N. Vemuri, E. Altman, D. P. Sangurdekar, A. B. Khodursky, and M. A. Eiteman. Overflow Metabolism in Escherichia coli during Steady-State Growth: Transcriptional Regulation and Effect of the Redox Ratio. *Applied and Environmental Microbiology*, 72(5):3653–3661, 2006.

[70] Amit Varma and Bernhard O Palsson. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Applied and environmental microbiology*, 60 (10):3724–31, 1994.

[71] Eliane Fischer, Nicola Zamboni, and Uwe Sauer. High-throughput metabolic flux analysis based on gas chromatographymass spectrometry derived 13C constraints. *Analytical Biochemistry*, 325(2):308–316, 2004.

[72] Kenneth Segers, Hugo Klaassen, Anastasios Economou, Patrick Chaltin, and Jozef Anné. Development of a high-throughput screening assay for the discovery of small-molecule SecA inhibitors. *Analytical Biochemistry*, 413(2):90–96, 2011.

[73] Hiroshi Nikaido and Emiko Y Rosenberg. Porin channels in Escherichia coli: studies with liposomes reconstituted from purified proteins. *Journal of bacteriology*, 153(1):241–52, 1983.

[74] Etsuko Sugawara and Hiroshi Nikaido. Pore-forming activity of OmpA protein of Escherichia coli. *The Journal of biological chemistry*, 267(4):2507–11, 1992.

[75] Ian C. West and Malcolm G.P. Page. When is the outer membrane of Escherichia coli rate-limiting for uptake of galactosides? *Journal of Theoretical Biology*, 110(1):11–19, 1984.

[76] John L Ingraham. Escherichia coli and Salmonella typhimurium: cellular and molecular biology. ASM, 1987.

# Chapter 5

# Effects of micronutrients on growth

## 5.1 Introduction

Trace metals are essential for all living organisms, for they are required for catalytic processes essential to energy conservation, metabolism, replication, and maintenance. Yet metals pose a unique challenge in constraint-based models of metabolism (*i.e.*, M-models) as they are neither produced nor consumed biochemically [1]; instead, metals in M-models are generally treated as a lumped sum in the biomass objective function rather than be integrated into the network [2]. In M-models, metal availability and growth rate are linearly correlated even though there is contrary experimental evidence [3]. In iHN637, the *Clostridium ljungdahlii* M-model, seven of ten metals ($Ca^{2+}$, $Cu^{2+}$, $Mg^{2+}$, $Mn^{2+}$, $Mo^{2+}$, $Ni^{2+}$, $Zn^{2+}$ + $Co^{2+}$, $Fe^{2+}$, $Na^{+}$) could only be imported or exported (in addition to their inclusion in the biomass objective function, which represents the total composi-

tion of the cell [4, 2], and only $Co^{2+}$ was predicted to participate in flux-carrying reactions that were not a transport reaction or biomass production. Thus, most metal ions were not associated to the reactions they help catalyze. This represents a general fact for M-models [2].

The next generation of constraint-based genome-scale models change this paradigm. Metabolic and gene expression models (ME-models) cover the proccesses of transcription, translation, and metabolism, which can also include protein modifications. Protein modifications can account for the presence of metals in biochemical reactions and thus enable predictions of the optimal distribution of resources in response to limited metal availability. Therefore, ME-models provide a robust, genome-wide approach to define how transition metals affect an organisms functional network, which addresses the articulated need to bridge chemistry and biology in a coherent and systematic way [1, 5]. The detailed representation of cofactors and prosthetic groups will enable us to manipulate the cofactor dependency of heterologous pathways to maximize energy conservation, subsequently optimizing chemical production by *C. ljungdahlii*.

For acetogens like *C. ljungahlii*, understanding the role of trace metals is particularly important, as metals are crucial for the Wood-Ljungdahl pathway (WLP), responsible for *C. ljungdahlii's* autotrophic growth, and more [5, 6]. Insights into such requirements provide an opportunity to rationally manipulate the WLP and other pathways for improved biotechnological outcomes [7]. Here, we focus on two metals: Nickel, which is a required ion in the WLP protein carbon monoxide dehydrogenase:Acetyl-CoA synthase, and zinc, an essential metal for organisms across all domains of life [8, 9]. The *C. ljungdahlii* ME-model, iJL965-

ME, predicted that nickel and zinc, essential metals for *C. ljungdahlii*, would affect the systematic network differently, as one metal is involved with a single pathway and the other with multiple processes. Results regarding nickel availability were examined *in vivo* as well.

## 5.2 Results

### 5.2.1 Nickel controls phenotype through Wood-Ljungdahl activity

iJL965-MEs only nickel-containing proteins, CODH4 and carbon monoxide dehydrogenase:Acetyl-CoA synthase (CODH_ACS), are part of the WLP, which afforded the possibility of controlling this pathway through changes in media composition both *in silico* and *in vivo*. Due to *C. ljungdahlii*s reliance on WLP for autotrophic growth, nickel was predicted to be essential for CO-growth. Although true essentiality could not be tested due to trace nickel in the media, the amount of additional nickel (added as multiples of 0.10 mM) significantly influenced *in vivo* growth rate in a quadratic fashion as predicted (Fig 5.1A). According to iJL965-ME, the non-linear effects of nickel limitations were caused by an uneven distribution of metal resources between CODH_ACS and CODH4, resulting in different rates of decreasing protein activity (Fig 5.1B). In turn, the other reactions in WLP were correlated to either CODH_ACS, like MTHFR5 and methyltetrahydrofolate corrinoid/iron-sulfur protein methyltransferase (METR), or CODH4 (Fig 5.2). Finally, iJL965-ME predicted that while nickel availability affected growth rate, protein activity, and acetate and ethanol yield, the acetate-to-ethanol production rate

would not change. The acetate:ethanol production rate ratio remained constant at 1.4 for different nickel concentrations (Fig 5.3A). HPLC measurements confirmed that acetate:ethanol production rate was unchanged with a ratio of 1.48±0.34 (Fig 5.3E), regardless of the nickel concentrations used (0x, 1x, and 5x [10x excluded due to carbon depletion], Fig 5.3D).

iJL965-ME predicted that nickel limitations would have different effects on fructose-grown cells. Removal of nickel was not predicted to affect growth rate or fructose uptake significantly (△gr=98%, △fructose=99%, Fig 5.4A). However, there was no CODH_ACS or METR activity under nickel depletion, which reduced the WLP activity (Table 5.1) and eliminated acetate secretion. Instead, the model predicted that only ethanol secretion would occur (Fig 5.4B, C). To test this prediction, *C. ljungdahlii* was grown either without added nickel (0x) or with high nickel concentrations (10x). Both cultures consumed the same amount of fructose (p=0.26) and produced identical amounts of ethanol (p=0.95), but exhibited different growth rates (p=0.062) and final concentrations of acetate (p=2.2e-4) (Fig 5.4D-G). Increased acetate secretion rate (p=0.016, Fig 5.5A) and final acetate concentrations in the 10x condition were due to the nickel-stimulated WLP consuming more CO2.

### 5.2.2 Zinc affects multiple cellular processes

Unlike nickel, which was incorporated into two proteins, zinc was a cofactor for twenty four proteins in iJL965-ME. Since these proteins were required for multiple processes, zinc was a predicted essential metal for *C. ljungdahlii* in both autotrophic and heterotrophic growth conditions (Fig 5.6A, B). Despite decreases

in growth rate, ethanol secretion rate was predicted to increase when zinc was limiting (Fig 5.6C, D).

Zinc-containing proteins were grouped by their biochemical activities; fourteen proteins catalyzed metabolic reactions, seven were involved in translation and protein formation, and three were involved in transcription. As zinc availability decreased, the activities of these proteins also decreased (Fig 5.7), but not at the same rates. While the decreasing rate of transcription stayed constant, zinc-containing proteins related to metabolism decreased at a faster rate than zinc-containing proteins in translation until the two processes were nearly equal in protein activity ($\text{zn\_uptake}_{CO} = 62\%$, $\text{zn\_uptake}_{frutose} = 70\%$), after which point they decreased at the same rate (Fig 5.7).

Zinc-containing proteins were involved in four flux-carrying metabolic reactions in CO conditions, three in fructose conditions (Fig 5.8). These reactions were dihydroorotase (DHORTS), the first step in the pyrimidine biosynthetic pathway, histidinol dehydrogenase (HISTD), the last metabolic step to produce histidine, acetyl-CoA carboxylase (ACCOAC), the first step towards fatty acid synthesis, and glycerol dehydrogenase (GLYCDx), which breaks down glycerol so that it can eventually enter glycolysis or gluconeogensis. The last reaction was predicted to be active in CO conditions, but not fructose. While all four reactions were affected by zinc-limitations, the most affected was ACCOAC (decreasing rate, *i.e.*, shadow price = -0.017, while the other reactions had rates <-0.001), at least until zinc uptake reached 62% in CO conditions and 70% in fructose conditions. This single reaction was why the metabolic activity of zinc-containing proteins initially decreased faster than translation and transcription (Fig 5.7).

By initially decreasing flux through ACCOAC, iJL965-ME first sacrificed fatty acid production for other functions. However, due to iJL965-ME's membrane constraint, the amount of lipids that could be removed was limited. The protein-to-lipid ratio of membrane surface area (SA) increased until the constraint (protein:lipid SA $\leq$ 1:1) was reached, which occurred at 62% of zinc uptake in CO-growth conditions and 70% in fructose-growth conditions (Fig 5.9, see Methods in chapters 4 & 2 for more details on the constraint). After decreasing the maximum amount of lipid production allowed, iJL965-ME predicted that the lack of zinc would cut equally into protein production and metabolic reactions and would decrease growth rate faster (shadow price, Fig 5.9).

## 5.3 Discussion

Through iJL965-ME, the intersection of trace metals and metabolism under both autotrophic and heterotrophic (*i.e.*, mixotrophic) conditions was examined in more depth.

The potential of controlling WLP activity through media composition was explored. Although the lack of CODH_ACS activity (achieved by removing nickel from the media) may not cease WLP activity entirely, it may stop acetate production (as *in vivo* nickel depletion results suggest), leading to ethanol production as the main fermentation end product (Fig 5.4). However, the discrepancy between *in silico* and *in vivo* growth rates of nickel-depleted cells grown on fructose implied that WLP was more important than predicted for maximizing growth in mixotrophic conditions (Fig 5.4). In contrast, nickel was essential for CO-growth,

but had no effect on the acetate:ethanol ratio (Fig 5.1).

Although some studies have investigated zinc-limitation in prokaryotic organisms, proper understanding of the effects of zinc-depletion are hindered by prokaryotes' effective mechanisms to increase uptake and leach zinc from unlikely sources like glass [10]. Using ME-models, the potential systems effects of zinc-depletion were also examined for future follow-ups. Unlike nickel, zinc was integral to metabolism, translation, and transcription, thereby making the metal essential for *C. ljungdahlii* in all nutrient conditions. Interestingly, iJL965-ME predicted that the first affected processes would involve membrane restructuring. Under zinc-limitations, the zinc uptake regulator (Zur) induces expression of membrane proteins, including high-affinity zinc ABC transporters, to encourage zinc transport in and out of the cell [10, 11, 12]. Furthermore, transcriptomic analysis of zinc-deprived *Escherichia coli* cells also highlighted that b1193, a membrane-bound transglycosylase E involved in murein hydrolysis, was significantly upregulated [10]. Together, the *in vivo* observations of increased membrane protein and cell wall hydrolysis expression and *in silico* observations of decreased lipid production indicates that zinc-depletion may affect prokaryotic membrane composition and size in a manner yet to be studied.

The combination of metabolism, multi-omics predictions, and cofactor integration in iJL965-ME is an important milestone for a holistic understanding of metals in metabolism. Although nickel and zinc were the only trace metals to be investigated here, iJL965-ME invites further studies elucidating specific effects of concurrent metal limitations and genetic perturbations. The ME-model represents an inclusive method that unites analysis and integration of multiple data types.

## 5.4 Methods

For reconstruction, usage, and analysis of iJL965-ME, see Methods from chapter 2.

### 5.4.1 Bacterial growth conditions

Clostridium ljungdahlii (ATCC 55383) was grown under anaerobic conditions containing PETC medium (ATCC medium 1754) at 37°C. Fructose cultures were grown in 125 mL serum bottles containing 100 mL of medium plus 28 mM fructose, CO in 125 mL serum bottles containing 25 mL of media and bottles were pressurized once with CO to 18 PSI. Medium contained 0.10 mM of $NiCl_2$*6 H2O (*i.e.*, 1x). For testing the effect of nickel, final concentrations of 0 mM (0x), 0.50 mM (5x) and 1.0 mM (10x) of nickel was added to the media from an anoxic stock solution before autoclaving. Growth was routinely determined by measurement of OD600. Concentrations of fructose, acetate, ethanol, and glycerol were determined by high-performance liquid chromatography (Waters) as previously described [13]. Detection was performed by UV absorption at 410 nm.

## 5.5 Figures and tables

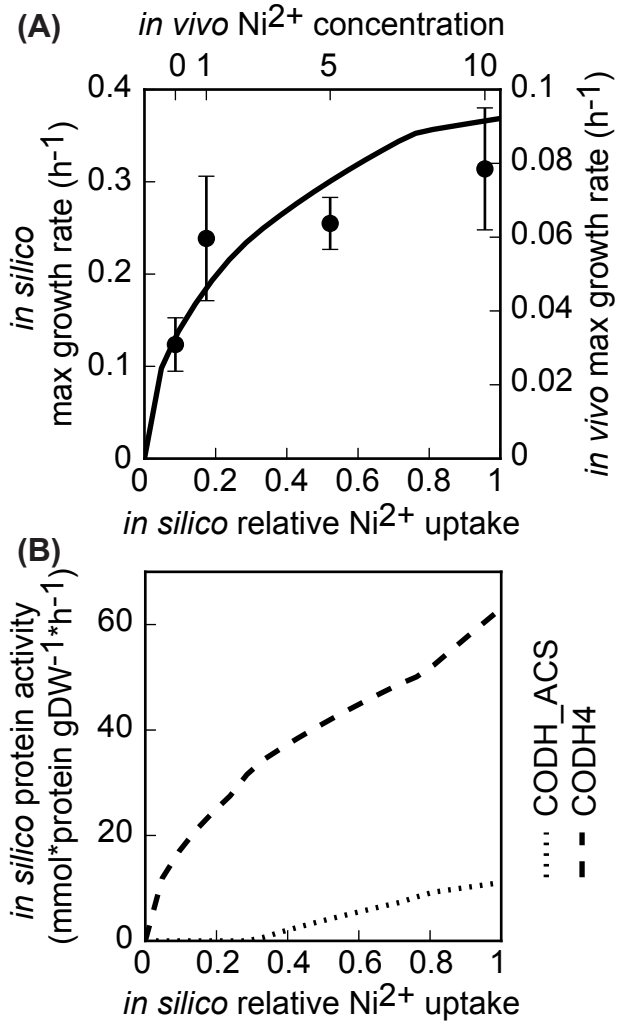



**Figure 5.1**: **Effects of nickel availability on CO-grown *C. ljungdahlii*.** (A) Maximum predicted growth rate was plotted against relative nickel uptake (line), and *in vivo* maximum growth rate verses the concentration of added nickel was plotted on the opposite axes (dot, ±std, n=3). (B) Predicted protein activity of the nickel-containing enzymes, carbon monoxide dehydrogenase (CODH4) and carbon monoxide dehydrogenase:acetyl-CoA synthase (CODH_ACS), was plotted against relative nickel uptake.

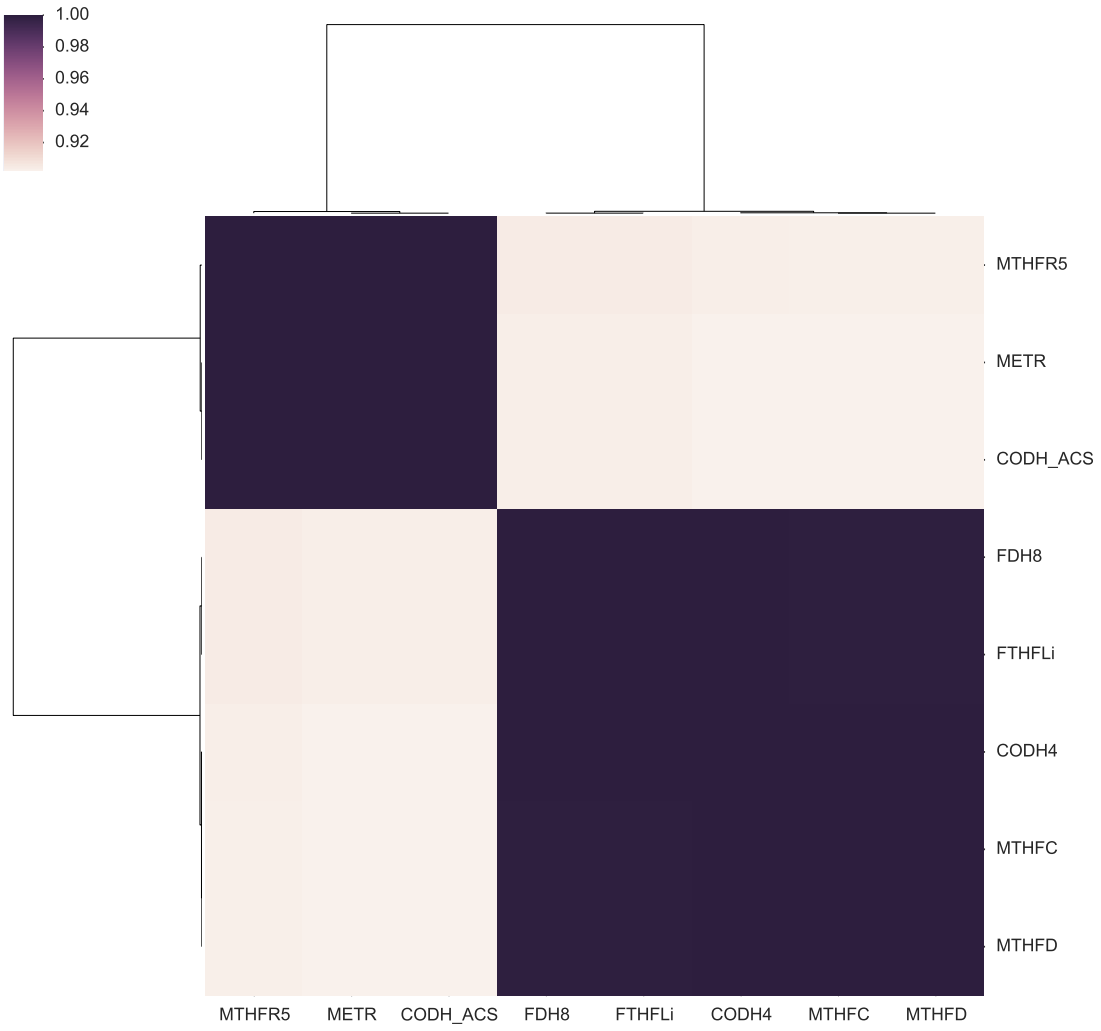



**Figure 5.2**: **Heatmap of Pearson correlations of WLP reaction fluxes over nickel availability of CO-grown cells and clustered by Euclidean distance.** Abbreviations: MTHFR5 = 5,10-methylenetetrahydrofolate reductase, METR = methyltetrahydrofolate corrinoid/iron-sulfur protein methyltransferase, CODH_ACS = carbon monoxide dehydrogenase:acetyl-CoA synthease, FDH8 = formate dehydrogenase, CODH4 = carbon monoxide dehydrogenase, MTHFC = methenyltetrahydrofolate cyclohydrolase, MTHFB = methylentetrahydrofolate dehydrogenase.

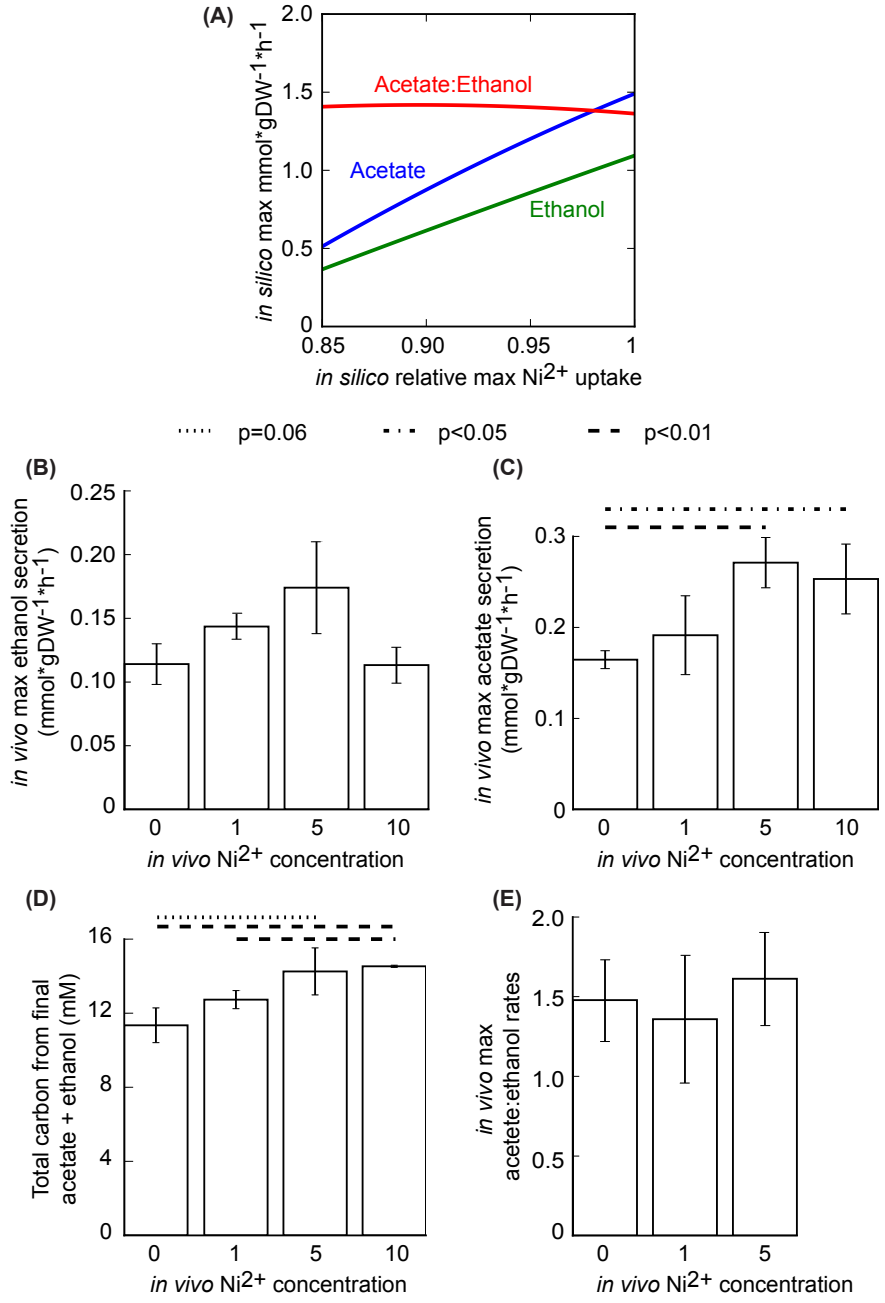



**Figure 5.3**: **Predicted and measured acetate and ethanol secretion rates of CO-grown *C. ljungdahlii* with varying nickel availability.** (A) Predicted maximum acetate and maximum ethanol secretion rates as well as the acetate-to-ethanol ratio were plotted against relative maximum constrained nickel uptake. Bar graphs of measured (B) maximum ethanol secretion rates, (C) maximum acetate secretion rates, (D) total carbon from final concentration of acetate and ethanol at t = 116 h, and (E) acetate-to-ethanol ratio were plotted for 4 different concentrations of added nickel (± std, n=3). Horizontal lines indicate significant differences.

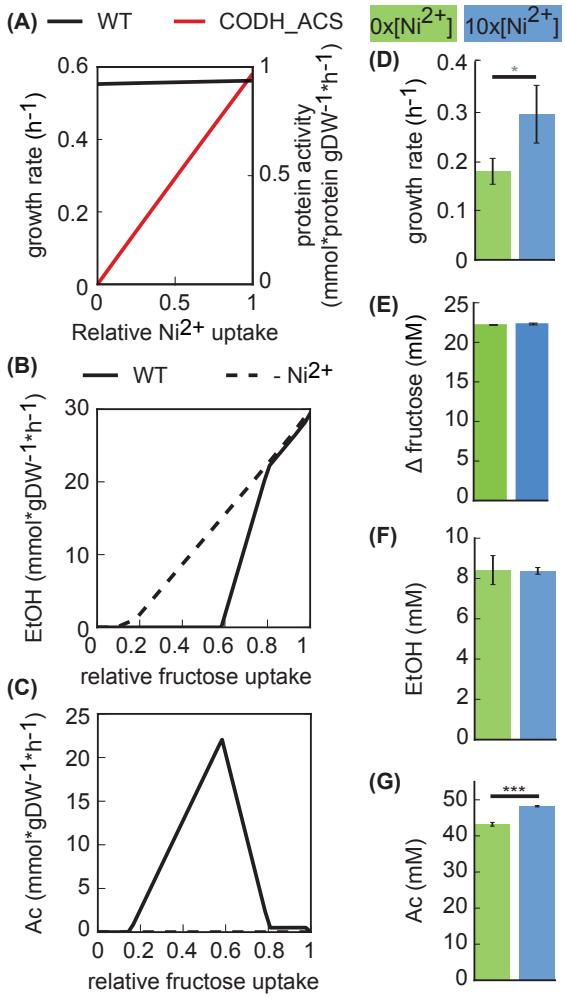



**Figure 5.4**: **Effects of nickel availability on fructose-grown *C. ljungdahlii*.** (A) Predicted growth rate and protein activity of CODH_ACS were plotted against relative nickel uptake (mmol*gDW$^{-1}$*h$^{-1}$). (B) Predicted ethanol (EtOH) secretion at optimal nickel uptake (WT) and no available nickel (-Ni$^{2+}$) were plotted against relative fructose uptake (mmol*gDW$^{-1}$*h$^{-1}$). (C) Predicted acetate (Ac) secretion at optimal nickel uptake and no available nickel were plotted against relative fructose uptake (mmol*gDW$^{-1}$*h$^{-1}$). Measured (D) growth rate, (E) fructose consumption, (F) final ethanol concentration, and (G) final acetate concentration of fructose-grown *C. ljungdahlii* without added nickel and with ten times the concentration of nickel were plotted (±std, n=3). Gray asterisk indicates difference significance is p=0.06, and three black asterisk indicates significance of p<0.001.

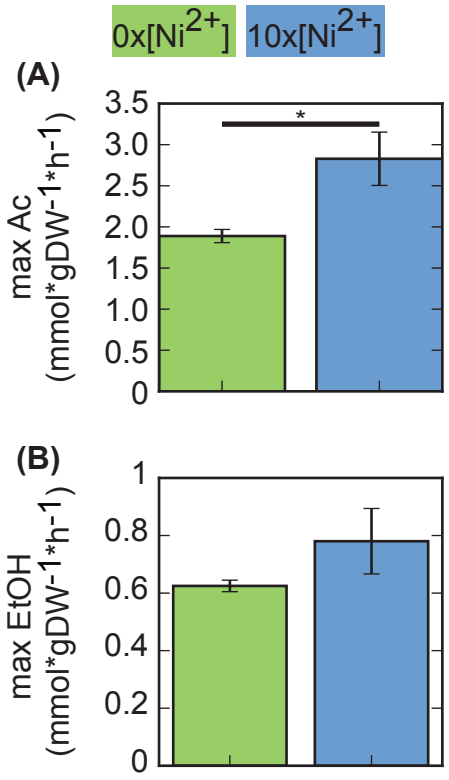



**Figure 5.5**: **Measured secretion rates of acetate and ethanol of fructose-grown *C. ljungdahlii* with and without nickel.** Bar graphs of measured (A) maximum acetate (Ac) secretion rates and (B) maximum ethanol (EtOH) secretion rates of fructose-grown *C. ljungdahlii* without added nickel (0x) and concentrated nickel (10x) (±std, n=3). Black asterisk indicates significance of $p<0.05$.

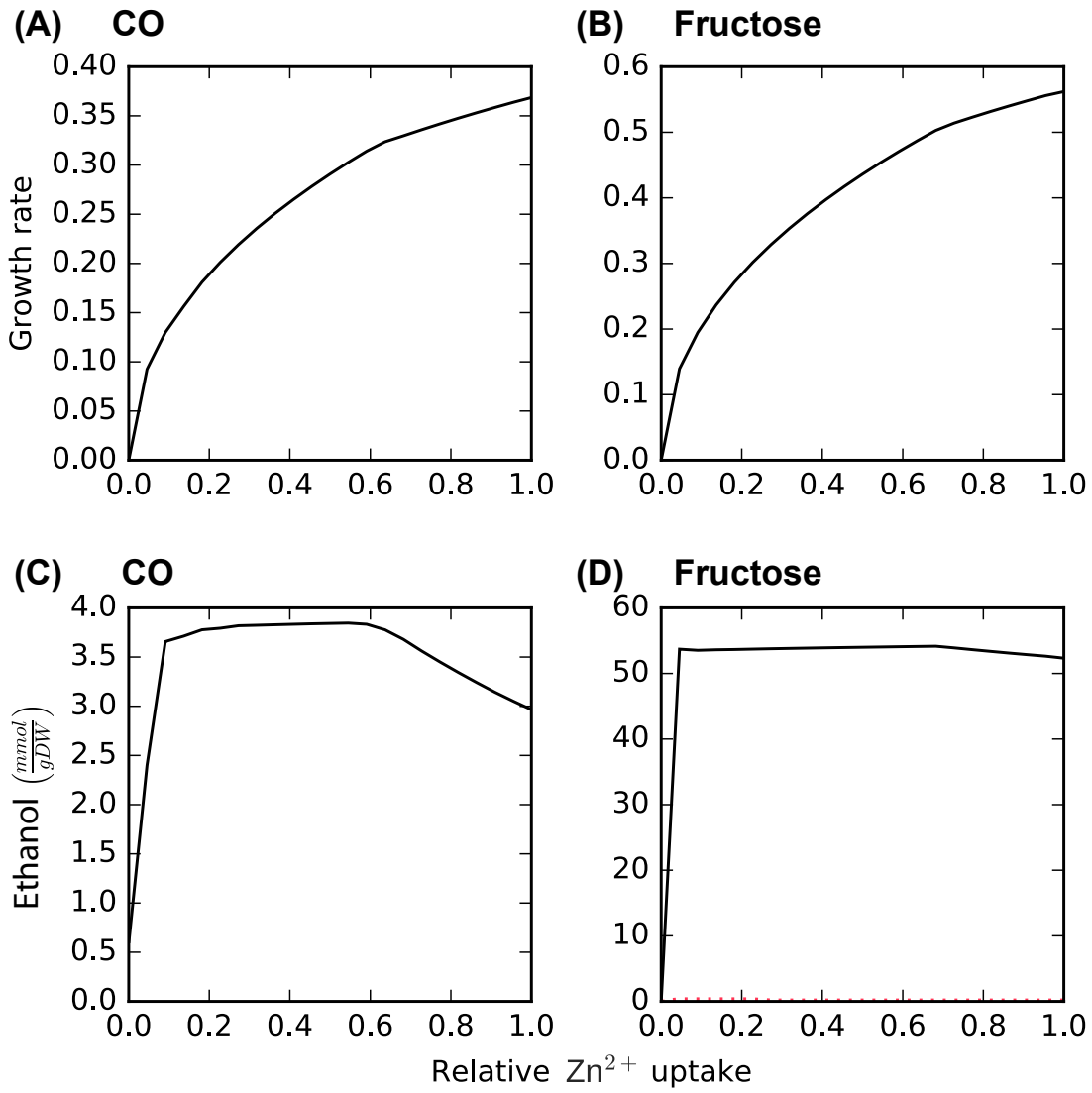



**Figure 5.6**: **Predicted growth rates and ethanol secretion rates under zinc limitations.** Relative zinc uptake was plotted against growth rate ($h^{-1}$) for both (A) CO growth conditions and (B) fructose growth conditions. Relative zinc uptake was also plotted against ethanol secretion for (C) CO growth conditions and (D) fructose growth conditions.

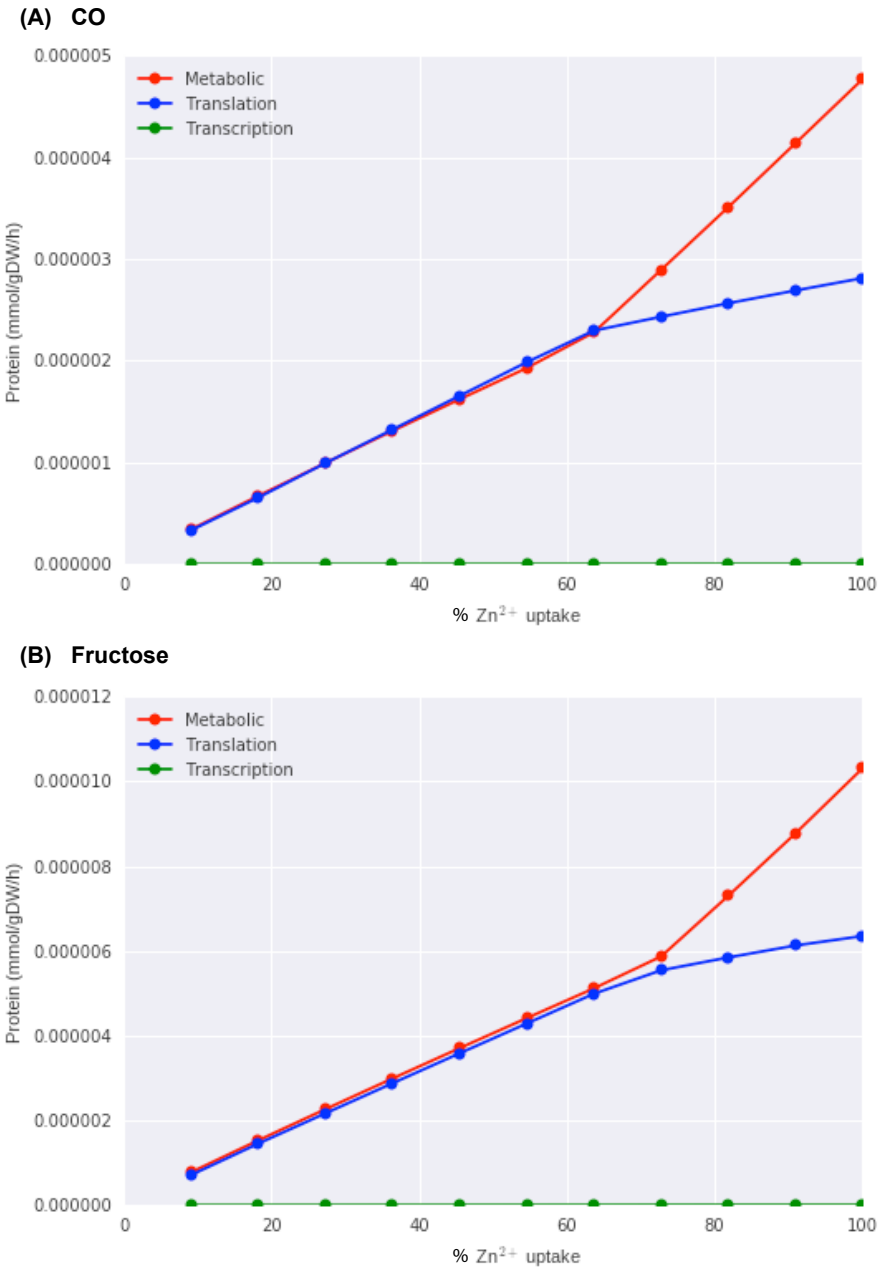



**Figure 5.7**: **Predicted protein activity levels grouped by transcription, translation, or metabolic under zinc limitations.** All twenty four proteins that required zinc were categorized by their associated reactions into transcription, translation, or metabolism. The percentage of maximum zinc uptake was plotted against protein activity for each process for (A) CO growth conditions and (B) fructose growth conditions. Abbreviations: HISTD = histidinol dehydrogenase, DHORTS = dihydroorotase, GLYCDx = glycerol dehydrogenase, ACCOAC = acetyl-CoA carboxylase.

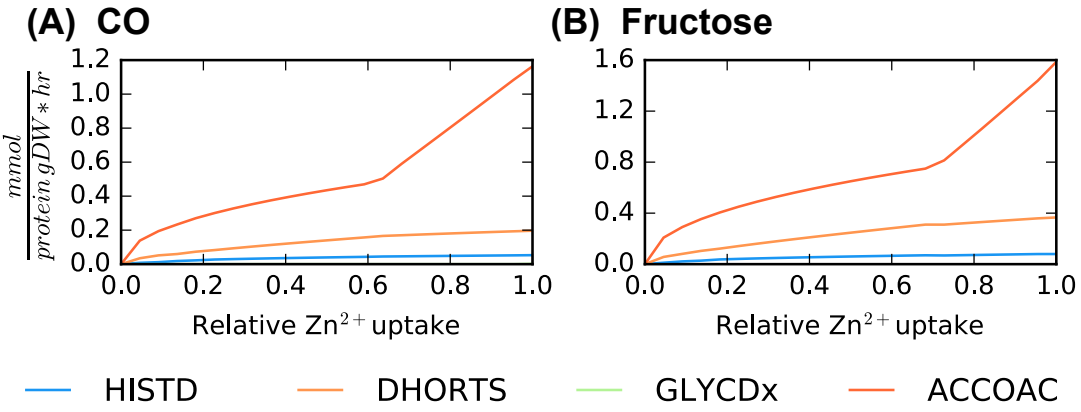



**Figure 5.8**: **Predicted protein activity levels of zinc-required metabolic enzymes under zinc limitations.** Relative zinc uptake was plotted against activities of the four metabolic reactions dependent on zinc for biochemical activity for (A) CO growth conditions and (B) fructose growth conditions.

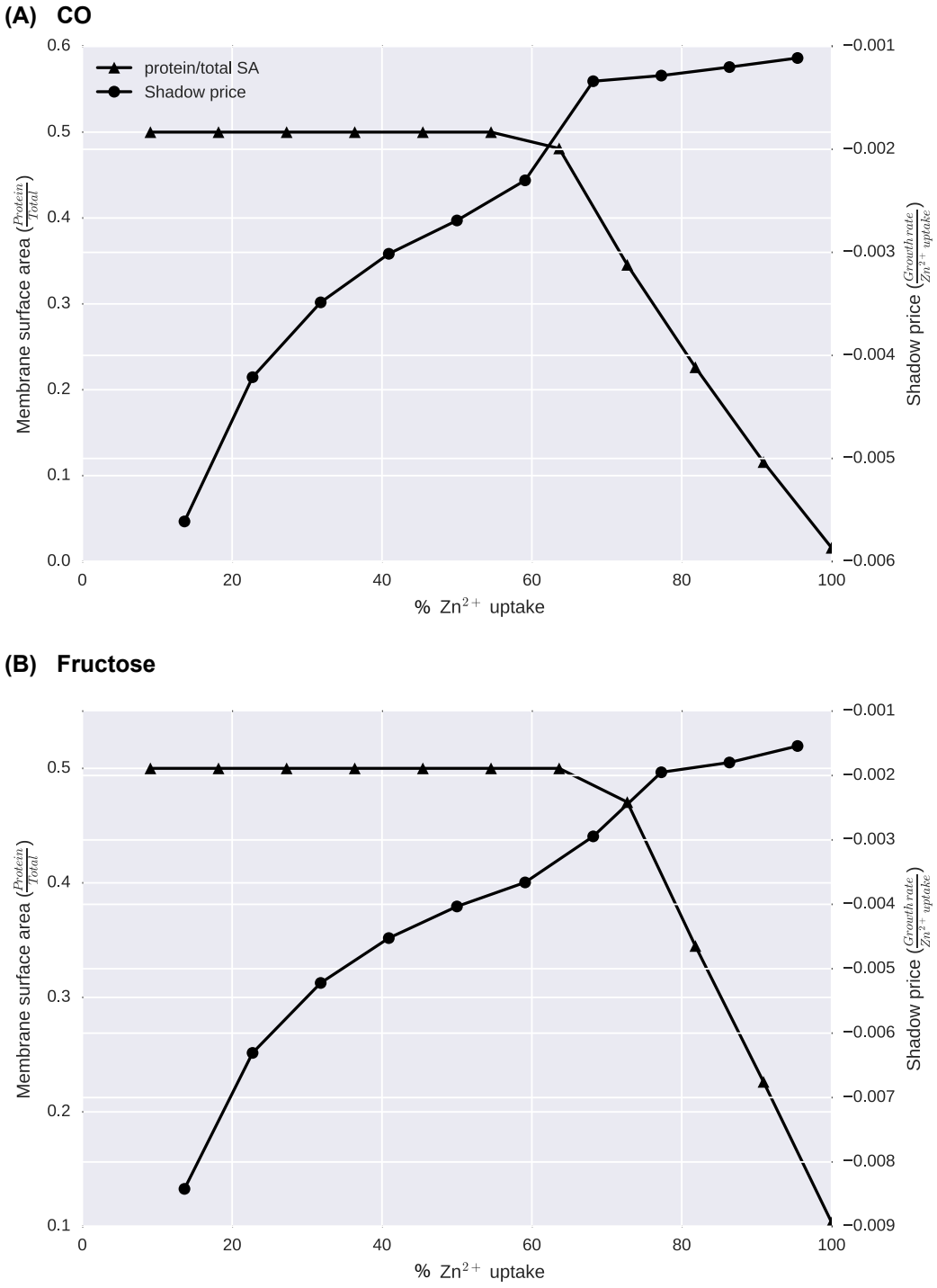



**Figure 5.9**: **The ratio of protein membrane surface area to total membrane surface area changes as zinc availability decreases.** The graph also shows the corresponding shadow price of how much Zn uptake effects the maximum growth rate.

**Table 5.1**: **Predicted changes in WLP reaction fluxes when nickel was removed from fructose-grown cells.**

| Reaction | Name | %△ expression (-Ni/WT) |
|---|---|---|
| FDH8 | formate dehydrogenase | -2.4 |
| FTHFLi | formate tetrahydrofolate ligase | -2.4 |
| MTHFC | methenyl tetrahydrofolate cyclohydrolase | -1.9 |
| MTHFD | methylenetetrahydrofolate dehydrogenase | -1.9 |
| MTHFR5 | methylenetetrahydrofolate reductase | -90 |
| METR | methyltetrahydrofolate corrinoid/iron-sulfur protein methyltransferase | -100 |
| CODH_ACS | carbon monoxide dehydrogenase:acetyl-CoA synthease | -100 |

## 5.6 Acknowledgments

Chapter 5, in part, is currently being prepared for submission for publication of the material. Joanne Liu, Ali Ebrahim, Mahmoud Al Bassam, Colton Lloyd, Ji-Nu Kim, Connor Olson, and Karsten Zengler. "A systems biology approach to investigate proteome control of acetate and ethanol production in *Clostridium ljungdahlii*" (working title). The dissertation author was the primary investigator and author of this material. We are also thankful to Cameron Martino, Kristine Ly, and Kevin Tang for assisting with growth experiments, and to Nathan Lewis, Cristal Zuñiga and Lívia Zaramela for fruitful discussions and input. Without them, this study would not have been possible.

## 5.7 References

[1] Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature biotechnology*, 28(3):245–8, 2010.

[2] Adam M Feist and Bernhard O Palsson. The biomass objective function. *Current Opinion in Microbiology*, 13(3):344–349, 2010.

[3] Jyotisna Saxena and Ralph S. Tanner. Effect of trace metals on ethanol production from synthesis gas by the ethanologenic acetogen, *Clostridium ragsdalei*. *Journal of Industrial Microbiology & Biotechnology*, 38(4):513–521, 2011.

[4] Harish Nagarajan, Merve Sahin, Juan Nogales, Haythem Latif, Derek R Lovley, Ali Ebrahim, and Karsten Zengler. Characterizing acetogenic metabolism using a genome-scale metabolic reconstruction of *Clostridium ljungdahlii*. *Microbial cell factories*, 12(1):118, 2013.

[5] Tudor I Oprea, Alexander Tropsha, Jean-Loup Faulon, and Mark D Rintoul. Systems chemical biology. *Nature chemical biology*, 3(8):447–50, 2007.

[6] Stephen W Ragsdale and Elizabeth Pierce. Acetogenesis and the Wood-Ljungdahl pathway of CO(2) fixation. *Biochimica et biophysica acta*, 1784 (12):1873–98, 2008.

[7] William F. Martin. Hydrogen, metals, bifurcating electrons, and proton gradients: The early evolution of biological energy conservation. *FEBS Letters*, 586(5):485–493, 2012.

[8] Dayle K. Blencowe and Andrew P. Morby. Zn(II) metabolism in prokaryotes. *FEMS Microbiology Reviews*, 27(2-3):291–311, 2003.

[9] W. Maret. Zinc Biochemistry: From a Single Zinc Enzyme to a Key Element of Life. *Advances in Nutrition: An International Review Journal*, 4(1):82–91, 2013.

[10] Alison I Graham, Stuart Hunt, Sarah L Stokes, Neil Bramall, Josephine Bunch, Alan G Cox, Cameron W McLeod, and Robert K Poole. Severe zinc depletion of Escherichia coli: roles for high affinity zinc binding by ZinT, zinc transport and zinc-independent proteins. *The Journal of biological chemistry*, 284(27):18377–89, 2009.

[11] S. I. Patzer and K. Hantke. The Zinc-responsive Regulator Zur and Its Control of the znu Gene Cluster Encoding the ZnuABC Zinc Uptake System in *Escherichia coli*. *Journal of Biological Chemistry*, 275(32):24321–24332, 2000.

[12] Tara K Sigdel, J Allen Easton, and Michael W Crowder. Transcriptional response of Escherichia coli to TPEN. *Journal of bacteriology*, 188(18):6709–13, 2006.

[13] Vasiliy A Portnoy, Markus J Herrgård, and Bernhard Ø Palsson. Aerobic fermentation of D-glucose by an evolved cytochrome oxidase-deficient *Escherichia coli* strain. *Applied and environmental microbiology*, 74(24):7561–9, 2008.

# Chapter 6

# Conclusions

## 6.1 Summary

Metabolic and gene expression models (ME-models) represent a milestone in our ability to mechanistically describe the link between genotype and phenotype. By identifying the genome-scale metabolic capabilities of an organism and systematically associating macromolecules required to enable such activity, we have provided a unified approach to model simultaneously molecular and cellular phenotypes based on the systematic needs of the organism. Due to the wide scope of predictions enabled, ME-models also provide a framework to concurrently analyze multiple data types, from transcriptomics to growth rates, as demonstrated through the reconstruction and validation of a *Clostridium ljungdahlii* ME-model named iJL965-ME. With iJL965-ME, we established that a ME-model predicted growth rate and acetate secretion rates more accurately than an M-model of the same organism. Additionally, a ME-model was able to simulate intrinsically the production of both known (ethanol) and novel (glycerol) products, which an M-

model could not do without *ad hoc* constraints defined by the model user. Finally, due to the incorporation of gene expression, ME-models were also able to predict the required RNA and protein abundances for optimal growth rate, which were comparable to RNA-seq data for *C. ljungdahlii*.

Because the capabilities of a ME-model are extensive, a variety of features can be interrogated using the ME-model framework. Here, three levels of cell organization were examined.

First, simulated shuffled tRNA locations and co-expressions suggested that genome architecture may be under selection to satisfy the growth conditions of the organism. While tRNA operon structures have evolved to minimize expression costs and maximize usage, optimized tRNAs by amino acid differ between *E. coli* and *C. ljungdahlii*. Furthermore, by examining tRNA-containing operons in these organisms, two strategies were identified and defined: fragmentation (the minimization of co-expressed tRNAs) and modularization (the co-expression of the majority of tRNAs). Due to fragmentation, *E. coli* was able to produce near-optimal values for rRNA expression, tRNA efficiency, and growth rate regardless of tRNA location, while the specific tRNA arrangement in *C. ljungdahlii* was optimized for tRNA efficiency, which implies that at the gene level, *E. coli* has maximized its output and *C. ljungdahlii* has maximized its resources.

Second, reconstruction of protein translocation and compartmentalization in the *E. coli* ME-model emphasized the impact of membrane formation. Not only did growth rate drop significantly due to the inclusion of translocation, but constraining the protein membrane space to measured levels recapitulated acetate production during growth on glucose.

Third, controlling nickel and zinc availability highlighted how metalloproteins drive performance in *C. ljungdahlii*. Corroboration between *in silico* and *in vivo* nickel limitations revealed that unmetabolized micronutrients, like metal ions, can strongly influence growth rate and product secretion rates dependent on other nutrients, like carbon source.

Because of the E-matrix, genome-scale constraint-based models can account for transcription units, proteins, and cofactors. Without ME-models, the three analyses described above would not have been possible. Thus, ME-models have significantly broadened the scope of microbial systems biology to not only examine evolutionary implications at the molecular level but also identify potential media combinations for cellular phenotypes.

## 6.2 Future possibilities

Although ME-models have been reported in several publications, they are still a newly developed technology and have not been robustly tested by the scientific community. As of writing, ME-models have been validated using experimental measurements (including gene knockouts, RNA-seq, ribosome profiling, proteomics of the cytoplasm, tRNA profiles, consumption and production rates, and growth rates) with accurate results [1, 2, 3]. However, no study has simultaneously compared multiple large-scale datasets (e.g., transcriptomics, ribosome profiling, proteomics, fluxomics, and metabolomics at once) from a single organism to the direct *in silico* equivalents. Such a study would be be insightful for understanding the whole cell organism and highlighting areas for further research.

A recurring issue with ME-models that continually arises is the $k_{eff}$ parameter, which represents the average turnover rate of a metabolic enzyme [4]. The $k_{eff}$ is set to 65 $s^{-1}$ in *E. coli* and 25 $s^{-1}$ in *C. ljungdahlii*. However, not all proteins operate at the average rate. For example, in *in vitro* conditions, enzymes in the Wood-Ljungdahl pathway could be four-fold more reactive than 25 $s^{-1}$ [5]. Such vast differences in potential enzyme activities hinder our ability to fully understand the systems response to both environmental and genetic cues, since we may under- or over-predict protein requirements and metabolic fluxes for certain pathways. Such a case may have happened in 4, when iJL1678-ME under-estimated the expression of inorganic ion, cofactor, and prosthetic related proteins. Defined universal speeds, perhaps normalized to average turnover rate, for key individual proteins or pathways may mitigate turnover rate problems from being carried over into future ME-models.

In addition to inaccurate turnover rates, the incorrect predictions of inorganic ion, cofactor, and prosthetic related proteins may be due to the lack of fully integrated gene regulatory pathways. Although incorporated feedback loops would compound the number of variables that need to be solved, a simple binary check of media composition availability before solving may be the next step in modeling regulation in ME-models. Not only could the expression of certain transcription units be turned on or off in response to the environment, but the turnover rates of transporter enzymes could be singularly adjusted to match the needs of the organism (e.g., if zinc availability was below a defined threshold, the zinc transporter's maximum uptake rate would increase in association with the expression of efficient zinc-binding proteins as defined by a pre-check algorithm) [6].

Many regulatory elements involve proteins in the membrane that sense molecules in either the external or internal environment. However, the current iteration of protein compartmentalization does not make a distinction between the inner and outer leaflets of the lipid bilayer, nor does the specific location of a protein make a difference for functional activity, even though the distance of a ribosome from an available mRNA changes the total time required for protein production. Thus, compartmentalization could be improved to account for subcellular organization like clustering of proteins or membrane nanodomains [7, 8], in addition to molecular overcrowding, which will provide insight into space usage for efficient function. Understanding these constraints will have both evolutionary implications as well as impacts on strain designs that require protein over-expression.

ME-models were used to explore genome architecture constraints on tRNA expression, but further understanding of the evolution of tRNA-containing transcription units could be obtained using constraint-based approaches that do not require ME-model reconstructions. The approach requires two sets of data, tRNA transcription units and the cell's amino acid composition (potentially identified through RNA-seq), and returns maximum possible tRNA efficiency (Fig 6.1). This model formulation could be applied to a variety of organisms, which will enhance our ability to trace the lineage of tRNA co-expression evolution and optimization. Thus, with this potential to compare a large number of tRNA operon structures from across the tree of life, we have the opportunity to further define or identify new strategies for tRNA organization.

- Input:
  - codon frequency (from RNA-seq, etc)
  - tRNA-to-codon dictionary (can be partially assumed)
  - tRNA-containing operon structure

**min Σ(unused_tRNA)** : objective
$0 \rightarrow tRNA\_operon_x$ : tRNA operons as EX reactions
$tRNA\_operon_x \rightarrow tRNA_a + tRNA_b + \ldots$ : only tRNAs are considered
$tRNA_a \rightarrow codon_1$ : tRNAs are linked to specific codons
$tRNA_a \rightarrow codon_2$
...
$tRNA_a \rightarrow unused\_tRNA$ : remove expressed tRNA but unused
...
$X\ codon_1 + Y\ codon_2 + \ldots \rightarrow 0$ : codon frequency
$unused\_tRNA \rightarrow 0$

- Output:
  - Most efficient tRNA expression for the given condition (disregarding regulation, costs, biomass, co-expression, etc)

**Figure 6.1**: A constraint-based approach to modeling tRNA operon structure.

### 6.2.1 Broader implications

ME-models serve as a theoretical baseline to understand the phenotypic response to environmental and genetic perturbations. Because we account for both RNA and protein abundances, both minute (*e.g.*, fluxes in gene expression) and coarse-grain (*e.g.*, growth rate) responses can be predicted. As a result, the potential applications of the ME-model extend to multiple fields. Since ME-models can simulate gene knock-outs and knock-ins, they can be applied readily to metabolic engineering tasks. Furthermore, ME-models can be used to eliminate undesirable media combinations specific to the strain being used. ME-models will also be useful for deeper understanding of interactions between organism and ecological community, because ME-models can simulate proteomic responses to changes in niche

composition. The niche in question is not limited to environmental studies, but may also come from the medical field, since mechanistic modeling of pathogenicity through incorporation of secretion systems can be achieved. The possibilities for ME-model usage extend far beyond what is listed here, but what is certain is that ME-models open exciting new avenues to interpret and predict biological functions for strong societal benefits.

## 6.3 References

[1] Joshua A. Lerman, Daniel R. Hyduke, Haythem Latif, Vasiliy A. Portnoy, Nathan E. Lewis, Jeffrey D. Orth, Alexandra C. Schrimpe-Rutledge, Richard D. Smith, Joshua N. Adkins, Karsten Zengler, and Bernhard O. Palsson. In silico method for modelling metabolism and gene product expression at genome scale. *Nature Communications*, 3:929, 2012.

[2] Edward J O'Brien, Joshua A Lerman, Roger L Chang, Daniel R Hyduke, and Bernhard Ø Palsson. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Molecular systems biology*, 9: 693, 2013.

[3] Laurence Yang, Ding Ma, Ali Ebrahim, Colton J. Lloyd, Michael A. Saunders, and Bernhard O. Palsson. solveME: fast and reliable solution of nonlinear ME models. *BMC Bioinformatics*, 17(391), 2016.

[4] Colton J Lloyd, Ali Ebrahim, Laurence Yang, Zachary Andrew King, Edward Catoiu, Edward J O'Brien, Joanne K Liu, and Bernhard O Palsson. COBRAme: A Computational Framework for Building and Manipulating Models of Metabolism and Gene Expression. *bioRxiv*, 2017.

[5] Johanna Mock, Yanning Zheng, Alexander P. Mueller, San Ly, Loan Tran, Simon Segovia, Shilpa Nagaraju, Michael Köpke, Peter Dürre, and Rudolf K. Thauer. Energy Conservation Associated with Ethanol Formation from H2 and CO2 in *Clostridium autoethanogenum* Involving Electron Bifurcation. *Journal of Bacteriology*, 197(18):2965–2980, 2015.

[6] Wolfgang Maret. Zinc Biochemistry: From a Single Zinc Enzyme to a Key Element of Life. *Advances in Nutrition: An International Review Journal*, 4 (1):82–91, 2013.

[7] Jan Spitzer and Bert Poolman. How crowded is the prokaryotic cytoplasm? *FEBS Letters*, 587(14):2094–2098, 2013.

[8] Yuanqing Ma, Elizabeth Hinde, and Katharina Gaus. Nanodomains in biological membranes. *Essays In Biochemistry*, 57:93–107, 2015.