UNIVERSITY OF CALIFORNIA SAN DIEGO

Multi-scale modeling to elucidate biological network structure and properties

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

in

Bioengineering

by

Bin Du

Committee in charge:

Professor Bernhard O. Palsson, Chair
Professor Pedro J. Cabrales Arevalo
Professor Jeff Hasty
Professor Christian M. Metallo
Professor Robert K. Naviaux

2019

The dissertation of Bin Du is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California San Diego

2019

DEDICATION

To my parents and friends who made all this possible

EPIGRAPH

True wisdom is knowing what you do not know.

—Confucius

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I have to thank so many people that have helped me tremendously over the course of my PhD study. Without them, it would have not been possible to get to where I am today.

I would like to first thank Professor Palsson for being a great mentor and role model. His passion for science and continuous desire to explore the unknown has been a great motivation for me. He gives me a lot of freedom to explore different opportunities and is always supportive, whether they are scientific projects, scholarship applications, internship opportunities and job applications. I am really grateful for all of the help and support from him.

Next I would like to thank Daniel Zielinski for all the help and mentorship. Dan worked closely with me for almost five years and we had a very productive time together. I still remember the time when I first joined the lab and was looking around for projects. Dan offered me a lot of great ideas and patiently guided me through all stages of scientific research. He would left tons of comments on my manuscript, pointing out what the problems are and letting me figure out how to address them. It was a painful learning process but I really appreciate it looking back. We coauthored five publications together and all these are not possible without Dan's help.

I would like to thank Laurence Yang for all the help and guidance. Laurence is very knowledgeable and patient at answering all my questions on ME models. He makes everything looks easy; and big congratulations at starting a faculty position and stepping into a new chapter of life! I would also like to thank Ke Chen for all the help on ME models, and being a great friend. I appreciate the help from my peers in the same year: James Yurkovich, Justin Tan, Colton Llyod and Jared Broddrick. And I want to acknowledge the tremendous help from all the current and previous members in SBRG, and to name a few: Andreas Dräger, Ali Ebrahim, Edward O'Brien, Erol Kavvas, Zhen Zhang, Kayla Ruggiero, Garri Arzumanyan, Sharon Grubner, Jonathan Monk,

Connor Olson, Anand Sastry, Xin Fang, Patrick Phaneuf, Muyao Wu, Richard Szubin, Julia Xu, Ye Gao, Ying Hefner, Adam Feist, Nathan Mih, David Heckman.

Also many thanks to Aarash Bordbar and Iman Famili for providing me the internship opportunity at Sinopia Biosciences. It was a great industry experience where I can apply the skills and trainings from my PhD study. We have worked on several fun projects and I am lucky to see how a biotech startup operates from day to day.

I would like to thank my committee: Professor Pedro Cabrales Arevalo, Professor Jeff Hasty, Professor Christian Metallo, Professor Robert Naviaux, for their helpful suggestions and support. All the work here is not possible without the funding from Novo Nordisk Foundation (NNF10CC1016517) and National Institute of General Medical Sciences of the National Institutes of Health (R01GM057089). Additionally, I want to thank the Graduate Student Association for the travel grant that allowed me to visit Denmark for a conference.

Last but not least, I want to say a big thank you to both my parents for all the love and support. It is a long journey away from home and they are always there for me. I would like to thank all my friends that are in my life over the years; and especially my best friend Xin Fang, for all the love and support, and for being with me through the good and bad times in life <3.

Chapter 1 in full is a reprint of material published in: **Bin Du**, Daniel C. Zielinski, Erol S. Kavvas, Andreas Dräger, Justin Tan, Zhen Zhang, Kayla E. Ruggiero, Garri A. Arzumanyan, Bernhard O. Palsson. 2016. "Evaluation of rate law approximations in bottom-up kinetic models of metabolism." *BMC Systems Biology*, 10(1), 40. The dissertation author was the primary author.

Chapter 2 in full is a reprint of material published in: **Bin Du***, Daniel C. Zielinski*, Bernhard O. Palsson. 2017. "Topological and kinetic determinants of the modal matrices of

dynamic models of metabolism.“ *PLoS One*, 12(12), e0189880. The dissertation author was the primary author (equally contributing with Daniel Zielinski).

Chapter 3 in full is a reprint of material published in: **Bin Du**, Daniel C. Zielinski, Bernhard O. Palsson. 2018. “Estimating metabolic equilibrium constants: progress and future challenges.“ *Trends in Biochemical Sciences*, 43(12), 960-969. The dissertation author was the primary author.

Chapter 4 in full is a reprint of material published in: **Bin Du**, Zhen Zhang, Sharon Grubner, James T. Yurkovich, Bernhard O. Palsson, Daniel C. Zielinski. 2018. “Temperature-dependent estimation of Gibbs energies using an updated group-contribution method.“ *Biophysical Journal*, 114(11), 2691-2702. The dissertation author was the primary author.

Chapter 5 in full is a reprint of material published in: **Bin Du**, Daniel C. Zielinski, Jonathan M. Monk, Bernhard O. Palsson. 2018. “Thermodynamic favorability and pathway yield as evolutionary tradeoffs in biosynthetic pathway choice.“ *Proceedings of the National Academy of Sciences*, 115(44), 11339-11344. The dissertation author was the primary author.

Chapter 6 in full is a reprint of the material: **Bin Du***, Connor A. Olson*, Anand V. Sastry, Xin Fang, Patrick V. Phaneuf, Ke Chen, Muyao Wu, Richard Szubin, Julia Xu, Ye Gao, Ying Hefner, Adam M. Feist, Bernhard O. Palsson. “Adaptive laboratory evolution of Escherichia coli under acid stress.“ *Submitted*. The dissertation author was the primary author (equally contributing with Connor Olson).

Chapter 7 in full is a reprint of the material: **Bin Du**, Laurence Yang, Colton J. Lloyd, Xin Fang, Bernhard O. Palsson. “Genome-scale model of metabolism and gene expression provides a multi-scale description of acid stress responses in *Escherichia coli.*“ *Submitted*. The dissertation author was the primary author.

VITA

2013	Bachelor of Science in Bioengineering, Biotechnology, University of California San Diego
2019	Doctor of Philosophy in Bioengineering, University of California San Diego

PUBLICATIONS

Yongsung Hwang, Samuel Suk, Susan Lin, Matthew Tierney, **Bin Du**, Timothy Seo, Aaron Mitchell, Alessandra Sacco, Shyni Varghese. 2013. "Directed *in vitro* myogenesis of human embryonic stem cells and their *in vivo* engraftment." *PLoS One*, 8(8), e72023.

Yongsung Hwang, Samuel Suk, Yu-Ru Vernon Shih, Timothy Seo, **Bin Du**, Yun Xie, Ziyang Li and Shyni Varghese. 2014. "WNT3A promotes myogenesis of human embryonic stem cells and enhances *in vivo* engraftment." *Scientific Reports*, 4, 5916.

Laurence Yang, Justin Tan, Edward J. OBrien, Jonathan M. Monk, Donghyuk Kim, Howard J. Li, Pep Charusanti, Ali Ebrahim, Colton J. Lloyd, James T. Yurkovich, **Bin Du**, Andreas Dräger, Alex Thomas, Yuekai Sun, Michael A. Saunders, Bernhard O. Palsson. 2015. "Systems biology definition of the core proteome of metabolism and expression is consistent with high-throughput data." *Proceedings of the National Academy of Sciences*, 112(34), 10810-10815.

Bin Du, Daniel C. Zielinski, Erol S. Kavvas, Andreas Dräger, Justin Tan, Zhen Zhang, Kayla E. Ruggiero, Garri A. Arzumanyan, Bernhard O. Palsson. 2016. "Evaluation of rate law approximations in bottom-up kinetic models of metabolism." *BMC Systems Biology*, 10(1), 40.

Bin Du*, Daniel C. Zielinski*, Bernhard O. Palsson. 2017. "Topological and kinetic determinants of the modal matrices of dynamic models of metabolism." *PLoS One*, 12(12), e0189880.

Ningzi Guan, **Bin Du**, Jianghua Li, Hyundong Shin, Rachel R. Chen, Guocheng Du, Jian Chen, Long Liu. 2018. "Comparative genomics and transcriptomics analysisguided metabolic engineering of *Propionibacterium acidipropionici* for improved propionic acid production." *Biotechnology and Bioengineering*, 115(2), 483-494.

Bin Du, Zhen Zhang, Sharon Grubner, James T. Yurkovich, Bernhard O. Palsson, Daniel C. Zielinski. 2018. "Temperature-dependent estimation of Gibbs energies using an updated group-contribution method." *Biophysical Journal*, 114(11), 2691-2702.

Bin Du, Daniel C. Zielinski, Bernhard O. Palsson. 2018. "Estimating metabolic equilibrium constants: progress and future challenges." *Trends in Biochemical Sciences*, 43(12), 960-969.

Bin Du, Daniel C. Zielinski, Jonathan M. Monk, Bernhard O. Palsson. 2018. "Thermodynamic favorability and pathway yield as evolutionary tradeoffs in biosynthetic pathway choice." *Proceedings of the National Academy of Sciences*, 115(44), 11339-11344.

Xin Fang, Jonathan M. Monk, Nathan Mih, **Bin Du**, Anand V. Sastry, Erol Kavvas, Yara Seif, Larry Smarr, Bernhard O. Palsson. 2018. "*Escherichia coli* B2 strains prevalent in inflammatory bowel disease patients have distinct metabolic capabilities that enable colonization of intestinal mucosa." *BMC Systems Biology*, 12(1), 66.

Hao Luo, Anne Sofie L. Hansen, Lei Yang, Konstantin Schneider, Mette Kristensen, Ulla Christensen, Hanne B. Christensen, **Bin Du**, Emre zdemir, Adam M. Feist, Jay D. Keasling, Michael K. Jensen, Markus J. Herrgrd, Bernhard O. Palsson. 2019. "Coupling *S*-adenosylmethionine-dependent methylation to growth: Design and uses." *PLoS Biology*, 17(3), e2007050.

Bin Du*, Connor A. Olson*, Anand V. Sastry, Xin Fang, Patrick V. Phaneuf, Ke Chen, Muyao Wu, Richard Szubin, Julia Xu, Ye Gao, Ying Hefner, Adam M. Feist, Bernhard O. Palsson. "Adaptive laboratory evolution of *Escherichia coli* under acid stress." *Submitted.*

Bin Du, Laurence Yang, Colton J. Lloyd, Xin Fang, Bernhard O. Palsson. "Genome-scale model of metabolism and gene expression provides a multi-scale description of acid stress responses in *Escherichia coli.*" *Submitted.*

* equal contribution

ABSTRACT OF THE DISSERTATION

Multi-scale modeling to elucidate biological network structure and properties

by

Bin Du

Doctor of Philosophy in Bioengineering

University of California San Diego, 2019

Professor Bernhard O. Palsson, Chair

Characterization of complex cellular behaviors on a molecular scale requires detailed understanding of the components and properties of the biological system. With the availability of genome sequences and high-throughput data, the mathematical and computational modeling of biological system has made tremendous advances in describing system-level behaviors and properties. For example, the dynamic analysis on kinetic models of metabolism has contributed to the understanding of temporal hierarchy of dynamic events, as well as the elucidation of fundamental dynamic structures of the network. Thermodynamic analysis on the biochemical reaction networks has revealed the fundamental constraints governing various cellular processes

and interactions. Additionally, constraint-based analysis in the context of genome-scale models of metabolism and gene expression has been used to compute the optimal metabolic flux state and proteome allocation for the given phenotype. In this dissertation, I am interested in applying multi-scale modeling to characterize the biological network structure and components. First, the dynamic network structures and properties at different timescales are explained through kinetic modeling of metabolism. Next, the evolutionary tradeoffs due to thermodynamic favorability and pathway yield in biosynthetic pathway choice of different organisms are revealed through the combination of thermodynamic analysis and genome-scale metabolic models. Last, the adaptive response of *E. coli* under acid stress is examined through laboratory evolution and such response is characterized through a mechanistic approach using genome-scale model of metabolism and gene expression.

Chapter 1

Evaluation of rate law approximations in bottom-up kinetic models of metabolism

1.1 Abstract

1.1.1 Background

The mechanistic description of enzyme kinetics in a dynamic model of metabolism requires specifying the numerical values of a large number of kinetic parameters. The parameterization challenge is often addressed through the use of simplifying approximations to form reaction rate laws with reduced numbers of parameters. Whether such simplified models can reproduce dynamic characteristics of the full system is an important question.

1.1.2 Results

In this work, we compared the local transient response properties of dynamic models constructed using rate laws with varying levels of approximation. These approximate rate laws were: 1) a Michaelis-Menten rate law with measured enzyme parameters, 2) a Michaelis-Menten rate law with approximated parameters, using the convenience kinetics convention, 3) a thermodynamic rate law resulting from a metabolite saturation assumption, and 4) a pure chemical reaction mass action rate law that removes the role of the enzyme from the reaction kinetics. We utilized in vivo data for the human red blood cell to compare the effect of rate law choices against the backdrop of physiological flux and concentration differences. We found that the Michaelis-Menten rate law with measured enzyme parameters yields an excellent approximation of the full system dynamics, while other assumptions cause greater discrepancies in system dynamic behavior. However, iteratively replacing mechanistic rate laws with approximations resulted in a model that retains a high correlation with the true model behavior. Investigating this consistency, we determined that the order of magnitude differences among fluxes and concentrations in the network were greatly influential on the network dynamics. We further identified reaction features such as thermodynamic reversibility, high substrate concentration, and lack of allosteric regulation, which make certain reactions more suitable for rate law approximations.

1.1.3 Conclusions

Overall, our work generally supports the use of approximate rate laws when building large scale kinetic models, due to the key role that physiologically meaningful flux and concentration ranges play in determining network dynamics. However, we also showed that detailed mechanistic models show a clear benefit in prediction accuracy when data is available. The work here

should help to provide guidance to future kinetic modeling efforts on the choice of rate law and parameterization approaches.

1.2 Background

Kinetic models of biochemical networks continue to grow in scope and scale [1–7]. The promise of these models is to serve as in silico platforms for prediction of complex system behavior and corroboration of experimental results. Specifically within metabolism, kinetic models have the potential to elucidate the control mechanisms underlying metabolic homeostasis and regulatory responses [8–10], as well as to identify flux bottlenecks impeding optimal performance of production strains [11]. To date, these models have been used to study such problems as the systemic effect of enzyme mutations [12, 13], metabolic bistability [10], and the coupling of signaling between metabolism and transcriptional regulation [3].

The primary challenge in kinetic modeling of metabolism is dealing with the frequent cases where data to construct detailed kinetic models is lacking [14]. This challenge is commonly addressed in part by selecting kinetic rate laws with particular approximations that reduce the number of parameters to be specified [15, 16]. If the assumptions made are valid across the conditions of interest, a consistent and predictive system should be obtainable by fitting parameters to available data [17]. Established examples of kinetic assumptions applied to enzyme reactions [5] include the quasi-steady state assumption utilized in Michaelis-Menten-type rate laws [4, 6, 18] and the lin-log approximation [2, 19] rooted in thermodynamic intuition. The degree to which these types of approximated systems represent the true system is a primary concern when choosing a modeling approach.

Here, we construct a set of kinetic models of red blood cell (RBC) metabolism using

various approximate rate laws, such that their parameters are equivalent to those of the fully-described enzyme mechanistic model. We choose the red blood cell due to the large amount of available data, enabling us to use physiological enzyme kinetic parameters, reaction fluxes, metabolite concentrations, and reaction equilibrium constants. Thus, we can examine the practical importance of rate law approximations against the backdrop of a realistic system.

We utilize these models to study the effect of simplifying assumptions to the rate laws on system dynamics through simulating the network response to small transient perturbations. We additionally discuss theoretical differences in the kinetic behavior of these rate laws. Finally, we iteratively replace approximate rate laws with mechanistic enzyme kinetics to examine whether we can anticipate general dynamic effects of certain types of approximations. We purposefully chose a simple perturbation approach with mathematical response properties as output metrics, as opposed to physiological prediction accuracy, in order to simplify the task of understanding any observed correlations or lack of correlations.

1.3 Results

1.3.1 Assumptions underlying rate law approximations

In preparation for investigating rate law effects through model simulation, we first discuss the assumptions underlying the different approximate rate laws. Perhaps the most well-known kinetic assumption is the QSS assumption, normally associated with Michaelis-Menten kinetics but originated by Briggs and Haldane [20]. This assumption states that all intermediate enzyme forms do not change concentrations over time (Figure 1.1a middle). Michaelis-Menten kinetics normally require Michaelis-Menten constants (K_ms) and catalytic constants (k_{cat}s) to param-

eterize the system, as well as metabolomics data, K_{eq}s of biochemical reactions, and enzyme concentrations. The conditions for validity of the assumptions underlying this rate law have been examined in great detail [21–27].

If sufficient kinetic data is lacking, but reproducing enzyme saturation behavior is desired, an additional assumption can be made to approximate the K_ms values. Previously it has been shown experimentally that enzyme K_ms values tend to be similar to the *invivo* concentrations of corresponding metabolites [28]. To determine whether this trend can be exploited to fill in unknown parameters, we examined the dynamic effect of using a $K_m s = x$ assumption to parameterize rate laws. If we additionally lack of knowledge about the enzyme reaction mechanism as well, the form of the QSS rate law equation into which parameters will be inserted is unclear. To deal with this, we can add a further assumption that the reaction follows a rapid equilibrium random order mechanism [29], following the previously suggested convenience kinetics formalism. We term this rate law with assumed rather than measured enzyme parameters as a Michaelis-Menten rate law with approximated properties (Figure 1.1a bottom right).

Another way to address cases where enzyme-specific data is lacking is to combine the QSS assumption with a different assumption that substrates are saturated relative to their binding constants, while products and inhibitors are of negligible concentration (i.e., $K_m s \ll x$ for substrates and activators while $K_m s \gg x$ for products and inhibitors). This assumption effectively removes enzyme-specific parameters from the rate law and leads to a thermodynamics-driven rate law similar to what has been termed Q-linear kinetics (Figure 1.1a bottom left) [30]. However, we note that Q-linear kinetics treats the mass action ratio Q as a thermodynamic variable while we treat the involved metabolites as separate variables. This Q-linear kinetics-like rate law is fully specified using only metabolomics, fluxomics, and K_{eq} data.

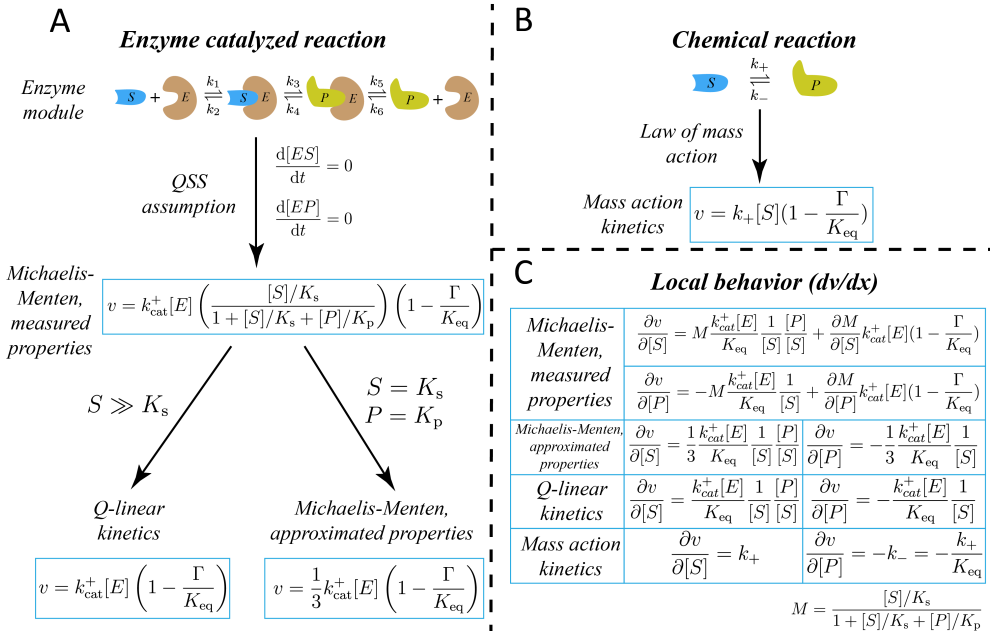


Figure 1.1: Comparison of rate laws and their resulting first derivatives. a) Formulation of Michaelis-Menten kinetics with measured properties, Q-linear kinetics and Michaelis-Menten kinetics with approximated properties from the enzyme module with different layers of assumptions [31]. b) Formulation of mass action kinetics based on the law of mass action for a pure chemical reaction. c) First derivatives (reaction sensitivities) calculated from the four approximate rate laws. K_s and K_p are the Michaelis-Menten constants for the substrate and product. Γ is denoted as the mass-action ratio, which is the ratio of product concentrations over reactant concentrations in a steady state raised to the exponent of their stoichiometric coefficients. K_{eq} is the equilibrium constant of the reaction. k^+_{cat} is the enzyme turnover rate constant. k_+, as defined in MASS models, is the pseudo-elementary rate constant in the forward direction.

The benefit of requiring fewer parameters is the major motivation for applying these simplified rate laws; however, before using them, we carefully examine whether they are able to accurately capture the dynamics of a model constructed of detailed enzyme modules. We might expect two general cases where rate law approximations should be successful. First, in cases where the underlying assumptions are valid, the rate law approximations should show accurate behavior provided that the assumptions are not violated substantially throughout the simulation. Second, if the rate laws are not the most important factor determining the dynamic behavior of the network, we would expect the use of an approximation to have little negative effect. For example, some of the rate laws may behave similarly near to equilibrium. In the course of this investigation, we will seek to identify both the degree to which approximate rate laws can reproduce the behavior of the true model, as well as the causes of this agreement or lack thereof.

1.3.2 Differences in mathematical behavior between rate laws

To place the subsequent results of simulating the various kinetic models in theoretical context, we briefly discuss differences between the analytical structures of the various rate laws. We focus on two key points: 1) the ability of the rate law to exhibit the saturation behavior that is characteristic of enzyme kinetics, and 2) the properties of the first derivative of the rate law, which defines the local dynamic behavior of the system.

Each rate law exhibits different behavior as metabolite concentrations approach infinity. For example, the Michaelis-Menten kinetics with measured properties exhibit the well-known saturation behavior due to the hyperbolic form, such that $v = v_{max}$ as x approaches infinity. A mass action enzyme module exhibits the same behavior due to the constant total enzyme, placing a constraining relationship between the fluxes of individual reaction steps. The manner in which

saturation is achieved between a full mass action enzyme module and the Michaelis-Menten kinetics is thus mathematically different.

In contrast to Michaelis-Menten kinetics with measured properties and enzyme module of mass action rate laws, the non-module mass action and Q-linear rate laws do not exhibit saturation behavior. Mass action kinetics will approach positive or negative infinity as substrate or product concentrations, respectively, approach infinity. Meanwhile, Q-linear kinetics exhibit asymmetrical saturation properties. The flux v will correctly have a maximum of v_{max} if the substrate concentration is maximized, but will incorrectly have a minimum of negative infinity if the product concentration is maximized. This asymmetry is known and proponents of the rate law suggest that the rate law only be used in a range near equilibrium [19], which is not possible to guarantee in real perturbations. For this reason, it is expected that the Q-linear kinetics and mass action kinetics will exhibit large deviations from the true mass action module system when perturbation of the saturation state of the enzyme is an important feature of the dynamic response.

Examining the first derivatives of the reactions is a straightforward analytical approach to anticipating dynamic differences between the rate laws (Figure 1.1c). From the analytical form of the rate law first derivatives, it is clear that the local dynamics between each type of rate law will be potentially substantially different, with numerical values dominated by different parameters in each case. The expressions for gradients obtained from the Michaelis-Menten kinetics with measured properties are complicated and multiple parameters play a role in affecting the numerical gradient values. The Michaelis-Menten kinetics with approximated properties and Q-linear kinetics rate laws have almost the same composition of their first derivatives, determined by enzyme turnover rate constant, the equilibrium constant and substrate and product concen-

trations. On the other hand, the local dynamic gradient in mass action kinetics is determined by the pseudo-elementary rate constants and equilibrium constant.

1.3.3 Construction and general properties of mass action modules for ten enzymes

We first constructed enzyme 'modules,' consisting of full mass action descriptions of enzymatic reaction mechanisms, for ten key enzymes in RBC central metabolism utilizing measured data for these enzymes (Figure 1.2, Table 1.1). An enzyme module consists of mass action rate laws for all known reaction steps such as substrate binding, catalytic conversion, and product release, as well as any activator or inhibitor binding (Figure 1.1a top). An enzyme module describes the detailed mechanism of enzyme catalysis and characterizes the dynamics of the enzymatic reaction subject only to certain basic assumptions such as deterministic behavior and a well-mixed solution [32]. The enzyme module requires a large number of parameters, including metabolomics data, equilibrium constants (K_{eq}s), enzyme concentrations, and rate constants of individual enzymatic reaction steps, to fully describe the dynamics of the system. We used these ten enzyme modules as a 'gold standard' for later comparison with approximate rate laws.

1.3.4 Construction of approximate rate laws

In this study, we examined four approximate rate laws to compare to the fully-described enzyme modules. Those four rate laws are: 1) Michaelis-Menten kinetics based on the quasi-steady state (QSS) assumption for the true enzyme module with measured enzyme parameters, 2) an assumed rapid-equilibrium random-order Michaelis-Menten rate law ignoring regulation and with K_m values being approximated as equal to the concentrations of corresponding metabolites,

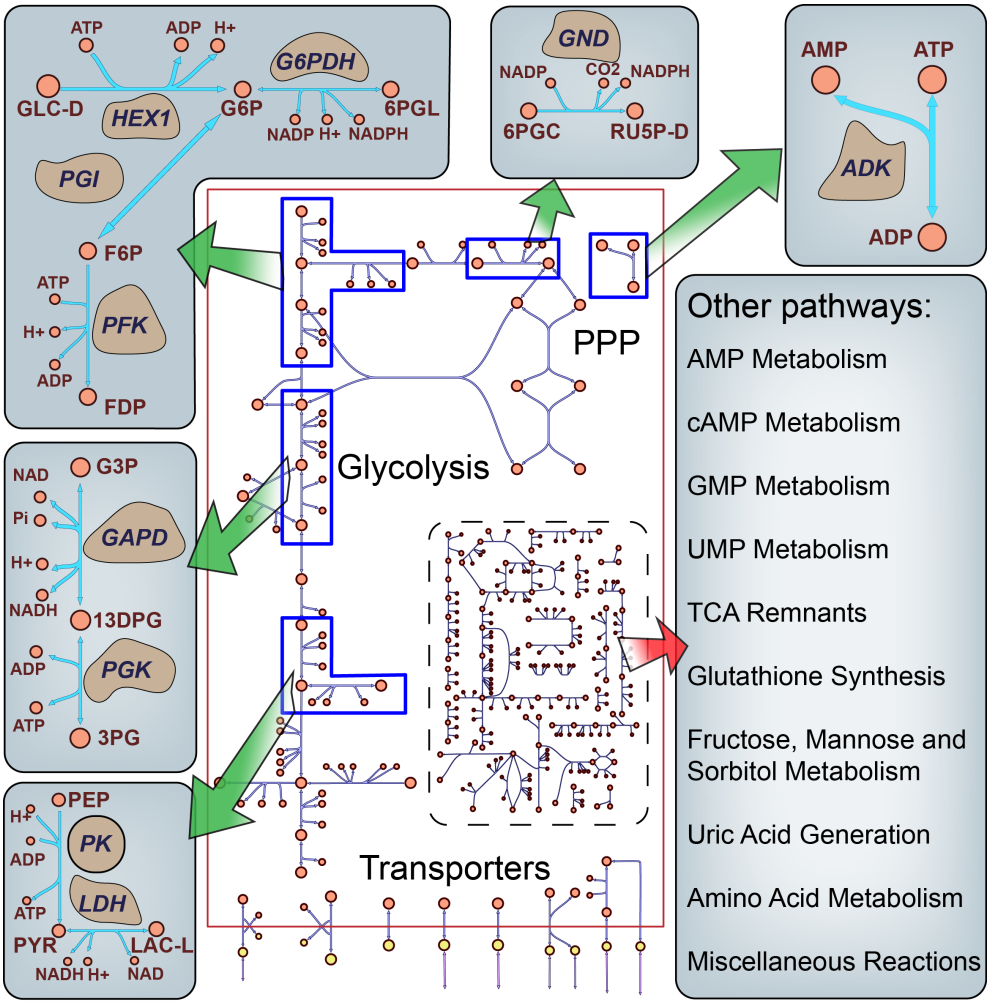


Figure 1.2: Schematic of the enzyme modules incorporated into the RBC metabolic network. The RBC metabolic network is based on a previous reconstruction [33]. The ten modules constructed were primarily located in glycolysis and the pentose phosphate pathway. Other pathways were included as Q-linear kinetics approximations.

Table 1.1: General description of the constructed enzyme modules

Enzyme name	Module size (metabolites × reactions)	Regulators (mechanism of action)
Phosphogluconate dehydrogenase (*GND*)	13 × 9	NADPH (PI)
Lactate dehydrogenase (*LDH*)	10 × 6	N/A
Glucose-6-phosphate dehydrogenase (*G6PDH*)	12 × 7	ATP (CI), NADPH (PI)
Glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*)	27 × 27	3PG (AI), G3P (AI)
Hexokinase (*HEX1*)	10 × 6	23DPG (CI)
Pyruvate kinase (*PK*)	30 × 34	FDP (AA), ATP (PI, AI)
Phosphofructokinase (*PFK*)	40 × 44	ADP (PI), ATP (AI), AMP (AA)
Phosphoglycerate kinase (*PGK*)	13 × 9	ATP (PI), 3PG (PI), 23DPG (CI)
Adenylate kinase (*ADK*)	8 × 5	N/A
Glucose-6-phosphate isomerase (*PGI*)	5 × 3	N/A

PI product inhibitor, *AI* allosteric inhibitor, *AA* allosteric activator, *CI* competitive inhibitor

to simulate the effect of unknown data, mechanisms, and regulation, 3) a rate law previously, termed Q-linear kinetics [30], containing only thermodynamic effects that results from a further metabolite saturation assumption, and 4) a rate law based on the mechanism of chemical mass action that effectively ignores the role of the enzyme in the reaction [5].

1.3.5 Construction of an approximate rate law scaffold model

We first constructed a cell-scale model of RBC metabolism using approximate Q-linear rate laws to serve as a scaffold model for analysis. Our approach was to insert the ten constructed enzyme modules into this scaffold, and compare this model behavior to that of models generated with different approximate rate laws substituted into those same ten reactions. The model was constructed using steady-state metabolite levels from plasma and intracellular erythrocyte metabolomics data from a fasting state [33]. The model contains 169 metabolites and

143 reactions, covering glycolysis, the pentose phosphate pathway, amino acid metabolism, and other pathways.

1.3.6 Designing a simulation-based kinetic analysis workflow

A straightforward way to estimate the similarity of behavior between different rate laws is to simulate the response of each model to perturbation. A perturbation in this case denotes the change of certain metabolite concentrations at time t = 0, after which the system is allowed to simulate through a long enough time such that the original steady state is once again reached. For example, we perturbed the concentrations of ATP, ADP and P_i at the same time to simulate the hydrolysis of ATP in the system.

Two key decisions in such an analysis are the choice of perturbation and the choice of output variable to observe. In this study, we perturbed both metabolites directly involved in as well as distant from the constructed enzyme modules. The list of perturbations can be found in Figure 1.3. To define output variables of interest, we created two metrics, the maximum perturbation (MP) and the relaxation time (RT). The MP is largest percent change in concentration compared to the steady state concentration that occurred during the simulation. Then, to calculate the RT of a metabolite, we identify the last time point at which the deviation from the steady state concentration is at least 5% of the MP.

One final decision in the simulation workflow is the size of the perturbation to use. As mentioned previously, the rate laws chosen differ in both saturation properties, which are non-linear features of the rate laws, and local dynamic properties, which are linear features of the rate law. It appeared to be a trivial result that saturating and non-saturating rate laws will exhibit very different behavior for large deviations where non-linear effects play a significant role.

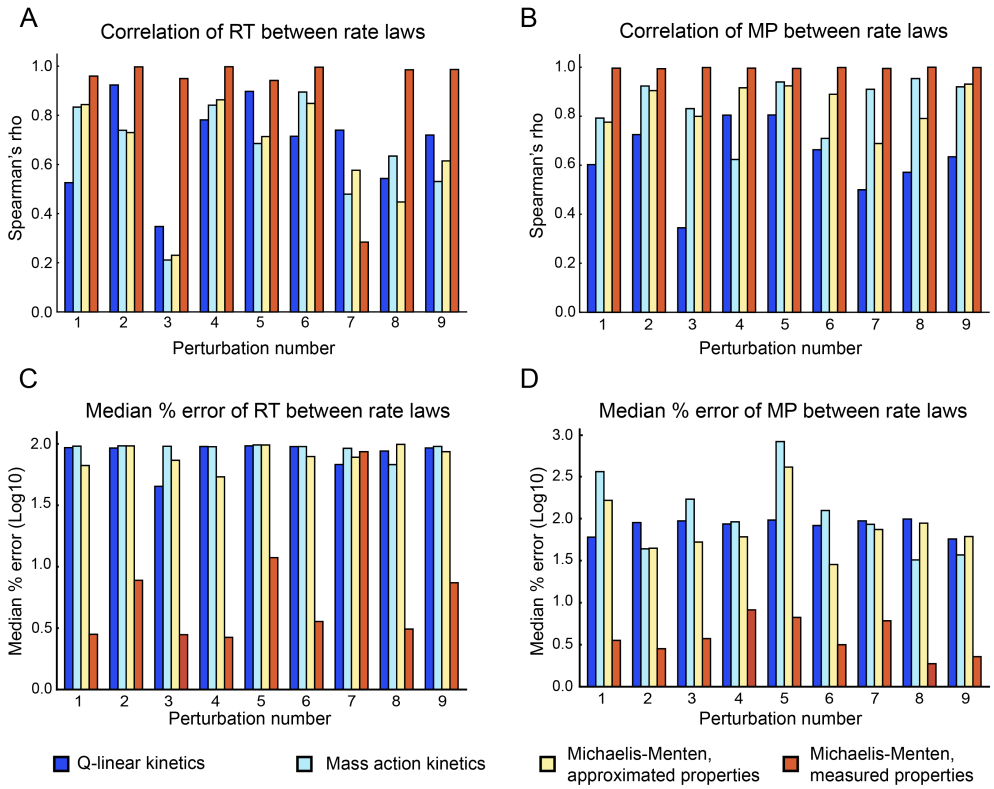


Figure 1.3: Simulation comparison of four simplified rate laws against a reference module containing detailed enzyme mechanism kinetics (enzyme modules). The responses of metabolites under different perturbations were compared between four simplified rate laws and the enzyme module. a) Correlation of metabolite relaxation time. b) Correlation of metabolite maximum perturbation. c) Median percent errors of metabolite relaxation time. d) Median percent errors of metabolite maximum perturbation. Nine different perturbations labeled from 1 to 9 were performed. 1, ATP, ADP and P_i perturbation; 2, NAD and NADH perturbation; 3, 23DPG perturbation; 4, 3PG perturbation; 5, PYR perturbation; 6, FDP perturbation; 7, PRPP perturbation; 8, MAN6P perturbation; 9, R5P perturbation. Spearman's rho: Spearman's rank correlation coefficient. The simulations were performed on the whole-cell kinetic model of erythrocyte constructed by Bordbar et al [33].

However, understanding the origin and nuances of such deviations is complex, and we sought to achieve a simpler goal as a baseline investigation. To avoid such obvious effects dominating our findings, we intentionally chose small perturbations to minimize saturation effects and instead focus on determining the importance of the linear/local differences between rate laws.

1.3.7 Numerical comparison of rate laws

The final workflow was to perform nine different small perturbations on the system with different rate laws and characterized the response of metabolites in terms of RT and MP (Figure 1.3a-b). Calculating the Spearman correlation for MP and RT of module metabolites between rate laws, we found that the Michaelis-Menten kinetics with measured properties behaved substantially better on both metrics compared to other rate laws. Median percent errors for MP and RT of module metabolites confirmed this trend (Figure 1.3c-d). Additionally, we found that the Michaelis-Menten rate law with approximated properties performed no better than the Q-linear kinetics and mass action kinetics. This indicates that the $K_m = x$ assumption (x being the concentration of the corresponding ligand) is not sufficiently correct to capture the dynamics of the original enzyme module. Notably we did not include known regulation of these enzymes in this approximate rate law, and further investigation of the behavior of models with the addition of these regulatory events with an analogous $K_d = x$ assumption may be warranted. We note that these conclusions regarding the suitability of approximate rate laws are not due to the choice of model underlying the analyses.

We repeated these analyses on a previously published model of the red blood cell, smaller scale but composed entirely of mechanistic enzyme mechanisms [34]. We iteratively substituted in different approximate rate laws and verified the identified trends, where Michaelis-Menten with

measured properties performs substantially better than the other approximations but all approximations retain positive correlation to the true model. We also verified the results using larger perturbations, suggesting that non-linearity of the perturbation response does not strongly affect the trends. However, as an exception to the general trends, we did identify rare perturbations where Michaelis-Menten rate laws with measured enzyme properties performed noticeably worse than more approximated rate laws. We attribute these cases to slow internal dynamics within the enzyme module, causing the quasi-steady state assumption to become invalid. However, these effects were difficult to isolate and we did not investigate these cases further due to their infrequency.

One key control in the study is to determine whether uncertainty in parameters significantly impacts the conclusions of analyses. To address this, we conducted Markov chain Monte Carlo (MCMC) convex sampling of steady-state fluxes given physiological ranges on metabolite uptakes and secretions [33]. Similarly, we conducted MCMC sampling of metabolite concentrations subject to a constraint on the feasibility of the concentrations with respect to the 2^{nd} law of thermodynamics [35]. We then combined sampled fluxes and concentrations and calculated rate constants for mass action rate laws for each reaction. It is found that the variation in rate constants due to flux and concentration uncertainty is small compared to the variation between rate constants of different reactions in the majority of cases. We also performed several simulations on models with these sampled rate constants, and found little variation in the RT or MP of metabolites across sampled models. Thus, it appears that experimental uncertainty in fluxes and concentrations, and the resulting uncertainty on estimated rate constants for simplified rate laws, is not a major concern in making claims about the dynamics of the network.

Since the simplified rate laws introduces noticeable discrepancies in dynamic behavior,

we wanted to determine whether these discrepancies would continue to increase as simplified rate laws are applied to more reactions until the correlation completely disappears, or whether the approximate model behavior would stabilize at some positive correlation to the true model. Based on the previous observation that Michaelis-Menten kinetics with measured properties closely resembled the true model, we set up a simple test case with as many reactions specified with Michaelis-Menten kinetics as possible (38 out of 168 reactions [33]) and then iteratively replaced them with mass action kinetics. We compared the RT and MP of the substrates and products of these reactions when a random set of reactions had their rate laws changed from Michaelis-Menten to mass action kinetics. We found that the correlation of RT and MP of metabolites between Michaelis-Menten and mass action kinetics stabilized as more reactions had their rate laws substituted (Figure 1.4). Since the discrepancy ceases to grow after a certain point, it appears likely that models with constructed entirely of simplified rate laws still be useful approximations of the real system, at least for small perturbations.

1.3.8 Effects of flux and concentration steady-state on network dynamics

We then investigated the source of the positive correlation between fully approximate models and the true model. As both models share the same initial steady state, in terms of reaction fluxes, metabolite concentrations, and reaction equilibrium constants, we sought to determine whether these values were essential to the dynamic consistency we observed across rate laws. The flux and concentration state of the cell play a role in determining the dynamic structure of the network. For example, large metabolite pools will be changed slowly by small fluxes, and vice versa, giving some expectation of fast and slow dynamics within the network. We wanted to investigate the degree to which network dynamics are determined by the initial flux

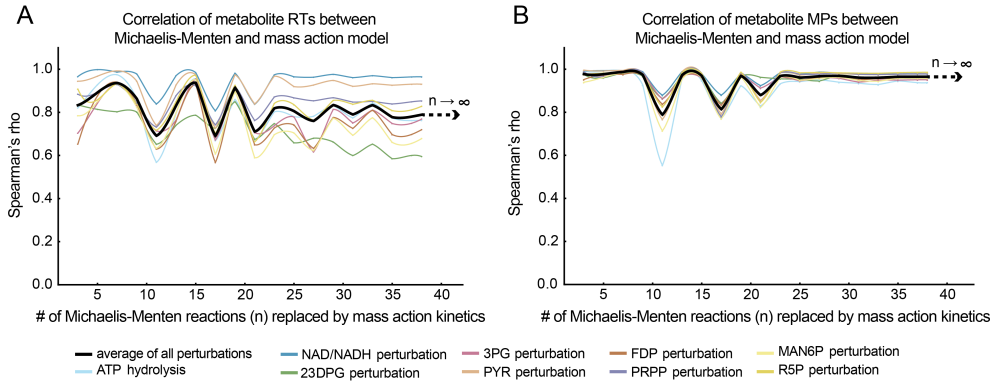


Figure 1.4: Iterative replacement of Michaelis-Menten kinetics with measured properties by mass action kinetics. An increasing number of Michaelis-Menten kinetics rate laws with measured parameters were replaced by mass action kinetics, and the RT and MP of affected metabolites were calculated. The correlation of metabolite RT and MP between Michaelis-Menten kinetics and mass action kinetics fluctuated initially but gradually stabilized as more reactions were replaced with mass action kinetics. The black line is the average correlation of all nine perturbations performed. a) Correlation of metabolite RTs between Michaelis-Menten and mass action model. b) Correlation of metabolite MPs between Michaelis-Menten and mass action model. Spearman's rho: Spearman's rank correlation coefficient. The simulations were performed on the whole-cell kinetic model of erythrocyte constructed by Bordbar et al [33].

and concentration state, as opposed to the choice of rate law. To this end, we sampled reaction fluxes and metabolite concentration within physiological ranges, and then in wider ranges. In contrast to changing rate laws, we found that widening the sampling range on fluxes and concentrations greatly impacted the dynamic response of metabolite throughout the network. For example, metabolite MP and RT subject to ATP hydrolysis perturbation showed weaker correlations within models sampled with wider concentration and flux ranges compared to those from models sampled with physiological concentration and flux ranges (Figure 1.5a-b). We also found that the distribution of metabolite RT and MP under ATP hydrolysis perturbation spanned a much larger range for models sampled with wider concentration and flux ranges (Figure 1.5c-d). Thus, it appears that the origin of the dynamic consistency across rate laws does indeed lie within the order of magnitude differences across reaction fluxes and metabolite concentrations throughout the network.

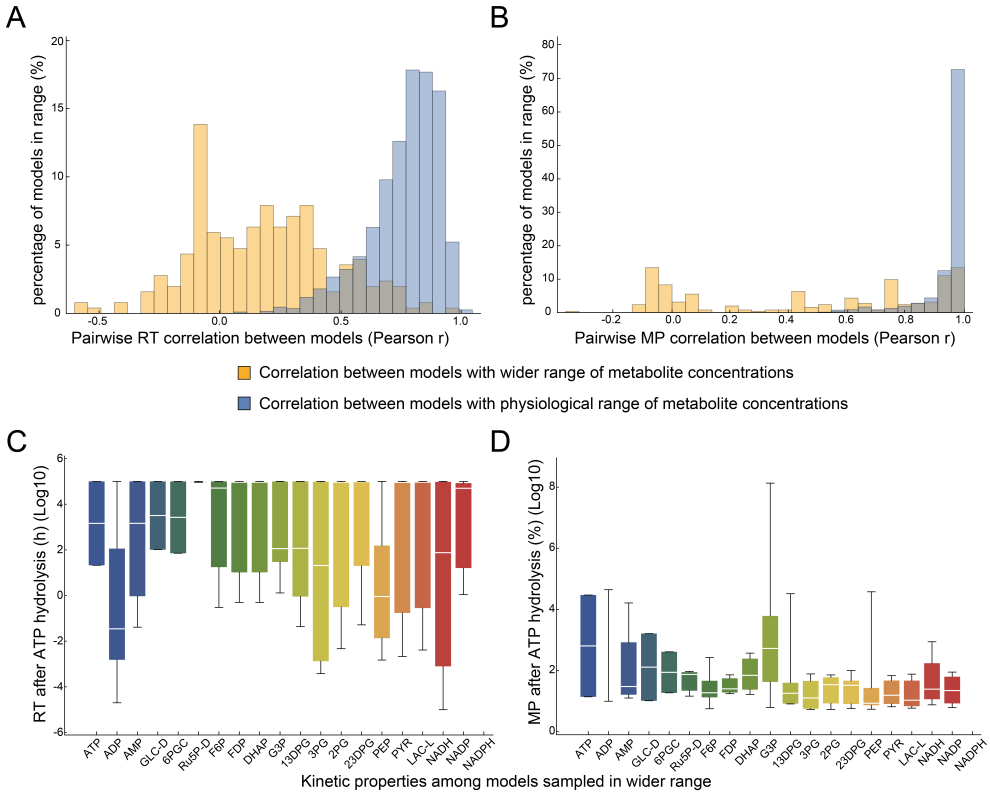


Figure 1.5: Kinetic properties of models sampled with physiological concentrations and fluxes compared to models sampled in wider ranges of concentrations and fluxes. First, 63 models were built with metabolite concentrations and fluxes sampled from physiologically relevant range. Then, 23 models were constructed with a wider range of metabolite concentrations (10^{-8} to 10^{5} mM) and fluxes. ATP hydrolysis was chosen as a reference perturbation as the perturbation on all models and RT and MP of the metabolites was calculated. a) Distribution of pair-wise Pearson correlation coefficients of metabolite RTs for models sampled with wider concentration and flux ranges and models sampled with physiologically relevant ranges. b) Distribution of pair-wise Pearson correlation coefficients of metabolite MPs for models sampled with wider concentration and flux ranges and models sampled with physiologically relevant ranges. c) Distribution of metabolite RTs for models sampled with wider concentration and flux ranges. d) Distribution of metabolite MPs for models sampled with wider concentration and flux ranges. The sampling and simulations were performed on the whole-cell kinetic model of erythrocyte constructed by Bordbar et al [33].

1.3.9 Dependence of the effect of rate laws approximations on reaction properties

We have showed that, while models constructed with approximate rate laws still hold valuable dynamic information due to the constraining effects of physiological flux and concentration differences, there is still a substantial increase in model accuracy from inclusion of additional kinetic information such as in a Michaelis-Menten rate law with measured properties. However, the question is still open of whether certain reactions are more necessary to model accurately than others. To probe this question, we began with a fully-defined mechanistic model [34], substituted each reaction in turn with a mass action approximation, and determined the effect on network dynamics. Clear trends emerged. First, reactions farther from equilibrium showed a larger effect from rate law approximation (Figure 1.6a). This is intuitive as irreversible reactions tend to be regulated allosterically, but the trend existed even for non-regulated enzymes. Second, certain reactions with metabolites that have high concentration tend to show a smaller effect by substitution of rate law approximation as well. For example, the enzymes *DPGASE* and *DPGM* are thermodynamically in an irreversible state but the high concentration of 23DPG creates a large slow moving pool that causes the dynamics of the network to be insensitive to the choice of rate law for these enzymes (Figure 1.6). However, there remain some unexplained cases, where reactions have one or both of these properties but rate law approximations result in effects outside of the general trend previously observed. For example, the enzymes *PGLASE* and *GSSGR* are clear outliers. This suggests that additional properties exist, such as network context given particular perturbations of interest, that may provide additional cases where rate law approximations work well.

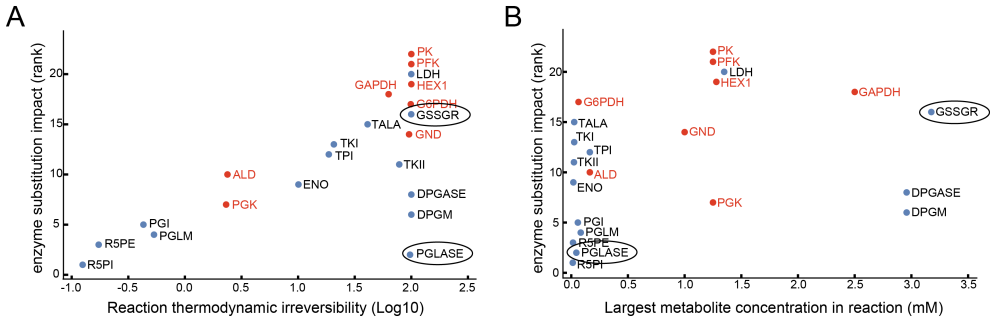


Figure 1.6: Reaction properties affecting the impact of reaction rate law approximations. a) Enzyme substitution impact (rank) against reaction thermodynamic irreversibility (Log10). Reaction thermodynamic irreversibility is calculated as (reaction equilibrium constant - mass action ratio)/reaction equilibrium constant. Lower rank score meant less change in dynamic response when the module is replaced by mass action kinetics. Reactions highlighted in red indicate presence of regulation. Circled reactions are outliers of the general trends. *PGLASE* is irreversible but shows low impact upon reaction rate law approximation. *GSSGR* has a large substrate concentration, yet still shows significant impact upon reaction rate law approximation. b) Enzyme substitution impact (rank) against largest metabolite concentration in the reaction. Red and circled reactions are the same as in panel (a). The simulations were performed on the model constructed based on Mulquiney et al [34].

1.3.10 Evaluating the consistency of effects of single enzyme mechanism substitutions throughout the network

One natural question to arise is whether it is possible to anticipate the changes to dynamic properties that occur when introducing enzyme mechanisms with particular features, such as allosteric regulation or a location upstream of a metabolite of interest, in place of an approximate rate law. For example, there exist some rules of thumb when dealing with small feedback networks, such as the role of negative feedback in increasing system response time, that might be applicable in these networks. However, we did not find such rules of thumbs to be reliable in the cases we examined.

In the case of the importance of network localization, for the nearby enzymes *PK* and *PGK*, there was no general trend observed in metabolite MPs and RTs under ATP hydrolysis perturbation following single module addition of *PK* or *PGK*. For example, the addition of the

PGK module slightly decreased the MP of lactate compared to no module while the addition of the *PK* module caused an increase in the MP of lactate, while from a structural standpoint we might expect the lactate node to have similar responses to the introduction of either enzyme mechanism. Along the same lines, the RT of 23DPG increased when adding the *PGK* module but decreased when adding the *PK* module. In addition to looking at the effect of different enzyme substitutions for a particular perturbation, we also looked across different perturbations for the same enzyme substitution. Specifically, we characterized the response of metabolite PYR under different perturbations upon the addition of the *PK* module and did not observe any general trend in the change of response.

As a case study for the effect of adding allosteric regulation, we chose the *HEX1* enzyme module, which contains 23DPG as a feedback inhibitor. We performed multiple perturbations on *HEX1* module with and without regulation and characterized the change in the dynamic response of the substrates and products of the enzyme. We found that G6P showed an increase in RT following addition of the feedback inhibitor, indicating that G6P relaxes more slowly following the addition of the inhibitor. The increase was also observed in metabolites downstream of the module. Meanwhile, G6P and F6P specifically showed an increase in MP with the addition of feedback inhibition. These observations appear contrary to the effect of feedback inhibition in simple feedback loops, where RT and MP decrease due to the effect of the inhibition [36]. This contradiction might be due to other interactions within the model, where metabolic reactions are usually nonlinear due to metabolites shared across multiple reactions. We performed the same analysis on the *GAPDH* module with 3PG as a feedback inhibitor. However, in this case we found a decrease in RT on FDP and G3P when the feedback inhibition was added, as well as a decrease in MP on 3PG and PEP. The two case studies above showed that the feedback

inhibition can cause quite different responses in different modules and the effect of regulatory mechanisms should be carefully considered on a case by case basis.

We also analyzed the effect of feedforward activation as an additional example of regulation. The example we studied the *PK* module with FDP as a feedforward activator. We found a decrease in RT for PYR in the *PK* module as well as a few metabolites upstream of the *PK* module, such as G6P, F6P, FDP, G3P and 3PG. Those metabolites also had a decrease in the MP (except 3PG and PYR). Again, this is contrary to the commonly observed effect of a simple feedforward loop, where RT and MP subsequently increase following addition of a feedforward activator [36]. Similar to the feedback inhibition, such contradiction may be attributable to more complex interactions within the metabolic network.

Overall, we showed that module addition can qualitatively affect the dynamics of related metabolites, but the quantitative effect can vary from case to case, possibly due to associated reaction and network connectivity properties. Therefore, it is difficult to predict any kind of consistent change moving from an accurate mechanistic description of enzyme catalyzed reactions to more approximate rate laws in specific cases.

1.3.11 Physiological and enzyme activity perturbations

Finally, while results so far were generated using perturbations of largely academic purpose, such as spontaneous internal metabolite changes, we sought to verify our results on perturbations of greater physiological meaning. First, we performed several simulations on decreased enzyme activity, in the form of a lower enzyme concentration or lowered catalytic rate constant, for the enzymes *G6PDH*, *PGK*, and *PK*, and verified the rate law trends identified thus far (see Methods). For example, the relative metabolite concentrations across different levels of *G6PDH*

activity were the same between enzyme module and Michaelis-Menten rate law with measured properties, while other rate laws showed noticeable differences. We made similar observations on relative metabolite level change across *PK* or *PGK* activity change, except that in *PGK* all rate laws behaved closely to the enzyme module. Then, we mimicked a previous study on an oxygen deprivation perturbation [37], and found that Michaelis-Menten rate law with measured properties was able to match exactly the dynamics of enzyme module, outperforming other approximated rate laws. However, none of the models quantitatively matched the experimental data well, suggesting confounding parameterization or model scope issues.

1.4 Discussion

In this work, we constructed a kinetic model of RBC metabolism with a mechanistic description of ten enzymatic reactions and compared the dynamic properties of the mechanistic model with those of several commonly proposed simplifying assumptions. We found that the Michaelis-Menten kinetics with measured properties yields a consistently good approximation of the full system, while the Q-linear kinetics and mass action kinetics can show substantial discrepancies. Furthermore, we formulated another Michaelis-Menten-type rate law in an attempt to simplify the Michaelis-Menten kinetics given limited data available, based on a $\mathrm{K_m} = \mathrm{x_{ss}}$ assumption with a rapid-equilibrium random order binding reaction scheme. However, this approach failed to show improved agreement in dynamics with the enzyme modules over other approximations. We attribute the positive correlation of even the most approximate rate laws with the true model as due to the important effect that reaction flux and metabolite concentration differences play in the network dynamics.

Obtaining enzyme kinetic parameters continues to be a core issue hindering the devel-

opment of practical large-scale kinetic models of metabolism. Databases such as BRENDA [38] continue to aggregate studies on the kinetic properties of enzymes for various organisms. However, not only are the collections of the most common kinetic parameters (K_ms and k_{cat}s) often incomplete and measured under non-physiological conditions, but there is a separate issue with the additional parameters that are required to parameterize a mass action mechanistic description of a reaction (which we term an enzyme module). Full specification of kinetic parameters is experimentally intensive but theoretically possible, and some enzymes such as *PFK* have been characterized in great detail in particular organisms, including pH and temperature dependence of parameters. However, the difficulty in determining these parameters and uncertain immediate value of the data, evidenced by lack of practical applications of resulting kinetic models, is likely the main reason these data are not routinely being generated. In this study, we show both the value of fully-defined enzyme mechanism as well as rate law approximations, and thus it appears that the appropriate rate law to use should continue to be determined by the goals of the modeler.

On the note of the design of this study, we note that kinetic models can be analyzed from numerous angles. Much work thus far has focus on the dynamic control of metabolic states. This goal is of great importance, but due to the non-linear and complex nature of such control, we targeted our investigation on a simpler task of understanding transient responses to small perturbations in the metabolic network. Experimentally measuring such transients, i.e., dynamics of metabolite concentrations, is challenging and fundamentally limited by sampling frequency and metabolism quenching time. However, we chose to focus on these perturbations as they are the most simple to understand mathematically. Further studies looking at the effect of rate law approximations on more intricate dynamic properties, such as the non-linear control of steady-state changes following enzyme inactivation, are extremely desirable if they can be

conducted in a rigorous way.

In our comparison of rate laws, we showed that the Michaelis-Menten kinetics with measured properties gives a good approximate of the full system when comparing the relaxation time and maximum perturbation of the metabolites. Thus, discrepancies due to ignoring dynamics of individual enzyme forms do not appear to be a significant issue. This success in approximation is likely due to the combination of the small concentrations of most enzyme forms relative to metabolite concentrations, a requirement for the validity of the QSS assumption [21], as well as the relatively large rate constants for reactions involved in enzyme regulation (effector binding) and structural transitions. For enzymes with larger concentrations and slow regulatory enzyme motions, there would likely be substantial discrepancies from using a QSS assumption. We also found that additional approximations from assuming saturation or neglecting enzyme behavior entirely cause substantial dynamic and structural issues. While these methods are attractive due to obviating the need for enzyme-specific parameters, the potential drawbacks may preclude their use. As an alternative, assumptions about enzyme parameters can be made in place of assumptions about rate laws. For example, one study has shown that metabolite concentrations tend to hover around the K_ms for corresponding enzymes [28], which could be a useful assumption for modeling in lieu of sufficient data. However, in practice, we found this assumption to be insufficient to recapitulate enzyme kinetic behavior, as deviations of the real data from this assumption were sufficiently large to induce substantial differences in behavior.

We showed that adding a module can bring qualitative effects to the dynamics of related metabolites. However, the quantitative effects have to be examined in a context specific manner, possibly due to the associated reaction property or network connectivity. We also showed that the addition of regulations, such as feedback inhibition and feedforward activation, can cause

dynamic behavioral changes different from those of simple genetic circuits. Taken together, we would advise a detailed mechanistic description for enzyme catalyzed reaction is likely a necessity for predicting system dynamics with reasonable accuracy.

There are two additional possible issues associated with modeling enzyme kinetics using an enzymatic mass action approach. The first is the estimation of kinetic parameters within the module. The current available experimental data on the enzyme include K_ms, v_{max} and K_ds. However, those data are not sufficient to solve for the rate constants of specific enzymatic steps in the module. Thus, a good fitting approach is necessary to obtain a set of rate constants that accurately recapitulate the existing experimental data. The second problem is associated with the simulation of the system containing multiple modules. A possible stiffness issue can occur when integrating the ODE equations during dynamic simulations. This might be due to the large difference in orders of magnitude between metabolite concentrations and enzyme intermediate concentrations. In this case, we would advise normalizing the enzyme concentrations to the same level as metabolite concentrations and adjust the corresponding rate constants. However, one needs to be careful with the magnitude of change in enzyme concentrations as we found that different changes can cause different dynamic responses. Looking forward, addressing these issues will be essential to make progress toward bottom-up construction of kinetic models of metabolism.

1.5 Conclusion

The work here explored the validity of using approximate rate laws with varying levels of assumptions in the context of a cell-scale RBC kinetic model. We found that the Michaelis-Menten rate law based on quasi-steady state assumption was able to recapitulate the dynamic

behaviors of the mechanistic model consistently as long as measured parameters were used. Rate laws that are derived from further approximations on Michaelis-Menten kinetics or ignore the role of the enzyme showed substantial discrepancies in dynamic behaviors compared to the mechanistic model. However, we found that the errors associated in these approximate models appeared to stabilize as more reactions were replaced by approximate rate laws, suggesting that even fully approximate models can contain useful information. This appears to be due to the dominant effect that the order of magnitude differences in reaction fluxes and metabolite concentrations have on the dynamic structure of the network. Still, we also found that replacing approximate models with the detailed mechanistic enzyme module can bring unpredictable quantitative effects to the system, suggesting a clear benefit of constructing mechanistically detailed enzyme modules when possible. The work here should aid the choice of rate laws and parameterization approaches in future kinetic modeling efforts.

1.6 Methods

All work was done in Mathematica. We used the MASS Toolbox kinetic modeling package (https://github.com/opencobra/MASS-Toolbox) for model construction and simulation. The RBC metabolic network with enzyme modules incorporated is available in Mathematica file format.

1.6.1 Construction of enzyme modules

The mass action rate law was used for reactions in enzyme modules, and the formulation can be found in Jamshidi et al.[5]. The steps for constructing enzyme modules are as follows: 1) Define elementary reactions and obtain their equilibrium constants from literature; 2) Formulate

the steady state mass balances for enzyme forms and solve them symbolically in terms of parameters of the reactions; 3) Substitute the symbolic enzyme forms into the equation of total enzyme concentration and approximate the rate constants of the reactions given a particular flux state; 4) Calculate concentrations of individual enzyme forms given the estimated rate constants.

For enzyme module with regulation, an additional enzymatic step was added in which the effector molecule (activator or inhibitor) is bound to a particular enzyme form.

1.6.2 Simulation of the network with the incorporated enzyme modules

The constructed modules were added into the RBC metabolic network [33] for further analysis. For incorporation of a specific module (e.g., *PFK* module), all the reactions in the module were added into the metabolic network and the original metabolic reaction (*PFK* reaction) was removed.

Before dynamic simulations, the steady state metabolite concentrations were set as the initial conditions of the system. For a particular perturbation, a change on certain metabolite concentrations were applied at time 0 and the subsequent simulation was conducted through numerical integration of the ODE equations. The system was allowed to simulate to over 100,000 h to regain the steady state concentrations.

1.6.3 Calculation of maximum perturbation and relaxation time

Given a concentration profile from simulation, the maximum perturbation is the largest percent change in concentration compared to the steady state concentration for a particular metabolite. The relaxation time is defined as the last time point at which the deviation from the steady state concentration is 5% of the maximum perturbation. Specifically, when calculating

the relaxation time, we traced backwards by starting from the concentration at a 'long enough' time (e.g., 100,000 h) and calculated the difference between the concentration at a particular time and the steady state concentration until the relaxation time was identified.

1.6.4 Constructing a model full of enzyme modules

We used the scope (Mulquiney et al [34] Scheme 1) and kinetic data (Mulquiney et al [34] Appendix) to construct a model full of enzyme modules. Specifically, the model contains 22 modules, mainly falling in glycolysis and pentose phosphate pathway. The enzyme modules were constructed based on the method previously described. We also added in the enzyme module for hemoglobin, which can be loaded from MASS Toolbox kinetic modeling package (https://github.com/opencobra/MASS-Toolbox). There are extra 13 reactions in the model that we did not build enzyme modules for. They are export/import reactions, generic metabolic reaction without specific reference to an enzyme and reactions with zero flux. Specifically, they are AMP export reaction, AMP import reaction, CO2 export reaction, glucose import reaction, proton export reaction, water export reaction, lactate export reaction, O2 export reaction, pyruvate export reaction, ATP hydrolysis reaction, glutathione redox reaction, NADH redox reaction, adenylate kinase reaction.

1.6.5 Enzyme activity simulation

The metabolic state of the system was simulated with different levels of enzyme activities, for the three enzymes *PK*, *PGK* and *G6PDH*. To simulate changing activity in the enzyme module, the total enzyme concentration was multiplied by a certain fraction. To simulate changing enzyme activity in simplified rate laws, the rate law equation was multiplied by a certain fraction.

After changing enzyme activities, the new steady state was obtained by simulating the system for a long enough time. The metabolite concentrations and associated metabolic states (e.g., inhibited hemoglobin level) were compared across rate laws and verified against physiological studies. All simulations were performed on the model constructed based on Mulquiney et al [34].

1.6.6 Iterative substitution of approximate rate laws in place of enzyme modules

We started with the model constructed based on Mulquiney et al [34] (containing 22 enzyme modules) and iteratively replaced the modules with four different simplified rate laws. We iteratively increased the number of modules replaced by rate laws, at intervals of 1, 2, 3, 6, 9, 12, 15, 18 and 22. Together with the original model consisting entirely of enzyme modules, we built a total of 37 models with different rate laws. We then performed 18 different perturbations on those models. The perturbations fell into three main categories: local metabolite perturbations where change of metabolite concentration is less than 10%, non-linear metabolite perturbations where change of metabolite concentration is greater than 10%, perturbations through rate constant where the rate constant of a particular reaction was altered. Models with replaced rate laws were compared against model containing all enzyme modules through correlation and percent error in metabolite RT and MP.

1.6.7 Single module replacement

To test the effect of replacing single module on the network dynamics, we started with the model constructed based on Mulquiney et al [34] (containing 22 enzyme modules) and built 22 different models by replacing each of the enzyme modules with mass action kinetics in a single

model. We then compared those 22 models against the original model consisting entirely of enzyme modules through correlation of metabolite RT across 18 different perturbations. We ranked each model based on its metabolite RT correlation with the original model in a perturbation. We then summed up the rank scores for each model across 18 different perturbations to obtain their final rank score. Lower rank score meant less change in dynamic response when the module is replaced with mass action kinetics. We compared the final rank against two factors that could determine the impact of simplified rate law replacing the enzyme module. One factor is reaction thermodynamic irreversibility, which is calculated as (reaction equilibrium constant - mass action ratio)/reaction equilibrium constant. The other is the largest metabolite concentration in the reaction.

1.6.8 Parameter sampling

We used the model constructed by Bordbar et al [33] for parameter sampling. The range of metabolite concentrations were based on the physiologically measured concentrations from 24 healthy individuals [33]. For unmeasured metabolites whose concentrations were taken from literature, their range was set based on the average standard error of measured metabolite concentrations. The sampled metabolite concentrations were constrained by the second law of thermodynamics, where equilibrium constants of the reaction were derived from eQuilibrator [39, 40]. We then used gpSampler in cobratoolbox to obtain 1000 sets of metabolite concentrations that fell in the physiologically relevant range and satisfied the thermodynamic constraint [35, 41]. The sampled fluxes of the model were obtained directly from Bordbar et al [33]. The rate constants of the reactions were then calculated from equilibrium constants, sampled metabolite concentrations and sampled fluxes. As a result, a total of 300 models were constructed from the

sampled parameters, concentrations and fluxes.

To compare the dynamic behavior of models with different sets of parameters, concentrations and fluxes, we performed three different perturbations on the 300 sampled models. The three perturbations were: changing ATP, ADP, P_i concentrations, changing NAD/NADH concentrations and changing FDP concentration. It was worth noting that only 63 models were able to achieve stable steady states after the perturbations. The RT and MP of the metabolites in those models were calculated from the perturbation profiles. We then selected metabolites with MP over 5% and compared the dynamic response across models.

1.6.9 Parameter sampling

We used the model constructed based on Mulquiney et al [34] (containing 22 enzyme modules) for physiological simulation. The physiological condition we chose was the hypoxia state of erythrocytes, and we simulated such a state by changing the external concentration of oxygen to 30% of its original level. Due to the known role of Band III (BIII) protein in erythrocytes under hypoxia condition, we added binding reactions of BIII to hemoglobin, *PFK*, *GAPDH* and *ALD* [37]. We replaced the rest of the modules with different approximate rate laws, simulated the models under hypoxia condition for long enough time until steady state was reached, and compared the time profiles of metabolites across rate laws.

Abbreviations

13DPG, 3-Phospho-D-glyceroyl phosphate; 23DPG, 2,3-diphosphoglycerate; 2PG, 2-Phospho-D-glycerate; 3PG, 3-Phospho-D-glycerate; 6PGC, 6-Phospho-D-gluconate; 6PGL, 6-phospho- D-glucono-1,5-lactone; *ADK*, adenylate kinase; ADP, adenosine diphosphate; ALD,

aldolase; AMP, adenosine monophosphate; ATP, adenosine triphosphate; CO2, carbon dioxide; DHAP, dihydroxyacetone phosphate; *DPGASE*, bisphosphoglycerate phosphatase; *DPGM*, bisphosphoglycerate mutase; *ENO*, enolase; F6P, D-Fructose 6-phosphate; FDP, D-Fructose 1,6-bisphosphate; G3P, glyceraldehyde 3-phosphate; G6P, D-Glucose 6-phosphate; *G6PDH*, glucose 6-phosphate dehydrogenase; *GAPDH*, glyceraldehyde 3-phosphate dehydrogenase; GLC-D, D-Glucose; *GND*, phosphogluconate dehydrogenase; *GSSGR*, glutathione reductase; H+, hydrogen ion; *HEX1*, hexokinase; LAC-L, L-Lactate; *LDH*, lactate dehydrogenase; MAN6P, D-Mannose 6-phosphate; MP, maximum perturbation; NAD, Nicotinamide adenine dinucleotide; NADH, Nicotinamide adenine dinucleotide - reduced; NADP, nicotinamide adenine dinucleotide phosphate; NADPH, nicotinamide adenine dinucleotide phosphate - reduced; O2, oxygen; PEP, phosphoenolpyruvate; *PFK*, phosphofructokinase; *PGI*, glucose 6-phosphate isomerase; *PGK*, phosphoglycerate kinase; *PGLASE*, 6-phosphogluconolactonase; *PGLM*, phosphoglycerate mutase; Pi, orthophosphate; *PK*, pyruvate kinase; PRPP, 5-Phospho-alpha-D-ribose 1-diphosphate; PYR, pyruvate; QSS, quasi-steady state; R5P, alpha-D-Ribose 5-phosphate; *R5PE*, ribulose 5-phosphate epimerase; *R5PI*, ribulose 5-phosphate isomerase; RBC, red blood cell; RT, relaxation time; RU5P-D, D-Ribulose 5-phosphate; *TALA*, transaldolase; *TKI/TKII*, transketolase; *TPI*, triose phosphate isomerase

Acknowledgments

We acknowledge support by the German Research Foundation (DFG) and the Open Access Publishing Fund of the University of Tuebingen. This work was supported by NIH grant GM068837. AD is funded by a Marie Curie International Outgoing Fellowship within the EU 7^{th} Framework Program for Research and Technological Development (project AMBiCon, 332020).

Chapter 1 in full is a reprint of material published in: **Bin Du**, Daniel C. Zielinski, Erol S. Kavvas, Andreas Dräger, Justin Tan, Zhen Zhang, Kayla E. Ruggiero, Garri A. Arzumanyan, Bernhard O. Palsson. 2016. "Evaluation of rate law approximations in bottom-up kinetic models of metabolism." *BMC Systems Biology*, 10(1), 40. The dissertation author was the primary author.

Chapter 2

Topological and kinetic determinants of the modal matrices of dynamic models of metabolism

2.1 Abstract

Large-scale kinetic models of metabolism are becoming increasingly comprehensive and accurate. A key challenge is to understand the biochemical basis of the dynamic properties of these models. Linear analysis methods are well-established as useful tools for characterizing the dynamic response of metabolic networks. Central to linear analysis methods are two key matrices: the Jacobian matrix ($\mathbf{J}$) and the modal matrix ($\mathbf{M}^{-1}$) arising from its eigendecomposition. The modal matrix $\mathbf{M}^{-1}$ contains dynamically independent motions of the kinetic model near a reference state, and it is sparse in practice for metabolic networks. However, connecting the

structure of $\mathbf{M}^{-1}$ to the kinetic properties of the underlying reactions is non-trivial. In this study, we analyze the relationship between $\mathbf{J}$, $\mathbf{M}^{-1}$, and the kinetic properties of the underlying network for kinetic models of metabolism. Specifically, we describe the origin of mode sparsity structure based on features of the network stoichiometric matrix $\mathbf{S}$ and the reaction kinetic gradient matrix $\mathbf{G}$. First, we show that due to the scaling of kinetic parameters in real networks, diagonal dominance occurs in a substantial fraction of the rows of $\mathbf{J}$, resulting in simple modal structures with clear biological interpretations. Then, we show that more complicated modes originate from topologically-connected reactions that have similar reaction elasticities in $\mathbf{G}$. These elasticities represent dynamic equilibrium balances within reactions and are key determinants of modal structure. The work presented should prove useful towards obtaining an understanding of the dynamics of kinetic models of metabolism, which are rooted in the network structure and the kinetic properties of reactions.

2.2 Background

In recent years, kinetic models of metabolism have become increasingly detailed, comprehensive, and consistent with the underlying biochemistry and genetics [1, 2, 5, 6, 33, 42]. These models can address a number of questions that are difficult to analyze directly with constraint-based or statistical models [43–45]. For example, kinetic models have shown utility in the study of: 1) regulatory mechanisms controlling the cellular metabolic network [9, 10], 2) complex dynamic behavior such as bistability [46], 3) intracellular signal transduction [47], and 4) the effect of enzyme mutations on a network scale [12, 13]. Furthermore, predictive kinetic models are desirable in metabolic engineering to improve production, substrate utilization, and product quality [14, 48].

A grand challenge moving forward is to analyze the dynamic properties of these models to obtain a deeper understanding of the structure and function of the metabolic network. A number of studies have made theoretical and practical headway in this regard by analyzing the linear properties of the dynamic system around a steady state. These linear analysis methods have helped to provide insight into metabolic flux control [49, 50], elucidate the temporal hierarchy of dynamic events [51], and describe the fundamental dynamic structure of the network [52].

At the core of these linear analysis methods is the modal matrix ($\mathbf{M}^{-1}$) resulting from the Jacobian matrix ($\mathbf{J}$) of the mass balance equation. The modal matrix contains dynamically decoupled motions of the metabolic network, called modes. For real metabolic networks, the modal matrix has a sparse structure [51], the interpretation of which can yield biological insight into dynamics occurring on particular time scales. However, while $\mathbf{M}^{-1}$ is a numerically-calculated matrix, $\mathbf{J}$ can be represented symbolically in terms of derivatives of the reaction rate laws ($d\mathbf{v}/d\mathbf{x}$) in the network. Thus, obtaining an understanding of the structure of $\mathbf{M}^{-1}$ in terms of the structure of J would allow us to connect the dynamics of the network to the kinetic properties of single reactions, providing insight into the origin of the network dynamic structure. Linear analysis is well-known in classical chemical reaction kinetics literature and has been applied to metabolic networks specifically in the form of metabolic control analysis (MCA) [53], which focuses on a scaled gradient ($d\mathbf{v}/d\mathbf{x}$) matrix $\mathbf{G}$. However, less work has been performed on modal ($\mathbf{M}^{-1}$) analysis of metabolic networks, and specifically very little has been discussed about why the modes of metabolic networks have particular sparsity structures.

In this study, we present results on the biochemical origin of the modal sparsity structure of kinetic models of metabolism, using the metabolic network of the human red blood cell (RBC) [54]. This model consists of ten enzyme mechanisms represented by mass action kinetics inserted

in a background of 133 approximated rate law reactions [5, 30], parameterized with measured metabolite concentrations and enzyme kinetic constants. It is essential that this analysis be performed on a real metabolic network rather than toy models, because the metabolic network topology as well as order of magnitude differences in reaction fluxes, metabolite concentrations, and reaction rate constants are essential features in determining the dynamics of the network [54].

Using both numerical and theoretical arguments, we demonstrate how the dynamic structure of the modal matrix $\mathbf{M}^{-1}$ forms due to specific properties of the Jacobian $\mathbf{J}$ matrix. Using Gershgorin circle theorem, we first show that simple dynamic structures often emerge due to the kinetic parameter scaling in metabolic networks. Then, we use the matrix power iteration algorithm to show how modes with more complicated sparsity structures arise from topologically connected elements of $\mathbf{J}$ that have similar magnitude. Furthermore, we describe how such complicated mode structures arise due to similar dynamic equilibrium ratios of connected reactions.

We focus on demonstrating general principles through a set of case studies on the concentration Jacobian matrix and the mode structures associated with metabolite groups. These principles also apply to the flux Jacobian matrix and the relate flux modal structures, which are characterized in terms of the flux variables and describe the dynamic properties of the reaction groups [55].

2.2.1 Linear analysis on dynamic structures of the metabolic network

We first briefly introduce the basic established theory for linear analysis of metabolic networks. In a biochemical reaction network, the dynamic mass balances for all m concentrations $\mathbf{x}$ are given in the form of a matrix equation:

$$d\mathbf{x}/dt = \mathbf{S} \cdot \mathbf{v}(\mathbf{x}, \mathbf{k}) \tag{2.1}$$

where $\mathbf{S}$ is the $m \times n$ stoichiometric matrix, $\mathbf{x}$ is the $m \times 1$ vector of metabolite concentrations, and $\mathbf{v}$ is the $n \times 1$ vector of reaction fluxes. The formulation of $\mathbf{v}$ depends on the reaction rate law used and the mass action rate law is expressed as a function of the concentrations $\mathbf{x}$ and kinetic parameters $\mathbf{k}$.

Linearizing around a particular steady state $\mathbf{x}_0$ (i.e., $\mathbf{S} \cdot \mathbf{v}(\mathbf{x}_0, \mathbf{k}) = 0$ yields,

$$d\mathbf{x}'/dt = \mathbf{J} \cdot \mathbf{x}' \tag{2.2}$$

where $\mathbf{x}' = \mathbf{x} - \mathbf{x}_0$ are the concentration deviation variables from the steady state and $\mathbf{J} = \mathbf{S} \cdot \mathbf{G}$ is the concentration Jacobian matrix [51]. $\mathbf{G}$ ($= d\mathbf{v}/d\mathbf{x}$) is the gradient matrix obtained from linearization of the reaction rates [30]. It is the same matrix as the non-normalized elasticity matrix from metabolic control analysis [50, 56].

An eigen-decomposition of the Jacobian matrix yields a different representation of the same linearized system, with dynamically independent motions of metabolites grouped into modes within the modal matrix [51].

$$\mathbf{J} = \mathbf{M} \cdot \Lambda \cdot \mathbf{M}^{-1} \tag{2.3}$$

where $\mathbf{M}^{-1}$ is the modal matrix and Λ is the diagonal matrix of eigenvalues. During eigen-decomposition, we can append the left null space vectors of the Jacobian matrix to the modal matrix and assign those vectors zero eigenvalues. This operation makes both modal matrices full rank since a rank deficient matrix is not invertible. The modes are defined as $m = \mathbf{M}^{-1} \cdot \mathbf{x}$.

Substituting Eq 2.3 into Eq 2.2, and based on the mode definitions, we have,

$$dm/dt = \Lambda m \tag{2.4}$$

As defined in Eq 2.4, the eigenvalues and modes give information on the dynamically independent motions of metabolite groups [51].

The rows of the modal matrix, which correspond to modes, are left eigenvectors of $\mathbf{J}$ ($u\mathbf{J} = \lambda u$). Each mode is associated with an eigenvalue and represents the dynamic motion in a characteristic time scale defined by the eigenvalue. These characteristic time scales describe the approximate time it takes for the mode to relax (return near its original reference state) when the system is perturbed from steady state. Our focus in this work is to examine the sparsity structure of the modes and determine how this structure is connected to properties of the Jacobian matrix.

2.3 Results

2.3.1 Half-reaction equilibria resulting from linearization of bilinear mass action rate laws are key dynamic features of G

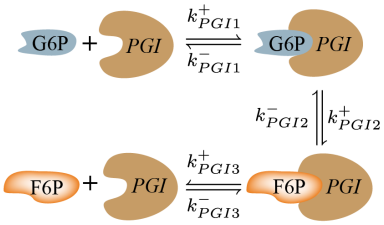
To aid in later discussions on mode sparsity structure, we first introduce the key concept of half-reaction equilibria, which appear in $\mathbf{G}$ due to linearization of mass action reactions. For mass action reactions, the $d\mathbf{v}/d\mathbf{x}$ derivatives comprising the gradient matrix $\mathbf{G}$ ($= d\mathbf{v}/d\mathbf{x}$) have a specific mathematical form and biochemical interpretation (Figure 2.1). The form of the mass action rate law for an example bilinear reaction between metabolite A and enzyme form E where $\mathrm{A} + \mathrm{E} \leftrightarrow \mathrm{EA}$ is $v = k^+[\mathrm{A}][\mathrm{E}] - k^-[\mathrm{EA}]$, and the three resulting $d\mathbf{v}/d\mathbf{x}$ terms in $\mathbf{G}$ for the reaction are $k^+[\mathrm{A}]$, $k^+[\mathrm{E}]$, $-k^-$. From these three terms, we can see that certain reactant/product terms

are eliminated when calculating the reaction sensitivities (derivatives in the form of $d\mathbf{v}/d\mathbf{x}$) in $\mathbf{G}$. This mathematical operation can be interpreted as splitting the original reaction into ***half reactions*** in a biochemical context. In the case of bilinear kinetics of enzymatic binding/release reactions, the half reaction describes the binding/release process for one reactant, which is held constant.

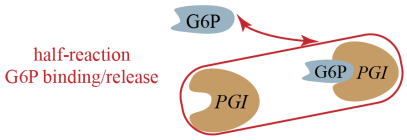
For a full reaction, the distance from equilibrium is defined as Γ/K_{eq}, where Γ is the mass action ratio and K_{eq} is the equilibrium constant. Thus, for the example bilinear reaction mentioned above, its distance from equilibrium can be expressed as $k^-[\mathrm{EA}]/k^+[\mathrm{A}][\mathrm{E}]$. Similarly, the distance from equilibrium for the half reaction associated with binding/release of A can be expressed as the ratio between the reaction sensitivities of E ($k^+[\mathrm{A}]$) and EA (k^-). This ratio can be simplified into $[\mathrm{A}]/\mathrm{K}_{\mathrm{d,A}}$, where $\mathrm{K}_{\mathrm{d,A}}$ equals k^-/k^+ and represents the dissociation constant for binding/release of A. In cases where there is only one reactant on both sides of the reaction, the half-reaction equilibrium is equivalent to the equilibrium of the reaction itself (since the resulting dynamic ratio is k^+/k^-).

As a specific example, we present a case study on the glucose 6-phosphate isomerase (PGI) enzyme module (Figure 2.1a) from a whole-cell kinetic model of RBC metabolism [54]. An enzyme module describes the individual reaction steps of an enzyme-catalyzed biochemical reaction, and each step is represented by a mass action rate law. Using $PGI1$ reaction as an example, the half reaction of interest is the binding/release of glucose 6-phosphate (G6P) (Figure 2.1b red). The comparison of the sensitivities of PGI (G6Pk^+_{PGI1}) with PGI&G6P ($-k^-_{PGI1}$) (& denotes PGI bound with metabolite G6P) in magnitude is equivalent to the comparison of G6P concentration with $1/\mathrm{K}_{eq,PGI1}$. This comparison effectively results in determining the distance from equilibrium for G6P binding/release half reaction. It is worth noting that the full equilibrium ratio would

A



B



$\frac{dv_{PGI1}}{dPGI} = \text{G6P}\, k^{+}_{PGI1}$ $\quad \frac{dv_{PGI1}}{dPGI\&\text{G6P}} = -k^{-}_{PGI1}$

C

Stoichiometric matrix **S**

	*PGI*1	*PGI*2	*PGI*3
PGI	-1	0	1
PGI&F6P	0	1	-1
F6P	0	0	1
G6P	-1	0	0
PGI&G6P	1	-1	0

Symbolic Gradient matrix **G**

	PGI	*PGI*&F6P	F6P	G6P	*PGI*&G6P
*PGI*1	G6P k^{+}_{PGI1}	0	0	*PGI* k^{+}_{PGI1}	$-k^{-}_{PGI1}$
*PGI*2	0	$-k^{-}_{PGI2}$	0	0	k^{+}_{PGI2}
*PGI*3	-F6P k^{-}_{PGI3}	k^{+}_{PGI3}	-*PGI* k^{-}_{PGI3}	0	0

D

Symbolic concentration Jacobian matrix **J**

	PGI	*PGI*&F6P	F6P	G6P	*PGI*&G6P
PGI	- G6P k^{+}_{PGI1} - F6P k^{-}_{PGI3}	k^{+}_{PGI3}	- *PGI* k^{-}_{PGI3}	- *PGI* k^{+}_{PGI1}	k^{-}_{PGI1}
PGI&F6P	F6P k^{-}_{PGI3}	$-k^{+}_{PGI3}-k^{-}_{PGI2}$	*PGI* k^{-}_{PGI3}	0	k^{+}_{PGI2}
F6P	- F6P k^{-}_{PGI3}	k^{+}_{PGI3}	- *PGI* k^{-}_{PGI3}	0	0
G6P	- G6P k^{+}_{PGI1}	0	0	- *PGI* k^{+}_{PGI1}	k^{-}_{PGI1}
PGI&G6P	G6P k^{+}_{PGI1}	k^{-}_{PGI2}	0	*PGI* k^{+}_{PGI1}	$-k^{-}_{PGI1}-k^{+}_{PGI2}$
	4.55×10^{6}	3.15×10^{7}	0	6.19×10^{3}	-8.22×10^{7} -9.39×10^{7}

Figure 2.1: *PGI* enzyme module and its associated matrices. a) A schematic diagram of individual reaction steps associated with *PGI* enzyme module and its stoichiometric matrix. The enzyme form *PGI* is in italic. We use an "&" notation to denote that the enzyme form is bound with metabolite(s). b) Graphical representation of the concept of half reaction. Here we demonstrate the half reaction associated with the binding/release process of G6P, which is held constant. To determine the equilibrium state of this half reaction, we are comparing the sensitivities associated with *PGI* (G6Pk^{+}_{PGI1}) and *PGI*&G6P ($-k^{-}_{PGI1}$). c) The gradient matrix of the *PGI* enzyme module. The gradient matrix ($= d\mathbf{v}/d\mathbf{x}$) is obtained from linearization of the reaction rates and represents reaction sensitivities to metabolite concentrations. d) The cause of diagonal dominance demonstrated through the symbolic concentration Jacobian matrix of the *PGI* enzyme module. Using row 5 as a case study, we observe that, in the case of mass action rate law, diagonal dominance is determined by the distance from half-reaction equilibrium for individual half-reactions. When comparing the terms associated with *PGI1* reaction between diagonal and off-diagonal positions, we are comparing the sensitivity of G6P (*PGI*k^{+}_{PGI1}) and sensitivity of *PGI* (G6Pk^{+}_{PGI1}) with that of *PGI*&G6P ($-k^{-}_{PGI1}$). In the current case, we can see that the absolute sum of off-diagonal elements in a column is always at least as large as the absolute diagonal element, meaning that diagonal dominance does not occur across columns.

include the enzyme forms that have been removed by differentiation and therefore do not influence the above comparison; thus, the distinct definition of a half-reaction equilibrium ratio is helpful.

As we will show later, the sparsity of a mode is dependent on the distance from equilibrium of connected half reactions defined by these sensitivities in $\mathbf{G}$. Half reactions that are far from equilibrium result in simple mode structures while those near equilibrium together form complex modes.

2.3.2 Diagonal dominance and the Gershgorin circle theorem applied to the Jacobian matrix

Now that basic definitions have been established, we can begin to examine the sparsity structure of the dynamic modes of kinetic models of metabolism. The modes are defined by

$$m_i =< u_i|\mathbf{x} > \tag{2.5}$$

where u_i is the left eigenvector and $\mathbf{x}$ is the steady state concentration vector. The bracket notations refer to the inner product of two vectors. The relative magnitudes of the elements of u_i determine the effective sparsity of a mode when low contributing elements are truncated. However, since the modes are calculated through a numerical algorithm, it is usually not straightforward to link a mode composition to particular elements of the Jacobian matrix, unless the Jacobian matrix has certain structural properties. One such property is diagonal dominance of the rows or columns of the Jacobian, which occurs when the magnitude of a diagonal element is greater than the sum of the magnitudes of off-diagonal elements in the same row (in the case of row dominance)

$$|\mathbf{J}_{ii}| > \sum_{k \neq i} |\mathbf{J}_{ik}| \tag{2.6}$$

or column (column dominance), see Figure 2.2a. We focus on row dominance in this work, as column dominance does not occur in the concentration Jacobian matrix due to the structure of the mass action rate law, as demonstrated in Figure 2.1d.

The degree of diagonal dominance of a row number i can be quantitatively described by a metric we term the diagonal fraction, defined as the ratio between the sum of the absolute values of off-diagonal elements and the absolute value of the diagonal element:

$$f_i = \frac{\sum_{k \neq i} |\mathbf{J}_{ik}|}{|\mathbf{J}_{ii}|} \tag{2.7}$$

Diagonal dominance of a row of the Jacobian matrix gives information about its corresponding eigenvalue. This relationship is made clear using Gershgorins circle theorem [57], which constrains an eigenvalue to be within a certain radius, based on the sum of the off-diagonal elements in a particular row/column, of the diagonal element. The theorem is particularly useful in confining eigenvalues within Gershgorin circles when strong diagonal dominance (a small f_i value) occurs, as the eigenvalue will be close to the diagonal element of the dominant row.

2.3.3 Diagonal dominance in the Jacobian matrix underlies simple mode structures

To investigate the occurrence and impact of diagonal dominance in a real metabolic network, we use the RBC kinetic model mentioned earlier to draw the Gershgorin circles and the eigenvalues from **J** (Figure 2.2b along x-axis). As highlighted in Figure 2.2b, for the selected set

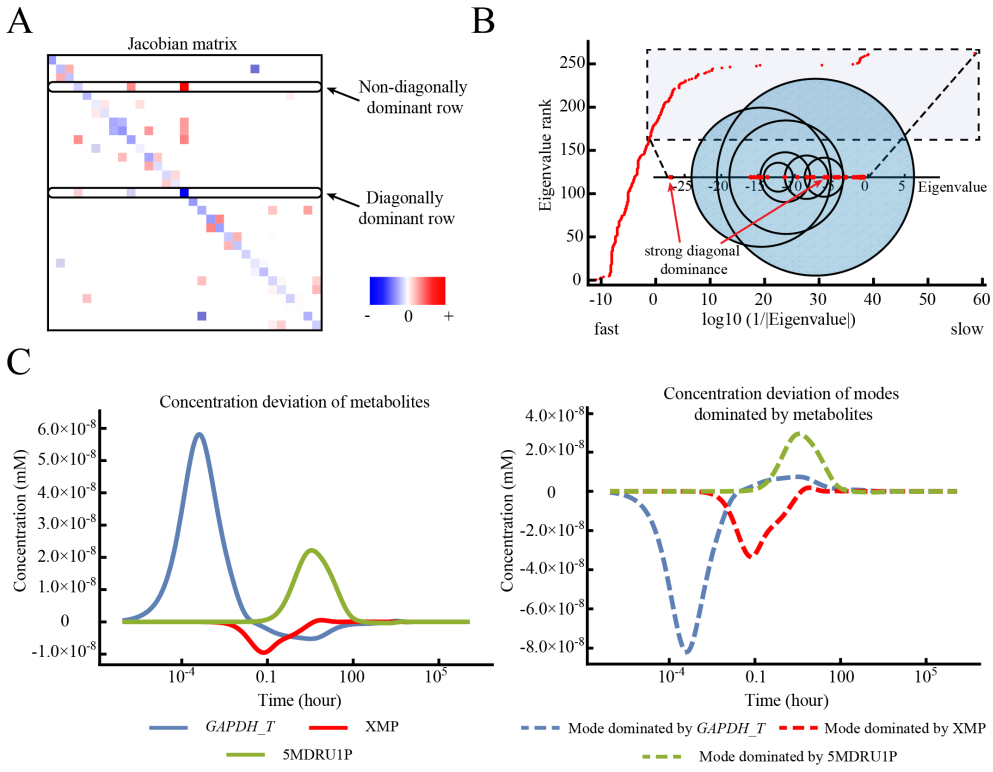


Figure 2.2: Diagonal dominance in the Jacobian matrix explains simple mode structures and corresponding eigenvalues with the help of Gershgorin circle theorem. a) Example Jacobian matrix of the RBC metabolic network [54] with different degrees of diagonal dominance. The Jacobian matrix of the metabolic network has a sparse structure, and the diagonal elements of the matrix are always negative due to the structure of the rate laws used. The matrix was extracted from the full concentration Jacobian matrix for illustrative purposes. b) The entire set of eigenvalues of the Jacobian matrix is shown in the larger plot, with x-axis denoting the inverse of absolute eigenvalues at the log10 scale. In the inset, selected Gershgorin circles of the Jacobian matrix with circle centers ranging from -27 to -5 are shown for illustrative purposes. Eigenvalues greater than -27 are drawn together with the selected circles. The Gershgorin circles from rows with strong diagonal dominance have centers at -26.2 and -5.26 as shown, and the eigenvalues inside are -26.3 and -5.33. All eigenvalues are negative as the system is dynamically stable. The imaginary components of the eigenvalues are small and therefore are neglected. c) The dynamic response of *GAPDH*_T, XMP, 5MDRU1P, compared to the respective modes dominated by these metabolites/enzymes, under an ATP hydrolysis perturbation. The dynamics of the mode dominated by a single metabolite coincide with the dynamics of that metabolite. These modes occur at fast, intermediate and slow timescales, showing that diagonal dominance can occur at any time as long as the structural properties of the Jacobian matrix allow.

of Gershgorin circles, there are two cases where the circle resulting from the strongly diagonally dominant row is very constrained and a unique eigenvalue falls inside the circle. In those cases, the eigenvalue is very closely approximated by the diagonal element.

In addition to providing information about the eigenvalues, diagonal dominance in **J** also causes a simple sparsity structure within modes corresponding to these eigenvalues. When a row has strong diagonal dominance ($f < 0.1$), the diagonal metabolite usually is the only significant non-zero element in the mode. For example, the enzyme form $GAPDH\_T$ (glyceraldehyde 3-phosphate dehydrogenase at tense state) has a very small diagonal fraction value, and is the only element in the mode at its corresponding time scale. The underlying reaction that causes its dominance is the transition step from enzyme form $GAPDH$ at relaxed state to tense state $GAPDH \leftrightarrow GAPDH\_T$, where the sensitivity of $GAPDH\_T$ $(-k^{-}_{GAPDH\_transition\_step})$ contributes the most to its diagonal element in **J** . When a mode contains only the diagonally dominant metabolite, the dynamic motion of the mode drives that metabolite back to its reference state on a timescale determined by the eigenvalue. For example, under ATP hydrolysis perturbation, the dynamics of $GAPDH\_T$ match closely with the dynamics of the mode in which $GAPDH\_T$ is dominant (Figure 2.2c). When diagonal dominance becomes weaker ($f > 0.1$), the diagonally dominant metabolite shares modes with other metabolites. In those cases, the ratio between those metabolites in the mode is similar to that in the diagonally dominant row of the Jacobian matrix. Overall, in the RBC metabolic model used in this work, the structure of 38 out of 244 (15.6%) concentration modes can be explained by diagonally dominant metabolites.

As another effect of diagonal dominance, there exists an important relationship between diagonal dominance in **J** and system dynamic stability, which is characterized by the sign of eigenvalues of **J** in that any positive eigenvalues result in the steady state being unstable. Negative

diagonal elements in **J** strongly support system stability, and this effect is further magnified by diagonal dominance.

2.3.4 Dependence of diagonal dominance on the parameters of the metabolic network

Having established that diagonal dominance is an important property of kinetic models of metabolism for real networks, we now describe the origin of diagonal dominance in terms of the kinetic and physiological parameters of the system. To understand how diagonal dominance in **J** is manifested through reaction properties, we can examine the association of elements between **J** and **G** . We can see that for each diagonally dominant metabolite (diagonal fraction < 1), its diagonal element in **J** can be matched with a specific reaction sensitivity element for that metabolite similar in absolute value in **G** . Such an element is the largest in absolute value for the flux-concentration derivatives ($d\mathbf{v}/d\mathbf{x}$) associated with that metabolite. Therefore, a single term in **G** dominates the resulting diagonal term in **J** . Furthermore, single reaction sensitivities in the form of $d\mathbf{v}/d\mathbf{x}$ in **G** can determine the dynamic behavior of the system in terms of the resulting eigenvalues when diagonal dominance occurs. This correspondence can also be extended to metabolites with non-diagonal dominance, indicating the interpretable connection between **J** and **G** .

As a case study, we examine the cause of diagonal dominance in **J** of the *PGI* enzyme module. We see that, in the enzyme module, diagonal dominance in **J** is determined by a particular half-reaction equilibrium ratio, as defined above. We demonstrate this by examining the enzyme form *PGI*&G6P in the 5^{th} row of **J** (Figure 2.1d). The diagonal term of **J** for *PGI*&G6P shows that the enzyme form is associated with two reactions, *PGI*1 and *PGI*2.

Specifically, reaction $PGI1$ can be split into two half reactions, related to G6P binding/release and PGI binding/release processes. The comparison of the diagonal term ($-k^-_{PGI1}$) with the off-diagonal terms (G6Pk^+_{PGI1} and $PGIk^+_{PGI1}$) related to $PGI1$ reaction is effectively examining the associated half-reaction equilibrium ratios, which are G6P/$\mathrm{K_{d},}_{PGI1}$ and $PGI/\mathrm{K_{d},}_{PGI1}$($\mathrm{K_{d},}_{PGI1} = k^-_{PGI1}/k^+_{PGI1}$). The term G6P$k^+_{PGI1}$ is smaller than $-k^-_{PGI1}$ on the diagonal position in magnitude while $PGIk^+_{PGI1}$ term is negligible compared to $-k^-_{PGI1}$, due to the small concentration of the PGI enzyme form. For reaction $PGI2$, the term k^+_{PGI2} at the diagonal position is much greater than k^-_{PGI2}, with the consumption of PGI&G6P favored. As a result, the diagonal term of **J** for PGI&G6P is greater than the sum of off-diagonal terms in the same row, resulting in diagonal dominance.

To summarize, diagonal dominance can be understood based on the distance from half-reaction equilibrium, by comparing metabolite concentrations to the reaction equilibrium constant. In the case of a single reactant on each side of the reaction, the equilibrium constant alone affects the degree of diagonal dominance. This type of analysis can also be applied to other enzyme forms in **J** .

2.3.5 Power iteration connects mode structure to the structure of the Jacobian matrix

Diagonal dominance explains the structure of most of the highly sparse modes, but cannot address mode structures that are complicated by more than one or two significant elements. We now show how more complicated mode structures form mathematically from specific elements of the Jacobian matrix. We demonstrate that examining the modes of the Jacobian matrix from the perspective of the matrix power iteration algorithm is illustrative in describing how complicated

mode structures arise.

Matrix power iteration is an algorithm to calculate the leading eigenvalue and eigenvector of a matrix (or left eigenvectors in the case of the modes) [58]. In the power iteration algorithm, the Jacobian matrix is left multiplied by a random vector (u_i), the resulting vector is normalized, and this process is repeated until the vector converges (Figure 2.3a). If the eigenvalue with the largest magnitude is well separated from the other eigenvalues, the final vector will converge to the corresponding leading eigenvector. The Euclidean norm of $u\mathbf{J}$ in the last iteration will be the associated leading eigenvalue λ, where $u\mathbf{J} = \lambda u$. During the iteration process, the elements of the Jacobian matrix that contribute to the modes will stretch the vector through multiplication in the direction of the leading eigenvector. The advantage of using this algorithm is that when run for a restricted number of iterations, the power iteration algorithm gives a simple approximation of the modes that enables the identification of mode-determining elements of the Jacobian matrix. Given the fact that the Jacobian matrix is sparse, the power iteration algorithm can help us understand eigenvector structure by inspecting how the Jacobian elements stretch the vector to ultimately result in the eigenvector.

To illustrate the process of vectors converging to the leading eigenvector through power iteration, we perform power iteration algorithms on 1000 random starting vectors using the full Jacobian matrix (292×292). We then perform principal component analysis (PCA) on all the iteration vectors (Figure 2.3c). The random starting vectors quickly converge in the dimension of the first principal component (71.2% contribution), representing the eigenvector, and stabilize in the dimension of the rest of components (second principal component shown only, contributing a very minor percentage) after around 10 to 20 iterations.

As a technical detail of the implementation, a limitation of the power iteration algorithm

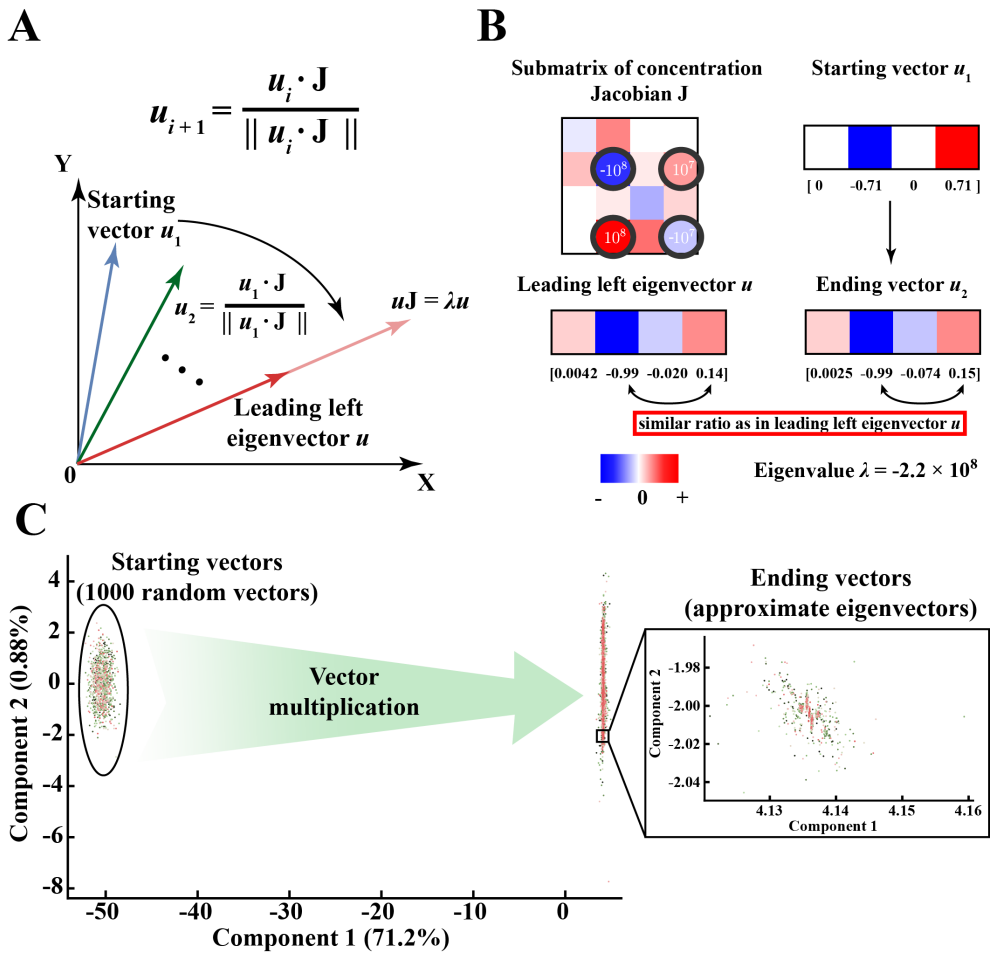


Figure 2.3: The power iteration algorithm demonstrates how complicated dynamic structures arise from topologically connected elements of similar magnitude within the Jacobian matrix. a) Power iteration can be used to calculate the dominant left eigenvector of the Jacobian matrix. The left eigenvectors are the modes of the metabolic network. The algorithm left multiplies the Jacobian matrix by a random vector (u_i), normalizes the resulting vector and repeats the process until the vector converges to the eigenvector. b) Topologically connected Jacobian elements of similar magnitude determine complicated eigenvector structure. In this case study, we extracted a submatrix of **J** that corresponds to the nonzero elements of a certain eigenvector, which contains *G6PDH* enzyme forms. The four Jacobian elements (also the largest) that are key in determining this eigenvector structure are located in the 2nd and 4th rows, circled in black. c) Principal component analysis on all power iteration vectors starting with 1000 different random vectors. We randomly picked 1000 starting vectors and multiplied them with the full Jacobian matrix (292 $\times$ 292). The starting vector is multiplied through several iterations (10 to 20) until it converges to the eigenvector (the dot product of the ending vector and the eigenvector is no greater than 1.0001 and no less than 0.9999). We then performed principal component analysis on all iteration vectors (including the starting vectors) and plotted each vector in terms of the contribution from the first two principal components. The first principal component corresponds to the leading eigenvector of the Jacobian matrix while the rest of components (less than 1% contribution each, only component 2 shown here) together explain the variation of the vector from the eigenvector.

is that it only calculates the leading eigenvalue and eigenvector. To calculate the next largest eigenvalue and the associated eigenvector, we must modify **J** to eliminate the impact of the previous eigenvector and eigenvalue at each step. Such elimination can be accomplished with the Hotelling deflation method [59], which returns a modified **J** , with the leading eigenvector and eigenvalue removed, that can be used for a new round of eigenvector and eigenvalue calculations using power iteration (see Methods).

2.3.6 A case study on using power iteration to understand complicated mode structure

We now use the power iteration method to demonstrate how the eigenvectors with more complicated structures form in a set of specific numerical examples on the RBC metabolic network. In this section, we show that that the topological connection of elements of similar orders of magnitude in **J** is critical in determining the sparsity structure of the eigenvectors. This similar order of magnitude tends to lie around the eigenvalue (Figure 2.3b).

As a case study, we extract a submatrix of **J** (4×4) corresponding to the positions of nonzero elements (see Methods for cutoff) of a particular eigenvector, which is associated with *G6PDH* enzyme forms of the RBC metabolic network. When **J** is pre-multiplied by a pseudo-random starting row vector, we see that the ending vector matches closely with the actual eigenvector (Figure 2.3b). It is clear upon inspection that the largest values in the submatrix are also the largest values in the mode. The four key **J** elements (also largest in the submatrix) determining eigenvector formation are located in the 2^{nd} and 4^{th} rows (Figure 2.3b black circles). These rows both have similar structures to the eigenvector, where the ratio between the 2^{nd} and 4^{th} elements in the row is the same as that in the eigenvector. This shows that the matrix

structure is reflected in the eigenvector structure.

To explore how the 2nd and 4th rows both contribute to eigenvector formation, we can perturb the starting vector such that it interacts with these rows specifically, such as (0, -1, 0, 0) and (0, 0, 0, 1), to examine each rows effect individually. As a result, starting from either vector leads to a structure similar to the original eigenvector. Thus, it seems that both rows have similar contributions to the structure of the eigenvector in this case, although their magnitude is different. Together, the four elements in those two rows (Figure 2.3b black circles) form a topologically connected structure and interact with each other symmetrically to determine the eigenvector structure. The other large element at position (4, 3) is not involved with this symmetric interaction and thus has a smaller contribution to eigenvector formation.

Next, to demonstrate the interplay of the submatrix elements, we show how modifying the four key elements of the sub-matrix changes the eigenvector. First, to examine the impact of the largest diagonal element in the submatrix at position (2, 2), we modify the diagonal element at position (4, 4) to have the same value as the element at (2, 2) (Figure 2.4b). The resulting vector has a different ratio between its elements compared to the original J eigenvector, with a larger value in the 4th element, reflecting the larger value in the (4, 4) position of the submatrix. We then further change the off-diagonal element of **J** at (2, 4) to be the same as the element at (4, 2) to create a more symmetric structure (Figure 2.4c). The resulting vector now has the same value on both the 2nd and 4th positions, showing that the off-diagonal elements modify the weightings on the eigenvector, and a fully symmetric Jacobian structure will result in an equally weighted eigenvector structure. These perturbations show that how the relative values of the dominant elements in a submatrix are clearly reflected in the corresponding mode structure.

The power iteration algorithm is a useful tool to analytically understand the structure

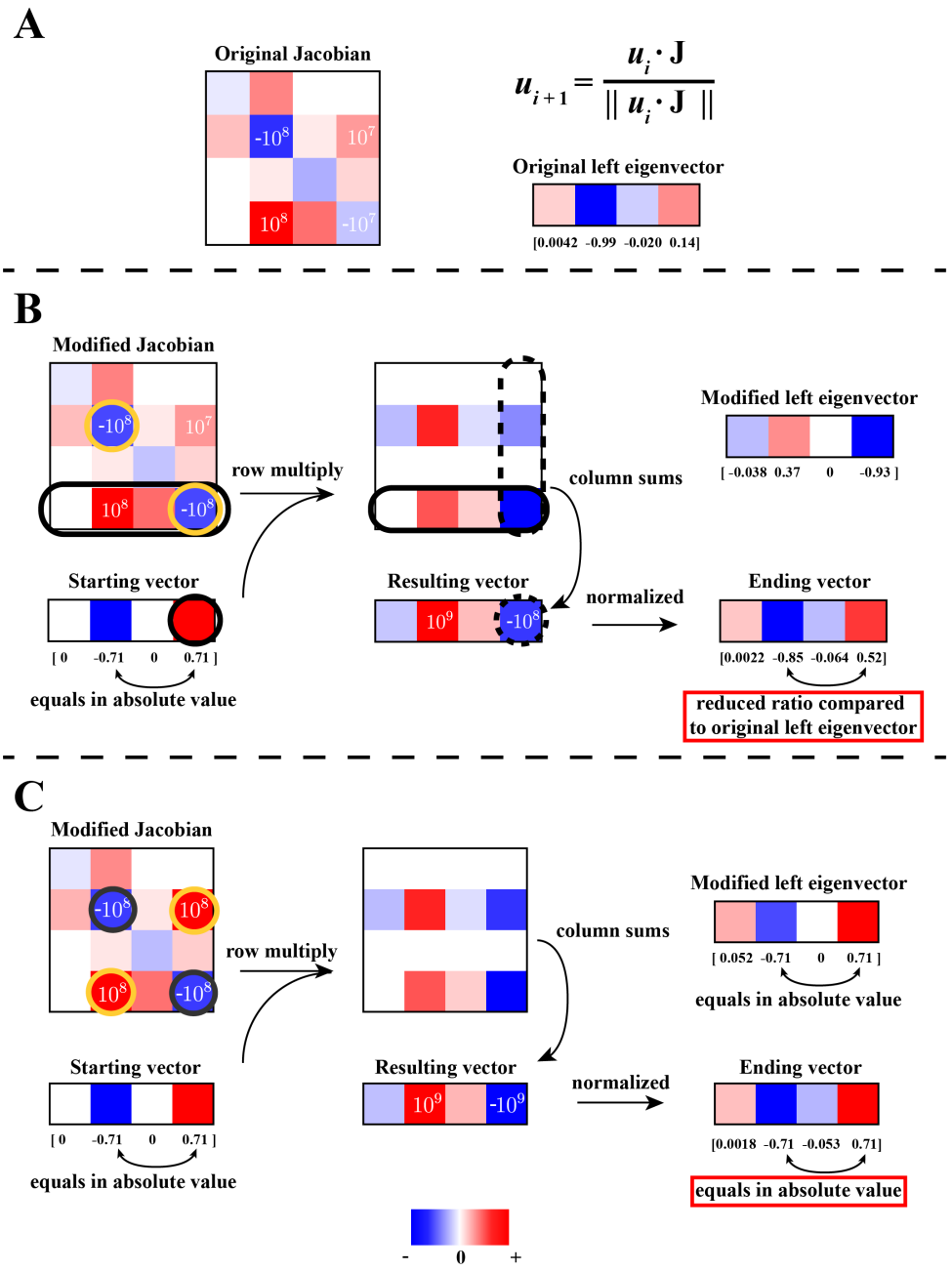


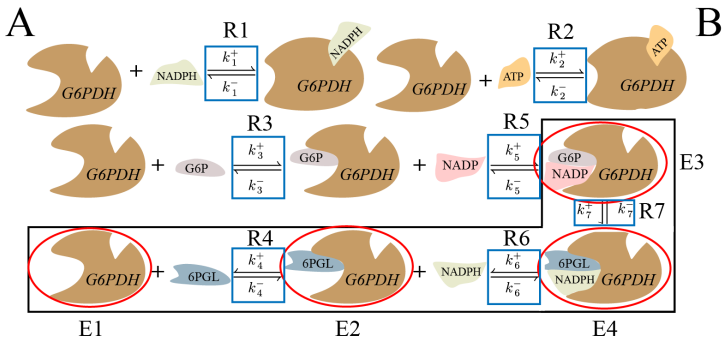
Figure 2.4: Analysis of complicated mode structure through power iteration with modified Jacobian matrix. a) The original Jacobian matrix and its leading left eigenvector. b) Starting vector multiplied with the modified Jacobian matrix. We modified the Jacobian element at position (4, 4) to be the same value as the element at position (2, 2). c) Starting vector multiplied with a different modified Jacobian matrix. We further changed the modified Jacobian matrix in panel A to create a more symmetric structure, where the element at position (2, 4) is same as the element at position (4, 2).

of complicated eigenvectors of a real system. We have demonstrated that the modes form from a network of topologically connected values of similar magnitude in the Jacobian matrix, and the relative ratio between these values influences the structure of the eigenvector. These trends, where an eigenvector can be linked to particular topologically-connected elements of **J** of similar magnitude, are generally applicable beyond this case study. The Jacobian modifications demonstrate that the eigenvector of the matrix can be altered in a predictable manner by changing either diagonal or off-diagonal Jacobian elements along the same order of magnitude.

2.3.7 Complicated mode structure arises from connected reactions with similar dynamic sensitivities in G

Power iteration helps to show numerically how complicated modes arise due to particular structures in **J** . For metabolic networks constructed with mass action rate laws, these numerical values have clear biological interpretations. Next, we describe the origin of complicated mode structure in terms of specific metabolite and reaction properties of the system. The goal of this section is to obtain a biochemical interpretation of the numerical results obtained in the previous section.

We use the same case study presented in the previous section, regarding the mode and submatrix of **J** for *G6PDH* enzyme forms. The mode contains four *G6PDH* enzyme forms (red circles in Figure 2.5a), with *G6PDH*&6PGL and *G6PDH*&NADPH&6PGL being the most dominant elements. The mode structure is largely determined by the sensitivities of reaction 6 in **G** (k_6^+, NADPHk_6^-) (Figure 2.5c). This reaction releases NAPDH and its elements in **G** dominate the topologically connected **J** elements at positions (2,2), (2,4), (4,2) and (4,4) (Figure 2.5d). The two most dominant mode elements mentioned above are associated with reaction



B

Stoichiometric matrix **S**

	R1	R2	R3	R4	R5	R6	R7
E1	-1	-1	-1	1	0	0	0
E2	0	0	0	-1	0	1	0
E3	0	0	0	0	1	0	-1
E4	0	0	0	0	0	-1	1

C

Symbolic Gradient matrix **G**

	E1	E2	E3	E4
R1	NADPH k_1^+	0	0	0
R2	ATP k_2^+	0	0	0
R3	G6P k_3^+	0	0	0
R4	-6PGL k_4^-	k_4^+	0	0
R5	0	0	$-k_5^-$	0
R6	0	- NADPH k_6^-	0	k_6^+
R7	0	0	k_7^+	$-k_7^-$

Numerical Gradient matrix **G**

	E1	E2	E3	E4
R1	65796.60	0	0	0
R2	1.79×10^6	0	0	0
R3	26522.50	0	0	0
R4	-1.72×10^6	5.00×10^7	0	0
R5	0	0	-1.30×10^6	0
R6	0	-3.13×10^8	0	5.00×10^7
R7	0	0	5.05×10^7	-623.04

D

Symbolic Jacobian matrix **J**

	E1	E2	E3	E4
E1	X	k_4^+	0	0
E2	6PGL k_4^-	$-k_4^+$ - NADPH k_6^-	0	k_6^+
E3	0	0	$-k_5^-$ - k_7^+	k_7^-
E4	0	NADPH k_6^-	k_7^+	- k_6^+ - k_7^-

Numerical Jacobian matrix **J**

	E1	E2	E3	E4
E1	-3.61×10^6	5.00×10^7	0	0
E2	1.72×10^6	-3.63×10^8	0	5.00×10^7
E3	0	0	-5.18×10^7	623.04
E4	0	3.13×10^8	5.05×10^7	-5.00×10^7

X = - NADPH k_1^+ - ATP k_2^+ - G6P k_3^+ - 6PGL k_4^-

E1= *G6PDH* E2= *G6PDH*&6PGL E3= *G6PDH*&NADP&G6P E4= *G6PDH*&NADPH&6PGL

Figure 2.5: The origin of complicated mode structure associated with *G6PDH* enzyme forms demonstrated through the associated matrices. a) The reaction steps for the biochemical reaction catalyzed by *G6PDH* enzyme. b) The stoichiometric matrix **S** for the four enzyme forms in the mode and their associated reactions. The **S** matrix describes the network topology of the enzyme forms and determines how they interact in the Jacobian matrix. c) The symbolic and numerical gradient matrix **G** for the four enzyme forms in the mode and their associated reactions. The key reaction sensitivities determining the two largest elements in the mode are associated with reaction 6 and its corresponding enzyme forms. d) The symbolic and numerical Jacobian matrix **J** for the four enzyme forms in the mode. We found that the elements of reaction 6 in **G** dominate the topologically connected Jacobian elements that determine the mode structure. These elements are located at positions (2,2), (2,4), (4,2) and (4,4). Reaction 6 is connected to reaction 4 and 7, whose reaction sensitivities are much smaller in magnitude compared to that of reaction 6, resulting in very small coefficient for their associated elements in the mode (*G6PDH* and *G6PDH*&NADP&G6P).

6. Their corresponding **J** elements contain k_6^+ and NADPHk_6^-, which are close numerically, meaning that NADPH concentration is similar to the equilibrium constant of the half reaction for NAPDH binding/release, where the term 'half reaction' is used as defined above. The ratio between NADPHk_6^- and k_6^+ (NADPH/$\mathrm{K_{d,6}}$, where $\mathrm{K_{d,6}} = \mathrm{K_{eq,6}}$) defines a half-reaction equilibrium ratio that is the key in determining the eigenvector structure. If NADPH concentration is higher, reaction 6 will become more sensitive to the concentration of the released form *G6PDH*&6PGL, compared to that of bound form *G6PDH*&NADPH&6PGL. This change will cause enzyme form *G6PDH*&6PGL to become more dominant in the mode, due to its greater diagonal dominance in **J**. Additionally, reaction 7 has the same order of magnitude sensitivity in the forward direction (k_7^+) as reaction 6, but has a much smaller sensitivity when interacting with *G6PDH*&NADP&G6P in the reverse direction, thus resulting in a much smaller contribution to this enzyme form in the mode. Finally, the unbound *G6PDH* enzyme form, although topologically connected to other enzyme forms through reaction 4, is not prominently featured in the mode, since its sensitivities in **G** are at a smaller order of magnitude.

Overall, only a few reaction sensitivities in **G** contribute to the mode structure in this case study, thus allowing us to determine the specific reactions that control the dynamics of the mode. For significant elements in the complicated mode structure, the associated half-reaction equilibrium constant is close to the metabolite concentration, thus creating dynamic interplay between multiple elements in the reactions. On the other hand, in the case of simple mode structure governed by diagonal dominance, the half-reaction equilibrium ratio associated with the diagonal metabolite is usually far from equilibrium. The analysis approach presented exploits the fact that dynamic features in **J** are an integration of the features in **S** and **G** , thus allowing us to understand modal structure in terms of both reaction sensitivities in **G** and network topology

in **S** .

2.3.8 Power iteration converges to eigenvector subspaces when eigenvalues are similar in magnitude

As an important technical aside, we note that the power iteration procedure works well when the eigenvalue is much larger in magnitude than the others; however, special behaviors arise when eigenvalues do not separate well. Specifically, when we reach modes where eigenvalues are close in magnitude, the power iteration algorithm converges to different ending vectors depending on the starting vectors. In this case, the starting vector is influenced by multiple eigenvectors comprising a subspace of dynamics active around this time scale, making the ending vector difficult to predict. The ending vectors overlap significantly with an "eigenvector subspace" (Figure 2.6a), as these vectors are influenced by multiple eigenvectors simultaneously. Also, the approximated eigenvalues overlap significantly with the actual eigenvalue cluster (Figure 2.6b), showing that the approximated eigenvalues settle in the range of the set of similarly leading eigenvalues. Overall, this analysis demonstrates how multiple eigenvectors influence dynamic response for time scales that are associated with multiple eigenvalues at similar magnitude.

2.4 Discussion

In this study, we developed an understanding of how the sparsity structures of the dynamic modes of kinetic models of metabolism are linked to specific properties of mass action reaction rate laws. 1) We showed that the diagonal dominance in rows of the Jacobian matrix is a common occurrence due to the order-of-magnitude scaling of kinetic constants, metabolite concentrations, and reaction fluxes. This diagonal dominance results in simple mode structures

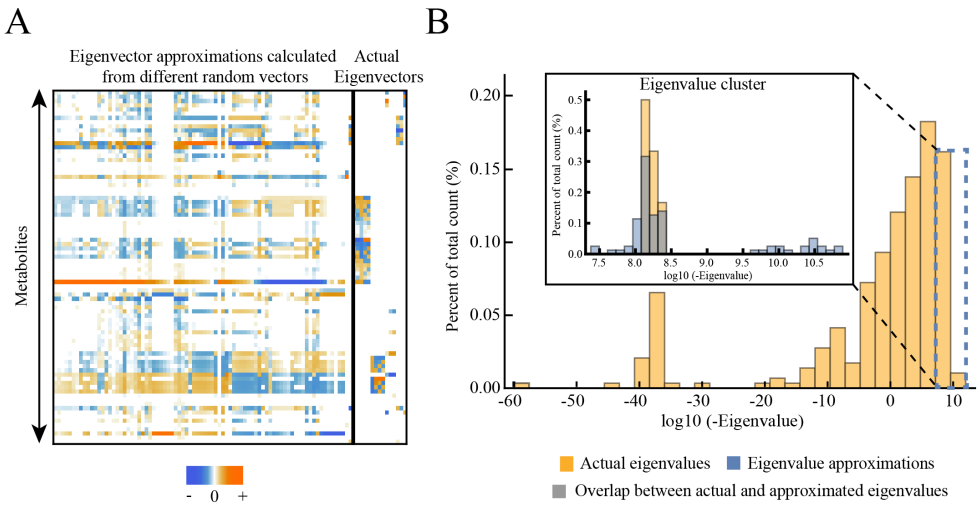


Figure 2.6: Eigenvalue and eigenvector approximations calculated from power iteration in cases where eigenvalues do not separate well. We selected a cluster of close eigenvalues (with a time scale around 0.016 milliseconds), reduced **J** using Hotelling's deflation method until this time scale was reached (see Methods), and calculated approximated eigenvalues and eigenvectors using power iteration with different starting vectors. a) Eigenvector approximations calculated during power iteration from different starting vectors, compared to the actual eigenvectors with eigenvalues in the selected range. We calculated the approximated 100 eigenvectors from 100 different random vectors with 100 iterations each and obtained vectors that are linearly independent with each other (see Methods). The left part of the matrix shown is the eigenvector approximations while the right part of the matrix shown is the actual eigenvectors, separately by the black bold vertical line. We found that the subspace formed by eigenvector approximations overlaps significantly with the actual eigenvector subspace. b) The selected eigenvalue cluster is compared to the eigenvalue approximations calculated from power iteration. The selected eigenvalues and eigenvalue approximations are shown in the inset plot. We obtained the eigenvalue approximations from the same set of power iterations performed in panel A. The cluster of eigenvalue approximations overlaps significantly with the cluster of actual eigenvalues, showing that the eigenvalue approximations settle in the range of the set of similarly dominant eigenvalues.

where single metabolites relax back to their references states driven by particular eigenvalues. 2) For more complicated mode structures, we used the power iteration algorithm to show that these complicated mode structures form from topologically connected values of similar orders of magnitude in the Jacobian matrix. 3) We showed that a key feature underlying mode structure is the reaction sensitivities in the gradient matrix $\mathbf{G}$, which can be interpreted as the distance from equilibrium of half reactions defined by linearization of bilinear mass action equations.

Diagonal dominance of the Jacobian matrix as described by Gershgorin circle theorem gives information about certain eigenvalues. This property results in simple mode structures, which can occur on time scales that span different orders of magnitude. A simple structure dominated by a single element indicates that the concentration variable relaxes to its reference state after its characteristic timescale and does not interact with others on this timescale. Thus, if rows of the Jacobian are diagonally dominated, there are fewer dynamic connections in the resulting modes, since these modes will have few nonzero elements. As these non-zero elements will correspond to the diagonally dominant metabolites, the dynamics on those timescales are 'local' or heavily influenced by local equilibria. The degree of diagonal dominance that we observe in a real metabolic Jacobian matrix indicates that these local dynamics are prevalent, and thus metabolism has a relatively disconnected dynamic structure on many timescales. This modular structure should simplify the challenge of predicting the dynamic behavior of the entire system. We note that the core theorem used in this analysis, Gershgorin circle theorem, is well-known in classical engineering applications, and also has previously been applied to the Jacobian matrix of metabolic networks to analyze system stability [60, 61].

We have shown that topologically connected elements of the Jacobian matrix at similar magnitude underlie complex mode structures. Here we used the power iteration algorithm to

demonstrate how eigenvectors arise from certain elements of the Jacobian matrix. The power iteration algorithm gives a sparse approximation of the modes that enables the identification of mode-determining elements of the Jacobian matrix. This contrasts to the more standard eigenvector calculation algorithms such as QR decomposition. While other algorithms also yield the eigenvalues and eigenvectors, often in a numerically more efficient manner, our goal was to understand how the elements of the Jacobian determine the eigenvectors. For this purpose, we found that the power iteration algorithm is well-suited to suits our needs. Using power iteration, it is possible to observe how particular elements of the Jacobian matrix influence a random vector and 'move' it in the direction of the eigenvector. This process is how we connect the structure of the Jacobian matrix to the structure of its modes, i.e. the left eigenvectors of the Jacobian matrix. Examining key Jacobian elements that determine eigenvector structure shows that they originate from a few reaction sensitivities of topologically connected reactions. These reaction sensitivities are at different orders of magnitude, resulting in well-separated dynamics for the metabolites/enzyme forms involved. In a physiologically relevant perturbation, these fast dynamics are not likely to be excited, leaving the slow ones to be main interest of study.

It would be remiss in any work on the linearized dynamics of metabolic networks to fail to mention the relation of the work to the foundational body of theory in Metabolic Control Analysis (MCA) [53]. The gradient matrix $\mathbf{G}$ ($d\mathbf{v}/d\mathbf{x}$) that we use to calculate the Jacobian matrix $\mathbf{J}$ is the same matrix that appears in MCA as the unscaled elasticity matrix [56]. However, the majority of MCA relationships involve the use of scaled matrices, the properties of which we have not yet examined in the context of the dynamic modes of the system. Additionally, frequent questions arising in MCA include the control and parameter sensitivity of the system fluxes. As they are rooted in the same matrices and dynamic properties of the reactions, it is likely that

the modal structure of the system is intricately connected to the local control properties of the system.

When examining the origin of mode structure, we have introduced a concept of a half reaction, which involves only a subset of the substrates and products of a particular reaction that dynamically respond on a particular timescale. We showed that the distance from equilibrium of topologically-connected half reactions is a determinant of the complexity of the mode structure. The half reaction definition arises from linearization of the mass balance equation, where certain reactant/product term has been removed due to differentiation. In a bilinear enzymatic reaction, the reaction sensitivities associated with the substrates/products are often at different orders of magnitude, resulting in half of the reaction responds at a particular time scale while the other half relaxes. This phenomenon is a key feature for the bilinear kinetics occurring in metabolic networks.

2.5 Methods

2.5.1 Software

All work was done in Mathematica 10. We used a package called the MASS Toolbox (https://github.com/opencobra/MASS-Toolbox) for model simulation and analysis. The models are available in SBML and Mathematica formats and can be found in online Supporting Materials (https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0189880).

2.5.2 Model simulation and perturbation

The model used in this study is a whole-cell kinetic model of red blood cell (RBC) metabolism consisting of 133 mass action reactions with 10 enzyme modules incorporated [54]. An

enzyme module describes the detailed reaction steps of an enzyme-catalyzed reaction, including substrate binding, catalytic conversion, product release and regulatory actions. The 10 enzyme modules are mainly located in glycolysis and the pentose phosphate pathway.

We used measured steady state metabolite concentrations as the starting state of the system before the perturbation. The perturbation used in this study was to simulate ATP hydrolysis in RBC. At time 0, the ATP concentration was decreased by 0.1 mmol/L while ADP and Pi concentrations were increased by 0.1 mmol/L. We then simulated the subsequent concentration and flux changes through numerical integration of the ODE equations. We gave the system enough time (10^6 hours) to regain the steady state concentrations. The dynamic response of a specific metabolite or a combination of metabolites over time was visualized using the plotting functions in MASS Toolbox.

2.5.3 Mode structure interpretation and dominant mode selection

To simplify the mode structure for interpretation, we neglected metabolites whose absolute coefficient values are less than 5% of the maximum absolute coefficient. We found that generally metabolites with small coefficients do not substantially contribute to the dynamic response of the mode, and 5% serves as a useful cutoff value for purposes of analysis.

When selecting modes that can be explained by diagonal dominance alone, we applied the following criteria to both concentration modes and flux modes. When examining a particular mode, we first neglected elements whose absolute coefficient values are less than 5% of the maximum absolute coefficient. If there is only one element left in the mode and it is diagonally dominant, the mode is explained by diagonal dominance. For modes with multiple elements, we selected the mode where its largest coefficient is at least twice as large as the next one and

corresponds to the most diagonally dominant element in the mode.

2.5.4 Power iteration and Hotelling's deflation

Since the modes are left eigenvectors of the Jacobian matrix, we left multiplied the Jacobian matrix by the vector during power iteration. We started with a random vector, obtained a new vector after matrix multiplication and normalized against the Euclidean norm. We kept running this iteration until the length of the ending vector converges. The algorithm is demonstrated as follows,

$$u_{i+1} = \frac{u_i \cdot \mathbf{J}}{\|u_i \cdot \mathbf{J}\|} \tag{2.8}$$

where i is the number of iterations, u_i is the starting vector and u_{i+1} is the ending vector in each iteration.

Since power iteration only calculates the leading eigenvalue and eigenvector of the Jacobian matrix, we used Hotelling's deflation to remove the impact of the leading eigenvector and calculated the next leading eigenvector [59]. The algorithm thus results

$$\mathbf{J}_{t+1} = \mathbf{J}_t - u_t u_t^T \mathbf{J}_t u_t u_t^T \tag{2.9}$$

where $\mathbf{J}_{t+1}$ is the Jacobian matrix after the leading eigenvector u_t of the previous Jacobian matrix $\mathbf{J}_t$ is removed.

In cases where the eigenvalues are clustered together, different starting vectors will result in different eigenvectors at the end of iteration. To compare the approximated eigenvectors from power iteration with the actual eigenvectors, we picked the eigenvalue cluster with time

scale around 0.016 milliseconds and reduced **J** using Hotelling's deflation method until this time scale was reached. We started with 100 random vectors and multiplied them by **J** through 100 iterations, which we found to be large enough for the vector to converge in practical cases. To obtain the set of linearly independent vectors out of the 10^4 vectors, we started with one of the vectors, added another vector (from the 10^4 vectors), and calculated the rank of the matrix formed by the current vector space. We kept adding the vector one at a time for all the ones we calculated. If the matrix rank increases, the added vector is linearly independent with the earlier vectors and will be kept in the final vector set. Otherwise, it will not be included. We also calculated the norms of all vectors during iterations as eigenvalue approximations for comparison with the eigenvalue cluster.

Acknowledgments

This work was funded by The Novo Nordisk Foundation Grant Number NNF10CC1016517.

Chapter 2 in full is a reprint of material published in: **Bin Du***, Daniel C. Zielinski*, Bernhard O. Palsson. 2017. "Topological and kinetic determinants of the modal matrices of dynamic models of metabolism." *PLoS One*, 12(12), e0189880. The dissertation author was the primary author (equally contributing with Daniel Zielinski).

Chapter 3

Estimating Metabolic Equilibrium Constants: Progress and Future Challenges

3.1 Abstract

Reaction equilibrium constants (K_{eq}s) are key parameters that impose thermodynamic constraints on the function of a metabolic network. An important approach for K_{eq} estimation is the group contribution method, which utilizes chemical moiety-based estimates of compound formation energies. In this Opinion, we delineate a number of current challenges with the group contribution method, specifically: (i) problems related to the completeness and quality of data necessary for reliable estimation; and (ii) inadequacies of the method to represent the physical properties of compounds. We then highlight a number of promising approaches to deal with

the limitations of group contribution methods. Further advancements should lead to more accurate prediction of equilibrium constants and a better representation of cellular function under biophysical constraints.

3.2 How Are Free Energies Estimated? The Fundamentals of Group Contribution Theory

Thermodynamics plays an essential role in the function of metabolic networks and puts constraints on cellular functions. A wide variety of thermodynamic analyses of the metabolic network have been explored, including examining pathway choices due to the thermodynamic efficiency of the proteome [62], analyzing the feasibility of network flux states [63, 64], and determining the thermodynamic feasibility of measurements of quantitative metabolite concentrations [28, 65]. Reaction equilibrium constants (K_{eq}s) are important parameters for thermodynamic analysis. The Keq of a reaction can be measured experimentally by adding the active enzyme and the substrates or products to a solution and allowing the reaction to proceed to equilibrium; at which point the substrate and product concentrations are measured. A large number of reactions have been studied in this manner, and the resulting data have been collected in databases such as the National Institute of Standards and Technology (NIST) Thermodynamics of Enzyme-catalyzed Reactions database (TECRdb).

However, K_{eq}s for the majority of biochemical reactions have not been experimentally measured. The existing set of measured K_{eq}s can be used to calculate unmeasured reactions using the standard Gibbs free energies of formation of the compounds ($\Delta_f G^\circ$) instead. The standard Gibbs free energy of reaction ($\Delta_r G^\circ$) and its involved compounds have a simple relationship

$$\Delta_{\mathrm{r}}G^{\circ} = \sum s_i \Delta_{\mathrm{f}}G_i^{\circ} \tag{3.1}$$

where s_i is the stoichiometric coefficient of the corresponding compound. The $\Delta_{\mathrm{f}}G^{\circ}$ of compounds can be inferred indirectly by forming complete reaction cycles and utilizing principles from the First Law of Thermodynamics to set up a system of equations and solve for $\Delta_{\mathrm{f}}G^{\circ}$. This procedure, also known as reactant contribution, has been performed on existing data and tabulations of resulting $\Delta_{\mathrm{f}}G^{\circ}$ data have been assembled [66, 67]. However, $\Delta_{\mathrm{f}}G^{\circ}$ values for the majority of compounds still cannot be inferred from available $\mathrm{K_{eq}}$ data due to lack of sufficient data coverage, making computational estimation necessary.

The primary approach to estimate unknown $\Delta_{\mathrm{f}}G^{\circ}$ in cases where data are insufficient for reactant contribution is through the group contribution method. Group contribution, based on the principle of additivity, approximates the compound energy as the sum of the energies of the chemical moieties (groups) comprising the compound.

$$\Delta_{\mathrm{f}}G^{\circ} = \sum c_i \Delta_{\mathrm{g}}G_i^{\circ} \tag{3.2}$$

Here, $\Delta_{\mathrm{g}}G_i^{\circ}$ is the group energy and c_i is the number of each group present in the compound. Given a set of measured $\mathrm{K_{eq}}$s, $\Delta_{\mathrm{g}}G^{\circ}$ can be estimated by regression and used to predict $\mathrm{K_{eq}}$s of unmeasured reactions, as long as their involved compounds can be decomposed into the same sets of groups. Since the original application of group contribution to estimate $\mathrm{K_{eq}}$s in aqueous solutions [68], a number of developments have been made, including increasing the scope of predictions, handling pH and temperature consistently and globally across reactions, and incorporating the First Law constraints from reactant contribution into estimations [69–72] (Figure 3.1a).

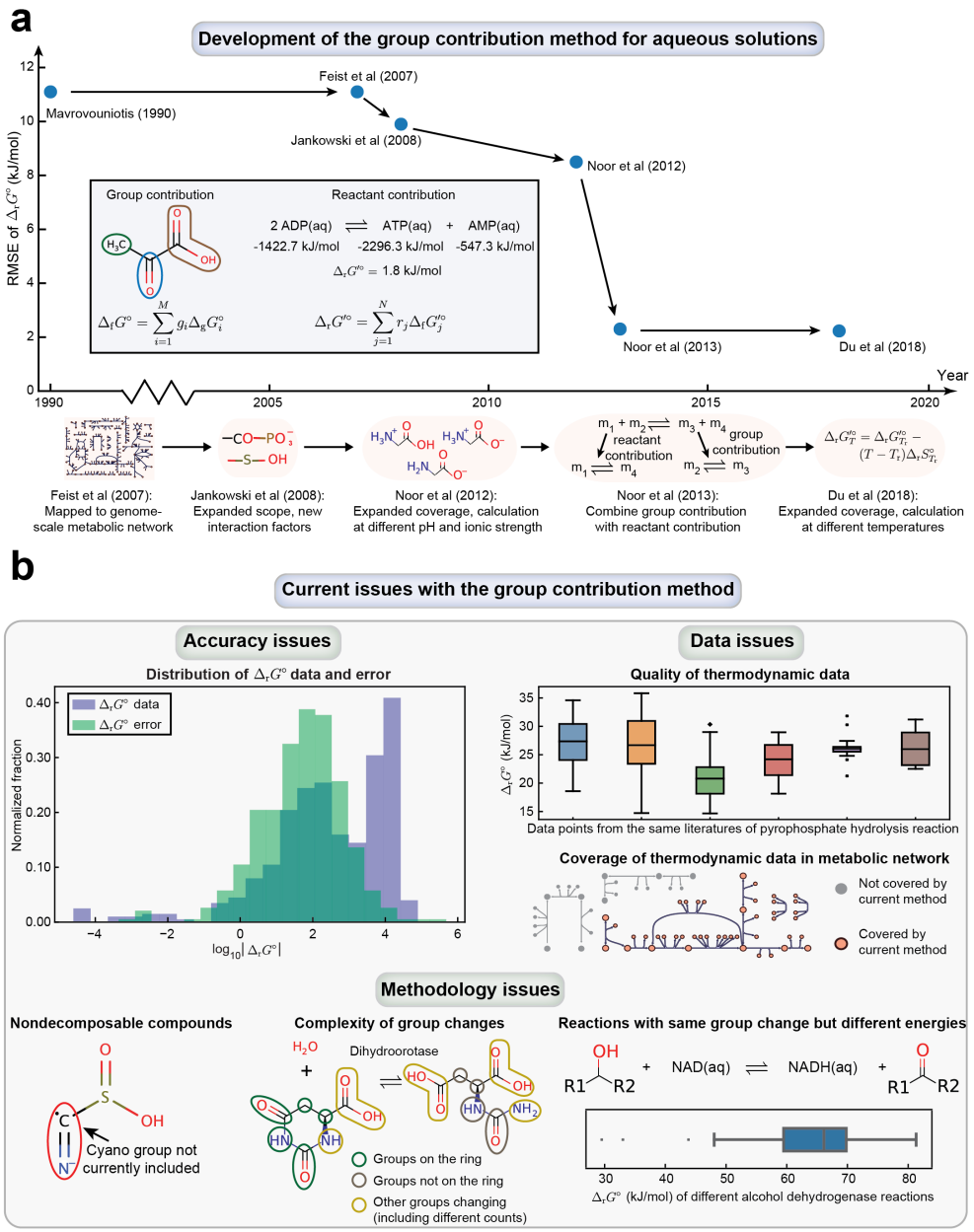


Figure 3.1: Overview of progress and challenges in estimation of reaction equilibrium constants in metabolism. a) Root mean square error (RMSE) of group contribution methods developed for aqueous biochemical reactions over time. The major improvements of each updated method are summarized below the timeline. b) Current issues associated with the group contribution method. They mainly include (i) accuracy issues where the values of $\Delta_r G^\circ$ data and their errors fall in the same order of magnitudes based on the latest method; (ii) data issues including inconsistent data with large variations for the same reaction and limited coverage of data in metabolic networks; and (iii) methodology issues including nondecomposable compounds due to insufficient coverage of compound groups, the complexity of group changes in reactions, and the inability to differentiate reactions with the same group change but different Gibbs reaction energies. See also [68–73].

Despite their extensive use, it has become clear that group contribution methods have a number of substantial challenges that have persisted throughout their development (Figure 3.1b). While accuracy continues to improve (Figure 3.1a; root mean square error of $\Delta_r G^\circ$ now at 2.3 kJ/mol), the error in estimating $\Delta_r G^\circ$ in the worst cases can be as large as 30 kJ/mol, corresponding to a 10^5-fold range in possible K_{eq} values. In this Opinion, we cover several issues that limit the accuracy of group contribution estimates. We further discuss recent progress that may help to overcome these limitations, focusing on the advancements in the last 5 years.

3.3 Key Limitations in Thermodynamic Data Available for Group Contribution Model Training

Assessing the thermodynamic data available to train group contribution estimates, we identify four primary issues related to the data consistency, data depth, data coverage, and variation in experimental conditions.

Data inconsistency refers to the variation in $\Delta_r G^\circ$ of the same reaction from different studies. This inconsistency can become a bottleneck in improving the accuracy of $\Delta_r G^\circ$ estimation. Utilizing TECRdb, we examined the agreement of $\Delta_r G^\circ$ values calculated from measured K_{eq}s of the same reaction across different studies. Ideally, these $\Delta_r G^\circ$ values should be identical. While most reactions have data that agree reasonably well, a number of reactions (5%) have greater than 10-kJ/mol variations in $\Delta_r G^\circ$ (56-fold change in K_{eq}) (Figure 3.2a). A case study on the creatine kinase reaction from different studies (Figure 3.2b) shows that the source of discrepancies in this particular case is the large range of total magnesium and calcium concentrations reported. The interactions of these ions with the compounds in reaction can significantly affect

the measured K_{eq}s and have not yet been accounted for accurately [72]. Additionally, we found that a number of $\Delta_f G^\circ$ values (calculated from experimental $\Delta_r G^\circ$) of the same compounds from different sources [66, 67] vary by more than 10 kJ/mol (Figure 3.2a).

Data depth refers to the issue where particular chemical moieties have poor data coverage, thus having high uncertainties in energy estimates. We thus examined the set of reactions with measured data in the TECRdb and resulting coverage of decomposed compound groups. Groups with a breadth of measured reactions give high confidence in their $\Delta_g G^\circ$ estimates. We found that a quarter of groups within TECRdb reactions only appear in a limited number of reactions (<3) (Figure 3.2c). As demonstrated by a bootstrapping analysis, groups appearing in fewer reactions (<3) tend to have substantially larger uncertainties in $\Delta_g G^\circ$ than those appearing in more reactions (≥ 3) (average median absolute deviation of 4.4 kJ/mol vs 2.6 kJ/mol) (Figure 3.2d). Therefore, when studying a reaction of interest, one should understand the coverage of the involved groups in TECRdb reactions and evaluate the uncertainties accordingly. Encouragingly, these types of uncertainty calculations are now commonly coupled to group contribution efforts [71, 72].

Data coverage is related to the availability of data for real metabolic networks. As a test case, we assessed the coverage of group contribution on the latest *Escherichia coli* genome-scale metabolic reconstruction iML1515 [74] (Figure 3.2e). We found that 78 of 88 groups that change in all iML1515 reactions are covered by TECRdb reactions, with the majority of the covered groups (83%) appearing in more than three unique TECRdb reactions. Assessing the coverage on iML1515 reactions based on metabolic functions, lipid metabolism and cofactor metabolism are clearly the most poorly covered. Compounding this problem, when looking at measured data in TECRdb directly, lipid metabolism is almost entirely unmeasured, indicating that $\Delta_g G^\circ$

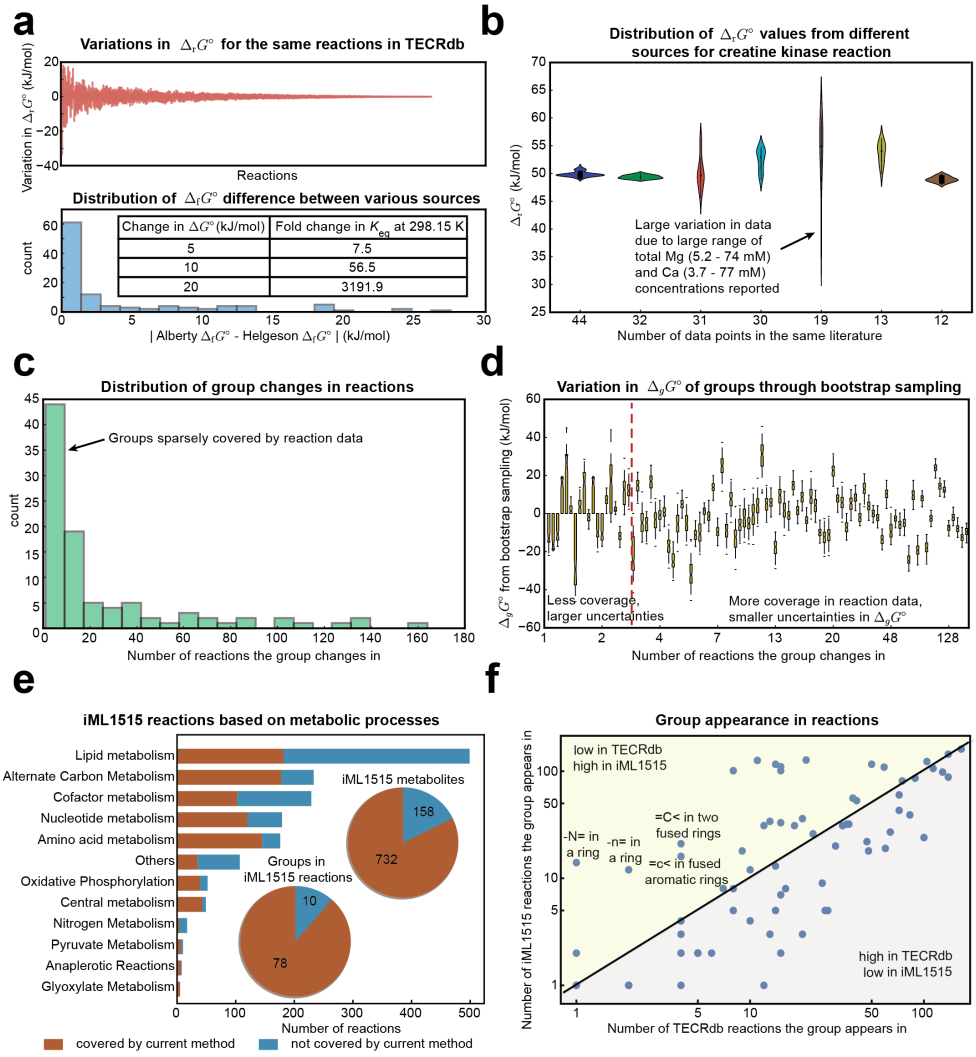


Figure 3.2: Quality and coverage issues with thermodynamic data used to parameterize the group contribution method. a) Consistency of thermodynamic data between studies and data collections. b) Distribution of $\Delta_r G^\circ$ data from the same literature for creatine kinase reaction. The large variation in one study (30-65 kJ/mol) is primarily due to the large range of total magnesium and calcium concentrations utilized in the study, which are not accounted for by the current group contribution method. c) Distribution of group changes in reactions from TECRdb. A large number of groups only appear in limited number of reactions (<10). d) Variation in group energies calculated from bootstrap sampling. We found that groups with fewer appearances in reactions tend to have larger uncertainties. e) Coverage of the current group contribution method on the latest *Escherichia coli* metabolic model iML1515. f) Comparison of group appearance in reactions for TECRdb and iML1515. We divide the plot into two areas, one on groups with low appearance in TECRdb but high in iML1515 and another on groups with high appearance in TECRdb but low in iML1515.

of groups comprising lipids are mostly parameterized based on estimates of nonlipid reactions. Such uncertainty around lipid energies is especially problematic as lipids dissolved in the cellular membrane likely have substantially different thermodynamic properties than metabolites in the aqueous phase.

To identify promising targets for future experimental measurement to maximize benefit to group contribution estimates, we compared the group appearance in reactions between TECRdb and iML1515 and identified groups poorly covered in TECRdb measurements but appearing frequently in IML1515 reactions (Figure 3.2f top left area). Reactions containing undermeasured groups were diverse, including fumarate reductase, murein polymerizing transglycosylase, and imidazole-glycerol-3-phosphate synthase. Experimental measurements of K_{eq}s of reactions containing these groups would be high priority for reducing uncertainties and improving coverage.

Variations in experimental conditions also affect group contribution estimations in ways that are not yet fully accounted for by current implementations. The estimation of $\Delta_g G^\circ$ using reaction data requires all $\Delta_r G^\circ$ values to be transformed to standard conditions. Thus, the transformations of K_{eq}s measured under different conditions to standard conditions $\Delta_r G^\circ$ need to be accurate and complete. Condition-dependent effects that are not accounted for by current methods include the following. (i) **Ionic strength**. Recent work has accounted for the activities of the reactants in dilute aqueous solutions in the presence of ions using the extended Debye-Hückel equation [70–72]. However, a more complete description accounting for specific ionic interactions and higher ionic strength has not been implemented due to parameterization challenges [75, 76]. Additionally, corrections of neutral compound activities due to ionic strength have not been included [77]. (ii) **Metal concentration**. A theory has been developed to correct Keq at different metal concentrations [66], but metal binding constants as the key parameters

are not available broadly. (iii) **Macromolecular environment**. Nonspecific interactions of the reactants and products with their surroundings in living cells can significantly affect K_{eq}s [78]. For example, handling of the activity change with solutes in concentrated protein solutions [79] could become relevant to *in vivo* estimation. Given these numerous effects, one should take the *in vivo* conditions into account and evaluate the possible variations in K_{eq} values accordingly when using K_{eq}s measured *in vitro*.

3.4 Methodological Challenges with Group Contribution Estimation

Beyond data issues, several fundamental issues with the group contribution method itself have persisted that may become the bottlenecks for further improvement in prediction accuracy. These issues are related to: (i) the completeness of the group definitions; (ii) the complexity of group changes in reactions; and (iii) the validity of the additivity assumptions underlying group contribution. While the specific observations shown here depend on the exact group definitions from the latest group contribution framework and might differ in different contexts, the issues identified should be applicable across different frameworks using the method.

3.4.1 Completeness of Group Definitions

Due to limitations in the representation of chemical moieties, the group contribution method fails to decompose the chemical structures of a substantial number of metabolites in iML1515 (87) (Figure 3.3a). Besides those containing undefined R groups, the metabolites that cannot be decomposed contain certain groups not covered by the current method (e.g., aromatic carbon atom with double bond) (Figure 3.3b). Therefore, continued efforts to include new group

definitions are important to make the method comprehensive. In addition, current methods define inorganic small molecules as individual groups (e.g., ammonia, phosphate, and oxygen) and directly set their energies based on existing data due to accuracy issues with their group-based estimations. While these compounds cover the most commonly seen cases, the properties of other inorganic compounds not covered by the framework (e.g., inorganic triphosphate) cannot be estimated accurately. Additionally, cofactors such as NAD/NADH can either be broken down explicitly or treated as individual groups. The former approach ensures consistency in group breakdowns across reactions including those synthesizing the specific cofactor, while the latter approach enforces more accurate redox potentials even if group energies would not accurately predict the cofactor energies.

3.4.2 Complexity of Group Changes in Reactions

The success of group contribution lies in the principle that for each substrate of a reaction, only a few reacting chemical groups change, the energies of which can then be determined through regression against measured data. To prevent this regression problem from being highly underdetermined, the number of groups changing in a reaction should be comparable to the number of reactions containing these groups with measured data. However, using the current group definitions, we observe that a subset of reactions have a large number of changing groups (26 reactions with >10 group changes) (Figure 3.3c). For example, reactions containing many compounds or involving formation and alteration on rings (e.g., dihydroorotase and cyclohydrolase) can result in a large number of group change. Due to overfitting, the uncertainties in estimating $\Delta_{\mathrm{g}}G^{\circ}$ of a large number of groups based on few $\Delta_{\mathrm{r}}G^{\circ}$ values can be significant.

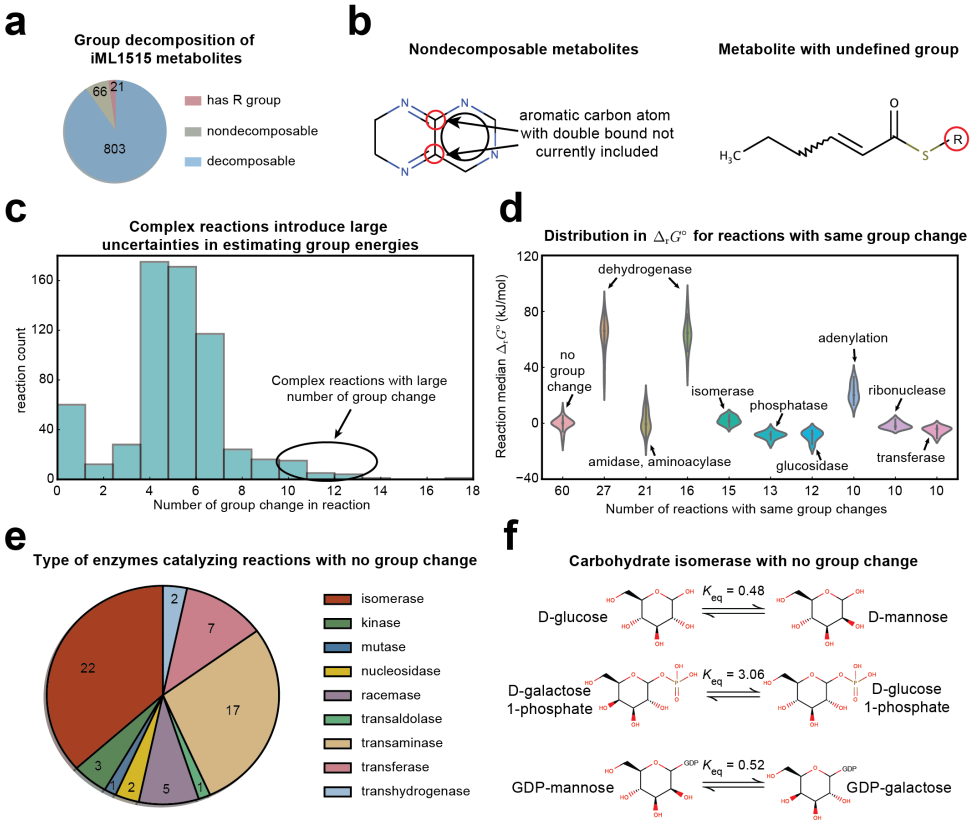


Figure 3.3: Common problems when using group contribution methods to estimate thermodynamic properties of compounds and reactions. a) Inadequacy of the group contribution method to decompose metabolites in a real metabolic network. A number of metabolites in the latest *Escherichia coli* metabolic model iML1515 cannot be decomposed by the current group contribution method. b) Examples of metabolites that cannot be decomposed due to groups that are undefined or not currently included in the method. c) Complex reactions with large number of group change exist and introduce difficulty in determining group energies accurately. d) Distribution in $\Delta_r G^\circ$ for reactions with the same group change. While for several reaction categories $\Delta_r G^\circ$ values are consistent (transferase and isomerase), a number of them show large variations in $\Delta_r G^\circ$ (e.g., dehydrogenase and amidase). e) Reaction types with no group change. Theoretically, such reactions should have zero $\Delta_r G^\circ$. However, we found $\Delta_r G^\circ$ of several reaction classes (e.g. isomerase, kinase, and transaminase) to deviate significantly from 0. f) Examples of reactions, for a set of carbohydrate isomerases, with no group change but nonzero $\Delta_r G^\circ$. Group contribution would predict a K_{eq} of 1 for each reaction, not taking into account small differences in physical properties between the reactants and products in each case.

3.4.3 Validity of the Additivity Assumption

According to the additivity principle, groups should have a constant energy regardless of structural context, and thus reactions with the same group change should have the same energies while those reactions with no group change should have zero $\Delta_r G^\circ$. However, in many cases, the available data tells a different story. Specifically, a number of reactions with the same group change have large variations in $\Delta_r G^\circ$ values (up to 40 kJ/mol) based on measured data, most notably within the dehydrogenase, amidase, and adenylation reaction classes (Figure 3.3d). Taking reactions catalyzed by dehydrogenases for example, the variation in $\Delta_r G^\circ$ values can likely be attributed to other chemical attributes of the reacting moiety (e.g., alcohol, carboxylic acid, or fatty acid) oxidized/reduced in different reactions.

As another telling example, several reactions with no group change have substantially nonzero $\Delta_r G^\circ$ values, notably within isomerase, kinase, and transaminase reaction classes (Figure 3.3e). A closer look at those reactions shows that other molecular properties are clearly contributing to the differences in compound energies, for example, different geometric conformations for structural isomers of sugar molecules (Figure 3.3f) [80]. Furthermore, different tautomers of the compounds exist and have not been accounted for by current methods. These cases demonstrate how group contribution often fails to consider more nuanced molecular interactions that can drive the difference in energy between reactants and products.

3.5 Opportunities for Improvement

Recent developments promise to address many of the challenges outlined above. To improve data quality, further curation and processing of existing data are needed [72]. Specifically, calculation of free metal ion concentrations, rather than the more commonly reported total

concentrations, and a better model to estimate metal binding constants may enable more accurate correction of K_{eq}s at different metal concentrations [72]. Furthermore, developing more detailed representations of ionic interactions and nonionic effects may yield better estimates for complex ionic solutions [81–83].

To improve data coverage, additional thermodynamic data can be incorporated. While significant experimental efforts are necessary to obtain more K_{eq} measurements, it is possible to use other types of data with certain adaptations. For example, gas phase thermodynamic data ($\Delta_f G^\circ(g)$) (https://webbook.nist.gov/chemistry/) and the compound hydration energies ($\Delta_h G^\circ$) from various sources and estimates [84, 85] can be used to calculate aqueous $\Delta_f G^\circ$ [86].

$$\Delta_f G^\circ(aq) = \Delta_f G^\circ(g) + \Delta_h G^\circ \tag{3.3}$$

Also, to address the lack of data around reactions in the lipid phase, existing information on compound lipid partition coefficients or distribution coefficients (LogP, LogD) might be used to estimate the aqueouslipid solvation energy ($\Delta_{sol-lipid} G^\circ$) [87, 88], which relates to the Gibbs formation energy in the lipid phase [89].

$$\Delta_f G^\circ(lipid) = \Delta_f G^\circ(aq) + \Delta_{sol-lipid} G^\circ \tag{3.4}$$

Additionally, computational models of the lipid membrane have been developed that could lead to inroads here [90, 91].

Possible solutions to handle reactions with many changing groups could involve manually adjusting group decompositions for these reactions or redefining the group decomposition rules more generally. Additionally, algorithmic improvements to the statistical models used in group

contribution may also be fruitful. For example, we have found that linear regression with L2 regularization can result in reduced cross-validation errors for $\Delta_r G^\circ$ estimation. The underlying rationale here is that group energies are unlikely to be very large, and this type of regularization prevents unrealistically large values being estimated for groups with little presence in the training reaction data. Additionally, recent developments on estimating $\Delta_r G^\circ$ based on reaction similarity [92] and by combining reactant contribution and group contribution [71] are further examples of promising algorithmic variants of the method.

To address issues related to the validity of the additivity assumption in group contribution, more sophisticated approaches are necessary. One promising approach is to incorporate additional molecular properties such as polarity or partial charge into the estimation workflow, as commonly performed in quantitative structureproperty relationship (QSPR) modeling [93, 94]. Such properties can be estimated with cheminformatics software, including ChemAxon (http://www.chemaxon.com), RDkit (http://www.rdkit.org) and Dragon (http://www.talete.mi.it). Increasing the complexity further, molecular dynamics simulations of the compound can be used either to estimate compound thermodynamic properties directly [95, 96] or to calculate other molecular properties that will later be incorporated into QSPR models [97, 98].

3.6 Concluding Remarks

Thermodynamic analysis of metabolic networks is a burgeoning area of research empowered by genome-scale estimations of reaction K_{eq}s. While group contribution has been steadily improving our ability to estimate K_{eq} with greater scope and accuracy, there remains a variety of challenges in data and methodology that can be addressed to advance the field. As outlined

above, new and more sophisticated approaches are on the horizon that may mitigate the most prominent shortcomings of current implementation of group contribution methods. While substantial challenges remain (see Outstanding Questions), this open discussion of limitations of current methods should spur further development in parameter estimation for thermodynamic modeling in metabolism.

Acknowledgments

This work was supported by the Novo Nordisk Foundation Grant Number NNF10CC1016517.

Chapter 3 in full is a reprint of material published in: **Bin Du**, Daniel C Zielinski, Bernhard O Palsson. 2018. "Estimating metabolic equilibrium constants: progress and future challenges." *Trends in Biochemical Sciences*, 43(12), 960-969. The dissertation author was the primary author.

Chapter 4

Temperature-dependent estimation of Gibbs energies using an updated group contribution method

4.1 Abstract

Response to acid stress is critical for *Escherichia coli* to successfully complete its life-cycle by passing through the stomach to colonize the digestive tract. To develop a fundamental understanding of this response, we established a molecular mechanistic description of acid stress mitigation responses in *E. coli* and integrated them with a genome-scale model of its metabolism and macromolecular expression (ME-model). We considered three known mechanisms of acid stress mitigation: 1) change in membrane lipid fatty acid composition, 2) change in periplasmic protein stability over external pH and periplasmic chaperone protection mechanisms, and 3)

change in the activities of membrane proteins. After integrating these mechanisms into an established ME-model, we could simulate their responses in the context of other cellular processes. We validated these simulations using RNA sequencing data obtained from five E. coli strains grown under external pH ranging from 5.5 to 7.0. We found: i) that for the differentially expressed genes accounted for in the ME-model, 80% of the upregulated genes were correctly predicted by the ME-model, and ii) that these genes are mainly involved in translation processes (45% of genes), membrane proteins and related processes (18% of genes), amino acid metabolism (12% of genes), and cofactor and prosthetic group biosynthesis (8% of genes). We thus established a quantitative framework that describes, on a genome-scale, the acid stress mitigation response of E. coli that has both scientific and practical uses.

4.2 Background

The First and Second Laws of Thermodynamics connect reaction flux directions, metabolite concentrations, and reaction equilibrium constants. An increasing number of systems biology methods have begun to take advantage of the intimate connection between thermodynamics and metabolism to obtain insights into the function of metabolic networks. These methods have been used in a number of applications including the calculation of thermodynamically-feasible optimal states [99, 100], the identification of thermodynamic bottlenecks in metabolism [101, 102], and the constraint of kinetic constants via Haldane relationships [103].

To perform thermodynamic analyses on metabolic networks, it is necessary to have values for the equilibrium constants of reactions carrying flux in the network. Experimentally, the equilibrium constant of a reaction is determined by calculating the mass action ratio (the ratio of product to substrate concentrations), also called the reaction quotient, when the reac-

tion is at equilibrium. A collection of experimentally measured equilibrium constants for over 600 reactions has been published in the NIST Thermodynamics of Enzyme-Catalyzed Reactions database (TECRdb) [104]. However, the equilibrium constants of the majority of known metabolic reactions are still unmeasured, making computational estimation necessary. The most commonly-used approach for estimating thermodynamic constants in aqueous solutions is the group contribution method [69, 105]. This method is based on the simplifying assumptions that the Gibbs energy of formation ($\Delta_{\mathrm{f}}G^{\circ}$) of a compound is based on the sum of the contributions of its composing functional groups, which are independent of each other. The contribution of each group can be estimated through linear regression, using existing data on $\Delta_{\mathrm{f}}G^{\circ}$ and Gibbs energies of reactions ($\Delta_{\mathrm{r}}G^{\circ}$).

Recent iterations of group contribution methods for reactions in aqueous solutions have incorporated pH corrections into estimations of equilibrium constants [106] and improved accuracy by taking advantage of fully-defined reaction stoichiometric loops forming First Law energy conservation relationships within the training data [107]. These methods also have begun to take advantage of computational chemistry software to estimate the pK_{a}s of compounds as part of thermodynamic parameter estimation. However, a number of issues remain for thermodynamic estimation of reaction equilibrium constants in metabolic networks, including: 1) significant estimation errors in many cases, which may be attributed to a number of factors including missing or erroneous reaction conditions, and 2) the lack of an established method to handle correction of thermodynamic data with respect to temperature changes across conditions. Additionally, existing group contribution methods have not taken into account the substantial metal ion binding of many metabolites at physiological ion concentrations, although established theory exists to correct reaction equilibrium constants for metal ion binding when ion dissociation constants are

available [66].

The geochemistry field has developed sophisticated theory to handle thermodynamic variables as a function of temperature for a wide variety of compounds in aqueous solutions [108–114]. The parameters used to calculate thermodynamic transformation across temperature are specific for different compounds. However, the available literature only covers less than half of the compounds in NIST TECRdb. Therefore, the estimation of a large number of compound-specific parameters is required. It is possible to use a group contribution approach by incorporating these parameters into the formulation of $\Delta_{\mathrm{r}}G'^{\circ}$ and fitting them against experimental data at different temperatures. However, due to the lack of data in necessary depth and resolution, the parameter estimation procedure on the fully-parameterized thermodynamic model can suffer large errors from overfitting of parameters. Therefore, a simplified approach with fewer parameters to transform $\Delta_{\mathrm{r}}G'^{\circ}$ across temperature is desirable.

In this study, we extend the capabilities of computational estimation of reaction equilibrium constants for metabolic networks. We first curate the NIST TECRdb of reaction equilibrium constants to obtain missing reaction conditions and correct any other errors. We further incorporate additional thermodynamic data, including $\Delta_{\mathrm{f}}G^{\circ}$ and data related to proton and metal ion binding, from a number of other sources [66, 115–119]. The equilibrium constants and $\Delta_{\mathrm{f}}G^{\circ}$ values are commonly used as the training data for group contribution. The proton and metal ion binding data are required to transform $\Delta_{\mathrm{r}}G'^{\circ}$ across different pH and metal ion concentrations. To enable calculation of equilibrium constants as a function of temperature, we adapt the thermodynamic theory from the geochemistry literature [108–114] given certain simplifying assumptions. The thermodynamic parameters required for such calculation, $\Delta_{\mathrm{f}}S^{\circ}$ of aqueous species, are estimated through the regression model using various molecular descriptors. Next,

to fill gaps in magnesium binding correction of equilibrium constants, we estimate magnesium binding constants for 618 compounds using molecular descriptors and magnesium binding groups defined based on known magnesium binding compounds. Finally, we incorporate these new data and functionalities into the most recently published group contribution framework, termed the component contribution [107], to obtain a new group contribution estimator for reaction equilibrium constants with expanded capabilities.

4.3 Methods

4.3.1 Workflow for estimation of equilibrium constants

We first introduce the workflow for estimation of equilibrium constants illustrated in Figure 4.1a. The following sections expand upon the workflow in greater detail. We collected and curated 4298 equilibrium constants (K') for 617 unique reactions measured under different conditions (temperature, pH, ionic strength, metal ion concentrations) as the training data set for the current group contribution method (Figure 4.1b). We also collected $\Delta_{\mathrm{f}}G^{\circ}$ values from multiple sources as the training data [66, 115, 116]. We collected and curated stability constants of metal-ion complexes from The IUPAC Stability Constants Database and $\Delta_{\mathrm{f}}S^{\circ}$ from various literature sources and online databases [66, 115, 116] (Figure 4.1c). To complete the necessary thermodynamic transformations to reference conditions, we estimated different thermodynamic properties for compounds where data were not available. We estimated $\mathrm{p}K_{\mathrm{a}}$ values using ChemAxon (http://www.chemaxon.com). We used regression models to estimate magnesium binding constants ($\mathrm{p}K_{\mathrm{Mg}}$) and $\Delta_{\mathrm{f}}S^{\circ}$ based on collected data.

First of all, we transformed all measurements to the same reference conditions at 298.15

K, pH 7, 0 M ionic strength and no metal concentration. We applied a Legendre transform to account for the different ion binding states of each compound as in the previous component contribution method [107]. The transformation of Gibbs free energy of reaction across pH and ionic strength is also based on the previous method. However, we used the Davies equation rather than the extended Debye-Hückel equation to calculate activity coefficients of electrolyte solutions, as the Davies equation was used in the previous work for thermodynamic transformations across temperature [108–111]. The transformation of Gibbs free energy of reaction across different metal concentrations is based on the formulation described by Alberty [66, 120]. The transformation of Gibbs free energy of reaction across temperature is based on adapted thermodynamic theory from the geochemistry literature [108–111] with simplifying assumptions.

Using $\Delta_{\mathrm{r}}G^{\circ}$ and $\Delta_{\mathrm{f}}G^{\circ}$ data at reference conditions, we applied the component contribution method by Noor et al [107] and obtained estimates of $\Delta_{\mathrm{r}}G^{\circ}$ and $\Delta_{\mathrm{f}}G^{\circ}$ at reference conditions. Using these values, as well as the estimated $\Delta_{\mathrm{r}}S^{\circ}$ to transform $\Delta_{\mathrm{r}}G'^{\circ}$ across temperature (more details in Results) and other thermodynamic transformations applied in the previous work [107], we are able to calculate the equilibrium constant of a given reaction at defined temperature, pH and ionic strength.

4.3.2 Curation of The IUPAC Stability Constants Database

The IUPAC Stability Constants Database (SC-database) contains ion binding data, i.e. dissociation/binding/stability constants, under various conditions from primary literature. Additionally, the database contains several different annotations for binding of protons and metal ions to specific aqueous species. When the ligand is a proton, the related dissociation constant is a $\mathrm{p}K_{\mathrm{a}}$ constant, while when the ligand is a metal ion such as magnesium, the dissociation constant

is a pK_{Mg} (modified to the specific ion) constant. For each compound of interest, we categorized the available binding data specific to each ion bound state. We then corrected binding data to 0 M ionic strength using the Davies equation [121]. For each ion binding reaction, we calculated the median of all available binding data as the value utilized in the fitting.

4.3.3 Features and data used in regression models to estimate pK_{Mg} and $\Delta_f S^\circ$

For estimation of pK_{Mg}, we included a total of 140 data points and 128 molecular descriptors as features for regression models. The molecular descriptors included magnesium binding groups identified from existing pK_{Mg} data, the charge of the compound excluding any magnesium binding groups, sums of partial charge and numbers of different types of atoms, and several additional molecular descriptors from ChemAxon and RDkit. For estimation of $\Delta_f S^\circ$, we included 762 data points and 195 features including group decompositions, sums of partial charge and numbers of different types of atoms, and molecular descriptors from ChemAxon and RDkit. The molecular descriptors of compound were estimated with Calculator Plugins, Marvin 16.11.21, 2016, ChemAxon (http://www.chemaxon.com) and RDKit: Open-source cheminformatics (http://www.rdkit.org).

4.3.4 Comparison of regression methods using nested 10-fold cross-validation

We tested six different regression methods to estimate pK_{Mg} and $\Delta_f S^\circ$. These methods are ridge regression, lasso regression, elastic net regularization, random forests, extra trees and gradient boosting. We applied nested 10-fold cross-validation to compare the performance of these regression methods. The specific implementation of nested 10-fold cross-validation involves generating an outer loop and inner loop of cross-validation. The outer loop separates the whole

dataset into 10 folds, with one fold for testing and the rest for training in each iteration. The training data in each iteration is further separated into 10-folds, and cross-validation is performed in the inner loop to select the optimal model hyperparameters through grid search. We repeated the nested 10-fold cross-validation on each regression method five different times by splitting the data into different subdivisions.

We then assessed model performance through the median absolute residual of testing errors calculated from the outer loop, for a total of 50 folds (10 folds $\times$ 5 repetitions). The testing errors calculated here also reflect how well the model generalizes on unseen data and are thus used as a metric to evaluate model performance. We also evaluated model stability by calculating the relative standard deviation (RSD = standard deviation/mean) of hyperparameters selected by the inner loop, for a total of 50 folds (10 folds $\times$ 5 repetitions). We evaluated both testing error and RSD of hyperparameters when selecting the final regression model to use. For every fitting procedure, we applied standardization on both the training and testing set using the mean and standard deviation of features calculated from the training set.

The regression models, including linear models, tree-based methods and gradient boosting, were implemented using the python package scikit-learn 0.19.1 [122].

4.3.5 Lasso regression for estimation of $\mathrm{p}K_{\mathrm{Mg}}$ and $\Delta_{\mathrm{f}}S^{\circ}$

Based on the evaluation of different regression methods through nested 10-fold cross-validation (more details in Results), we used lasso regression as the model to estimate $\mathrm{p}K_{\mathrm{Mg}}$ and $\Delta_{\mathrm{f}}S^{\circ}$. Specifically, the objective function to minimize is

$$\min_{w} \frac{1}{2n_{\text{samples}}} \|y - Xw\|^2 + \alpha \|w\|_1 \tag{4.1}$$

where y is the vector of data with length n_{samples}, X is the matrix with features in the row corresponding to each data point, w is the vector of coefficients of the model, and α is a constant that tunes the degree of the l_1 penalty.

We repeated 10-fold cross-validation 100 times on $\text{p}K_{\text{Mg}}$ and $\Delta_{\text{f}}S^{\circ}$ datasets respectively to find the optimal α values that lead to the lowest testing errors. We then constructed a lasso regression-based estimator for each $\text{p}K_{\text{Mg}}$ and $\Delta_{\text{f}}S^{\circ}$ dataset, using the selected α value and applying standardization on the dataset.

4.3.6 Comparison of previous and current group contribution method

We compared how the previous [107] and the current group contribution method perform at different temperatures. Since the previous group contribution method does not involve an explicit term to correct for $\Delta_{\text{r}}G'^{\circ}$ at different temperatures, we were only able to substitute different temperatures in thermodynamic transformations and Legendre transform as the temperature transformation on $\Delta_{\text{r}}G'^{\circ}$. On the other hand, the current method includes an explicit term ($\Delta_{\text{r}}S^{\circ}$) besides the RT term to calculate $\Delta_{\text{r}}G'^{\circ}$ at different temperatures. Using the two methods, we calculated $\Delta_{\text{r}}G'^{\circ}$ values of all the TECRdb data measured at different temperatures and the absolute residual of the estimated $\Delta_{\text{r}}G'^{\circ}$ values against experimental data.

We then performed 10-fold cross-validation on the 432 reactions that overlapped between the previous and the current group contribution method. Specifically, we first transformed experimentally measured $\Delta_{\text{r}}G'^{\circ}$ data to the reference state $\Delta_{\text{r}}G^{\circ}$ (298.15 K, pH 7, 0 M ionic strength), with different sequential modifications on this procedure (based on the previous method). These modifications include updated media conditions, the Davies equation to correct for the effect of ionic strength, new compound groups, temperature correction and metal correction. For each

set of $\Delta_r G^\circ$ values obtained, we calculated the median $\Delta_r G^\circ$ of all data points in each unique reaction, and performed 10-fold cross-validation on those 432 $\Delta_r G^\circ$ values. We repeated this procedure 100 times by splitting the data into different subdivisions. We then calculated the median absolute residual of 100 repetitions for each reaction.

Additionally, we also compared how well the two methods perform on the 185 new reactions collected in this work. The first method is based on the previous work by Noor et al [107], while the second method in current work is similar to the first but has several modifications, including updated media conditions, the Davies equation, new compound groups and the temperature correction. We fit the group contribution model using both methods with $\Delta_r G^\circ$ values of the original 432 overlapping reactions as training data, and calculated the absolute residual in predicting $\Delta_r G^\circ$ for the 185 new reactions as the testing set.

4.3.7 Calculation of standard entropy change of formation

The standard entropy change of formation ($\Delta_f S^\circ$) of the compound is not directly available. Given the type of data available, it can be calculated either from $\Delta_f G^\circ$ and the standard enthalpy of formation ($\Delta_f H^\circ$) of the compound

$$\Delta_f S^\circ = (\Delta_f H^\circ - \Delta_f G^\circ)/T \tag{4.2}$$

or from the standard molar entropy (S°) of the compound

$$\Delta_f S^\circ = S^\circ - \sum_{i=1}^{N_e} n_e S_e^\circ \tag{4.3}$$

where S_e° is the standard molar entropy of the element N_e composing the compound and n_e is

the number of atoms for the element N_e.

4.3.8 Implementation and availability of source code

The updated group contribution method has been implemented in python 2.7.6. The source code is available on GitHub (https://github.com/bdu91/group-contribution), together with detailed instructions on how to install and examples using the package.

4.4 Results

4.4.1 Collection and curation of thermodynamic data

The workflow for estimating reaction equilibrium constants under given pH, temperature, ionic strength and metal ion concentrations is demonstrated in Figure 4.1a (Methods). To obtain the necessary data for this estimation, we curated a number of databases and primary literature sources. First of all, from the NIST TECRdb (https://randr.nist.gov/enzyme) [104], we obtained measured equilibrium constants (K') and enthalpies of reactions ($\Delta_{\mathrm{r}}H'^{\circ}$) for 617 and 207 unique reactions, respectively. Noticing a number of gaps in experimental conditions and other minor issues, we curated a total of 4298 measured K' data from NIST TECRdb. This curation effort resulted in 48.9% corrected data entries, including updated experimental media conditions (35.78%), addition of new data (5.12%), correction of K' values (3.49%), removal of problematic data (3.33%) and correction of reaction formulae (1.14%) (Figure 4.1b).

Next, we collected data on standard Gibbs free energies of formation ($\Delta_{\mathrm{f}}G^{\circ}$), standard enthalpies of formation ($\Delta_{\mathrm{f}}H^{\circ}$) and standard entropy of formation changes ($\Delta_{\mathrm{f}}S^{\circ}$) for 312, 254 and 499 unique compounds, respectively (Figure 4.1c). $\Delta_{\mathrm{f}}S^{\circ}$ data are usually not directly measured but instead are calculated from either $\Delta_{\mathrm{f}}G^{\circ}$ and $\Delta_{\mathrm{f}}H^{\circ}$ data or standard molar entropy

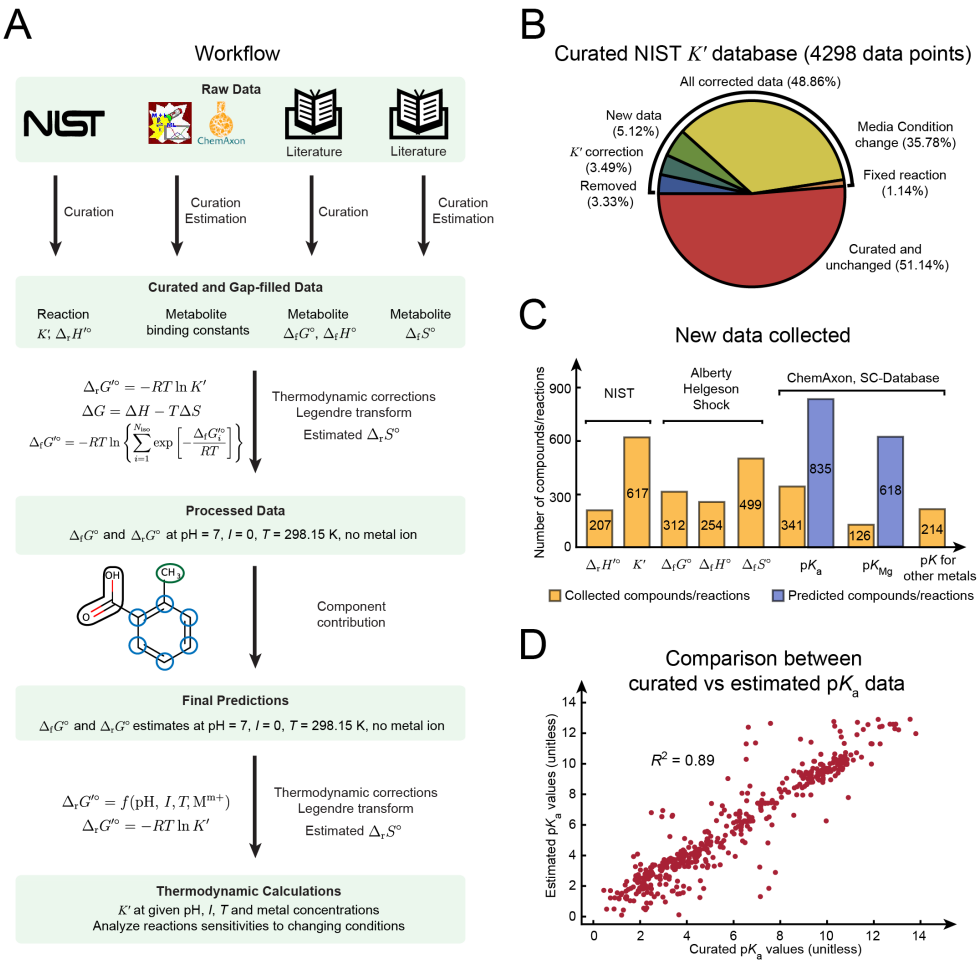


Figure 4.1: Estimation of reaction equilibrium constants. a) Workflow of data curation and parameter fitting for equilibrium constant estimation. b) The results of curation of equilibrium constants from the NIST Thermodynamics of enzyme-catalyzed reactions database (TECRdb). c) New thermodynamic properties generated, either collected from sources shown or computationally estimated. d) Comparison between curated pK_a data from The IUPAC Stability Constants Database (SC-database) as well as literature with computationally estimated pK_a values from ChemAxon.

(S°) of the compound (Methods). The above data are from multiple sources: *Thermodynamics of Biochemical Reactions* by Alberty [66], the SUPCRT92 database [115] and the Organic Compounds Hydration Properties Database [116].

Lastly, we collected and curated pK_a data for 341 compounds, magnesium binding constants for 126 compounds, and other metal type binding constants for 214 compounds (including cobalt, iron, zinc, sodium, potassium, manganese, calcium, lithium) from The IUPAC Stability Constants Database (SC-database) and primary literature [117–119] (Figure 4.1c). We also predicted pK_a data for 835 compounds using ChemAxon (http://www.chemaxon.com) (Figure 4.1c). We compared the collected pK_a data and the predicted values from ChemAxon for the same compounds (Figure 4.1d). We found that the differences between the collected and predicted pK_a values can be as large as 5.84 (unitless), with a median of 0.42 (unitless). This error is a large enough difference to substantially alter the major protonation states for metabolites containing groups with pK_as around physiological pH. We examined the specific cause of the largest discrepancies and found that they are due to issues such as assignment of the pK_a value to the wrong charged form by ChemAxon (e.g. 4-oxo-L-proline) or error in calculating pK_as related to particular molecular moieties, such as nitrogenous bases and nitrogen atoms on unsaturated rings (e.g. 2′-deoxyguanosine 5′-monophosphate, xanthine-8-carboxylate, deaminocozymase). We thus used measured pK_a data when available.

4.4.2 Thermodynamic parameters for transformation of $\Delta_\mathrm{r}G'^\circ$ across temperature

We then sought to develop the capability to calculate standard transformed Gibbs energy of reaction ($\Delta_\mathrm{r}G'^\circ$) as a function of temperature. Specifically, we adapted theory from the

geochemistry literature under constant enthalpy and entropy assumptions [108–111], as well as the assumption that the contribution of heat capacity to change in Gibbs energy over temperature is negligible compared to the contribution of entropy. Thus, we obtained a simple linear formulation of $\Delta_{\mathrm{r}}G'^{\circ}$ at a given temperature T using the standard entropy change of reaction $\Delta_{\mathrm{r}}S^{\circ}$ at a reference T_r (298.15 K):

$$\Delta_{\mathrm{r}}G'^{\circ}_{T} = \Delta_{\mathrm{r}}G'^{\circ}_{T_{\mathrm{r}}} - (T - T_{\mathrm{r}})\Delta_{\mathrm{r}}S^{\circ}_{T_{\mathrm{r}}} \tag{4.4}$$

As $\Delta_{\mathrm{r}}S^{\circ}_{T_{\mathrm{r}}}$ (we use $\Delta_{\mathrm{r}}S^{\circ}$ in later references since T_r is the only condition of interest, same for $\Delta_{\mathrm{f}}S^{\circ}$) of reactions can be calculated from $\Delta_{\mathrm{f}}S^{\circ}$ of the compounds involved, we sought to construct a regression model to estimate $\Delta_{\mathrm{f}}S^{\circ}$ values. Besides collecting 669 $\Delta_{\mathrm{f}}S^{\circ}$ values for 499 compounds at different protonation states as training data, we also collected $\Delta_{\mathrm{r}}S^{\circ}$ values from multiple sources. These $\Delta_{\mathrm{r}}S^{\circ}$ values are effectively linear combinations of $\Delta_{\mathrm{f}}S^{\circ}$ values and can also be used as training data for $\Delta_{\mathrm{f}}S^{\circ}$ estimation. From NIST TECRdb, we selected reactions with K' data measured under at least 4 different temperatures. We then calculated $\Delta_{\mathrm{r}}S^{\circ}$ of each reaction using the $\Delta_{\mathrm{r}}G'^{\circ}$ of the reaction at different temperatures based on Equation 4.4, obtaining 51 $\Delta_{\mathrm{r}}S^{\circ}$ values. Next, we picked reactions in NIST TECRdb with both $\Delta_{\mathrm{r}}G^{\circ}$ and $\Delta_{\mathrm{r}}H^{\circ}$ data available and calculated their $\Delta_{\mathrm{r}}S^{\circ}$ values, obtaining 41 additional data points. Together, we obtained a total of 762 data points for $\Delta_{\mathrm{f}}S^{\circ}$ estimation.

4.4.3 Estimation of standard entropy change of formation $\Delta_{\mathrm{f}}S^{\circ}$

We found that simple molecular descriptors, notably the number of atoms in the compound and the compound charge, were highly useful as predictors for $\Delta_{\mathrm{f}}S^{\circ}$. Specifically, we found $\Delta_{\mathrm{f}}S^{\circ}$ data to be highly correlated simply with the total numbers of atoms in the compound, with

an R^2 of 0.89 (Figure 4.2a). The $\Delta_f S^\circ$ data as a function of atom number are separated into two main clusters, one of which contains aqueous species with large atom numbers and large absolute $\Delta_f S^\circ$ values (NAD, NADH, NADP, NADPH). The other cluster contains a wide variety of aqueous species, with a few categories labeled in Figure 4.2a. We noticed clear separations among aqueous species with -5, -4, -3 and -2 charge, but less so for those with -1, 0 and +1 charge (Figure 4.2a). We found the trend between $\Delta_f S^\circ$ and number of atoms exists even more strongly among compounds within the same homologous series, where the compound structures differ only by the number of CH_2 units in the main carbon chain. Specifically, $\Delta_f S^\circ$ value decreases by approximately 0.11 kJ/K/mol with every additional CH_2 unit. This trend was observed in a number of homologous series including alkanes, alkenes, alkynes, aldehydes, single carboxylic acids, amines, amides, and thiols. However, the change in $\Delta_f S^\circ$ with respect to number of atoms across different homologous series is inconsistent, thus requiring additional molecular descriptors.

As an additional descriptor, we found that partial charge of atoms can help distinguish $\Delta_f S^\circ$ from different homologous series. For example, the carbon atoms in glycerol (alcohol containing multiple hydroxyl groups) have larger partial charges than those in methanol (alcohol containing a single group). The prediction of glycerol $\Delta_f S^\circ$ from methanol $\Delta_f S^\circ$ based on their difference in atom numbers yielded a smaller absolute $\Delta_f S^\circ$ value than the actual glycerol data (Figure 4.2b). The correlation of larger partial charges of carbon atoms with larger absolute $\Delta_f S^\circ$ is also observed in other pairs in Figure 4.2b (deoxyribose vs. ribose, methanol + formic acid vs. glycolic acid, benzene + formic acid vs. benzoic acid). Besides carbon atoms, we also found differences in partial charges of oxygen atoms to be associated with $\Delta_f S^\circ$ differences, as shown between formic acid and oxalic acid (Figure 4.2b). Following these observations, we included the sums of absolute partial charge of each type of atom as molecular descriptors for the regression

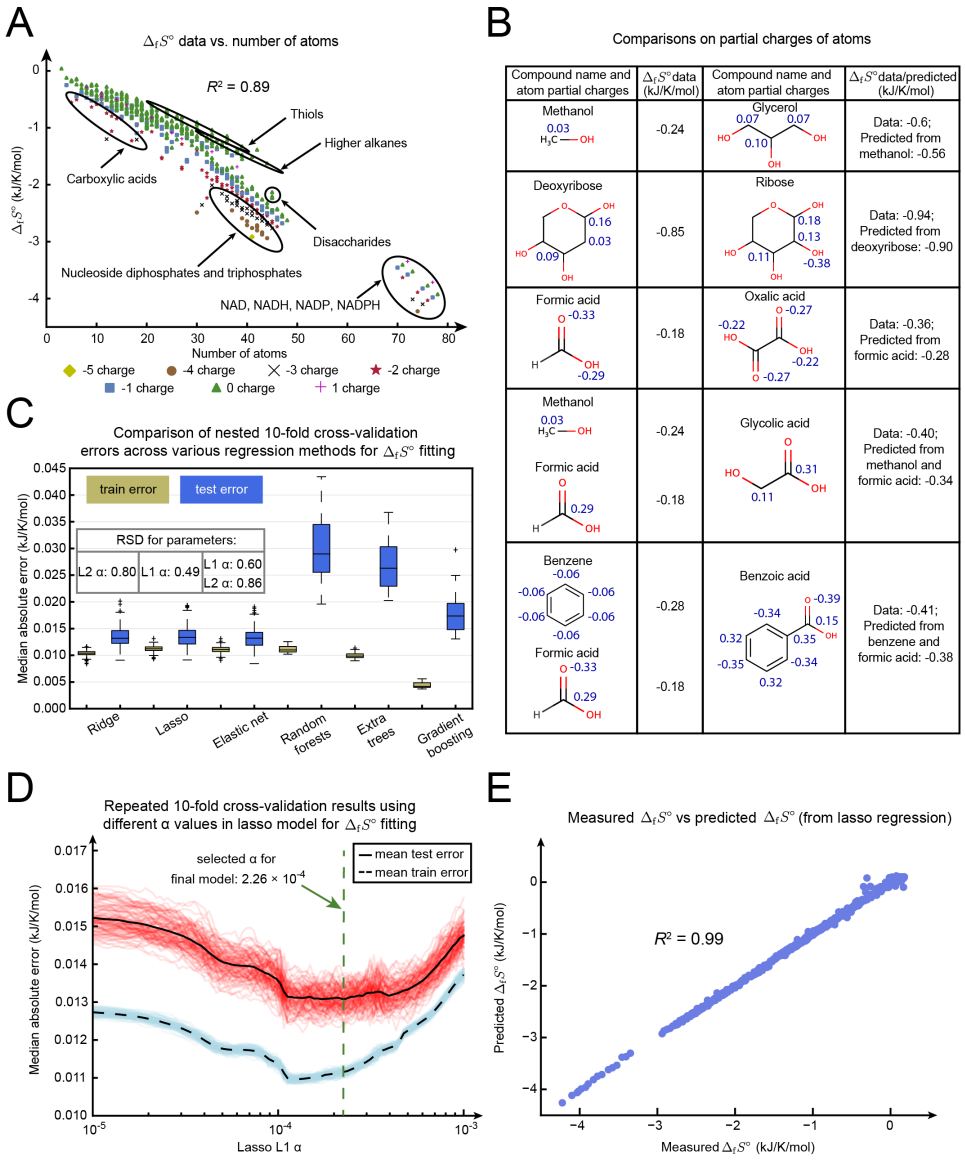


Figure 4.2: Estimation of standard entropy change of formation ($\Delta_f S°$). a) $\Delta_f S°$ vs. number of atoms. b) Comparisons of partial charges of atoms between compounds. Each row contains a pair of compounds and their $\Delta_f S°$ data. In each pair, $\Delta_f S°$ of the latter compound can be predicted from that of the former compound(s) based on difference in atom number. The partial charges of atoms that are different within each pair are marked in blue. c) Training and testing errors of nested 10-fold cross-validation (repeated 5 different times) on $\Delta_f S°$ data using 6 different regression methods. d) Selection of parameters in the lasso regression using 10-fold cross-validation on all $\Delta_f S°$ data. We repeated 10-fold cross-validation 100 times and calculated training (blue) and testing (red) errors at α from 10^{-5} to 10^{-3}. The mean training and testing errors are shown in dashed and solid black lines. e) Comparison of 762 measured $\Delta_f S°$ training data vs predicted $\Delta_f S°$ values from the final lasso regression model.

model.

In addition to partial charge, we also considered a number of other molecular descriptors from ChemAxon and RDkit (Methods). We obtained a total of 195 features and 762 $\Delta_f S^\circ$ data for regression models. We performed nested 10-fold cross-validation to compare between multiple regression models (Figure 4.2c). We selected lasso regression as the final model to use since it has significantly smaller testing errors compared to more complex methods and the least variation in parameters selected from cross-validation compared to other linear regression methods (Figure 4.2c). Using parameters selected from cross-validation on the entire $\Delta_f S^\circ$ dataset (Figure 4.2d), we constructed a lasso regression model and predicted 672 $\Delta_f S^\circ$ values. We obtained 121 predictive variables from the final lasso model, including the number of carbon, hydrogen and oxygen atoms, the partial charge of hydrogen and oxygen atoms, the formal charge of the compound, the presence of phosphate groups, and the solvent accessible surface area. The median absolute residual of the lasso regression model for $\Delta_f S^\circ$ estimation is 0.013 kJ/K/mol (Figure 4.2c). Since $\Delta_r S^\circ$ values are linear combinations of $\Delta_f S^\circ$ values, we used the final lasso regression model to estimate the $\Delta_r S^\circ$ values for all 617 reactions in TECRdb.

4.4.4 Evaluation of temperature-dependent estimation of $\Delta_r G'^\circ$

We next evaluated the performance of our method in estimating $\Delta_r G'^\circ$ at different temperatures. We calculated $\Delta_r G'^\circ$ values of all the K' data measured at different temperatures in TECRdb, using the current method with estimated $\Delta_f S^\circ$ values and the previous group contribution method [107]. We calculated the absolute residuals of $\Delta_r G'^\circ$ estimation and compared the two methods across temperature. We found that our method resulted in smaller residuals than the previous method in all temperature ranges (Figure 4.3a). This result is also confirmed

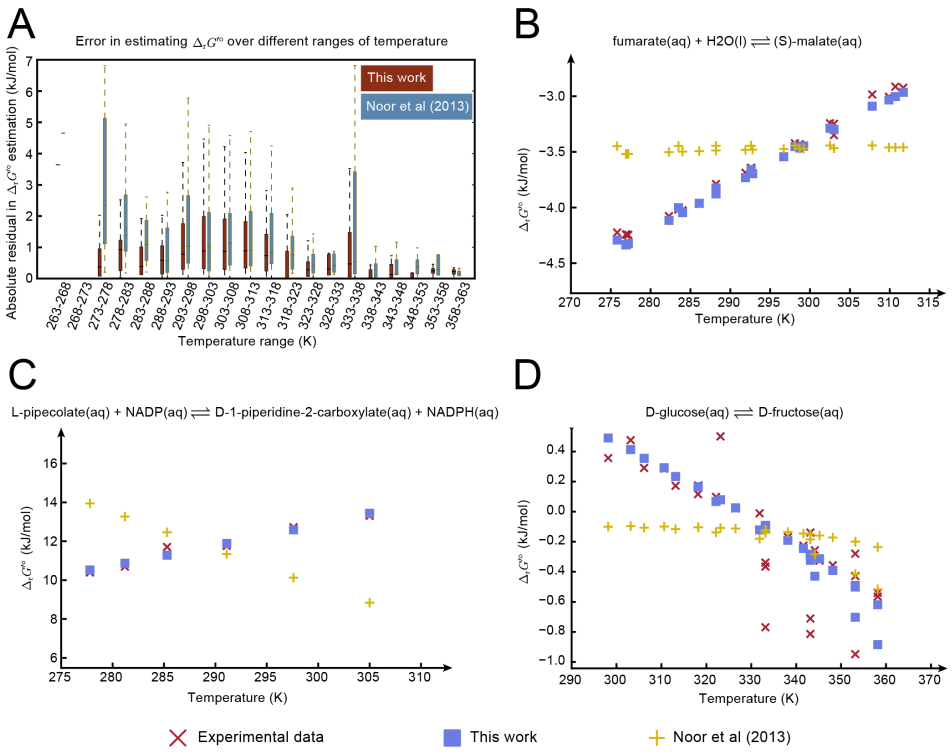


Figure 4.3: Evaluation of temperature-dependent estimation of $\Delta_r G'^{\circ}$. a) Comparison of absolute residuals on estimating $\Delta_r G'^{\circ}$ at different temperatures between the previous group contribution method [107] and the current method. For all the TECRdb data measured at different temperatures, we estimated the $\Delta_r G'^{\circ}$ values using the previous method and the current methods and calculated the absolute residual against experimental data. For clarity in comparison, we divided the entire temperature range into windows with 5K difference. b) Estimated $\Delta_r G'^{\circ}$ values for fumarate hydratase reaction at different temperatures using the previous method and the current method. c) Estimated $\Delta_r G'^{\circ}$ values for 1-piperidine-2-carboxylate reductase reaction at different temperatures using the previous method and the current method. d) Estimated $\Delta_r G'^{\circ}$ values for xylose isomerase reaction at different temperatures using the previous method and the current method.

in different reactions where we identified series of K' data measured at different temperatures. In all those cases, our estimated $\Delta_r G'^\circ$ across temperature agreed well with the experimental data, in contrast to the estimations by the previous method (Figure 4.3b-d). Additionally, we found the temperature-dependent estimation of $\Delta_r G'^\circ$ to be quite robust in the temperature range of available data in TECRdb (0 - 90°C), which covers the living conditions of most organisms. Examining reactions whose $\Delta_r G^\circ$ values are predicted to be sensitive to change in temperature (large $\Delta_r S^\circ / \Delta_r G^\circ$ ratio), a number of interesting cases in central metabolism were identified, including malate dehydrogenase, amino acid transaminase and transketolase.

4.4.5 Estimation of unknown magnesium binding constants

In addition its dependence on temperature, the standard transformed Gibbs free energy of the compound ($\Delta_f G'^\circ$) can also depend on pH and the concentrations of metal ions, due to the presence of different protonation states and various metal bound species. Specifically, $\Delta_f G'^\circ$ can be calculated based on the standard transformed Gibbs energies of its different ion bound states ($\Delta_f G'^\circ_1$, $\Delta_f G'^\circ_2$, etc) through Legendre transform [66].

$$\Delta_f G'^\circ = -RT \ln \left\{ \sum_{i=1}^{N_{\text{iso}}} \exp \left[-\frac{\Delta_f G'^\circ_i}{RT} \right] \right\} \tag{4.5}$$

The equation can be rewritten as

$$\Delta_f G'^\circ = \Delta_f G'^\circ_1 - RT \ln \left\{ 1 + \exp \left(\frac{\Delta_f G'^\circ_1 - \Delta_f G'^\circ_2}{RT} \right) + \exp \left(\frac{\Delta_f G'^\circ_1 - \Delta_f G'^\circ_3}{RT} \right) + ... \right\} \tag{4.6}$$

where $\Delta_f G'^\circ_1$ is the Gibbs energy of a particular ion bound state (typically with the least hydrogens

and metal ions bound). The Gibbs energy of a specific ion bound state ($\Delta_f G_i'^\circ$) can then be written in terms of $\Delta_f G_1'^\circ$ and the binding polynomial P_i

$$\Delta_f G_i'^\circ = \Delta_f G_1'^\circ - RT \ln P_i \tag{4.7}$$

where P_i is expressed in terms of the proton concentration and metal ion concentration, as well as the binding constants of successive proton and metal ion binding steps to obtain the i^{th} ion bound state [66]. Therefore, metal binding constants are important parameters that affect $\Delta_f G'^\circ$ and subsequently reaction equilibrium constants.

We focused on magnesium binding since the magnesium ion is well known to bind to various metabolites, and its binding to ATP and other phosphate-containing compounds has been characterized experimentally [123, 124]. However, magnesium binding data is still lacking for a large number of compounds that contain similar structural groups to those known to bind magnesium, suggesting that many more compounds may have substantial magnesium binding than have been measured.

Based on the structures of compounds with known magnesium binding, we determined 31 magnesium binding groups, most of which are phosphate and carboxyl groups. We were unable to determine the specific binding groups for certain categories of compounds that were measured to complex with magnesium, including nucleobases, ribonucleosides, and purine derivatives. To try to capture metabolite properties responsible for Mg binding in these cases, we added molecular properties (Methods) as additional descriptors. Together we used 128 features and 140 measured magnesium binding constants to construct several candidate regression models for the prediction of magnesium dissociation constants. We performed nested 10-fold cross-validation to compare between multiple regression models. We selected lasso regression as the best predictor due to its

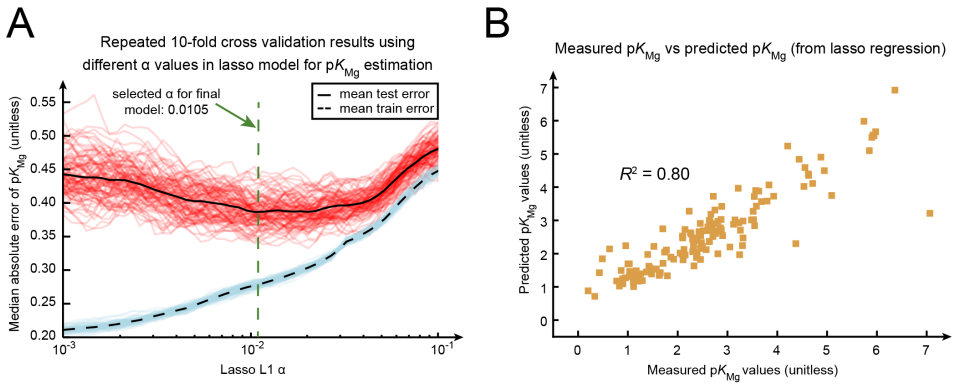


Figure 4.4: Estimation of compound magnesium binding constants (pK_{Mg}). a) Selection of parameters in the lasso regression using 10-fold cross-validation on all pK_{Mg} data. We repeated 10-fold cross-validation 100 times and calculated training (blue) and testing (red) errors at α from 10^{-3} to 10^{-1}. The mean training and testing errors are shown in dashed and solid black lines. The selected α at the lowest mean testing error is 0.0105 (unitless). b) Comparison of 140 measured pK_{Mg} training data vs predicted pK_{Mg} values from the final lasso regression model.

superior generalizability compared to more complex methods and stable model parameters across cross-validation replicates compared to other linear methods. Using 140 measured magnesium binding constants as training data, we constructed a lasso regression model with parameters tuned through cross-validation (Figure 4.4a) and predicted 1707 magnesium binding constants for aqueous species from 618 compounds. We obtained 35 predictive variables from the final lasso model, including the formal charge, the solvent accessible surface area, the presence of various phosphate groups for magnesium binding, the partial charge of nitrogen atoms, the compound charge excluding its magnesium binding groups and dipole moment of the molecule. We found 34 of the 618 compounds are predicted to bind to magnesium at physiological concentrations (2 to 3 mM) [125]. The median absolute residual of the lasso regression model for magnesium binding constant estimation is 0.39 (unitless), as calculated by the nested 10-fold cross-validation.

4.4.6 Estimation of standard Gibbs free energy of reaction

Utilizing the curated and estimated datasets mentioned above, as well as the estimation of $\Delta_r S^\circ$, we adapted the most recent group contribution-based method, termed component contribution [107], to calculate reaction equilibrium constants for a set of 617 unique reactions in NIST TECRdb. Besides the addition of transformation of $\Delta_r G'^\circ$ across temperature, we also included 17 novel group definitions to account for compounds with new functional groups not covered by the previous component contribution method. Additionally, we used the Davies equation [121] rather than the extended Debye-Hückel equation (used in the previous component contribution method [107]) to correct for the effect of ionic strength, as the Davies equation was used in the previous work on temperature-dependent thermodynamic calculations [108–111]. We also showed that the Davies equation was slightly more effective in correcting data at high ionic strength compared to the extended Debye-Hückel equation. On top of the new functionalities, we also added additional $\Delta_r G^\circ$ values for 185 reactions and $\Delta_f G^\circ$ values for 178 compounds over the dataset used in the previous method.

We compared the accuracy of the updated component contribution method with the previous work using repeated 10-fold cross-validation (Methods) for a set of 432 overlapping reactions [107]. We applied the modifications mentioned above sequentially on the framework to examine how each new functionality affects the estimation error globally. We first noted that the updated media conditions increased the median absolute residual of $\Delta_r G^\circ$ estimation (6.21 kJ/mol), which we found to be due to the addition of data at high ionic strength (> 0.5M, beyond the working range of the Davies equation). Removal of those data resulted in similar errors as in the previous work (5.95 kJ/mol). We found modest decrease in median absolute residual with the additional group definitions (5.82 kJ/mol) and capability to transform Gibbs energy

of reaction across temperature (5.71 kJ/mol). Surprisingly, we observed a considerable increase in error (6.47 kJ/mol) after applying the correction on magnesium concentration globally. We investigated this issue in detail and found that problems related to inconsistency in measured K' data (involving magnesium binding) and report of total magnesium concentration can be major sources of error, even though the correction works with well curated data. Therefore, we proceed by omitting the global correction on magnesium concentration from our procedure.

Additionally, we compared our method to the most recent method by predicting $\Delta_{\mathrm{r}}G^{\circ}$ for 185 new reactions collected in this work, using the 432 overlapping previous reactions as training data. We found the median absolute residual from the current method (8.17 kJ/mol) is notably smaller than that from the previous work (11.47 kJ/mol).

To summarize, we included the Davies equation, new group definitions and temperature transformation capabilities, but not the magnesium correction, in our final group contribution framework. We used the equilibrium constants from TECRdb and the collected $\Delta_{\mathrm{f}}G^{\circ}$ values as the training data. Additionally, we used the collected pK_{a} data from the SC-database when possible and estimated the rest using ChemAxon. Overall, our method led to improved performance compared to the most recent group contribution method, while adding the capability to correct equilibrium constants with respect to temperature and substantially expanding the scope of predictions and thermodynamic datasets used in estimation.

4.5 Discussion

In this work, we expanded the scope of thermodynamic calculations to more compounds and reactions with both curated and estimated data, and also extended the group contribution methods for estimating reaction equilibrium constants to account for variations in temperature.

We first collected and curated thermodynamic data, including K', $\Delta_r H'^\circ$, $\Delta_f G^\circ$, $\Delta_f H^\circ$, $\Delta_f S^\circ$ and various ion binding constants, from a number of databases. We then applied existing thermodynamic theory with simplifying assumptions to enable the calculation of Gibbs free energy of reaction across temperature and estimated the necessary parameters ($\Delta_f S^\circ$) using a linear regression model. We also estimated magnesium binding constants for 618 compounds using molecular descriptors and magnesium binding groups based on existing binding data. With new capabilities and new data, we utilized an updated group contribution method to calculate equilibrium constants with improved accuracy over previous work.

The curation of NIST TECRdb revealed that fully-specified media conditions, which influence the ionic strength and metal ion concentration corrections, were often lacking. Surprisingly, curating the literature and filling in media conditions did not improve the resulting fit on the estimation of equilibrium constants, with one possible cause that we added data at high ionic strengths that exceed the intended range of the Davies and Debye-Hückel models for chemical activity. Another possible source of error could be related to the relatively simple model used to account for the effect of ionic strength on activity coefficients of aqueous electrolytes. The Davies equation fails to account for specific interactions between various ions present in solution and is unable to calculate activity coefficients at temperatures other than 298.15 K. Equations with a more comprehensive handling of these thermodynamic theories are established [75, 76, 108–111], but require substantially more data than is currently available for the vast majority of compounds.

Utilizing reasonable assumptions of constant enthalpy and entropy over the range of biological interest, we formulated a simplified approach to calculate temperature transformation of Gibbs energy of reaction and reduced the number of parameters needed for estimation drastically.

With the incorporation of temperature transformation capabilities, we obtained similar errors in estimating $\Delta_r G^\circ$ compared to the previous method [107]. Such similar errors seem to be largely due to the fact that most of the data were measured not far from 298.15 K (83.5% of the data were measured under 295.15 K to 313.15 K), resulting in minor change in correction of K' to the reference conditions. However, we do predict large change in Gibbs energy of many reactions at high temperatures (approaching 373 K), which thus may be significant for high-interest thermophilic organisms such as those living in hot springs and hydrothermal vents.

The compound-specific parameters required for temperature transformation in our simplified model is $\Delta_f S^\circ$, which is missing for a large number of compounds in TECRdb. Using a regression model, we predicted $\Delta_f S^\circ$ of a comprehensive collection of compounds with high accuracy by identifying key chemical properties such as number of atoms and partial charge. The linear correlation of other thermodynamic properties (e.g. standard molar entropy, standard partial molal volume, $\Delta_f G^\circ$) with number of atoms has been demonstrated in previous work [126–129], but only for compounds in the same homologous series. We found the partial charge of atoms to be useful to distinguish $\Delta_f S^\circ$ from different homologous series, possibly due to the fact that the partial charge of atoms of the aqueous species influence its interaction with surrounding water molecules. The regression model was unable to clearly differentiate $\Delta_f S^\circ$ of compounds within certain categories, however, such as monosaccharides and disaccharides. For example, the differences in $\Delta_f S^\circ$ for fructose, mannose and sorbose are around 10 to 20 J/K/mol, while the model only predicts up to 5 J/K/mol difference, due to their similar chemical properties. Such error is not evident when evaluating the accuracy of $\Delta_f S^\circ$ estimation, as $\Delta_f S^\circ$ of monosaccharides are around 1000 J/K/mol. However, when calculating $\Delta_r S^\circ$ of the isomerization reaction between monosaccharides, we found that the errors of $\Delta_f S^\circ$ prediction, though small compared

to $\Delta_f S^\circ$ values, are significant compared to the calculated $\Delta_r S^\circ$ values. We observed this issue to be prevalent for a number of reactions in NIST TECRdb. Thus, identification of new molecular properties or additional features describing group interactions to more accurately differentiate these complex carbohydrates can be a productive next step to improve $\Delta_f S^\circ$ estimation. Additionally, the error in $\Delta_f S^\circ$ estimation can be incorporated into the calculation of confidence intervals developed by the previous method [107], offering the capability to assess the error in estimating $\Delta_r G'^\circ$ at different temperatures.

We demonstrated that magnesium binding groups (specifically the phosphate groups) that could be identified from known magnesium binding compounds are useful features to estimate magnesium binding constants with good accuracy. However, we found a number of compounds that complex with magnesium do not contain the binding groups we defined. These compounds include nucleobases, ribonucleosides, deoxyribonucleosides, purine derivatives and small chemicals such as ammonia, thiocyanate and urea. Currently, we use molecular properties to describe their binding to magnesium. Such issue in identifying the chemical moiety responsible for magnesium binding can still make it difficult to extend our predictions to new compounds with similar structures as the compounds described above. The approach of estimating magnesium binding constants can also be applied to other metals. However, we did not perform such predictions here due to the scarcity of binding data available for other metals.

We found the overall errors in estimating $\Delta_r G^\circ$ increase with the incorporation of magnesium correction using curated and predicted magnesium binding data. We identified inconsistency in K' data (with magnesium binding involved) to be one primary source of error. Another source of error can be due to the uncertainty in estimation of magnesium binding constants and missing binding data for other metals. Additionally, most measurements only reported total metal

ion concentrations, while the metal correction formulation uses free metal ion concentrations. Therefore, additional effort is necessary to calculate free metal ion concentrations from measured data. Due to the lack of binding data and uncertainty in estimated data, an iterative approach might be taken where free metal ion concentrations calculated using the current binding data are applied to optimize the binding data, which are then fed into calculation of free metal ion concentrations.

The current work expands opportunities toward an understanding of thermodynamic factors underlying metabolic network and function in biological systems. This area has generated a number of exciting results, such as the discovery that amino acid biosynthesis, which is endergonic at surface conditions, is exergonic under the conditions of life in hydrothermal vents [130]. Another recent effort proposed proteomic constraints due to thermodynamic bottlenecks as a critical factor underlying metabolic pathway choice [102]. As methods for estimating the thermodynamic properties of metabolic networks continue to improve, these efforts are likely to be increasingly fruitful in uncovering the physical constraints driving the function and evolution of metabolic networks.

4.6 Conclusion

The work here provides an updated group contribution method with an expanded set of thermodynamic data and extended capabilities to calculate equilibrium constants as a function of temperature. We collected and curated thermodynamic data for compounds and reactions from a number of databases and primary literature sources. We established a simple yet well-justified framework, which included formulations derived from existing theory and the necessary parameters ($\Delta_f S^\circ$), to calculate equilibrium constants as a function of temperature. We also

used molecular properties and magnesium binding groups defined from existing data to estimate magnesium binding constants for 618 compounds through a linear regression model. Taken together, this work fills a gap in previous group contribution methods to calculate equilibrium constants to temperature conditions and better correct for magnesium ion binding. These efforts should facilitate the growing number of applications to apply thermodynamic principles to better understand cell metabolism.

Acknowledgments

We would like to thank Nikolaus Sonnenschein for valuable discussions. This work was supported by the Novo Nordisk Foundation Grant Number NNF10CC1016517.

Chapter 4 in full is a reprint of material published in: **Bin Du**, Zhen Zhang, Sharon Grubner, James T Yurkovich, Bernhard O Palsson, Daniel C Zielinski. 2018. "Temperature-dependent estimation of Gibbs energies using an updated group-contribution method." *Biophysical Journal*, 114(11), 2691-2702. The dissertation author was the primary author.

Chapter 5

Thermodynamic favorability and pathway yield as evolutionary tradeoffs in biosynthetic pathway choice

5.1 Abstract

The structure of the metabolic network contains myriad organism-specific variations across the tree of life, but the selection basis for pathway choices in different organisms is not well understood. Here, we examined the metabolic capabilities with respect to cofactor use and pathway thermodynamics of all sequenced organisms in the Kyoto Encyclopedia of Genes and Genomes Database. We found that (i) many biomass precursors have alternate synthesis

routes that vary substantially in thermodynamic favorability and energy cost, creating trade-offs that may be subject to selection pressure; (*ii*) alternative pathways in amino acid synthesis are characteristically distinguished by the use of biosynthetically unnecessary acyl-CoA cleavage; (*iii*) distinct choices preferring thermodynamic-favorable or cofactor-useefficient pathways exist widely among organisms; (*iv*) cofactor-useefficient pathways tend to have a greater yield advantage under anaerobic conditions specifically; and (*v*) lysine biosynthesis in particular exhibits temperature-dependent thermodynamics and corresponding differential pathway choice by thermophiles. These findings present a view on the evolution of metabolic network structure that highlights a key role of pathway thermodynamics and cofactor use in determining organism pathway choices.

5.2 Background

Metabolism has historically been viewed as a highly conserved network across all branches of life [131]. However, as a greater number of organisms are sequenced and characterized [132], there is an increasing appreciation of the diversity of organism-specific metabolic differences [133–135]. Diverse organism living conditions, including nutrient availability, electron acceptors, temperature, pH, pressure, and salt concentrations [136], can create environmental niches that have specific metabolic requirements [137–139]. How these environmental conditions impact metabolic diversity remains an important question [136].

Metabolic network reconstructions are highly curated knowledge bases of metabolic function that provide a way to systematically investigate the differences in metabolic capabilities among various organisms [140, 141]. Metabolic network models, derived from metabolic reconstructions, are mathematical representations of the metabolic capabilities of an organism that can

be used to compute organism phenotypes. Recent efforts reconstructing genome-scale metabolic networks for various organisms have offered a quantitative route to begin to understand the principles underlying metabolic diversity across the tree of life [142, 143].

Metabolic network reconstructions enable a number of powerful computational analyses. First, flux balance analysis (FBA) of metabolic models can calculate the flow of metabolites through the metabolic network by utilizing optimization principles [144]. FBA can be used for a number of calculations such as product yields and substrate utilization efficiency at a network level [145, 146]. Second, FBA can be integrated with thermodynamic equilibrium constants to calculate additional network properties such as thermodynamically feasible optimal states [99] and thermodynamic bottlenecks [102]. These methods thus allow us to evaluate the properties of metabolic pathways in an organism-specific context and provide the basis toward understanding pathway choice among various organisms.

In this study, we utilized metabolic network analysis to evaluate alternative pathway choice in terms of the underlying physicochemical properties of pathways in organisms with diverse lifestyles. We first collected all available information on de novo biosynthesis pathways for biomass precursors and identified organisms containing these pathways. We focused on the biosynthetic pathways for five amino acids with differential use of acyl-CoA cleavage (lysine, arginine, cysteine, isoleucine, and methionine). We examined the basis for preference of *Escherichia coli* for alternative pathway choice in amino acid biosynthesis using in vivo metabolite and protein concentration measurements. We also identified clusters of organisms with distinct pathway choices related to a tradeoff between thermodynamic favorability and cofactor-use efficiency. Lastly, we focused on two specific cases, isoleucine and lysine biosynthesis, to investigate how organisms lifestyles relate to the choice of biosynthetic pathways.

5.3 Results

5.3.1 Identifying biosynthetic pathway alternatives found in sequenced genomes

First, we collected the gene content of 5,203 organisms from the Kyoto Encyclopedia of Genes and Genomes (KEGG) Database of genome annotations [147]. The organisms spanned three domains of life, with major phyla including Proteobacteria (n = 2,167), Firmicutes (n = 908), Actinobacteria (n = 575), Bacteroidetes (n = 234), Euryarchaeota (n = 179), Tenericutes (n = 134), Chlamydiae (n = 118), Chordata (n = 108), and Cyanobacteria (n = 102). Based on the available genome annotations in KEGG, we obtained a total of 8,247 genes with KEGG orthology identifiers from all organisms. The list of genes corresponds to specific Enzyme Commission numbers and includes both metabolic functions and cellular processes such as assembly of macromolecules, signal transduction, etc. We found that organisms cluster by relationship on the phylogenetic tree based on their gene content (Figure 5.1a). For example, organisms in the Archaea and Eukaryota domains each belong to individual clusters, while organisms in the major phyla of the Bacteria domain (Proteobacteria, Actinobacteria, and Firmicutes) fall into separate clusters.

We then identified alternative pathways for de novo synthesis of biomass precursors using the KEGG PATHWAY and MetaCyc databases [147, 148]. The list of biomass precursors examined included amino acids, nucleotides, lipids, and certain small molecules such as vitamins and polyamines. We classified the precursors based on the types of alternative biosynthetic pathways present (Figure 5.1b). Specifically, the pathways either (i) have only one biosynthetic route for the precursor, (ii) start from the same metabolite and use the same cofactors but with different

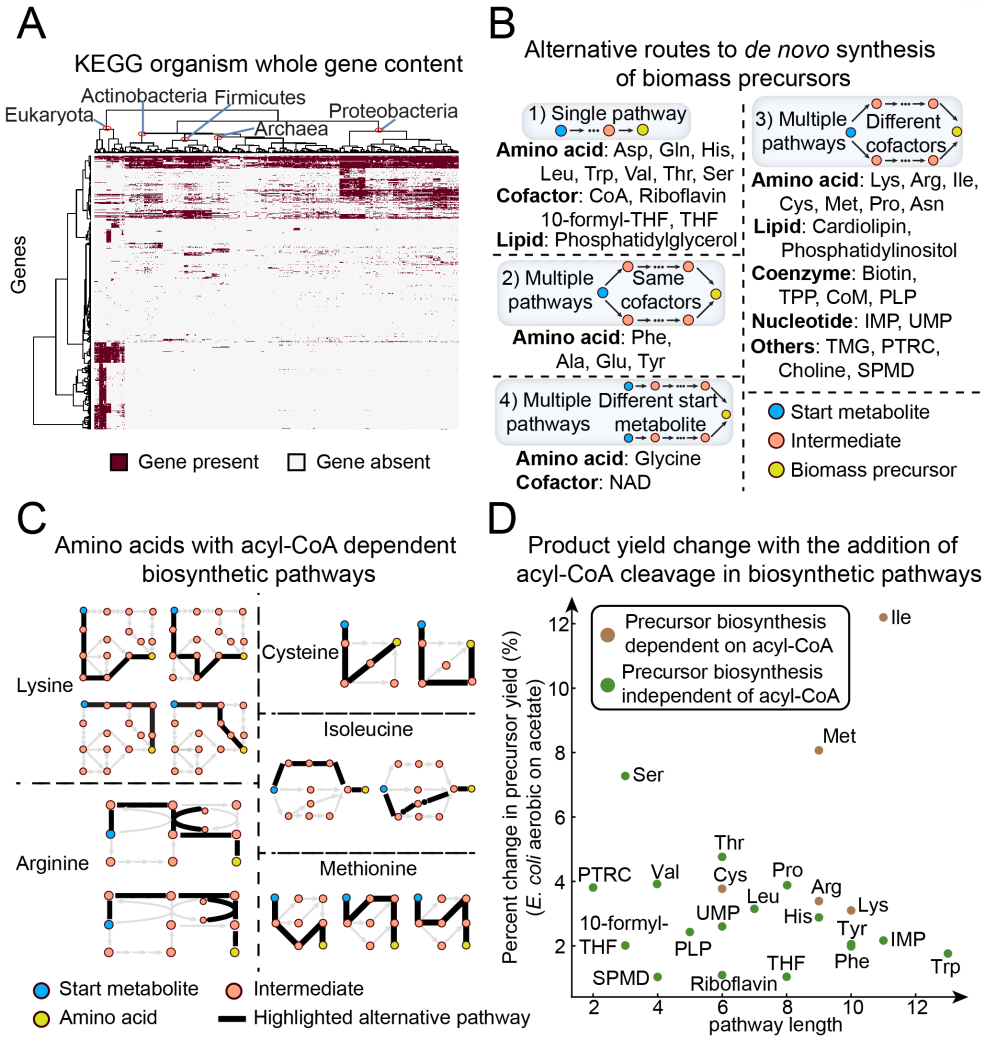


Figure 5.1: Alternative biosynthetic routes of biomass precursors. a) Hierarchical clustering on 5,203 KEGG organisms, with 8,247 genes having Enzyme Commission annotations. b) Categories of alternative routes to de novo biosynthesis of biomass precursors. For amino acids, we provided their three letter codes for simplicity. c) Amino acids with acyl-CoAdependent alternative biosynthetic pathways. d) Product yield change due to the addition of acyl-CoA cleavage in biosynthetic pathways of biomass precursors. 10-formyl-THF, 10-formyltetrahydrofolate; CoM, coenzyme M; PLP, pyridoxal phosphate; PTRC, putrescine; SPMD, spermidine; THF, tetrahydrofolate; TMG, trimethylglycine; TPP, thiamine diphosphate.

intermediate metabolites, (*iii*) start from the same metabolite and use different cofactors, or (*iv*) start from different metabolites and use different cofactors. For precursors with multiple alternative routes, we attempted to trace the pathways back until they intersect at a common starting metabolite. However, for alternative routes that reach central metabolic pathways (e.g., glycolysis and TCA cycle) but have not converged to a common starting metabolite, we considered them as having different starting points.

We found that while some biomass precursors have only a single de novo biosynthetic pathway, a large number display multiple pathways (Figure 5.1b). We distinguished between pathways that share common starting metabolites and pathways that start from different metabolites. Pathways that start from the same metabolite but have alternate routes with different cofactor usage include those for a number of amino acids (arginine, asparagine, cysteine, lysine, and methionine), nucleotides (IMP and UMP), and essential small metabolites (biotin, putrescine, spermidine, and thiamine diphosphate). These alternative pathways allowed us to control for any possible factors associated with concentrations or thermodynamics of the starting metabolites themselves when evaluating alternatives. Lastly, pathways starting from alternate metabolites were those for glycine (from 3-phosphoglycerate, glyoxylate, or oxaloacetate via threonine) and NAD (from tryptophan or aspartate).

5.3.2 Alternative pathways in amino acid biosynthesis differ by acyl-CoA cleavage and show distinct yield differences

We examined the thermodynamics [72] and cofactor use of the alternative biosynthetic pathways for biomass precursors. Pathways with lower standard transformed reaction Gibbs energies ($\Delta_{\mathrm{r}}G'^{\circ}$) are considered more thermodynamically favorable than those with higher energies.

We found that alternative pathways can vary substantially in thermodynamic favorability due to their differences in cofactor use. Examining the common cofactors involved, we found that certain cofactor pairs are prevalent in biosynthetic pathways, including those providing energy (ATP hydrolysis), those serving as the oxidizing/reducing agent (NADH/NAD and NADPH/NADP), and those donating the amino group (glutamate/α-ketoglutarate and glutamine/glutamate).

However, the use of acyl-CoA cleavage to drive biosynthetic pathways is present for only a subset of amino acids, including lysine, arginine, cysteine, isoleucine, and methionine (Figure 5.1c). Interestingly, these five amino acids have both acyl-CoA-dependent and -independent pathways present. We found the acyl-CoA-dependent pathways of these amino acids to be identical with the other alternatives in cofactor use, except for the additional acyl-CoA cleavage, which for lysine, arginine, and cysteine results in more favorable pathway thermodynamics. On the other hand, the acyl-CoA-independent pathway in isoleucine biosynthesis through threonine has lower energy than the acyl-CoA-dependent route, because it is coupled to a greater energetic cost of hydrolysis of three ATP molecules and oxidation of three equivalent NADH molecules per isoleucine produced.

We then investigated why these five amino acids have alternative biosynthetic pathways that differ by acyl-CoA use while the other biomass precursors have only acyl-CoA-independent pathways. We identified two factors contributing to the presence of acyl-CoA-dependent pathways: the pathway length in terms of reaction number and the change in precursor yield from using pathways with the additional acyl-CoA cleavage. First of all, alternative pathways are unlikely to arise when the production of the precursor takes very few steps (e.g., a single step for alanine, aspartate, glutamate, and glutamine synthesis). Additionally, a large difference in precursor yield due to the additional acyl-CoA cleavage in the pathway may benefit organisms

with certain lifestyles, thus motivating the presence of alternative pathways differing in acyl-CoA use.

We obtained biosynthetic pathway length for all biomass precursors (Figure 5.1b) from the MetaCyc database and calculated the median length for precursors with multiple alternative routes. We also compared the difference in precursor yield due to the use of acyl-CoA-dependent versus acyl-CoA-independent pathways, through simulations with organism-specific genome-scale metabolic networks (Methods). For precursors without acyl-CoA-dependent pathways present, we created pseudo pathways similar to their original pathways but with the additional acyl-CoA cleavage. Taking *E. coli* grown on acetate aerobically, for example (Figure 5.1d), we found that the five amino acids (brown dots) with acyl-CoA-dependent pathways present generally have longer pathway length and larger yield change from using the acyl-CoA-dependent pathways, with isoleucine and methionine showing the largest yield change. This trend can also be extended to a number of organisms [149] under different conditions, for which we found the yield change for the five amino acids to be significantly higher than other precursors in 24 of 43 conditions examined ($P < 0.05$). Notably, for organisms grown on acetate, we found a significant difference in 11 of 13 conditions examined. In contrast, when examining yield change in pathways differing by the use of ATP hydrolysis, we found only 6 of 43 conditions to have a significantly higher yield change for precursors with ATP-dependent alternative routes compared to those without such alternates ($P < 0.05$). Therefore, we demonstrate that alternative pathways differing by the use of acyl-CoA show significant yield differences, as is the case for the five amino acids.

5.3.3 *E. coli* uses thermodynamically-favorable but cofactor-use-inefficient amino acid biosynthetic pathways

We sought to further compare the acyl-CoA-dependent and -independent alternative pathways for the five amino acids using *E. coli* , taking advantage of its well-curated metabolic network and abundant quantitative physiological data available. Specifically, we focused on two aspects: pathway thermodynamic favorability and cofactor-use efficiency. *E. coli* uses acyl-CoA-dependent pathways for biosynthesis of four of the five targeted amino acids, with the exception being isoleucine. Compared with pathways that are not present in *E. coli* , we found that the pathways used by *E. coli* are thermodynamically more favorable in each case in terms of intrinsic pathway energy (i.e., lower standard Gibbs energy, $\Delta_{\mathrm{r}}G'^{\circ}$, which does not take into account metabolite concentrations). We then calculated transformed Gibbs energy ($\Delta_{\mathrm{r}}G'$) values for each pathway using measured quantitative metabolomics data of *E. coli* [28, 150, 151] and verified that these pathways are indeed substantially further from equilibrium (more negative $\Delta_{\mathrm{r}}G'$) (Figure 5.2a).

We next calculated the ATP equivalent cost of the pathways to evaluate the cofactor-use efficiency of the pathways (Methods). A high ATP cost of the pathway corresponds to a low efficiency in cofactor use. We found that the ATP equivalent costs of the pathways used by *E. coli* (using glucose aerobically) are greater than those of the alternative pathways in four of five cases (Figure 5.2b), indicating that *E. coli* uses cofactor-use-inefficient pathways. This result was further confirmed by the fact that pathways present in *E. coli* have lower product yield than those not present in *E. coli* (Figure 5.2c).

Thermodynamically favorable pathways can be beneficial in terms of protein cost, as the enzyme level required to achieve a given flux can increase dramatically for reactions near equi-

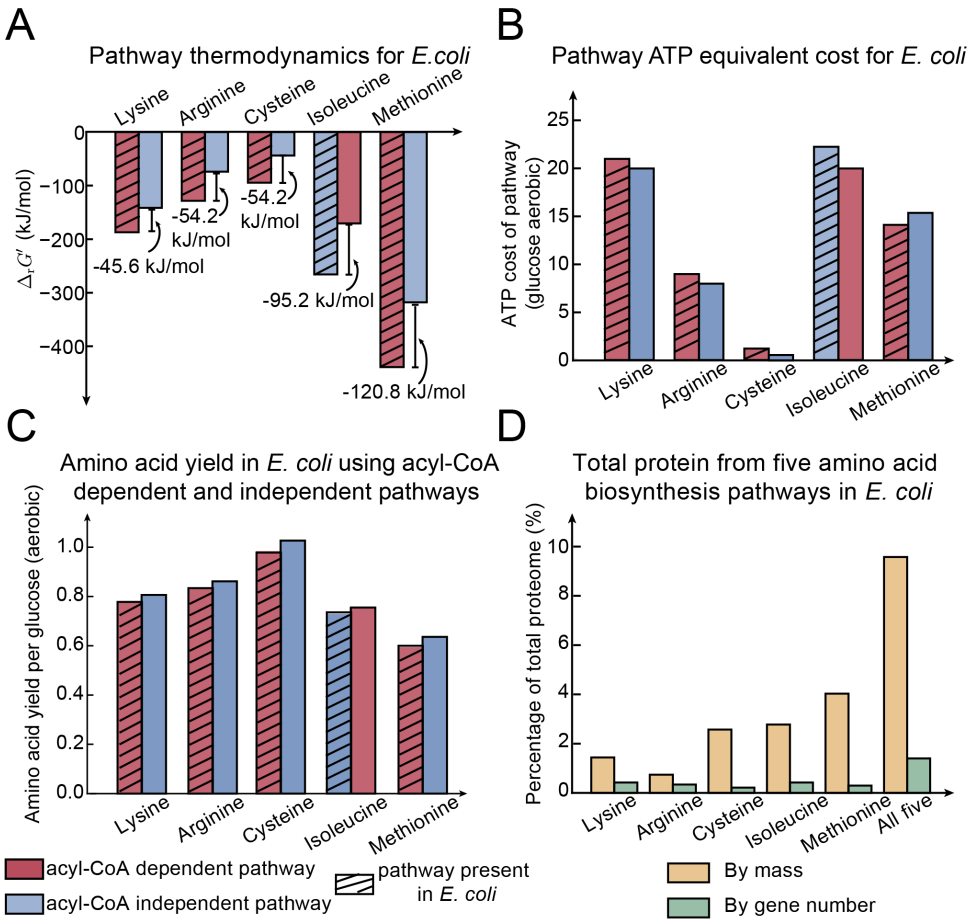


Figure 5.2: Thermodynamics and cofactor-use efficiency of alternative biosynthetic pathways in *E. coli* . a) Gibbs energies of reaction ($\Delta_r G'$) of acyl-CoA-dependent and -independent alternative pathways using *E. coli in vivo* metabolite concentrations. b) ATP equivalent cost of acyl-CoA-dependent and -independent pathways calculated from an *E. coli* metabolic model grown on glucose aerobically. c) Amino acid yield from acyl-CoA-dependent and -independent pathways simulated using the *E. coli* metabolic model grown on glucose aerobically. d) The proteins from five amino acid biosynthesis pathways in terms of fraction in mass and gene number of the total proteome in *E. coli* .

librium [102]. Therefore, it is possible that organisms already with significant resources invested in synthesizing pathway proteins select the thermodynamically favorable routes for efficiency in protein use. We found evidence supporting this hypothesis using *E. coli* proteomics data [152], wherein the proteins required for biosynthesis of each of the five amino acids occupy a higher fraction of the whole *E. coli* proteome by mass compared to gene number (Figure 5.2d). Together, the proteins from all five amino acid biosynthesis pathways occupy 10% of the proteome by mass, while only 2% by number of genes.

To summarize, we found that tradeoffs between thermodynamic favorability and cofactor-use efficiency exist in pathway alternatives. In the case of *E. coli* , the use of thermodynamically more favorable pathways may improve the efficiency of pathway protein use.

5.3.4 Distinct acyl-CoA-dependent pathway choices exist among organisms

To understand the underlying factors for alternative pathway choice, we first clustered the organisms based on their presence/absence information of alternative pathways for the five amino acids. We found that the organism clusters did not separate cleanly by phylogeny (Figure 5.3a), suggesting that factors other than phylogenetics may underlie the choice of alternative pathways. However, we observed interesting patterns when examining the use of acyl-CoA-dependent pathways for the five amino acids among organisms. For each amino acid, we separated the KEGG organisms into three categories: (*i*) those containing only acyl-CoA-dependent pathway(s); (*ii*) those containing only acyl-CoA-independent pathway(s); and (*iii*) those containing both acyl-CoA-dependent and -independent pathways. For methionine biosynthesis, we labeled the pathway using two acyl-CoA molecules as acyl-CoA dependent, and the pathways using only one as acyl-CoA independent.

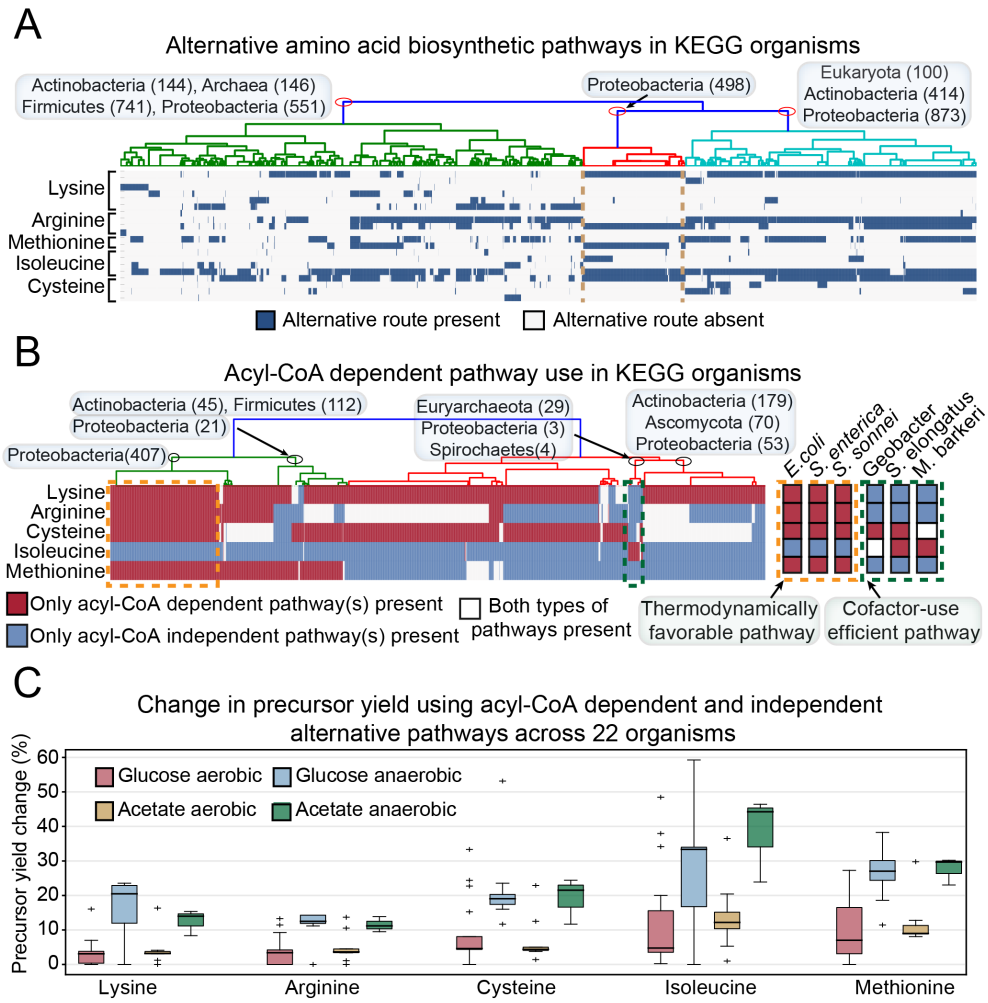


Figure 5.3: Alternative amino acid biosynthetic pathways in organisms. a) Hierarchical clustering on 5,203 KEGG organisms based on the presence of alternative biosynthetic pathways for the five amino acids. b) Hierarchical clustering of KEGG organisms based on their use of acyl-CoA-dependent pathways. We show two clusters of organisms: one mostly uses thermodynamically favorable pathways (yellow dashed box), and the other mostly uses cofactor-efficient pathways (green dashed box). (C) Change in precursor yield using the acyl-CoA-dependent and -independent alternative pathways under different conditions for the five amino acids. We predicted the product yield change across 22 organisms, using their available genome-scale metabolic models.

Examining patterns of pathway use within organisms, we did not find any organism choosing acyl-CoA-dependent pathways for all five amino acid biosynthesis pathways, but for only a selection of them. As we clustered the organisms based on the type of acyl-CoA pathways used for the five amino acids, we found that the pathway choice did not break down cleanly by phylogeny. Further analysis on the metabolic genes related to the use of acyl-CoA pathways also shows complex traits in metabolic functions, indicating that nonspecific factors, such as lifestyle or organism history, may underlie the acyl-CoA pathway use broadly.

On the other hand, we identified groups of organisms with distinct pathway choices. We found one cluster containing *E. coli* and other Gammaproteobacteria (Figure 5.3b, yellow box) for which the choice of acyl-CoA-dependent pathways is the same as in *E. coli* . This cluster represents a set of organisms choosing thermodynamically favorable pathways. We also identified a different cluster of organisms that select cofactor-use-efficient pathways instead, including *Geobacter metallireducens*, the methanogen *Methanosarcina barkeri*, and the cyanobacterium [*Synechococcus elongatus* (Figure 5.3b, green box)]. These two opposing clusters indicate that tradeoffs between efficiency in product yield and proteome cost in biosynthesis may widely exist in organisms' pathway choice, perhaps with similar underlying principles to the recent observation that cellular overflow metabolism results from the balance between efficient pathway yield and efficient protein use [153].

A closer look at the lifestyles of these two organism clusters shows that organisms favoring thermodynamic-favorable pathways generally depend on complex carbon sources with aerobic respiration, while organisms favoring cofactor-use-efficient pathways depend on simple carbon sources with anaerobic respiration. To understand how different lifestyles affect the product yield, we attempted to compare the yield change from using the alternative acyl-CoA-dependent

pathways under four growth conditions: glucose aerobic, glucose anaerobic, acetate aerobic, and acetate anaerobic. We used the curated genome-scale metabolic models of a total of 22 organisms for simulations [149, 154, 155]. It is worth noting that not all 22 organisms examined here fall into the two organism clusters identified above. We found that anaerobic respiration results in the most significant change in precursor yield for all five amino acids (Figure 5.3c), possibly due to the different cofactor cost and availability under different respirations. On the other hand, the carbon sources do not seem to significantly affect the product yield change.

5.3.5 Trade-off between pathway thermodynamic favorability and efficiency of cofactor use underlies organisms' pathway choice for isoleucine biosynthesis

To understand the choice of alternative isoleucine biosynthesis pathways among various organisms, we focused on two alternative pathways and compared their properties in terms of thermodynamic favorability and cofactor-use efficiency. The first pathway uses threonine as the intermediate (Figure 5.4a, green) and is present in a large number of organisms from the Bacteria and Eukarya domains. The second pathway uses citramalate as the intermediate (Figure 5.4a, red) and is typically present in Archaea but is also found in bacteria from the Spirochaetes phylum [156, 157]. A recent study showed that both pathways are present in *Geobacter* spp., which primarily uses the one through citramalate [158]. We selected organisms from each category described above, including *E. coli* (contains pathway through threonine), *M. barkeri* and *Methanosarcina acetivorans* (contain pathway through citramalate), and *Geobacter sulfurreducens* and *G. metallireducens* (contain both pathways but mainly use the one through citramalate).

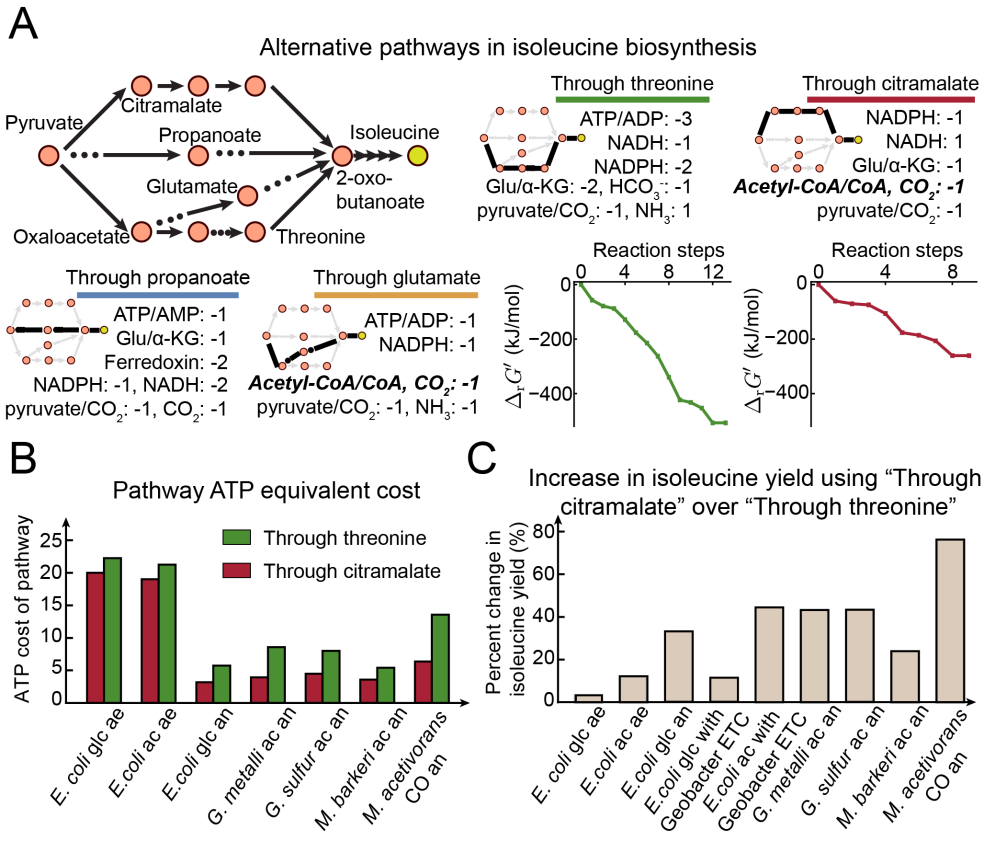


Figure 5.4: Alternative pathways for isoleucine biosynthesis. a) Sketches of four alternative isoleucine biosynthetic pathways with their cofactor usage. We also included the Gibbs energies ($\Delta_r G'$), considering metabolite concentrations for each reaction step in the pathways through threonine and citramalate. b) Cofactor use in terms of ATP cost for "Through threonine" and "Through citramalate" pathways in different organisms. The organism name, the substrate used, and the respiration type are labeled for each simulated organism condition. c) Increase in isoleucine yield using the "Through citramalate" pathway compared with "Through threonine" for different organisms. ac, acetate; ae, aerobic; an, anaerobic; CO, carbon monoxide; glc, glucose; Glu/α-KG, glutamate/α-ketoglutarate.

Although the standard energies ($\Delta_r G'^{\circ}$) of the pathway through threonine are significantly lower than those of the citramalate pathway across different conditions, we further compared the $\Delta_r G'$ of both pathways by taking metabolite concentrations into account. Using the quantitative metabolomics data of *E. coli* [28, 150, 151], we calculated the reactionwise energy profile for each pathway (Methods) and confirmed that the overall $\Delta_r G'$ of the pathway through threonine is much lower than that through citramalate (Figure 5.4a).

Using the available genome-scale metabolic models [74, 154, 155, 159, 160], we calculated the ATP equivalent cost of the two pathways for five organisms with their respective carbon sources and types of respiration (Figure 5.4b) (Methods). We used both glucose and acetate as the substrates for *E. coli* ; the latter is a common carbon source for the other four organisms. We also allowed both aerobic and anaerobic respirations for *E. coli* , although only the latter is possible for the other four organisms. We found that the pathway through threonine is always more costly in cofactor use compared with the one through citramalate (Figure 5.4b), while being thermodynamically more favorable.

To examine the possible benefit of using the cofactor-efficient pathway in *E. coli* , we inserted the citramalate pathway into the *E. coli* metabolic model. We observed marginal improvement (3.6%) in isoleucine yield per mole of glucose when comparing the two pathways for *E. coli* (Figure 5.4c). On the other hand, we observed a relatively large improvement in isoleucine yield using the citramalate pathway for organisms dependent on simple carbon sources (such as acetate) and grown anaerobically (*Geobacter* spp., 43.2% and 43.4%; methanogens, 23.9% and 76.2%) (Figure 5.4c). To examine whether the carbon source or anaerobic respiration contributes the most to such a large difference in yield change between *E. coli* and the other four organisms, we performed the same calculations on *E. coli* grown on acetate aerobically, glucose anaerobi-

cally, glucose using the electron transport chain (ETC) from *Geobacter* spp., and acetate using the ETC from *Geobacter*. We found the yield change to be most significant under anaerobic respiration (34.7% for *E. coli* glucose anaerobic). Interestingly, the isoleucine yield change is almost identical for *E. coli* with the *Geobacter* ETC and grown on acetate. We obtained similar results for a larger set of organisms shown in Figure 5.3c, whereby under anaerobic respiration, the product yield changes the most between acyl-CoA-dependent and -independent pathways.

Thus, the citramalate pathway is more beneficial for organisms under anaerobic respiration and leads to much greater isoleucine yield. In contrast, for *E. coli* that can utilize aerobic respiration, the threonine pathway with greater thermodynamic favorability is selected over the citramalate pathway, which brings marginal benefit in terms of product yield.

5.3.6 Lysine biosynthesis in thermophiles shows differential temperature dependence of thermodynamics

Examining the change of pathway thermodynamics with respect to temperature, we found that $\Delta_r G'^{\circ}$ values of the diaminopimelate pathways (M00016, M00525, M00526, and M00527) generally increase with temperature, while $\Delta_r G'^{\circ}$ values of the 2-aminoadipate pathways (M00030 and M00031) decrease with respect to temperature (Figure 5.5b). These trends lead to the intersection of $\Delta_r G'^{\circ}$ values at high temperature between pathway M00031 and pathways M00016 and M00526. Notably, organisms with pathway M00031 have culture temperature around the crossing point (80 to 100 °C) (Figure 5.5c), at which the thermodynamics between the three lysine biosynthesis branches are almost equivalent. These organisms living at high temperature are called thermophiles. In contrast, organisms with pathways M00016 and M00526 have much lower culture temperature and are typically termed mesophiles (Figure 5.5c). We also note that

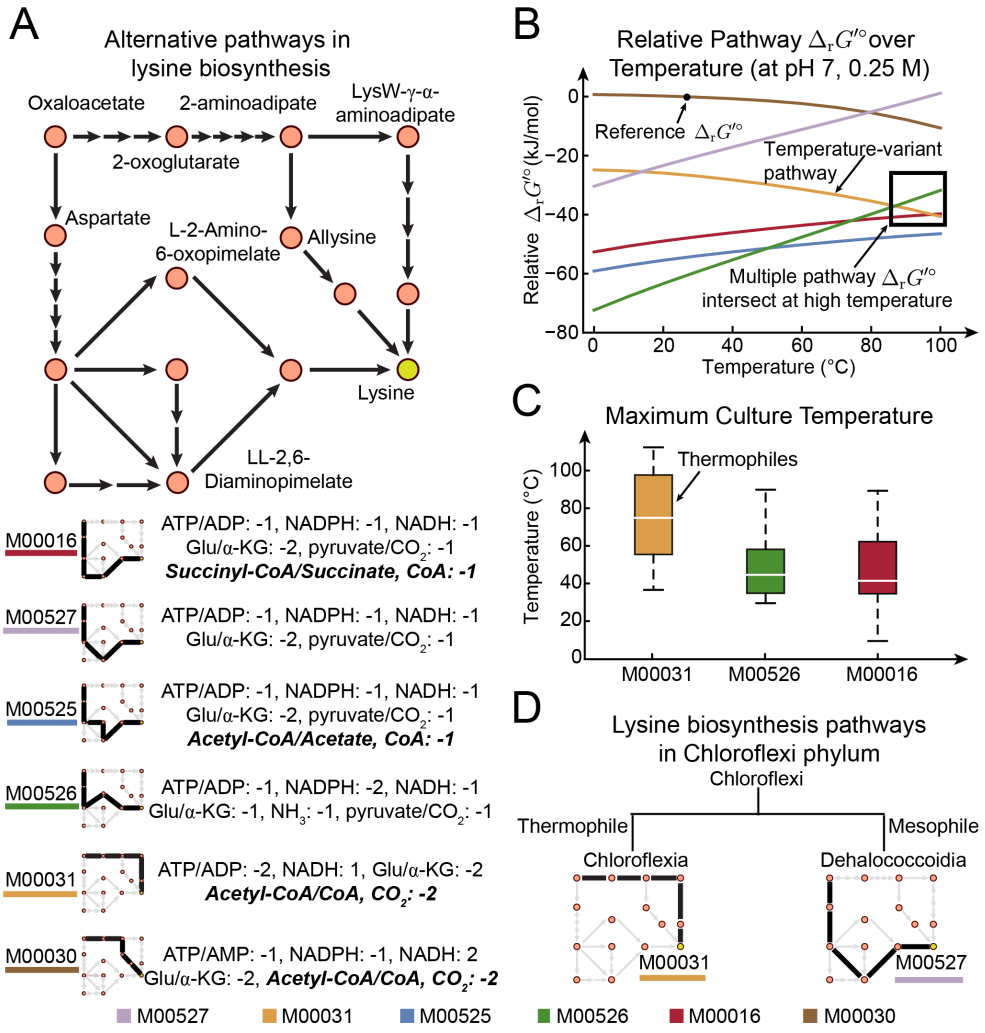


Figure 5.5: Alternative pathways for lysine biosynthesis. a) Sketches of six alternative lysine biosynthetic pathways (with KEGG identifiers), along with their cofactor usage. b) Thermodynamics of alternative pathways as a function of temperature. The pathway $\Delta_r G'^\circ$ values relative to that of the pathway with the highest $\Delta_r G'^\circ$ at 298.15 K, pH 7, and ionic strength 0.25 M (reference $\Delta_r G'^\circ$). (C) Maximum culture temperature of organisms containing pathways M00031, M00526, and M00016. (D) Alternative lysine biosynthetic pathways in Chloroflexi phylum. Glu/α-KG, glutamate/α-ketoglutarate.

the organisms containing these three pathways did not show an appreciable difference in culture pH, with medians around pH 7. Therefore, at high temperature, pathway M00031 might become less disadvantageous in terms of thermodynamics compared with the other pathways and more likely to be selected by thermophiles. On the other hand, mesophiles still favor pathways with greater thermodynamic favorability, such as M00016 and M00526.

Interestingly, we found that organisms in the same phylum select different lysine biosynthesis pathways, and such choice is correlated with organism culture temperature. Specifically, for organisms in the Chloroflexi phylum, those from the Chloroflexia class are thermophiles with the M00031 pathway, while those from the Dehalococcoidia class are mesophiles with the M00527 pathway (Figure 5.5d). We found that these two pathways have relatively similar thermodynamics at low temperature, but the M00031 pathway is much more favorable at high temperature than the M00527 pathway (Figure 5.5b). This difference in thermodynamic favorability may explain the choice of thermophiles and mesophiles between these two pathways. Additionally, the results demonstrate how factors other than phylogeny can affect pathway choice, whereby organisms that are close in phylogenetic distance select different lysine biosynthesis pathways due to their different environmental conditions.

5.4 Discussion

In this work, we examined the basis for the presence of alternative biosynthetic routes for biomass precursors. We showed that acyl-CoA-dependent biosynthetic pathways are only present for five amino acids and investigated the possible factors related to the presence of acyl-CoA-dependent alternative routes. We evaluated the tradeoff between thermodynamic favorability and cofactor-use efficiency of the biosynthetic pathways and identified two clusters of organisms

with distinct pathway choices. We found that organism living environment, rather than inherent metabolic capabilities, was the driving factor for alternative pathway choice. Specifically, organisms normally under aerobic respiration benefit from the thermodynamically more favorable routes, while organisms under anaerobic respiration benefit from cofactor-efficient routes, which are usually more advantageous in product yield. Additionally, we showed that organisms living at different temperatures can select alternative lysine biosynthesis pathways.

Examination of the thermodynamics of alternative pathways revealed that many pathways show high- and low-energy routes that are both prevalent among organism genomes. It may be argued that these pathways could have come about through evolution with insufficient selection to distinguish between these energetic alternatives. However, the presence of high- and low-energy alternatives with consistent cofactor structure (e.g., using acyl-CoA cleavage in amino acid biosynthesis), would suggest that these energetic alternatives may exist by selection. Because studies have demonstrated the advantage of thermodynamic favorability in increasing efficiency in protein use [102], a hypothesis emerges for why this might be the case. Presumably, the cell can choose to "waste" additional cofactor for higher thermodynamic favorability, reducing the downstream resources allocated for unnecessary pathway protein synthesis. While the current study examined only *E. coli* due to the limited organism-specific proteomics data available, more quantitative proteomics data across organisms and conditions may be used to probe this hypothesis.

We have found that organism lifestyle, such as type of respiration and temperature, can affect the alternative pathway choice for amino acid biosynthesis. Organisms in poor/anaerobic environments often choose cofactor-use-efficient pathways, which lead to significant yield improvement under those conditions. While those pathways may or may not depend on acyl-CoA

cleavage, it is interesting to note that acyl-CoA is the distinguishing cofactor between the pathway alternatives. The underlying reason might be related to the cost of acyl-CoA under different conditions. Furthermore, while the improvement in growth rate can be small when using the cofactor-use-efficient pathways, such a benefit may not be relevant to organisms living under nutrient-poor and anaerobic conditions. For example, *Geobacter* spp. may rarely encounter favorable conditions to achieve maximum growth rate and, thus, would not select thermodynamically favorable routes to achieve better efficiency in protein use. In another specific case, thermophiles select pathways that become thermodynamically favorable at higher temperature in lysine biosynthesis. This observation suggests the possibility that certain alternatives may become viable when they become thermodynamically equivalent to other pathways under certain environmental conditions. This work complements known adaptations to high temperature, such as increase in protein stability [161] and alteration in membrane compositions [162].

Together, these results show how alternative pathway choice can be related to organism lifestyle, due to the tradeoff in thermodynamic favorability and cofactor-use efficiency. This study is one of a number of recent efforts aimed at discovering connections between thermodynamics and constraints on metabolic pathways. For example, one study showed that autotrophic amino acid synthesis was exergonic under the conditions in hydrothermal vents, rather than endergonic at surface conditions [130]. Another recent effort looked at thermodynamic bottlenecks and proteomic constraints underlying the use of the Entner-Doudoroff pathway [102], indicating that similar tradeoffs in protein efficiency can be observed in central metabolism. As methods for estimating the thermodynamic properties of metabolic networks continue to be refined and as genome annotations continue to improve, these efforts are likely to continue to reveal the physical constraints underlying the adaptation and evolution of metabolic networks to meet organisms'

lifestyles.

5.5 Methods

The specific procedure for collecting the information of KEGG organisms and alternative biosynthetic pathways is described in online SI Appendix, Supplementary Information Text (https://www.pnas.org/content/115/44/11339). The workflow for calculation of pathway thermodynamics and product yield using metabolic network reconstructions can also be found in online SI Appendix, Supplementary Information Text.

Acknowledgments

This work was supported by the Novo Nordisk Foundation Grant NNF10CC1016517.

Chapter 5 in full is a reprint of material published in: **Bin Du**, Daniel C Zielinski, Jonathan M Monk, Bernhard O Palsson. 2018. "Thermodynamic favorability and pathway yield as evolutionary tradeoffs in biosynthetic pathway choice." *Proceedings of the National Academy of Sciences*, 115(44), 11339-11344. The dissertation author was the primary author.

Chapter 6

Adaptive laboratory evolution of *Escherichia coli* under acid stress

6.1 Abstract

The ability of *Escherichia coli* to tolerate acid stress is important for its survival and colonization in the human digestive tract. Here, we performed adaptive laboratory evolution of the laboratory strain *E. coli* K-12 MG1655 at pH 5.5 in glucose minimal medium. By 800 generations, six independent populations under evolution reached 18.0% higher growth rates than their starting strain at pH 5.5, while maintaining comparable growth rates to the starting strain at pH 7. We characterized the evolved strains to find that: (1) whole genome sequencing of isolated clones from each evolved population revealed mutations in *rpoC* appearing in 5 of 6 sequenced clones; (2) gene expression profiles, using RNA-seq on two selected acid-adapted strains, revealed different strategies to mitigate acid stress, that are related to amino acid metabolism and energy production and conversion; Thus, a combination of adaptive laboratory evolution,

genome resequencing, and expression profiling reveals, on a genome-scale, the strategies that *E. coli* deploys to mitigate acid stress.

6.2 Background

As a commonly found enteric bacteria species in the human digestive tract, *Escherichia coli* is known to withstand various levels of acid stress [163–168]. For example, *E. coli* can survive several hours under pH 2 [163], which is within the range of the extremely acidic stomach (pH 1.5 to 3) that serves as the barrier against most bacteria [169]. Additionally, *E. coli* has been shown to grow under mild acid stress [166–168], which is typically found in the human intestinal tract [169, 170]. Such adaptability to low pH environments has raised wide interest in understanding the underlying mechanisms that protect *E. coli* from acid stress. Furthermore, studying the acid resistance mechanisms of *E. coli* has important implications in the food and health care industry. For example, treatment strategies can be developed to target specific acid resistance mechanisms in the case of a pathogenic *E. coli* infection.

The acid resistance mechanisms of *E. coli* have been studied extensively. To maintain the intracellular pH homeostasis, *E. coli* has developed various strategies including cytoplasmic buffering [171], proton-consuming systems [172–175], adjustment of cellular metabolism [176, 177], and physiological responses [178–182]. The buffering capacity of the cytoplasm mainly comes from inorganic phosphates, amino acid side chains, polyphosphates, and polyamines [171]. The proton-consuming systems include four types of amino acid decarboxylase systems that function under different pHs and formate hydrogen lyase that is active under anaerobic conditions [183]. The metabolic responses under acid stress include the up-regulation of components in the electron transport chain and metabolism of sugar derivatives that have decreased acid production

compared to glucose [176, 177]. The physiological responses include the activation of periplasmic chaperones HdeA and HdeB [178], adjustment of membrane lipid compositions [179, 180], and blockage of outer membrane porins [181, 182].

Adaptive laboratory evolution (ALE) is an important scientific approach for understanding the adaptive response of microorganisms under particular environments or after specific perturbations [184]. During an ALE experiment, the microorganism is cultured under defined conditions for an extended period of time. ALE allows the selection of improved phenotypes, typically the growth rates, under certain growth environments. Furthermore, the advancement of next-generation sequencing technology makes it convenient to obtain the genotypes underlying the favorable traits over the course of evolution [185]. A previous study also investigated the adaptive evolution of *E. coli* under acid stress, where *E. coli* K-12 W3110 was evolved in a nutrient rich environment (LBK medium) buffered at pH range 4.6 - 4.8 for 2000 generations [186, 187]. Here, we are interested in the adaptive evolution of *E. coli* under acid stress in a nutrient limited environment, where glucose is the only carbon source.

In this study, we perform ALE on *E. coli* at pH 5.5 in glucose minimal medium. For the evolved strains, we use whole genome sequencing to identify genetic mutations that arise over the course of evolution. Additionally, to examine the change in gene activity after evolution, we perform RNA-seq to characterize the gene expression profile of the evolved endpoints when growing under different pHs. We then identify the differentially expressed genes (DEGs) of the evolved endpoints at different pHs. We also uncover new cellular processes that emerge over the adaptive evolution under acid stress, using DEGs identified in the starting strain across pH as a reference.

6.3 Results

6.3.1 Laboratory evolution and acid-adapted endpoint strains

We used wild-type *Escherichia coli* K-12 MG1655 that had been previously evolved on M9 glucose minimal medium as the starting strain for evolution and refer to it as GLU strain [188]. We used GLU as the starting strain to isolate changes due to adaptation to acid stress from those caused by adaptation to the culture medium. The genetic mutations of the GLU strain against *E. coli* K-12 MG1655 is also documented in ALEdb (aledb.org) under the experiment name GLU. Six independent cultures were established under pH 5.5 in glucose minimal medium, buffered with 150 mM 2-(N-morpholino) ethanesulfonic acid (MES) (pKa = 6.1) [189]. In addition, we lowered magnesium (Mg) concentration in the media to 0.2 mM to minimize precipitations. We refer to six acid-adapted strains as AA1 to AA6, respectively. To account for the possible effects due to changes in media composition, we also set up two independent cultures under pH 7 with 150 mM MES buffer (MES1, MES2) and lowered Mg concentration (LM1, LM2), respectively. All of the strains used in this study and their relationships can be found in Figure 6.1a.

We performed ALE using an automated system, which tracked culture growth rates and passed the cells to fresh media when OD600 measurements reached 0.3 to ensure selection at exponential-phase growth. Additionally, we periodically measured the pH of the clean media and recently passaged cultures to ensure proper buffering. The culture pH remained relatively stable for strains evolved in MES buffer under pH 5.5 and 7. For strains evolved under lowered Mg concentration and with only phosphate buffers ($pK_a = 7.2$) in glucose minimal medium, the culture pH dropped significantly at the end of the culture, likely due to the secretion of organic acids during growth. The laboratory evolution process lasted 35 days for strains AA1 to AA6

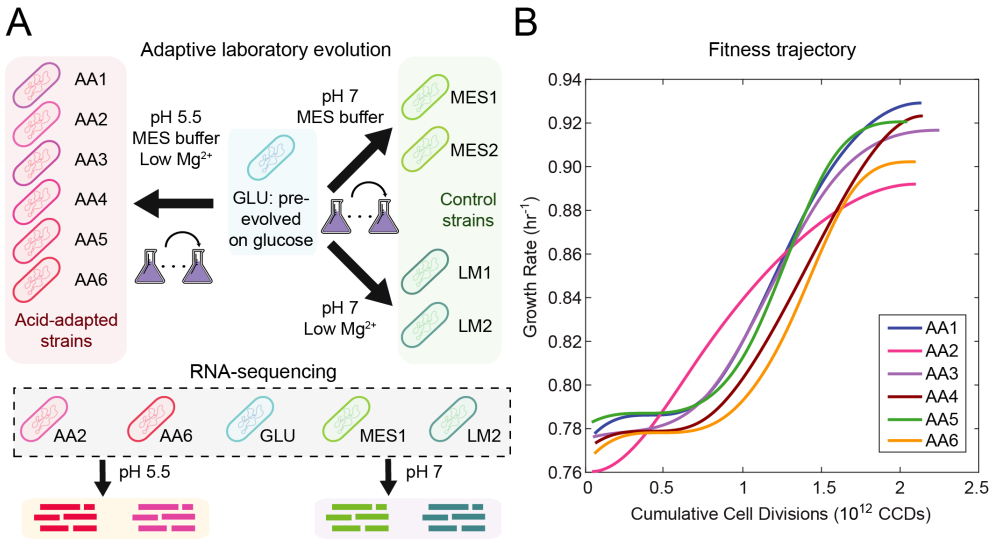


Figure 6.1: Adaptive laboratory evolution (ALE) of *E. coli* under acid stress. a) Schematic for ALE process and strains used for RNA-sequencing. Starting from the GLU strain, we performed ALE under pH 5.5 to obtain six acid-adapted (AA) strains and under pH 7 to obtain four control strains. Using RNA sequencing, we obtained the gene expression profiles of five selected strains under pH 5.5 and 7. b) Smoothed fitness trajectories of six acid-adapted strains. We show here the change of growth rate over cumulative cell divisions through the evolution process. The average growth rate improvement is 18%.

under pH 5.5, corresponding to 800 generations and 2.1×10^{12} cumulative cell divisions (CCD). The fitness trajectories of the evolved strains are shown in Figure 6.1b. We observed the growth rates to continuously improve over CCDs and approach stable values at the end of the evolution. Overall, we found the evolved endpoints to have an average of 18.0% improvement in growth rate (from 0.77 ± 0.01 hour^{-1} to 0.91 ± 0.01 hour^{-1}) over their starting strain.

To evaluate the fitness of acid-adapted strains against the starting GLU strain, we obtained the growth rates of the strains under pH 5.5 and 7 in a separate experiment. We found the acid-adapted strains to have increased fitness under pH 5.5, with growth rate at 0.83 ± 0.01 hour^{-1} compared to the growth rate at 0.67 ± 0.02 hour^{-1} of the GLU strain. We also found the growth rates under pH 7 for acid-adapted strains to be 1.00 ± 0.01 hour^{-1}, slightly higher than that of the GLU strain at 0.94 ± 0.03 hour^{-1}.

Table 6.1: Converged mutations identified in the clones of acid-adapted strains under pH 5.5

Gene	Mutation	Protein change	Flask number	AA1	AA2	AA3	AA4	AA5	AA6
rho	C→A	R102S (CGC→AGC)	87	X					
rho	C→T	R102C (CGC→TGC)	**111**						X
rpoC	C→A	A397E (GCG→GAG)	**111**		X				
rpoC	G→C	G444A (GGT→GCT)	88,**118**			X			
rpoC	Δ7 bp	coding (4106-4112 nt)	**113**				X		
rpoC	Δ1 bp	coding (4111 nt)	**111**					X	
rpoC	C→T	S539F (TCT→TTT)	**111**						X
nagA	G→C	S90* (TCA→TGA)	84				X		
nagA	C→T	R149H (CGT→CAT)	83						X

6.3.2 Genetic mutations of the evolved strains

To understand the genetic basis of the observed phenotypic change, we performed whole genome sequencing on individual clones picked from acid-adapted strains AA1 to AA6, as well as the control strains MES1, MES2, LM1, and LM2. We identified the genetic mutations of the evolved strains by comparing them to the reference genome using *breseq* computational pipeline v0.33 (Methods) [190]. We reported the converged mutations of acid-adapted strains in Table 6.1. Converged mutations are mutations on the same gene identified across multiple strains from independent cultures.

Overall, we found a total of 22 mutations in all acid-adapted strains, including those of clones picked from the endpoints (Table 6.1 bold flask number) and midpoints of the evolution (Table 6.1). Notably, we observed mutations in *rpoC* to appear in 5 of the 6 endpoint clones. The gene product of *rpoC* is a subunit of RNA polymerase, which is known to act as a global regulator for gene expressions [191, 192]. The mutations on *rpoC* include both SNPs and deletions. We found these mutations to be located on the interaction interfaces of the protein product. Specifically, mutation with protein change A397E (Table 6.1) is located at the exit gate of the newly synthesized RNA strand. The region with mutation G444A interacts with rpoB

subunit and the region with mutation S539F interacts with rpoA subunit (Table 6.1). The two deletion mutations are located at the interaction interface with rpoZ subunit (Table 6.1). Mutations in *rpoC* in *E. coli* have been found in several previous ALE experiments, covering a variety of experimental conditions or perturbations, e.g. high temperature, alternating substrate, gene knockouts [193–195]. Several studies have suggested mutations in *rpoC* to be mainly associated with improvement in metabolic efficiency and growth rate [196–198].

The other converged mutations are found in *rho* (transcription regulation) and *nagA* (metabolism of N-acetyl-D-glucosamine) (Table 6.1). The mutations on these two genes are all SNPs. Unlike *rpoC* where mutations are found in endpoint clones, mutations in *rho* appear in the midpoint clone of strain AA1 and endpoint clone of strain AA6. Mutations in *nagA* appear in midpoint clones of strain AA4 and AA6 and are not found in the endpoint clones of these two strains. For the rest of the mutations observed in acid-adapted strains, each of them appear only in a single strain. These mutations appeared in both the coding regions and intergenic regions. The types of mutations include SNPs, deletions, and insertion elements.

We found distinct patterns when examining mutations in control strains evolved under different conditions. We discovered SNPs on *oxyR* gene to be the converged mutations in strains LM1 and LM2. On the other hand, the converged mutations for strains MES1 and MES2 are found in the intergenic region of *ilvL* and *ilvX*. Notably, the exact same mutation between *ilvL* and *ilvX* is also found in the endpoint clone of acid-adapted strain AA5, confirming the possible effect of the MES buffer during the evolution process.

6.3.3 Differential gene expression of the evolved endpoints at different pHs

To understand how the mutations can affect gene products in evolved strains, we used RNA-seq to examine the gene expressions of ALE endpoints. We selected two acid-adapted strains for RNA-seq, strain AA2 which has a single mutation in *rpoC* and strain AA6 which has the most number of mutations among all AA strains (Table 6.1). We selected strains MES1 and LM2 for different control conditions. We performed RNA-seq and obtained gene expression profiles of the selected strains as well as the starting GLU strain grown under pH 7 and pH 5.5 (Figure 6.1a; Methods). We then analyzed the expression profiles on the level of individual genes and their related cellular processes. Specifically, we performed statistical tests to identify the DEGs of the same strain growing under pH 5.5 compared to pH 7 (Methods).

To ensure the DEGs identified for acid-adapted strains across pH are only due to the effect of adaptive evolution under acid stress, we first need to understand the response to acid stress of the starting strain and also control for possible variations in culture medium during the evolution process. Therefore, we examined the DEGs across pH for the GLU strain, as well as the MES1 strain and LM2 strain to account for the possible effects due to MES buffer and lowered magnesium concentration. We found significant overlap of upregulated genes involved in cell wall/membrane biogenesis and translation processes among the acid-adapted strains and the control strains, implicating these two cellular processes as the common acid resistance mechanisms in *E. coli* .

We then examined DEGs in the acid-adapted strains under pH 5.5 compared to pH 7, after removing DEGs also found in GLU and the control strains. We found 183 genes to be differentially expressed for strain AA2 (111 upregulated and 72 downregulated) and 40 genes for strain AA6 (38 upregulated and 2 downregulated). Of those genes, we found 14 upregulated

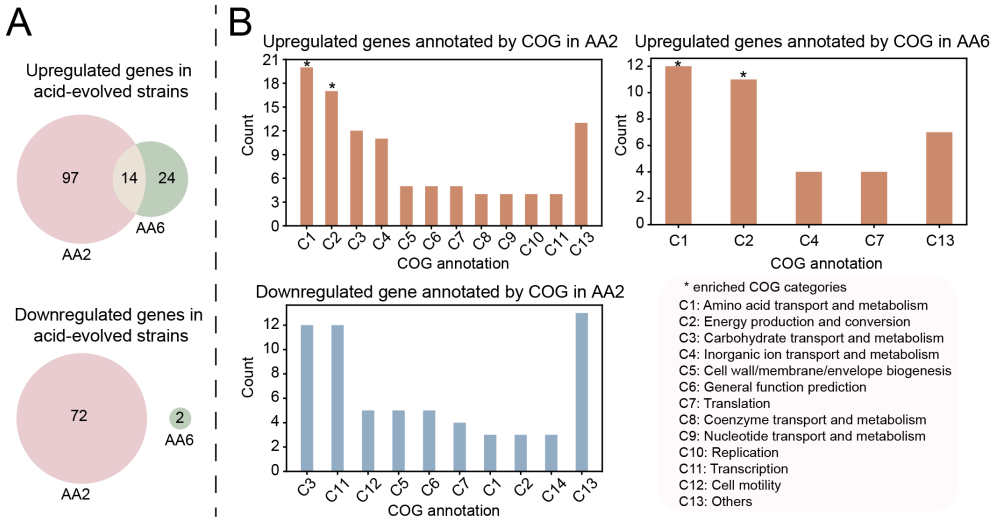


Figure 6.2: Differentially expressed genes (DEGs) of acid-adapted strains at different pHs. The DEGs are calculated for the same strain by comparing its gene expression profiles when growing under pH 5.5 and pH 7. a) Number of upregulated (top panel) and downregulated genes (bottom panel) in acid-adapted strains AA2 and AA6. We found AA2 and AA6 to share 14 upregulated genes. b) COG categories in the upregulated and downregulated genes in acid-adapted strains. The asterisk sign on top of the bar means that the COG category is enriched, as calculated using a hypergeometric test (Methods). We did not show the COG categories for downregulated genes in strain AA6 since there are only two of them.

genes that appeared in both acid-adapted strains (Figure 6.2a). Based on cluster of orthologous group (COG) annotation, the 14 genes were found to be mainly involved in energy production and conversion (e.g., TCA cycle, respiratory chain, ATP synthase) and amino acid transport and metabolism (e.g., biosynthesis of glutamate).

We next examined the specific COG categories of the DEGs across pH in each acid-adapted strain. For strain AA2, the upregulated genes are associated with more than ten COG categories (Figure 6.2b), with the largest number of genes on amino acid transport and metabolism (e.g., biosynthesis of histidine, threonine), energy production and conversion (e.g., nitrite/nitrate reductase, succinate dehydrogenase, TCA cycle, etc.), carbohydrate transport and metabolism (e.g., glycolysis), and inorganic ion transport and metabolism (e.g., transport of iron, zinc, nitrite, nitrate). Among the processes identified, the upregulated genes are enriched

in amino acid transport and metabolism and energy production and conversion based on a hypergeometric test (Figure 6.2b stars) (Methods). The downregulated genes of strain AA2 are found mostly in carbohydrate transport and metabolism (e.g., secondary carbon sources such as xylose, arabinose) and transcription. No COG categories are found to be enriched in those downregulated genes. For the other acid-adapted strain, AA6, the upregulated genes are mainly enriched in amino acid transport and metabolism (e.g., biosynthesis of glutamate, arginine) and energy production and conversion (e.g., TCA cycle, respiratory chain, ATP synthase) based on a hypergeometric test (Figure 6.2b stars). Again, no COG categories are enriched in downregulated genes in strain AA6.

Overall, we observed the DEGs to be enriched in similar general COG categories for two acid-adapted strains, AA2 and AA6. However, the specific underlying cellular processes still differ, indicating different strategies developed by *E. coli* over the course of evolution under acid stress. Such differences cover different processes that are upregulated in amino acid biosynthesis (e.g., histidine for AA2 and glutamate for AA6) and energy production and conversion (e.g., anaerobic respiration found in AA2 but not AA6) between two acid-adapted strains. Additionally, we found the downregulated genes to be involved in different COG processes between these two strains.

6.4 Discussion

In this work, we used ALE to study the adaptation of *E. coli* K-12 MG1655 under acid stress in glucose minimal medium. Using whole genome sequencing, we identified mutations on *rpoC* , *rho*, and *nagA* to be the converged mutations in acid-adapted strains. We then used RNA-seq to examine the gene expression profiles of acid-adapted strains across pH and compared

them to those of the starting GLU strain and the control strains. Through analysis of DEGs, we identified cellular processes acquired by *E. coli* through the adaptive evolution under acid stress.

We found five of the six acid-adapted strains to have mutations in *rpoC* , which functions as a subunit of RNA polymerase. The specific mutations include substitutions that change the encoded amino acids (AA2, AA3, and AA6) and deletions in the coding region that lead to shifts in reading frame (AA4 and AA5). Based on the Pfam database, the substitutions occurred in the protein domains that contain the active site and the pore region that allows the entrance of nucleotides to the active site [199–201]. It is worth noting that the deletions occurred at the end of *rpoC* gene (base pair 4106 and 4111 out of 4224 base pairs) and likely did not result in significant disruption of the gene function. A previous study on the evolution of *E. coli* under acid stress by Harden et al. also observed missense mutations in subunits of RNA polymerase (rpoBCD) for all of the acid-adapted strains [186]. The authors in that study proposed several mechanisms to explain how mutations on the RNA polymerase complex might enhance fitness under acid stress. Here, however, we consider the mutations on *rpoC* to be associated with inducing faster growth rather than acid resistance. A comprehensive analysis on the 278 gene expression datasets of *E. coli* across diverse conditions has revealed that mutations on genes related to RNA polymerase typically lead to improved growth rate and reduced stress-related gene expression [202].

Other mutations found cover a range of cellular processes. However, none of the processes are directly related to the commonly known acid resistance mechanisms. A previous study by Harden et al. identified mutations in genes related to the amino acid decarboxylase systems, and these mutations result in loss or downregulation of amino-acid decarboxylase activities [186]. The different mutations observed from the two studies are likely due to the different culture

media in which the evolution took place. The culture medium used by Harden et al. is LBK medium, which is rich in amino acids. The activation of the amino acid decarboxylase systems require the presence of amino acids in the medium [172–175]. According to Harden et al., amino acid decarboxylase systems protect *E. coli* from acid stress upon early exposure to the acidic environment, but incur fitness costs over the long term, where *E. coli* has developed other strategies to maintain the non-stress physiology. In our study, the culture medium is glucose minimal medium, where the sole carbon source is glucose. Therefore, the amino acid decarboxylase systems are never activated under this condition. Rather, we see mutations in genes that might be related to general cellular responses under stress conditions, e.g., transcription regulation (*rho*) and cellular physiology (*csgD/csgB*, *yiaA*).

For the two acid-adapted strains with gene expression profiles available, they share several common general COG categories for upregulated genes under acid stress. However, besides sharing 14 DEGs in processes such as TCA cycle, respiratory chain, ATP synthase, and glutamate biosynthesis, the two strains have a number of DEGs with different cellular functions (Figure 6.2). Both strains have SNPs in the *rpoC* gene, but in different protein domains according to Pfam database as mentioned earlier. Additionally, mutations on other genes in strain AA6 can possibly contribute to the different strategies used by *E. coli* under acid stress to adjust the level of gene transcripts. Similarly, Harden et al. also observed different patterns of gene expression across four acid-adapted strains [186]. These two studies together demonstrate that regardless of the level of acid stress (pH 4.6 - 4.8 by Harden et al. and pH 5.5 in this study) and nutrient availability (LBK medium and glucose minimal medium), the evolutionary pressure can drive *E. coli* to develop different strategies against acid stress.

Overall, we studied the adaptive evolution of *E. coli* under acid stress, linking the im-

proved phenotype to the underlying genotypes and levels of gene expression. The study here provides a novel perspective on acid resistance mechanisms, as the commonly known acid resistance systems depend on rich medium or specific amino acids [203, 204]. In addition to the analysis of genetic mutations and DEGs, further analysis can be performed to understand the change in regulatory actions using a recently developed approach [202]. Such analysis can be helpful in understanding the response to acid stress at the level of transcriptional regulation and revealing potential drivers behind the global adjustment of cellular response against acid stress.

6.5 Methods

6.5.1 Culture medium

The M9 glucose minimal medium was prepared by adding the following to Milli-Q water: 0.1mM $CaCl_2$, 0.2 mM $MgSO_4$, 1X trace elements solution, 1X M9 salt solution, and 4 g/L D-glucose. Trace elements solution (4000X) was prepared in concentrated HCl with 27 g/L $FeCl_3 \cdot 6H_2O$, 1.3 g/L $ZnCl_2$, 2g/L $CoCl_2 \cdot 6H_2O$, 2 g/L $Na_2MoO_4 \cdot 2H_2O$, 0.75 g/L $CaCl_2$, 0.91 g/L $CuCl_2 \cdot 2H_2O$, and 0.5 g/L H_3BO_3. M9 salt solution (10X) was prepared by dissolving 68 g/L Na_2HPO_2, 30 g/L KH_2PO_4, 5 g/L NaCl, and 10 g/L NH_4Cl in Milli-Q water. It is worth mentioning that the concentration of $MgSO_4$ is 10 times lower than used previously [188], as higher concentration of magnesium ion led to precipitation issues. To maintain the pH around 5.5 during cell culture, the culture medium was supplemented with 150 mM MES buffer from a 500 mM stock prepared in Milli-Q water. After mixing all components of the medium, the pH was adjusted using 2 M H_2SO_4 and 4 M KOH. All stock solutions as well as the final medium were sterile filtered through a 0.22 μM PVDF membrane.

6.5.2 Adaptive laboratory evolution process

Cultures were initiated from isolated colonies of an *Escherichia coli* K-12 MG1655 strain (ATCC 47076), which had previously been evolved for approximately 1013 CCDs on M9 minimal medium supplemented with 4 g/L of glucose [188]. The cultures were first grown overnight and then placed in tubes on a platform that performed passage automatically. The working culture volume was 15 mL, and the culture temperature was maintained at 37 °C. The culture medium was magnetically stirred at 1100 rpm to ensure a well mixed and aerobic growth environment.

From the start of the culture to the next passage, on average 4 samples of 100 μL culture medium were taken and the optical density measurements at a wavelength of 600 nM (OD600) were performed in a spectrophotometer (Tecan Sunrise). To maintain the cells in exponential growth phase, the culture medium containing *E. coli* (100 μL) was passaged to a tube containing fresh medium when OD600 of the original medium approached 0.3. The growth rate was determined for each culture using a least-squares fit on ln(OD600) versus time. Growth trajectories were generated by fitting a monotonically increasing cubic-interpolating-spline to the calculated growth rate values versus CCDs, as described previously [188]. Glycerol stocks of the cultures were taken periodically by mixing 800 μL of sterile 50% glycerol with 800 μL of culture and storing at -80 °C.

Throughout the course of the evolution, the culture medium pH was constantly measured to ensure proper buffering. Specifically, the pH values of the fresh medium and the culture medium before the next passage were measured. The culture medium was filtered through 0.22 μM membranes, and the pH was measured using a meter (Fisher Scientific Accumet AB15). Additionally, OD600 measurements of the culture medium were taken before the next passage to assess the possible effect of cell density on culture medium pH.

6.5.3 Whole genome sequencing and analysis of genetic mutations

Genomic DNA was isolated using bead agitation as described previously in Marotz et al. [205]. Whole genome DNA sequencing libraries were generated using a Kapa HyperPlus library prep kit (Kapa Biosystems). The libraries were then run on an Illumina HiSeq 4000 platform with a HiSeq SBS kit and 150/150 paired-end reads. The raw DNA sequencing reads in fastq format were processed using the *breseq* computational pipeline v0.33 [190]. Specifically, the workflow includes quality control [206], alignment to the *E. coli* genome (NCBI accession NC_000913.3) to identify mutations and annotation of the mutations. It is worth mentioning that genomic DNA was extracted for individual clones taken at different CCDs of the evolution and at the end of the evolution. The mutations identified in the clones at the end of the evolution were reported and those found at earlier stages were used to track how different mutations emerge or disappear throughout the course of the evolution.

6.5.4 RNA sequencing

RNA-sequencing data were generated from cell cultures under exponential growth phase at pH 5.5 and pH 7. The culture conditions at the specific pH were the same as used in ALE experiments mentioned above. Cells were stabilized with Qiagen RNA-protect Bacteria Reagent. Cell pellets were stored at -80 °C before RNA extraction. Then, frozen cell pellets were thawed and incubated with lysozyme, protease K, SuperaseIN, and 20% sodium dodecyl sulfate for 30 minutes at 4 °C. Total RNA was isolated and purified using Qiagen's RNeasy Plus Mini Kit based on the manufacturer protocol. The total RNA quality was checked using the RNA 6000 Nano kit from Agilent Bioanalyzer. For gram-negative bacteria, ribosomal RNA was removed using Ribo-Zero rRNA removal kit from Epicentre. Single-end, strand-specific RNA-seq libraries

were generated using KAPA RNA HyperPrep Kit from Kapa Biosystem. RNA-seq libraries were run on an illumina NextSeq platform using a 75 cycle mid-output kit.

6.5.5 Analysis of DEGs on RNA sequencing data

Raw sequencing reads in fastq format were first mapped to the reference genome (NCBI accession NC_000913.3) using bowtie v1.2.2 [207]. The abundance of the transcript was obtained using the summarizeOverlaps function from the GenomicAlignments package in R [208]. From the transcript abundance, the DEGs between two conditions were identified through the DESeq2 package in Bioconductor [209]. The output for the DEGs include log2(fold change) and the corresponding p-values (FDR-adjusted). DEGs with log2(fold change) greater than 1 and p-value smaller than 0.01 were considered to be significantly changed between the two conditions compared. The RNA-seq data is available in the Gene Expression Omnibus (GEO) database with accession number.

6.5.6 Enrichment analysis for COG categories

The set of DEGs between two different conditions were annotated using COG categories. The hypergeometric test was then performed for the set of upregulated genes and downregulated genes, respectively. To calculate the enrichment of each COG category in the gene set, four values were obtained to perform the test: the total number of genes mapped in RNA-seq data, the number of genes in the current set, the number of genes with the current COG category out of all genes, the number of genes with the COG category out of the current gene set. The FDR correction was applied on the p-values of the COG categories in the gene set. COG category with corrected p-value smaller than 0.05 was considered enriched in the gene set.

Acknowledgments

We would like to thank Laurence Yang for valuable discussions. This work was supported by National Institute of General Medical Sciences of the National Institutes of Health Grant R01GM057089 and Novo Nordisk Foundation Grant NNF10CC1016517.

Chapter 6 in full is a reprint of the material: **Bin Du***, Connor A. Olson*, Anand V. Sastry, Xin Fang, Patrick V. Phaneuf, Ke Chen, Muyao Wu, Richard Szubin, Julia Xu, Ye Gao, Ying Hefner, Adam M. Feist, Bernhard O. Palsson. "Adaptive laboratory evolution of Escherichia coli under acid stress." *Submitted*. The dissertation author was the primary author (equally contributing with Connor Olson).

Chapter 7

Mechanistic description of acid stress responses in *Escherichia coli* using genome-scale model of metabolism and gene expression

7.1 Abstract

Response to acid stress is critical for *Escherichia coli* to successfully complete its life-cycle by passing through the stomach to colonize the digestive tract. To develop a fundamental understanding of this response, we established a molecular mechanistic description of acid stress mitigation responses in *E. coli* and integrated them with a genome-scale model of its metabolism and macromolecular expression (ME-model). We considered three known mechanisms of acid

stress mitigation: 1) change in membrane lipid fatty acid composition, 2) change in periplasmic protein stability over external pH and periplasmic chaperone protection mechanisms, and 3) change in the activities of membrane proteins. After integrating these mechanisms into an established ME-model, we could simulate their responses in the context of other cellular processes. We validated these simulations using RNA sequencing data obtained from five *E. coli* strains grown under external pH ranging from 5.5 to 7.0. We found: i) that for the differentially expressed genes accounted for in the ME-model, 80% of the upregulated genes were correctly predicted by the ME-model, and ii) that these genes are mainly involved in translation processes (45% of genes), membrane proteins and related processes (18% of genes), amino acid metabolism (12% of genes), and cofactor and prosthetic group biosynthesis (8% of genes). We thus established a quantitative framework that describes, on a genome-scale, the acid stress mitigation response of *E. coli* that has both scientific and practical uses.

7.2 Background

Multiple studies have focused on the ability of *Escherichia coli* to tolerate acid stress [163–168]. *E. coli* has been shown to survive under extreme acid stress at pH 2 for several hours and to grow under acid stress above pH 4.5 [163, 166–168]. The ability to tolerate acid stress is critical for *E. coli* to complete its life cycle as an enteric bacteria. For colonization in the human digestive tract, it has to pass through the stomach with pH 1.5 to 3, and then metabolize and proliferate at around pH 5 to 6 in the intestinal tract [169, 170]. A fundamental understanding of the acid resistance mechanisms of *E. coli* thus has important implications in the food and health care industry, e.g., the development of effective strategies against pathogenic *E. coli* by targeting specific acid resistance mechanisms.

Various acid resistance mechanisms exist that protect *E. coli* under acid stress and are found across different cellular compartments. In the cytoplasm, mechanisms that actively consume protons include four types of amino acid decarboxylase systems and formate hydrogen lyase [172–175, 183]. Metabolism of secondary carbon sources and sugar derivatives are upregulated as these carbon sources produce fewer acids compared to glucose when metabolized [176, 177]. Additionally, cytoplasmic buffering from inorganic phosphates, amino acid side chains, polyphosphates and polyamines helps to maintain intracellular pH homeostasis [171]. When cytoplasmic pH drops under extreme acid stress, cytoplasmic chaperones such as Hsp31 bind and protect unfolded protein intermediates; DNA-binding proteins bind and protect DNA [210–212]. On the inner membrane, activities of electron transport chain components and composition of membrane lipids change under acid stress [176, 177, 179, 180]. In the periplasmic space, periplasmic chaperones HdeA and HdeB are activated under acid stress to bind and protect unfolded protein intermediates [178]. Lastly, outer membrane porins are bound by polyphosphate or cadaverine to reduce proton influx [181, 182].

While there have been extensive studies describing the response of *E. coli* under acid stress, research to elucidate how different acid resistance mechanisms function together to protect *E. coli* against low pH environment is lacking. Such an explanation will require a detailed characterization of different acid resistance mechanisms of *E. coli* . The genome-scale metabolic model (M-model) of *E. coli* provides a mathematical representation of its metabolic capabilities and serves as an ideal framework to describe the acid stress response of *E. coli* [74]. Recently, M-models have been extended to include the synthesis of the gene expression machinery (called ME-models) [213, 214]. In addition to computing the optimal metabolic flux state of the organism, ME-model computes the optimal proteome allocation for a given phenotype [214, 215],

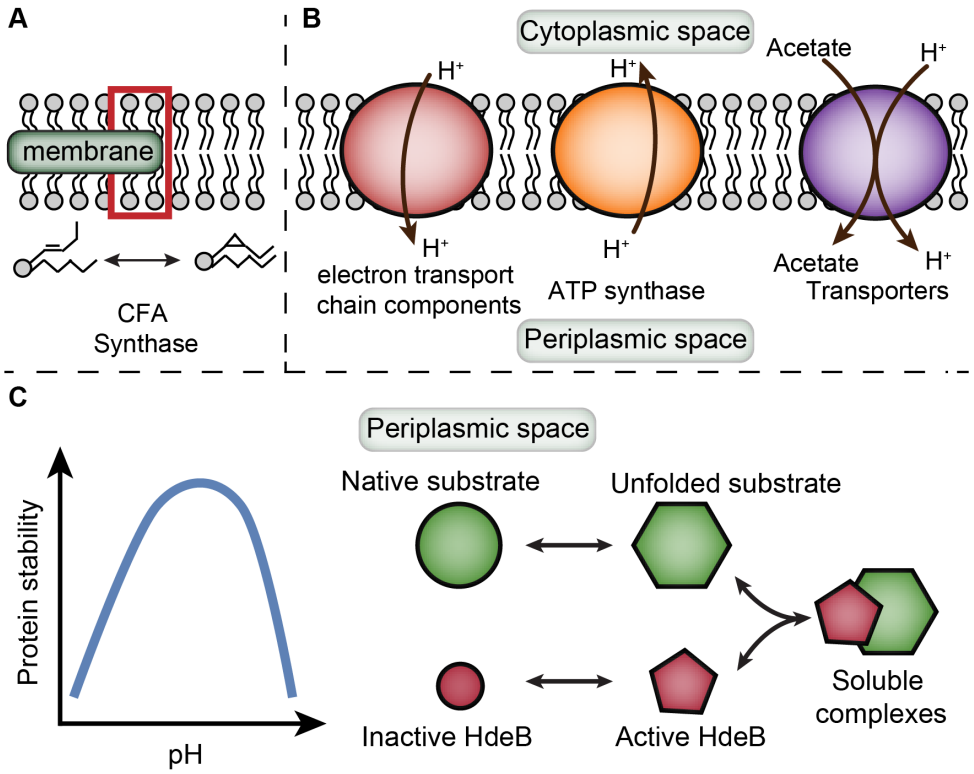


Figure 7.1: Illustrations of three different stress response mechanisms of *E. coli* under acid stress. a) Adjustment of membrane lipid fatty acid composition. b) Change in periplasmic protein stability and periplasmic chaperone protection. c) Activity change of membrane proteins.

thus providing additional information on the cellular processes as a whole. Furthermore, the calculation on proteome allocation can be validated with RNA sequencing data, which can be conveniently obtained with the advancement of next-generation sequencing technology.

In this work, we characterize the growth of *E. coli* under mild acid stress using the ME-model framework (Figure 7.1). Mild acid stress can be found under a variety of conditions, including the intestinal tract and fermented food, where the pH is around 5 to 6 [169, 216]. We first incorporate the change in fatty acid composition of membrane lipids into the ME-model, based on experimental measurements under mild acid stress. Next, we model the change in periplasmic proteins under acid stress, specifically on protein stability and periplasmic chaperone protection. We also model the change in activity for proteins located in the inner membrane of

E. coli , including ATP synthase, electron transport chain components, and transporters. We integrate all these modifications into the ME-model and compare the simulations with RNA sequencing data of *E. coli* grown under neutral pH and mild acid stress. Specifically, we examine the upregulated and downregulated genes, as well as the change in cellular processes based on cluster of orthologous group (COG) annotation [217].

7.3 Results

7.3.1 Adjustment of *E. coli* membrane lipid fatty acid composition under acid stress

The *E. coli* membrane serves as a barrier between the intracellular space and the external environment by controlling the entry and exit of ions and molecules of different sizes. The components of the membrane have been shown to actively respond to changes in the external environment [218]. Specifically, membrane lipids are important components in maintaining membrane function and integrity under environmental perturbations. Several studies have demonstrated that the membrane lipid composition of *E. coli* changes under acid stress, resulting in the change of membrane fluidity that potentially reduces the leakage of protons into the cytoplasm [179, 180, 219]. Here, we will recapitulate this response in the context of the *E. coli* ME-model framework.

The current ME-model provides a detailed description of the proteins and lipids that constitute the inner and outer membranes of *E. coli* [214]. However, it does not include the constraint that the membrane surface area is completely occupied by proteins and lipids. Therefore, we need to add this constraint into the current ME-model to describe this acid stress response. Our incorporation of the membrane area constraint was able to reproduce the results of similar

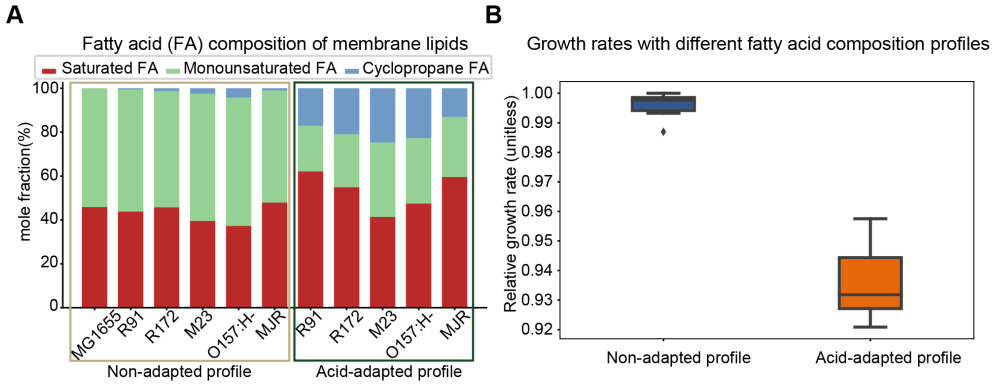


Figure 7.2: Fatty acid composition of membrane lipids under different pH conditions. a) Comparison of calculated acid-adapted (AA) against non-adapted (NA) fatty acid composition profiles for different *E. coli* strains. The fatty acid composition profiles are calculated based on published data [179]. b) Comparison of simulated *E. coli* growth rates with different fatty acid composition profiles incorporated into the ME-model. The use of the experimentally determined changes in membrane composition under acid stress leads to around 6% decrease in the computed growth rate.

earlier work [220].

Earlier study showed that the composition of fatty acid tails on the membrane lipids of *E. coli* changes during adaptation to acid stress [179]. Specifically, the mole fraction of monounsaturated fatty acids decreased during adaptation, while the proportion in saturated fatty acids and cyclopropane fatty acids increased. This trend is consistently observed across all five *E. coli* strains examined. Notably, the composition in cyclopropane fatty acids increased significantly (from an average of 1.57% to 19.6% out of the total fatty acid content) during acid adaptation. We obtained a total of 11 profiles of membrane lipid fatty acid composition of *E. coli* strains from the study by Brown et al. [179] and the existing M-model reconstruction [74]. We grouped the profiles into two categories: the group with an acid-adapted profile where *E. coli* was grown under acidic pH and the group with a non-adapted profile where *E. coli* was grown under neutral pH ((Figure 7.2a).

We incorporated the change in membrane lipid fatty acid composition into the *E. coli*

ME-model, while maintaining consistency on the biomass composition and membrane surface area constraints [73, 220]. Specifically, the mole fractions of membrane lipids with different fatty acid tails are transformed to their relative fractions in biomass following the procedures in a previous work [73], with units in millimole per gram dry weight of biomass. The calculated lipid biomass fractions are used as the coefficients of lipids in the ME-model reaction on biomass function [214].The ME-model predicted the group with the acid-adapted profile to have lower relative growth rates (0.94 ± 0.01) compared to the group with the non-adapted profile (1.00 ± 0.01) (p-value 5.93×10^{-6}) (Figure 7.2b).

Based on model simulation, we found the *cfa* gene to have the largest change in expression level between the acid-adapted profile and the non-adapted profile. The product of the *cfa* gene, cyclopropane fatty acyl phospholipid synthase, catalyzes the transfer of the methyl group from S-adenosyl-L-methionine to convert unsaturated fatty acids to cyclopropane fatty acids. The other genes with the largest computed change in expression levels are mainly associated with the recycling of S-adenosyl-L-methionine and cover a variety of cellular processes including methionine metabolism (*luxS*, *metK*, *metE*), nucleotide metabolism (*purN*, *deoD*), and folate metabolism (*metF*, *folD*).

7.3.2 Periplasmic protein stability as a function of pH and periplasmic chaperone protection

Under mild acid stress, *E. coli* maintains intracellular pH within a narrow range (7.4 - 7.6) [171, 221]. However, the pH of the periplasm is close to the external pH when *E. coli* is exposed to an acidic environment [222]. The acidic pH in the periplasm poses a challenge to the periplasmic proteins. *E. coli* has developed strategies to protect periplasmic proteins from

acid-induced damage, using molecular chaperones HdeA and HdeB that bind to native substrates to reduce protein denaturation and aggregation [178]. Here, we focus on modeling the change in periplasmic protein stability and the protection by molecular chaperones on periplasmic proteins under acid stress.

Protein stability as a function of pH depends on the pK_{a} and protonation states of the amino acid side chains of the protein [223–226]. Specifically, protein stability can be described using folding energy ($\Delta G_{folding}$), which is the difference between the folded state and unfolded state of the protein. For the same protein, a more negative folding energy indicates greater stability. An empirical approach has been developed that calculates $\Delta G_{folding}$ based on the number of amino acids of the protein [223, 227]. To account for the change in $\Delta G_{folding}$ as a function of pH, Ghosh and Dill [223] expressed $\Delta G_{folding}$ as the sum of two terms,

$$\Delta G_{folding} = \Delta G_{neutral} + \Delta G_{electric} \tag{7.1}$$

where $\Delta G_{neutral}$ is the energy term that does not consider any charge effect and $\Delta G_{electric}$ accounts for electrostatic interactions and is a function of pH. The term $\Delta G_{electric}$ is protein-specific and depends on the charge and radius of gyration of the folded and unfolded states (Methods).

We calculated the profiles of $\Delta G_{folding}$ as a function of pH for 86 of 93 periplasmic proteins in the ME-model. Folding energies of the other 7 proteins could not be calculated due to issues associated with protein charge calculation (Methods). We also compared the folding energies of the periplasmic proteins under pH 7 and pH 5.5. We found that proteins under pH 7 generally have lower $\Delta G_{folding}$ than those under pH 5.5 (Figure 7.3a), indicating greater stability for proteins under neutral pH. Notably, all periplasmic proteins examined are favorable towards

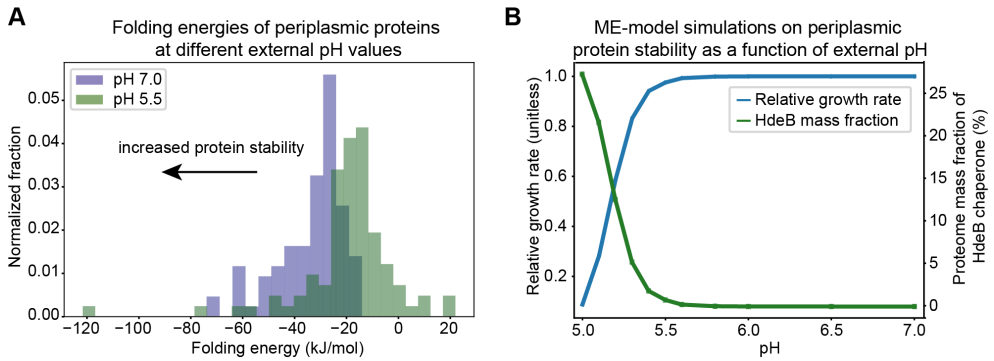


Figure 7.3: Periplasmic protein stability is reflected in protein folding energies. a) Comparison of calculated folding energies of *E. coli* periplasmic proteins at external pH 7.0 and pH 5.5. Proteins with lower folding energies are generally more stable. Therefore, the periplasmic proteins are found to be more stable at external pH 7 compared to pH 5.5. b) ME-model simulations on relative growth rate and HdeB mass fraction at different external pH conditions. We calculated the folding energies of periplasmic proteins as a function of pH and modeled the relative ratio of folded and unfolded states of each protein in the ME-model (Methods). We also included the binding of HdeB chaperone to the unfolded states. We then simulated the change in *E. coli* growth rate due to change in protein stability under different external pH conditions. We also showed the change of HdeB mass fraction of the total proteome as a function of external pH.

folding under pH 7 (Figure 7.3a). We also determined the optimal pH for each protein under study, where $\Delta G_{folding}$ is the lowest and the protein is most stable under the optimal pH. We found that while most proteins have optimal pH around 7, a large number of them have optimal pH around 12 and some have optimal pH around 3. Interestingly, we found a few proteins with increased expression levels under their optimal pH, e.g., FodG (optimal pH 3.5), Slt (optimal pH 13), PotD and HisJ (optimal pH 11.5) [177], indicating that protein stability might be an underlying factor that influences the protein expression in the periplasm.

We describe the relationship between the folded and unfolded states of the protein in the form of a ME-model reaction, similar to the approach in the previous work [228]. Specifically, the ratio between the folded and unfolded states of the protein can be calculated from

$$\Delta G_{folding} = -RT\ln([\text{Folded}]/[\text{Unfolded}]) \tag{7.2}$$

where R is the ideal gas constant, T is the temperature, [Folded] and [Unfolded] are the concentrations of the folded and unfolded peptide states. The ratio is expressed as the metabolite coefficient in the ME-model reaction, connecting the folded and unfolded states of the protein (Methods). Next, to model periplasmic chaperone protection, we focus on the mechanisms of HdeB, since HdeB has an optimal activation pH from 4 to 5, while HdeA is most active under pH 2 to 3 [229]. We described HdeB protection on the protein in the form of a ME-model reaction, in which the HdeB protein binds to the unfolded state of the protein to form a chaperone-protein complex (Methods).

Incorporating the description on periplasmic protein stability and HdeB protection in the ME-model, we simulated the response of *E. coli* under different external pH conditions. We found the relative growth rate to decrease slowly as pH decreases from 7 to 5.5, but drops quickly when pH decreases beyond 5.5 (Figure 7.3b). Similarly, we observed the mass fraction of HdeB of the total proteome to change slowly before pH decreases to 5.5 and increases significantly as pH decreases from 5.5 to 5. The stability change of LptA protein was found to be the major factor causing the drop in growth rate and increase in HdeB mass fraction. LptA protein is involved in the transport of (KDO)2-lipid IVA, which contributes to *E. coli* biomass [230]. Based on ME-model simulations, genes with the largest change in expression levels as a result of decreasing pH are *hdeB* (periplasmic chaperone), *lptA* (lipopolysaccharide Biosynthesis), *rpoE* (transcription), and *secBDEFGY* (Sec translocation processes). The Sec complexes are responsible for translocating the LptA protein from the cytoplasm into the periplasmic space [220].

7.3.3 Membrane protein activity as a function of pH

Under mild acid stress, *E. coli* maintains pH homeostasis in the cytoplasm (pH around 7.4) while its periplasmic pH is close to that of the external acidic environment [171, 221, 222]. Thus, the difference in proton concentration across the inner membrane results in a large proton motive force [203, 231]. For membrane proteins involved in proton import/export processes, their activities can be significantly affected by the change in proton motive force at different external pH conditions. These proteins include ATP synthase, electron transport chain components and various membrane transporters. Here, we model the change of their activities under mild acid stress and integrate these changes into ME-model simulations.

We first model the activity change of ATP synthase under mild acid stress using an existing kinetic model [232]. Specifically, the model consists of a series of elementary reactions that describes the proton transport and the rotation of the rotor subunit in ATP synthase. The rate of ATP synthesis is expressed in terms of the proton concentrations in the cytoplasm and periplasm, as well as the kinetic parameters of the elementary reactions (Methods). It is worth mentioning that ATP synthesis rate also depends on the membrane potential [232, 233] and different sets of kinetic parameters are needed when the membrane potential changes under different external pH values. Thus, we fitted the experimental data by Fischer and Gräber [233] on ATP synthesis rate as a function of transmembrane pH difference at three different transmembrane potentials and obtained three parameter sets for rate calculation. The calculated ATP synthesis rates at different external pH values can be found in Figure 7.4a.

Next, we examined the activity change of the electron transport chain components and various membrane transporters. There is not much evidence available on these reaction mechanisms in terms of the detailed elementary steps. Thus, we modeled their reaction rates based on

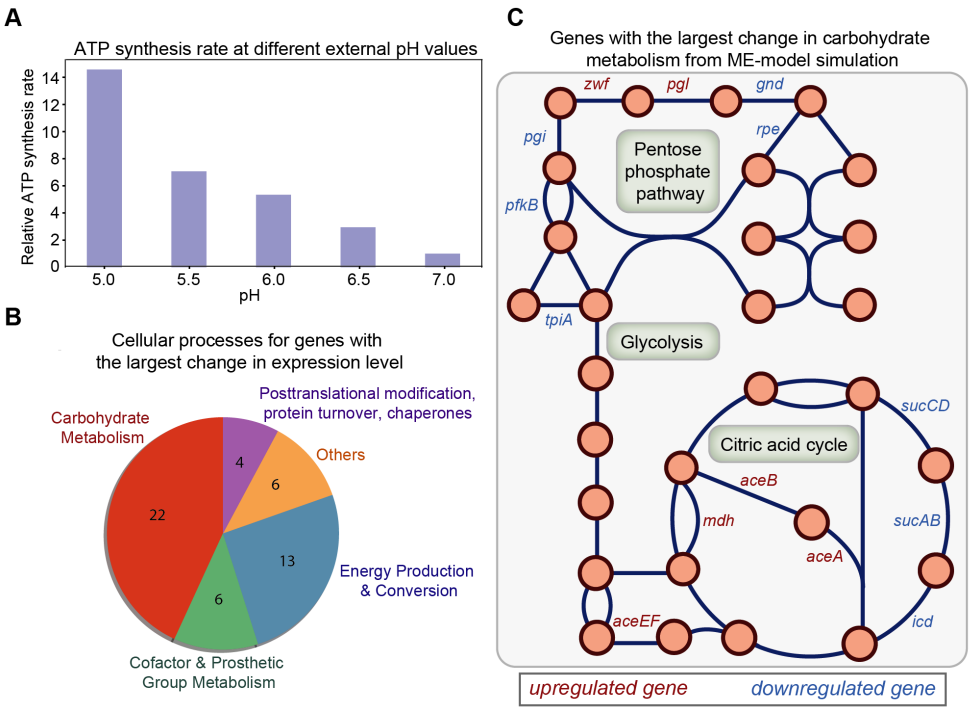


Figure 7.4: Change in ATP synthesis rate at different external pH values and the effect on cellular processes simulated using the ME-model. a) Relative ATP synthesis rates calculated at different external pH values. The relative ATP synthesis rate at pH 7 is set to 1. b) Genes with the largest computed change in expression level at pH 5.5 compared to pH 7. The top 51 genes with the largest change in expression level are grouped based on their assignment to the indicated cellular processes. c) Genes with the largest change in carbohydrate metabolism from ME-model simulations. Most of the genes are involved in three main processes in central metabolism (glycolysis, pentose phosphate pathway and citric acid cycle) and are displayed on the metabolic network map. Genes predicted to be upregulated are colored in red and genes predicted to be downregulated are colored in blue.

the theory of nonequilibrium thermodynamics [234]. Specifically, the rate is expressed in terms of the reaction energy, the membrane potential, periplasmic and cytoplasmic proton concentrations, as well as the concentrations of metabolites involved (Methods). The calculations on the electron transport chain components show that their rates remain almost unchanged from neutral pH to acidic pH. We were unable to calculate the reaction rates for most of the membrane transporters, due to missing metabolite concentration data. However, we found that the change of their activities had minimal impact on cellular growth rate and processes (¡1%) through the sensitivity analysis using the ME-model.

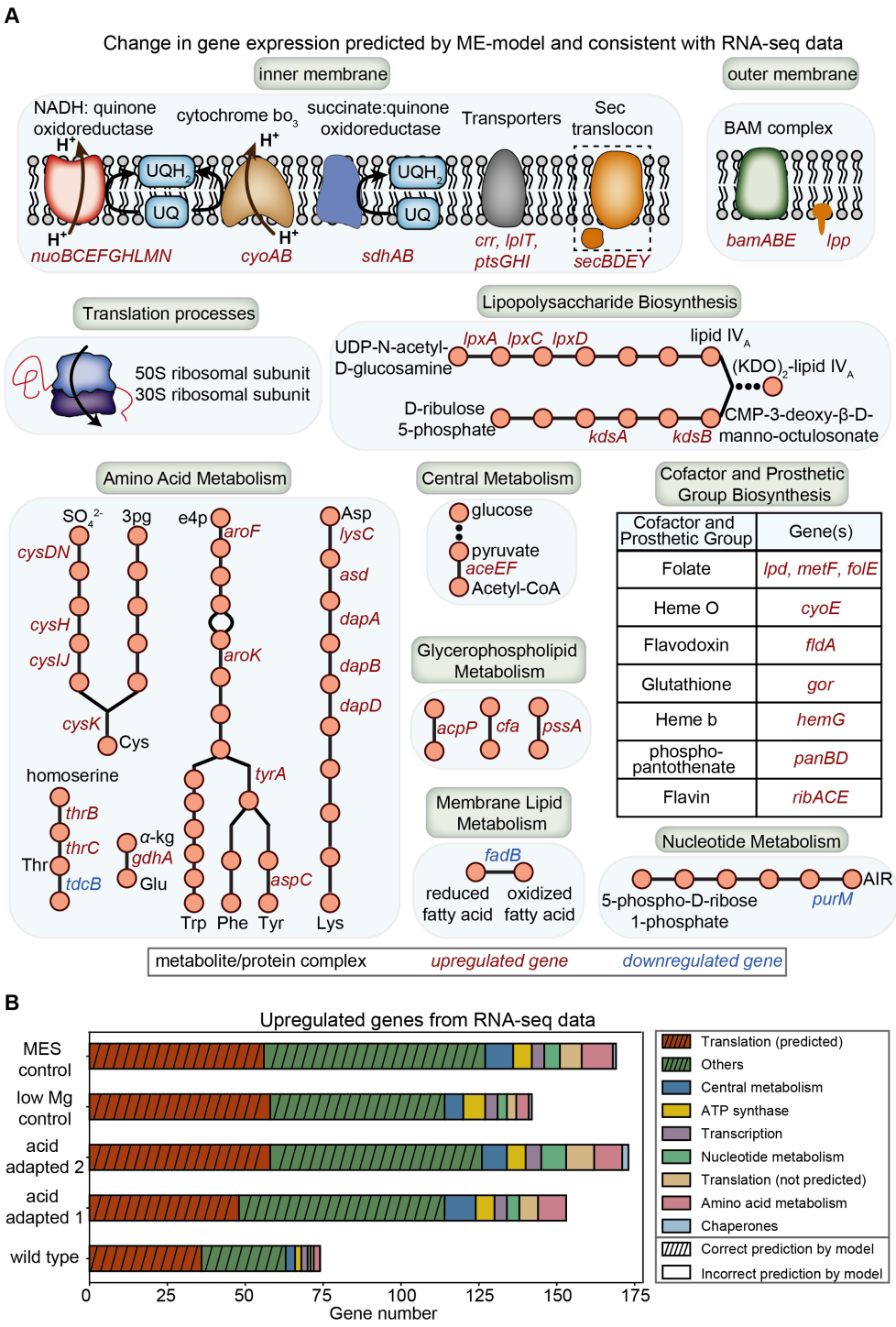
Based on the analysis on the activity change of different membrane proteins across pH, we modeled the change of ATP synthesis rate at different external pH values by modifying the effective turnover rate (k_{eff}) of the reaction catalyzed by ATP synthase in the ME-model [214] (Methods). Considering possible errors due to parameter fitting, we performed sensitivity analysis and found the change in cellular processes at different ATP synthesis rates to be similar. Using the calculated ATP synthesis rate at pH 5.5 as an example (Figure 7.4a), the top 50 genes with the largest change in expression levels are mainly involved in carbohydrate metabolism (e.g., citric acid cycle, glycolysis/gluconeogenesis, pentose phosphate pathway) and energy production and conversion (oxidative phosphorylation related to ATP synthase) (Figure 7.4b). We also showed the genes with the largest change in carbohydrate metabolism in the context of a metabolic network map (Figure 7.4c).

7.3.4 ME-model with integrated mechanisms explains the acid stress response of *E. coli*

We integrated the description of the three pH stress mitigation mechanisms (membrane lipid fatty acid composition, periplasmic protein stability and periplasmic chaperone protection, and the activity change of membrane proteins) into the ME-model, and then simulated its response under neutral pH and mild acid stress (pH 5.5). We compared the simulations to RNA sequencing data of K-12 MG1655 *E. coli* strains grown under pH 7 and pH 5.5 in glucose minimal medium from a previous study [235]. The *E. coli* strains from which the RNA-seq data were obtained include: 1) the wild type strain, 2) two strains adapted to pH 5.5 through adaptive laboratory evolution [184], and 3) two control strains adapted to specific media conditions. Since the acid-adapted strains were evolved in glucose minimal medium with lowered magnesium concentration and MES buffer, the two control strains (one for lowered magnesium concentration and one for MES) were necessary to account for the possible effects due to these two changes in media composition.

We compared ME-model simulations and RNA-seq data in terms of the differentially expressed genes (DEGs) due to acid stress (growth under pH 5.5 versus pH 7). We grouped the DEGs found in RNA-seq data into three categories: 1) DEGs currently not active in the ME-model, 2) DEGs correctly predicted by the ME-model, and 3) DEGs incorrectly predicted by the ME-model.

We found a large number of genes in the first category to be associated with membrane proteins and transporters and their related cellular processes. For example, one of the reported acid stress responses involves the blockage of outer membrane porins by secreted cadaverine [236]. Therefore, these DEGs are currently outside the ME-model's predictive capabilities. To include



Cofactor and Prosthetic Group	Gene(s)
Folate	*lpd, metF, folE*
Heme O	*cyoE*
Flavodoxin	*fldA*
Glutathione	*gor*
Heme b	*hemG*
phospho-pantothenate	*panBD*
Flavin	*ribACE*

Figure 7.5: Comparison of ME-model simulations, accounting for the three acid stress mechanisms, against RNA-seq data from *E. coli* . a) Differentially expressed genes (DEGs) due to acid stress found to be consistent with model predictions and RNA-seq data. b) Upregulated genes in RNA-seq data compared to ME-model simulations.

such descriptions in the ME-model, quantitative measurements on cadaverine binding to outer membrane porin and the corresponding change under acid stress is required.

For genes in the second category, we found that, on average, 80% of the upregulated genes in the RNA-seq data to be correctly predicted. These correctly predicted DEGs are mainly involved in the translation process (45% of genes), membrane proteins and related processes (18% of genes), amino acid metabolism (12% of genes), and cofactor and prosthetic group biosynthesis (8% of genes). Additionally, we found a limited number of downregulated genes to be active in the ME-model, as shown in Figure 7.5a. For genes in the third category, those found to be upregulated in the data but predicted to be downregulated in the ME-model are grouped by COG categories and shown in Figure 7.5b. Genes found to be downregulated in the data but predicted to be upregulated by the ME-model are discussed in more detail below.

We grouped the correctly predicted DEGs by COG categories (Methods) and summarized them by the underlying mechanisms in Figure 7.5a. We found a large number of upregulated genes to be related to the translation process (Figure 7.5b red). Additionally, we found upregulated expression for a number of proteins on the inner and outer membranes of *E. coli* . These proteins include the electron transport chain components, transporters (uptake of sugar, lysophospholipid), Sec translocase, and BAM complex responsible for outer membrane assembly. Furthermore, the DEGs in amino acid metabolism cover processes related to cysteine, threonine, lysine, glutamate, and aromatic amino acids (tryptophan, tyrosine, phenylalanine). We found cofactor and prosthetic group biosynthesis to be another major category with a number of upregulated genes. Lastly, due to changes in membrane lipid fatty acid composition and periplasmic proteome predicted by the ME-model, we found upregulated genes in RNA-seq data to be related to membrane lipid metabolism, lipopolysaccharide biosynthesis, and glycerophospholipid

metabolism.

We then examined the incorrectly predicted DEGs. We found a few genes to be downregulated in the RNA-seq data but predicted to be upregulated (*rlmC*, *glcD*, *hisI*, *erpA*, *nadB*). Upon examining the reactions catalyzed by these gene products, we found proton generation to be involved in three reactions, with the corresponding genes being *rlmC*, *hisI*, *nadB*. The proton generation in the reaction explains the downregulation of these genes, as *E. coli* tends to minimize proton production under acid stress. Genes found to be upregulated in the data but downregulated in ME-model predictions were grouped based on the COG categories (Figure 5B). These incorrectly predicted DEGs suggest ways to further develop the modeling of acid stress response. For example, the arginine-dependent acid resistance system has been shown to play a role under acid stress [174], but the corresponding genes were not correctly predicted by the ME-model. A possible way to improve model predictions is to fine-tune model parameters related to arginine metabolism based on RNA-seq data. We also found genes related to cytoplasmic chaperones to be upregulated in RNA-seq data but not predicted by the ME-model. A previous reconstruction of the cytoplasmic chaperone network in the ME-model exists [228] and its incorporation can potentially improve predictions of the use of chaperone related processes.

7.4 Discussion

In this study, we described the response of *E. coli* under acid stress using the ME-model framework. We first modified the membrane lipid fatty acid composition based on experimental data, with the addition of the constraint on total membrane surface area. Second, we modeled the pH-dependent periplasmic protein stability and periplasmic chaperone protection mechanisms. Third, we characterized the activities of membrane proteins under low pH. Lastly, we integrated

these descriptions of stress mitigation mechanisms into the ME-model and compared the simulations of the integrated model with measured RNA sequencing data. We demonstrated that the ME-model was able to recapitulate DEGs under acid stress in a number of cellular processes, including amino acid metabolism, cofactor and prosthetic group biosynthesis, processes related to membrane proteins, and translation process. The effects of acid stress mitigation on these cellular processes can now be understood at the systems level and quantitatively computed. We also suggested a few areas for further model development, based on model predictions that were inconsistent with the RNA-seq data.

The work here describes the change in the cellular state of *E. coli* between two distinct conditions, the mild acidic condition and the neutral condition. A continuous profile of the change in cellular processes as the pH decreases from neutral to acidic can provide more insights into how *E. coli* adjusts its cellular resource allocation when facing increased acid stress. However, such an effort is currently limited due to the lack of relevant experimental data. For example, the current data on fatty acid composition of membrane lipids of *E. coli* are only measured under pH 5 and 7. Possible steps forward include acquiring more experimental data at the intermediate pH values between 5 and 7 or making simplifying assumptions about how the fatty acid composition profile changes over pH.

ME-model simulations predicted only a few of the periplasmic proteins to be active. The main reason for the inactivation of other proteins is the lack of description of their downstream processes or metabolic reactions they catalyze in the ME-model. The addition of relevant processes could help provide a more complete picture of the periplasmic protein response under acid stress, as the stability profiles for most of the periplasmic proteins are available from this work. Furthermore, adding these descriptions can uncover more periplasmic proteins that significantly

affect the growth rate and cellular processes, and potentially improve the predictions on acid stress response.

After integrating the three acid stress mitigation mechanisms considered into the ME-model framework and performing the simulations, we revealed changes in cellular processes of *E. coli* under mild acid stress. There are several known acid resistance mechanisms that were not activated in the ME-model framework. First, although various amino acid decarboxylase systems play important roles in maintaining pH homeostasis, we did not model the change of their activities due to the context of acid stress response described here. We used the ME-model to describe the adapted response of *E. coli* under mild acid stress. On the other hand, the optimal activities of decarboxylases were shown to occur below neutral pH, indicating that the decarboxylases are typically active when there is a large influx of proton into the cytoplasmic space or when *E. coli* is under extreme acid stress when the intracellular pH drops to around 4 to 5 [183]. Another related acid resistance mechanism is cytoplasmic buffering. The description of this mechanism requires a detailed characterization of the metabolites and amino acid side chains at different protonation states, which is currently out of the model's scope and can be an area of further ME-model development. Additionally, the descriptions of DNA-binding proteins and HdeA activation are not currently included in the ME-model, but will be more relevant in terms of the response under extreme acid stress [183].

The ME-model framework here enables predictions of how different interventions affect the acid stress tolerance of *E. coli* . For example, we can design intervention strategies on the recycling of S-adenosyl-L-methionine, which is an important cofactor responsible for the adjustment of membrane lipid fatty acid composition under acid stress. As another example, the effect of *hdeB* knockout can be simulated using the ME-model and compared with experimental

data. Discrepancies between model simulations and the data can potentially lead to discoveries of novel periplasmic chaperone protection mechanisms [237].

Taken together, the work here describes acid stress mitigation responses in *E. coli* through a mechanistic approach and provides insights into the resulting changes to its cellular processes. It is worth noting that the current description focuses on the acid stress response of *E. coli* under the aerobic growth condition with glucose as the sole carbon source. In practice, *E. coli* faces more complicated nutrient environments and can be subjected to anaerobic respiration. The response to acid stress differs due to different environmental conditions (e.g., activation of formate hydrogen lyase under anaerobic acid stress [183]). Thus, descriptions of additional acid resistance mechanisms can be added to expand the scope of ME-model predictions. The study here is a first step towards a complete characterization of the wide array of acid stress responses of *E. coli* .

7.5 Methods

7.5.1 ME-Model and simulations

The ME-model framework is based on the work by Lloyd et al [214], with no change on the parameters used other than the inclusion of acid stress mitigation responses described in the text. A quad-precision NLP solver was used to obtain the ME-model solutions [238]. The source code for model construction and integration of the acid stress mitigation mechanisms is available on GitHub (https://github.com/bdu91/acidify-ME). All work here is in implemented in Python 2.7.6.

7.5.2 Stability of periplasmic proteins as a function of pH

As mentioned in the main text, protein stability can be quantified by the folding energy $\Delta G_{folding}$, which is the sum of $\Delta G_{neutral}$ and $\Delta G_{electric}$ based on equation 7.1. The change in pH affects the value of $\Delta G_{electric}$, which can be expressed as

$$\Delta G_{electric} = kT \left(\frac{Q_{folded}^2 l_b}{2R_{folded}(1+\kappa R_{folded})} - \frac{Q_{unfolded}^2 l_b}{2R_{unfolded}(1+\kappa R_{unfolded})} \right) \tag{7.3}$$

where Q_{folded} and $Q_{unfolded}$ are the protein charges in the folded and unfolded states, R_{folded} and $R_{unfolded}$ are radius of gyration of the folded and unfolded states, k is the Boltzmann constant, T is the temperature, l_b is the Bjerrum length and $\kappa = 2cl_b$ (c as the salt concentration, set as 0.25 M here) [223].

The charge of the unfolded state of the given protein can be calculated based on the pKas and charges of the individual amino acid side chains. The charge of the folded state can be obtained through a method called multi-conformation continuum electrostatics (MCCE), which calculates the pK_{a}s and charges of the amino acid side chains of the folded state [239]. The MCCE method requires the PDB structures of the folded proteins, which were obtained from the latest genome-scale metabolic network reconstruction of *E. coli* [74]. It is worth mentioning that the charge of 7 periplasmic proteins cannot be calculated due to failed delphi runs in the MCCE method. The radius of gyration of the folded protein R_{folded} is calculated through the Bio3d package in R [240], using the PDB structure of the folded protein. The radius of gyration of the unfolded protein $R_{unfolded}$ is obtained by fitting empirical data and the relationship between the number of amino acid residues N and $R_{unfolded}$, where $R_{unfolded} \propto N^{0.588}$ [241].

Finally, as $\Delta G_{folding}$ at neutral pH can be calculated based on the number of amino acids of the protein [223, 227], $\Delta G_{folding}$ at different pH values can be obtained from the change of $\Delta G_{electric}$ over pH using equation 7.3.

7.5.3 Periplasmic chaperone protection by HdeB in the ME-model

We first modeled the formation of HdeB protein, including steps on transcription, translation, translocation from the cytoplasm to the periplasm and formation of HdeB dimer [229]. The details of each step have been defined in the COBRAme framework by Lloyd et al [214]. We then modeled the protection of HdeB on unfolded proteins. We defined a spontaneous folding reaction for each periplasmic protein, using the coupling constraint defined by Ke et al [228]. Specifically, we have

$$K[HdeB] + (1 + K + \mu/k_{folding})[Unfolded] \leftrightarrow K[HdeB - unfolded - complex] + [Folded] \quad (7.4)$$

where $[HdeB]$ is the HdeB protein, $[Unfolded]$ and $[Folded]$ are the folded and unfolded states of the protein, $[HdeB - unfolded - complex]$ is the complex formed by HdeB bound to the unfolded state, K is the ratio between the unfolded state and the folded state and can be obtained from Gfolding under the given pH, μ is the growth rate in the ME-model, $k_{folding}$ is the kinetic folding rate and can be calculated based on the work by Gromiha et al [242]. For proteins where $\Delta G_{folding}$ cannot be obtained, we assume the protein is favorable towards folding under all conditions and set $\Delta G_{folding}$ to -100 kJ/mol.

7.5.4 Activity of ATP synthesis rate as a function of external pH in the ME-model

We used the kinetic model by Jain and Nath [232] to describe the mechanism of ATP synthase through a list of elementary steps, including proton transport and rotor rotation. The rate of ATP synthesis can be expressed in terms of the cytoplasmic and periplasmic proton concentrations, as well as the kinetic parameters.

$$v = \frac{k_1}{1 + \frac{k_2 H^+_{cytoplasm}}{H^+_{periplasm}} + \frac{k_3}{H^+_{periplasm}}} \tag{7.5}$$

It is worth mentioning that parameters k_1, k_2 and k_3 are composite terms. Each term consists of various kinetic parameters of the elementary steps.

We used the experimental data from Fischer and Gräber [233], where the rate of *E. coli* ATP synthase was measured as a function of transmembrane pH difference at three different transmembrane potentials (80 mV, 108 mV, 152 mV). Based on equation 7.5, we obtained three sets of kinetic parameters at different membrane potentials by fitting the experimental data through a non-linear least-squares minimization procedure [243].

To calculate the rate of ATP synthesis under a specific external pH, we first calculated the cytoplasmic pH, using the relationship between the cytoplasmic pH and the external pH derived by Slonczewski et al [221]. We next calculated the membrane potential of *E. coli* under the given external pH based on the experimental measurements by Felle et al [244]. From the three fitted parameter sets at different membrane potentials, we selected the set with the closest membrane potential. Using the selected parameter set and the calculated pH values, we calculated the rate of ATP synthesis under different external pH conditions. To standardize the calculated rates, we

defined the rate under pH 7 as 1 and expressed the rates under other pH values as the fold change relative to it. To incorporate the change in ATP synthesis rate under the specific external pH in the ME-model, we adjusted the effective turnover rate (k_{eff}) of ATP synthase in the ME-model according to the calculated fold change under the given external pH [214].

7.5.5 Activity of electron transport chain components as a function of pH

For electron transport chain components, we described the rate as a function of pH using the derivation by Jin and Bethke [234], based on the theory of nonequilibrium thermodynamics. Specifically, the rate is expressed as,

$$v = v_+ \left(1 - \exp\left(\frac{-nF\Delta E^\circ + mF\Delta\psi}{RT}\right)\left(\frac{[H^+_{periplasm}]^m[D^+]^{v_{D^+}}[A^-]^{v_{A^-}}}{[H^+_{cytoplasm}]^m[D]^{v_D}[A]^{v_A}}\right)\right) \tag{7.6}$$

where v_+ is the forward reaction flux, n is the number of electrons transferred, ΔE° is the difference in standard redox potential between the donating and accepting half-reactions, m is the number of protons transported across the membrane, $\Delta\psi$ is the membrane potential, F is Faraday's constant, R is the ideal gas constant, T is the temperature, $[D^+]$ and $[D]$ are the concentrations of the oxidized and reduced forms of the electron-donating half reaction, $[A]$ and $[A^-]$ are the concentrations of the oxidized and reduced form of the electron-accepting half reaction. Since we were only interested in the relative change of activity for the electron transport chain components, we focused on calculating the term after v_+ in equation 7.6. The difference in standard redox potential as termed ΔE° is calculated based on the standard redox potential of the half-reactions from multiple sources [71, 245, 246]. The membrane potential $\Delta\psi$ at the specific external pH is calculated based on the experimental measurements by Felle et al [244].

The concentrations of the electron donors and acceptors are obtained from the experimental measurements by Bennett et al [28].

7.5.6 Comparison of DEGs between ME-model predictions and RNA sequencing data

We computed the amount of individual proteins expressed in the ME-model and determined the relative change of each protein expression from neutral pH to acidic pH. We compared the change in protein expression to the DEGs in the RNA sequencing data in terms of the direction of change. For a more systematic comparison of DEGs, we grouped the *E. coli* genes into cellular processes based on COG annotation. Different *E. coli* strains have different sets of DEGs under acid stress in the RNA-seq data, with a small set of DEGs overlapping. Thus, we compared the DEGs found in each strain against the DEGs predicted by the ME-model and grouped the correctly and incorrectly predicted DEGs by COG categories. To obtain the set of genes consistent between model predictions and RNA-seq data, we obtained the list of COG categories commonly found across all five *E. coli* strains in which the correctly predicted genes fall. For each COG category, we then summarized the list of correctly predicted genes from all five *E. coli* strains.

Acknowledgments

We thank Ke Chen for the valuable discussions. This work was supported by National Institute of General Medical Sciences of the National Institutes of Health Grant R01GM057089 and Novo Nordisk Foundation Grant NNF10CC1016517.

Chapter 7 in full is a reprint of the material: **Bin Du**, Laurence Yang, Colton J. Lloyd,

Xin Fang, Bernhard O. Palsson. "Genome-scale model of metabolism and gene expression provides a multi-scale description of acid stress responses in *Escherichia coli*." *Submitted*. The dissertation author was the primary author.

Chapter 8

Conclusion

In this dissertation, I have demonstrated the application of multi-scale modeling to address biological problems from different angles. Using kinetic models, we are able to interpret the dynamic motions of different components over time. The use of thermodynamics in the context of genome-scale models provide an interesting view towards the evolution of metabolic network structure in different organisms. Furthermore, the detailed description of macromolecular components by ME-models allows us to characterize the response of *E. coli* under acid stress at the molecular level. I would like to expand a bit further on some possible future work regarding the different types of multi-scale modeling techniques mentioned above.

In terms of the work on dynamic analysis of kinetic models, we mainly focused on the analysis of concentration Jacobian matrix (chapter 2). In fact, the metabolic network is composed of both the concentration and flux components. Thus, the analysis on the flux Jacobian matrix will lead to similar results as on the concentration Jacobian matrix. It is interesting to explore how the concentration mode and flux mode connect at the specific timescale. The analysis of concentration Jacobian matrix and flux Jacobian matrix together can be used to interpret

the dynamics of more eigenvalues than would be possible with either variable set alone, thus simplifying the overall interpretation of the dynamics of the system.

Thermodynamic analysis on metabolic network benefits from accurate measurement or estimation of equilibrium constants. While more measurements of equilibrium constants are underway, the accurate estimation of equilibrium constants should also be emphasized and actively developed. We have discussed several existing issues in chapter 3. To address them, we can focus on various aspects including improving data quality, expanding the scope of available data, developing more detailed modeling for ion interactions, adding more features (e.g. molecular properties) to train for machine learning models, or even developing novel approaches other than the group contribution method (e.g. quantum thermodynamics). A lot of exciting work to be done in this field.

While ME-models are capable of recapitulating the macromolecular machinery and metabolic flux state, existing ME-models are still in great need of development, as 40% of the proteome in mass are not accounted for in the current *E. coli* ME-model. One such area that needs improvement is the description of cellular processes in the periplasmic space. Only a limited number of periplasmic proteins are actively expressed based on the current *E. coli* ME-model simulations. The other periplasmic proteins are never active due to the lack of description of the downstream processes or metabolic reactions they catalyze. As periplasmic space serves as the buffer layer between the cell and the environment, the characterization of related processes can provide useful insights into the interaction of the cell with the environment, especially in cases where conditions deviate from normal growth environments, e.g. acid stress, high pH stress, oxidative stress, osmotic shock, etc.

Bibliography

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

2. Smallbone, K., Simeonidis, E., Swainston, N. & Mendes, P. Towards a genome-scale kinetic model of cellular metabolism. *BMC Syst. Biol.* **4,** 6 (Jan. 2010).

3. Kotte, O., Zaugg, J. B. & Heinemann, M. Bacterial adaptation through distributed sensing of metabolic fluxes. *Mol. Syst. Biol.* **6,** 355 (Mar. 2010).

4. Joshi, A. & Palsson, B. O. Metabolic dynamics in the human red cell. Part I–A comprehensive kinetic model. *J. Theor. Biol.* **141,** 515–528 (Dec. 1989).

5. Jamshidi, N. & Palsson, B. Ø. Mass action stoichiometric simulation models: incorporating kinetics and regulation into stoichiometric models. *Biophys. J.* **98,** 175–185 (Jan. 2010).

6. Takahashi, K., Ishikawa, N., Sadamoto, Y., Sasamoto, H., Ohta, S., Shiozawa, A., Miyoshi, F., Naito, Y., Nakayama, Y. & Tomita, M. E-Cell 2: multi-platform E-Cell simulation system. *Bioinformatics* **19,** 1727–1729 (Sept. 2003).

7. Bordbar, A., Nagarajan, H., Lewis, N. E., Latif, H., Ebrahim, A., Federowicz, S., Schellenberger, J. & Palsson, B. O. Minimal metabolic pathway structure is consistent with associated biomolecular interactions. *Mol. Syst. Biol.* **10** (July 2014).

8. Salter, M., Knowles, R. G. & Pogson, C. I. Metabolic control. *Essays Biochem.* **28,** 1–12 (1994).

9. Liu, Y.-Y., Slotine, J.-J. & Barabási, A.-L. Controllability of complex networks. *Nature* **473,** 167–173 (May 2011).

10. Bier, M., Bakker, B. M. & Westerhoff, H. V. How Yeast Cells Synchronize their Glycolytic Oscillations: A Perturbation Analytic Treatment. *Biophys. J.* **78,** 1087–1093 (Mar. 2000).

11. Fell, D. A. Metabolic control analysis: a survey of its theoretical and experimental development. *Biochem. J* **286 (Pt 2),** 313–330 (Sept. 1992).

12. Jamshidi, N. & Palsson, B. Ø. Systems biology of SNPs. *Mol. Syst. Biol.* **2,** 38 (July 2006).

13. Jamshidi, N., Wiback, S. J. & Palsson B, B. Ø. In silico model-driven assessment of the effects of single nucleotide polymorphisms (SNPs) on human red blood cell metabolism. *Genome Res.* **12,** 1687–1692 (Nov. 2002).

14. Almquist, J., Cvijovic, M., Hatzimanikatis, V., Nielsen, J. & Jirstrand, M. Kinetic models in industrial biotechnology - Improving cell factory performance. *Metab. Eng.* **24,** 38–60 (July 2014).

15. Heijnen, J. J. Approximative kinetic formats used in metabolic network modeling. *Biotechnol. Bioeng.* **91,** 534–545 (Sept. 2005).

16. Hadlich, F., Noack, S. & Wiechert, W. Translating biochemical network models between different kinetic formats. *Metab. Eng.* **11,** 87–100 (Mar. 2009).

17. Dräger, A., Kronfeld, M., Ziller, M. J., Supper, J., Planatscher, H., Magnus, J. B., Oldiges, M., Kohlbacher, O. & Zell, A. Modeling metabolic networks in C. glutamicum: a comparison of rate laws in combination with various parameter optimization strategies. *BMC Syst. Biol.* **3,** 1–24 (2009).

18. Chassagnole, C., Noisommit-Rizzi, N., Schmid, J. W., Mauch, K. & Reuss, M. Dynamic modeling of the central carbon metabolism of Escherichia coli. *Biotechnol. Bioeng.* **79,** 53–73 (July 2002).

19. Visser, D. & Heijnen, J. J. Dynamic simulation and metabolic re-design of a branched pathway using linlog kinetics. *Metab. Eng.* **5,** 164–176 (July 2003).

20. Briggs, G. E. & Haldane, J. B. A Note on the Kinetics of Enzyme Action. *Biochem. J* **19,** 338–339 (1925).

21. Palsson, B. O. & Lightfoot, E. N. Mathematical modelling of dynamics and control in metabolic networks. I. On Michaelis-Menten kinetics. *J. Theor. Biol.* **111,** 273–302 (Nov. 1984).

22. Schnell, S. Validity of the Michaelis-Menten equation–steady-state or reactant stationary assumption: that is the question. *FEBS J.* **281,** 464–472 (Jan. 2014).

23. Segel, L. A. & Slemrod, M. The Quasi-Steady-State Assumption: A Case Study in Perturbation. *SIAM Rev.* **31,** 446–477 (1989).

24. Tzafriri, A. R. Michaelis-Menten kinetics at high enzyme concentrations. *Bull. Math. Biol.* **65,** 1111–1129 (Nov. 2003).

25. Sanft, K. R., Gillespie, D. T. & Petzold, L. R. Legitimacy of the stochastic Michaelis–Menten approximation. *IET Syst. Biol.* **5,** 58–69 (Jan. 2011).

26. Segel, L. A. On the validity of the steady state assumption of enzyme kinetics. *Bull. Math. Biol.* **50,** 579–593 (1988).

27. Palsson, B. O. On the dynamics of the irreversible Michaelis-Menten reaction mechanism. *Chem. Eng. Sci.* **42,** 447–458 (1987).

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

29. Liebermeister, W. & Klipp, E. Bringing metabolic networks to life: convenience rate law and thermodynamic constraints. *Theor. Biol. Med. Model.* **3,** 41 (2006).

30. Canelas, A. B., Ras, C., ten Pierick, A., van Gulik, W. M. & Heijnen, J. J. An in vivo data-driven framework for classification and quantification of enzyme kinetics and determination of apparent thermodynamic data. *Metab. Eng.* **13,** 294–306 (May 2011).

31. Noor, E., Flamholz, A., Liebermeister, W., Bar-Even, A. & Milo, R. A note on the kinetics of enzyme action: A decomposition that highlights thermodynamic effects. *FEBS Lett.* **587,** 2772–2777 (Sept. 2013).

32. Grima, R. & Schnell, S. Modelling reaction kinetics inside cells. *Essays Biochem.* **45,** 41–56 (2008).

33. Bordbar, A., McCloskey, D., Zielinski, D. C., Sonnenschein, N., Jamshidi, N. & Palsson, B. O. Personalized Whole-Cell Kinetic Models of Metabolism for Discovery in Genomics and Pharmacodynamics. *cels* **1,** 283–292 (Oct. 2015).

34. Mulquiney, P. J., Bubb, W. A. & Kuchel, P. W. Model of 2,3-bisphosphoglycerate metabolism in the human erythrocyte based on detailed enzyme kinetic equations: in vivo kinetic characterization of 2,3-bisphosphoglycerate synthase/phosphatase using 13C and 31P NMR. en. *Biochem. J* **342 Pt 3,** 567–580 (Sept. 1999).

35. Schellenberger, J. & Palsson, B. Ø. Use of Randomized Sampling for Analysis of Metabolic Networks. *J. Biol. Chem.* **284,** 5457–5461 (Feb. 2009).

36. Alon, U. *An introduction to systems biology: design principles of biological circuits* (CRC press, 2006).

37. Kinoshita, A., Tsukada, K., Soga, T., Hishiki, T., Ueno, Y., Nakayama, Y., Tomita, M. & Suematsu, M. Roles of hemoglobin Allostery in hypoxia-induced metabolic alterations in erythrocytes: simulation and its verification by metabolome analysis. *J. Biol. Chem.* **282,** 10731–10741 (Apr. 2007).

38. Scheer, M., Grote, A., Chang, A., Schomburg, I., Munaretto, C., Rother, M., Söhngen, C., Stelzer, M., Thiele, J. & Schomburg, D. BRENDA, the enzyme information system in 2011. *Nucleic Acids Res.* **39,** D670–6 (Jan. 2011).

39. Zamboni, N., Kümmel, A. & Heinemann, M. anNET: a tool for network-embedded thermodynamic analysis of quantitative metabolome data. *BMC Bioinformatics* **9,** 1–11 (2008).

40. Flamholz, A., Noor, E., Bar-Even, A. & Milo, R. eQuilibrator–the biochemical thermodynamics calculator. *Nucleic Acids Res.* **40,** D770–5 (Jan. 2012).

41. Becker, S. A., Feist, A. M., Mo, M. L., Hannum, G., Palsson, B. Ø. & Herrgard, M. J. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat. Protoc.* **2,** 727–738 (2007).

42. Khodayari, A. & Maranas, C. D. A genome-scale Escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. en. *Nat. Commun.* **7,** 13806 (Dec. 2016).

43. Steuer, R. Computational approaches to the topology, stability and dynamics of metabolic networks. en. *Phytochemistry* **68,** 2139–2151 (Aug. 2007).

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

45. Blair, R. H., Trichler, D. L. & Gaille, D. P. Mathematical and statistical modeling in cancer systems biology. en. *Front. Physiol.* **3,** 227 (June 2012).

46. Mannan, A. A., Toya, Y., Shimizu, K., McFadden, J., Kierzek, A. M. & Rocco, A. Integrating Kinetic Model of E . coli with Genome Scale Metabolic Fluxes Overcomes Its Open System Problem and Reveals Bistability in Central Metabolism. *PLoS One* **10,** e0139507 (Oct. 2015).

47. Barua, D., Faeder, J. R. & Haugh, J. M. Structure-based kinetic models of modular signaling protein function: focus on Shp2. en. *Biophys. J.* **92,** 2290–2300 (Apr. 2007).

48. Villaverde, A. F., Bongard, S., Mauch, K., Balsa-Canto, E. & Banga, J. R. Metabolic engineering with multi-objective optimization of kinetic models. en. *J. Biotechnol.* **222,** 1–8 (Mar. 2016).

49. Qian, H., Beard, D. A. & Liang, S.-D. Stoichiometric network theory for nonequilibrium biochemical systems. en. *Eur. J. Biochem.* **270,** 415–421 (Feb. 2003).

50. Heuett, W. J., Beard, D. A. & Qian, H. Linear analysis near a steady-state of biochemical networks: control analysis, correlation metrics and circuit theory. *BMC Syst. Biol.* **2,** 1–11 (2008).

51. Jamshidi, N. & Palsson, B. Ø. Top-down analysis of temporal hierarchy in biochemical reaction networks. *PLoS Comput. Biol.* **4,** e1000177 (Sept. 2008).

52. Steuer, R., Gross, T., Selbig, J. & Blasius, B. Structural kinetic modeling of metabolic networks. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 11868–11873 (Aug. 2006).

53. Hofmeyr, J.-H. S. *Metabolic control analysis in a nutshell* in *Proceedings of the 2nd International conference on systems biology* (2001), 291–300.

54. Du, B., Zielinski, D. C., Kavvas, E. S., Dräger, A., Tan, J., Zhang, Z., Ruggiero, K. E., Arzumanyan, G. A. & Palsson, B. O. Evaluation of rate law approximations in bottom-up kinetic models of metabolism. *BMC Syst. Biol.* **10,** 1–15 (2016).

55. Jamshidi, N. & Palsson, B. Ø. Flux-concentration duality in dynamic nonequilibrium biological networks. *Biophys. J.* **97,** L11–3 (Sept. 2009).

56. Reder, C. Metabolic control theory: A structural approach. *J. Theor. Biol.* **135,** 175–201 (Nov. 1988).

57. Varga, R. S. *Geršgorin and His Circles:* (Springer Berlin Heidelberg, 2004).

58. *The Algebraic Eigenvalue Problem* (ed Wilkinson, J. H.) (Oxford University Press, Inc., New York, NY, USA, 1988).

59. Mackey, L. W. in *Advances in Neural Information Processing Systems 21* (eds Koller, D., Schuurmans, D., Bengio, Y. & Bottou, L.) 1017–1024 (Curran Associates, Inc., 2009).

60. Flach, E. H. & Schnell, S. Stability of open pathways. en. *Math. Biosci.* **228,** 147–152 (Dec. 2010).

61. Ivanov, O., van der Schaft, A. J. & Weissing, F. J. Stability of metabolic pathways with irreversible reactions. *Proceedings of the 21st International Symposium on Mathematical Theory of Networks and Systems (MTNS 2014).* **Groningen, 2014,** 890–893.

62. Flamholz, A., Noor, E., Bar-Even, A., Liebermeister, W. & Milo, R. Glycolytic strategy as a tradeoff between energy yield and protein cost. *Proceedings of the National Academy of Sciences* **110,** 10039–10044 (June 2013).

63. Hamilton, J. J., Dwivedi, V. & Reed, J. L. Quantitative assessment of thermodynamic constraints on the solution space of genome-scale metabolic models. en. *Biophys. J.* **105,** 512–522 (July 2013).

64. Soh, K. C. & Hatzimanikatis, V. Constraining the flux space using thermodynamics and integration of metabolomics data. en. *Methods Mol. Biol.* **1191,** 49–63 (2014).

65. Park, J. O., Rubin, S. A., Xu, Y.-F., Amador-Noguez, D., Fan, J., Shlomi, T. & Rabinowitz, J. D. Metabolite concentrations, fluxes and free energies imply efficient enzyme usage. en. *Nat. Chem. Biol.* **12,** 482–489 (July 2016).

66. Alberty, R. A. *Thermodynamics of biochemical reactions* (Massachusetts Institute of Technology, Cambridge, MA, 2003).

67. Johnson, J. W., Oelkers, E. H. & Helgeson, H. C. SUPCRT92: A software package for calculating the standard molal thermodynamic properties of minerals, gases, aqueous species, and reactions from 1 to 5000 bar and 0 to 1000 C. *Comput. Geosci.* **18,** 899–947 (Aug. 1992).

68. Mavrovouniotis, M. L. Group contributions for estimating standard gibbs energies of formation of biochemical compounds in aqueous solution. *Biotechnol. Bioeng.* **36,** 1070–1082 (Dec. 1990).

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

70. Noor, E., Bar-Even, A., Flamholz, A., Lubling, Y., Davidi, D. & Milo, R. An integrated open framework for thermodynamics of reactions that combines accuracy and coverage. en. *Bioinformatics* **28,** 2037–2044 (Aug. 2012).

71. Noor, E., Haraldsdóttir, H. S., Milo, R. & Fleming, R. M. T. Consistent estimation of Gibbs energy using component contributions. en. *PLoS Comput. Biol.* **9,** e1003098 (July 2013).

72. Du, B., Zhang, Z., Grubner, S., Yurkovich, J. T., Palsson, B. O. & Zielinski, D. C. Temperature-Dependent Estimation of Gibbs Energies Using an Updated Group-Contribution Method. en. *Biophys. J.* **114,** 2691–2702 (June 2018).

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

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

75. Pitzer, K. S. & Kim, J. J. Thermodynamics of electrolytes. IV. Activity and osmotic coefficients for mixed electrolytes. *J. Am. Chem. Soc.* **96,** 5701–5707 (Sept. 1974).

76. Elizalde, M. P. & Aparicio, J. L. Current theories in the calculation of activity coefficients-II. Specific interaction theories applied to some equilibria studies in solution chemistry. en. *Talanta* **42,** 395–400 (Mar. 1995).

77. Ni, N. & Yalkowsky, S. H. Prediction of Setschenow constants. *Int. J. Pharm.* **254,** 167–172 (Mar. 2003).

78. Minton, A. P. How can biochemical reactions within cells differ from those in test tubes? en. *J. Cell Sci.* **119,** 2863–2869 (July 2006).

79. Vilker, V. L., Colton, C. K. & Smith, K. A. The osmotic pressure of concentrated protein solutions: Effect of concentration and ph in saline solutions of bovine serum albumin. *J. Colloid Interface Sci.* **79,** 548–566 (Feb. 1981).

80. Rhys, N. H., Bruni, F., Imberti, S., McLain, S. E. & Ricci, M. A. Glucose and Mannose: A Link between Hydration and Sweetness. en. *J. Phys. Chem. B* **121,** 7771–7776 (Aug. 2017).

81. Guggenheim, E. A. & Turgeon, J. C. Specific interaction of ions. en. *Trans. Faraday Soc.* **51,** 747–761 (Jan. 1955).

82. Pitzer, K. S. Thermodynamics of electrolytes. I. Theoretical basis and general equations. *J. Phys. Chem.* **77,** 268–277 (Jan. 1973).

83. Grenthe, I. & Wanner, H. Guidelines for the Extrapolation to Zero Ionic Strength. *Nuclear Energy Agency* (2000).

84. Mobley, D. L. Experimental and Calculated Small Molecule Hydration Free Energies. *https://escholarship.org/uc/item/6sd403pz* (Jan. 2013).

85. Chamberlin, A. C., Cramer, C. J. & Truhlar, D. G. Predicting aqueous free energies of solvation as functions of temperature. en. *J. Phys. Chem. B* **110,** 5665–5675 (Mar. 2006).

86. Plyasunov, A. V. & Shock, E. L. Thermodynamic functions of hydration of hydrocarbons at 298.15 K and 0.1 MPa. *Geochim. Cosmochim. Acta* **64,** 439–468 (Feb. 2000).

87. Bannan, C. C., Calabró, G., Kyu, D. Y. & Mobley, D. L. Calculating Partition Coefficients of Small Molecules in Octanol/Water and Cyclohexane/Water. *J. Chem. Theory Comput.* **12,** 4015–4024 (Aug. 2016).

88. Genheden, S. Solvation free energies and partition coefficients with the coarse-grained and hybrid all-atom/coarse-grained MARTINI models. en. *J. Comput. Aided Mol. Des.* **31,** 867–876 (Oct. 2017).

89. Garrido, N. M., Queimada, A. J., Jorge, M., Macedo, E. A. & Economou, I. G. 1-Octanol/Water Partition Coefficients of n-Alkanes from Molecular Simulations of Absolute Solvation Free Energies. *J. Chem. Theory Comput.* **5,** 2436–2446 (Sept. 2009).

90. Matubayasi, N., Shinoda, W. & Nakahara, M. Free-energy analysis of the molecular binding into lipid membrane with the method of energy representation. *J. Chem. Phys.* **128,** 195107 (May 2008).

91. Xiang, T.-X. & Anderson, B. D. A Computer Simulation of Functional Group Contributions to Free Energy in Water and a DPPC Lipid Bilayer. *Biophys. J.* **82,** 2052–2066 (Apr. 2002).

92. Rother, K., Hoffmann, S., Bulik, S., Hoppe, A., Gasteiger, J. & Holzhütter, H.-G. IGERS: inferring Gibbs energy changes of biochemical reactions from reaction similarities. en. *Biophys. J.* **98,** 2478–2486 (June 2010).

93. Grover, M., Singh, B., Bakshi, M. & Singh, S. Quantitative structure–property relationships in pharmaceutical research – Part 1. *Pharm. Sci. Technolo. Today* **3,** 28–35 (Jan. 2000).

94. Yousefinejad, S. & Hemmateenejad, B. Chemometrics tools in QSAR/QSPR studies: A historical perspective. *Chemometrics Intellig. Lab. Syst.* **149,** 177–204 (Dec. 2015).

95. Jinich, A., Rappoport, D., Dunn, I., Sanchez-Lengeling, B., Olivares-Amaya, R., Noor, E., Even, A. B. & Aspuru-Guzik, A. Quantum Chemical Approach to Estimating the Thermodynamics of Metabolic Reactions. *Sci. Rep.* **4,** 7022 (Nov. 2014).

96. Lin, S.-T., Maiti, P. K. & Goddard, W. A. Two-Phase Thermodynamic Model for Efficient and Accurate Absolute Entropy of Water from Molecular Dynamics Simulations. *J. Phys. Chem. B* **114,** 8191–8198 (June 2010).

97. Tang, X., Huston, K. J. & Larson, R. G. Molecular Dynamics Simulations of Structure–Property Relationships of Tween 80 Surfactants in Water and at Interfaces. *J. Phys. Chem. B* **118,** 12907–12918 (Nov. 2014).

98. Meng, E. C. & Kollman, P. A. Molecular Dynamics Studies of the Properties of Water around Simple Organic Solutes. *J. Phys. Chem.* **100,** 11460–11470 (Jan. 1996).

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

100. Hamilton, J. J., Dwivedi, V. & Reed, J. L. Quantitative assessment of thermodynamic constraints on the solution space of genome-scale metabolic models. en. *Biophys. J.* **105,** 512–522 (16 7 2013).

101. Kümmel, A., Panke, S. & Heinemann, M. Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. en. *Mol. Syst. Biol.* **2,** 2006.0034 (20 6 2006).

102. Noor, E., Bar-Even, A., Flamholz, A., Reznik, E., Liebermeister, W. & Milo, R. Pathway Thermodynamics Highlights Kinetic Obstacles in Central Metabolism. *PLoS Comput. Biol.* **10,** e1003483 (20 2 2014).

103. Beard, D. A., Vinnakota, K. C. & Wu, F. Detailed Enzyme Kinetics in Terms of Biochemical Species: Study of Citrate Synthase. *PLoS One* **3,** e1825 (19 3 2008).

104. Goldberg, R. N., Tewari, Y. B. & Bhat, T. N. Thermodynamics of enzyme-catalyzed reactions—a database for quantitative biochemistry. *Bioinformatics* **20,** 2874–2877 (Jan. 2004).

105. Mavrovouniotis, M. L. Group contributions for estimating standard gibbs energies of formation of biochemical compounds in aqueous solution. en. *Biotechnol. Bioeng.* **36,** 1070–1082 (May 1990).

106. Noor, E., Bar-Even, A., Flamholz, A., Lubling, Y., Davidi, D. & Milo, R. An integrated open framework for thermodynamics of reactions that combines accuracy and coverage. en. *Bioinformatics* **28,** 2037–2044 (Jan. 2012).

107. Noor, E., Haraldsdóttir, H. S., Milo, R. & Fleming, R. M. T. Consistent Estimation of Gibbs Energy Using Component Contributions. *PLoS Comput. Biol.* **9,** e1003098 (Nov. 2013).

108. Helgeson, H. C. & Kirkham, D. H. Theoretical prediction of the thermodynamic behavior of aqueous electrolytes at high pressures and temperatures; I, Summary of the thermodynamic/electrostatic properties of the solvent. *Am. J. Sci.* **274,** 1089–1198 (1974).

109. Helgeson, H. C. & Kirkham, D. H. Theoretical prediction of the thermodynamic behavior of aqueous electrolytes at high pressures and temperatures; II, Debye-Huckel parameters for activity coefficients and relative partial molal properties. *Am. J. Sci.* **274,** 1199–1261 (1974).

110. Helgeson, H. C. & Kirkham, D. H. Theoretical prediction of thermodynamic properties of aqueous electrolytes at high pressures and temperatures. III. Equation of state for aqueous species at infinite dilution. *Am. J. Sci.* **276** (1976).

111. Helgeson, H. C., Kirkham, D. H. & Flowers, G. C. Theoretical prediction of the thermodynamic behavior of aqueous electrolytes by high pressures and temperatures; IV, Calculation of activity coefficients, osmotic coefficients, and apparent molal and standard and relative partial molal properties to 600 C and 5kb. *Am. J. Sci.* **281,** 1249–1516 (Jan. 1981).

112. Shock, E. L. & Helgeson, H. C. Calculation of the thermodynamic and transport properties of aqueous species at high pressures and temperatures: Correlation algorithms for ionic species and equation of state predictions to 5 kb and 1000 C. *Geochim. Cosmochim. Acta* **52,** 2009–2036 (Jan. 1988).

113. Plyasunov, A. V. & Shock, E. L. Correlation strategy for determining the parameters of the revised Helgeson-Kirkham-Flowers model for aqueous nonelectrolytes. *Geochim. Cosmochim. Acta* **65,** 3879–3900 (Jan. 2001).

114. Plyasunov, A. V., O'Connell, J. P., Wood, R. H. & Shock, E. L. Semiempirical equation of state for the infinite dilution thermodynamic functions of hydration of nonelectrolytes over wide ranges of temperature and pressure. *Fluid Phase Equilib.* **183,** 133–142 (Jan. 2001).

115. Johnson, J. W., Oelkers, E. H. & Helgeson, H. C. SUPCRT92: A software package for calculating the standard molal thermodynamic properties of minerals, gases, aqueous species, and reactions from 1 to 5000 bar and 0 to 1000 C. *Comput. Geosci.* **18,** 899–947 (Jan. 1992).

116. Plyasunova, N. V., Plyasunov, A. V. & Shock, E. L. Database of Thermodynamic Properties for Aqueous Organic Compounds. en. *Int. J. Thermophys.* **25,** 351–360 (Jan. 2004).

117. Pettit, L. D. & Powell, K. J. The IUPAC Stability Constants Database. *Chemistry International – Newsmagazine for IUPAC* **28** (2006).

118. Kortüm, G. & Andrussow, K. *Dissociation constants of organic acids in aqueous solution* (Butterworths, 1961).

119. Perrln, D. D. *Dissociation constants of organic bases in aqueous solutions* 1965.

120. Alberty, R. A. Effect of pH and Metal Ion Concentration on the Equilibrium Hydrolysis of Adenosine Triphosphate to Adenosine Diphosphate. *J. Biol. Chem.* **243,** 1337–1343 (Oct. 1968).

121. Davies, C. W. 397. The extent of dissociation of salts in water. Part VIII. An equation for the mean ionic activity coefficient of an electrolyte in water, and a revision of the dissociation constants of some sulphates. *J. Chem. Soc.* 2093–2098 (1938).

122. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).

123. Goldberg, R. N. & Tewari, Y. B. Thermodynamics of the disproportionation of adenosine 5'-diphosphate to adenosine 5'-triphosphate and adenosine 5'-monophosphate. I. Equilibrium model. *Biophys. Chem.* **40,** 241–261 (July 1991).

124. Larson, J. W., Tewari, Y. B. & Goldberg, R. N. Thermochemistry of the reactions between adenosine, adenosine 5'-monophosphate, inosine, and inosine 5'-monophosphate; the conversion of d-histidine to (urocanic acid+ammonia). *J. Chem. Thermodyn.* **25,** 73–90 (Jan. 1993).

125. Cayley, S., Lewis, B. A., Guttman, H. J. & Record Jr, M. T. Characterization of the cytoplasm of Escherichia coli K-12 as a function of external osmolarity. Implications for protein-DNA interactions in vivo. *J. Mol. Biol.* **222,** 281–300 (20 11 1991).

126. Helgeson, H. C. Calculation of the thermodynamic properties and relative stabilities of aqueous acetic and chloroacetic acids, acetate and chloroacetates, and acetyl and chloroacetyl chlorides at high and low temperatures and pressures. *Appl. Geochem.* **7,** 291–308 (1992).

127. Shock, E. L. Organic acids in hydrothermal solutions: standard molal thermodynamic properties of carboxylic acids and estimates of dissociation constants at high temperatures and pressures. en. *Am. J. Sci.* **295,** 496–580 (May 1995).

128. Schulte, M. D. & Rogers, K. L. Thiols in hydrothermal solution: standard partial molal properties and their role in the organic geochemistry of hydrothermal environments. *Geochim. Cosmochim. Acta* **68,** 1087–1097 (2004).

129. Schulte, M. D. & Shock, E. L. Aldehydes in hydrothermal solution: standard partial molal thermodynamic properties and relative stabilities at high temperatures and pressures. en. *Geochim. Cosmochim. Acta* **57,** 3835–3846 (1993).

130. Amend, J. P. & Shock, E. L. Energetics of amino acid synthesis in hydrothermal ecosystems. en. *Science* **281,** 1659–1662 (Nov. 1998).

131. Peregrín-Alvarez, J. M., Sanford, C. & Parkinson, J. The conservation and evolutionary modularity of metabolism. en. *Genome Biol.* **10,** R63 (June 2009).

132. Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H.,

Pfannkoch, C., Rogers, Y.-H. & Smith, H. O. Environmental genome shotgun sequencing of the Sargasso Sea. en. *Science* **304,** 66–74 (Apr. 2004).

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

134. Hatzimanikatis, V., Li, C., Ionita, J. A., Henry, C. S., Jankowski, M. D. & Broadbelt, L. J. Exploring the diversity of complex metabolic networks. en. *Bioinformatics* **21,** 1603–1609 (Apr. 2005).

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

136. Horikoshi, K., Antranikian, G., Bull, A. T., Robb, F. T. & Stetter, K. O. *Extremophiles Handbook* (eds Horikoshi, K., Antranikian, G., Bull, A. T., Robb, F. T. & Stetter, K. O.) 1271 (Springer, Aug. 2010).

137. Barrie Johnson, D. & Hallberg, K. B. Carbon, iron and sulfur metabolism in acidophilic micro-organisms. en. *Adv. Microb. Physiol.* **54,** 201–255 (2009).

138. Steunou, A.-S., Bhaya, D., Bateson, M. M., Melendrez, M. C., Ward, D. M., Brecht, E., Peters, J. W., Kühl, M. & Grossman, A. R. In situ analysis of nitrogen fixation and metabolic switching in unicellular thermophilic cyanobacteria inhabiting hot spring microbial mats. en. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 2398–2403 (Feb. 2006).

139. Falb, M., Müller, K., Königsmaier, L., Oberwinkler, T., Horn, P., von Gronau, S., Gonzalez, O., Pfeiffer, F., Bornberg-Bauer, E. & Oesterhelt, D. Metabolism of halophilic archaea. en. *Extremophiles* **12,** 177–196 (Mar. 2008).

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

141. Santos, F., Boele, J. & Teusink, B. A practical guide to genome-scale metabolic models and their analysis. en. *Methods Enzymol.* **500,** 509–532 (2011).

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

143. Ma, H. & Zeng, A.-P. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* **19,** 270–277 (Jan. 2003).

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

145. Blank, L. M., Ebert, B. E., Bühler, B. & Schmid, A. Metabolic capacity estimation of Escherichia coli as a platform for redox biocatalysis: constraint-based modeling and experimental verification. *Biotechnol. Bioeng.* **100,** 1050–1065 (2008).

146. Campodonico, M. A., Andrews, B. A., Asenjo, J. A., Palsson, B. O. & Feist, A. M. Generation of an atlas for commodity chemical production in Escherichia coli and a novel pathway prediction algorithm, GEM-Path. en. *Metab. Eng.* **25,** 140–158 (Sept. 2014).

147. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. en. *Nucleic Acids Res.* **28,** 27–30 (Jan. 2000).

148. Caspi, R., Altman, T., Dreher, K., Fulcher, C. A., Subhraveti, P., Keseler, I. M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Pujar, A., Shearer, A. G., Travers, M., Weerasinghe, D., Zhang, P. & Karp, P. D. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. en. *Nucleic Acids Res.* **40,** D742–53 (Jan. 2012).

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

150. Kukko, E. & Heinonen, J. The intracellular concentration of pyrophosphate in the batch culture of Escherichia coli. en. *Eur. J. Biochem.* **127,** 347–349 (Oct. 1982).

151. Asplund-Samuelsson, J., Janasch, M. & Hudson, E. P. Thermodynamic analysis of computed pathways integrated into the metabolic networks of E. coli and Synechocystis reveals contrasting expansion potential. en. *Metab. Eng.* **45,** 223–236 (Jan. 2018).

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

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

154. Mahadevan, R., Bond, D. R., Butler, J. E., Esteve-Nuñez, A., Coppi, M. V., Palsson, B. O., Schilling, C. H. & Lovley, D. R. Characterization of metabolism in the Fe(III)-reducing organism Geobacter sulfurreducens by constraint-based modeling. en. *Appl. Environ. Microbiol.* **72,** 1558–1568 (Feb. 2006).

155. Benedict, M. N., Gonnerman, M. C., Metcalf, W. W. & Price, N. D. Genome-Scale Metabolic Reconstruction and Hypothesis Testing in the Methanogenic Archaeon Methanosarcina acetivorans C2A. *J. Bacteriol.* **194,** 855–865 (Feb. 2012).

156. Ekiel, I., Smith, I. C. P. & Sprott, G. D. Biosynthesis of isoleucine in methanogenic bacteria: a carbon-13 NMR study. *Biochemistry* **23,** 1683–1687 (Apr. 1984).

157. Charon, N. W., Johnson, R. C. & Peterson, D. Amino acid biosynthesis in the spirochete Leptospira: evidence for a novel pathway of isoleucine biosynthesis. en. *J. Bacteriol.* **117,** 203–211 (Jan. 1974).

158. Risso, C., Van Dien, S. J., Orloff, A., Lovley, D. R. & Coppi, M. V. Elucidation of an Alternate Isoleucine Biosynthesis Pathway in Geobacter sulfurreducens. *J. Bacteriol.* **190,** 2266–2274 (Apr. 2008).

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

160. Feist, A. M., Nagarajan, H., Rotaru, A.-E., Tremblay, P.-L., Zhang, T., Nevin, K. P., Lovley, D. R. & Zengler, K. Constraint-Based Modeling of Carbon Fixation and the Energetics of Electron Transfer in Geobacter metallireducens. *PLoS Comput. Biol.* **10,** e1003575 (Apr. 2014).

161. Razvi, A. & Scholtz, J. M. Lessons in stability from thermophilic proteins. en. *Protein Sci.* **15,** 1569–1578 (July 2006).

162. Koga, Y. Thermal adaptation of the archaeal and bacterial lipid membranes. en. *Archaea* **2012,** 789652 (Aug. 2012).

163. Small, P., Blankenhorn, D., Welty, D., Zinser, E. & Slonczewski, J. L. Acid and base resistance in Escherichia coli and Shigella flexneri: role of rpoS and growth pH. en. *J. Bacteriol.* **176,** 1729–1737 (Mar. 1994).

164. Lin, J., Lee, I. S., Frey, J., Slonczewski, J. L. & Foster, J. W. Comparative analysis of extreme acid survival in Salmonella typhimurium, Shigella flexneri, and Escherichia coli. en. *J. Bacteriol.* **177,** 4097–4104 (July 1995).

165. Lin, J., Smith, M. P., Chapin, K. C., Baik, H. S., Bennett, G. N. & Foster, J. W. Mechanisms of acid resistance in enterohemorrhagic Escherichia coli. en. *Appl. Environ. Microbiol.* **62,** 3094–3100 (Sept. 1996).

166. Conner, D. E. & Kotrola, J. S. Growth and survival of Escherichia coli O157:H7 under acidic conditions. en. *Appl. Environ. Microbiol.* **61,** 382–385 (Jan. 1995).

167. Dlamini, B. C. & Buys, E. M. Survival and growth of acid adapted Escherichia coli strains in broth at different pH levels. *J. Food Saf.* **29,** 484–497 (Aug. 2009).

168. Vivijs, B., Aertsen, A. & Michiels, C. W. Identification of Genes Required for Growth of Escherichia coli MG1655 at Moderately Low pH. en. *Front. Microbiol.* **7,** 1672 (Oct. 2016).

169. Evans, D. F., Pye, G., Bramley, R., Clark, A. G., Dyson, T. J. & Hardcastle, J. D. Measurement of gastrointestinal pH profiles in normal ambulant human subjects. en. *Gut* **29,** 1035–1041 (Aug. 1988).

170. Ibekwe, V. C., Fadda, H. M., McConnell, E. L., Khela, M. K., Evans, D. F. & Basit, A. W. Interplay between intestinal pH, transit time and feed status on the in vivo performance of pH responsive ileo-colonic release systems. en. *Pharm. Res.* **25,** 1828–1835 (Aug. 2008).

171. Slonczewski, J. L., Fujisawa, M., Dopson, M. & Krulwich, T. A. Cytoplasmic pH measurement and homeostasis in bacteria and archaea. en. *Adv. Microb. Physiol.* **55,** 1–79, 317 (2009).

172. Castanié-Cornet, M.-P., Treffandier, H., Francez-Charlot, A., Gutierrez, C. & Cam, K. The glutamate-dependent acid resistance system in Escherichia coli: essential and dual role of the His-Asp phosphorelay RcsCDB/AF. en. *Microbiology* **153,** 238–246 (Jan. 2007).

173. Richard, H. & Foster, J. W. Escherichia coli glutamate- and arginine-dependent acid resistance systems increase internal pH and reverse transmembrane potential. en. *J. Bacteriol.* **186,** 6032–6041 (Sept. 2004).

174. Diez-Gonzalez, F. & Karaibrahimoglu, Y. Comparison of the glutamate-, arginine- and lysine-dependent acid resistance systems in Escherichia coli O157:H7. *J. Appl. Microbiol.* **96,** 1237–1244 (2004).

175. Kashiwagi, K., Suzuki, T., Suzuki, F., Furuchi, T., Kobayashi, H. & Igarashi, K. Coexistence of the genes for putrescine transport protein and ornithine decarboxylase at 16 min on Escherichia coli chromosome. en. *J. Biol. Chem.* **266,** 20922–20927 (Nov. 1991).

176. Hayes, E. T., Wilks, J. C., Sanfilippo, P., Yohannes, E., Tate, D. P., Jones, B. D., Radmacher, M. D., BonDurant, S. S. & Slonczewski, J. L. Oxygen limitation modulates pH regulation of catabolism and hydrogenases, multidrug transporters, and envelope composition in Escherichia coli K-12. en. *BMC Microbiol.* **6,** 89 (Oct. 2006).

177. Maurer, L. M., Yohannes, E., Bondurant, S. S., Radmacher, M. & Slonczewski, J. L. pH regulates genes for flagellar motility, catabolism, and oxidative stress in Escherichia coli K-12. en. *J. Bacteriol.* **187,** 304–319 (Jan. 2005).

178. Hong, W., Wu, Y. E., Fu, X. & Chang, Z. Chaperone-dependent mechanisms for acid resistance in enteric bacteria. en. *Trends Microbiol.* **20,** 328–335 (July 2012).

179. Brown, J. L., Ross, T., McMeekin, T. A. & Nichols, P. D. Acid habituation of Escherichia coli and the potential role of cyclopropane fatty acids in low pH tolerance. en. *Int. J. Food Microbiol.* **37,** 163–173 (July 1997).

180. Chang, Y. Y. & Cronan Jr, J. E. Membrane cyclopropane fatty acid content is a major factor in acid resistance of Escherichia coli. en. *Mol. Microbiol.* **33,** 249–259 (July 1999).

181. delaVega, A. L. & Delcour, A. H. Cadaverine induces closing of E. coli porins. en. *EMBO J.* **14,** 6058–6065 (Dec. 1995).

182. Rowbury, R. J., Goodson, M. & Wallace, A. D. The PhoE porin and transmission of the chemical stimulus for induction of acid resistance (acid habituation) in Escherichia coli. en. *J. Appl. Bacteriol.* **72,** 233–243 (Mar. 1992).

183. Kanjee, U. & Houry, W. A. Mechanisms of acid resistance in Escherichia coli. en. *Annu. Rev. Microbiol.* **67,** 65–81 (May 2013).

184. Dragosits, M. & Mattanovich, D. Adaptive laboratory evolution – principles and applications for biotechnology. en. *Microb. Cell Fact.* **12,** 64 (July 2013).

185. Buermans, H. P. J. & den Dunnen, J. T. Next generation sequencing technology: Advances and applications. en. *Biochim. Biophys. Acta* **1842,** 1932–1941 (Oct. 2014).

186. Harden, M. M., He, A., Creamer, K., Clark, M. W., Hamdallah, I., Martinez, K. A., Kresslein, R. L., Bush, S. P. & Slonczewski, J. L. Acid-adapted strains of Escherichia coli K-12 obtained by experimental evolution. *Appl. Environ. Microbiol.* (Jan. 2015).

187. He, A., Penix, S. R., Basting, P. J., Griffith, J. M., Creamer, K. E., Camperchioli, D., Clark, M. W., Gonzales, A. S., Chavez Erazo, J. S., George, N. S., Bhagwat, A. A. & Slonczewski, J. L. Acid evolution deletes amino-acid decarboxylases and reregulates catabolism of Escherichia coli K-12. *Appl. Environ. Microbiol.* (Apr. 2017).

188. LaCroix, R. A., Sandberg, T. E., O'Brien, E. J., Utrilla, J., Ebrahim, A., Guzman, G. I., Szubin, R., Palsson, B. O. & Feist, A. M. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. en. *Appl. Environ. Microbiol.* **81,** 17–30 (Jan. 2015).

189. Wood, E. J. Data for Biochemical Research (third edition). *Biochem. Educ.* **15,** 97 (Apr. 1987).

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

191. Helmann, J. D. RNA polymerase: a nexus of gene regulation. en. *Methods* **47,** 1–5 (Jan. 2009).

192. Klein-Marcuschamer, D., Santos, C. N. S., Yu, H. & Stephanopoulos, G. Mutagenesis of the bacterial RNA polymerase alpha subunit for improvement of complex phenotypes. en. *Appl. Environ. Microbiol.* **75,** 2705–2711 (May 2009).

193. Sandberg, T. E., Pedersen, M., LaCroix, R. A., Ebrahim, A., Bonde, M., Herrgard, M. J., Palsson, B. O., Sommer, M. & Feist, A. M. Evolution of Escherichia coli to 42 C and subsequent genetic engineering reveals adaptive mechanisms and novel mutations. en. *Mol. Biol. Evol.* **31,** 2647–2662 (Oct. 2014).

194. Sandberg, T. E., Lloyd, C. J., Palsson, B. O. & Feist, A. M. Laboratory Evolution to Alternating Substrate Environments Yields Distinct Phenotypic and Genetic Adaptive Strategies. en. *Appl. Environ. Microbiol.* **83** (July 2017).

195. McCloskey, D., Xu, S., Sandberg, T. E., Brunk, E., Hefner, Y., Szubin, R., Feist, A. M. & Palsson, B. O. Evolution of gene knockout strains of E. coli reveal regulatory architectures governed by metabolism. en. *Nat. Commun.* **9,** 3796 (Sept. 2018).

196. Conrad, T. M., Frazier, M., Joyce, A. R., Cho, B.-K., Knight, E. M., Lewis, N. E., Landick, R. & Palsson, B. Ø. RNA polymerase mutants found through adaptive evolution reprogram Escherichia coli for optimal growth in minimal media. en. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 20500–20505 (Nov. 2010).

197. Cheng, K.-K., Lee, B.-S., Masuda, T., Ito, T., Ikeda, K., Hirayama, A., Deng, L., Dong, J., Shimizu, K., Soga, T., Tomita, M., Palsson, B. O. & Robert, M. Global metabolic network reorganization by adaptive mutations allows fast growth of Escherichia coli on glycerol. en. *Nat. Commun.* **5,** 3233 (2014).

198. Wytock, T. P., Fiebig, A., Willett, J. W., Herrou, J., Fergin, A., Motter, A. E. & Crosson, S. Experimental evolution of diverse Escherichia coli metabolic mutants identifies genetic loci for convergent adaptation of growth rate. en. *PLoS Genet.* **14,** e1007284 (Mar. 2018).

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

200. Severinov, K., Mustaev, A., Kukarin, A., Muzzin, O., Bass, I., Darst, S. A. & Goldfarb, A. Structural modules of the large subunits of RNA polymerase. Introducing archaebacterial and chloroplast split sites in the beta and beta' subunits of Escherichia coli RNA polymerase. en. *J. Biol. Chem.* **271,** 27969–27974 (Nov. 1996).

201. Cramer, P., Bushnell, D. A. & Kornberg, R. D. Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. en. *Science* **292,** 1863–1876 (June 2001).

202. Sastry, A. V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., Choudhary, K. S., Yang, L., King, Z. A. & Palsson, B. O. P. *The Escherichia coli Transcriptome Consists of Independently Regulated Modules*

203. Foster, J. W. Escherichia coli acid resistance: tales of an amateur acidophile. *Nat. Rev. Microbiol.* **2,** 898–907 (Nov. 2004).

204. Lund, P., Tramonti, A. & De Biase, D. Coping with low pH: molecular strategies in neutralophilic bacteria. *FEMS Microbiol. Rev.* **38,** 1091–1125 (Nov. 2014).

205. Marotz, C., Amir, A., Humphrey, G., Gaffney, J., Gogul, G. & Knight, R. DNA extraction for streamlined metagenomics of diverse environmental samples. en. *Biotechniques* **62,** 290–293 (June 2017).

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

207. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. en. *Genome Biol.* **10,** R25 (Mar. 2009).

208. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T. & Carey, V. J. Software for computing and annotating genomic ranges. en. *PLoS Comput. Biol.* **9,** e1003118 (Aug. 2013).

209. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. en. *Genome Biol.* **15,** 550 (2014).

210. Lee, J., Page, R., García-Contreras, R., Palermino, J.-M., Zhang, X.-S., Doshi, O., Wood, T. K. & Peti, W. Structure and function of the Escherichia coli protein YmgB: a protein critical for biofilm formation and acid-resistance. en. *J. Mol. Biol.* **373,** 11–26 (Oct. 2007).

211. Hobbs, E. C., Astarita, J. L. & Storz, G. Small RNAs and small proteins involved in resistance to cell envelope stress and acid shock in Escherichia coli: analysis of a bar-coded mutant collection. en. *J. Bacteriol.* **192,** 59–67 (Jan. 2010).

212. Choi, S. H., Baumler, D. J. & Kaspar, C. W. Contribution of dps to acid stress tolerance and oxidative stress tolerance in Escherichia coli O157:H7. en. *Appl. Environ. Microbiol.* **66,** 3911–3916 (Sept. 2000).

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

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

215. Yang, L., Yurkovich, J. T., Lloyd, C. J., Ebrahim, A., Saunders, M. A. & Palsson, B. O. Principles of proteome allocation are revealed using proteomic data and genome-scale models. en. *Sci. Rep.* **6,** 36734 (Nov. 2016).

216. Jordan, K. N., Oxford, L. & O'Byrne, C. P. Survival of low-pH stress by Escherichia coli O157:H7: correlation between alterations in the cell envelope and increased acid tolerance. en. *Appl. Environ. Microbiol.* **65,** 3048–3055 (July 1999).

217. Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Expanded microbial genome coverage and improved protein family annotation in the COG database. en. *Nucleic Acids Res.* **43,** D261–9 (Jan. 2015).

218. Russell, N. J., Evans, R. I., ter Steeg, P. F., Hellemons, J., Verheul, A. & Abee, T. Membranes as a target for stress adaptation. en. *Int. J. Food Microbiol.* **28,** 255–261 (Dec. 1995).

219. Yuk, H.-G. & Marshall, D. L. Adaptation of Escherichia coli O157:H7 to pH alters membrane lipid composition, verotoxin secretion, and resistance to simulated gastric fluid acid. en. *Appl. Environ. Microbiol.* **70,** 3500–3505 (June 2004).

220. Liu, J. K., O'Brien, E. J., Lerman, J. A., Zengler, K., Palsson, B. O. & Feist, A. M. Reconstruction and modeling protein translocation and compartmentalization in Escherichia coli at the genome-scale. en. *BMC Syst. Biol.* **8,** 110 (Sept. 2014).

221. Slonczewski, J. L., Rosen, B. P., Alger, J. R. & Macnab, R. M. pH homeostasis in Escherichia coli: measurement by 31P nuclear magnetic resonance of methylphosphonate and phosphate. en. *Proc. Natl. Acad. Sci. U. S. A.* **78,** 6271–6275 (Oct. 1981).

222. Wilks, J. C. & Slonczewski, J. L. pH of the cytoplasm and periplasm of Escherichia coli: rapid measurement by green fluorescent protein fluorimetry. en. *J. Bacteriol.* **189,** 5601–5607 (Aug. 2007).

223. Ghosh, K. & Dill, K. A. Computing protein stabilities from their chain lengths. en. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 10649–10654 (June 2009).

224. Alexov, E. Numerical calculations of the pH of maximal protein stability. The effect of the sequence composition and three-dimensional structure. en. *Eur. J. Biochem.* **271,** 173–185 (Jan. 2004).

225. Schaefer, M., Sommer, M. & Karplus, M. pH-Dependence of Protein Stability: Absolute Electrostatic Free Energy Differences between Conformations. *J. Phys. Chem. B* **101,** 1663–1683 (Feb. 1997).

226. Yang, A. S. & Honig, B. On the pH dependence of protein stability. en. *J. Mol. Biol.* **231,** 459–474 (May 1993).

227. Dill, K. A., Ghosh, K. & Schmit, J. D. Physical limits of cells and proteomes. en. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 17876–17882 (Nov. 2011).

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

229. Ding, J., Yang, C., Niu, X., Hu, Y. & Jin, C. HdeB chaperone activity is coupled to its intrinsic dynamic properties. en. *Sci. Rep.* **5,** 16856 (Nov. 2015).

230. Sperandeo, P., Cescutti, R., Villa, R., Di Benedetto, C., Candia, D., Dehò, G. & Polissi, A. Characterization of lptA and lptB, two essential genes implicated in lipopolysaccharide transport to the outer membrane of Escherichia coli. en. *J. Bacteriol.* **189,** 244–253 (Jan. 2007).

231. Krulwich, T. A., Sachs, G. & Padan, E. Molecular aspects of bacterial pH sensing and homeostasis. en. *Nat. Rev. Microbiol.* **9,** 330–343 (May 2011).

232. Jain, S. & Nath, S. Kinetic model of ATP synthase: pH dependence of the rate of ATP synthesis. en. *FEBS Lett.* **476,** 113–117 (July 2000).

233. Fischer, S. & Gräber, P. Comparison of DeltapH- and Delta***φ***-driven ATP synthesis catalyzed by the H(+)-ATPases from Escherichia coli or chloroplasts reconstituted into liposomes. en. *FEBS Lett.* **457,** 327–332 (Sept. 1999).

234. Jin, Q. & Bethke, C. M. Kinetics of electron transfer through the respiratory chain. en. *Biophys. J.* **83,** 1797–1808 (Oct. 2002).

235. Du, B., Olson, C. A., Sastry, A. V., Fang, X., Phaneuf, P. V., Chen, K., Wu, M., Szubin, R., Xu, S., Gao, Y., Hefner, Y., Feist, A. M. & Palsson, B. O. *Adaptive laboratory evolution of Escherichia coli under acid stress* en. June 2019.

236. Samartzidou, H., Mehrazin, M., Xu, Z., Benedik, M. J. & Delcour, A. H. Cadaverine inhibition of porin plays a role in cell survival at acidic pH. en. *J. Bacteriol.* **185,** 13–19 (Jan. 2003).

237. Denoncin, K., Schwalm, J., Vertommen, D., Silhavy, T. J. & Collet, J.-F. Dissecting the Escherichia coli periplasmic chaperone network using differential proteomics. en. *Proteomics* **12,** 1391–1401 (May 2012).

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

239. Georgescu, R. E., Alexov, E. G. & Gunner, M. R. Combining conformational flexibility and continuum electrostatics for calculating pK(a)s in proteins. en. *Biophys. J.* **83,** 1731–1748 (Oct. 2002).

240. Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A. & Caves, L. S. D. Bio3d: an R package for the comparative analysis of protein structures. en. *Bioinformatics* **22,** 2695–2696 (Nov. 2006).

241. Wall, F. T. Principles of Polymer Chemistry. Paul J. Flory. Cornell Univ. Press, Ithaca, New York, 1953. 688 pp. Illus. $8.50. en. *Science* **119,** 555–556 (Apr. 1954).

242. Gromiha, M. M., Thangakani, A. M. & Selvaraj, S. FOLD-RATE: prediction of protein folding rates from amino acid sequence. en. *Nucleic Acids Res.* **34,** W70–4 (July 2006).

243. Newville, M., Stensitzki, T., Allen, D. B., Rawlik, M., Ingargiola, A. & Nelson, A. Lmfit: Non-Linear Least-Square Minimization and Curve-Fitting for Python. en. *Astrophysics Source Code Library* (2016).

244. Felle, H., Porter, J. S., Slayman, C. L. & Kaback, H. R. Quantitative measurements of membrane potential in Escherichia coli. en. *Biochemistry* **19,** 3585–3590 (July 1980).

245. Schultz, B. E. & Chan, S. I. Thermodynamics of electron transfer in Escherichia coli cytochrome bo3. en. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 11643–11648 (Sept. 1998).

246. Ohnishi, T. & Trumpower, B. L. Differential effects of antimycin on ubisemiquinone bound in different environments in isolated succinate . cytochrome c reductase complex. en. *J. Biol. Chem.* **255,** 3278–3284 (Apr. 1980).