UNIVERSITY OF CALIFORNIA, SAN DIEGO

The Systems Biology of Red Cell Metabolism: Physiology under Storage Conditions

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

in

Bioinformatics and Systems Biology

by

James T. Yurkovich

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Jeff Hasty, Co-Chair
Professor Philip Gill
Professor Nathan E. Lewis
Professor Larry Smarr

2018

The dissertation of James T. Yurkovich is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2018

DEDICATION

To:

Mama Bear and Papa Bear

EPIGRAPH

To learn is not to know;

there are the learners and the learned.

Memory makes one,

philosophy the other.

—Alexandre Dumas

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am indebted to many people for the successes I have enjoyed during the course of my doctoral studies—I am where I am and who I am because of the generosity, leadership, and guidance of others. I would like to use this space to acknowledge several individuals for their role in my education to this point.

First and foremost, I would like to thank my doctoral adviser, Bernhard Palsson. He has served as an outstanding mentor and friend to me throughout my studies, providing me with unparalleled resources and opportunities to develop both professionally and personally. The first time I met him, he told me that graduate school is all about finding that thing about which you are passionate, that thing that makes you want to get out of bed every morning and get to work. He has helped me to find and mold my intellectual passions, many of which are detailed in these pages.

Aside from Prof. Palsson, I have been fortunate to study under the guidance several other individuals. I have had a lifetime of outstanding teachers, too numerous to name here (I wonder if this dissertation would have received that elusive 20/20 in Scott Pharion's AP English?). My first real taste of research came in my sophomore year at Notre Dame (back when I was pre-medical student!) with Prof. Sylwia Ptasinska in the area of biophysics. She gladly welcomed me into her lab as an undergraduate researcher, even providing me the opportunity to publish two papers with her. I also had the privilege of studying under Prof. M Vidyasagar FRS for a summer—a truly invaluable experience. Prof. Sagar provided several opportunities for me that summer and has continued to do so since then; I cannot adequately express my gratitude for all the opportunities he has provided me. At UCSD, I would like to thank the members of my doctoral committee: Prof. Hasty, Prof. Gill, Prof. Smarr, and especially Prof. Nate Lewis.

Nate has been an extremely important mentor for me throughout my graduate career, directing one of my research rotations, providing numerous insights on my many research projects, and offering career advice. I have also been fortunate to have collaborations with individuals around the world, especially with Óli Sigurjónsson and Andreas Dräger. Óli hosted me on a wonderful trip to Iceland and has been a great mentor for me. Andreas started in our group here at UCSD and has since moved to Germany, where he has continued to mentor me and work with me on several projects.

I have been fortunate to work very closely with several outstanding scientists during my time in the Systems Biology Research Group (SBRG). In particular, several senior researchers have been outstanding mentors to me: Aarash Bordbar, Dan Zielinski, and Laurence Yang. Aarash was the corresponding author on my first research project in graduate school and provided invaluable guidance throughout the course of that project and well beyond. Dan took me under his wing when I first joined the lab and quickly became a great resource for ... well, almost anything. Laurence is the Mozart to my Rachmaninoff—both in research and in our weekly jam sessions. He is always up for a stimulating intellectual discussion on the way to find caffeine in the afternoon and is an unparalleled wealth of knowledge.

I have also had the privilege of working with some of the best PhD students in the world during my time in SBRG, several of whom have become very good friends. Daily (and sometimes hourly) Tea Time with Jared Broddrick has kept me sane through it all; I will fondly look back on those discussions as some of the best parts of graduate school. Bin and I were always the first ones into lab every morning and—along with Jared—led the lunch charge at 11:00am sharp every day. I would like to thank the SBRG collectively, especially Marc, Helder, and Zak; it has been a tremendous environment for me to spend time learning.

Graduate school is not always easy, but I am blessed to have had very good friends that have been there to celebrate the good times and to offer support during the bad times. P.J., Tim, Archer, Joe, Ian, Suzy, and Liz—I could not have made it through without you guys. Weekly poker nights with Tim, Nick, Nate, Afsheen, Nikos, and Ian helped keep things in perspective over the last few years. Thanks guys.

My brother, B.J., has been one of my biggest inspirations over the years and one of the best resources for which I could ever ask. I even had the opportunity to coauthor an article with him as a result of the two of us spending one entire Thanksgiving holiday huddled around a laptop coding (sorry, Mom). I have been truly blessed to have a brother who is not only accomplished and knowledgeable in so many areas, but who also cares about me and has spent countless hours to help me without any promise of reward for himself.

Last but certainly not least, I would like to thank my parents, to whom I have dedicated this dissertation. Mom, I could not have done this without all your love and support throughout my whole life. Dad, you are the greatest teacher I have ever or will ever have. There will never be anything that I can do to repay both of you for all the amazing opportunities, advice, and love you have provided me.

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

- **JT Yurkovich** and BO Palsson. 2016. "Solving Puzzles with Missing Pieces: The Power of Systems Biology." *Proceedings of the IEEE*, 104(1):2-7. The dissertation author was the primary author.

- **JT Yurkovich**, A Bordbar, ÓE Sigurjónsson, and BO Palsson. 2018. "Systems biology as an emerging paradigm in transfusion medicine." *BMC Systems Biology*, 12:31. The

dissertation author was the primary author.

Chapter 2 in part is a reprint of material published in: **JT Yurkovich**, DC Zielinski, L Yang, G Paglia, O Rolfsson, E Sigurjnsson, JT Broddrick, A Bordbar, K Wichuk, S Brynjolfsson, S Palsson, S Gudmundsson, and BO Palsson. 2017. "Quantitative time-course metabolomics in human red blood cells reveal the temperature dependence of human metabolic networks." *Journal of Biological Chemistry*, 292(48):19556-19564. The dissertation author was the primary author. Chapter 3 in part is a reprint of material published in:

- **JT Yurkovich***, L Yang*, and BO Palsson. 2017. "Biomarkers are Used to Predict Quantitative Metabolite Concentration Profiles in Human Red Blood Cells." *PLOS Computational Biology*, 13(3):e1005424. The dissertation author was one of the two primary authors.

- **JT Yurkovich**, L Yang, and BO Palsson. 2017. "Utilizing biomarkers to forecast quantitative metabolite concentration profiles in human red blood cells." Proceedings of the IEEE Conference on Control Technology and Applications (CCTA), Kohala Coast, HI (August 27-30, 2017). The dissertation author was the primary author.

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

- **JT Yurkovich***, MA Alcantar*, ZB Haiman, and BO Palsson. "Network-level allosteric effects are elucidated by detailing how ligand-binding alters the catalytic potential." Submitted January 2018 (Under review, *PLOS Computational Biology*). The dissertation author was one of the two primary authors.

- A Bordbar*, **JT Yurkovich***, G Paglia, O Rolfsson, ÓE Sigurjónsson, and BO Palsson. 2017. "Elucidating metabolic physiology." *Scientific Reports*, 7 (46249). The dissertation

author was one of the two primary authors.

- **JT Yurkovich**, A Bordbar, ÓE Sigurjónsson, and BO Palsson. 2018. "Systems biology as an emerging paradigm in transfusion medicine." *BMC Systems Biology*, 12:31. The dissertation author was the primary author.

Chapter 5 in part is a reprint of material published in: L Yang*, **JT Yurkovich***, CJ Lloyd, A Ebrahim, MA Saunders, and BO Palsson. 2016. "Principles of proteome allocation are revealed using proteomic data and genome-scale models." *Scientific Reports*, 6:36734. The dissertation author was one of the two primary authors. Chapter 6 in part is a reprint of material published in:

- **JT Yurkovich** and BO Palsson. 2018. "Quantitative -omics data empowers bottom-up systems biology." (*Current Opinion in Biotechnology*), 51:130-136. The dissertation author was the primary author.

- **JT Yurkovich**, BJ Yurkovich, A Dräger, BO Palsson, and ZA King. 2017. "A Padawan Programmer's Guide to Developing Software Libraries." *Cell Systems*, 5(5):431-437. The dissertation author was the primary author.

- **JT Yurkovich**, A Bordbar, ÓE Sigurjónsson, and BO Palsson. 2018. "Systems biology as an emerging paradigm in transfusion medicine." *BMC Systems Biology*, 12:31. The dissertation author was the primary author.

VITA

2013	Bachelor of Science in Electrical Engineering: Biosystems, University of Notre Dame
2018	Doctor of Philosophy in Bioinformatics and Systems Biology, University of California, San Diego

PUBLICATIONS

S Ptasinska, I Tolbatov, P Bartl, **JT Yurkovich**, B Coffey, DM Chipman, C Leidlmair, H Schöbel, P Scheier, and NJ Mason. 2012. "Electron impact on N_2/CH_4 mixtures in He droplets—probing chemistry in Titan's atmosphere." *RSC Advances*, 2:10492-10495.

I Tolbatov, P Bartl, **JT Yurkovich**, P Scheier, DM Chipman, S Denifl, and S Ptasinska. 2014. "Monocarbon cationic cluster yields from N_2/CH_4 mixtures embedded in He nanodroplets and their calculated binding energies." *The Journal of Chemical Physics*, 140:034316.

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

JT Yurkovich and BO Palsson. 2016. "Solving Puzzles with Missing Pieces: The Power of Systems Biology." *Proceedings of the IEEE*, 104(1):2-7.

D Waltemath, JR Karr, FT Bergmann, V Chelliah, M Hucka, M Krantz, W Liebermeister, P Mendes, CJ Myers, P Pir, B Alaybeyoglu, NK Aranganathan, K Baghalian, AT Bittig, PEP Burke, M Cantarelli, YH Chew, RS Costa, J Cursons, T Czauderna, AP Goldberg, HF Gmez, J Hahn, T Hameri, DFH Gardiol, D Kazakiewicz, I Kiselev, V Knight-Schrijver, C Knpfer, M Knig, D Lee, A Lloret-Villas, N Mandrik, JK Medley, B Moreau, H Naderi-Meshkin, SK Palaniappan, D Priego-Espinosa, M Scharm, M Sharma, K Smallbone, NJ Stanford, JH Song, T Theile, M Tokic, N Tomar, V Tour, J Uhlendorf, TM Varusai, LH Watanabe, F Wendland, M Wolfien, **JT Yurkovich**, Y Zhu, A Zardilis, A Zhukova, and F Schreiber. 2016. "Toward Community Standards and Software for Whole-Cell Modeling." *IEEE Transactions on Biomedical Engineering*, 63(10):2007-2014.

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

JT Yurkovich*, L Yang*, and BO Palsson. 2017. "Biomarkers are Used to Predict Quantitative Metabolite Concentration Profiles in Human Red Blood Cells." *PLOS Computational Biology*, 13(3):e1005424.

A Bordbar*, **JT Yurkovich***, G Paglia, O Rolfsson, ÓE Sigurjónsson, and BO Palsson. 2017. "Elucidating metabolic physiology." *Scientific Reports*, 7 (46249).

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

JT Yurkovich, BJ Yurkovich, A Dräger, BO Palsson, and ZA King. 2017. "A Padawan Programmer's Guide to Developing Software Libraries." *Cell Systems*, 5(5):431-437.

JT Yurkovich, L Yang, and BO Palsson. 2017. "Utilizing biomarkers to forecast quantitative metabolite concentration profiles in human red blood cells." Proceedings of the IEEE Conference on Control Technology and Applications (CCTA), Kohala Coast, HI (August 27-30, 2017).

JT Yurkovich, DC Zielinski, L Yang, G Paglia, O Rolfsson, ÓE Sigurjónsson, JT Broddrick, A Bordbar, K Wichuk, S Brynjolfsson, S Palsson, S Gudmundsson, and BO Palsson. 2017. "Quantitative time-course metabolomics in human red blood cells reveal the temperature dependence of human metabolic networks." *Journal of Biological Chemistry*, 292(48):19556-19564.

L Yang, **JT Yurkovich**, ZA King, and BO Palsson. 2018. "Modeling the multi-scale mechanisms of macromolecular resource allocation." *Current Opinion in Microbiology*, 45:8-15.

JT Yurkovich and BO Palsson. 2018. "Quantitative -omics data empowers bottom-up systems biology." *Current Opinion in Biotechnology*, 51:130-136.

E Kavvas, Y Seif, **JT Yurkovich**, C Norsigian, S Poudel, WW Greenwald, S Ghatak, BO Palsson, and J Monk. 2018. "Updated and standardized genome-scale reconstruction of Mycobacterium tuberculosis H37Rv, iEK1011, simulates flux states indicative of physiological conditions." *BMC Systems Biology*, 12:25.

JT Yurkovich, A Bordbar, ÓE Sigurjónsson, and BO Palsson. 2018. "Systems biology as an emerging paradigm in transfusion medicine." *BMC Systems Biology*, 12:31.

* equal contribution

ABSTRACT OF THE DISSERTATION

The Systems Biology of Red Cell Metabolism:
Physiology under Storage Conditions

by

James T. Yurkovich

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2018

Professor Bernhard Ø. Palsson, Chair
Professor Jeff Hasty, Co-Chair

The human red blood cell (RBC) is a logical starting point for the development and application of systems biology methods because of its simplicity, intrinsic experimental accessibility, and importance in human health. New "-omics" technologies have been used to study the biochemical and morphological changes that occur in red blood cells during cold storage, collectively referred to as the "storage lesion." Here, we extend these previous efforts by using systems biology to examine the metabolic physiology of RBCs under storage conditions. We first characterized the temperature dependence of the storage process using previously identified

storage-age biomarkers as a representation of systems-level trends, showing that the metabolic state of the RBC is conserved but accelerated with increasing temperature. We then questioned whether these biomarkers—which had been shown to be excellent qualitative markers of systemic behavior—held any potential to provide quantitative information about the system. Using simple linear statistical models, we showed that a subset of the biomarkers could be used to predict the quantitative concentration profiles of other metabolites in the RBC network. We expanded these efforts by integrating network structural information into these statistical models to forecast future values of these concentration profiles after measurements made during only the first eight days of storage. Next, we used multiple first principles modeling approaches to understand the underlying mechanisms and temporal dynamics of the observed behaviors and developed a method for the integration of metabolomics data into cell-scale mathematical models. Finally, we developed a method for the integration of quantitative proteomics data into cell-scale models using *Escherichia coli* as a test case. Collectively, these results provide empirical proof that the RBC metabolome can be represented in a low-dimensional space and offer the starting point for a whole-cell model of the RBC. More broadly, we detail the development and use of systems biology methods on the human RBC, providing a starting point from which we can expand these efforts to other, more complicated cellular systems.

Chapter 1

A Systems View of Biology

Life is a program written in DNA. Starting in 1995, genome sequences detailing this program have ushered in a new point of view in biology: a true systems-level, or "genome-scale," perspective. The genome sequence for an organism is analogous to having a component list for a circuit, except that many connections between components, and even some of the component functions themselves, are unknown. So how do we solve a puzzle when there are pieces missing? Enter systems biology. The combination of full genome sequences with over half a century of research in genetics, molecular biology, and biochemistry has enabled the genome-scale reconstruction of networks underlying well-studied cellular functions, such as metabolism. A quality-controlled reconstruction process effectively produces a circuit diagram of the metabolic network encoded in an organisms genome that can be modeled mathematically. Thus, a first principles "bottom-up" approach to systems biology rooted in fundamental mechanisms has arisen, and the quest to reveal the program that DNA encodes is underway. This article will familiarize you with some of the engineering concepts, methods, and applications in systems biology.

1.1 Systems biology as a paradigm

The scientist and philosopher Thomas Kuhn proposed that scientific advances occur through periodic shifts in prevailing paradigms [1]. Prevailing paradigms are explored through periods of "normal science," extremely productive phases in which scientists solve puzzles that arise within such paradigms. When shortcomings of the prevailing paradigm are identified, the field shifts from one set of paradigms to another, restarting the process anew. There are different drivers that lead to such paradigm shifts. The development of new technologies represents a driver for change, as demonstrated by the "-omics" revolution which we are still witnessing in the life sciences. Another driver of change is the application of methodologies and approaches from different disciplines. Bottom-up systems biology is emerging as a way to integrate disparate -omics data types based on first principles, providing detailed mechanistic descriptions as a basis for -omics data analysis.

The bottom-up approach to systems biology aligns with engineering thinking embodied in systems science. Systems biology aims to understand how all the molecules that make up a cell interact to form coherent physiological functions. Metabolic networks are made up of thousands of biochemical reactions that can now be "reconstructed" and converted into mathematical formats amenable to modeling. Because these models are built from first principles, they are able to describe the functional states of networks and therefore the systems-level behavior of the cell. Bottom-up systems biology is helping to unravel and understand the "genotype-phenotype relationship" on a genome-scale basis. The "genotype" of an organism (the collection of all genetic elements on a genome) contains the information that determines its form and function (the "phenotype"). Defining, understanding, and using this relationship is fundamental to systems biology.

The genotype-phenotype relationship is multi-scale (Fig. 1.1). At the smallest scale, molecular biology and biochemistry give us an understanding of DNA and how information in the form of genes and other genetic elements is encoded in it. These genetic elements have various structural and regulatory functions and encode the proteins that catalyze and facilitate biochemical reactions. At larger scales, these reactions form ever more complicated modules of biochemical functions in a network setting that together manifest an overall cellular behavior, or phenotype. Phenotypic states can thus be viewed as the result of running the "program" that is encoded in the DNA.

Taking a few liberties, we construct a simple analogy to relate the genotype-phenotype relationship to a familiar electrical engineering concept (Fig. 1.1). In an electrical circuit, we can see a "genotype-phenotype" relationship emerge as we go from an atomic level to a component level to an engineering application. At the lowest scale of system complexity, we have the atoms that make up P-N junctions described by semiconductor physics. Using this information allows for the construction of both active and passive circuit elements. Together, these circuit elements can be arranged in a network from which the "phenotype" or systems function emerges: capturing and converting a signal to music through an amplifier and a speaker. As in an organism, the form and function of the radio is defined by the properties of its "genotype."

A fundamental paradigm for the implementation of systems biology on the genome-scale has arisen [2], driven by the recent ability to generate data describing the many levels of biological complexity. First, all components within a cell (proteins, biochemical reactions, etc.) are enumerated and annotated. These components are then connected and used to reconstruct the network map. These reconstructed networks are translated into mathematical formats that describe the underlying biological knowledge. Finally, testable predictions are made using the

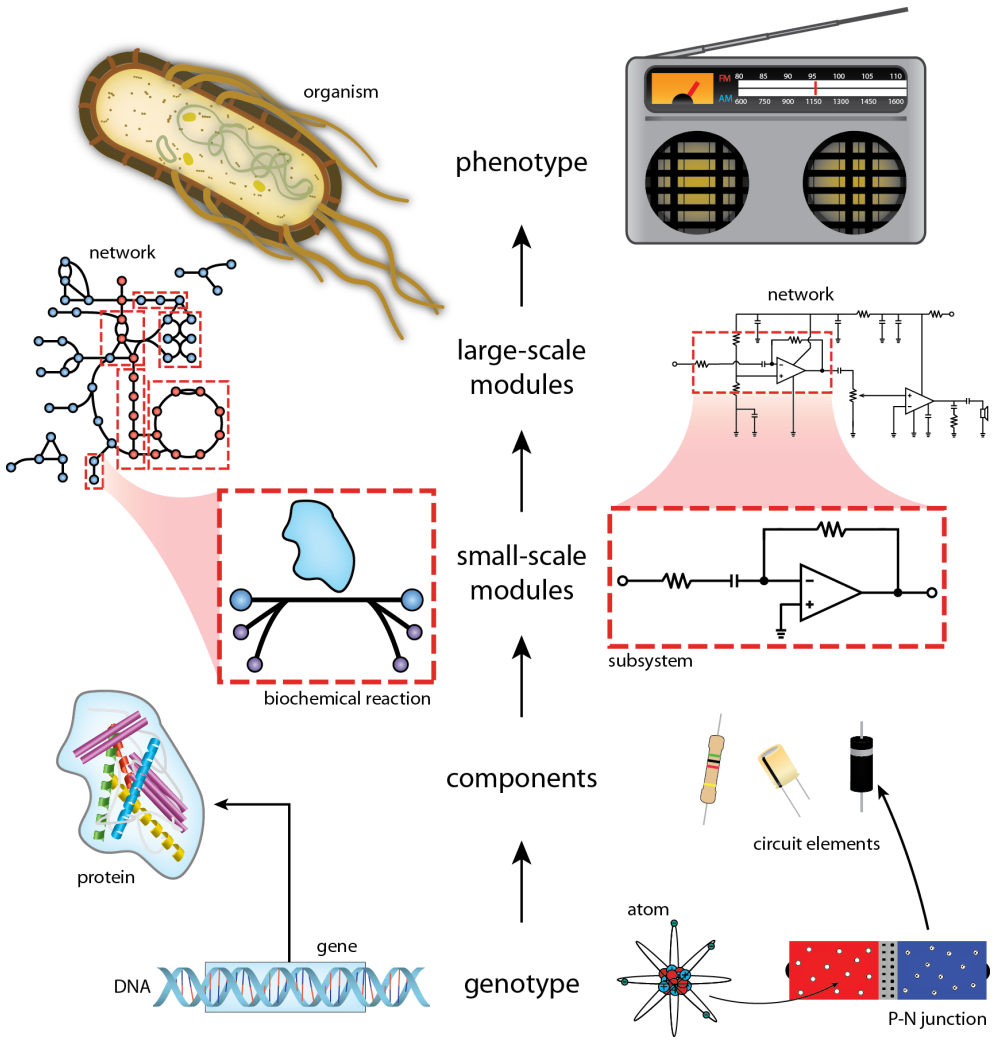

Figure 1.1: Biology is multi-scale. The genotype-phenotype relationship in biology (left) and an analogous view in an electrical system (right) allows for the modularization of a system over different complexity scales.

mathematical models that describe the network. This process must be repeated for each new organism of interest. Well-studied model organisms such as *Escherichia coli* (*E. coli*) or the human red blood cell (RBC) can serve as a Rosetta Stone for inferring the genetic content and function of poorly characterized organisms.

Modeling Methods and Techniques

Once a network is reconstructed, it is translated into a mathematical format using fundamental physical and chemical principles. As mentioned before, these reconstructed networks are incomplete. Finding the missing pieces and accounting for incomplete network structure requires the use of mathematics familiar to engineers to help construct a network that still carries biological meaning. Here, we describe some of the standard modeling tools and techniques, highlighting some of the differences that arise between biological systems and other engineering systems.

Reconstructing a network The reconstruction process is a system identification problem aimed at reverse engineering and inferring biochemical network structure from first principles. Biologists and chemists study organisms on a molecular level to identify individual connections between molecules (reactions) and characterize the inputs and outputs. This data is stored in large repositories and is not always organism specific. Therefore, it must be manually curated to identify which reactions occur in a given network (i.e., some reactions may not be capable of occurring in a given organism and should not be included). This curation becomes a time-intensive process as even relatively simple bacteria such as *E. coli* have thousands of reactions.

Connections between components are modular and can be linked to form the larger network (Fig. 1.1). The reconstruction process has been reduced to a standard operating procedure that can be followed to derive the network structure for new organisms [3]. The resulting networks are inherently incomplete because we simply do not have knowledge of all of the compounds and connections. Standard system identification techniques are used to expand models and infer missing content (referred to as "gap filling").

The major limiting factor for a reconstruction is the time it takes to complete—a recon-

struction of the human metabolic network took a team of six people over two years! Further, the human aspect of the curation process leads to concerns regarding consistency and quality control between reconstructions. Thus, there is a real need for tools that could automate the reverse engineering of biochemical networks to yield reconstructions; such tools would allow for the elucidation of the network structure of new organisms of interest without devoting hundreds of hours to the process.

Translating into a mathematical format Once the biochemical network has been reconstructed, it must be translated into a mathematical format that is amenable to modeling. As is common in simplified modeling of many engineering systems, the entire network can be captured in a matrix that represents the inputs and outputs (in this case, the stoichiometry of all reactions in the network). This stoichiometric matrix, $\mathbf{S}$, is an incidence matrix with rows representing nodes and columns representing links (Fig. 1.2). Because a given compound only participates in a handful of reactions, $\mathbf{S}$ is sparse. Further, almost all of the nonzero entries are either $+1$ or -1 (outputs are positive and inputs are negative by convention). The simple mathematical structure of $\mathbf{S}$ allows for manageable computation and compression of large networks. The formulation of a biochemical network as a connectivity matrix represents a huge leap forward because it enables the use of familiar systems engineering tools like loop analysis (Fig. 1.2).

Dynamic description of biochemical reactions An important feature of $\mathbf{S}$ is the bilinearity of the reactions it represents. Two chemical components can react to produce a third, leading to more than two nonzero entries in the corresponding column of $\mathbf{S}$. The content of all the nodes in a system like metabolism (i.e., the concentrations of the corresponding compounds) can be represented by a system of ordinary differential equations (ODEs). These systems of ODEs can

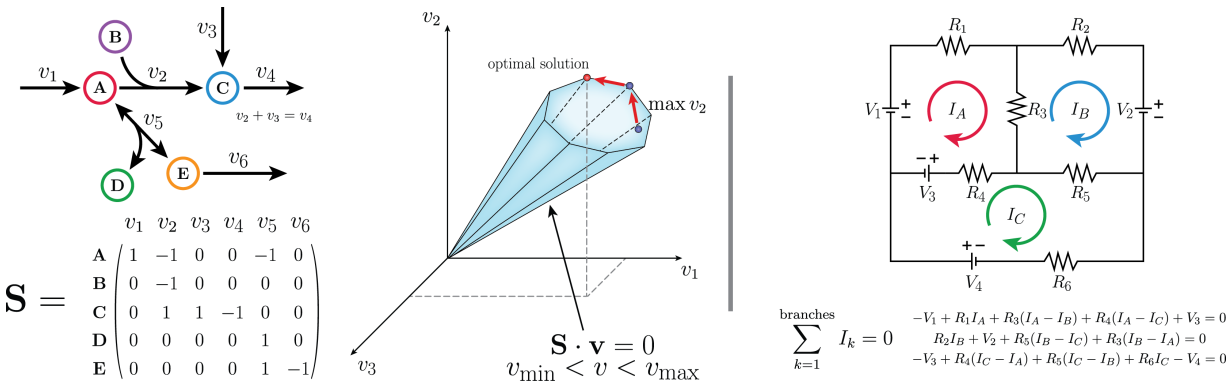

Figure 1.2: Loop and network analysis. Loop and network analysis in systems biology (left) and electrical circuits (right). Some of the most powerful tools for analyzing systems-level models follow directly from Kirchhoff's laws.

be numerically solved to provide an idea of the systems state at a given time [4]. Having a model that is able to predict the concentrations of metabolic nodes is powerful because the nodes in metabolism represent some of the primary targets for therapeutic drugs [5]. In a biological network, it is possible for individual nodes to accumulate or lose mass; that is, the flow of mass into the node may not match that of the outflow. The rates of intracellular node accumulation can impact steady-state models because of the timescales on which some reactions occur. This characteristic is often neglected in simplified modeling techniques, such as loop analysis.

Modeling at the genome-scale In systems biology, it is generally impractical—if not impossible—to build true genome-scale ODE models. This impracticality is due in part to the lack of parameterization for individual reactions (initial conditions and rate constants are not known for every reaction). However, like in many engineering applications, modeling the system at steady-state is often a good proxy for the most interesting biological state. Thus, other methods for genome-scale network analysis have been developed, benefiting from the application of well-developed systems tools such as hidden Markov processes [6] and convex optimization [7]. Some of these other modeling approaches do not require the extensive parameterization of ODE

models and are therefore more amenable to modeling large, incompletely characterized systems.

One of the more successful methods for modeling genome-scale metabolic networks is to use constraint-based modeling to represent the flow of mass (the "flux") through every reaction in the network. Termed "flux balance analysis," this modeling strategy comes directly from the network structure and therefore bypasses the need for extensive parameterization [8]. With the network described in the form of a matrix, a simple matrix equation is used to model the system at steady-state:

$$\mathbf{S} \cdot \mathbf{v} = 0 \tag{1.1}$$

where $\mathbf{S}$ is the stoichiometric matrix and $\mathbf{v}$ is a vector that represents the flux of each reaction in the network. Solving this simple matrix equation results in a solution space where each point in the space is a possible flux state of the steady-state system. Thus, a family of candidate solutions—rather than a single solution—is obtained. Constraints representing the properties of the biological machinery involved in each reaction are then imposed, resulting in a constrained solution space (Fig. 1.2). The balanced network obeys Kirchhoffs laws: the flux around a metabolic loop must add to zero and the sum of the flux into a node must equal the sum of the fluxes leaving that node.

Computing the Genotype-Phenotype Relationship

Constraint-based models make use of COnstraint-Based Reconstruction and Analysis (COBRA) methods [9] to simulate, analyze, and predict phenotypes. The variety of biological constraints applied to biochemical network reconstructions has grown from simply placing bounds on individual reaction fluxes to now include compartmentalization of molecules, mass conservation, and thermodynamic directionality of reactions. These additions have vastly in-

creased the scope of biological questions that can be addressed using COBRA methods [10]. Applications often focus on perturbing individual components through gene deletion or addition and understanding how the effects propagate throughout the system. Such studies lead to novel predictions about the genotype-phenotype relationship that now can be tested experimentally on a large scale given the advances in genome editing [11].

The solution space of these constraint-based models is typically convex and can be characterized in several ways. One of the key capabilities of constraint-based models is the ability to construct an optimization problem to find the minimum or maximum flux for a reaction of interest (Fig. 1.2). Systematically optimizing each reaction in the model to find the minimum and maximum feasible fluxes can therefore be used to characterize the solution space. This capability, for instance, allows for direct molecular engineering applications, such as coupling the production of valuable biomolecules with vital growth pathways of the organism. The suite of COBRA methods has led to a number of applications [5], the development of computational toolboxes [9, 12], and a series of scientific meetings focused on method development and applications.

The first-generation constraint-based models of metabolism incorporate what knowledge we have of how specialized proteins ("enzymes") facilitate and catalyze biochemical reactions. When constraints are placed on the flux through reactions in the network, these models provide useful and accurate phenotypic predictions [10]. The ability to optimize a model for a specific phenotype is useful for representing organisms possessing a clear biological objective; in many bacterial species, that objective is to maximize growth (the faster an organism can grow, the more successful it will be). To compute the growth rate, an objective function is defined and added to the model in the form of a reaction; that reaction can then be maximized as a function of important system inputs (Fig. 1.3). These predictions can then be experimentally tested and

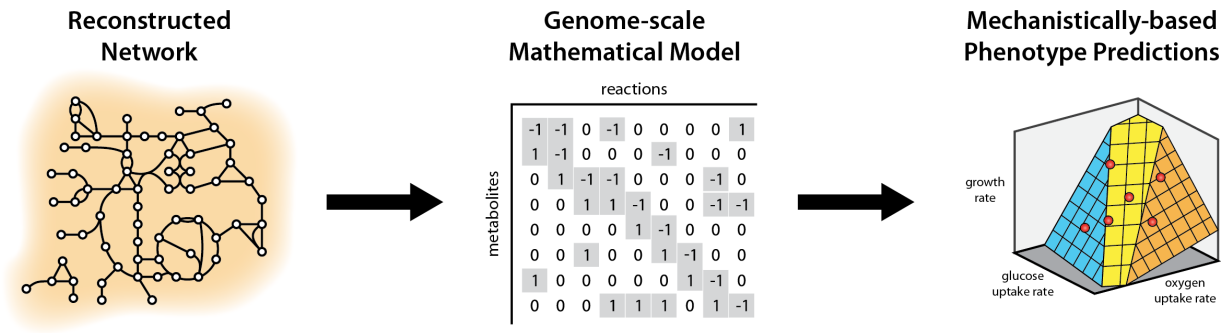

Figure 1.3: Computing phenotypic states. Genome-scale mathematical models allow for explicit computational modeling of the genotype-phenotype relationship. Optimizing ME-Models for an objective such as maximum growth rate provides mechanistic knowledge behind optimal functional states of both metabolic fluxes and expression levels as a function of important inputs such as glucose or oxygen [14].

have been successfully used in the design process of industrial production organisms [13].

While it is powerful to have the ability to find optimal states for industrial engineering applications (e.g., production of valuable biomolecules), defining an accurate objective function is very challenging. Life does not necessarily operate at an optimal growth state; instead, we often see that organisms operate at near optimal growth states because they are trying to optimize for other functions (such as readiness for unforeseen environmental stresses). Thus, other engineering tools have been integrated into the suite of COBRA methods to expand the analytical capabilities so that additional states can be calculated. One such tool uses a Markov chain Monte Carlo method to sample the solution space, computing possible flux states of the network. There is currently a lot of interest in computing these non-optimal growth states, either through the development of new COBRA methods or by adding additional constraints.

Recently, constraint-based models have been extended to account for the mechanistic detail of gene expression and protein synthesis (the processes of "transcription" and "translation," respectively). In other words, models can now account for more than just the network structure itself—they can compute both the cost of synthesizing all the machinery required for a partic-

ular state (the protein and enzyme demand) and the cost of running that particular state (how much flux is required for each reaction in the network). These next-generation "ME-Models" (Metabolism and Expression) allow genome-scale reconciliation of molecular biology and biochemistry by explicitly accounting for the biological machinery responsible for gene expression and protein synthesis. The increasingly comprehensive ME-Models are able to provide better predictions for biological objectives, such as growth (Fig. 1.3). As data for new organisms is generated and analyzed, ME-Models can be constructed for new organisms and systems, like human metabolism. Much exciting work lies ahead to automate the generation of ME-Models, to include new data types and biological knowledge, and to use the models to solve fundamental and applied problems in the life sciences.

1.2 RBC storage for transfusion medicine

RBCs account for over 84% of the native cells in the human body by number, making them the most numerous cell type by a large margin [15]. The transfusion of RBCs has long been an integral part of modern healthcare [16, 17] over 112 million RBC units collected for blood transfusions worldwide annually [18]. The storage of RBCs in non-physiological conditions (i.e., packed in plastic blood bags in a static environment at 4°) leads to many changes in the biochemical and physiological properties of RBCs. Over the past several decades, the transfusion medicine has made great progress in defining a central paradigm that outlines the biochemical and morphological changes—the so-called RBC "storage lesion" (RSL)—that red cells undergo during cold storage [19–22]. Such changes include a decrease in 2,3-diphosphoglycerate (2,3-DPG) levels, a decrease in nitric oxide (NO) metabolism, an increase in endothelial adherence, and morphological modifications to the shape and structure of the cells. Some of these changes are

reversible upon transfusion (e.g., 2,3-DPG levels), while morphological changes and alterations in NO metabolism can be irreversible.

Transfusion medicine is now in the early stages of a paradigm shift that embraces the benefits derived from the application of systems biology approaches [23]. In recent years, the use of -omics technologies has been deployed to gain a better understanding of RSL [24]. In particular, metabolomics data have become a central part of the effort to better understand RSL [25]. Profiling the metabolic state of the cell is an important approach that allows a functional interpretation of cellular biochemistry [26]. With the availability of such data, systems biology methods can be applied to study and understand RSL in considerable detail. While correlations are important for the practice of medicine, an actionable and mechanistic understanding of relevant physiological phenomena is desired [27, 28]. Systems approaches have already proven valuable through the evaluation of drug therapies [29, 30], identification of biomarkers for cancer [31], and the prediction of oncogenes in cancer conditions [32]. Here, we discuss how the study of RSL is being added to this list.

Three key ingredients for systems biology

The systems biology approach is an inherently iterative process of refinement that unites three key ingredients: data collection, analysis, and computational modeling (Fig. 1.4). The first ingredient is data collection. Working in conjunction with blood banks to ensure that standard quality controls are met is vital for generating high-quality data. Absolutely quantified metabolomics data—while more costly—can yield greater benefits since it can be integrated with quantitative, mechanistic models. The data sets described here include exo-metabolomic, endo-metabolomic, and other hematological measurements routinely performed in blood banks

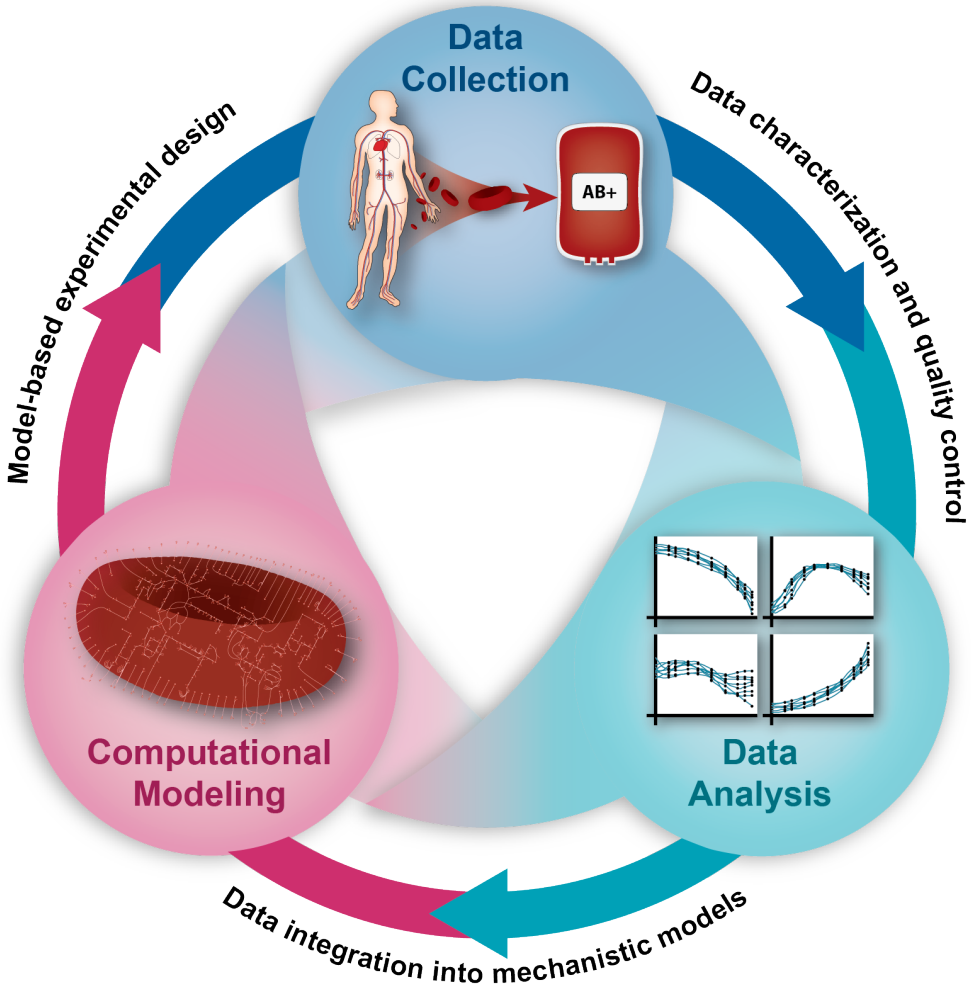

Figure 1.4: Three key ingredients for systems biology. Three key ingredients come together to form a workflow capable of extracting knowledge from -omics data.

(e.g., pH, pO_2). These measurements were made as part of time courses, resulting in these measurements at each time point.

Time-resolved metabolomics data have yielded important insights into metabolic physiology provided the time grid overlaps with the time scale of key metabolic changes. This time scale is faster than a week, which is the commonly used time increment in sampling stored RBC bags. In the experiments discussed below, data is collected every three to four days, or 14 times over the 42 day storage period. Such data sets represent a significant departure from historical norms

in this field. A continued exploration of various perturbations to the standard storage conditions will help further elucidate RBC metabolic biochemistry. Such perturbation experiments might be informed by previous experiments or by computational models [27].

The second ingredient is the application of multivariate data analysis to the large data sets generated. Multivariate statistical analyses can reveal the overall structure of the data sets and subtle trends within the data. In particular, methods like principal component analysis (PCA) [33], partial least squares discriminant analysis (PLS-DA) [34], and independent component analysis (ICA) [35] have been used effectively to analyze complex metabolomics data sets. Care must be taken when choosing a methodstatistical methods have specific applications and cannot be blindly applied to raw data; fortunately, there are several excellent resources that provide guidance for this process [33, 36]. Although the data sets generated and analyzed here are large compared to the history of the field, they do not qualify as "Big Data." In the future, genetic information and other parameters may enrich this data, as has been demonstrated by the generation of personalized RBC models [37].

The third ingredient is a computational, mechanistic metabolic network model capable of integrating disparate data types. Such models incorporate the results of statistical analyses to generate biological insights and testable hypotheses. A metabolic network specific to the RBC has been generated by mapping multiple proteomic data sets onto the reconstruction of the global human metabolic network [38, 39]. This mapping has resulted in a functional metabolic network of the RBC containing 283 metabolic reactions [40]. This mapping contains contiguous known pathways and revealed the presence of previously unidentified pathways. The specifics of this network have been further delineated through a comprehensive manual curation of the literature (the "bibliome"). This reconstructed metabolic network inherently includes available information

about the genome, metabolome, proteome, and bibliome [3], truly representing multi-omic data integration.

Several years ago, we proposed the use of systems biology in the transfusion medicine field to extend the lifetime of stored RBC units [23]. Since then, we have used the principles outlined above to study RBC storage from a systems perspective. Here, we review the outcome of several of these efforts and try to contextualize the studies of other groups that have pursued similar goals.

The three-phase metabolic decay in stored RBCs

Our first goal was to characterize and understand the baseline RBC metabolic behavior during storage [41]. We thus collected RBC units from 20 individuals and stored them in SAGM media. We absolutely quantified 142 metabolites and hematological variables (e.g., hematocrit, pH) at 14 time points over 45 days of storage. Such fine resolution on the time series (measurements taken every three to four days) allowed for the construction of an intensive grid of data points that was able to capture previously unobserved behaviors. Further, the quantitative nature of the measurements allowed for a better characterization of previously determined qualitative concentration shifts, such as the increase in hypoxanthine to higher-than-physiological levels.

In order to generate an initial global characterization of the data, principal component analysis (PCA) was performed on the raw metabolomics data. PCA is a multivariate statistical method that is commonly employed on metabolomics data for dimensionality reduction; the principal components identified through this analysis represent the relative contribution of each measurement to the variability observed in the data. PCA on the metabolomics data revealed

three distinct metabolic shifts that occur during the 42 day life of a stored RBC unit. These shifts in metabolic state, occurring at days 10 and 17, showed that RBCs do not undergo a simple linear decay process. Danish transfusion records were consulted to determine whether there was any correlation between seven-day mortality and the age of the transfused RBC unit. These shifts were shown to be potentially clinically relevant, with blood stored past day 10 representing the most significant association. This same three-phase metabolic decay profile has been validated by other groups and in different storage media [42, 43].

We observed the previously reported high concentration of 2,3-DPG that depletes over time, as well as the initial increase and subsequent decrease in ATP levels after the first shift at day 10. We observed that the "metabolic inflection points" (i.e., the points in time at which the metabolic shifts in the PCA plots occur) coincide with the depletion of extracellular adenine and accumulation of hypoxanthine and xanthine in the storage medium. One notable observation from the quantitative metabolomics data was the existence of a large intracellular malate pool (greater than 1 mM).

Perturbing the storage conditions

Having characterized the baseline metabolic behavior of RBCs under cold storage, the next step was to determine whether we can perturb the storage conditions to affect the metabolic decay process. We identified four perturbations that posed interesting questions (Fig. 1.5): (1) does the three-phase decay pattern manifest itself only in SAGM media, or is it present in other storage media types used in transfusion medicine?; (2) does supplementing the bag with additional carbon sources support ATP levels?; (3) is the depletion of adenine the cause of the metabolic shifts?; and (4) is there a subset of measurements that is representative of the metabolic

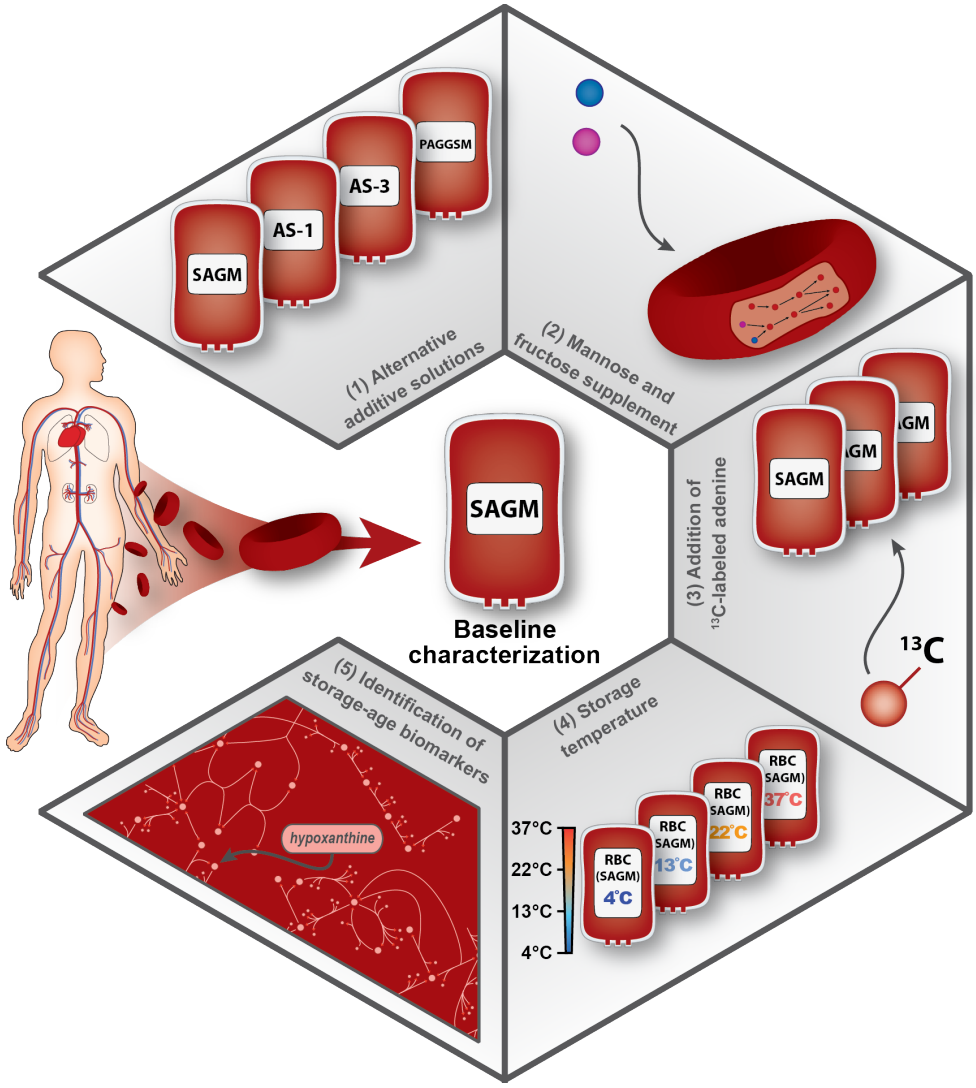

Figure 1.5: Perturbations to the storage conditions. Baseline characterization and perturbation experiments on stored RBCs. Perturbation experiments examined the effect of (I) alternative media formulations, (II) supplementation with additional sugars, and (III) addition of adenine to the media. (IV) The final experiments yielded a set of storage-age biomarkers.

state of the RBC that could be biomarkers?

Does the storage media affect cellular metabolism?

With the baseline behavior in SAGM media now well characterized, we set out to determine whether RBCs stored in other additive solutions exhibited similar metabolic behavior [44].

RBC units from 12 individuals were stored in SAGM [45,46], AS-1 [47], AS-3 [47,48], and PAGGSM [49] for 45 days; samples were collected and metabolically profiled at 14 time points during storage. These media types were chosen because they represent the most widely-used additive solutions in Europe (SAGM, PAGGSM) and the United States (AS-1, AS-3) [50].

Several changes in basic metabolic behavior was observed across the four additive solutions. Notably, citrate uptake and metabolism was increased in AS-3 and PAGGSM compared to that of SAGM and AS-1. Corresponding changes in intracellular malate concentrations suggest that citrate uptake impacts malate utilization. Labeled citrate added to the bag prior to storage in SAGM showed that citrate is taken up and converted to intracellular malate, contributing to the large pool previously observed in the baseline characterization. This behavior has been shown to occur in other media [51] and was computationally predicted and validated in the baseline data [52]. Statistical analyses indicated that the difference in citrate uptake and metabolism impacted glycolytic function. In particular, fructose-6-phosphate and glucose-6-phosphate were identified as locations within the network on which the metabolic alterations were focused.

Do other sugars better support ATP levels?

The baseline data showed that fructose and mannose, found in the plasma collected with the RBCs from the donor, are rapidly metabolized and depleted during the first metabolic phase. Mannose and fructose have been shown to be metabolized through different pathways than glucose in RBCs [53,54], thus providing potential benefit over glucose as the primary energy source for metabolism. Further, fructose—while known to have adverse effects on human physiology [55]—has also been shown to act as a protectant against oxidative damage [56]. Following in the footsteps of work by Beutler and Duron [57] and by Dawson and colleagues [53,58],

we supplemented RBC units with mannose and fructose to better characterize alternate sugar metabolism during storage [59]. These units were metabolically profiled at 14 time points over 25 days over storage in SAGM media.

These experiments showed that the metabolism of mannose and fructose at 4° is reflects their metabolism at 37°. The timing of the metabolic inflection points was altered slightly with the supplemented sugars, with the observed changes primarily centered in glycolysis. The hypothesized protective effect of fructose was not observed. The additives failed to maintain ATP and 2,3-DPG levels under the tested experimental conditions, although this was likely due to the presence of glucose; a better characterization of the metabolism of these sugars could be obtained by replacing glucose with mannose and/or fructose (instead of supplementing) and examining the resulting metabolomics measurements.

While fructose is known to be taken up through the GLUT5 transporter [60], the exact mechanism for mannose incorporation has yet to be clearly elucidated. However, it is believed that mannose is transported into the cell via the GLUT1 transporter [54]. GLUT1 is also used by glucose, leading to competition for uptake of the two compounds. The ^{13}C labeling results from this perturbation study support the hypothesis that mannose is taken up by GLUT1. More importantly, these results imply that mannose is preferentially taken up and oxidized over glucose.

Is the depletion of adenine the cause of the metabolic shift?

Following the identification of the three-phase metabolic decay observed in SAGM [41], it was observed that the depletion of adenine coincided with the metabolic inflection points observed in the PCA plots. We therefore hypothesized that these metabolic shifts were due in part to the depletion of adenine. To test this hypothesis, we labeled adenine in both normal and

double concentrations in SAGM media [61]. We took metabolomics measurements at 10 time points over 31 days of storage.

We observed that the RBCs consumed approximately 1.5 mg/L adenine per day over the first eight days of storage, almost depleting the total adenine concentration in the bag toward the end of the first metabolic phase. During this first phase, adenine was converted into inosine and IMP but not ATP. By day 18 (the end of the second phase), the extracellular adenine was completely depleted.

Having characterized this behavior more completely, we doubled the initial concentration of adenine. Surprisingly, we observed the identical consumption rate of adenine until day 18, at which point adenine was no longer taken up by the RBCs. In other words, it appears that the perfect amount of adenine is added in SAGM media; adding any more would result in adenine sitting in the extracellular media without being taken up by the cells. One possible explanation for this intriguing result is that this behavior was previously characterized during the development of SAGM but never published (and has now been re-discovered years later). One notable observation was that the higher levels of adenine resulted in a buildup of 5-methylthioadenosine. The conclusion of this study was that adenine is not responsible for the observed metabolic shifts, but there is another internal process that leads to termination of adenine uptake.

Is there a subset of measurements that can be used to define RBC metabolic health?

One obstacle to the routine use of metabolomics data is the cost of generating it. With the increasing amount of metabolomics data already available for RBCs under storage conditions [25] and relative invariance of the metabolome composition during decay, it is logical to ask if we can identify biomarkers that describe the decay process through simple measurements. Have we

reached a point where there is a critical mass of data available for true systems analysis leading to the identification of robust biomarkers?

Thus, we set out to identify a set of metabolites that could define the trends that had been observed in the studies discussed above. We were searching for a small number of extracellular metabolites because of the ease, cost, and reliability of such measurements. Through detailed statistical analysis of existing data sets, we identified eight extracellular metabolites (adenine, hypoxanthine, glucose, lactate, malate, nicotinamide, 5-oxoproline, and xanthine) that can differentiate between the three metabolic states elucidated through PCA [62]. These "storage-age" biomarkers robustly represent the RBC metabolome throughout the storage process. Initially identified in SAGM media, these storage-age biomarkers were validated in AS-3 and independently verified in a separate laboratory with a different analytical setup and different sample sets [62].

Glucose, lactate, 5-oxoproline, and adenine represent the primary metabolic inputs and outputs can effectively serve as "clocks" for storage time. Further, the large malate pool is related to a major component in the buffers used during processing: citrate. The potentially more interesting biomarkers are nicotinamide, hypoxanthine, and xanthine that are directly indicative of the metabolic state. Nicotinamide is one of the components of major cofactors (NAD+/NADH and NADP+/NADPH) and is released from RBCs after approximately ten days of storage. The toxic effects of hypoxanthine and xanthine are well known [63].

1.3 Outline of the dissertation

With so much data available on the various storage media perturbations described above and a set of robust storage-age biomarkers identified, we set out to find a perturbation that

could provide meaningful information regarding the RBC metabolic network itself. As mentioned previously, the artificial storage environment is one of the primary reasons we see the effects of RSL. In particular, it is a shock for cells to be kept 33°C below the *in vivo* temperature. Thus, we decided to investigate how temperature affects the metabolic network by using metabolomics measurements to study RBCs stored at four different temperatures.

Having observed several of these behaviors, we asked whether we could predict these behaviors. We applied linear statistical modeling approaches to metabolomics data. These efforts focused on extending the utility of the storage-age biomarkers by showing their capacity to act as quantitative biomarkers of systemic behaviors. To make these models more directly applicable to blood banking, we reformulated these models to forecast future values of other metabolites in the system using only measurements taken up until day eight of storage.

The next objective was to attempt to understand some of these behaviors using mechanistic models. We used ordinary differential equations to model the kinetics and temporal dynamics of glycolytic kinases. Enzymatic rate laws have historically been used to simulate the dynamics of complex metabolic networks. Regulated reactions are typically by allosteric rate laws. Here, we use detailed elementary reaction descriptions of regulatory enzymes allowing for the explicit computation of the fraction of the enzyme molecules that are in a catalytically active state. The fraction of the enzyme that is in the active state represents the time dependent utilization of its catalytic potential, and thus reflects the fundamental result of enzyme regulation.

Having described explicit mechanistic phenomena on the scale of a single pathway, we wanted to scale these efforts up to include the entire metabolic network. Thus, we used constraint-based modeling to describe the full RBC metabolic network. In order to better describe the dynamics that are observed during storage, we devised a novel computational method that al-

lows for the integration of exo- and endo-metabolomics data into constraint-based models. The resulting models are able to more accurately describe metabolic physiology.

To this point, our modeling efforts have accounted for the structure of the metabolic network using genomics, specific information on individual aspects of the network using the bibliome, and snapshots of the metabolic state using metabolomics. The final part of this thesis integrates information regarding the proteome through the integration of quantitative proteomics data into constraint-based models. Due to the availability of high-quality data and a genome-scale model, we developed a method for the integration of quantitative proteomics data into an existing model of *Escherichia coli*. This method provides the starting point for the development of a whole-cell model of the RBC that will include information regarding the metabolome and the proteome.

Acknowledgements

This research was supported by the European Research Council (ERC 232816), the National Heart, Lung, and Blood Institute (NHLBI R43HL123074), and the Landspitali University Hospital Research Fund.

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

- **JT Yurkovich** and BO Palsson. 2016. "Solving Puzzles with Missing Pieces: The Power of Systems Biology." *Proceedings of the IEEE*, 104(1):2-7. The dissertation author was the primary author.

- **JT Yurkovich**, A Bordbar, ÓE Sigurjónsson, and BO Palsson. 2018. "Systems biology as an emerging paradigm in transfusion medicine." *BMC Systems Biology*, 12:31. The

dissertation author was the primary author.

Chapter 2

A Systems Analysis of Perturbed RBC Storage Conditions

The temperature dependence of biological processes has been studied at the levels of individual biochemical reactions and organism physiology (e.g., basal metabolic rates) but has not been examined at the metabolic network level. Here, we used a systems biology approach to characterize the temperature dependency of the human RBC metabolic network between 4°C and 37°C through absolutely quantified exo- and endo-metabolomics data. We used an Arrhenius-type model (Q_{10}) to describe how the rate of a biochemical process changes with every 10°C change in temperature. Multivariate statistical analysis of the metabolomics data revealed that the same metabolic network-level trends previously reported for RBCs at 4°C were conserved but accelerated with increasing temperature. We calculated a median Q_{10} coefficient of 2.89 ± 1.03 for 48 individual metabolite concentrations, within the expected range of 2-3 for biological processes. We then integrated these metabolomics measurements into a cell-scale metabolic model to study

pathway usage, calculating a median Q_{10} coefficient of 2.73 ± 0.75 for 35 reaction fluxes. The relative fluxes through glycolysis and nucleotide metabolism pathways were consistent across the studied temperature range despite the non-uniform distributions of Q_{10} coefficients of individual metabolites and reaction fluxes. Together, these results indicate that the rate of change of network-level responses to temperature differences in RBC metabolism is consistent between 4°C and 37°C. More broadly, we provide a baseline characterization of a biochemical network given no transcriptional or translational regulation that can be used to explore the temperature dependence of metabolism.

2.1 The temperature dependence of RBC metabolism

The rate of biological processes increases with increasing temperature. The dependence of biochemical rates on temperature has been studied since the late 19th century using an Arrhenius-type approach [64–69]. The metric used for evaluating such temperature dependence is the Q_{10} value. If $Q_{10} = 2$ for a given process, then the rate of that process increases by a factor of 2 for every 10°C increase in temperature. The Q_{10} can be calculated from the slope of a rate vs. temperature plot, which is approximately linear over the biologically-relevant temperature range of 0°C to 40°C [70]. Individual enzymes have different Q_{10} coefficients that are generally expected to be in the 2 to 3 range [66, 67, 69].

Temperature dependency at the physiological level is determined using phenomenological measurements (such as growth rate) to study overall physiological changes [68, 70–73]. Changes at the physiological level depend on more than changes in the underlying individual biochemical reaction rates [74, 75]. For instance, various regulatory mechanisms (e.g., transcriptional, post-translational, allosteric) determine how cells respond to temperature shock [76]. The existence

of extra layers of regulation complicates the effects of temperature change on the biochemical network. Biology is inherently multi-scale, and the gap between observing the temperature dependence at the scale of an individual reaction and at the physiological level can be addressed through methods of systems biology [77].

RBCs represent an ideal cell type to study the temperature dependence of network-level metabolic biochemistry due to the absence of a nucleus and genetic material. This absence results in a lack of complicated transcriptional or translational regulation on metabolic enzyme activity. Allosteric and other regulation of enzymatic reaction rates is still present in RBCs, representing enzyme kinetic mechanisms and thus direct biochemical functions.

In this study, we investigated the temperature dependence of metabolism in the RBC at the network-level by examining the rate of change of metabolite concentrations and metabolic reactions rates. We measured exo- and endo-metabolomics profiles in human RBCs stored at four different temperatures that span the range between the *ex vivo* (storage) and *in vivo* (body) temperatures: 4°C, 13°C, 22°C, and 37°C. We regressed the concentration profile of each metabolite across the measured temperature range to calculate its Q_{10} value. We integrated these measurements with a cell-scale network reconstruction of RBC metabolism [40] that contains 216 metabolites (43% of which are measurable by quantitative metabolomic profiling) to calculate Q_{10} coefficients for reaction fluxes and to observe pathway usage. By examining metabolite profiles in the context of a cell-scale metabolic model, we were able to assess temperature dependence on a network level, thus bridging the gap between studies at the reaction and physiological levels.

2.1.1 Results

Measurement of the temperature dependence of RBC metabolism *ex vivo*

RBCs were collected using standard collection procedures and stored in SAGM media [46] at 4°C, 13°C, 22°C, and 37°C. Starting one day after RBC collection (taken as time zero in figures), metabolomic measurements were made in biological triplicate over time at each temperature (Fig. 2.3A). The data set was composed of 97 metabolites. In addition, standard blood bank quality control and assurance (QC/QA) measurements (e.g., hemolysis, pH) were made at multiple time points over 21 days (4°, 13°, and 22°) and 7 days (37°); all measured profiles are presented in Fig.s S6-S16.

As part of the baseline characterization, we observed the same metabolic changes that have previously been reported in the literature for RBC storage at 4°C. We observed the same previously-documented accumulation of lactate, 5-oxoproline, and hypoxanthine [18, 41, 42], and the depletion of AMP and the phosphoglycerate pool [41, 42]. We measured the same high levels of intracellular malate previously reported in SAGM at 4°C [41].

The measured hemolysis was almost identical at 4°C and 13°C within the accepted range (<0.8% cells), while the threshold was exceeded at 22°C and 37°C as the storage time progressed. The activity of lactate dehydrogenase and the concentration of free hemoglobin closely matched this trend. The pH fell from approximately 7 at all temperatures to around 6.4 before rising. The average mass of hemoglobin per cell was calculated and, although noisy, was approximately the same across all temperatures.

In glycolysis and the pentose phosphate pathway (PPP), glucose (both intracellular and extracellular), intracellular oxidized glutathione, 6-phosphogluconate, glucose 6-phosphate, fructose 1,6-bisphosphate (FBP), the phosphoglycerate pool, and phosphoenolpyruvate were all ob-

served to deplete, while lactate (both intracellular and extracellular) accumulated at a rate that increased consistently with temperature. The temperature vs. rate plots for glucose and glutathione displayed similar qualitative behavior with slightly nonlinear shapes, while fructose 1,6-bisphosphate, lactate, glucose 6-phosphate, and phosphoenolpyruvate all displayed highly linear behavior.

Several key metabolites from the nucleotide synthesis and salvage pathways were measured. There was an accumulation of intracellular xanthine which drastically increased with temperature, leading to one of the highest Q_{10} values found (4.84). There was also an accumulation of hypoxanthine (both intracellular and extracellular), extracellular xanthine, and uridine (both intracellular and extracellular). Intracellular GMP and IMP both depleted, with the latter showing an initial spike that increased in magnitude with increasing temperature. The behavior displayed by extracellular uridine at 37°C was the most noticeable here as there is no corresponding jump at 22°C. Three of these measurements—GMP, intracellular hypoxanthine, and extracellular hypoxanthine—had Q_{10} values of approximately 1.5, while the other measurements (with the exception of xanthine) had values that were all approximately 3.

The partial pressure of carbon dioxide rose initially but fell drastically at 13°C, 22°C, and 37°C, leading to a nonlinear shape and reduced R^2 value in the temperature vs. rate plot. 6-phosphogluconate, the phosphoglycerate pool, phosphoenolpyruvate, and glucose 6-phosphate all had Q_{10} values around 3, while the majority of the other measurements had values closer to 2. Some metabolites in adjacent reactions, like the phosphoglycerate pool and phosphoenolpyruvate, had very similar Q_{10} values, while others (e.g., intracellular vs. extracellular lactate) had different Q_{10} values. 2,3-Disphosphoglycerate (2,3-DPG) has previously been implicated in RBC storage lesion [78]. While we did not measure 2,3-DPG, the similar trends observed in intracel-

lular S-Adenosylmethioninamine (SAM), hypoxanthine, and intracellular oxidized glutathione to previous measurements [24] indicate that the 2,3-DPG trend is also consistent.

In the glutathione synthesis pathway, there was an accumulation of 5-oxoproline (both intracellular and extracellular), extracellular glutamate, and extracellular serine (Fig. 2.1). We observed a depletion of both intracellular and extracellular reduced glutathione depleted, except for the extracellular at 37°C increased after an initial depletion; no later increase in concentration was reported at 4°C for extracellular reduced glutathione through 42 days of storage [41]. Intracellular oxidized glutathione depleted with increased temperature, while the extracellular measurement displayed a large spike at 22°C and 37°C that was not observed at low temperatures; this spike was not observed through 42 days of storage at 4°C [41]. Extracellular glutamine showed large initial spikes in concentration at 22°C and 37°C that were not observed at low temperatures through 21 days.

We measured 15 amino acids. Many of these measurements were considered too noisy to be included in the Q_{10} calculations, although general qualitative trends were still visible. We observed increasing concentrations of extracellular L-glutamate, intracellular and extracellular L-lysine, intracellular and extracellular L-phenylalanine, extracellular L-serine, and intracellular L-tryptophan.

At high temperatures, both intracellular and extracellular L-histidine increased while concentrations remained fairly steady at low temperatures. The same behavior was observed for the intracellular and extracellular L-isoleucine/L-leucine pools. Several concentrations remained rather steady, including intracellular L-arginine, intracellular L-asparagine, intracellular L-aspartate, and intracellular L-threonine. Intracellular L-glutamate decreased steadily at 4°C and 13°C but later increased at high temperatures. Intracellular and extracellular L-glutamine

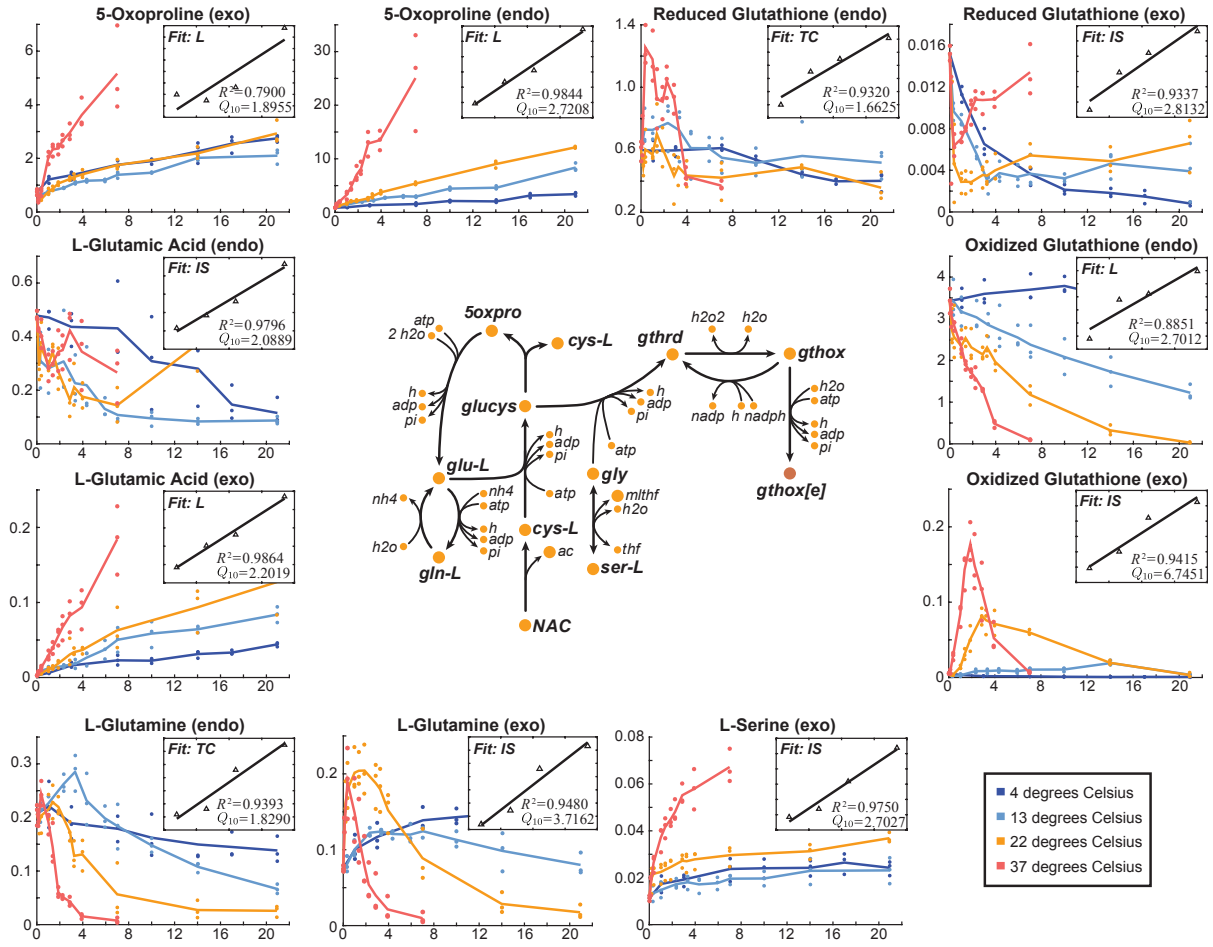

Figure 2.1: Metabolic map of glutathione synthesis. The y-axis for metabolites is concentration (mM); the x-axis for metabolites is time (days). Day 0 here is taken to be 1 day after the beginning of the storage period. The y-axis for inset plots is $\log_2$(rate) where rate in units of concentration per time; the x-axis for inset plots is temperature (°C).

both showed an initial spike followed by a depletion. The depletion was more pronounced at high temperatures. The measurements for intracellular L-serine, L-arginine, L-asparagine, the L-isoleucine/L-leucine pool, L-histidine, L-phenylalanine, L-threonine, L-tryptophan, L-tyrosine, and L-valine as well as the extracellular L-isoleucine/L-leucine pool were too noisy to determine meaningful trends.

One of the more interesting results was the behavior of the measured ions: extracellular chloride, potassium, and sodium (Fig. 2.2). The trends observed in potassium and sodium at

4°C and 13°C were similar, with small quantitative differences. At 22°C and 37°C, the same changes were observed to be more pronounced. It can be expected that the magnitude of change in potassium concentration is greater than that of sodium [79], a behavior which was observed here across all temperatures. Sodium was among the slowest-scaled measurements ($Q_{10} = 1.63$), while chloride and potassium were closer to the center of the distribution ($Q_{10} = 2.44$ and $Q_{10} = 2.14$, respectively).

At low temperatures, the membrane ATPase in RBCs undergoes a reversible inhibition which shuts down active Na/K transport and instead allows for a steady leak of cations across the membrane [79, 80]. Consequently, extracellular potassium is expected to rise and extracellular sodium is expected to fall at low temperatures since the Na/K pump is not actively pumping ions back across the membrane. At high temperatures when the pump is functioning properly, we can expect to observe steady cation concentrations. However, our results here do not coincide with these expectations, as we see the same fall of sodium levels and rise of potassium levels at all temperatures (Fig. 2.2). One contributing factor to these results is the depletion of intracellular ATP [41], which is required for the Na/K pump to function. As ATP depletes, the Na/K pump will function slower, resulting in a rise in extracellular potassium and fall of extracellular sodium as observed here. Another possible explanation for the unexpected cation behavior involves the role of magnesium in membrane ATPase activity. It has been previously suggested that the citrate additive meant as an anticoagulant could bind free magnesium ions, further inhibiting ATPase activity [79]. We observed increasing amounts of intracellular citrate at 22°C and 37°C but steady, low concentrations at 4°C and 13°C. Thus, if the large amounts of intracellular citrate are bound to magnesium at high temperatures, the irregular behavior of citrate and the cations could be related. Ultimately, the implications of the cation behavior seen here are uncertain and

require further investigation.

A few measurements exhibited qualitatively different concentration profiles across the four temperatures. The most drastic difference was observed in extracellular oxidized glutathione (Fig. 2.2). Using the initial slope to fit this nonlinear profile allowed us to account for the magnitude of the spike observed at higher temperatures; this resulted in a calculated Q_{10} value of 6.75, the highest among all measurements. Intracellular citrate showed a similar concentration spike at high temperatures that was not observed at low temperatures (Fig. 2.2). Previous studies reported a decreasing concentration of citrate at 4°C over 42 days of storage in SAGM, with a slight spike just before day 40 [41], which matches the trend observed here at 4°C. Intracellular inosine exhibited no clear trend across temperature, with a substantial spike at 37°C that is not present at the other three temperatures (Fig. 2.2). Even at 37°C, however, the concentration of inosine was very low most likely due to its toxicity [81].

In order to determine temperature dependence, we first needed to determine how relative storage time scaled across temperature (FDA regulations set the maximum storage time for RBCs at 42 days). To determine a network-level Q_{10}, we used multivariate statistical analysis on the metabolomics data to assess the impact of temperature changes on systemic metabolism. The network-level Q_{10} was used to determine the time points that represented the same metabolic phase at each temperature. Once these time periods were defined, we calculated Q_{10} coefficients for each measured metabolite using linear regression. The metabolomics data was then integrated into a mechanistic cell-scale model to calculate the rate of each reaction (i.e., the flux) in the network at each temperature; these calculated reaction rates were used to determine Q_{10} coefficients for reactions and to assess pathway usage across temperature.

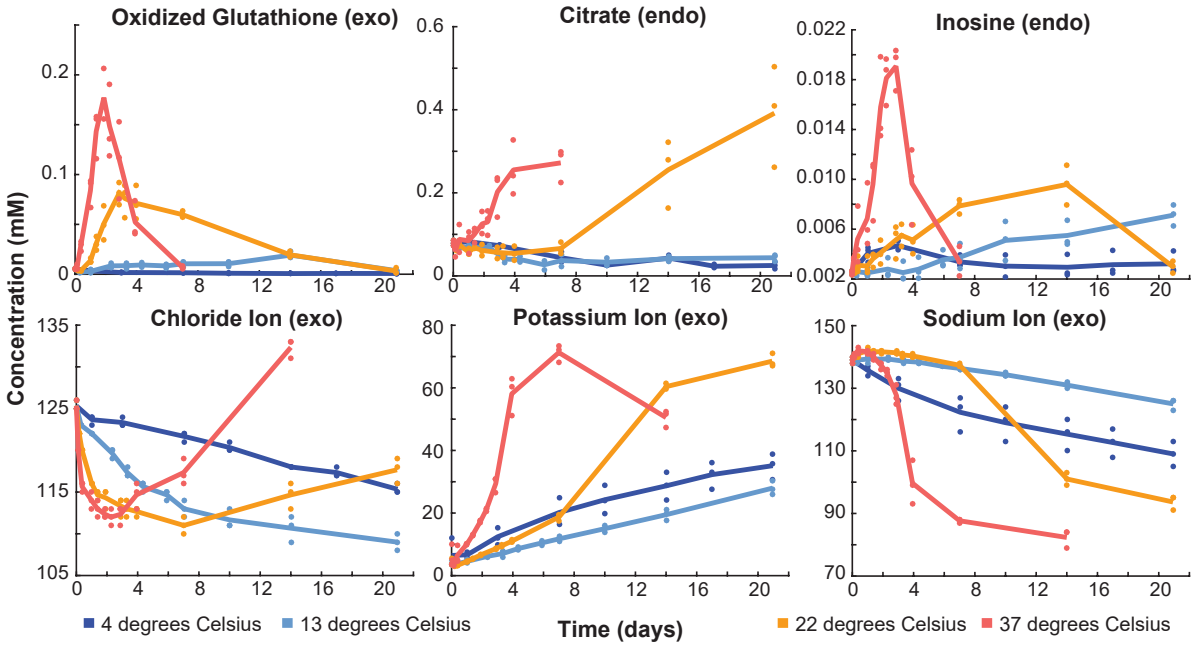

Figure 2.2: Measurements with no clear dynamic trend. Measurements with qualitative time course differences at higher temperatures and measured ions. Day 0 here is taken to be 1 day after the beginning of the storage period.

Network-level temperature dependence

Following published reports for RBC storage at 4°C [41,42,62], principal component analysis (PCA) was performed in order to obtain a global characterization of the dataset. PCA is a multivariate statistical method that reduces the dimensionality of a complex data set by calculating the relative contribution of each measurement to the overall variability observed in the data. For each temperature, we performed PCA (Fig. 2.3B) on the time-series concentration profiles of eight recently identified extracellular metabolites (adenine, glucose, hypoxanthine, lactate, malate, nicotinamide, 5-oxoproline, and xanthine) that robustly represent the RBC metabolome under storage conditions [62]. These metabolites serve as qualitative biomarkers for the age of stored RBCs and have been shown to also be good quantitative predictors for other systemic metabolite concentrations [82,83]. In order to make an accurate comparison across temperatures,

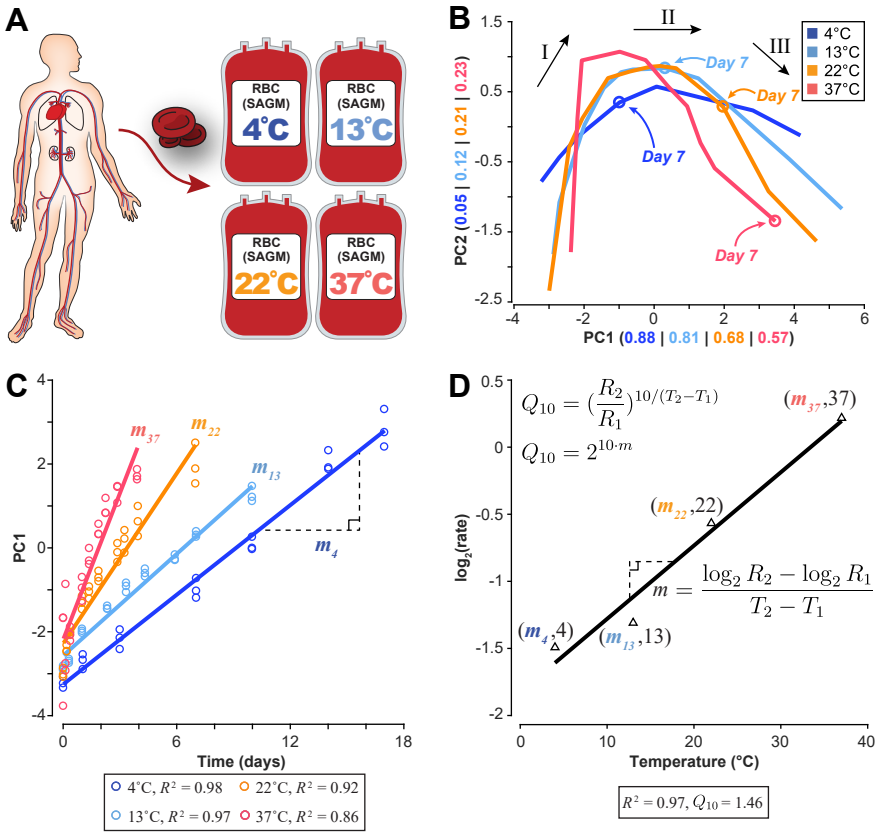

Figure 2.3: Data generation and analysis workflow. (**A**) Human red blood cells were collected, stored in SAGM media at 4°C, 13°C, 22°C, and 37°C, and metabolically profiled across multiple time points. (**B**) Principal component analysis (PCA) of the eight extracellular biomarkers (same loading coefficients applied to data at each temperature). Overlaying these plots on the same axes shows that the shape of the three-phase metabolic decay is conserved but accelerated with increasing temperature, as evidenced by the location of the Day 7 time point. The numbers in parentheses represent the amount of variance explained by each component. Black arrows and numerals label the three metabolic shifts that occur over the storage period. (**C**) The first principal component was plotted against the time vector at each temperature to determine the relative storage time at each temperature. Linear regression was used to estimate the rate of change, showing strong correlation between PC1 and time at each temperature. (**D**) These rates of change were then used to estimate the change in metabolic rate for every 10° (Q_{10}) from an Arrhenius-type $\log_2$(rate) vs. temperature plot.

the same loading coefficients were applied to the data at each temperature (see Experimental Procedures for full details).

PCA revealed the same metabolic "shifts" that have been previously observed at 4°C [41, 42]. These shifts separate three distinct metabolic states that can be reliably determined from

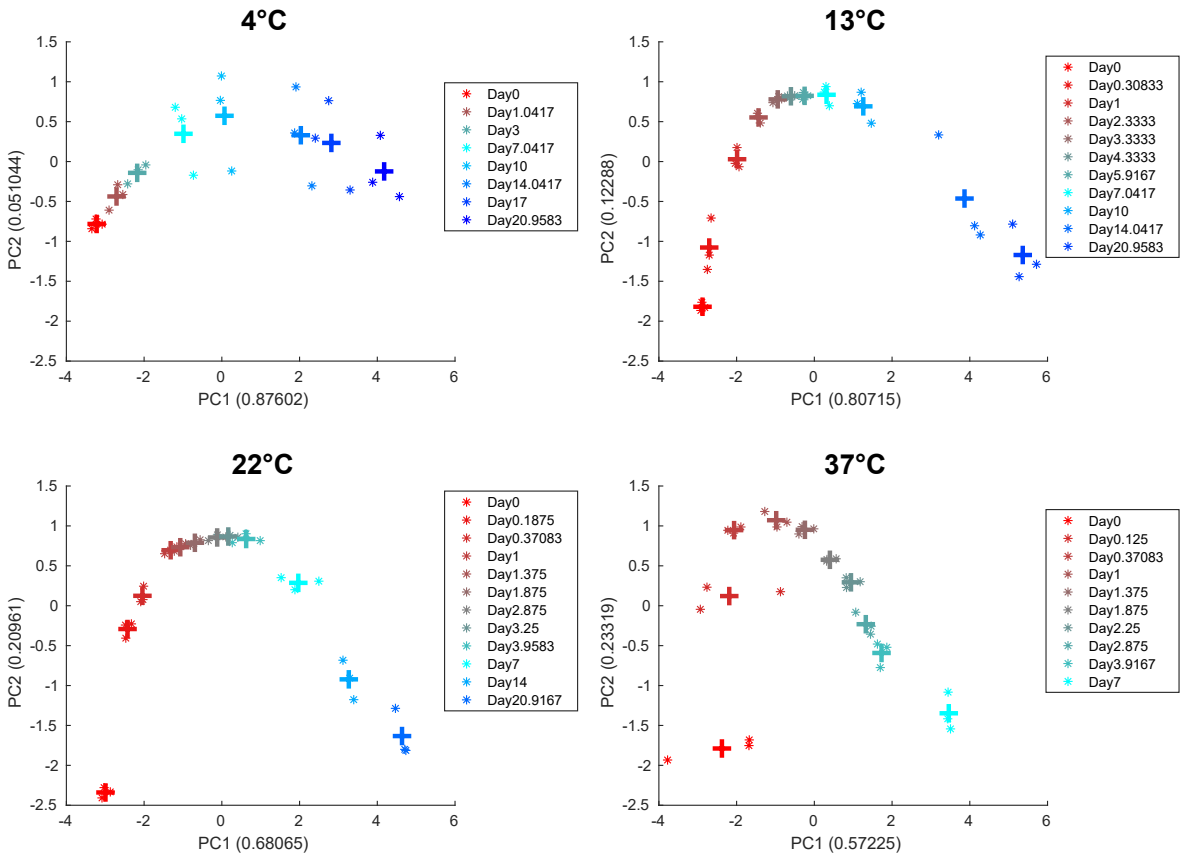

Figure 2.4: PCA for biomarkers. Principal component analysis of the eight extracellular biomarkers. PCA of the 4°C data yielded loading coefficients that were then applied to the data at the other temperatures.

the profiles of the biomarkers [62]. During storage at 4°C, the two shifts in the PCA plot occur approximately at Days 10 and 17. Here, these same metabolic states were observed to be conserved but notably accelerated with temperature as evidenced by the location of the Day 7 time point at each temperature (Fig. 2.3B and Fig. 2.4). We identified the duration of the first metabolic state at each temperature. We then used these time points to determine the starting and ending points for the linear regression that would be used to calculate individual metabolite and reaction Q_{10} coefficients.

The first principal component at each temperature was highly correlated with time

(Fig. 2.3C), which yielded a "network-level" Q_{10} of 1.46 ($R^2 = 0.97$) that describes how the system proceeds in storage time. The third metabolic state at 4°C is primarily characterized by a general loss of function as the RBC undergoes severe morphological changes [20, 24], often leading to complications for transfusion patients [84, 85]. Thus, we only used measurements from the first two metabolic states to calculate the network-level Q_{10}. These results show an overall three-state metabolic decay that is observed to accelerate with increased temperature.

Metabolite-level temperature dependence

In order to determine the temperature dependence of individual metabolites, we used the data from the first metabolic state at each temperature (identified from the PCA results in Fig. 2.3B). We made this choice since the data is believed to be the most accurate as the cells are still intact and metabolism is functioning the closest to its normal physiological state. We linearly regressed the concentration profile of each metabolite at each temperature and used these rates to calculate a Q_{10} value (Fig. 2.3C and 2.3D). Not all metabolite profiles could be accurately fit with a linear curve during the first state; to account for this, we did not include metabolites with an $R^2 < 0.50$. Q_{10} coefficients for the 48 metabolites whose profiles could be estimated well with a linear fit are reported in Table 2.1. The calculated Q_{10} coefficients span 1.28 (extracellular 5-oxoproline) to 5.89 (intracellular hypoxanthine).

The calculated metabolite Q_{10} coefficients generally fall in the expected range of 2 to 3 for biochemical reactions [66, 67, 69], with a median of 2.89 (Fig. 2.5). The standard deviation of the metabolite Q_{10} coefficients was 1.03, indicating that although the distribution was centered in the expected range for biological measurements, the temperature scaling was not uniform across the network. Interestingly, the biomarker pools that have been shown to robustly define

Table 2.1: Q_{10} coefficients for extracellular and intracellular metabolites. Q_{10} coefficients for extracellular (exo) and intracellular (endo) metabolites. * denotes previously reported biomarker [62].

Metabolite	Q_{10}	R^2
*5-Oxoproline (exo)	1.28	0.69
L-Glycerate (endo)	1.84	0.52
S-Adenosylmethioninamine (endo)	1.94	0.83
L-Glutamate (endo)	2.01	0.94
cis-Aconitate (endo)	2.06	0.91
L-Glutamate (exo)	2.15	0.99
L-Aspartate (endo)	2.18	0.73
*Nicotinamide (exo)	2.19	0.91
Mannitol (exo)	2.22	0.86
Uridine (exo)	2.27	0.94
Citrate (exo)	2.32	0.99
L-Glutamine (endo)	2.34	0.90
Reduced Glutathione (exo)	2.39	0.99
Choline (endo)	2.41	0.85
GMP (endo)	2.42	0.97
5-Oxoproline (endo)	2.56	0.98
Lactate (endo)	2.61	0.98
L-Acetylcarnitine (endo)	2.62	0.72
Xanthine (endo)	2.65	0.84
cis-Aconitate (exo)	2.75	0.91
Phosphorylcholine (endo)	2.75	0.89
*Lactate (exo)	2.78	0.99
Uridine (endo)	2.79	0.96
*Glucose (exo)	2.81	0.98
AMP (endo)	2.96	0.79
*Adenine (exo)	2.98	0.93
L-Serine (exo)	2.99	0.99
S-Adenosylhomocysteine (endo)	3.01	0.83
Malate (endo)	3.02	0.97
Oxidized Glutathione (endo)	3.03	0.95
L-Carnitine (exo)	3.04	0.97
5-MTA (endo)	3.07	0.73
Adenine (endo)	3.17	0.95
Glucose 6-Phosphate (endo)	3.19	0.97
L-Glutamine (exo)	3.23	0.99
L-Phenylalanine (exo)	3.33	0.75
IMP (endo)	3.45	1.00
L-Lysine (exo)	3.46	0.96
Chloride Ion (exo)	3.94	0.99
ATP (endo)	3.95	0.92
L-Histidine (exo)	4.07	0.89
ADP (endo)	4.48	0.84
6-Phosphogluconate (endo)	4.84	0.79
Oxidized Glutathione (exo)	4.99	0.95
*Malate (exo)	5.05	0.99
L-Lysine (endo)	5.18	0.97
Reduced Glutathione (endo)	5.61	0.96
Hypoxanthine (endo)	5.89	0.95

the metabolic decay process [62] represented almost the full range of calculated Q_{10} coefficients, from 1.28 (5-Oxoproline) to 5.05 (malate); extracellular xanthine and hypoxanthine did not have calculated Q_{10} coefficients due to the R^2 cutoff. Several metabolite Q_{10} coefficients fell below 2.00 or above 3.00, including extracellular 5-oxoporline (1.28), ATP (3.95), ADP (4.48), and

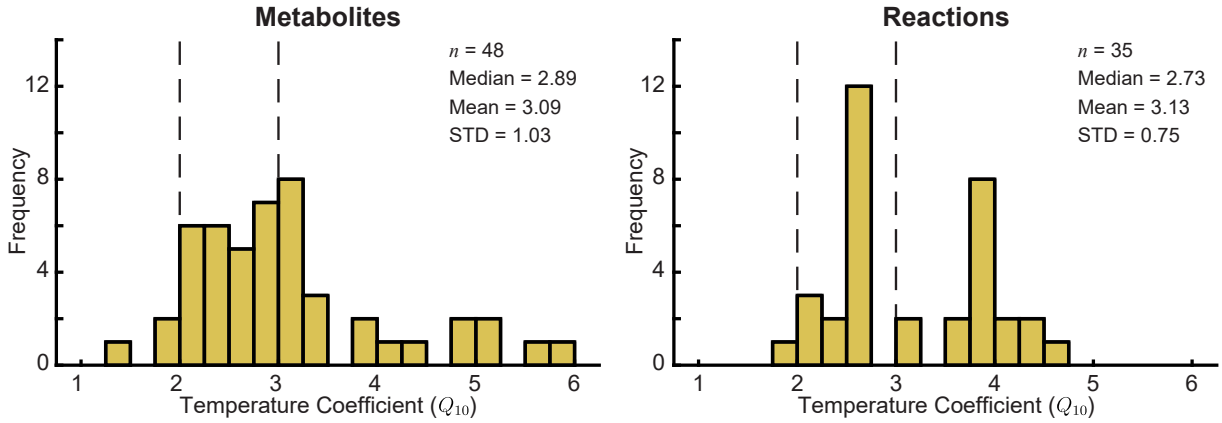

Figure 2.5: Distribution of Q_{10} coefficients for metabolites and reactions. Q_{10} coefficients for metabolites were calculated based on the observed change in metabolite concentration across temperature. The vertical dashed lines at $Q_{10} = 2$ and $Q_{10} = 3$ represent the typical estimated range of Q_{10} coefficients for biological processes.

extracellular malate (5.05).

Reaction-level temperature dependence

Next, we investigated the temperature dependence of biochemical reactions in the metabolic network. Previous studies have investigated the temperature dependence of individual reactions [64–69], but our goal was to use systems biology approaches to determine the temperature dependence of all reactions in the network together. To this end, we used a mechanistic cell-scale model of the RBC [40] to calculate the flux state of the network (i.e., the flux through each reaction in the system). The flux through a reaction (a rate with units mmol/hr) was calculated at each temperature; these values were then used to calculate a Q_{10} for each reaction using the same procedure shown in Fig. 2.3D. We tailored the model to the physiology at each temperature by integrating the metabolomics measurements for the first metabolic state into the model according to [52]. See Experimental Procedures for full details on flux modeling and metabolomics integration; the full method is presented in Chapter 4.

Table 2.2: Q_{10} coefficients for reaction fluxes. Reaction and metabolite abbreviations are BiGG identifiers.

Reaction	Q_{10}	R^2	Formula
GGCT	1.98	0.93	glucys → 5oxpro + cys-L
GLUCYS	2.06	0.93	atp + cys-L + glu-L → adp + glucys + h + pi
GTHS	2.21	0.93	atp + glucys + gly → adp + gthrd + h + pi
GLNS	2.23	0.91	atp + glu-L + nh4 → adp + gln-L + h + pi
ADK1	2.29	0.91	amp + atp ↔ 2 adp
PPA	2.40	0.94	h2o + ppi → h + 2 pi
PGK	2.53	0.95	3pg + atp ↔ 13dpg + adp
ENO	2.57	0.95	2pg ↔ h2o + pep
PGM	2.57	0.95	2pg ↔ 3pg
ADA	2.58	0.99	adn + h2o + h → ins + nh4
PYK	2.59	0.95	adp + h + pep → atp + pyr
HEX1	2.60	0.95	atp + glc-D → adp + g6p + h
AMPDA	2.62	0.99	amp + h2o + h → imp + nh4
GAPD	2.63	0.95	g3p + nad + pi ↔ 13dpg + h + nadh
PFK	2.65	0.96	atp + f6p → adp + fdp + h
FBA	2.65	0.96	fdp ↔ dhap + g3p
TPI	2.65	0.96	dhap ↔ g3p
PGI	2.72	0.96	g6p ↔ f6p
PEPCK	3.04	0.93	gtp + oaa → co2 + gdp + pep
NDPK1	3.04	0.93	atp + gdp ↔ adp + gtp
NTD11	3.60	0.91	h2o + imp → ins + pi
NTD7	3.70	0.91	amp + h2o → adn + pi
PRPPS	3.78	0.93	atp + r5p ↔ amp + h + prpp
PDE1	3.79	0.91	camp + h2o → amp + h
ADNCYC	3.79	0.91	atp → camp + ppi
DPGase	3.79	0.91	23dpg + h2o → 3pg + pi
GLUN	3.79	0.91	gln-L + h2o → glu-L + nh4
GUAPRT	3.81	0.91	gua + prpp → gmp + ppi
NTD9	3.81	0.91	gmp + h2o → gsn + pi
PUNP3	3.81	0.91	gsn + pi ↔ gua + r1p
PPM	4.04	0.95	r1p ↔ r5p
PUNP5	4.16	0.96	ins + pi ↔ hxan + r1p
OPAHir	4.26	0.92	5oxpro + atp + 2 h2o → adp + glu-L + h + pi
PC	4.34	0.94	atp + hco3 + pyr → adp + h + oaa + pi
HXPRT	4.54	0.96	hxan + prpp → imp + ppi

Model simulations yielded flux states for each temperature. We excluded transporters and reactions that carried flux at fewer than three temperatures to ensure the accuracy of the Q_{10} calculations. The calculated Q_{10} coefficients for 35 reactions are shown in Table 2.2 (only fits with $R^2 \geq 0.50$ were included in analysis). The distribution of reaction Q_{10} coefficients (Fig. 2.5) was tighter than that of the metabolite Q_{10} coefficients (STD of 0.75 for reactions vs. 1.03 for metabolites) but was still centered in the expected 2-3 range (median of 2.73). Several reactions in nucleotide metabolism and glutathione metabolism had Q_{10} coefficients above 3.5.

Pathway usage across temperature

We then examined pathway usage across the 33°C temperature range studied. Each reaction in the RBC reconstruction has previously been assigned to one of 18 different metabolic subsystems [40]. We first investigated whether the reactions in any of these pathways shared a similar dependence on temperature (i.e., pathways in which reactions had similar Q_{10} coefficients). We performed a pairwise comparison of the Euclidean distance between Q_{10} coefficients of reactions from a given subsystem (n reactions) versus 100,000 random permutations of n reactions from all calculated Q_{10} coefficients. The reactions in glycolysis ($n = 12$) and nucleotide metabolism ($n = 7$) both showed statistically significantly smaller distances between Q_{10} coefficients than would be expected due to random chance ($p < 7e\text{-}5$ and $p < 0.047$, respectively). Notably, these two metabolic subsystems contain reactions that interact with several of the storage age biomarkers (glucose, lactate, hypoxanthine, and xanthine). In particular, the temperature dependence of the glycolytic reactions is dictated by glucose, the major input to that linear pathway (Fig. 2.6).

While the enrichment of a given pathway for similar Q_{10} coefficients depends on the reaction fluxes, this analysis does not directly measure whether the flux states between two temperatures was similar. In order to investigate the conservation of pathway usage across temperature, we normalized the data at each temperature to glucose uptake and calculated the percent distance between the flux states (i.e., the flux through each reaction) at 13°C, 22°C, and 37°C and the flux state at 4°C. Reactions that were unused at every temperature were excluded ($n = 61$). Of the remaining 166 reactions, 33 (19.9%) had less than 50% difference at each of the three higher temperatures. All of the glycolytic reactions for which Q_{10} coefficients were calculated (Fig. 2.6) were present in this subset of reactions.

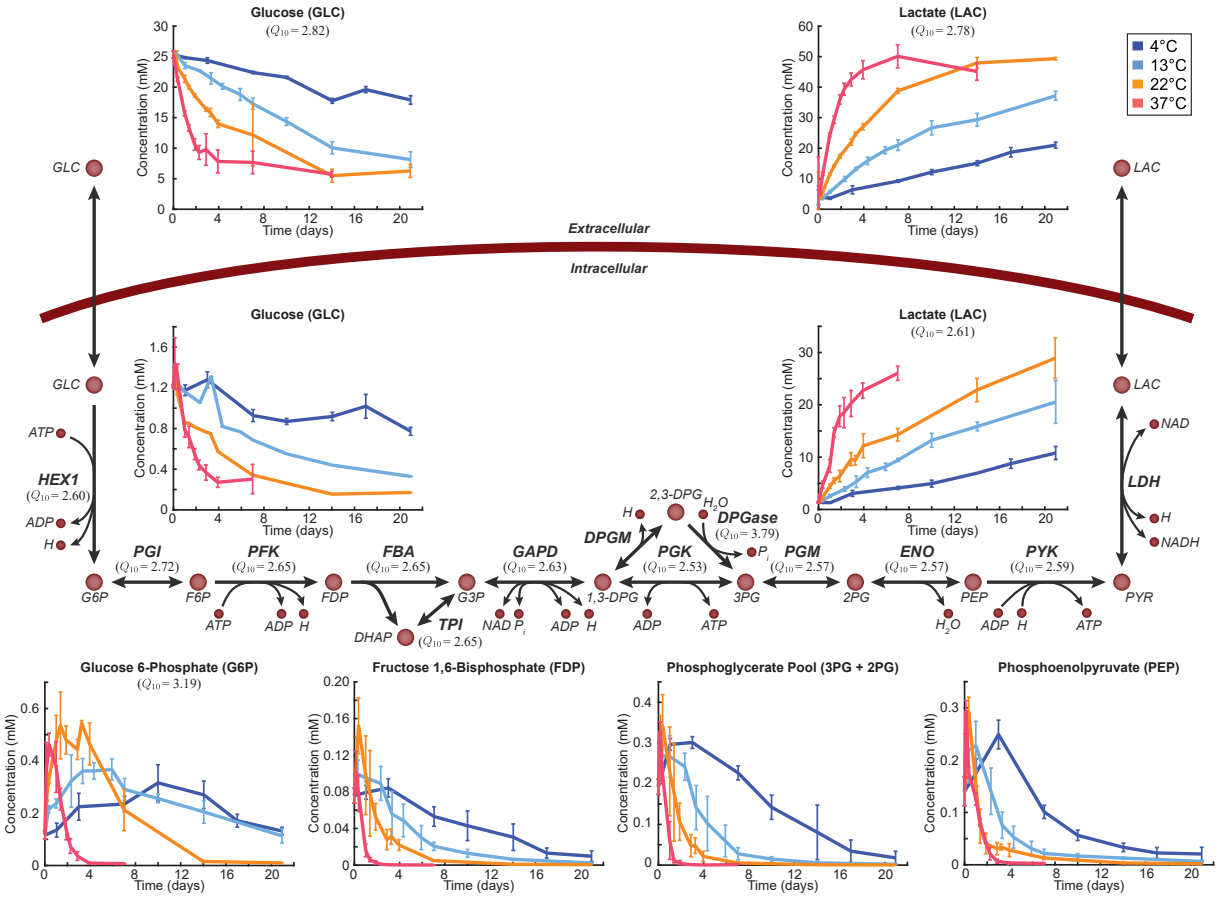

Figure 2.6: Metabolic map of glycolysis. Day 0 here is taken to be 1 day after the beginning of the storage period. Q_{10} coefficients are provided for those metabolites and reactions which could be calculated.

Organizational structure of the network

To this point, we have investigated how the metabolomics measurements could be used to determine the temperature dependence of individual metabolites and reactions and of the network. When we integrated the metabolomics measurements into the cell-scale model, the network structure at each temperature changed due to the accumulation and/or depletion of metabolites. These results, however, do not provide any explanation as to why we observe certain Q_{10} coefficients for certain reactions. To answer this question, we studied the organizational structure of the network in terms of the coupling of certain reaction fluxes. Two reaction fluxes

are said to be "flux coupled" if the ratio of one to the other is constant [86]; the flux coupling of a network is a property inherent to its topology. A set of coupled reactions is simply the linking of coupled reactions into a single pathway.

To determine whether the coupling of the network changed with temperature, we defined the coupling characteristic of a network to be the mean of the coupled reaction sets. We compared the base network structure of the RBC metabolic network to that of each temperature, revealing that the coupling characteristic of the network decreased approximately 5-10% at each temperature. This result prompted us to ask whether this decoupling was more or less than would be expected due to random chance. We ran a permutation test, determining that the decoupling of the networks observed with a change in temperature was significantly less than would be expected due to random chance ($p <$ 6e-3; see Experimental Procedures for details on the generation of random networks and the permutation test). Thus, the RBC metabolic network is robust against changes in temperature over the measured 33°C range.

2.1.2 Discussion

The temperature dependence of biological processes is of fundamental interest. Temperature coefficients (Q_{10} coefficients) have been used to characterize the temperature dependency of individual biochemical reactions and of organism-level behavior. While these studies have yielded valuable insights, there is a gap between studying temperature dependency at the biochemical and physiological levels [64–73]. Systems biology principles can be used to move past the individual reaction level and assess temperature dependence on the network level, effectively bridging the gap between the previous work on temperature dependency at the reaction and physiological levels. In this study, we used deep-coverage metabolomics of human red blood cells in storage

to investigate the temperature dependence of network-level biochemistry. We chose a range of temperatures that ranges from the storage temperature of RBCs (4°) to the body temperature (37°). By studying temperatures in this range, we provide baseline data with no transcriptional or translational regulation that can be used to begin to understand the temperature dependence of metabolism in broader contexts. Specifically, this data addresses the "passive" control that temperature has over network fluxes and metabolites in a system that has most of central carbon metabolism but is not growing. The results obtained here have several primary implications.

First, we determined that while the rate of change for each metabolite and reaction increased with increasing temperature, the network-level response was dampened (i.e., lower Q_{10}). This behavior is to be expected because of the difference in response times between individual reactions and a pathway. Notably, the scaling of metabolite and reaction temperature dependence was not observed to be uniform across the network, with high variability in Q_{10} coefficients. The fact that these network-level behaviors are conserved is a surprising result because certain enzyme inactivations might be expected to be qualitatively disruptive of network behavior at various temperatures. The observed behavior indicates that individual metabolite and reaction Q_{10} coefficients vary dramatically in order to preserve global network characteristics. This variability was particularly notable for the storage age biomarkers [62], an interesting result due to their ability to define the qualitative trend of the entire network. Overall, such variability can be expected due to the thermodynamics and kinetics associated with biochemical and enzymatic reactions. Approximately half of the calculated temperature coefficients fell in the 2 to 3 range (Fig. 2.5), which is generally accepted as the typical estimate for biological systems [66, 67, 69]. Several of the metabolites (e.g., malate, hypoxanthine, glutathione) and reactions (several glutathione synthesis and nucleotide metabolism reactions) which we calculated to have high

Q_{10} coefficients are relatively disconnected from the rest of the network and thus may not be as affected.

Second, the pathway analyses indicate that even without transcriptional or translational regulation, RBCs maintain consistent flux through glycolysis and nucleotide metabolism across temperature. The storage age biomarkers in glycolysis (glucose and lactate) represent the primary exchanges for the RBC metabolic network. The preservation of these exchanges indicates that sustaining these reactions is inherent to the network topology and necessary to maintain physiological functions. The change in pathway usage across the rest of the network is dependent upon kinetics and thermodynamics, properties that are subject to change over a 33°C temperature range.

Third, it is important to be aware that the temperature dependencies calculated here are based on *ex vivo* measurements—not single-enzyme *in vitro* assays. Such an assay would inherently be absent from any regulatory or network influences. Our results, while also generated in the absence of transcriptional or translational regulatory effects, suggest that the organizational structure of the network influences the temperature dependence of individual enzymes. Thus, we would not expect the systemic Q_{10} coefficients calculated here to correlate with previously reported *in vitro* values. For example, previous studies have reported Q_{10} coefficients pyruvate kinase (PYK) in the range of 3.3-4.2 for fish [68], 1.4-1.9 for bats [87], and 1.66-1.69 for turtles [88]; we calculated a systemic Q_{10} coefficient of 2.59 for PYK. The use of a cell-scale model to calculate the flux coupling of various reactions in the metabolic network provides an unambiguous explanation for why sets of reactions in the network have similar Q_{10} coefficients despite variation of individual Q_{10} coefficients. Additionally, the flux coupling results suggest that the *ex vivo* Q_{10} coefficients calculated here would be different from *in vitro* Q_{10} coefficients and from

each other because the network structure intrinsically constrains the temperature dependence of certain reactions. The networks at the measured temperatures displayed no significant loss of flux coupling compared with the base model, suggesting that the structural characteristics of the RBC network are robust despite the accumulation or depletion of intermediate metabolites.

We have described the temperature dependence of a human metabolic network over a temperature range of 33°C, from 4°C (the FDA-defined RBC storage temperature for transfusion) to 37°C (the *in vivo* temperature). While the RBC represents a simple cellular system, the lack of complex regulatory motifs allows for a direct interrogation of the systems biochemistry underlying metabolism. The use of systems biology methods empowered us to assess the temperature dependence of an *ex vivo* metabolic system at the network-level, helping to understand the relationship between the structure and function of the RBC metabolic network.

2.1.3 Experimental Procedures

Experimental methods

RBC unit preparation, data measurements, and metabolomics analyses were performed as previously reported [41, 89]. Metabolomics measurements were made over 21 days (4°, 13°, and 22°) and 7 days (37°).

Whole blood was collected from three healthy blood donors into 63 mL of CPD anticoagulant solution (Fenwall, Lake Zurich, IL, USA) and was held on butanediol plates for minimum of 2 hours at 4°C. After separation of plasma and buffy coat by centrifugation (800g, 11 minutes, 20°C), RBCs were suspended in 100 mL of SAGM additive solution (Fenwal) and leukodepleted using a SEPACELL Pure RC white cell reduction filter (Asahi Corporation, Tokyo, Japan). Each unit was split into 4 standard Pediatric storage containers (Fenwall) and samples were collected

via sterile connected clave valve collected into a syringe with a Luer-Lock connector on different time points. The National Bioethics Committee of Iceland and Icelandic Data Protection Authority approved the study.

RBC samples were first processed to separate supernatant and cells by centrifugation of 0.5 mL of RBC (1600g, 15 min, 4°C) and then prepared separately. Immediately after centrifugation, cell-free supernatant was removed and collected in separate tubes. RBC supernatant (80 μL) was processed by adding internal standard mixture (30 μL) and methanol (0.5 mL). The internal standard mixture contained the following standards: phenylalanine d2 (72 mg/L), succinate d4 (50 mg/L), glucose ^{13}C6 (2100 mg/L), carnitine d9 (20 mg/L), glutamic acid d5 (30 mg/L), lysine d4 (90 mg/L). Alanine d4 (300 mg/L), AMP ^{13}C1015N5 (50 mg/L) and 1 mL of -20°C methanol-water (7:3) were added to the cell pellets. Cells were lysed by two freeze and thaw steps. Samples were centrifuged (15000g, 20 min, 4°C) and supernatant was transferred into a new tube. 1 mL of 20°C methanol-water (7:3) (1 mL) was added to pellets and samples were vortexed for 1 min, centrifuged (15000g, 20 min, 4°C) and supernatant was added to the precedent. Samples were dried using a vacuum concentrator, reconstituted in 300 μL H_2O:ACN (50:50), and filtered to remove residual hemoglobin by centrifugation (Amicon Ultra 0.5 mLfilter, 15000g, 4°C, 60 min). The first sample was taken one day after storage had begun; this time point is taken as $t = 0$ in all figures.

At each time point, we monitored typical QC/QA hematological parameters of RBC physiology. A blood gas analyzer (ABL90FLEX, Radiometer, Copenhagen Denmark) was used to determine pH (37°C), pO2, and pCO2, total hemoglobin, K+, Na+, Cl- in the media. RBC concentration, mean RBC volume, hematocrit, RBC distribution width, and white blood cell count were assayed using an hematoanlyzer (CELLDYN Ruby, Abbot Diagnostics, Lake For-

est, IL, USA). Hemolysis was calculated using the following formula: % hemolysis = (supernatant Hb (g/L)/total Hb (g/L)) × (100-Hct (%)), where total Hemoglobin (Hb) and hematocrit was analyzed using a hematoanalyzer and supernatant Hb was measured with a HemoCue Plasma/Low Hb system (HemoCue, Angelshol, Sweden). Adenosine triphosphate (ATP) and 2,3-diphosphoglycerate (2,3-DPG) concentrations, employing the CellTiter-Glo kit (Promega) and the 2,3-DPG kit (Roche Diagnostics), respectively. Lactate dehydrogenase (LDH) activity was assessed by an LDH assay kit (ab102526, Abcam, Cambridge, UK).

The metabolomics analysis was performed using ultra performance liquid chromatography (UPLC), (Acquity, Waters, Manchester, UK) coupled with a quadrupole/time of flight mass spectrometer (MS) (Synapt G2, Waters). Chromatographic separation was achieved by working in hydrophilic interaction liquid chromatography (HILIC) mode using an Acquity amide column, 1.7 μm (2.1 × 150 mm) (Waters).

All RBC samples were analyzed three times: once in positive ionization mode using acidic chromatographic condition and twice in negative ionization mode using both acidic and basic chromatographic conditions. During acidic conditions, mobile phase A was 100% ACN and B was 100% H_2O, both containing 0.1% formic acid. The following elution gradient was used during acidic condition: 0 min 99% A; 7 min 30% A; 7.1 min 99% A; 10 min 99% A. Basic conditions employed ACN:sodium bicarbonate 10 mM (95:5) as mobile phase A and ACN:sodium bicarbonate 10 mM (5:95) as mobile phase B. During basic condition the following elution gradient was used: 0 min 99% A; 6 min 30% A; 6.5 min 99% A; 10 min 99% A.

In all conditions, the flow rate was set at 0.4 mL/min, column temperature was set at 45°C, and injection volume at 3.5 μL. The MS operated using a 1.5 kV capillary voltage, 30 V sampling cone and 5 V extraction cone. The cone and the desolvation gas flow were 50 L/h

and 800 L/h, respectively. The source and desolvation gas temperatures were 120° and 500°C, respectively. MS spectra were acquired in centroid mode from m/z 50 to 1000 using scan time of 0.3 s. Leucine enkephalin (2 ng/μL) was used as lock mass (m/z 556.2771 and 554.2615 in positive and negative experiments, respectively).

Identification of unexpected metabolites was achieved by integration, alignment, and conversion of MS data points into exact mass retention time pairs (MarkerLynx, v4.1, Waters). The identity of the unexpected metabolites was established by verifying peak retention time, accurate mass measurements, and tandem mass spectrometry against our in-house database and online databases, including HMDB [90] and METLIN [91]. TargetLynx (v4.1, Waters) was used to integrate chromatograms of targeted metabolites. Extracted ion chromatograms were extracted using a 0.02 mDa window centered on the expected m/z for each targeted compound. Quantitation was performed by external calibration with reference standards. Details regarding the quantitative analysis (including the linear range and LOD) are reported in Tables S1 and S2.

Bacterial testing was performed at the end of the study. 10 mL from each unit was injected into a BacT/Alert FA Plus flasks for aerobic bacteria and BacT/Alert FN Plus flasks for anaerobic bacteria. The flasks were cultured for 5 days on a BacT/ALERT 3D microbial detection system (BioMrieux, Marcy l'Etoile, France) and analyzed by a specialist in medical microbiology.

Multivariate statistical analysis

Principal component analysis (PCA) has previously shown three distinct metabolic shifts occurring over the 42-day storage period for RBCs at 4°C: days 1-10, days 10-17, and days 17-42 [41,42,62]. In the data presented in this study, RBCs were stored at 4°C for 21 days; in order

to ensure that all three shifts were captured in full, we used the metabolomics data from Bordbar et al. [41] to calculate the weightings for the principal components. PCA was performed on the Z-scores of the eight extracellular biomarkers [62]: adenine, glucose, hypoxanthine, lactate, malate, nicotinamide, 5-oxoproline, and xanthine. These weights were then used to transform the data presented here so that a more representative comparison could be made across all temperatures. The components were rotated in the transformed space such that the Day 0 measurement appears in the lower left corner of the plot.

Calculation of temperature coefficients

Each measurement was plotted against time in order to find the rate of change. The rate of change was calculated through simple linear regression according to

$$\hat{y} = \beta_0 + \beta_1 x + \epsilon \tag{2.1}$$

where $\hat{y}$ is the calculated response, β_0 is the y-intercept, β_1 is the regression coefficient (i.e., the slope), and ϵ is the error (Fig. 2.3C). In order to determine the goodness of fit, the coefficient of determination (R^2) was calculated by

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{2.2}$$

where $\hat{y}$ is the calculated value of y and $\bar{y}$ is the mean of y. Once a rate was obtained for each measurement, the temperature coefficient (Q_{10}) was calculated from the slope of the $\log_2$(rate) vs. temperature plot (Fig. 2.3D). This procedure is outlined in the following text.

The definition of the temperature coefficient (Q_{10}) is on the Arrhenius equation [92], which states that the rate of a process (k) is exponentially related to the temperature of that process:

$$k = Ae^{-E_a/RT} \tag{2.3}$$

where k represents the rate of the process, A is a constant factor that represents the frequency of molecular collision, E_a is the activation energy, R is the gas constant, and T is the temperature of the process [93]. The temperature coefficient is empirically defined as the ratio of the rates of two processes [69, 74, 92, 94–96]:

$$Q_{10} = (\frac{k_2}{k_1})^{10/(T_2-T_1)} \tag{2.4}$$

where k_1 and k_2 represent the rates of two processes, and T_1 and T_2 represent the temperature at which these processes occur. This relationship is the slope of a plot of $\log_2$(rate) vs. temperature (Fig. 2.3D). Using linear regression, the slope of the $\log_2$(rate) vs. temperature plots was calculated for each measurement and used to calculate the Q_{10} value from

$$Q_{10} = 2^{10 \cdot m} \tag{2.5}$$

where m is defined as

$$m = \frac{\log_2 k_2 - \log_2 k_1}{T_2 - T_1}. \tag{2.6}$$

Metabolites whose R^2 were less than 0.50 were excluded in the analysis. This cutoff was determined based on the distribution of R^2 values (Fig. 2.7); our goal was to maximize the amount of data captured while simultaneously minimizing the inclusion of noisy or poorly fit data. We used the same procedure for calculating the reaction Q_{10} coefficients. We only included reactions that carried flux in at least three of the temperatures; transport reactions and reactions whose R^2 was less than 0.50 were excluded in the analysis. This cutoff was determined based on the distribution of R^2 values (Fig. 2.7); our goal was to maximize the amount of data captured while simultaneously minimizing the inclusion of noisy or poorly fit data. For reactions, extended this cutoff to be based on the p value of the F statistic for the linear regression fit of the $\log_2$(rate) vs. temperature plot; only those reactions whose p value was less than 0.05 were included. The

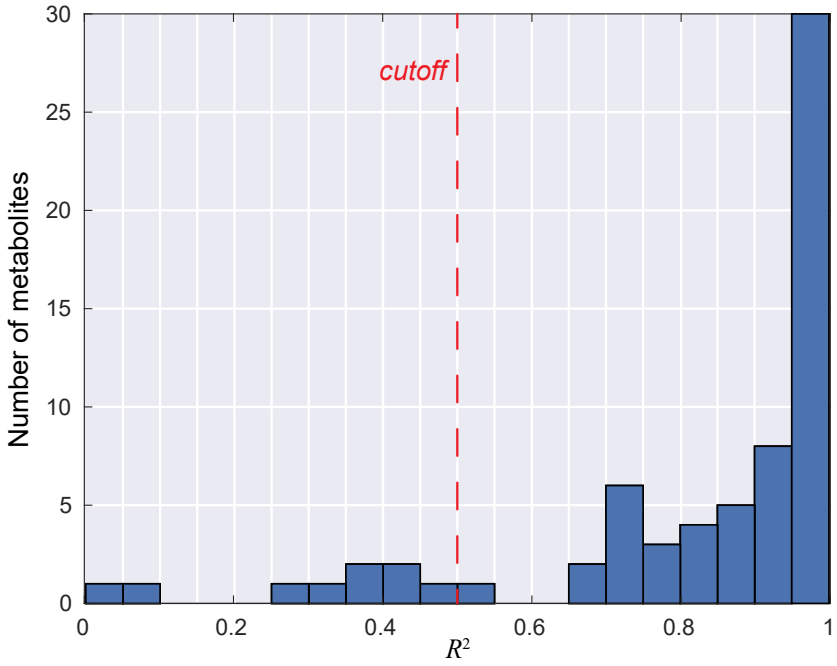

Figure 2.7: Metabolite R^2 distribution. Histogram of the the R^2 value for metabolites. The cutoff (at $R^2 = 0.50$) was used to exclude data that was poorly fit using linear regression.

R^2 for the reactions whose p values were less than 0.05 was greater than 0.80 for all reactions (i.e., no reactions were excluded based on the R^2 value from its linearly regressed fit).

Flux modeling

We used a modified version of the erythrocyte metabolic reconstruction iAB-RBC-283 [40], which was previously used for building personalized kinetic models [37]. We integrated the metabolite concentrations into this model in order to predict the flux state of the network at each temperature using unsteady-state flux balance analysis [52]. 2,3-DPG has previously been determined to be one of the more important metabolites in RBC physiology but was not measured here. In order to obtain the most accurate model of the entire network, we used existing metabolomics data [41] to predict the concentration profile of 2,3-DPG using the profiles of the

eight biomarkers as input according to the workflow described in Yurkovich et al. [82].

Flux coupling

We used F2C2 [97] to calculate the flux coupling of the metabolic networks. When the metabolomics data is integrated into the metabolic model, the structure of the network is altered through the addition of source and sink reactions [52]. In order to construct random networks, we needed to ensure that the models were feasible and mimicked the models created using the measured data. Therefore, we randomly chose nodes in the base RBC model that were "measured" as either accumulation (source) or depletion (sink); we used the measured data to determine (1) the number of randomly "measured" nodes, (2) the distribution of intracellular vs. extracellular nodes, (3) the distribution of sources vs. sinks, and (4) the bounds for the added sources and sinks.

The permutation test compared the difference in the flux coupling characteristic between the base network and the networks at the measured temperatures with the flux coupling characteristic between the base model and 1,000 randomly-generated networks. The p value was determined to be the fraction of random networks whose difference in the flux coupling characteristic was less than or equal to that of the measured temperature networks.

2.2 Conclusions

These results open the door for another interesting possibility: can the temperature dependences calculated here be used to run high throughput screens of additional perturbation experiments at higher temperatures? Our results suggest that an RBC unit stored at a temperature of only 13°C would only need to be stored for approximately 14 days to observe the

equivalent 42 day storage behaviors observed at 4°C. Since the global network-level changes were consistent at higher temperatures, any screens in which the three-phase decay pattern was disrupted could then be investigated in detail under the proper conditions. It is important to note, however, that there is still further evaluation required; the temperature-driven effects on ion homeostasis—which activates ion-dependent cascades (e.g., calcium-induced eryptosis, a phenomenon known to occur during prolonged storage [98, 99])—would not be taken into account by such measurements. If these considerations were properly addressed, a shorter experimental time would not only allow for high throughput screens, but it also yields the very practical consequence of reducing experimental costs. Experiments could be further accelerated if we could find biomarkers that characterize the three-phase decay without the need for full and expensive metabolomic data generation.

Acknowledgements

The following individuals contributed to this work: BO Palsson conceived the study; G Paglia, O Rolfsson, ÓE Sigurjónsson, and K Wichuk generated the data; S Brynjolfsson, S Palssson, and S Gudmundsson operationalized the study; JT Yurkovich, DC Zielinski, A Bordbar, and L Yang analyzed the results; JT Yurkovich, DC Zielinski, L Yang, JT Broddrick, and BO Palsson wrote the manuscript. The authors would like to thank Ke Chen, Bin Du, and Nathan Mih for their assistance in performing calculations for the thermodynamic properties and thermostability of enzymes. This work was supported by the European Research Council (ERC-232816), the US Department of Energy (DE-SC0008701), the National Heart, Lung, and Blood Institute (NHLBI-R43HL123074), and by the Novo Nordisk Foundation through the Center for Biosustainability at the Technical University of Denmark (NNF10CC1016517).

Chapter 2 in part is a reprint of material published in: **JT Yurkovich**, DC Zielinski, L Yang, G Paglia, O Rolfsson, ÓE Sigurjónsson, JT Broddrick, A Bordbar, K Wichuk, S Brynjolfsson, S Palsson, S Gudmundsson, and BO Palsson. 2017. "Quantitative time-course metabolomics in human red blood cells reveal the temperature dependence of human metabolic networks." *Journal of Biological Chemistry*, 292(48):19556-19564. The dissertation author was the primary author.

Chapter 3

Statistical Modeling of the RBC Metabolome

Deep-coverage metabolomic profiling has revealed a well-defined development of metabolic decay in human RBCs under cold storage conditions. A set of extracellular biomarkers has been recently identified that reliably defines the qualitative state of the metabolic network throughout this metabolic decay process. Here, we extend the utility of these biomarkers by using them to quantitatively predict the concentrations of other metabolites in the red blood cell. We are able to accurately predict the concentration profile of 84 of the 91 (92%) measured metabolites ($p < 0.05$) in RBC metabolism using only measurements of these five biomarkers. The median of prediction errors (symmetric mean absolute percent error) across all metabolites was 13%. The ability to predict numerous metabolite concentrations from a simple set of biomarkers offers the potential for the development of a powerful workflow that could be used to evaluate the metabolic state of a biological system using a minimal set of measurements.

3.1 Using biomarkers to predict systemic concentrations

While deep-coverage -omics data sets are allowing for more complete characterization of biological systems, there has been a concerted effort to identify a subset of measurements that are representative of qualitative network-level behavior. For some systems—like the human RBC—such biomarkers have already been identified. Using the concentration profiles of these biomarkers as input to a statistical model, we predict quantitative concentration profiles of other metabolites in the RBC network. These results demonstrate that if good biomarkers are available for a biological system, it is possible to use these measurements to gain insight into the quantitative state of the rest of the network.

The data generated from deep coverage -omics tools are becoming broadly available and thus their use is becoming more common [24, 100]. With this data, researchers have begun to identify metabolomics biomarkers that can be used to describe systemic behavior with only a few inexpensive and reliable measurements [62, 101–104]. In transfusion medicine, deep coverage metabolomics data sets for human RBCs in cold storage are rapidly accumulating [25] and have been used to characterize the state of the RBC metabolic network during storage [18, 41, 42, 61, 105].

Big data analysis of RBC metabolomics data has yielded a well-defined three-phase pattern of metabolic storage lesion that has fundamental consequences for blood storage [41, 42]. Recently, eight extracellular metabolic biomarkers have been identified that reliably define this three-phase decay process observed in RBCs [62]. These biomarkers (adenine, glucose, hypoxanthine, lactate, malate, nicotinamide, 5-oxoproline, and xanthine) recapitulate the qualitative trend of the entire metabolome. However, it has yet to be determined whether these biomarkers can be used to predict quantitative network behavior.

In this study, we determine that five of the eight biomarkers (glucose, hypoxanthine, lactate, malate, and xanthine) are not only excellent qualitative predictors, but also accurate quantitative predictors of metabolic concentrations in the rest of the metabolic network. Using a simple computational formulation [106] prevalent in a variety of fields [107–110], we extend the utility of these biomarkers by using them to quantitatively predict the concentration profiles of 91 other metabolites in the network. This added use of validated biomarkers offers the potential for a powerful workflow that utilizes five biomarkers to evaluate the state of RBC metabolism.

3.1.1 Results

For this study, we used the metabolomics data set from Bordbar et al. [41] that measured 96 intracellular and extracellular metabolites in human red blood cells under storage conditions. The data set measured 14 time points over a 45 day time period for 20 biological replicates. For the purposes of modeling, we randomly divided these 20 replicates into equal sized training and testing sets of 10 samples. We observed a high amount of variability in the extracellular glucose measurement at Day 31, a behavior which was not observed in the intracellular glucose measurement but was seen in other extracellular measurements at Day 31. In order to avoid bias arising from the inclusion of potentially erroneous data, we excluded the measurements from Day 31, resulting in 13 total time points spanning 45 days of storage.

We trained multiple polynomial models of varying complexity on the concentration profiles of the biomarkers and the concentration profile of the target metabolite (Fig. 3.1). The best performing model was a simple, linear Output-Error model [106]. Variation between blood bags is a known challenge, as both donor and technical factors contribute to sample heterogeneity [24]. Due to this variation, we noted that simply because these eight biomarkers are good

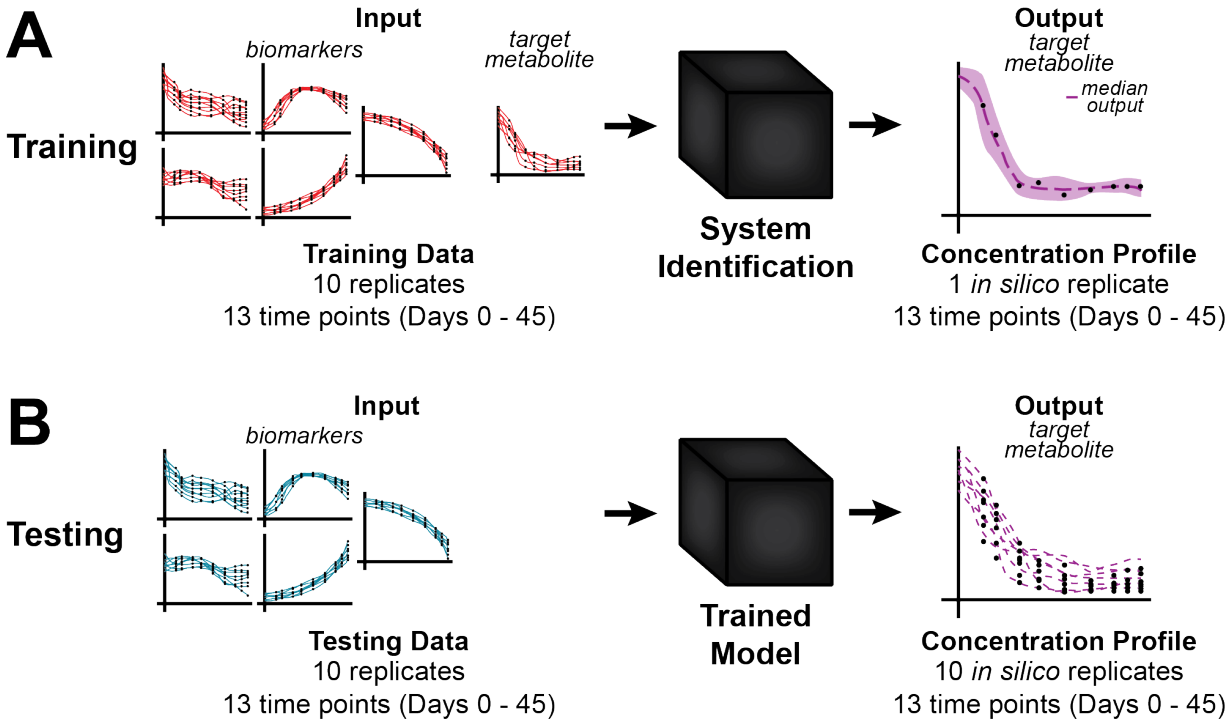

Figure 3.1: Prediction workflow. (**a**) The model is trained on the measured concentration profiles of the five biomarkers (glucose, hypoxanthine, lactate, malate, and xanthine) and the target metabolite. (**b**) The resulting ensemble of models (one for each replicate) can then be used to predict the concentration profile of the target metabolite given only the measured concentration profiles of the five biomarkers.

qualitative predictors of systemic behavior does not imply that they are also good quantitative predictors. We therefore performed a feature selection and cross validation within the eight biomarkers, determining that adenine, nicotinamide, and 5-oxoproline were not able to quantitatively predict systemic behavior as well as the other five biomarkers (see Methods). Thus, glucose, hypoxanthine, lactate, malate, and xanthine were used for the remaining analysis.

In order to generate a prediction for each metabolite, we trained the model using the five biomarkers and a measured profile for the target metabolite as input (Fig. 3.1). We used an ensemble modeling approach [111] to reduce bias arising from using either individual replicates or averaging replicates to train a single model. With 10 training replicates, this approach allowed

us to generate an ensemble of trained models that inherently includes the biological variation of the training data. We then used this trained ensemble computational model to predict a consensus concentration profile of a target metabolite, this time only using the biomarkers as input (Fig. 3.1).

We tested the model's capabilities by comparing the predicted profiles of the remaining 91 measured metabolites to their measured profiles (Fig. 3.2). We calculated the symmetric mean absolute percentage error (SMAPE) for each predicted concentration profile, resulting in a median error of 0.1340 ± 0.1505. To further validate our model, we compared against 10,000 profiles generated using a naive random walk for each metabolite. The naive random walk model assumes that metabolite concentration changes over time are independent of each other and are normally distributed. The random walk is a widely used benchmark for dynamic forecasting models [112]. When a significant number ($\geq$500/10,000, i.e., $\geq$5%) of random walks outperform a trained model for a metabolite, we conclude that the dynamics of that metabolite are indiscernible from noise for the data given (see Methods for details on the random walk comparison). Despite the complexity of RBC metabolism, we found that 84/91 (92%) of RBC metabolites were predicted more accurately than random walks using five biomarkers as input ($p < 0.05$).

In an effort to lend biological intuition to this surprising result, we viewed these results in the context of the complete RBC metabolic network (Fig. 3.2). The map highlights several points. First, the five biomarkers are largely distributed across key subsystems. Surprisingly, two biomarkers are adjacent in the network: xanthine and hypoxanthine. From inspection of the map, it becomes more intuitive that to unambiguously predict IMP levels (Fig. 3.2), both biomarkers need to be quantitatively measured.

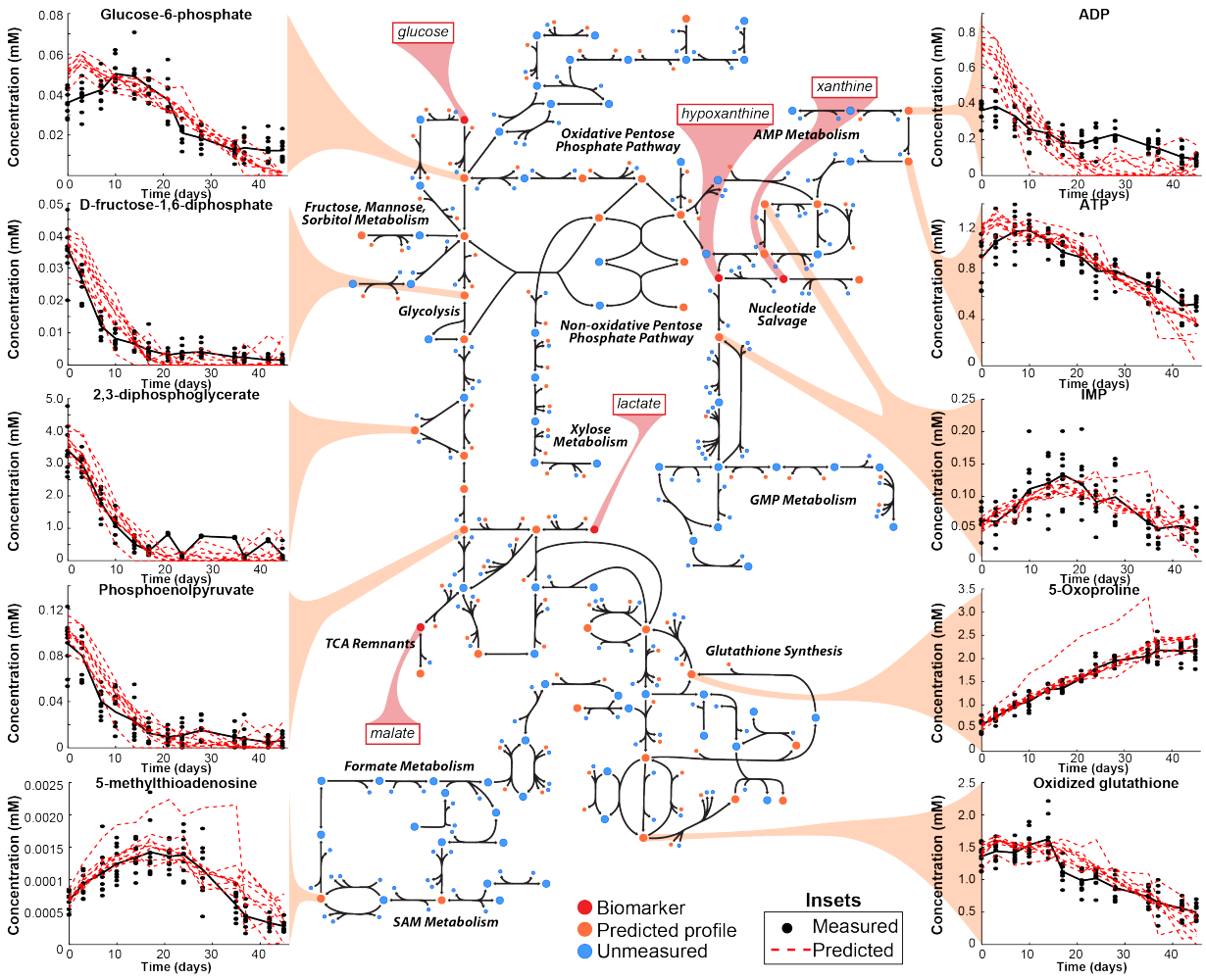

Figure 3.2: Predicted concentration profiles. Using the five biomarkers (highlighted in red), the concentration profiles for the remaining 91 measured metabolites were predicted (inset profile metabolites are highlighted in yellow).

3.1.2 Discussion

RBCs in storage undergo a series of morphological changes (commonly referred to as "storage lesion") that become more pronounced throughout the storage process [24, 78, 113]. Recent studies have shown that blood transfused after being stored for longer than five weeks is associated with post-transfusion complications [84, 85], indicating the serious clinical implications of metabolic decay in transfused blood. With the recent identification of eight extracellular

biomarkers that are able to define this decay, the field of transfusion medicine now has an opportunity to define the metabolic state of stored RBCs with just a few measurements. Thus, there is a need for predictive modeling methods that can extend the applicability of these biomarkers to provide deeper understanding of the metabolic state of RBCs collected and stored under blood banking conditions using current and future technologies (e.g., improved bags or storage solutions, pathogen reduction technologies).

In this study, we have developed a statistical model that uses these biomarkers to predict the time series concentration profiles of other metabolites in the RBC metabolic network. This powerful tool was rigorously validated to avoid overfitting through model (complexity) and feature selection, and comparing against a standard forecasting baseline model (i.e., naive random walk). As with any data modeling approach, the performance of a model is dependent upon the quality of the input data; this is no exception here. We see that certain metabolites (e.g., ADP, inosine) had higher prediction errors, which can be partially attributed to noise in the training data and to low concentrations.

The results presented here have two important implications. First, we have shown that if good biomarkers are available for a given system (like for the human RBC), then they can be used to make quantitative predictions about systemic behavior. Second, this provides the potential for a cost-effective workflow to monitor the metabolic state of a biological system since the only input under new conditions is the concentration profiles of biomarkers. Through the use of modeling and statistical analysis, the measured and predicted concentrations would enable a quantitative understanding of systems-level behavior.

Thus, we have demonstrated the predictive power of biomarkers through the use of a statistical model for RBCs in storage. This data-driven statistical modeling approach performed

remarkably well for the RBC system, even without a detailed kinetic model. These results are encouraging and provide a complementary approach for predicting metabolite dynamics in less characterized organisms. As our validation procedure indicates, a critical mass of high-quality data is required to extract meaningful signals from noise. Our workflow provides a valuable assessment on whether this critical mass has been satisfied; the results here indicate that as few as 20 biological replicates are sufficient to provide a training set capable of achieving >90% accuracy. Follow up studies should address the question of how many measurements need to be made during storage in order to provide a reliable assessment of the RBC metabolome during storage, as this question has direct clinical implications.

As biomarkers are identified for new systems, there will be a need to analyze -omics in an attempt to efficiently characterize complex biological systems using just these few informative measurements. Our workflow addresses this need by incorporating such biomarkers with a statistical model, offering broad utility in both the laboratory and the clinic.

3.1.3 Methods

All computations were performed in Matlab R2016b (Mathworks, Natick, MA).

System Identification

An Output Error (OE) model [106] predicts system dynamics from past values, measured inputs, and unmeasured disturbances as follows:

$$y(t) = \sum_{i=1}^{n} \frac{B_i(q)}{F_i(q)} u_i(t - nk_i) + e(t) \tag{3.1}$$

where $y(t)$ is the output at time t, u_i is an input (i.e., metabolite i), e is the unmeasured disturbance (i.e., system noise), and $B(q)$ and $F(q)$ are polynomials expressed in the time-shift

operator q as follows:

$$B(q) = b_1 + b_2 q^{-1} + ... + b_{nb} q^{-nb+1} \tag{3.2}$$

$$F(q) = 1 + f_1 q^{-1} + ... + f_{nf} q^{-nf}. \tag{3.3}$$

For this system, $n = 5$ (i.e., the five biomarkers), $nb = 1$, $nf = 0$, and there was no input delay ($nk = 0$). The B and F polynomials are estimated during the system identification step using least squares regression to minimize the difference between the measured signal and the predicted output.

This OE model performed better than more complex OE models having higher nb and more complex polynomial models. It also performed better than simpler linear regression—the OE model thus represents an optimal degree of complexity.

Model Evaluation

In order to evaluate the accuracy of the predicted concentration profiles for the various metabolites, we calculated the symmetric mean absolute percentage error (SMAPE), given by:

$$\text{SMAPE} = \frac{1}{n} \sum_{t=1}^{n} \frac{|y_t - \hat{y}_t|}{y_t + \hat{y}_t} \tag{3.4}$$

where n is the number of time points, y is the measured concentration profile, and $\hat{y}$ is the predicted concentration profile.

Quantitative Biomarker Selection

We trained the OE model using a recently published metabolomics data set of RBCs under storage conditions at 4°C with 20 biological replicates from Bordbar et al. [41]. In order to predict the concentration of target metabolites, we used the eight extracellular biomarkers [62] as input

since they are highly representative of the qualitative behavior of the rest of the system. In order to determine if these biomarkers are also good quantitative predictors, we performed a 10-fold cross validation on the set of 10 samples used for training the model to verify the generalization performance of the trained model. We ran our cross validation on all 56 combinations of five biomarkers (i.e., 8 choose 5); the five selected biomarkers had a mean SMAPE of 10.33%, which was within 1% of the top performing set of five biomarkers. Thus, we used glucose, hypoxanthine, lactate, malate, and xanthine as the final set of biomarkers input to the OE model.

Training an Ensemble of Models

We trained an ensemble of OE models using the five biomarker profiles and each of the 91 measured metabolite profiles. Thus, we trained 91 ensemble models (one ensemble for each metabolite). Each ensemble model consisted of 10 OE models, each trained on a biological replicate. We used Bags 1-10 as this training set. We combined the outputs of these 10 OE models into a single prediction for each metabolite by computing the median of the 10 predictions at each time point. This ensemble modeling approach captures the biological variability inherent among the samples used for training.

Predictions on Testing Data

We used Bags 11-20 as the testing data set. In order to assess the variability between the training and testing data, we performed a two-sample t-test at each time point for each metabolite. This showed that approximately 24% of the data rejected the null hypothesis (FDR-adjusted $p < 0.05$) that the two data sets came from the same distribution and also showed greater than a 20% difference in the mean concentrations at a given time point. For each test

replicate, the five biomarkers were input to the trained ensemble model.

Comparison to Naive Random Walk

In addition to the prediction error, as computed by SMAPE, we also evaluated our model by comparing its performance against a benchmark model. We chose as a benchmark the random walk model, which assumes that metabolite concentration changes over time are independent of each other and are normally distributed with zero mean. The random walk model is commonly used to benchmark dynamic forecasting models [112]. To ensure that the random walk was representative of the metabolite concentration changes, we estimated the standard deviation of random changes from all 10 testing replicates across all time points for each metabolite. We further ensured that the random walk was an appropriate benchmark by initializing with a realistic concentration. To do so, we randomly chose from the pool of the 10 measured starting points of the testing replicates for each metabolite.

We generated 10,000 of these random walk profiles for each metabolite. In order to compare these to our model predictions, our null hypothesis was that our trained model performed no better than the random profiles. We calculated the SMAPE for each of the random profiles and compared to the SMAPE for the predicted profiles; the given p value is the number of random profiles which had a lower SMAPE than the average of the predicted profiles for that metabolite.

3.2 Forecasting future concentration profiles

The primary limitation of this first study was that the models required the input of the biomarkers at a given time point in order to predict the concentration of another metabolite in the system. Here, we extend our previous study in two ways. First, we use the structure of

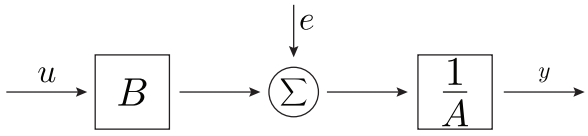

Figure 3.3: Linear black-box formulation for ARX model.

the network to inform which biomarkers should be used as input for a given target metabolite. Second, we reduce the amount of data needed to make a prediction by only using a subset of the timecourse as input and forecasting future values of a target metabolite. We show that 57 of the 70 metabolites measured in the RBC metabolic network (81%) can be forecasted after 8 days of storage (5 time points) with a median global error of 18.36%. The ability to forecast the concentration profiles of metabolites after only a few days of storage makes these methods immediately applicable in a clinical setting.

3.2.1 Methods

All data used for this study came from the metabolomics data set from Bordbar et al. [41] that measured intracellular and extracellular metabolites in human red blood cells under storage conditions. The data consists of 14 time points non-uniformly measured from days 0-45 days of storage. We excluded the day 31 measurement for all metabolites as previously reported [82]. In order to get better resolution on the measured timeseries, we used a spline interpolation function to calculate an even sampling of the measured data with sampling time $T_s = 2$ days. Thus, our data now span days 0-45 with measurements every two days. We randomly selected 10 of the samples (blood bags 3, 4, 7, 9, 11, 12, 13, 15, 17, 19) to use as training data and the remaining 10 samples to use as testing data for validation for all results reported in this study unless otherwise noted.

System Identification

An Auto-Regressive with eXogenous inputs (ARX) model [106] predicts system dynamics from past values, measured inputs, and unmeasured disturbances (see Fig. 3.3) given by

$$A(q)y(t-n_k) = B(q)u(t-n_k) + e(t-n_k) \tag{3.5}$$

where $y(t)$ is the output at time t, $u(t)$ is the input at time t, e is the unmeasured disturbance (i.e., system noise), and $A(q)$ and $B(q)$ are polynomials expressed in the time-shift operator q as follows:

$$A(q) = a_1 + a_2 q^{-1} + ... + a_{n_a} q^{-n_a} \tag{3.6}$$

$$B(q) = b_1 + b_2 q^{-1} + ... + b_{n_b} q^{-n_b+1}. \tag{3.7}$$

The orders of the polynomials $A(q)$ and $B(q)$ were determined individually for each target metabolite. To do this, we selected the model structure that minimized the Akaike information criterion (AIC) given by

$$\text{AIC} = \log(V) + \frac{2d}{N} \tag{3.8}$$

where V is the loss function, d is the number of model parameters, and N is the number of data points used for the estimation. The loss function V is the estimated error between the model output and the measured response. All combinations of model orders between 0 and 4 (output signal) and between 0 and 3 (input signals) were tested.Higher orders were not tested to avoid overfitting.

We included no time delay (i.e., $n_k = 0$) on the system because the sampling time $T_s = 2$ days, meaning that any time delay would have resulted in a multiple day shift which is not physiologically accurate for metabolite concentrations. The $A(q)$ and $B(q)$ polynomials are

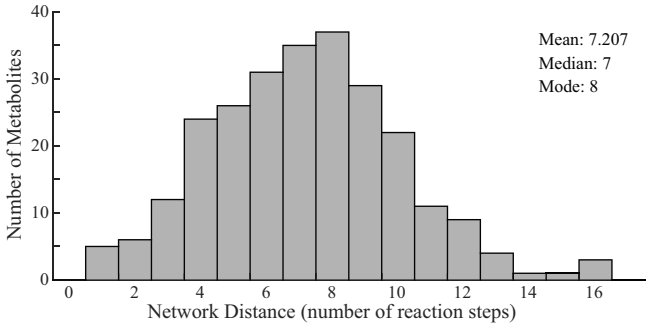

Figure 3.4: Distribution of biomarker-metabolite distances. The distance between two nodes in the network is defined as the number of reaction steps (i.e., edges) that separate them.

estimated during the system identification step using least squares regression to minimize the difference between the measured signal and the forecasted output at each time point.

For each metabolite, one model was trained for each of the 10 samples. The resulting models were merged into a single model by using the covariance matrices to determine the merged model parameters as a statistically weighted mean of the individual model parameters.

Calculating metabolite distances

In our previous work [82], we used five of the eight biomarkers identified by Paglia et al. [62]—extracellular glucose, hypoxanthine, lactate, malate, and xanthine—as input to train a model for all target metabolites. While this approximation produced positive results, it assumes that all metabolites are influenced by each of these five biomarkers regardless of whether or not a given metabolite is close to or even actually connected to a given biomarker in the network. Here, we hypothesized that distance within the metabolic network influences a biomarker's ability to inform quantitative predictions and should therefore be considered when constructing a model for each target metabolite.

The metabolic network is defined as an incidence matrix in which the stoichiometry (i.e., the inputs and outputs) of each reaction is represented in a matrix [114]. In order to calculate

the network distance, we converted the metabolic network into a directed graph that accounts for the directionality of each biochemical reaction. Edge weights were assumed to be uniform. We then calculated the distance (defined as the number of reaction steps separating two nodes in the network) between each of the five biomarkers and all 70 metabolite targets in the network (Fig. 3.4). We used the shortest paths algorithm in the Python NetworkX software [115] to compute these distances. In order to determine what the best distance cutoff is for deciding whether or not to use a biomarker as input to train a model for a given metabolite, we varied the reaction distance cutoff from 3 to 12 and calculated the error of the predicted output of each target metabolite; if a metabolite was further from all five biomarkers than the defined cutoff, then all five biomarkers were used as input. The predicted output for this analysis was a forecasted prediction starting at day 10 (5 time points); we used half of replicates from the training set to train the models and the remaining five samples for evaluation. Because of the differences in the orders of magnitude between various metabolites and the potential for zero terms, we used the symmetric mean absolute percent error (SMAPE) for all error calculations:

$$\text{SMAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|y_t - \hat{y}_t|}{|y_t| + |\hat{y}_t|} \tag{3.9}$$

where n is the number of time points, y is the measured concentration profile, and $\hat{y}$ is the predicted concentration profile at a given time point t. We computed the median of all errors for each metabolite, yielding a "global" error. This error was then used to determine the optimal distance cutoff. As we varied the distance cutoff, the error for various reactions were modulated (as the inputs to the ARX model changed). we determined that a distance of ten reaction steps produced the most accurate models (Fig. 3.5). The combination of biomarkers used and the corresponding model orders for each metabolite are shown in Fig. 3.6.

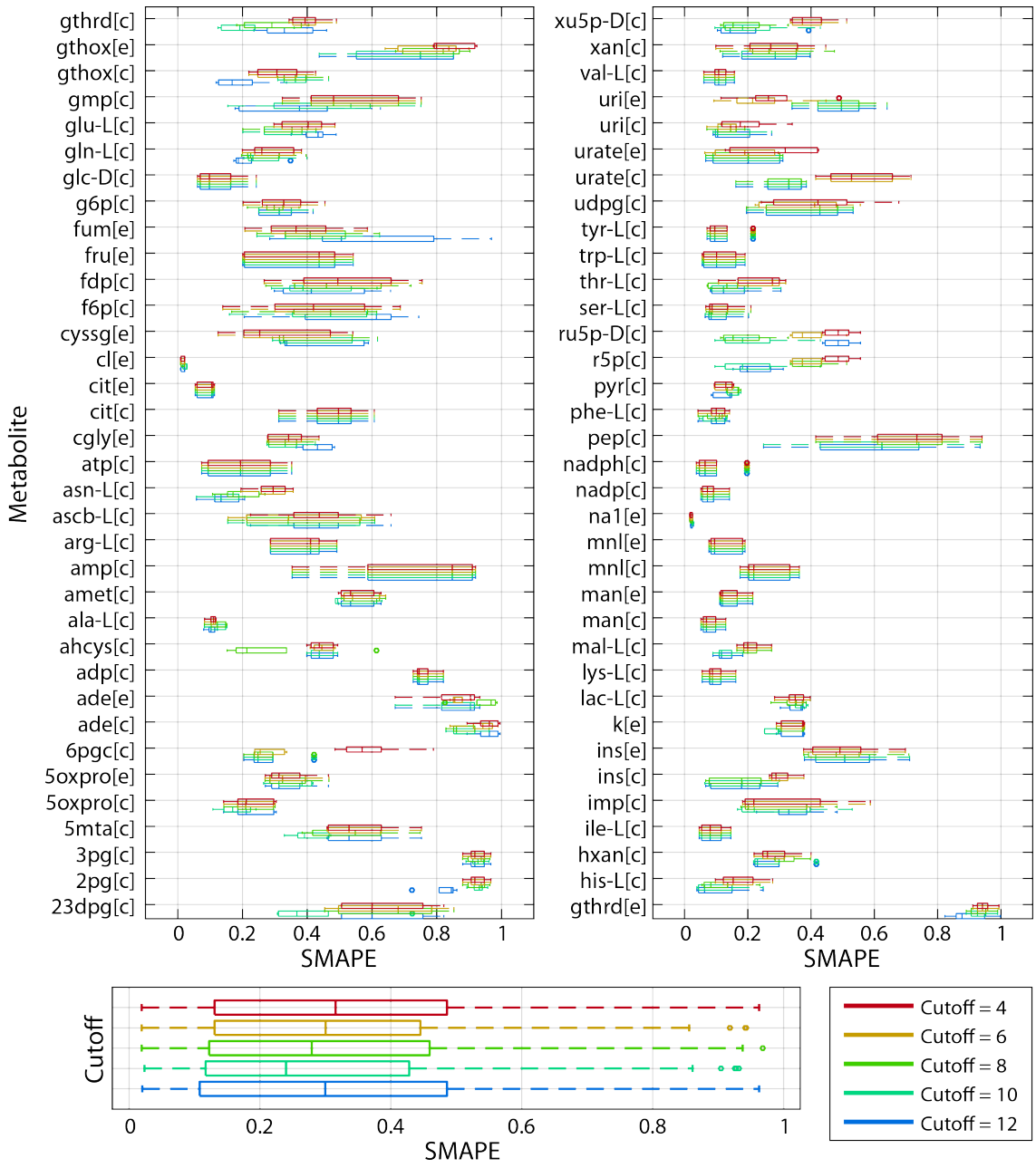

Figure 3.5: Sensitivity analysis for input selection. We varied the distance cutoff from three reaction steps to 10 reaction steps. We defined the distance between two nodes in the network as the number of reaction steps that separate them. We used the distance between each biomarker and each metabolite to determine whether or not to use that biomarker as one of the inputs to the model. Abbreviations are BiGG IDs.

3.2.2 Results

We constructed an ARX model for each metabolite, with the number of inputs and the model orders tailored to each metabolite specifically. This represents a more detailed and

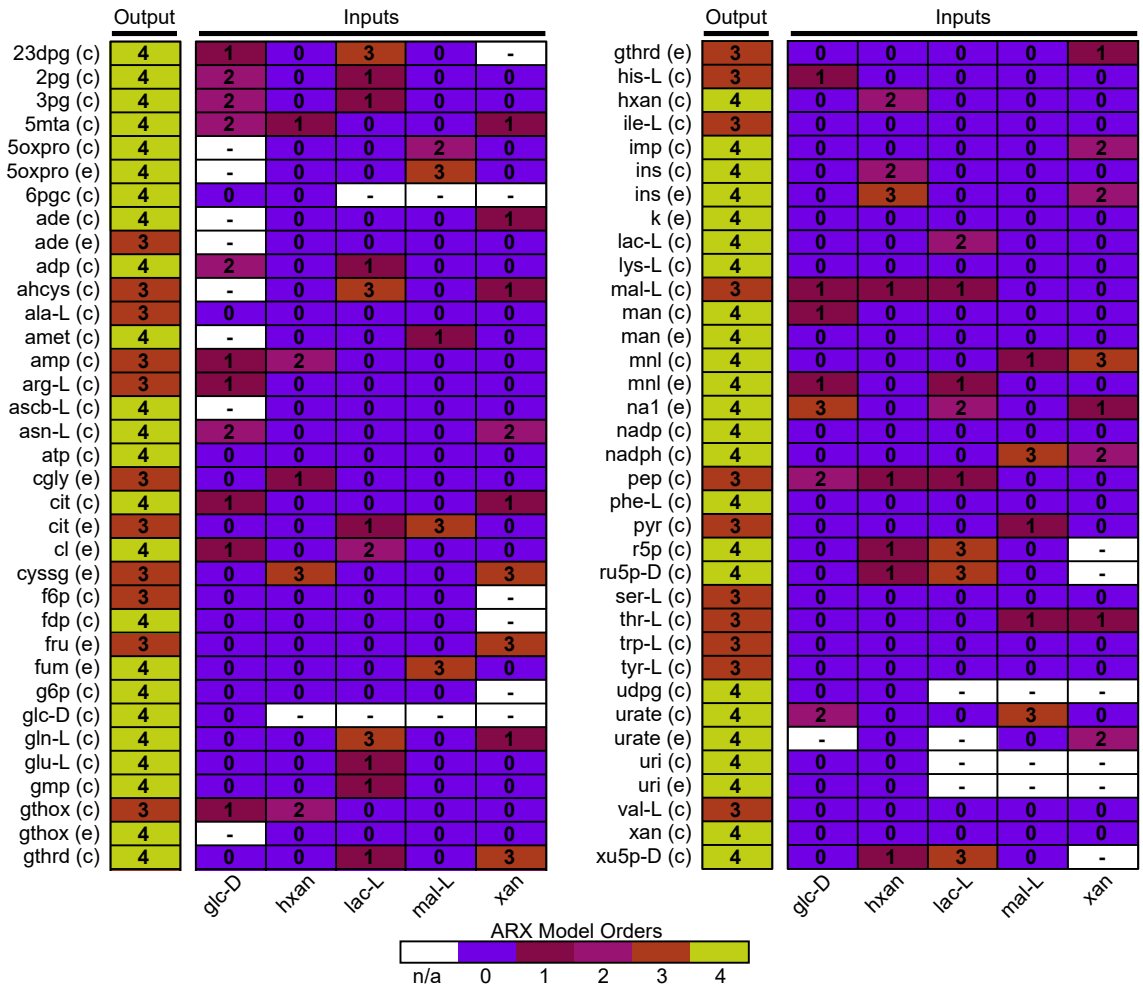

Figure 3.6: Orders for input/output signals of ARX models. White boxes indicate that the corresponding biomarker was not used as input for that metabolite. Intracellular metabolites are denoted by (c) and extracellular metabolites by (e); metabolite abbreviations are BiGG IDs.

physiologically-influenced system identification approach than was previously used [82]. The estimated models were used to forecast future values of the timeseries starting at day 10 (time point 6). In order to perform the forecasting, past values (i.e., measurements from days 0-8) of the target metabolite and the selected biomarkers (Fig. 3.6) were input; additionally, the median of all training replicates for the rest of the timecourse (i.e., days 10-45) was input.

In order to evaluate the results of the model forecasts, we used two metrics: (i) calcula-

Table 3.1: Coefficient of variation for biomarker profiles. Data for all ten training replicates.

Glucose	Hypoxanthine	Lactate	Malate	Xanthine
6.323%	194.18%	7.8027%	13.696%	28.956%

tion of the error (SMAPE) between the measured profile and the forecasted profile, and (*ii*) a comparison of the model forecasted profiles to a benchmark prediction. In this case, we chose to use the historical median (i.e., the median of all training replicates) as the benchmark [112, 116].

The evaluation of the forecasts for all 70 metabolites according to these two metrics are shown in Fig. 3.7. The global median error of the forecasted profiles was 0.1836 ± 0.0817, with a minimum error of 0.0251 (extracellular potassium) and a maximum error of 0.6885 (intracellular L-arginine). It is important to note the accuracy of the model predictions is limited by the variability in the data, both in the input data (Table 3.1) and in the profiles being forecasted. This variability can be attributed to both technical factors involved in making experimental measurements and to donor-to-donor heterogeneity [24].

The historical average (or median, in our case) benchmark is based on the hypothesis that the average metabolite concentration profile across historical samples adequately describes the profile of any new test sample. To be useful, a model should consistently produce better forecasts than this benchmark. This benchmark is commonly used in various fields and can be surprisingly difficult to beat for certain systems [112, 116].

Of the 70 metabolite profiles used, the ARX model forecasts outperformed the historical median for 57 metabolites (81.43%). Other benchmarking methods such as a naive random walk will be compared to the model forecasts in our future work to further evaluate the performance of the ARX models. Note that typically the historical average is used as the benchmark for

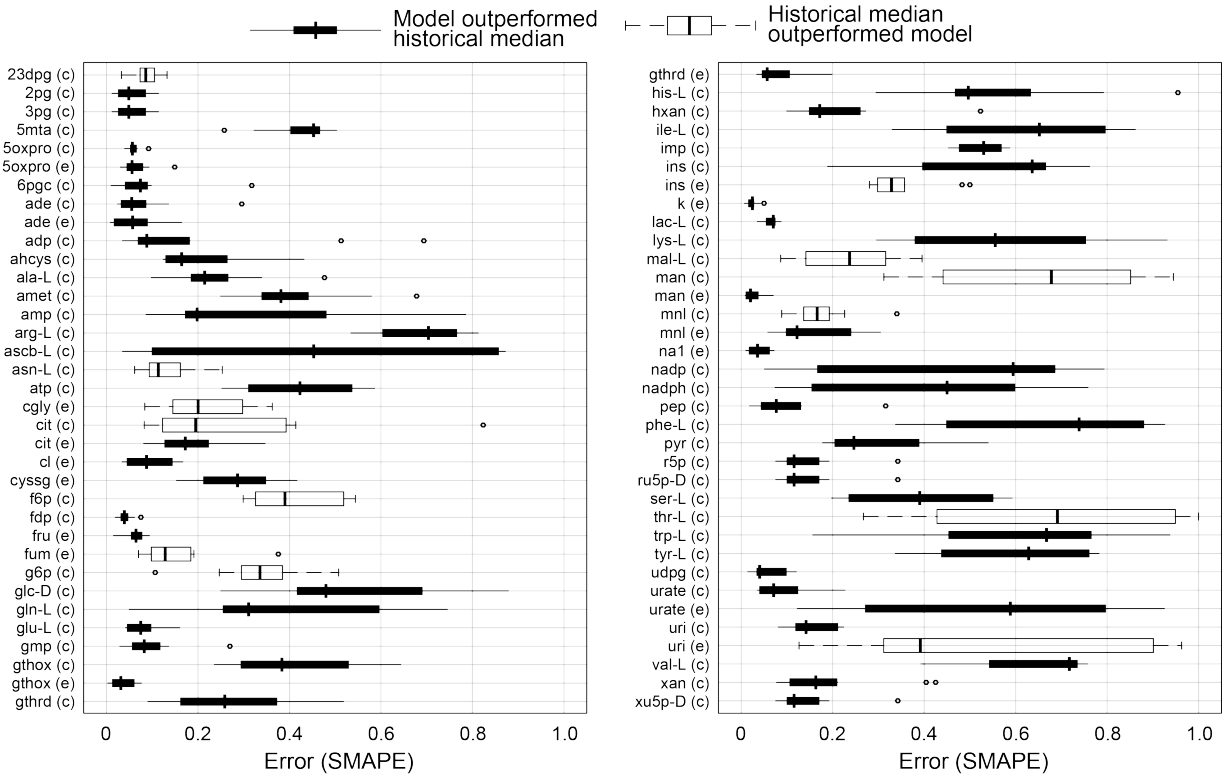

Figure 3.7: Forecasted model predictions. The distribution of errors for each of the ten testing replicates is shown. The filled boxes correspond to metabolites for which the model forecasts had a lower error (SMAPE) than the historical median for at least half of the testing replicates, while the outlined boxes correspond to metabolites for which the historical median had a lower error than the model forecasts. Intracellular metabolites are denoted by (c) and extracellular metabolites by (e); metabolite abbreviations are BiGG IDs.

forecasting [112, 116]; because of the biological variability within the data and the small sample size, we instead used the median of the historical data as the median is less influenced by outliers and therefore a stronger benchmark.

After evaluating the global performance of the models, we can explore the profiles of metabolites that have been previously implicated in the health of stored RBCs (Fig. 3.8), such as 2,3-disphosphoglycerate (2,3-DPG) [78], 5-oxoproline [62], ATP [78], and 5-methylthioadenosine (5-MTA) [61]. For each of the metabolites in Fig. 3.8, we show the measured profile and corresponding forecasted profile of three representative replicates (of the ten total testing replicates).

Of the four metabolites shown, the model forecasts for 5-MTA, 5-oxoproline, and ATP outperformed the historical median for at least half of the ten testing replicates; the model forecasts for 2,3-DPG did not outperform the historical median.

3.2.3 Discussion

Previously, we showed that other metabolites in the RBC network could be accurately predicted from a subset of just five extracellular measurements [82]. Here, we extend those results to demonstrate that future values of these metabolite profiles can also be predicted using measurements of the biomarkers from days 0-8 of storage. This ability to forecast future profiles after only 8 days of storage represents a significant step toward being able to identify potentially unhealthy blood bags.

As presented here, there are still obstacles that must be overcome. Due to the autoregressive nature of the models used, points that were previously forecasted to be zero heavily influence the trajectory of the forecasted profile. This is demonstrated in the forecasted profile of 2,3-disphosphoglycerate (Fig. 3.8), where the concentration starts high then quickly drops and is almost depleted by day 20 of storage. The models are able to correctly predict the shape of this curve, but as soon as a value of zero is predicted, the predictions tend to never recover to the true value.

The error rates achieved with the ARX models (0.1836 ± 0.0817) in this study are comparable but overall higher (0.1340 ± 0.1505) than what was previously achieved using an Output-Error (OE) model to predict the concentration profile of a given target metabolite taking as input the concentrations of all five biomarkers at each time point [82]. This is to be expected considering that the previous results used the biomarker concentrations at each time point, while the results

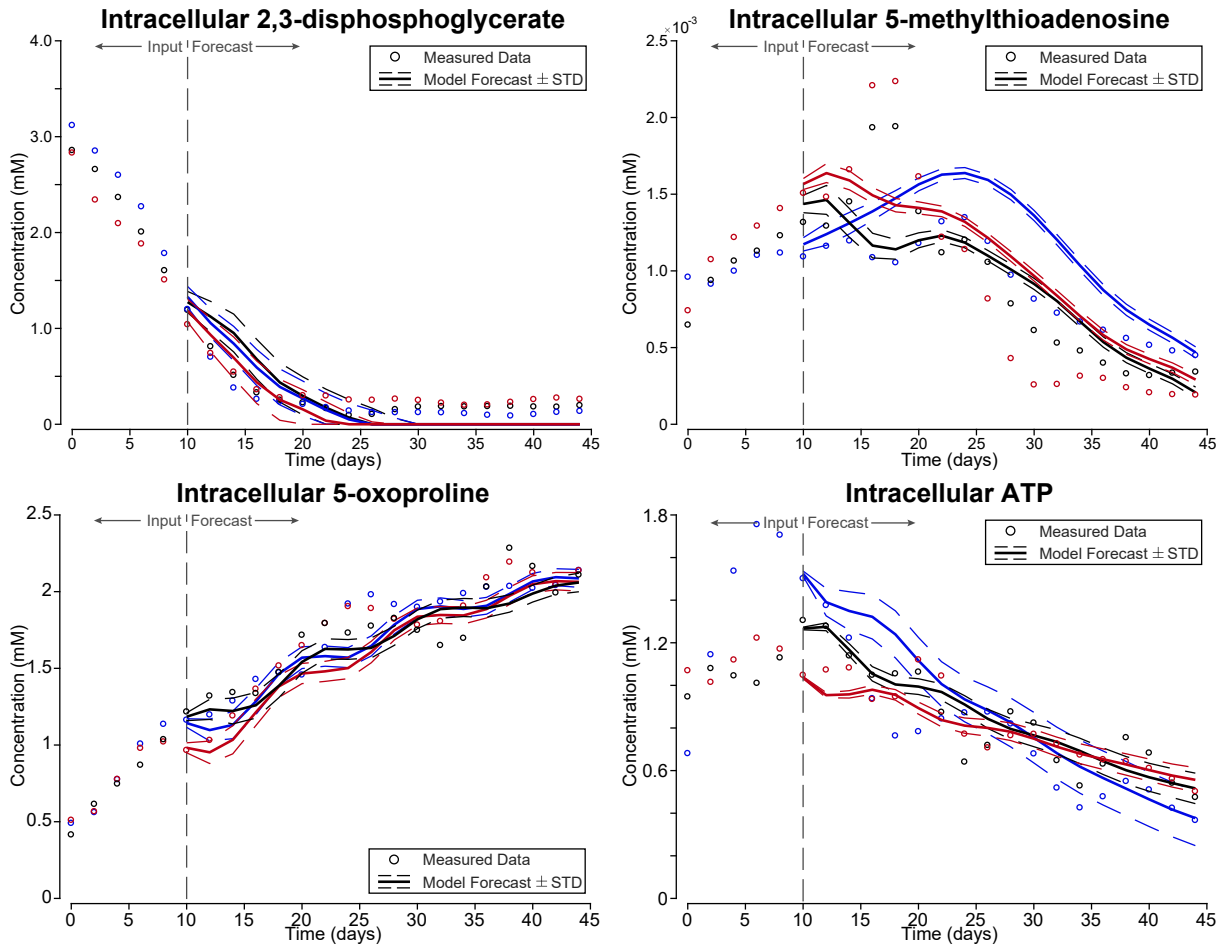

Figure 3.8: Sample forecasted profiles. Example forecasts for 3 representative samples are shown. The four metabolites shown represent important cellular quantities previously implicated in RBC storage health. Different colors represent three different testing replicates. Dashed lines on the profiles represent the predicted standard deviation of the forecast. The vertical dashed line denotes the time at which the forecasting begins (day 10).

here only used these values up to Day 8. One key difference between these two model designs is that our OE models only required the historical concentration profiles of a target metabolite in order to train the model, while the ARX models require the historical concentrations for training but also online measurements of the target metabolite up to day 8 for forecasting. This means that in an online forecasting scenario, experimental measurements must be obtained for more than just the extracellular biomarkers.

Since the primary application of the workflow presented here is a clinical setting (blood banks), having to perform deep-coverage metabolomics profiling in order to forecast future concentration profiles is a considerable limitation. One way to overcome this limitation is to use a model to predict the profile of a given target using only the biomarkers as input [82] for the first 8 days of storage, then inputting that predicted profile into the modeling structure detailed here to forecast future values of the profile. While this proposed framework would add an additional source for error in the forecasting results (due to using predicted instead of measured past data for the auto-regressive calculations), it would eliminate the need to measure more than just the five extracellular biomarkers. We are currently working to incorporate these efforts into the workflow presented here. The results of this work will appear in an upcoming article.

3.3 Conclusions

The ability to characterize the dynamic behavior of cellular metabolism using just a few informative measurements is a challenging problem due to the complexity of metabolic networks. In this study, we have applied standard system identification techniques to forecast future metabolite concentration profiles in human red blood cells using a pre-determined set of biomarkers as input. We have expanded upon our previous work by using the structure of the metabolic network to inform the estimation of metabolite-specific linear black-box models. This research demonstrates promising results directed toward the development of a computational workflow that can assess the state and health of metabolism of RBCs under storage conditions from just a subset of measurements.

Acknowledgements

The following authors contributed to this work: JT Yurkovich, L Yang, and BO Palsson conceived the study; JT Yurkovich performed the analysis; JT Yurkovich, L Yang, and BO Palsson wrote the manuscript. The authors gratefully acknowledge Prof. S Yurkovich for valuable discussions on black-box modeling methods. This research was supported by the US Department of Energy (DE-SC0008701) and the National Institute of General Medical Sciences of the National Institutes of Health (awards U01-GM102098 and R01-GM057089).

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

- **JT Yurkovich***, L Yang*, and BO Palsson. 2017. "Biomarkers are Used to Predict Quantitative Metabolite Concentration Profiles in Human Red Blood Cells." *PLOS Computational Biology*, 13(3):e1005424. The dissertation author was one of the two primary authors.

- **JT Yurkovich**, L Yang, and BO Palsson. 2017. "Utilizing biomarkers to forecast quantitative metabolite concentration profiles in human red blood cells." Proceedings of the IEEE Conference on Control Technology and Applications (CCTA), Kohala Coast, HI (August 27-30, 2017). The dissertation author was the primary author.

Chapter 4

Mechanistic Modeling of the RBC Metabolome

Allosteric regulation has traditionally been described by mathematically-complex allosteric rate laws in the form of ratios of polynomials derived from the application of simplifying kinetic assumptions. As an alternative approach that requires no simplifying assumptions, we explicitly describe the fraction of allosteric enzyme that is in an active form (i.e., its "catalytic potential") by developing a detailed reaction network containing all known ligand binding events to the enzyme. The catalytic potential is the fundamental result of multiple ligand binding that represents a "tug of war" between the various regulators and substrates. This formalism allows for the assessment of interacting allosteric enzymes and development of a network-level understand of regulation. First, we characterize the catalytic potential of three key kinases in RBC glycolysis: hexokinase, phosphofructokinase, and pyruvate kinase. Second, we compute their time-dependent interacting catalytic potentials in response to external disturbances and

found that increased regulation improves the system's ability to return to the homeostatic state. Third, we examine nine existing personalized RBC models to study the variability in glycolytic regulation among individuals and found that the catalytic potential allows for the identification of subtle but important differences. In addition, we develop a graphical representation of the dynamic interactions between the individual kinase catalytic potential that provides an easy way to understand and visualize how a robust homeostatic state is maintained. Together, these results represent a novel approach that enables the study of the network-level effects of interacting allosteric enzymes.

4.1 Modeling temporal dynamics with ODEs

The human RBC has historically been the target of complex kinetic model building of its metabolism due to its relative simplicity and the vast amounts of data and information available on its biochemistry and physiology. RBCs lack cellular compartments (e.g., nuclei, mitochondria) [117] and therefore certain cellular functions, such as transcriptional and translational regulation and the ability to use oxidative phosphorylation to produce energy [118]. As a result, glycolysis is the primary source of ATP generation for the RBC, a pathway that undergoes allosteric regulation at major control points. Glycolytic ATP production is thus largely directly in response to the rate of ATP utilization of known cellular functions, mostly the ATP-driven sodium/potassium transmembrane pump.

Mathematical models have been used to study the dynamics of RBC metabolism since the 1970s [119]. Constraint-based modeling methods have been used to explore the mechanisms underlying cellular metabolism [5, 120, 121], and specialized methods have been developed that allow for the study of system dynamics [52, 122, 123]. Kinetic models represent an approach that

has the potential to truly capture the temporal dynamics at short time scales [124]. The first whole-cell kinetic model of RBC metabolism was published in the late 1980s [125–128], with other such models produced since then [37,129,130]. More recently, so-called "enzyme modules" have been introduced and used to explicitly model detailed binding events of ligands involved in allosteric regulation as an alternative to the traditional use of allosteric rate laws [131,132]. These enzyme modules provide a fine-grained view of the activity and state of a regulated enzyme. Further, they open up many new possibilities in understanding the metabolic regulation that result from complex interactions of regulatory signals, as well as a way to explicitly represent biological data types such as sequence variation and protein structures.

Historically, the primary way to visualize the output from a kinetic model is to plot the time profiles of individual network components (e.g., metabolite concentrations, enzymatic reaction rates). While these quantities are informative, they fail to provide insight into systemic qualities of the network. Dynamic phase portraits have been explored as an alternative [4]. With the formulation of enzyme modules, there is a need to study alternative ways to visualize network dynamics to bring about a new understanding of integrated functions similar to what Bode plots [133] or root loci [134] achieved in the early days of the development of classical control theory. Enzyme modules allow for the explicit computation of the fraction of the regulatory enzyme that is in an active state and generates the reaction flux. The collective action of all the ligands binding to the enzyme—through the computation of the active enzyme fraction—fundamentally represent its regulation.

Here, we use enzyme modules to model hexokinase (HEX), phosphofructokinase (PFK), and pyruvate kinase (PYK), the three major regulatory points in RBC glycolytic energy generation. We compute the catalytic potential of these kinases as a measure of an enzyme's capacity

to influence the rest of the network, using the enzyme modules individually. We analyze the response of each enzyme module to perturbations in ATP utilization, simulating the impact of various physiological stresses on the RBC [135–137]. We then integrate all three enzyme modules into a single model of glycolysis and show that increasing the number of allosteric enzymes improves the disturbance rejection capabilities RBC glycolysis to perturbations in the ATP utilization rate. We further explore how RBCs maintain homeostasis by using nine personalized RBC models to examine the sensitivity of the network to these perturbations. Finally, we elucidate how a graphical representation of the three kinase catalytic potentials leads to an insightful way to visualize the state of RBC glycolysis.

4.1.1 Results

Properties of regulation by phosphofructokinase

PFK has been called the "pacemaker" of glycolysis [138], and it plays a major role in determining glycolytic flux. PFK converts fructose 6-phosphate (F6P) into fructose 1,6-bisphosphate (FDP). Here, we use a reaction mechanism (Fig 4.1A) where PFK binds first to ATP, forming a complex that then binds F6P and then converts the two bound substrates to FDP producing ADP in the process. The four binding sites operate independently, i.e., they do not "cooperate." The catalytic activity of PFK is controlled through allosteric regulation by AMP and ATP (Fig 4.1). AMP and ATP bind to an allosteric site, distal to the catalytic site, inducing a conformational change that modulates the activity of PFK.

For an enzyme allosterically regulated through effector molecules, we can define a quantity that relates the amount of enzyme in the active form to the total amount of enzyme. This

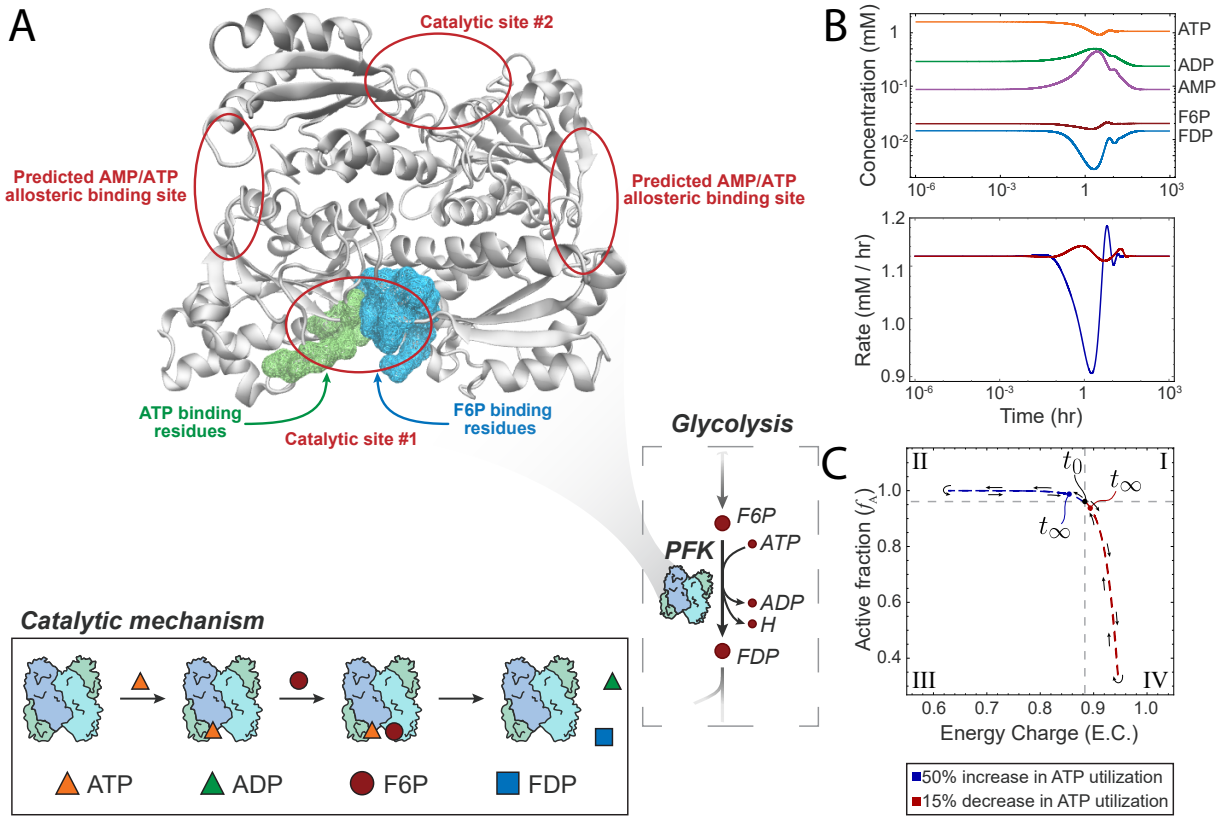

Figure 4.1: PFK mechanism and simulation results. (A) The structure of one of two PFK homomers along with the catalytic mechanism. The predicted allosteric binding sites for AMP/ATP are highlighted. (B) Concentration and reaction rate profiles for PFK regulatory module in glycolysis. The concentration profiles shown are for a 50 % increase in ATP utilization. (C) Phase portrait showing the catalytic potential of PFK. Two perturbations are shown: (1) a 50 % increase in ATP utilization, and (2) a 15 % decrease in ATP utilization. Roman numerals indicate comparisons with the steady-state: (I) more enzyme in active form and higher energy charge; (II) more enzyme in active form and lower energy; (III) more enzyme in inactive form and lower energy charge; and (IV) more enzyme in inactive form and higher energy charge.

catalytically active fraction (f_{A}) is given by

$$f_{\mathrm{A}} = \frac{\sum_{i=0}^{n} R_i + R_{i,A} + R_{i,AS}}{E_{total}} \tag{4.1}$$

where n is the number of enzymatic binding sites, R_i is the unbound enzyme in the active state (i.e., not bound to inhibitors), $R_{i,\mathrm{A}}$ is the enzyme bound to the cofactor, $R_{i,\mathrm{AS}}$ is the enzyme bound to the substrate and cofactor, and E_{total} is the total amount of enzyme. The subscript i

represents the amount of activators bound to allosteric sites; for tetrameric structures like PFK and PYK, i ranges between 0 and 4 [139, 140]. We term this ratio the "catalytic potential" of an enzyme because it provides a representation of an enzyme's effect on the rest of the system and its ability to maintain the homeostatic state.

In order to characterize the system response to perturbations, we modulated the ATP utilization by adjusting the rate of reaction for ATP. We modeled two perturbations that have been previously observed to fall within a physiologically-feasible response: a 50% increase and a 15% decrease in ATP utilization. The increase in ATP utilization is used to model a cell that is undergoing sheer-induced ATP release [136], which is a common phenomenon experienced by RBCs *in vivo*. Under hypoxic conditions, ATP utilization has been shown to increase by over five fold [137]. Therefore, we chose to model a 50% increase in ATP utilization to observe the trends of how the catalytic potential changes in order to maintain a homeostatic state. The decrease in ATP utilization models the reduced release of ATP under cancerous conditions [135].

We modeled these perturbations by modulating the rate of reaction for the hydrolysis of ATP; in order to allow the system flexibility to respond to these perturbations, phosphate was modeled as a variable quantity (see Methods for full details). Increasing this rate decreases the amount of available ATP and ADP, resulting in a decrease in the rate of PFK (Fig 4.1B). Conversely, lowering this value increases the rate of the PFK reaction. For both perturbations, the final rate value eventually returns to the same as the unperturbed system. We were interested in characterizing the enzymatic response to these energetic perturbations. The energetic state of a cell can be measured using the energy charge [141], which relates the amount high energy bonds available in the adenosine phosphate pool. The energy charge is given by:

$$\text{energy charge} = \frac{[\text{ATP}] + \frac{1}{2}[\text{ADP}]}{[\text{ATP}] + [\text{ADP}] + [\text{AMP}]} \tag{4.2}$$

where [AMP], [ADP], and [ATP] represent the concentrations of those respective metabolites. To evaluate how the regulatory state of an enzyme is related to the energy state of the system, we can plot the ratio of active to total enzyme as a function of the energy charge (i.e., the catalytic potential).

We characterized the glycolytic model first with only the PFK enzyme module. Dynamic simulation for the model was performed and then the catalytic potential was plotted as a function of the energy charge (Fig 4.1C). We observed an inverse relationship between the energy charge and the active fraction of the enzyme. An increase in ATP utilization lowers the energy charge and increases the active fraction to increase the glycolytic flux, while a decrease in ATP utilization leads to the opposite result. The inverse relationship between the energy charge and the active fraction—the catalytic potential—is an emergent property of the system. The interpretation of this graphical representation is that PFK senses the energy charge and adjusts the flux through PFK to reverse the change in the energy charge, returning the system to a homeostatic state.

Properties of regulation by hexokinase and pyruvate kinase

Having used an enzyme module to explicitly model and simulate the regulation of PFK, our next step was to expand the representation of allosteric regulation to include two other glycolytic kinases (HEX and PYR). We constructed enzyme modules for both enzymes using mechanisms that allowed the substrate to bind cofactors in any order. 2,3-diphosphoglycerate is involved in the regulation of HEX therefore requiring the inclusion of the Rapoport-Luebering Shunt and hemoglobin in the model [4] (see Methods for full details). We validated each enzyme module individually by performing the same ATP utilization perturbation (i.e., 50% increase and 15% decrease). The catalytic potential observed for the HEX module was in agreement with

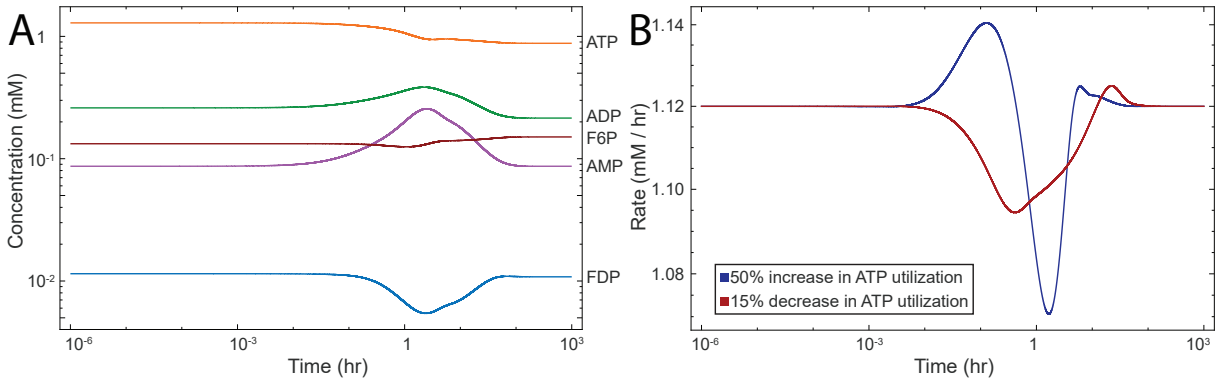

Figure 4.2: Classical representation of PFK simulation for TKRM. (A) Concentration time profiles shown for a 50% increase in ATP utilization. (B) Reaction rate time profiles for the reaction rate of PFK.

previously observed experimental evidence [142]. The PYK module exhibited a direct relationship between f_{A} and energy charge, conflicting with the inverse relationship previously observed *in vitro* [142].

Three kinase regulatory model (TKRM)

Once we had validated each kinase module individually, we built an expanded model of glycolysis that included all three enzyme modules and hemoglobin. We will refer to this model as the "three kinase regulatory model" (TKRM). The TKRM allows us to simulate how the three allosteric enzymes interact in determining the systems responses.

We first calculated the concentration profiles for PFK (Fig 4.2A), observing higher FDP levels than with just the PFK module (Fig 4.1B). This effect is likely due to the inclusion of the PYK module, in which FDP is an allosteric activator. We also examined the reaction rate profile of PFK (Fig 4.2B), which was inverted compared to that of the PFK model only (Fig 4.1B). This inversion arose from the addition of the HEX module, demonstrating the interplay among the various enzymes within a network.

In order to better capture this interplay among enzymes, we constructed phase portraits for the fraction of catalytically active enzyme (f_A) for each pairwise combination of enzyme modules in the TKRM (Fig 4.3A). These phase portraits show that as a greater fraction of PFK entered a more catalytically active state, a greater fraction of HEX become catalytically inactive; a similar behavior was observed for the PFK-PYK pair. This behavior was observed in each of the enzyme modules individually. We observed that HEX and PYK moved in tandem, with both enzymes moving into catalytically active or inactive states together. This behavior is likely due to the fact that these enzymes represent the boundaries of the system and therefore are linked in order to maintain system stability. Finally, we constructed catalytic potential plots for each enzyme module in the TKRM (Fig 4.3B). We observed that HEX and PYK exhibited primarily proportional relationships between energy charge and the fraction of catalytically active enzyme, while we observed an inverse relationship between these two quantities for PFK. This inverse relationship observed for PFK recapitulated the previously reported relationship between catalytically active enzyme fraction and energy charge [143].

Maintaining the homeostatic state: disturbance rejection properties

The inclusion of feedback and other regulatory mechanisms are designed to improve the disturbance rejection capabilities of a system [144]. For biological systems, regulatory mechanisms are expected to enable organisms to maintain a robust homeostatic state in the face of environmental perturbations. Having characterized our individual allosteric enzyme modules and the TKRM, our next goal was (1) to investigate the capacity for each of these models to help maintain the homeostatic state and (2) to examine how understanding the catalytic potentials help elucidate this ability. We simulated a 50% increase in ATP utilization and calculated the

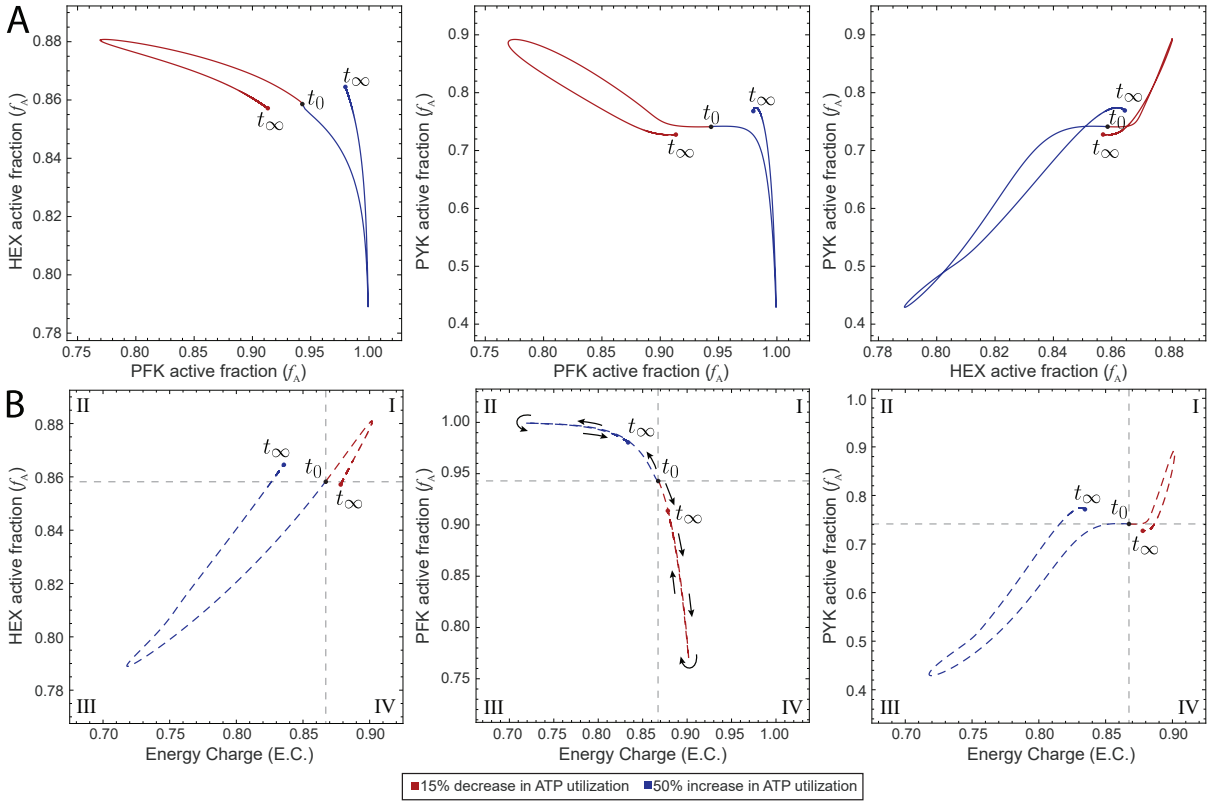

Figure 4.3: Dynamic response to perturbations in ATP utilization. (A) Phase portraits displaying all pairwise relationships between the catalytic potentials of two kinases. (B) Catalytic potential plots for each of the enzyme modules as a function of the energy charge. Roman numerals indicate comparisons with the steady-state: (I) more enzyme in active form and higher energy charge; (II) more enzyme in active form and lower energy; (III) more enzyme in inactive form and lower energy charge; and (IV) more enzyme in inactive form and higher energy charge.

total ATP flux in the network (i.e., total flux through ATP-producing reactions minus total flux through ATP-consuming reactions) for each of the models constructed (Fig 4.4). All systems were able to maintain a stable homeostatic state following the perturbation (Fig 4.4A). We calculated the sum of squared error (SSE) for each model in order to quantify the disturbance rejection capabilities of each model (Fig 4.4A). As expected, the models with little or no regulation performed the worst, while increased regulation generally lowered the SSE. The base glycolytic model with the PYK module performed the worst, while the model containing the PFK and HEX modules with hemoglobin performed the best. The final steady-state values for the energy charge differed

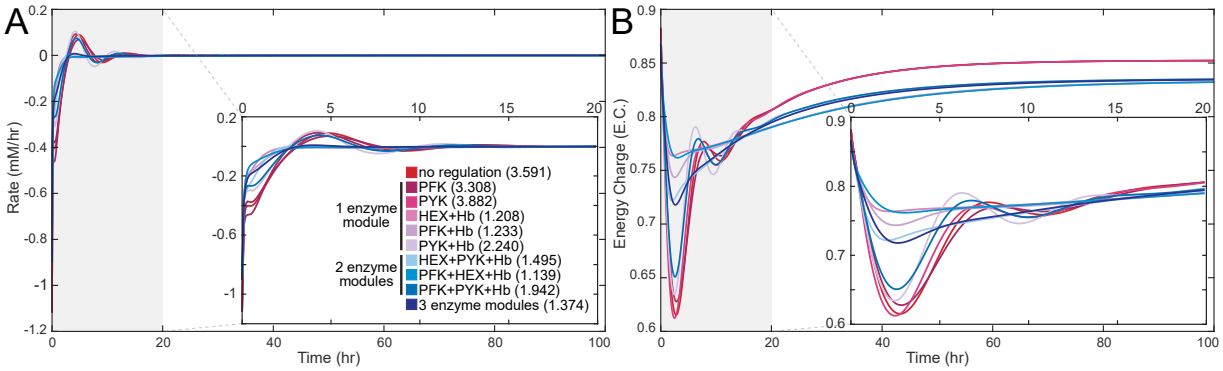

Figure 4.4: Disturbance rejection capabilities for various regulation models. (A) The net rate of ATP usage (i.e., total flux through ATP-producing reactions minus total flux through ATP-consuming reactions) is shown as a function of time; the inset zooms in on the 0 to 20 hour time range. The number in parentheses represents the SSE for each model, quantifying the total deviation of the output from the setpoint. (B) The energy charge is shown as a function of time; the inset zooms in on the 0 to 20 hour time range.

with the inclusion of hemoglobin in the model, although the magnitude of these differences was small (Fig 4.4B).

The baseline RBC glycolytic model used to construct the models here is based on nominal parameter values [4]. Genetic variation in the human population leads to varying RBC metabolic dynamics in different individuals, therefore requiring sensitivity analysis of the model parameters. RBC and plasma metabolite levels have been reported from a series of individuals, enabling the construction of "personalized" RBC models by calculating rate constants for kinetic descriptions of the reaction dynamics [37]. We thus constructed personalized models using glycolytic metabolite concentrations and equilibrium constants for nine individuals from a previous study [37]. Using personalized models provides a sensitivity analysis that examines physiologically-feasible parameter values. We performed our sensitivity analysis using the model that includes an enzyme module for PFK and hemoglobin due to simplicity for numerical simulation.

The general qualitative trend for the catalytic potential plot was similar to the one using literature values (Fig 4.1C and Fig 4.3B), but initial f_A values were significantly lower in the

personalized models (Fig 4.5A,B). In particular, the amount of active PFK for each individual reached a saturation point that was higher than the initial steady-state value in order to compensate for the increase in ATP utilization before returning to a final steady-state value. While we observe that there is little difference among the rate profiles (Fig 4.5C), we observe much greater differences in the catalytic potential plots (Fig 4.5A,C) and energy charge profiles (Fig 4.5D). Notably, the model for Individual #1 exhibited a much different response than the other eight personalized models (Fig 4.5A,B,D). Upon further examination, we determined that this difference stemmed from the fact that the rate constants for the binding of ATP and F6P to PFK were outliers with over 99% confidence according to the Dixon's Q test (see Methods for full details); these were the only rate constants that were deemed to be outliers out of all enzymatic reactions.

4.1.2 Discussion

The ability to mechanistically model cellular metabolism allows for the construction of predictive physiological models. However, the mechanistic results obtained from time-course plots can complicate the interpretation and analysis of systems-wide responses to relevant perturbations. To help provide a better method of elucidating this behavior, we built modularized glycolytic models with enzymes serving as regulators. These models were then validated against existing empirical data to understand the relationship between the catalytically active enzyme fraction and energy charge—the catalytic potential of an enzyme. Visualizing the catalytic potential allowed for the analysis of important systems behaviors. The results presented here have several primary implications.

First, we have studied glycolysis from a perspective in which enzymes are regulators.

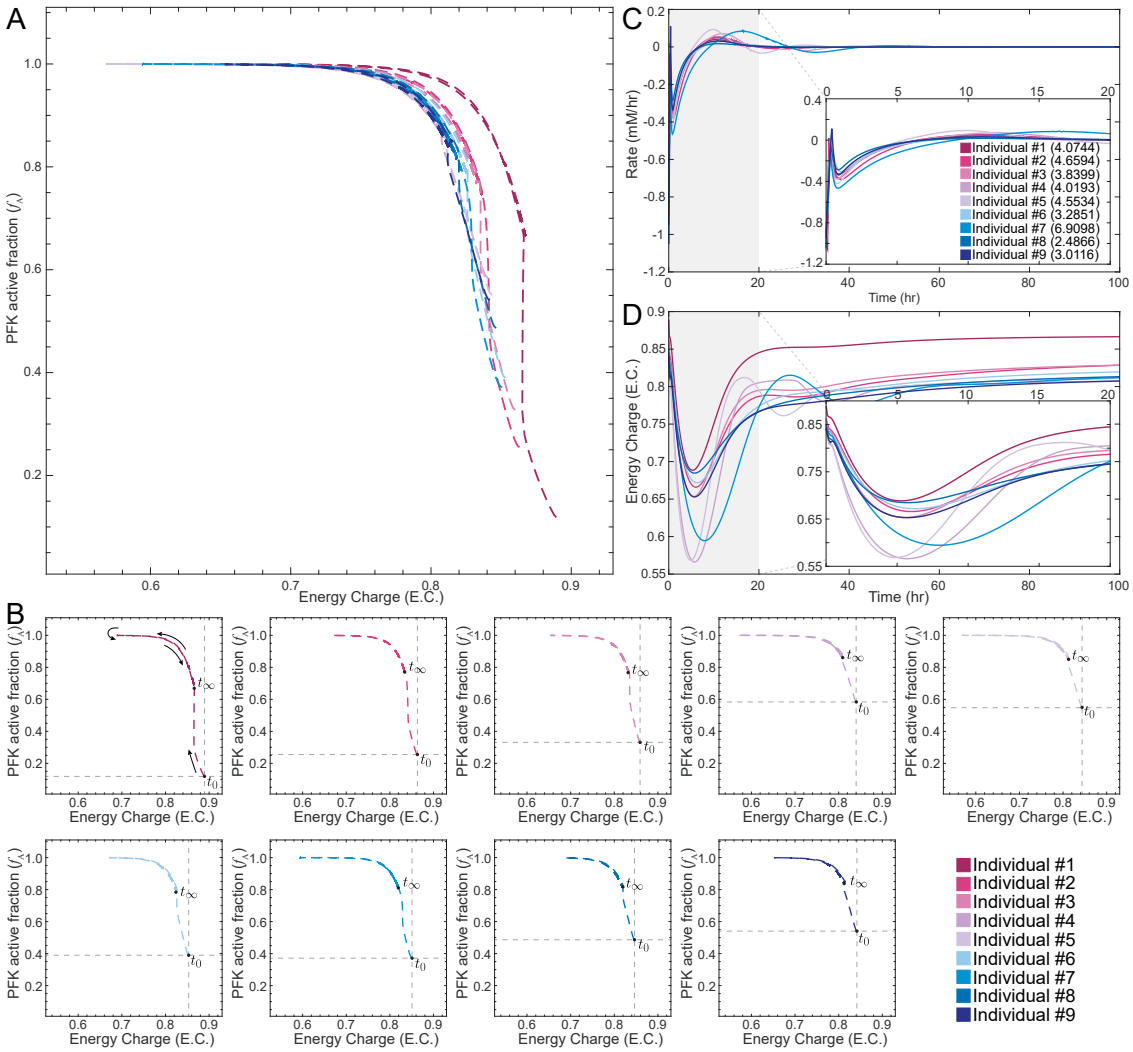

Figure 4.5: Disturbance rejection capabilities of personalized glycolytic models. (A) Superimposed catalytic potential plots for all personalized models. (B) Catalytic potential plots for each individual; the intersection of the gray lines denotes the initial steady-state value at time zero and helps show the differences among the population. (C) The net rate of ATP usage (i.e., total flux through ATP-producing reactions minus total flux through ATP-consuming reactions) is shown as a function of time; the inset zooms in on the 0 to 20 hour time range. The number in parentheses represents the SSE for each model, quantifying the total deviation of the output from the setpoint. (D) The energy charge is shown as a function of time; the inset zooms in on the 0 to 20 hour time range.

Individual kinases serve as tuning dials for the system. Adjusting these dials changes the response of the system, as demonstrated by examining individual parameterization of personalized models (Fig 4.5). Through an examination of the catalytic potential of PFK, we were able to gain

insight into how the regulator within a model is tuned in different individuals in order to maintain homeostasis (Fig 4.5A,B,D), a behavior that was not discernible through more typical metrics like rates of reaction (Fig 4.5C). Our observations of PFK and HEX were similar to those reported in the literature, but we observed differences in the behavior of PYK. There are several factors that could account for this discrepancy that focus on the scale and environmental factors of our model in comparison to the literature. The networks used in previous studies were on a much smaller scale than our network, negating the influence of other enzymes on PYK activity. Additionally, these assays did not contain FDP, which is a known activator of PYK. In our model, increasing the energy charge led to an initial increase in FDP concentration, which corresponded to an increase in the amount of PYK in the catalytically active form. While the present results are limited by the scope of the model (i.e., only glycolysis), expanding this framework to larger-scale models of RBC metabolism could provide similarly interesting results.

Second, the disturbance rejection capabilities of the models improved with the incorporation of additional regulatory mechanisms (Fig 4.4A). We simulated physiologically-relevant perturbations, observing that systems with regulation are improved over those with less regulation (i.e., fewer modules) as shown by quantifying the total deviation of the model output from the setpoint (i.e., the SSE). It is notable that models with hemoglobin and either HEX or PFK performed well despite not accounting for all regulatory mechanisms, indicating that kinetic models that do not account for the regulation in these important steps in glycolysis fail to capture important behaviors that affect the rest of the network. It is likely that the vast improvement of those two models over the model with PYK and hemoglobin is due to the fact that PYK is one of the last steps in glycolysis and therefore has a smaller impact on the rest of the system. We further investigated the disturbance rejection capabilities of the PFK and

hemoglobin model through a sensitivity analysis that used physiologically-relevant parameterizations instead of randomly-distributed parameter sets. This analysis helped elucidate subtle differences among individuals that were accessible only by studying the systems-level effects of regulation.

Finally, we have shown that the catalytic potential is a metric that can provide additional insight into how metabolic networks maintain a homeostatic state following physiologically-relevant perturbations. Using a small-scale model that explicitly accounted for the regulatory mechanisms of the three glycolytic kinases, we investigated the interplay between these three enzymes. When we applied this metric to examine the response of personalized models to ATP utilization perturbations, we observed differences that were not apparent simply from the rate profile. Upon further investigation, we were able to hypothesize that the catalytic potential for that individual was different than the others due to differences in the binding of ATP and F6P to PFK. Thus, the catalytic potential helped provide insight into how subtle differences among individuals can lead to differing systemic responses to perturbations that push the system away from the homeostatic state.

Red blood cells are networks consisting of well-studied metabolic pathways and their associated metabolites. However, it is often difficult to examine individual enzymes *in vivo* without using small scale assays [142, 143, 145, 146]. These assays are not comprehensive and, as a result, may not provide an accurate depiction of the interplay between multiple regulatory enzymes in a network like glycolysis. New methods of visualizing this behavior—such as the catalytic potential plot introduced here—can lead to new insights and discoveries. Viewing enzymes as regulators through which we can tune the system response opens the door for us to investigate what the optimal state might be and how that state helps maintain homeostasis.

4.1.3 Methods

All calculations were performed in Mathematica 11.1 [147]. Simulations were conducted using the Mass Action Stoichiometric Simulation (MASS) Toolbox kinetic modeling package (https://github.com/opencobra/MASS-Toolbox). Details for formulating a MASS model are found in Jamshidi et. al. [131]. All models used are available upon request.

Glycolysis and the Rapoport-Luebering shunt

The base glycolysis network included all 10 glycolytic enzymes and lactate dehydrogenase. Reaction rates were defined using mass action kinetics, representing enzyme catalysis as a single step. These simplified reactions were systematically replaced with enzyme modules following the procedure outlined by Du et al. [132]. Additionally, a phosphate exchange reaction was incorporated into the glycolytic network utilizing parameters obtained from Prankerd et al. [148]. Similarly, the Rapoport-Luebering shunt was included in some models to account for the presence of hemoglobin, whose binding to oxygen is regulated by 2,3-diphosphoglycerate (2,3-DPG). Incorporation of this shunt was accompanied by parameter changes as previously described [4].

Enzyme module construction

Regulation was manually incorporated into the enzyme reactions. Initial conditions from the glycolysis and hemoglobin MASS toolbox example data were used in conjunction with equilibrium constants which were obtained from from [149, 150]. These values were subsequently utilized to solve for new kinetic parameters. This procedure (outlined in [4]) adheres to the formula:

$$\frac{d\vec{x}}{dt} = \mathbf{S} \cdot \vec{v} = 0 \tag{4.3}$$

where $d\vec{x}/dt$ is the concentration rate of change with respect to time for metabolites, **S** is the stoichiometric matrix, and $\vec{v}$ is a vector containing reaction fluxes.

We constructed a total of ten different models with varying amounts of regulation, spanning from the base glycolytic model with no enzyme modules (and therefore no regulation) to the TKRM with three enzyme modules and the Rapaport-Luebering shunt. The remaining models represented each combination of the three kinase modules. Enzyme module incorporation was accompanied by the deletion of the original single-step reaction in order to avoid redundant reactions. Stability for all systems was verified by simulating the network and ensuring that a steady-state point was found for all metabolites.

Hexokinase (HEX)

HEX (EC 2.7.1.1) was modeled as a monomer to account for the fact that it contains only one active catalytic site. The previously specified mechanism was chosen to match that used by [132] because all kinetic parameters were obtained from this source. A hemoglobin module is necessary to include when the HEX module is included because it affects the level of 2,3-DPG, which serves as a regulatory molecule for HEX. The HEX module consisted of the following chemical reactions:

$$\textbf{HEX} + \text{ATP} \rightleftharpoons \textbf{HEX}\text{-ATP}$$

$$\textbf{HEX} + \text{GLC} \rightleftharpoons \textbf{HEX}\text{-GLC}$$

$$\textbf{HEX} + \text{G6P} \rightleftharpoons \textbf{HEX}\text{-G6P}$$

$$\textbf{HEX} + \text{ADP} \rightleftharpoons \textbf{HEX}\text{-ADP}$$

$$\textbf{HEX}\text{-ATP} + \text{GLC} \rightleftharpoons \textbf{HEX}\text{-ATP-GLC}$$

$$\textbf{HEX}\text{-GLC} + \text{ATP} \rightleftharpoons \textbf{HEX}\text{-ATP-GLC}$$

$$\textbf{HEX}\text{-GLC} + \text{23DPG} \rightleftharpoons \textbf{HEX}\text{-GLC-23DPG}$$

$$\textbf{HEX}\text{-ATP-GLC} \rightleftharpoons \textbf{HEX} + \text{ATP} + \text{GLC}$$

where the bold text represents the enzyme and dashes show bound species.

Phosphofructokinase (PFK)

PFK (EC 2.7.1.11) was modeled as a homotetramer to account for its four catalytic and allosteric binding sites [151]. The previously specified mechanism was chosen to match that used by [132] because all kinetic parameters were obtained from this source. The PFK module consisted of the following chemical reactions:

$$\textbf{PFK} + \text{ATP} \rightleftharpoons \textbf{PFK}\text{-ATP}$$

$$\textbf{PFK}\text{-ATP} + \text{F6P} \rightleftharpoons \textbf{PFK}\text{-ATP-F6P}$$

$$\textbf{PFK}\text{-ATP-F6P} \rightarrow \textbf{PFK} + \text{ADP} + \text{FDP} + \text{H}$$

where the bold text represents the enzyme and dashes show bound species. Additional reactions were included to account for the conversion between the tight and relaxed state, as well as the effector molecule binding.

Pyruvate kinase (PYK)

PYK (EC 2.7.1.40) was modeled to include allosteric regulation. Additional reactions were also included to account for the equilibration of both enzymes between the relaxed (R) and tense (T) state [140]. Additionally, PYK was modeled as a tetramer to account for the four catalytic and allosteric sites on each enzyme. Dissociation constants were obtained from [127] and rate constants were solved using equation 3. The PYK module consisted of the following chemical reactions:

$$\mathbf{PYK} + \text{PEP} \rightleftharpoons \mathbf{PYK}\text{-PEP}$$

$$\mathbf{PYK} + \text{ADP} \rightleftharpoons \mathbf{PYK}\text{-ADP}$$

$$\mathbf{PYK}\text{-ADP} + \text{PEP} \rightleftharpoons \mathbf{PYK}\text{-ADP-PEP}$$

$$\mathbf{PYK}\text{-PEP} + \text{ADP} \rightleftharpoons \mathbf{PYK}\text{-ADP-PEP}$$

$$\mathbf{PYK}\text{-ADP-PEP} \rightarrow \mathbf{PYK} + \text{ADP} + \text{PEP} + \text{H}$$

where the bold text represents the enzyme and dashes show bound species.

Personalized models

Personalized models were constructed by replacing all primary intracellular glycolytic metabolite concentrations and equilibrium constants with values reported by Bordbar et al. [37]. New pseudo-elementary rate constant (PERC) values were calculated using the personalized concentration data. The Rapoport-Luebering shunt was added to the RBC network and PFK enzyme modules were created for all individuals using the resulting concentration values after the addition of the Rapoport-Luebering pathway. Due to numerical issues when attempting to simulate, we only used 9/24 of the models available in [37]. Individuals #1-9 in our study

correspond to individuals 2, 4, 5, 6, 7, 8, 10, 16, and 18, respectively, from [37].

To identify outliers within the reaction PERCs compared with the other personalized models, we performed a Dixon's Q test [152]:

$$Q = \frac{\text{gap}}{\text{range}} \tag{4.4}$$

where the gap is the absolute difference between the point in question and the nearest value, and the range is the range of all values. For a set with nine samples, we can be 99% confident that a point is an outlier if the Q value is greater than 0.598; the Q values for the ATP and F6P binding steps had Q values of 0.84257 and 0.73164, respectively.

System analysis

Rate pools for enzymes were defined as the rate of at which enzyme produced product. This was accomplished by defining a pool from the product's ODE consisting solely of the terms contributing to product formation. In other words:

$$\text{rate}_{\text{enzyme}} = \sum v_{\text{formation}} \tag{4.5}$$

where $v_{\text{formation}}$ represents the forward rate of the enzyme reaction and possesses units of $\text{mmol/L} \cdot \text{s}$. Defining the rate pools in this manner neglected effects of reversible reactions contributing to the formation of product. Thus, this pool quantified the actual catalytic activity of the enzyme of interest.

Simulating ATP utilization perturbations

In order to mimic a physiologically-relevant perturbation away from the homeostatic state, we simulated a 50% increase in ATP utilization and a 15% decrease in ATP utilization [135–137].

Changes in ATP utilization were applied by changing the rate (k_{ATP}) associated with ATP hydrolysis:

$$\mathrm{ATP} + \mathrm{H_2O} \overset{k_{\mathrm{ATP}}}{\rightleftharpoons} \mathrm{ADP} + \mathrm{P_i} \tag{4.6}$$

where $\mathrm{P_i}$ represents inorganic phosphate. We calculated the sum of squared error (SSE) for each model in order to quantify the total deviation of the output from its setpoint, which is zero. The resulting quantity (i.e., the SSE) is compared between models, with a smaller value indicating better disturbance rejection capabilities.

4.2 Scaling up to network-level dynamics

As described in Ch. 1, one of the major limitations of kinetic modeling is the difficulty in parameterizing a cell-scale model. Thus, scaling up our efforts to study the RBC metabolome require a different modeling approach. Metabolic reaction networks can be represented mathematically and interrogated using ordinary differential equations, metabolic flux analysis, or constraint-based modeling. Constraint-based modeling formalizes biochemical, genetic, and genomic knowledge of cellular metabolism into a mechanistic model and is suited for understanding systems level metabolic physiology without the need for extensive parameterization [5]. Over the past decade, constraint-based modeling methods have developed to better constrain biological systems, in part by including new types of -omics data, especially transcriptomics [10]. However, fewer studies have integrated metabolomics data with such networks. To date, dynamic simulations have primarily integrated or modeled extracellular metabolite concentrations [153, 154], with some exceptions mentioned [155, 156]. Without including intracellular concentrations, models can overlook the impact of large intracellular metabolite pools on metabolic flux [157]. Thus, there is a need for methods that can integrate changes in intracellular metabolite data with

mechanistic models to accurately predict metabolic physiology under dynamic conditions.

Here, we present unsteady-state flux balance analysis (uFBA), a constraint-based modeling method and workflow that integrates time-course metabolomics data to predict metabolic flux states for dynamic systems. uFBA and steady-state FBA (henceforth referred to simply as FBA) models were constructed and compared for three dynamic systems: stored human RBCs, stored human platelets, and *Saccharomyces cerevisiae* during anaerobic batch fermentation and carbon starvation. In addition, one classical example of a static system was modeled: *Escherichia coli* during steady-state exponential growth. We find that for the dynamic systems, inclusion of intracellular metabolomics with uFBA provides different and more accurate predictions than FBA. In particular, uFBA predictions for RBC were experimentally validated with isotopic metabolic flux analysis. For the platelet and yeast systems, experimental data from the literature was used to benchmark the modeling results. Finally, the static *E. coli* system served as a negative control, with uFBA and FBA displaying similar predictions.

4.2.1 Results

Unsteady-state Flux Balance Analysis (uFBA)

The uFBA workflow integrates time-course absolute quantitative metabolomics data with a constraint-based model to predict metabolic flux states. Time-course metabolomics data is often non-linear. The first step of the workflow discretizes non-linear metabolite time profiles into time intervals of linearized metabolic states for piecewise simulation, resulting in a separate model for each metabolic state (Fig. 4.6). Principal component analysis (PCA) is applied to the time-course metabolomics data to identify the time intervals for which a model will be constructed. For example, the data in Fig. 4.6 is discretized into two time intervals.

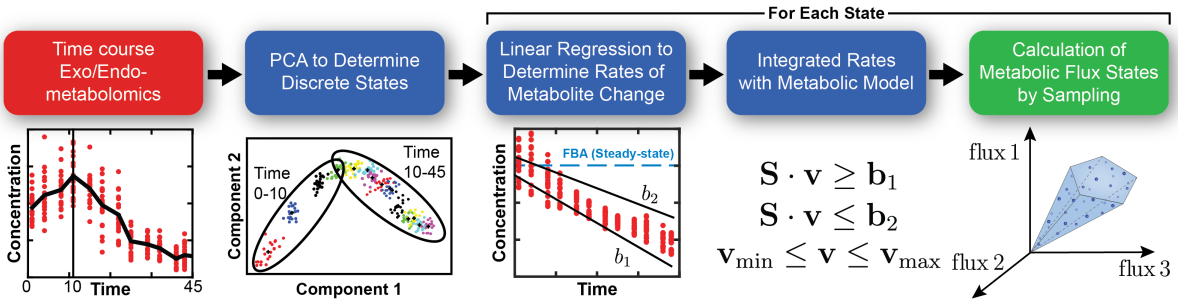

Figure 4.6: Overview of the uFBA workflow. First, extracellular (exo) and intracellular (endo) metabolite time profiles are split into discrete time intervals of linearized metabolic states using principal component analysis. For each metabolic state, the rate of change of each metabolite is calculated using linear regression, along with the 95% confidence interval ($\mathbf{b}_1$ and $\mathbf{b}_2$). If the metabolites rate of change is significant, the model is updated by changing the steady-state constraint $\mathbf{b} = 0$ to a range denoted by $\mathbf{b}_1$ and $\mathbf{b}_2$. uFBA differs from FBA in that elements of the $\mathbf{b}$ vector are known and can be used as constraints, but FBA in the absence of such information assumes that these elements are zero (i.e., at steady-state).

For each time interval, a parameterized model is constructed by (1) determining the rate of change of measured metabolites using linear regression, (2) integrating the calculated rates of change with the constraint-based model, (3) treating the model as a closed system, and (4) reconciling data measurement error and incompleteness through a metabolite node relaxation algorithm to produce a functional model.

The final two steps are different than traditional FBA. uFBA is a data driven approach and aims to make flux predictions solely from time-course metabolomics data. uFBA only allows changes to metabolite levels, including extracellular metabolites for exchange, if the metabolite is measured to be significantly increasing or decreasing in the extracellular (exo) or intracellular (endo) metabolomics data. In principle, if all metabolites are accurately measured over time, the model should simulate. However, due to the experimental infeasibility of data completeness (not all metabolites are measured), additional metabolites will need to accumulate or deplete for proper simulation. To predict which unmeasured metabolites are changing, we developed a relaxation algorithm that makes the model simulation feasible. After all exchange reactions

are removed (i.e., a closed system), the metabolite node relaxation algorithm determines the minimum number of metabolites that need to deviate from steady-state for a feasible model.

Once the uFBA model is constructed, most constraint-based modeling analyses can be used, including the maximization or minimization of an objective, flux variability analysis (FVA), and candidate flux sampling. For this study, we focused on Markov chain Monte Carlo (MCMC) sampling [9] to calculate the probability distribution of flux through every metabolic enzyme in the network. This allowed for the calculation of the most likely flux state of the system.

Global Differences between uFBA and FBA

To test and validate uFBA, the workflow was applied to four systems: (1) human RBCs stored up to 45 days at 4°C; (2) human platelets stored up to 10 days at 22°C; (3) two strains of *Saccharomyces cerevisiae* during anaerobic batch fermentation and carbon starvation, and (4) *Escherichia coli* during steady-state exponential growth as a negative control. These cases were chosen because they represent diverse systems in terms of physiological dynamics, metabolic network complexity, cell density, data coverage, and timescales.

Comprehensive absolute quantitative metabolomics data was obtained from the literature [41, 89, 158, 159]. All four datasets were generated using LC-MS methods, augmented with some extracellular metabolites being measured by HPLC or a blood gas analyzer. Using PCA, the metabolite time profiles for stored RBCs were discretized into three metabolic states, into two states for stored platelets, and into three states for *S. cerevisiae* during mixed glucose/xylose fermentation. The negative control, *E. coli* steady-state growth, was treated as one state. uFBA models were constructed and compared to steady-state FBA models for the corresponding systems and states. The steady-state models were constructed by integrating only the

exometabolomics data and allowing for extracellular exchange for unmeasured metabolites with the environment, currently a standard practice.

We assessed the global difference between calculated flux states of the uFBA and FBA models (Fig. 4.7). We observed considerable differences between uFBA and FBA predictions in the dynamic situations of long-term cell storage or batch fermentation, while fewer differences were observed for *E. coli* during steady-state growth. Further, the detected differences in flux estimates were not due to a uniform increase or decrease of flux across the network as evidenced by the significantly lower correlation of uFBA and FBA calculated fluxes as compared to controls (Fig. 4.7b). This indicates that the ordering of reactions, from high to low flux, had fundamentally changed in RBCs, platelets, and *S. cerevisiae*.

Reactions with significantly different flux estimates between uFBA and FBA were not uniformly distributed across the subsystems of the various models. Typically, higher metabolomics coverage of metabolites in a particular subsystem resulted in larger reported flux differences. Further, the number of significantly different reactions within a given subsystem was not constant across metabolic states. This difference is due to the nonlinearity of time-course data, where some metabolite concentrations change during one metabolic state but are at steady-state during another metabolic state.

Outside of MCMC sampling flux states, we assessed whether the size of the solution space had changed by uFBA using flux variability analysis (FVA). We applied FVA to each state of each test case to determine the flux range for each metabolic reaction (flux range = maximum flux − minimum flux). In the RBC and platelet cases, we found that by deviating intracellular metabolites from steady-state, flexibility in the system increased and certain reactions had larger flux ranges for uFBA than for FBA. Overall, however, we found that most of the reactions had

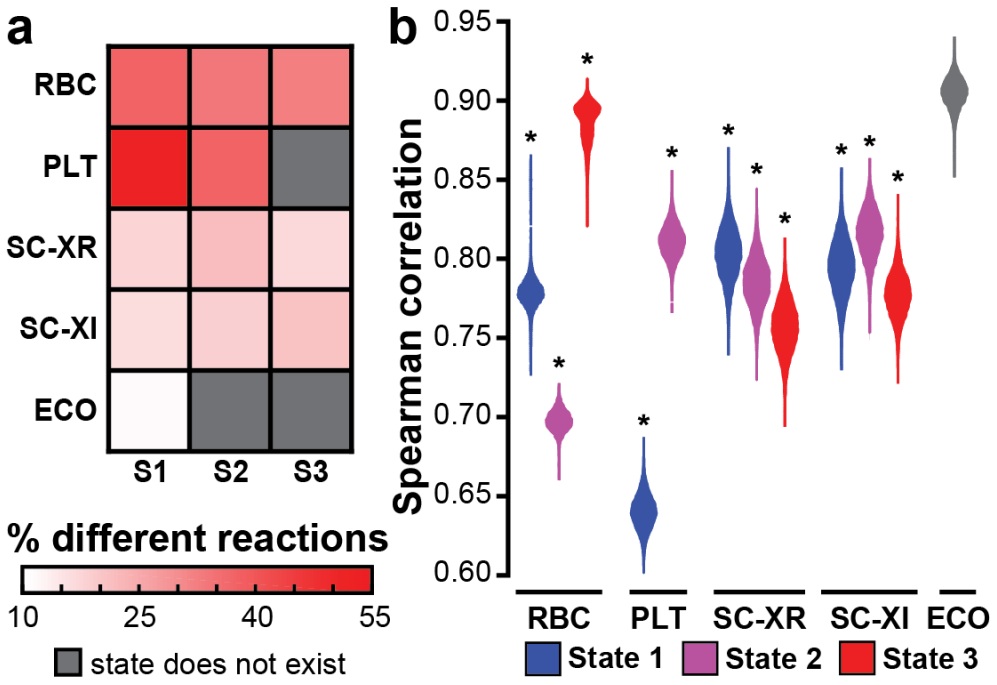

Figure 4.7: Comparison of uFBA and FBA flux states. (**a**) Percentage of reactions with significantly different fluxes between the uFBA and FBA models. (**b**) The Spearman correlation between uFBA and FBA flux states indicates that differences in flux estimates were not due to uniform increases or decreases in fluxes but a reordering of reactions from high to low flux. The null hypothesis is that the distribution of Spearman correlations is drawn from the same distribution when comparing candidate flux states from uFBA to uFBA or FBA to FBA. * - $p = 0.0$; *E. coli* $p = 0.995$.

equal or lower flux ranges in the uFBA formulation. The lower flux ranges are most likely due to uFBA constraining extracellular metabolite exchanges to only measured data, and not allowing free exchange out of the system.

After a global comparison of uFBA and FBA, we focused on key differences in reaction flux predictions that would have consequences on biological interpretation on metabolic physiology by those investigating the metabolomics data.

Red Blood Cells

There were considerable differences in flux predictions made by uFBA and FBA for RBC metabolism. To hone in on major discrepancies, we focused on metabolic reactions where uFBA and FBA predicted opposite directions of flux. We observed flux reversals in the cytosolic remnants of the TCA cycle reactions. These enzymes had been previously detected in RBC proteomic datasets [160] (Fig. 4.8a). Over storage, RBCs uptake citrate (2.32 μM/day during State 1) and secrete malate (1.96 μM/day during State 1) and fumarate (0.300 μM/day during State 1). FBA predicts that nearly all citrate is converted to malate and fumarate (Fig. 4.8c). This is a reasonable estimation because the flux into and out of the cell is roughly balanced, and the K_{eq} of malate dehydrogenase ($K_{\text{eq}} = 2.1 \times 10^5$) heavily favors metabolite flow in that direction. However, through intracellular metabolite profiling, we discovered that RBCs have a high concentration of intracellular malate (>1 mM). With this additional information, uFBA predicts that the secretion of malate and fumarate are due to the depletion of the large intracellular malate pool (26.1 μM/day, State 1). Further, uFBA predicts the shuttling of the majority of the intracellular malate and citrate into lower glycolysis through oxaloacetate, as well as production of glutamate from citrate (Fig. 4.8b). The network in Figure 4.8a shows all TCA remnant enzymes previously detected in RBCs in proteomic studies [160] or through literature curation [40].

We experimentally validated the fate of extracellular citrate by replacing the anticoagulant with fully labeled ^{13}C citrate. Measurements of isotope abundance were determined and analyzed using metabolic flux analysis (MFA) tools [161]. As intracellular metabolite levels are changing throughout and the labeling patterns are unstable, traditional "reverse" ^{13}C MFA calculations where fluxes are predicted based on isotopic labeling patterns are not applicable. Instead, we completed a "forward" MFA simulation where the isotopic labeling pattern is predicted based

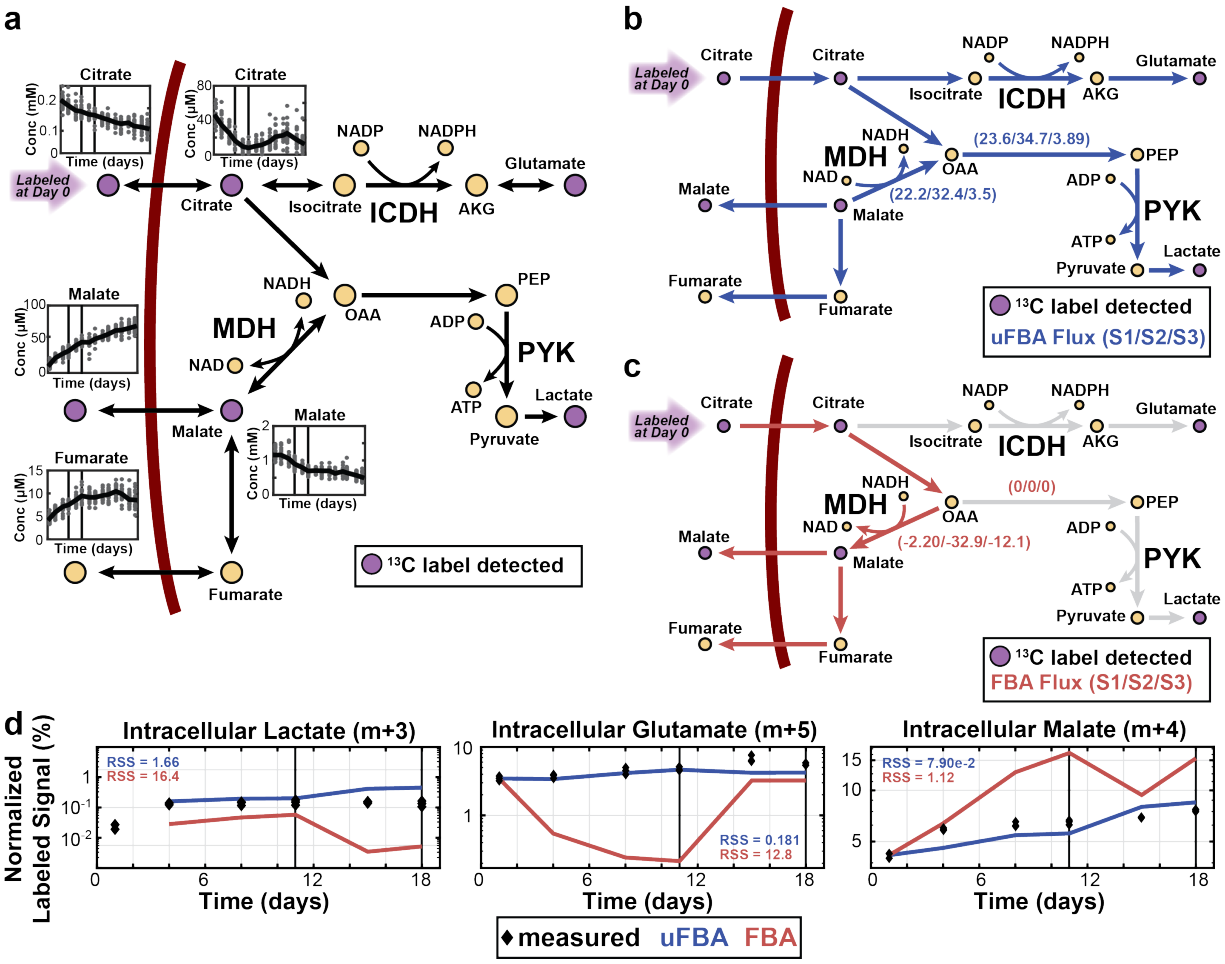

Figure 4.8: Experimental validation of *in silico* predictions. (**a**) TCA metabolites and pathways in the RBC metabolic model are shown, including changes in metabolite levels and metabolites found to be isotopically labeled after addition of fully labeled ^{13}C citrate. Cofactor producing reactions are shown, while other cofactors and reaction names are omitted. Concentrations are shown as μmol/L of bag or mmol/L of bag. Time spans 0-45 days for insets. (**b**) uFBA (blue arrow) predicts that the depletion of intracellular malate produces extracellular malate and fumarate, while driving lactate production. Extracellular citrate is used to produce glutamate and lactate. Fluxes shown in μM. (**c**) FBA (red arrow) predicts extracellular citrate is used only to produce malate and fumarate and that MDH proceeds in the opposite direction than in uFBA. (**d**) uFBA and FBA predicted fluxes were integrated with a ^{13}C MFA "forward" tracer simulation to simulate how labeled citrate would accumulate across the first two metabolic states. The residual sum of squares (RSS) of both uFBA and FBA simulations compared with the measured data is shown. Abbreviations: oaa: oxaloacetate; akg: alpha-ketoglutarate; pep: phosphoenolpyruvate; MDH: malate dehydrogenase; PYK: pyruvate kinase; ICDH: isocitrate dehydrogenase. Vertical lines on metabolite time profiles denote the time intervals of the three metabolic states.

on the initial isotopic pattern and the predicted fluxes by uFBA or FBA (see Methods). We compared how well the use of uFBA or FBA fluxes were able to match the isotopic labeling pattern of intracellular metabolites that were detected to be labeled and for which we had absolute quantitation. uFBA produced quantitatively more accurate predictions (lower residual sum of squares (RSS)) than FBA for the isotopic labeling pattern (Fig. 4.8d). The discrepancy between uFBA and FBA is predominantly due to the depletion of intracellular malate into oxaloacetate, which creates a gradient pushing flux from malate into oxaloacetate.

RBC citrate metabolism can proceed in two directions. First, alpha-ketoglutarate can be formed through aconitase and isocitrate dehydrogenase (IDH). Alpha-ketoglutarate can then form glutamate, which was found to have increasing percentages of isotopic labeling. In this process, NADPH is generated through IDH1, which is the only known isozyme proteomically detected in RBCs16. In the second direction, citrate forms oxaloacetate through ATP citrate lyase. An acetyl group is cleaved off during this process forming acetyl-CoA. We detected an increasing labeling percentage of acetylcarnitine (m+2), which is most probably created from acetyl-CoA. Oxaloacetate can then become aspartate, malate, and lactate, which we found all increasingly labeled.

Platelets

For the platelet storage data, the major discrepancy between uFBA and FBA models concerned the utilization of the electron transport chain (ETC), particularly in State 1. As this dataset did not have information on oxygen uptake, the metabolite node relaxation algorithm determined the necessary amount based on the rest of the metabolomics data. Based on the measured metabolites, the algorithm was able to accurately predict that oxygen was required

for both uFBA and FBA models, but the amount of oxygen needed was only quantitatively predicted on the right order of magnitude for uFBA. uFBA predicted that the ETC accounted for 90.2% and 88.8% of ATP generation for States 1 and 2, respectively. However, FBA predicted 0.23% and 64.1% in States 1 and 2, respectively, suggesting that platelets use only glycolysis for ATP generation (and not the ETC) in State 1. Previous experimental studies [162, 163] have shown oxygen uptake to be higher in State 1 than in State 2 and that approximately 85% of ATP production during storage is due to the ETC. The uFBA workflow quantitatively predicted the oxygen uptake rate and ATP production. When the FBA models were re-parameterized to not allow free exchange out of the system, the oxygen uptake rate and ETC usage was corrected. The discrepancy between the uFBA and FBA model predictions was determined to be caused by the standard practice in FBA models to allow free exchange of metabolites out of the system, In particular, the reason was free exchange of L-alanine out of the system.

Saccharomyces cerevisiae

The final dynamic case study used metabolomics data for two *S. cerevisiae* strains engineered to assimilate xylose and fermented in a mixed glucose/xylose culture [158]. One strain consumes xylose through an isomerase (strain XI), while the other consumes xylose through a reductase and a dehydrogenase (strain XR). The PCA identified three metabolic intervals. The first shift differentiated between mixed nutrient metabolism (glucose/xylose, State 1) and xylose as the sole carbon source (States 2 and 3). Using a constraint-based model for yeast metabolism, we found considerable differences between uFBA and FBA flux predictions (Fig. 4.7). We compared the flux predictions against recently generated ^{13}C MFA studies on XI and XR yeast strains [164, 165]. Although the experimental conditions were not identical, the uFBA models

predicted flux state resulted in a lower residual normalized error than did FBA when compared with the measured data for the three tested cases: XR growth on glucose (State 1) and xylose (State 3); XI growth on glucose (State 1). XI growth on xylose (State 3) was excluded for comparison as the growth rate of yeast in the ^{13}C MFA study was considerably different than in the study used for generating uFBA and FBA predictions.

Further, the uFBA flux predictions provide a systematic method to propose mechanisms for the observed time-course changes in metabolites, rather than use intuition. In particular, the authors postulated that the observed drain of 6-phosphogluconate (6PG) in the XI strain after glucose consumption is due to a reduced flux through the non-oxidative PPP. uFBA instead predicts 6PG decrease to be due to a significant decrease in oxidative PPP activity in State 2. The FBA model did not make a similar prediction. Further, the large decrease in flux through the oxidative PPP predicted by uFBA was confirmed with the ^{13}C MFA studies on yeast XI and XR strains noted above.

Escherichia coli

The *E. coli* data was used as a negative control because FBA has been traditionally successful in analyzing microbial growth processes based on exometabolomic data alone, in part due to *E. coli*s balanced growth nature. As expected, uFBA flux predictions deviated less from steady-state flux predictions than did the other test cases (Fig. 4.7). The similar results demonstrate that the differences observed in the other case studies are not artifacts of the uFBA workflow.

Though flux predictions are not substantially affected, the inclusion of endometabolomics measurements does impact gene essentiality predictions in *E. coli*. Overall results were very

similar for the 1,366 genes in the model, but uFBA predicted non-essentiality for 11 genes that FBA predicted to be essential in glucose minimal media. Like other linear programing based constraint-based methods, uFBA cannot explicitly account for metabolites as variables. uFBA incorporates the rate of depleting intracellular metabolites without accounting for the metabolite concentration, thus providing an infinite supply of the metabolite, increasing chances of predicting non-essentiality. This artifact of depleting pools is similar to potential reasons for conflicts in experimental gene essentiality results in minimal media. Experimentally, 17.8% of genes predicted to be essential by uFBA have conflicting experimental results across three studies [166–168] in glucose minimal media. The 11 differing predictions were enriched in conflicting experimental results (3.06× enrichment, $p = 0.0032$), suggesting that the measured drains in intracellular metabolites may play a role in conflicting experimental results. The genes were related to NAD and AMP biosynthesis, which were rescued by the measured depletion of the cofactor pools during growth. The observed discrepancies in gene essentiality calls may be due to differences in plating techniques, the time point for assaying growth, or the chosen growth/no growth threshold. Residual intracellular metabolite pools from LB media before plating may play a role in causing conflicting results. These results suggest that *E. coli* retains higher than required cofactor levels in anticipation of changing environmental conditions, a result consistent with the finding that intracellular concentrations of ATP and NAD pools in *E. coli* are an order of magnitude higher than K_{m}s of their associated enzymes [169].

4.2.2 Discussion

Metabolomics data provides a rich detailing of cellular biochemistry. Metabolomics data is becoming readily available, and there is still a need for tools that can integrate such data into

mechanistic models to provide a deeper understanding of systems level metabolic physiology. Statistical methods can pinpoint changes or associations but have difficulty elucidating mechanisms. Metabolic modeling techniques can predict which metabolic pathways or enzymes caused the observed statistical changes (i.e., whether upstream or downstream enzymes are more likely the cause of observed behavior), although more detailed, kinetic modeling and metabolic flux analysis are often difficult to construct and parameterize for large cellular networks. Constraint-based models are better suited for studying metabolism at cellular scale, but the steady-state assumption hinders studying dynamic states.

In this study, we present unsteady-state flux balance analysis, a constraint-based modeling method, to study dynamic cellular states. uFBA provides additional utility to existing constraint-based methods that integrate metabolomics data. We identified four test cases and for each, compared uFBA flux predictions with steady-state FBA in order to quantify the advantages of integrating intracellular metabolite concentrations. For the three dynamic systems, we found considerable differences in flux predictions. The size of the metabolic network and data coverage (i.e., what percentage of metabolites are measured) impacts the increased utility of uFBA over traditional FBA as evidenced by the significant differences in RBCs and less significant differences in *S. cerevisiae*. For a more traditional use of FBA (steady-state bacterial growth), we found less difference, confirming that the uFBA is not overly sensitive. Theoretically, in a system where no intracellular metabolites change over time, uFBA and FBA would predict identical flux distributions, indicating that the use of uFBA should be considered if metabolomics measurements are available.

The results presented here have two major implications. First, uFBA quantifies the impact of large and previously unmeasured intracellular metabolite concentration pools on network

flux calculations. We experimentally validated a notable uFBA prediction for RBCs using isotopic labeling and metabolic flux analysis. The unexpected complexity of TCA metabolism in human RBCs is biologically notable. RBCs utilize TCA intermediates to produce phosphoenolpyruvate (PEP), most probably through a PEP carboxylase-like mechanism mediated by hemoglobin [170] which ultimately results to ATP generation by pyruvate kinase. Further, RBCs produce NADH and NADPH outside of glyceraldehyde dehydrogenase and the pentose phosphate pathway through the cytosolic forms of malate dehydrogenase and isocitrate dehydrogenase, respectively. This finding may be of importance in transfusion medicine as some of the FDA approved media additives for RBC cold storage contain large amounts of citrate (20 mM), while other additives do not. The added citrate may affect the red blood cells ability to combat oxidative stress during RBC storage [24]. This discovery was only possible through absolute quantitative metabolomics and uFBA. The high levels of intracellular malate (>1 mM) had not been previously observed. Without the intracellular data, the uptake and secretion of metabolites by RBCs was still mass balanced so FBA predictions would have been inaccurate but still feasible. This example shows the importance of combining comprehensive intracellular metabolomics with network calculations. Further, as the labeling percentage of malate is less than 5% at day 1 (Fig. 4.8d), the concentration of malate is not due to RBC storage in citrate anti-coagulant.

Second, the common practice to allow free metabolic exchange of unmeasured metabolites out of the system can lead to erroneous predictions. For the platelet data, FBA predictions were inaccurate partly due to free exchange of L-alanine out of the system. For the yeast example, differences in FBA and uFBA predictions were partly due to the use of exchange reactions. uFBA aims to be fully data driven and only allows exchange of metabolites out of the system

if they were experimentally measured or if the optimization algorithm requires the metabolite to be secreted for model feasibility. With the increasing size of metabolic networks, it requires both intimate knowledge of the cellular system and of constraint-based modeling to identify the modeling inaccuracies caused from metabolic exchanges. The uFBA approach simplifies metabolomics data integration and accurately deals with data inconsistencies through the node relaxation algorithm.

Two other methods for systematically integrating intracellular metabolite concentrations with constraint-based models are available. TREM-Flux [156] and MetDFBA [155] have important methodological differences from uFBA. First, TREM-Flux estimates network flux between each two time points, making the approach extremely susceptible to data noise and outliers. The uFBA approach avoids this issue by defining time intervals that represent metabolic states, effectively lowering the chances that noise and outliers affect flux predictions. Second and more importantly, TREM-Flux accounts for data incompleteness by allowing all unmeasured metabolites to deviate from steady state up to the maximum change measured in the metabolomics data. Such an approach provides too much freedom in the optimization problem, making many fluxes inaccurate. uFBA deals with data incompleteness by modifying the fewest number of metabolites using the metabolite node relaxation algorithm.

MetDFBA integrates intracellular measurements with the traditional DFBA approach. In order to deal with the huge complexity of the necessary ordinary differential equations, MetDFBA lumps and removes the majority of metabolic reactions resulting in smaller, core networks. While focusing on specific cellular processes is common in kinetic modeling, it eliminates the ability to interrogate metabolism at the whole-cell level, which is possible with FBA methods. Though uFBA is not as detailed as MetDFBA, it allows for studying metabolism at a comprehensive

scale.

The integration of metabolomics data using uFBA has limitations. First, the metabolomics data used to constrain the model must be absolutely quantified using internal standards. Absolutely quantified metabolomics data is often difficult to generate accurately and is more expensive. Second, the increased accuracy of uFBA predictions are determined by how large a percentage of metabolites are measured in the network. In particular, it is important to have metabolite measurements for cofactors and high flux pathways. Third, like other constraint-based methods, uFBA does not explicitly model metabolites as variables and thus total concentrations are not captured. Fourth, organelle specificity of metabolite concentrations is often lacking, requiring modelers to make assumptions on metabolite location. Finally, metabolomics data is noisier than other -omics data types. To calculate significant metabolic rates of change, uFBA may require more than three replicates.

uFBA provides a systematic and standardized method to generate hypotheses for the causes of detected changes in metabolite levels over time. uFBA flux predictions are based on unlabeled metabolomics but provide high quantitative accuracy in flux estimates even during dynamic metabolic conditions. These findings are not evident from statistical analysis of the time-course metabolomics alone nor from standard analysis of FBA models. We anticipate that the use of uFBA and the associated workflow will aid in deeper analysis of metabolomics data while also increasing the predictive power of constraint-based models.

4.2.3 Materials and Methods

All analyses were performed in Matlab (Mathworks, Natick, MA) using the COBRA 2.0 Toolbox [9]. The uFBA method and associated workflow are available as an extension for

COBRA 2.0 at (https://opencobra.github.io/cobratoolbox/).

Data Preparation

Red blood cell metabolomics data from normal blood banking conditions using SAGM media was taken from [41]. Metabolomics data for platelets retrieved by apheresis during storage under normal blood banking conditions was taken from [89]. Metabolomics data for *S. cerevisiae* during mixed culture anaerobic fermentation was taken from [158]. Metabolomics data for *E. coli* during exponential growth was taken from [159]. Missing values were imputed using a k-nearest neighbor algorithm that takes the weighted average of the five nearest neighbors; this approach has been previously shown to be accurate for metabolomics data [171]. The intracellular and extracellular concentrations were adjusted so that concentrations were in mmol/L of total bag volume for RBCs and platelets, and mmol/gDW for *S. cerevisiae* and *E. coli*. For RBC and platelet datasets, glucose and lactate were measured using a blood gas analyzer. For RBC data, ATP and 2,3-DPG were measured using enzymatic assays. In Bergdahl et al. and McCloskey et al., changes in key extracellular metabolites in *S. cerevisiae* (glucose, xylose, xylitol, ethanol, and glycerol) and *E. coli* (glucose, acetate, succinate, and formate) were also measured using HPLC. The low-throughput and HPLC measurements were prioritized over mass spectrometry measurements for use in the model. Some metabolite pools were not resolved in the original data and were manually split into individual concentrations based on known physiological ratios (3-Phospho-D-glycerate/D-Glycerate 2-phosphate, pentose sugars, leucine/isoleucine). Due to the rigorous quality control standards within blood banks for those two publications, there was negligible RBC and platelet cell death during storage. Thus, cell death was not included during modeling.

Principal component analysis and linear regression

Principal component analysis (PCA) was performed on each metabolomics data set in order to objectively determine the time intervals of the discrete, linearized metabolic states. PCA was performed on the standardized Z-scores. Once the time intervals for each state were defined, linear regression was performed in order to estimate the rate of change for each metabolite during that particular state. The 95% confidence interval for each metabolites rate of change was calculated for integration with the model. If the 95% confidence interval for a particular metabolites rate of change crossed zero, the metabolite was deemed to be at steady-state as there is not enough statistical evidence for the metabolite to be changing. Rates for RBC and platelet were in mmol/L/h, while *S. cerevisiae* and *E. coli* rates were in mmol/gDW/h.

Constraint-based model integration

The endometabolomics data acquired from the literature was integrated with constraint-based models. The RBC data was integrated with a modified version of the erythrocyte model iAB-RBC-283 [40], which was previously used for building personalized kinetic models [37]. The platelet data was integrated with the platelet metabolic model iAT-PLT-636 [172]. The *S. cerevisiae* data was integrated with the *S. cerevisiae* metabolic model iMM904 [173]. The *E. coli* data was integrated with the *E. coli* metabolic model iJO1366 [166]. The measured growth rate was included as a constraint for *S. cerevisiae* and *E. coli* simulations. Platelet and *S. cerevisiae* metabolic models contain multiple compartments. If metabolites were known to be predominantly from a specific compartment, the metabolomics data was assigned as such. Specifically, tricarboxylic acid cycle metabolites were set to the mitochondria. If no information was available, metabolite concentrations were assumed to be in the cytosol.

Unsteady-state flux balance analysis (uFBA)

The significant rates of change for measured metabolites for each state (as determined by linear regression) are input to the uFBA method. All four test cases are treated as closed systems, and all exchange reactions were removed from the model. Subsequent steps (see next section) determine other metabolites that can enter or leave the system. The measured metabolomics data is integrated with the model by:

$$\mathbf{S} \cdot \mathbf{v} \geq \mathbf{b}_1 \tag{4.7}$$

$$\mathbf{S} \cdot \mathbf{v} \leq \mathbf{b}_2 \tag{4.8}$$

where $\mathbf{S}$ is the stoichiometric matrix, $\mathbf{v}$ is the calculated flux vector, and $[\mathbf{b}_1, \mathbf{b}_2]$ represent the 95% confidence interval for each significantly changing metabolite. All unmeasured metabolites are assumed to be at steady-state: $\mathbf{b}_1 = \mathbf{b}_2 = 0$.

Ideally, all model metabolites would be accurately measured. If so, the model would properly simulate, as all metabolic changes would be accounted for. However, due to experimental limitations, most metabolites cannot be measured, and often those that can be measured are done so unreliably. Optimally, one would measure metabolites that have the highest rate of change relative to the flux going through the associated pathways. However, knowing which metabolites these are may not be possible as such metabolites may change from condition to condition or the system being studied may not be well known. Thus, from any metabolomics dataset, there exists unmeasured intracellular or extracellular metabolites that are not at steady-state that are required to change in order to allow the model to simulate.

To deal with this data incompleteness and data quality issue, we developed an algorithm to reconcile the metabolomics data and the network structure. In brief, the algorithm tries to

parsimoniously allow unmeasured metabolites to deviate from steady-state in order to build a computable mode. The approach is detailed in the next section.

Further, a standard practice in FBA is to allow all extracellular metabolites to have free exchange out of the system. As uFBA is a metabolomics driven approach, uFBA only allows exchange of extracellular metabolites out of the system if (1) it is measured to be increasing in the exometabolomics data, or (2) it is required by the metabolite node relaxation algorithm for feasibility.

The remainder of the uFBA modeling formalism follows FBA principles. In particular, a biomass objective function is used in growing cells such as *S. cerevisiae* and *E. coli* to account for the generation of proteins, RNA, DNA, and lipids. Further, metabolite reserves are modeled using sinks. This is used in particular for glycogen stores for *S. cerevisiae* during carbon starvation.

From a mathematical standpoint, the uFBA approach changes the degrees of freedom (DOF) of the system. First, for each "unsteady" metabolite incorporated, the system gains a DOF. Next, all exchange reactions are removed, reducing the DOF by the number of reactions removed. At this point, the metabolic model is most likely infeasible as the system is now over-determined. Previous approaches (i.e., TREM-Flux) deal with this issue by allowing all unmeasured metabolites to deviate from steady-state, increasing the DOF by one for each unmeasured metabolite. The uFBA approach applies a metabolite node relaxation algorithm to parsimoniously deviate unmeasured metabolites from steady-state to minimize the increase of DOF. The systems increase in DOF is equivalent to the number of metabolites deviated by the algorithm.

Unmeasured metabolite relaxation from steady-state

Because not every metabolite in the network was measured, we developed an automated method for parsimoniously deviating unmeasured metabolites from steady-state to build a computable model. We term this estimation "metabolite node relaxation."

For each metabolite that was not measured, two sink reactions are added that allow each of these metabolites to both accumulate ("up" sink reaction) and deplete ("down" sink reaction). Then, an optimization problem determines the minimal number of sinks to retain while still having a computable model. This parsimonious method was chosen under the assumption that cellular systems typically aim to maintain homeostatic levels, which has been previously shown [174].

To exhaust potential methods, we implemented five different optimization approaches for relaxing the unmeasured metabolite nodes in the system (Eqs. 4.9-4.13) that all assume parsimony but in slightly different ways. The first technique is an MILP optimization that minimizes the number of unmeasured metabolites relaxed from steady-state (Eq. 4.9):

$$\min \sum_{i}^{m} 1_{\Delta x_i \neq 0} \tag{4.9}$$

where m is the number of unmeasured metabolites in the system and Δx is the deviation from steady-state of the unmeasured metabolites, which is defined as the flux through the sink reaction. Essentially, Case 1 minimizes the increase of the DOF. The second technique is an LP optimization that minimizes the sum of the magnitude of the rate of change of unmeasured metabolites (Eq. 4.10):

$$\min \sum_{i}^{m} |\Delta x_i|. \tag{4.10}$$

The third technique is also an LP optimization that minimizes both the sum of the magnitude of

the rate of change of the unmeasured metabolites as well as the reactions fluxes in the network (Eq. 4.11):

$$\min \sum_{i}^{m} |\Delta x_i| + \sum_{i}^{n} |v_j| \tag{4.11}$$

where n is the number of intracellular reactions in the system and v is the reaction flux. The fourth technique is a QP optimization that minimizes the sum of the square of the rate of change of the unmeasured metabolites (Eq. 4.12):

$$\min \sum_{i}^{m} \Delta x_i^2 \tag{4.12}$$

The fifth technique is also a QP optimization that minimizes the sum of the square of the rate of change of the unmeasured metabolites and the square of the intracellular reaction fluxes (Eq. 4.13):

$$\min \sum_{i}^{m} \Delta x_i^2 + \sum_{i}^{n} v_j^2. \tag{4.13}$$

As alternative optima may exist for these optimization problems, especially for Case 1, the given optimization problem is run multiple times for a user-specified number of iterations, not allowing for previous solutions using an integer cut method. For the larger models (platelet, *S. cerevisiae*, and *E. coli*), the termination criterion for optimization for each iteration was set to a relative gap tolerance of 1e−6 or a time limit of 45 seconds. These criteria were chosen based on convergence properties of the solutions as the two parameters were varied.

The multiple solutions are tallied, and a final optimization is run, preferentially weighting unmeasured metabolites that appeared more frequently in the multiple solutions. The sinks associated with the relaxed nodes are retained, while the remaining sinks are removed from the model. 100 iterations were used in this study. A comparison of the accuracy of the optimization approaches in determining necessary relaxations is discussed in the next section. Further, we

provide the user with the option to preferentially weight unmeasured extracellular metabolite nodes to be relaxed first, as such an approach is similar to the common practice of modifying extracellular exchanges in FBA models.

Once the final set of unmeasured metabolites to be relaxed from steady-state has been determined, the magnitude of relaxation is addressed. The flux through the retained sink reactions is minimized while allowing the model to simulate. We tested whether increasing this minimum amount affected flux simulations (scaling the minimum bound by 1×, 1.5×, 5×, 10×, or 100×). Sensitivity analysis of this parameter by re-sampling the reaction fluxes showed that the 1x scaling was too tightly constrained, yielding different results than the higher multiples. However, the higher multiples were very similar to each other. A scaling of 1.5× was used for all analyses in this study as it does not over constrain the network.

Accuracy of unmeasured metabolite relaxation from steady-state

We tested whether the various optimization approaches for parsimoniously relaxing unmeasured metabolites from steady-state was an accurate method for the data incompleteness issue. All five techniques were used to build uFBA models for the RBC metabolomics data. The relaxed nodes for each of the five approaches were compared to a qualitative dataset for the same condition that measured many more metabolites than our absolute concentration dataset. The larger dataset had qualitative information on the changes (or lack thereof) for 31 metabolites that were unmeasured in the absolute concentration dataset. Based on the time points measured in the qualitative dataset, only states 1 and 3 in the RBC could be compared. Metabolites that increased 2× or decreased 0.5× were deemed to be changing in the qualitative dataset.

The criterion for selecting the best optimization technique was minimizing the number

of cases where a metabolite was incorrectly relaxed from steady-state. This criterion was chosen as it is more detrimental to over relax than to under relax, because it allows the model to have many new feasible flux states that may be erroneous. We found the MILP formulation (Case 1) to be the best method for this criterion. Further, the MILP formulation was also found to have the overall best accuracy. For all analyses in the main text, the MILP formulation is used.

Construction of FBA models

The FBA models constructed for control comparison against the uFBA models followed the same workflow as the uFBA model, except that only the exometabolomics data was integrated. Further, in order to satisfy the intracellular steady-state requirements for FBA, models are typically allowed to have free extracellular metabolite exchange out of the system. This practice was used for the FBA models in this study. Still, the optimization algorithms for node relaxation were required and were used in the same manner as that for uFBA.

Markov chain Monte Carlo (MCMC) sampling and gene essentiality

The FBA and uFBA models were sampled using Markov chain Monte Carlo (MCMC) sampling methods [9]. We sampled 5,000 points for the RBC model and 10,000 points for the platelet, *S. cerevisiae*, and *E. coli* models; all models were sampled until the mixed fraction was below a threshold of 0.54. The sampling distributions for each reaction were deemed significantly different between modeling formulations (uFBA vs FBA) if the two distributions overlapped by less than 5%. The correlations which are represented as histograms in Fig. 2b of the main text were calculated using the Spearman correlation of two sampled flux vectors. The process was repeated 5,000 times for the RBC model and 10,000 times for the platelet, *S. cerevisiae*, and *E. coli*

models to account for all generated sample points. For all determining significant differences and Spearman correlations (including controls), only reactions that were in both models and were not involved in Type III loops were compared.

Gene essentiality predictions were completed with COBRA 2.0 [9]. The threshold for growth was set to 1% of the growth rate experimentally determined in the publication of the metabolomics dataset [159]. All bounds and constraints that conflict with the zero vector being a feasible solution were modified to allow for proper simulation. Metabolic exchange out of the system for the uFBA model was also allowed. Computational predictions were compared to four data sets of glucose minimal media in three publications [166–168].

^{13}C labeling citrate experiment

A healthy donor was recruited and, after obtaining informed consent, red cells were obtained using ISO 9001:2008 certified red cell isolation protocols at the Landspitali-University Hospital Iceland Blood Bank. The study and all associated methods were approved by The National Bioethics Committee of Iceland and the Icelandic Data Protection Authority. The methods were carried out in accordance with the guidelines and regulations outlined by those committees.

Isolated red cells were added to 65 ml of modified Citrate Phosphate Dextrose (CPD) solution containing 105 mM uniformly labeled ^{13}C citrate (78% m+6, 1.5% m+5, 20.5% m+0) in place of unlabeled citrate in regular CPD SAGM media. Sodium hydroxide was use to match ionic strength imposed by trisodium citrate in regular CPD SAGM (Fenwal, Lake Zurich, IL, USA). As a negative control, red cells were stored in unmodified CPD SAGM. RBC units were stored at 4°C. 5 mL samples were removed at ten time points and aliquoted accordingly for

subsequent quality control assessment and metabolomics analysis. The experiment was run for 31 days, as the shift in metabolic dynamics from State 2 to State 3 occurs much earlier (day 17).

Red cell blood banking quality control assessment was performed immediately by ABL90 FLEX blood gas analyzer (Radiometer, Copenhagen, Denmark) determining pH, pO2, pCO2, total hemoglobin, [K^+], [Na^+], [Ca^2+], [Cl^-], [Glucose], and [Lactate]. A XN-1000 hematology analyzer (Sysmex, Norderstedt, Germany) was used to record RBC count, white blood cell count, platelet count, hemoglobin, % hematocrit, total mean cell hemoglobin, mean cell hemoglobin concentration, and red cell distribution width. Labeled samples were found to have similar QC properties as control units.

Metabolomic analysis was performed using a previously reported method [89, 175] based on ultra high performance liquid chromatography (UHPLC) (Acquity, Waters, Manchester, UK) coupled with a quadrupole-time of flight mass spectrometer (Synapt G2, Waters). Chromatographic separation was achieved by hydrophilic interaction liquid chromatography (HILIC) using an Acquity amide column, 1.7 μm (2.1 $\times$ 150 mm) (Waters).

500 μL of RBC sample were used for metabolomics analysis and supernatant and cells were separated by centrifugation (1600g, 4°C, 15 min). Immediately after centrifugation, cell-free supernatant was removed from centrifuged tubes and collected in separates tubes and processed as previously described [89, 175].

Data integration of targeted compounds was achieved by using TargetLynx (Waters). Raw data was then corrected for natural abundance of ^{13}C isotopes using IsoCor (MetaSys, Toulouse, France) [176] that afforded the % isotopes that exceed natural abundance and the corrected isotopic distribution.

Metabolic Flux Analysis

Metabolic flux analysis (MFA) was completed using the INCA software suite [161] in order to compare the measured isotopic labeling patterns to the flux predictions of the uFBA and FBA models. An isotopic model was constructed for the MFA simulations in which all reactions were split into irreversible reactions and metabolic reactions not directly related to citrate metabolism or production of the labeled metabolites were omitted. For uFBA associated MFA simulations, accumulation or depletion of metabolites was modeled using sinks. If a labeled intracellular metabolite was decreasing, the associated sink was labeled at the percentage of the initial labeling pattern for that particular state.

The INCA software suite allows for two basic types of simulations: a "forward" simulation in which the labeling pattern is predicted for a given flux state, and a "reverse" simulation in which the flux state of the network is calculated based on fitting the experimentally measured labeling pattern. The traditional "reverse" MFA simulation cannot be used when both metabolite levels are changing and the labeling pattern is unsteady. Instead, the "forward" simulation was used, which takes as input: (1) measured intracellular metabolite concentrations, (2) the labeling pattern of the metabolites, and (3) the flux state of the network. The mean flux state of the network as calculated by MCMC sampling for each of the metabolic states for each of the uFBA/FBA models was used as the initial flux state labeling pattern on Day 1.

The "forward" MFA simulations were completed for the first two states, where the labeling pattern approaches steady-state. First, the initial labeling pattern (Day 1) and the estimated uFBA fluxes were inputted and the labeling pattern was simulated for the duration of State 1. Next, the final predicted labeling pattern and the uFBA fluxes for State 2 were simulated for the duration of State 2. The labeling pattern reached steady-state by State 3 and is not included as

there were little differences between uFBA and FBA. The same workflow was also applied to the FBA fluxes for comparison.

To assess goodness of fit between the measured data and the simulated data, the residual sum of squares (RSS) was calculated for each of the four metabolites for which we had absolute concentrations and labeling data (lactate, glutamate, malate, and citrate). The RSS was calculated by

$$\sum_{i=1}^{m}(y_i - f(x_i))^2 \tag{4.14}$$

where y is the experimental value, $f(x)$ is the tracer simulation predicted value and the range of i contains all measured time points and each replicate is treated as a separate value. We evaluated the relative difference in RSS between the uFBA and FBA simulations to demonstrate the difference in accuracy of predictions. All RSS calculations were done with $\log_{10}$ transformed abundances.

Comparison to experimental flux states

Candidate flux states for the uFBA and FBA models were determined using MCMC sampling. The mean sampling vector was compared to measured flux values for the XR strain growth on glucose and xylose [164] and for the XI strain growth on glucose [165]. Though the experimental conditions were not identical, comparisons were made when the yeast growth rate were similar across the respective experimental conditions. Fluxes from metabolic reactions involved in type III pathways that are known to result in erroneous flux predictions were omitted for the comparison to measured data. In order to evaluate the error of the predicted flux states, we calculated the normalized Euclidean distance given by:

$$\text{error} = \frac{||v_{\text{meas}} - v_{\text{pred}}||}{||v_{\text{meas}}||} \tag{4.15}$$

where v_{meas} is the flux vector from MFA measurements and v_{pred} is the flux vector from the uFBA and FBA models.

4.3 Conclusions

To this point, we have investigated an additional perturbation to the storage conditions (temperature variation) and used both statistical and mechanistic models to understand the RBC metabolic network. We have found that the utility of these storage-age biomarkers transcends their ability to distinguish among the three metabolic phases. We recently showed that the concentrations of these extracellular metabolites at a particular time point can be used to quantitatively predict the concentration of other metabolites in the network [82]. Additional follow up work demonstrated that certain combinations of these biomarkers (based on their location within the metabolic network) could be used to forecast the future values of other metabolites in the network [83]. The identification of these robust biomarkers is further important from a practical standpoint: they reflect the fact that the metabolic decay process is fairly invariant under the conditions examined, thus revealing the inherently low dimensionality of the dynamic decay process.

The data provided by this baseline study has proven to be quite useful beyond a basic characterization of storage. Since the baseline metabolomics data are absolutely quantified, they can be integrated into mechanistic, cell-scale models capable of making quantitative predictions [10, 52]. These models predicted that the large 2,3-DPG pool is utilized to generate ATP using 2,3-DPG as a proton buffer through the reversal of bisphosphoglycerate mutase. This hypothesis is quite important, as the catabolism of 1,3-DPG generates two ATP, while the expected dephosphorylation of 2,3-DPG to 3PG generates only one ATP. Given that the initial

2,3-DPG pool is high, the shift in degradation route has a large influence on the overall ATP generation during storage. The quantitative model also made predictions about the metabolic fate of citrate, a compound added to the storage medium as an anticoagulant during blood collection. A follow-up study experimentally validated these predictions, namely the fact that citrate is used to produce lactate and glutamate [52]. Thus, quantitative metabolomics data enables the identification of key changes in metabolic pathway usage.

Where are we now, and where do we go from here?

The next step for the community is to continue this exceedingly productive phase of "normal science," further characterizing different perturbation conditions using the systems biology approach outlined here to better understand red cell physiology. Other groups have begun to investigate the metabolism of high altitude stored blood [177–179], including how well blood from high altitude donors stores [180]. The studies outlined here show how robust the metabolome of the red cell is during storage; we see the same decay of the metabolome in different storage media and additives and at different temperatures.

As more new data types are generated, there is an increasing need for accessible data analytics that can concisely capture and extract information from the integration of disparate data types. The incorporation of systems biology principles into standard workflows allows us to gather meaningful knowledge from disparate data types. This knowledge can then be used in conjunction with mechanistic cell-scale *in silico* models to develop hypotheses that can subsequently be tested experimentally (Figure 1).

The transfusion community must therefore work in parallel with these experimental efforts to build new models that integrate new data types and can explain some of the anomalous

behaviors that current models cannot explain. In particular, we need to elucidate the governing constraints on the system. Much of the data available suggests that 2,3-DPG and ATP levels have the most influence on the system. D'Alessandro and colleagues combined metabolomics and proteomics analyses to elucidate the role of glyceraldehyde 3-phosphate dehydrogenase (GAPDH) in the third metabolic state [181]. Similar efforts have begun to explore the lipidome in great detail, providing further insights into the multiscale complexities of RSL [182].

Orthogonal data types will be included in future model analyses. Systems biologists need to start including information about the proteome to further inform and grow existing models, a feat that has already been accomplished in other organisms [183–186]. Recently, Bryk and Winiewski used quantitative proteomics to identify over 2,600 proteins in RBCs [187]. Rich, quantitative data sets such as this one have previously been used to calibrate computational models of bacteria [188]. Such comprehensive -omics studies will allow for the integration of this data into mechanistic models of the RBC. In addition, protein structures and their genetic bases will find their way into computational models [189].

Acknowledgements

The following authors contributed to the research on enzyme regulation: JT Yurkovich and BO Palsson conceived the study; JT Yurkovich, MA Alcantar, and ZB Haiman performed the analysis; JT Yurkovich, MA Alcantar, and BO Palsson wrote the manuscript. This work was supported by the Galletti Chair Funds and the Genentech Foundation Scholars Program (MAA). The authors gratefully acknowledge Nathan Mih, Laurence Yang, and Bin Du for valuable discussions.

The following authors contributed to the research on constraint-based modeling: A Bor-

dbar and JT Yurkovich completed all analyses; G Paglia, O Rolfsson, and ÓE Sigurjónsson completed the isotopic-labeling experiment; A Bordbar, JT Yurkovich, and BO Palsson wrote the manuscript; A Bordbar and BO Palsson conceived of the study. This work was supported by the National Heart Lung and Blood Institute (R43HL123074 and R43HL127843), the European Research Council (232816), and the U.S. Department of Energy (DE-SC0008701). The authors would like to thank A D'Alessandro for providing the validation metabolomics data for the relaxation optimization algorithm and D McCloskey, DC Zielinski, A Ebrahim, JM Monk, and NE Lewis for valuable discussions.

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

- **JT Yurkovich***, MA Alcantar*, ZB Haiman, and BO Palsson. "Network-level allosteric effects are elucidated by detailing how ligand-binding alters the catalytic potential." Submitted January 2018 (Under review, *PLOS Computational Biology*). The dissertation author was one of the two primary authors.

- A Bordbar*, **JT Yurkovich***, G Paglia, O Rolfsson, ÓE Sigurjónsson, and BO Palsson. 2017. "Elucidating metabolic physiology." *Scientific Reports*, 7 (46249). The dissertation author was one of the two primary authors.

- **JT Yurkovich**, A Bordbar, ÓE Sigurjónsson, and BO Palsson. 2018. "Systems biology as an emerging paradigm in transfusion medicine." *BMC Systems Biology*, 12:31. The dissertation author was the primary author.

Chapter 5

Toward a Whole-Cell RBC Model

Integrating -omics data to refine or make context-specific models is an active field of constraint-based modeling. In Chapter 4, we presented an algorithm for the integration of quantitative time-course metabolomics data into constraint-based models. While this method will expand the utility of metabolomics data for mathematical models, we are still faced with the challenging of integrating other disparate -omics data types into mechanistic models. Recently, the most comprehensive quantitative proteomics data set for the RBC was published [187], providing the opportunity to use these data to further improve our models of the RBC. The first step toward using proteomics data is devising a mathematical formalism for their inclusion in a model. In this Chapter, we use *E. coli* as a test case because of the availability of data under multiple conditions and the existence of multi-scale models.

Proteomics now cover over 95% of the *E. coli* proteome by mass. Genome-scale models of Metabolism and macromolecular Expression (ME) compute proteome allocation linked to metabolism and fitness in *E. coli*. Together, the availability of quantitative proteomics data and models with proteomics constraints integrated make *E. coli* the ideal system in which to test and

understand the integration of absolutely quantified proteomics data.

Defining a core functional proteome supporting the living process has importance for both developing fundamental understanding of cell functions and for synthetic biology applications. Comparative genomics has been the primary approach to achieve such a definition. Thus, we used genome-scale models to define a core proteome that computationally supports basic cellular function. This core proteome for metabolism and protein expression, defined through systems biology methods, is validated and characterized by using multiple disparate data types.

While the transcriptional regulatory network of *E. coli* has expanded considerably in recent years through new chromatin immunoprecipitation (ChIP) data, an open question remains: does the global transcriptional regulatory network, reconstructed by combining ChIP data for individual transcriptions factors, consistently explain observed differential gene expression? We have reconstructed a high-confidence TRN, determined its consistency with transcriptomics and predictive capabilities across multiple conditions, extracted 10 functional regulatory modules, and characterized this network at the sequence and structural levels. Our multi-omics algorithmic pipeline is expected to facilitate rigorous validation and prioritization of experiments to elucidate transcriptional regulatory networks in other bacteria.

Following these basic characterizations, we used proteomics data to formulate allocation constraints for key proteome sectors in the ME model. The resulting calibrated model effectively computed the "generalist" (wild-type) *E. coli* proteome and phenotype across diverse growth environments. Across 15 growth conditions, prediction errors for growth rate and metabolic fluxes were 69% and 14% lower, respectively. The sector-constrained ME model thus represents a generalist ME model reflecting both growth rate maximization and "hedging" against uncertain environments and stresses, as indicated by significant enrichment of these sectors for the general

stress response sigma factor σ^S. Finally, the sector constraints represent a general formalism for integrating -omics data from any experimental condition into constraint-based ME models. The constraints can be fine-grained (individual proteins) or coarse-grained (functionally-related protein groups) as demonstrated here. This flexible formalism provides an accessible approach for narrowing the gap between the complexity captured by -omics data and governing principles of proteome allocation described by systems-level models.

5.1 Using *E. coli* as a model

Genome-scale models have been used to conduct systems-level studies of cellular metabolism for over 15 years [190]. They can elucidate structures in large datasets that are not captured by purely statistical models [10]. In particular, the COnstraint-Based Reconstruction and Analysis (COBRA) field has a rich history of using -omics data to refine and improve predictions [191–193]. These methods have been useful for the systems biology community. For example, tissue-specific models could be generated for health applications [194] or computational strain design could be improved for metabolic engineering [195]. Despite many methods for -omics integration in COBRA, the general problem of relating gene expression to metabolic flux and cell physiology remains challenging. One challenge has been that metabolic models only indirectly relate expression to flux. ME (Metabolism and macromolecular Expression) models [183,184,196,197] now relate gene and protein expression directly to metabolic flux. Therefore, in theory it is possible to integrate transcriptomics and proteomics data directly into COBRA to refine predictions. However, the ME model is multiscale and spans multiple cellular processes including metabolism and protein expression machinery. The latest ME model [197] models the function of 1,678 genes described by nearly 80,000 reactions and 70,000 constraints involving bio-

chemical species and macromolecules spanning nearly 70 cellular subsystems. Therefore, it is not obvious how changes to one part of the system affect others. Specifically, it is not obvious how expression changes for multiple proteins will quantitatively affect growth rate or metabolic fluxes. As a first step, a recent study shows that growth rate predictions are indeed markedly improved when the overall fraction of unused protein (i.e., expressed but not actively used) is constrained using estimates from proteomics data [198]. A remaining question then is whether constraining specific functional protein groups can also improve growth and metabolic flux predictions.

Schmidt et al. recently published a proteomics data Resource [199] covering ∼55% of the predicted *E. coli* genes (> 95% by mass) under 22 experimental conditions. Using genome-scale models, we show how such proteomics resources can be used to reveal principles underlying proteome allocation. ME models compute growth optimal proteomes consistent with laboratory evolved strains accurately [200], but are unable to compute processes that are not directly related to growth (e.g., stress response, preparation for unfavorable conditions) [201]. In anticipation of environmental change, generalist (wild-type) *E. coli* allocate a fraction of the proteome to non-growth related functions. Collectively, such allocation can be viewed as "hedging" against unknown enviromental challenges [200], reflecting the evolutionary history of the organism and its successful survival strategy. Recent studies have estimated that 20% of the expressed proteome confers no direct fitness benefit [201].

Elucidating trade-offs between multiple cellular objective is an active field of systems biology research [202, 203]. Here we develop a pragmatic approach for modeling the proteome allocation resulting from such complex cellular objectives following in the spirit of -omics integration with genome-scale models. Namely, we define sector constraints using proteomics data. We then show that sector-constrained ME models can compute the proteome composition of

generalist *E. coli* in a variety of different growth environments. We then compare the "optimal" versus the "generalist" proteomes to reveal principles underlying proteome allocation.

5.1.1 Results

Modeling generalist *E. coli* proteome allocation and physiology

We first computed optimal proteome allocation that maximized growth rate. Measured proteome allocation differed notably from these computed optimal proteomes. While the interactions between thousands of individual proteins is complex, the *E. coli* proteome has been shown to exhibit relatively simple relationships when proteins were grouped into meaningful "sectors" (e.g., linear relations with growth rate) [204]. Similarly, we coarse-grained the proteome into functional sectors. Here, we specifically used Clusters of Orthologous Groups (COGs) [205] as they represent a reasonable trade-off between complexity (24 sectors) and protein function coverage.

We then identified proteome sectors whose measured mass fractions were greater (over-allocated) and smaller (under-allocated) compared to the optimal proteomes across growth conditions (Fig. 5.1a). For our analysis we focused on the 15 minimal medium growth conditions under batch and chemostat culture without additional pH, osmotic or temperature stresses. Six COG sectors had high measured mass fractions (5% or more) under all 15 conditions. The minimum mass fraction across conditions for these COGs ranged from 5.4% to 16%, totaling 58% (Table 5.1). Meanwhile, the optimal proteomes allocated between 13% to 54% of proteome to the sum of these sectors (Fig. 5.1a). Additionally, the computed growth rates corresponding to optimal proteomes were consistently higher than measured (Fig. 5.1b).

We hypothesized that by constraining the ME model to more accurately allocate proteome

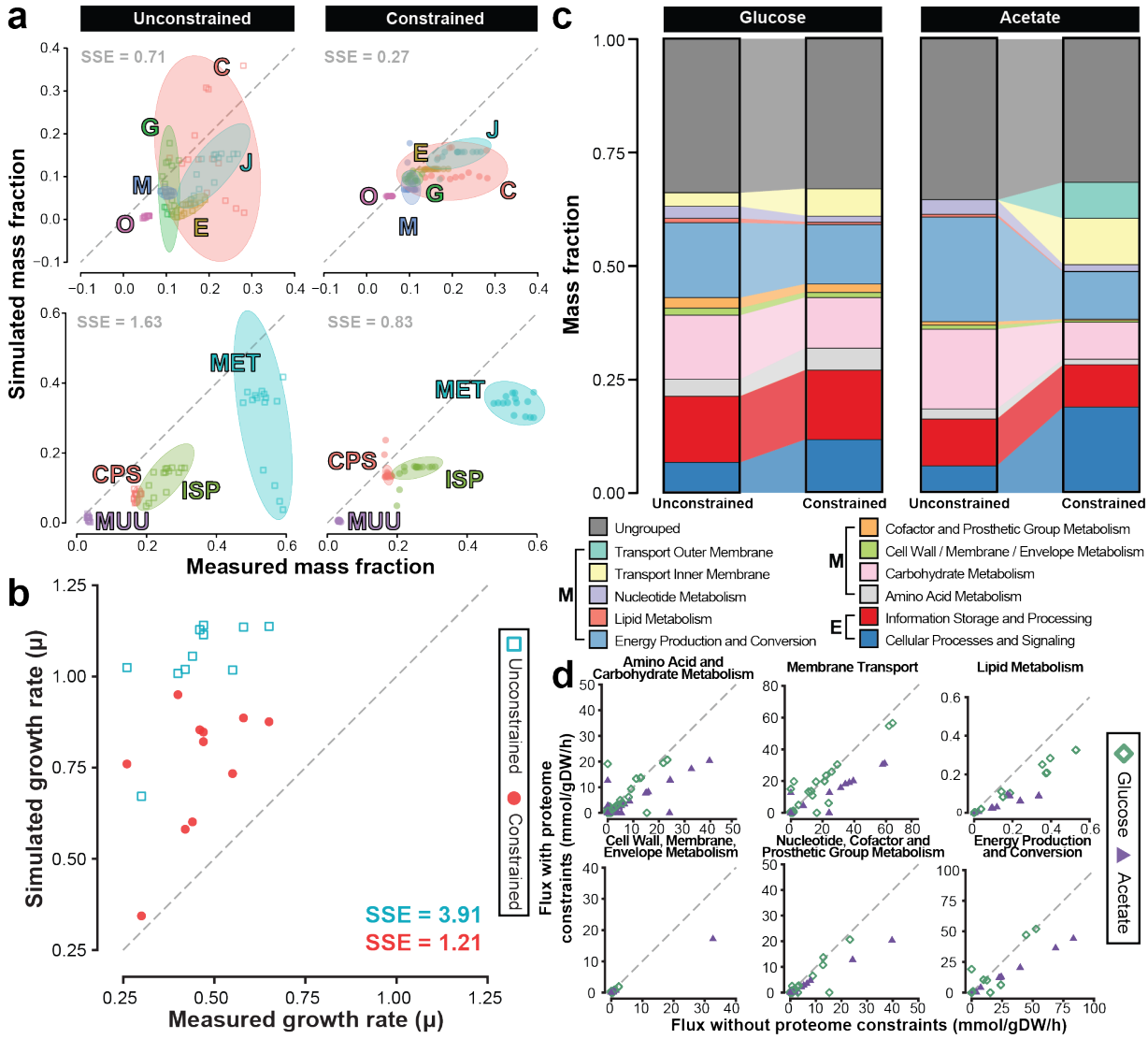

Figure 5.1: Model-based interpretation of proteomic data. (**a**) Predicted mass fractions are validated by proteins grouped by COG for optimal and generalist ME models. Ellipses show 95% confidence intervals. (**b**) Growth rate predictions improve due to proteome sector constraints. (**c**) The model predicts global proteome reallocation due to sector constraints for Metabolic (M) and Expression (E) systems. (**d**) The ME model computes global metabolic shifts due to proteome reallocation. Abbreviations: C, Energy production and conversion; E, Amino acid transport and metabolism; G, Carbohydrate transport and metabolism; J, Translation, ribosomal structure and biogenesis; M, Cell wall/membrane/envelope biogenesis; O, Posttranslational modification, protein turnover, chaperones; CPS, Cellular Processes and Signaling; ISP, Information Storage and Processing; MET, Metabolism; MUU, Mobilome, Unknown, and Ungrouped; SSE, sum of squared errors.

Table 5.1: Proteome sector constraints. Sectors and their mass fractions defined from proteomics data and used to constrain the ME model.

Sector	Mass fraction
Amino acid transport and metabolism	0.115
Carbohydrate transport and metabolism	0.089
Cell wall/membrane/envelope biogenesis	0.068
Energy production and conversion	0.096
Posttranslational modification, protein turnover, chaperones	0.054
Translation, ribosomal structure and biogenesis	0.156

to these large sectors, we could better predict growth rate and metabolism of the wild-type, generalist *E. coli*. To this end, we added new "sector constraints" to the ME model (constraint 5.5 in Methods). Here we constrained the sum of protein mass fractions within each of the six sectors; however, the formulation is general in that any individual protein or different sector definition (coarser or finer-grained) can be used. It is important to note that while our sector constraints involved 966 of the 1678 genes in the iJL1678 ME model, only six actual constraints were added to the model—only the sum of mass fraction of each sector was constrained. Therefore, the individual protein mass fractions were still computed by the ME model. Because our objective was to develop a generalist ME model, we used these coarse-grained sector constraints to prevent over-fitting the proteome to specific conditions.

The addition of the six sector constraints led to markedly improved agreement in growth rate predictions across all conditions, with 69% lower sum of squared error (SSE), overall (Fig. 5.1b). We also compared measured and computed proteome allocation for a functional grouping of proteins that was different from the COGs used for sector constraints. The SSE was

49% lower for proteome allocation (Fig. 5.1a).

We thus designated this sector-constrained ME model the Generalist ME model. Because the total proteome is limited in size, the increased allocation to certain sectors would lead to decreased resource allocation to at least one other sector. Thus, the sector constraints reflect costs of cellular decisions to over-allocate proteomic resources for purposes other than maximal growth on a minimal medium. For example, the COG categories "carbohydrate transport and metabolism," "energy production and conversion," and "translation, ribosomal structure and biogenesis" were enriched for control by the stress response sigma factor σ^{S} and could reflect preparation for unfavorable conditions [201].

Proteomic and metabolic consequences of proteome sector constraints

Next, we examined the resource allocation of each computed proteome by metabolic (M) and macromolecular expression (E) model subsystem (Fig. 5.1c). Note that these (M) and (E) subsystems differ from the gene categorization used to define proteome sector constraints (i.e., COGs) in Table 5.1. Allocation to membrane transport proteins (specifically amino acid and carbohydrate transport and metabolism) was increased according to the sector constraints, although they had no fitness benefit for growth on acetate according to the optimal model. These sectors included genes controlled by the stress response sigma factor σ^{S}. These sector constraints thus resemble foraging and stress response that intensifies in lower-quality substrates such as acetate [206]. The generalist acetate proteome was lowered in "energy production and conversion." This sector showed considerable decrease in enzymes catalyzing acetate consumption (acetate kinase, phosphotransacetylase) and energy metabolism (cytochrome oxidase bo3, ATP synthase), leading to a decreased growth rate. The optimal proteomes represent the minimal

proteomic resources needed to grow at sub-optimal growth rates. The extent to which *E. coli* allocates its proteome beyond the minimum required was surprisingly high: a growth rate of 0.12 h^{-1} could theoretically be supported with 95% less proteome allocated to all M and E sectors than measured.

The sector constraints also altered metabolic flux distributions (Fig. 5.1d). Specifically, proteome constraints induced statistically significantly smaller ($P < 0.05$) fluxes in 5 of 8 of the subsystems for glucose (Lipid Metabolism;, Cell Wall/Membrane/Envelope Metabolism; Nucleotide, Cofactor and Prosthetic Group Metabolism; Amino Acid and Carbohydrate Metabolism; Other), and all 8 metabolic subsystems for acetate.

The sector constraints also affected predicted protein fraction of cell dry weight. Compared to the Optimal model, the Generalist model predicted 5.5% to 24.8% higher protein fraction (Table 5.2), which was significant (Wilcoxon rank sum test, $P = 6.4 \times 10^{-9}$). Interestingly, the Generalist model showed a clear and significant negative correlation between total proteome size and growth rate, whereas the Optimal model did not). This linear trend in the Generalist model arises because many of the sector constraints force expression of unused protein [198] and is consistent with previously observed growth rate-dependent decrease in constitutively expressed proteins [207].

In addition, the Generalist model's proteome size of 72.7% dry weight at dilution rate of 0.12 h^{-1} agrees well with the measured value of 70.1% for this dilution rate [208]. However, in unlimited growth on glucose, the Generalist model overestimates proteome size (62.4% versus 51.1% measured [208]). This result is due to the sector constraints being defined based on multiple carbon sources, whereas *E. coli* is more adapted for growth on its preferred substrate, glucose [209].

Table 5.2: Predicted protein percent of cell dry weight.

Condition	Optimal		Generalist	
	Growth rate, h^{-1}	Protein, % dry weight	Growth rate, h^{-1}	Protein, % dry weight
Acetate	0.672	60.1	0.344	70.0
Chemostat.mu.0.12	0.120	58.2	0.120	72.7
Chemostat.mu.0.20	0.200	58.0	0.200	71.0
Chemostat.mu.0.35	0.350	57.9	0.350	68.8
Chemostat.mu.0.5	0.500	58.1	0.500	66.7
Fructose	1.140	57.1	0.876	62.5
Fumarate	1.020	59.4	0.581	67.5
Galactose	1.020	57.8	0.760	62.8
Glucosamine	1.130	57.0	0.854	62.7
Glucose	1.140	57.4	0.886	62.4
Glycerol	1.140	57.2	0.821	63.4
Mannose	1.110	57.2	0.848	62.7
Pyruvate	1.010	58.7	0.950	61.9
Succinate	1.060	59.6	0.601	67.7
Xylose	1.020	57.9	0.734	63.6

Validation of intracellular fluxes

We next validated intracellular flux predictions for the Optimal and Generalist models using metabolic flux analysis (MFA) data by Gerosa et al. [210] for 7 carbon sources: acetate, fructose, galactose, glucose, glycerol, pyruvate, and succinate. The Generalist model was more consistent with MFA for 5 of 7 conditions. In particular, acetate, succinate and glycerol predictions improved greatly, with 68%, 41%, and 15% lower RMSE (root mean squared error), respectively. RMSE was higher for the Generalist model for glucose and galactose conditions (8.0% and 30% higher RMSE, respectively). However, when we performed similar validation using a different MFA data set [211], we observed slightly (6.5%) lower RMSE on glucose and slightly (6.4%) higher RMSE for galactose. This discrepancy between MFA data sets partially arises because the MFA data relied on simplified models of central carbon metabolism, whereas the ME model considers the genome-scale metabolic network. The sector constraints most affected TCA cycle fluxes, with the Generalist model fluxes decreasing between 100% to 16.8%

across conditions. Therefore, respiratory capacity was predicted to be most strongly affected by allocating proteome toward hedging functions.

Sensitivity of predictions to parameter uncertainty

Due to proteome constraints, ME models exhibit essentially no flux variability at the proteome level at maximum growth rate [212]. However, optimal solutions can vary due to uncertainty in effective rate constants. Thus, we next assessed how sensitive growth rate and protein allocation predictions were to uncertainty in effective rate constants (k_{eff}). To this end, we randomly perturbed effective rate constants by $\pm 50\%$ of their nominal values and maximized growth rate. Due to the considerable computational burden of many ME simulations, we limited the sensitivity analysis to glucose and acetate conditions. We ended up with 80 simulations that were still feasible after perturbations. We found that on average, the sensitivity to k_{eff}uncertainty was greatest at the proteome mass fraction level, decreased for metabolism, and was smallest for the growth rate. For protein mass fractions, the coefficient of variation (CV) of proteins across randomly perturbed simulations ranged between 0.073 to 3.9 (median of 0.31) and 0.086 to 8.9 (median of 0.62) for the Optimal and Generalist models, respectively (Fig. 5.2a). The Optimal model was significantly less sensitive to k_{eff}perturbations (Wilcoxon rank sum test, $P < 2.2 \times 10^{-16}$). Metabolic fluxes showed similar variability with CVs ranging from 0.29 to 4.8 (median of 0.38) and 0.28 to 9.7 (median of 0.61) for the Optimal and Generalist models, respectively (Fig. 5.2b). Again, the Optimal model was less sensitive to k_{eff}perturbations (Wilcoxon rank sum test, $P = 1.0 \times 10^{-5}$). Finally, growth rates varied with median CVs of 0.28 and 0.31 for Optimal and Generalist models, respectively (Fig. 5.2c), which was not significantly different between the two models (Wilcoxon rank sum test, $P = 1$). Therefore, we conclude that while uncertainty in

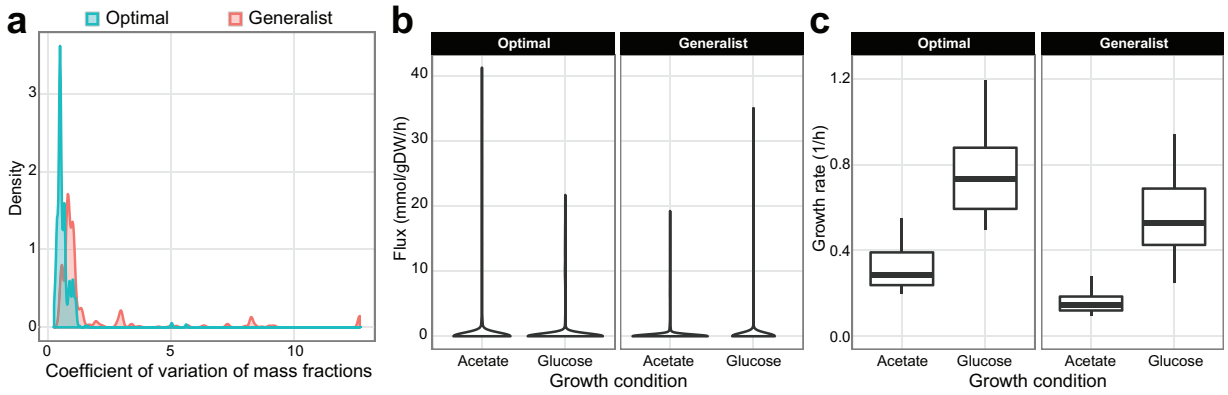

Figure 5.2: Sensitivity to model parameters. (**a**) Probability density of the coefficients of variation of simulated protein mass fractions across random perturbations of effective rate constants. (**b**) Variation in simulated metabolic fluxes upon perturbing effective rate constants. (**c**) Variation in simulated growth rate upon perturbing effective rate constants.

k_{eff}can lead to wide variability in the expression of individual proteins, the effects are attenuated for metabolic fluxes and eventually growth rate.

5.1.2 Discussion

One of the primary uses of ME models in previous studies has been to model optimal *E. coli* strains [213]. In particular, strains have been evolved in the laboratory while under environmental pressures designed to select for mutations that optimize for growth rate [214]. Such strains are highly useful in metabolic engineering applications where the goal is to produce a particular product efficiently [13]. However, the objective of all organisms is not necessarily to maximize growth rate. Many non-evolved laboratory strains or even pathogens have objectives that require allocation of proteomic and cellular resources to more than just growth rate [198,203]. To this end, we constrained a ME model using proteomics data collected under various conditions in order to better predict sub-optimal proteome allocation of "generalist" *E. coli*.

The ME model indicated that at growth rates as low as 0.1 h^{-1}, up to 95% of the generalist proteome was not beneficial for growth. Much of the apparently wasted proteome was

related to general stress response and "hedging" against environmental uncertainty. By integrating proteomics data into a ME model, we revealed the cellular cost of dedicating resources to maintaining this generalist proteome. We showed that proteomics data can be used to identify key proteome sector constraints and calibrate ME models. This approach can be extended in future work using sectors other than COGs, or determining novel sectors using the ME model itself. In particular, here we defined sector constraints capturing broad trends across many conditions, rather than fitting individual conditions. Yet, nearly all of the 15 minimal media examined showed improved predictions in terms of proteome allocation, growth rate, and metabolic fluxes. In future work, the sector constraints may be extended to include known regulation. For example, transcriptional regulation of carbon transport and utilization pathways has been well-studied for phosphotransferase system (PTS) sugars [215] and non-PTS sugars [203]. Thus, it may be possible to combine regulatory models (e.g., by cyclic AMP-CRP) with ME models to model dynamic sector constraints in response to environmental changes such as carbon source availability or other environmental signals.

Further, we believe that the efforts here provide a framework moving forward for using novel data sets to tailor models to represent cellular objectives other than maximum growth rate. For example, with the various growth conditions in the data provided by Schmidt et al. [199], we were able to place constraints on the proteome that allowed us to build a model for a generalist organism prepared for unfavorable conditions, as suggested by the significant enrichment of the constrained proteome sectors for the general stress response sigma factor σ^{S}. Thus, new data sets that describe other experimental conditions or environmental pressures could be used to place additional constraints on the proteome using a ME model. Such experiments might measure the proteome across multiple nutrients under various stresses such as low or high pH [216], iron

limitation [217], or exposure to reactive oxygen species [218]. This proteomics data could then be used to constrain the ME model's stress response, thereby incorporating stress response into the objective function. Furthermore, with the increasing coverage and resolution of transcriptional regulatory interactions from Chromatin ImmunoPrecipitation (ChIP) experiments, it may ultimately be possible to integrate regulatory networks to proteome re-allocation. Combining data from proteomics, ChIP, metabolic flux analysis, physiological (growth and exchange rates) and potentially metabolomics (at least metabolites involved in regulation) would provide the prerequisites for mechanistically modeling proteome allocation according to complex and multi-faceted cellular objectives.

One of the biggest challenges facing systems biology is to integrate new data types into genome-scale mathematical models to provide biologically meaningful phenotypic predictions. ME models provide mechanistic understanding of how metabolic flux states are linked to protein expression. Integrating data into a structured framework thus leads to an improved understanding of systems-level properties of organisms, suggesting a combined experimental and modeling approach to meet the "Big Data to Knowledge" challenge.

5.1.3 Methods

Simulating growth maximization using ME models

We used the latest published ME model of *E. coli* (iJL1678) [197] for simulations consisting of nearly 80,000 biochemical reactions describing metabolism, transcription, and translation processes. To maximize growth rate in each growth condition, we used bisection (binary search) as in [184] to maximize growth rate to six decimal points. Because ME-Models are ill-conditioned [184,196,219], we used the 128-bit (quad-precision) linear and nonlinear program-

ming solver qMINOS 5.6 [212, 220, 221]. (The soplex solver [222] is another viable option, as it uses iterative refinement to achieve the needed numerical precision.) All qMINOS runs were performed with feasibility and optimality tolerances of 10^{-20}. These tight tolerances were necessary because ME fluxes can vary by 15 orders of magnitude [212].

Computing generalist proteome allocation using sector constraints

The generalist ME model includes "sector constraints" in addition to the standard ME model formulation. The complete formulation of the optimization problem associated with the "generalist" ME model is the following:

$$\max_{v,\mu,p} \quad \mu, \tag{5.1}$$

$$\text{s.t.} \quad Sv = 0, \tag{5.2}$$

$$g(\mu)Av + Bv = 0, \tag{5.3}$$

$$p = \sum_j w_j v_j, \qquad \forall j \in \text{Translation}, \tag{5.4}$$

$$\sum_{j \in \text{Sector}(k)} w_j v_j - \phi_k \cdot p \geq 0, \qquad k = 1, \ldots, n^{\text{sector}}, \tag{5.5}$$

$$l \leq v \leq u, \tag{5.6}$$

where $g(\mu)$ is a function of growth rate μ, p is the total simulated proteome mass, Translation is the set of translation fluxes, n^{sector} is the number of constrained sectors, Sector(k) is the index set of translation reactions in sector k, w_j is the molecular weight for protein (in g/mmol) j, v_j is the translation flux for protein j (in mmol/gDW/h), ϕ_k is the mass fraction of sector k, and v, l and u are the vectors of reaction fluxes, lower and upper flux bounds, respectively. "Optimal" ME models are formulated similarly, except without constraints 5.4 and 5.5.

Constraint 5.5 is the "sector constraint," which constrains the summed mass fraction of a proteome sector to reach the specified amount. In this case, we constrained each sector by an inequality so that allocation to the constrained sector was greater than or equal to the measured mass fraction. The constraint is derived from the relation,

$$\frac{\sum_{j \in \text{Sector}(k)} w_j v_j}{\sum_{j \in \text{Translation}} w_j v_j} \geq \phi_k \text{ (g/g)} = \frac{\text{Steady-state synthesis rate of proteome Sector } k \text{ (g/gDW/h)}}{\text{Steady-state synthesis rate of total proteome (g/gDW/h)}}, \quad (5.7)$$

noting that the translation fluxes v are the rates of reactions that synthesize protein. The macromolecule Expression (E) matrix in the ME model enables the explicit computation of protein synthesis rate.

Our formulation can also be considered a more generalized extension of work by O'Brien et al. [198], who determined a global parameter for the "un-utilized" and "under-utilized" protein expression using ME models. Here, we divided the proteome into functional sectors to a specified resolution or level of granularity and imposed constraints on several key sectors at this level.

Effective catalytic rate constants (k_{eff}) were kept identical for both the Optimal and Generalist ME models, and were the same values as the original iJL1678 ME model [197]. Recall that effective rate constants $k_{\text{eff}}$and fluxes v are related as $v = k_{\text{eff}} E$, where E is the enzyme concentration. Therefore, the maximum flux of a reaction is affected by proteome allocation across conditions.

Comparing computed versus measured growth rates and proteomes

Computed growth rates μ were compared with those measured by Volkmer and Heinemann [223]. Computed proteome mass fractions were compared with those measured by Schmidt

et al. [199]. Measured mass fractions were calculated using the measured protein masses (fg/cell). ME model mass fractions f_i were computed by weighting the translation flux (v_i in mmol/gram-dry-weight/h) of each protein by its molecular weight (m_i): $f_i = m_i v_i / \sum_{j \in Translation} m_j v_j$, where $Translation$ is the index set of translation reactions.

Computing proteome size

ME models compute the protein fraction of dry cell weight, P (grams protein/gram dry weight). Because total protein synthesized was diluted by cell division, we have

$$\text{Protein synthesized} = \sum_j w_j v_{\text{trsl},i} = P\mu = \text{Protein diluted},$$

where w_j is the molecular weight of protein j, $v_{\text{trsl},j}$ is the translation flux of protein j, and μ is the growth rate (h^{-1}). Therefore, $P = \sum_j w_j v_{\text{trsl},j}/\mu$,

Analyzing sensitivity to uncertainty in effective rate constants

We analyzed the sensitivity of proteome mass fraction, metabolic flux, and growth rate predictions to uncertainty in 1322 effective rate constants (k_{eff}). We (uniformly) randomly perturbed these k_{eff} by $\pm 50\%$ of their original values in the published iJL1678 model [197]. Each randomly perturbed model was used to simulate growth maximization with and without sector constraints, as described in Methods.

Enrichment analysis

Enrichment analysis was performed using hypergeometric test p-values, with $P < 0.05$ considered statistically significant.

Proteome sector constraints

In this study, we used proteomics data from 15 minimal media conditions [199] to define our sector constraints at the level of COGs [205]. The 15 conditions were the 4 chemostat conditions (dilution rates of 0.12, 0.20, 0.35, 0.50 h^{-1}), and 11 minimal media (acetate, fructose, fumarate, galactose, glucosamine, glucose, glycerol, mannose, pyruvate, succinate, xylose). We chose the six sectors having at least 5% mass fraction across all conditions. The constrained mass fraction for each sector (ϕ_i in constraint 5.5) was the minimum mass fraction across all conditions.

5.2 Conclusions

In this Chapter, we have developed a computational method for the integration of quantitative proteomics data into a genome-scale mechanistic model of *E. coli*. Future work will involve applying this type of framework to the RBC. The formulation is slightly different due to the absence of transcription and translation machinery in RBCs, thus requiring a modified model.

Acknowledgements

The following authors contributed to this work: L Yang and BO Palsson conceived the study; L Yang, CJ Lloyd and A Ebrahim conducted the experiments; L Yang and JT Yurkovich analyzed the results; L Yang, JT Yurkovich and BO Palsson wrote the manuscript. This work was funded by the National Institute of General Medical Sciences of the National Institutes of Health (awards U01GM102098 and R01GM057089), the US Department of Energy (DE-SC0008701),

the National Science Foundation Graduate Research Fellowship (DGE-1144086), and the Novo Nordisk Foundation [NNF16CC0021858]. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the US Department of Energy (DE-AC02-05CH11231).

Chapter 5 in part is a reprint of material published in: L Yang*, **JT Yurkovich***, CJ Lloyd, A Ebrahim, MA Saunders, and BO Palsson. 2016. "Principles of proteome allocation are revealed using proteomic data and genome-scale models." *Scientific Reports*, 6:36734. The dissertation author was one of the two primary authors.

Chapter 6

A New Paradigm Emerges: Systems Transfusion Medicine

The large-scale generation of -omic data holds the potential to increase and deepen our understanding of biological phenomena, but the ability to synthesize information and extract knowledge from these data sets still represents a significant challenge. Bottom-up systems biology overcomes this hurdle through the integration of disparate -omic data types, and absolutely quantified experimental measurements allow for direct integration into quantitative, mechanistic models. The human red blood cell has served as a starting point for the application of systems biology approaches and has been the focus of a recent burst of generated quantitative metabolomics and proteomics data. Thus, the red blood cell represents the perfect case study through which to examine our ability to glean knowledge from the integration of multiple disparate data types.

6.1 The old paradigm

Over the last two decades, the life sciences have witnessed a paradigm shift brought on by the development of high-throughput -omic technologies. With the advent of these technologies, systems biology emerged as a way to holistically integrate the new data being generated. Integrative thinking was not something previously absent in molecular biology, it was just that high-throughput -omic technologies were making the scale of these integrative inquiries much larger [224]. With the availability of full genome sequences and other data, more and more researchers embraced the promise in systems biology and began to develop ways to bridge the gap between -omic data and computational modeling efforts [225].

Some of the first cell-scale computational models were published in the late 1980s [125]. These enzyme kinetic models detailed the known metabolic network of the human red blood cell (RBC). Why study the RBC? The reasoning was simple: if systems biology cannot be successfully applied to the simplest human cell, then why attempt to study more complex ones? Indeed, simple systems are the best starting point for the application of systems biology. The RBC is therefore a logical starting point for the development and application of systems biology methods because of its simplicity and intrinsic experimental accessibility. RBCs are also of great importance for our understanding of human health and physiology—over 84% of all human cells by count are RBCs [15]. Transfusion medicine represents an integral part of healthcare, with approximately 85 million RBC units transfused worldwide annually [226]. The systems biology analysis of the health of stored RBCs is thus a productive focus from a basic and applied standpoint.

Within the last several years, -omic technologies have been exploited to study RBCs under refrigerated storage for use in transfusion medicine [25, 227] in an attempt to understand and elucidate the underlying physiological changes that occur because of the artificial environ-

ment [20]. Concurrently, computational biologists have worked to develop new mathematical modeling frameworks that can use these data. Because of the inherent quantitative nature of these models, however, their utility is only fully realized with quantitative data. In this context, quantitative data implies the use of standards to absolutely quantify the abundance of measured species; the output is data with quantified units (e.g., g/L, mM), rather than qualitative data that have relative units (e.g., arbitrary units, relative signal). While there have been several important studies that have used qualitative data effectively, the future of systems biology modeling efforts will hinge upon the availability of high quality quantitative data.

6.2 A new paradigm emerges

Recent work in -omic data generation and corresponding computational methodologies have begun to move the field forward. The use of quantitative data aids modeling efforts and enables new questions to be asked. There are several -omic data types and new experimental techniques that will likely prove to be valuable for the field. As these experimental technologies are developed, a variety of computational modeling approaches that integrate these data types and have been developed and contributed important advances.

A variety of -omic data types describe cellular physiology

A cell is a system of interconnected complex systems described by a variety of -omic data types [228]. Metabolomics data provide a snapshot of the cellular biochemistry that details energy production [229]. Fluxomics measurements—the use of isotopic tracers (e.g., ^{13}C)—yield an understanding of the flux state of a metabolic network [230]. Proteomics data allow for an understanding of the abundance, localization, and interactions of proteins, the cellular machin-

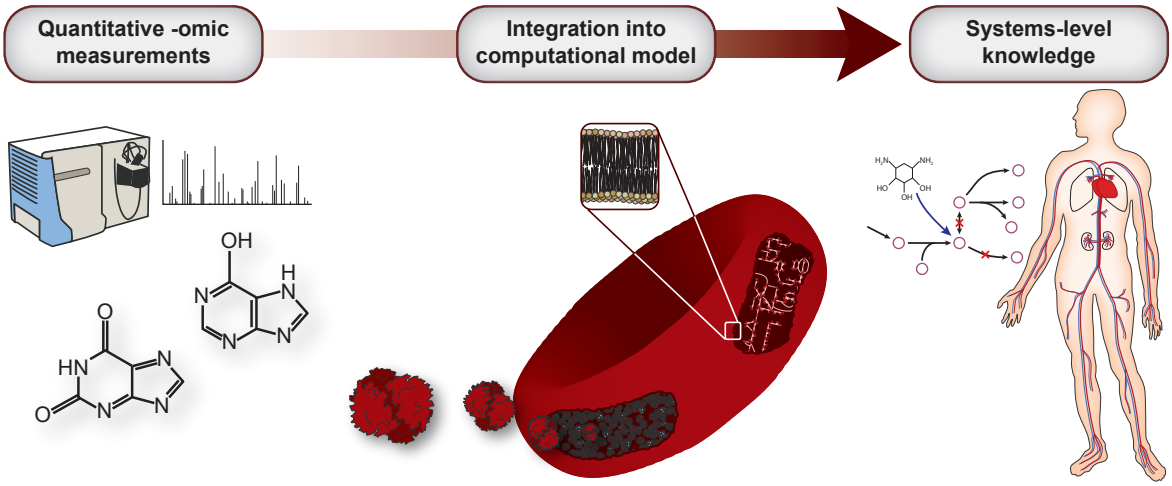

Figure 6.1: Workflow for integrating quantitative -omics data. Quantitative -omic data allows for integration into quantitative mechanistic models capable of generating phenotypic predictions.

ery underlying all metabolic processes and regulatory mechanisms [231]. Lipidomics technologies have enabled the in-depth characterization of the cellular membrane, including signaling, transport, and respiration mechanisms [232]. Researchers utilize one or more of these techniques to interrogate their system of interest, leading to rich information and important phenomenological observations. Recently, the RBC has been the source of much -omic data, yielding advances in analytic experimental techniques, rich data sets, and driving computational method development (Fig. 6.1).

Metabolomics

During the storage of RBCs in blood bags at 4°C, a variety of changes occur within the cells that impact their ability to carry oxygen and generate energy upon transfusion into a patient. These morphological and biochemical changes—collectively referred to as the "storage lesion" [22]—have been thoroughly explored through the use of metabolomics data over the last decade [25]. These studies have explored the impact of various perturbations to the storage media

on the metabolic function of the RBCs over the course of the 42 day storage period by taking weekly time points. A set of metabolites was identified that serves as storage-age biomarkers [62]. Several of these studies have provided very informative data sets [18, 42, 43, 61, 233–235], but their utility for systems biology modeling is limited due to their qualitative nature; while raw signals can provide meaningful statistical analyses [33, 35], they are inherently incapable of being integrated into quantitative models.

More recently, there has been an influx of quantitative data characterizing the RBC storage process. Perhaps the most complete characterization to date was produced by Bordbar et al. [41], providing a much finer resolution on the temporal dynamics observed during storage by taking time points every 3-4 days. The RBC community is embracing the trend of quantitative data generation, producing more absolutely quantified data sets [44, 59]. Similar data have also been produced in other cells and organisms, such as the human platelet [89, 175], *E. coli* [159], and *S. cerevisiae* [158].

With such rich data available, the onus has been on the modeling community to help realize the full potential of these data. As a result, there have been several different computational approaches that utilize quantitative metabolomics data. Some studies have relied on statistical methodologies, such as using trained models to predict the concentration of metabolites in the RBC based on only the storage-age biomarkers as input—approaches presented in Chapter 3 of this dissertation [82, 83]. Personalized kinetic models of RBC metabolism have been produced through the incorporation of metabolomics data into a cell-scale model [37]; we examined how explicitly modeling regulatory mechanisms affected the ability of these models to maintain a homeostatic state in Chapter 4 of this dissertation. Kinetic models rely heavily on parameterization, however, and are therefore limited in scope.

Other modeling approaches, such as constraint-based modeling [120], have become widely used for the computation of cellular flux states. In Chapter 4 of this dissertation, we described a new constraint-based modeling method and workflow was developed that allows for the integration of both endo- and exo-metabolomics data [52]. Using the RBC as a case study because of the availability of a cell-scale metabolic model [40] and detailed data [41], we computationally predicted and experimentally validated the finding that citrate metabolized to produce glutamate and lactate; these findings were validated and explored in more detail in a separate study [51]. This method will likely prove to be a valuable resource for the community as it enables modeling cell-scale dynamics without a kinetic model; it has already been used to explore the temperature dependence of the RBC's metabolic network through the integration of quantitative metabolomics data measured at different temperatures as presented in Chapter 2 of this dissertation [121]. With recent breakthroughs in analytical methodologies, like that of D'Alessandro and colleagues for a three-minute quantitative metabolomics and fluxomics characterization [236], we will undoubtedly see an avalanche of high quality quantitative data in the near future.

Fluxomics

The current state-of-the-art for experimentally validating these constraint-based flux models is isotopic labeling ("fluxomics") data. This specialized type of metabolomics data allows for the tracing of compounds through a metabolic network [237]. In the RBC, these data can be used to build kinetic models [238] and determine the metabolic fate of adenine [61] and citrate [51, 52]. While fluxomics data are limited to elucidating the flux through key reactions in cells and cannot determine the metabolic state of all reactions in the network, they can be integrated into genome-scale models [161]. These integrated modeling platforms allow for the

validation of model-based flux predictions [239].

Proteomics

The last several years have seen important steps forward for both the development of proteomics data and their subsequent integration into systems biology methods. Without the use of computational models, quantitative proteomics data have proven to be extremely informative: they have been used to identify novel proteins in already well-characterized systems like *E. coli* [240], identify and characterize post-translational modifications in rat muscle tissue [241], and provide insights into disease phenotypes [242]. In 2016, Schmidt et al. characterized the *E. coli* proteome across 22 different experimental conditions, doubling the previous number of quantified *E. coli* proteins [199]. These data yielded several important observations, including how protein abundances were dependent upon growth rate.

The great potential for such quantitative data was quickly realized by the computational community, and numerous modeling approaches have been produced that are able to integrate quantitative proteomics data. Shortly after the Schmidt et al. data were published [199], so-called Metabolism and Expression (ME) models were adapted to allow for the direct integration of quantitative proteomics data. The resulting method—presented in Chapter 5 of this dissertation [188]—calibrated an existing genome-scale model of *E. coli* to compute proteome allocation under the various conditions explored in the data, providing improved phenotypic predictions. Like ME-models, other modeling formalisms, such as GECKO [185], also provide a mechanistic approach for the computation of cellular flux states. The modeling framework proposed by Hui et al. [204] offers an alternative, statistically-based approach for the integration of quantitative proteomics, offering the ability to predict proteome composition in new environments.

The RBC has been the focus of in-depth proteomics analysis as well. Through the use of quantitative proteomics, over 2,500 proteins have been identified in the RBC, almost 1,900 of which occur at more than 100 copies per cell [187]. This study represents one of the most comprehensive characterizations of the RBC proteome to date, an advance made possible by quantitative proteomics analysis [243]. The most complete cell-scale mechanistic model of RBC metabolism [40] was informed by proteomics data [244–247], but this network reconstruction is in need of an update to correspond with the updated characterization of the RBC proteome. Now that the data is available, it is clear that the modeling community will move to not only update our current understanding of the metabolic network but will also work to integrate quantitative proteomics data using current modeling formalisms as blueprints.

Lipidomics

The last several years have seen a push in both the development of lipidomic technologies [248, 249] and data generation [250, 251]. In particular, the field has embraced the use of quantitative technologies to identify and characterize new lipids [252] and reveal systems-level phenotypic trends [253]. Because membrane lipids are involved in signaling processes, they act as biomarkers that can be used in the prediction of diseased phenotypes [254]. The RBC has been used as a model for detailing the structure and fluidity of the lipid bilayer [255], a critical step in understanding pathogenesis [256, 257]. Changes in the RBC membrane can lead to the development of sickled cells and inflammation [258] and have been implicated in the development and progression of Alzheimer's disease [259].

Nevertheless, there is still progress to be made in the generation of high quality quantitative lipidomics data [260]. There are also challenges associated with the integration of lipidomics

data into multi-omic models [261]. Like with other -omic technologies, the field will surely address and overcome these challenges. To date, several metabolic models have incorporated lipid maps [262], but it will be important for computational biologists to start integrating quantitative lipidomics data into their models as more data is generated. The development of such models will lead to more powerful predictions defining the role of lipids in areas such as pathogenesis.

6.3 What is on the horizon of systems biology?

There are many quantitative -omic data sets available for a variety of organisms and cell types. Such data will undoubtedly enable a new wave of systems biology models. However, it is not enough to simply generate quantitative data; the data itself must also be high quality. High quality data means generating an appropriate number of replicates to capture inherent biological variability, account for the batch effect, and provide statistical power. Any model, whether based on mechanisms or statistics, is only as good as the data on which it was built. There are ways to account for uncertainty in computational models [263, 264], but there is only so much that can be done on the computational side—large errors in measurements propagate and eventually lead to low quality models. Thus, it is important that the necessary time be taken to ensure that the experimental data being generated is of high quality.

The continued generation of quantitative data will provide progressively informative observations. As -omic data are continually generated, it will be increasingly important to use mechanistic models to interpret the data and generate testable hypotheses [27]. What is required for us to realize this important goal? It is clear that experimental technologies have opened the door for extremely informative quantitative data sets that probe a wide range of cellular behaviors. Researchers who generate these data need to embrace these technologies to produce quantitative

data. In order to use these data, however, we first need accurate, quantitative models.

In the preceding sections, we detailed various types of mathematical models that have been developed for different -omic data types (e.g., kinetic models account for metabolite concentrations and protein abundances). Now that computational methods for the integration of individual data types exist, we need to start to build models that merge these modeling types together to account for multiple data types within a single model. Quantitative -omic data are bringing us closer to the next generation of multi-omic models, of which only a few currently exist [265, 266].

However, biological systems are highly complex and involve many kinds of interactions (e.g., protein-protein) that are not necessarily explicitly accounted for by a particular modeling method. Great strides have been made to characterize the biophysical interactions occurring among various cellular components through methods like the yeast two-hybrid system [267]. Resources like the STRING database [268] catalogue this "interactome," providing the starting point for systems biology studies of these interaction networks [269]. As this information becomes available for cells like RBCs, the modeling community is poised to begin to incorporate the interactome into multi-omic models that will reveal new insights into cellular physiology [270].

To this point, modeling formulations have ignored several other important biophysical constraints placed on the cell and particularly the metabolome. Genome-scale metabolic models have become a dominant paradigm in systems biology by allowing the prospective calculation of reaction network flux states without the need for measured metabolite concentrations or parameterized reaction rate laws. Thermodynamic analysis has been developed to integrate compartment-specific metabolite concentration data and identify consistency of fluxes and concentrations with the Second Law of Thermodynamics. However, consistency with the second law

and compartment-specific metabolite concentration data are not the only criteria for a valid concentration set. Additionally, concentration sets must be osmotically balanced, charge balanced, and consistent with isomer pool and non-compartment-specific measurements, which are much more common than compartment-specific measurements. The next steps for systems biologists would be to develop a framework that integrates these physico-chemical constraints with the second law constraints to define the biophysically-constrained metabolite concentration space.

In particular, constraints that directly affect metabolites would be of interest. Metabolite concentrations are fundamental variables that, together with macromolecule concentrations, determine the functional state of the cell. Genome-scale metabolic network reconstructions have precisely detailed the metabolite composition for a broad range of organisms [271–273]. Metabolites vary several orders of magnitude in their concentration within the cell [274–276], and changes in metabolite concentrations often serve regulatory [277] or homeostatic [278] roles in response to environmental stimuli. In order to maintain a functioning cell, metabolites cannot have arbitrary levels; in other words, metabolic concentrations are constrained to be at levels that allow cell viability and accomplish cellular objectives. Defining these constraints allows for the calculation of the feasible space of metabolite concentrations and study how this space changes as cell state variables—such as cellular flux state or culture temperature—vary with changing environmental conditions.

One such constraint that has been previously examined arises from the Second Law of Thermodynamics, which states that the Gibbs energies of reactions are negative for reactions with positive net reaction rate [279]. These Gibbs energies are determined by metabolite concentrations, and thus the need of the cell to meet particular metabolic objectives, such as the production of ATP from glucose through glycolysis, places constraints upon metabolite concen-

trations. These Second Law constraints have been defined and utilized to analyze the metabolite concentration space under different flux conditions [280, 281]. Notably, these Second Law constraints alone are insufficient to explain the quantitative levels of metabolites [282], in large part because the Second Law defines relative constraints between metabolite levels but does not enforce absolute constraints upon their levels.

A number of other biophysical constraints on cells exist that are determined in large part by metabolite concentrations. For example, in order for cellular volume to be maintained, the volume of water within the cell must be maintained, which is determined by the relative osmolality of the intracellular and extracellular spaces. This osmolality is determined primarily by metabolite levels and thus places an additional constraint upon their levels. The next step is therefore to describe a number of biophysical as well as experimental constraints on metabolite concentrations (Fig. 6.2). The first challenge would be to efficiently search the constrained metabolic concentration space for arbitrary objectives, such as finding the maximum absolute concentration of a particular metabolite.

In Fig. 6.2, we visualize some of these constraints in *E. coli*. However, this framework would be applicable to any organism because of the fundamental biophysical nature of these constraints—they act on all living cells. Understanding these constraints and how they work together to maintain and regulate the cellular environment will undoubtedly provide insight into the development of primordial cells. One of the first cells to examine with these constraints will be the RBC; an abundance of available data combined with its simplicity makes the RBC an excellent starting point.

The RBC has been and will continue to be a leading platform in which to develop and utilize systems biology approaches: it has a host of available data and knowledge, it is the simplest

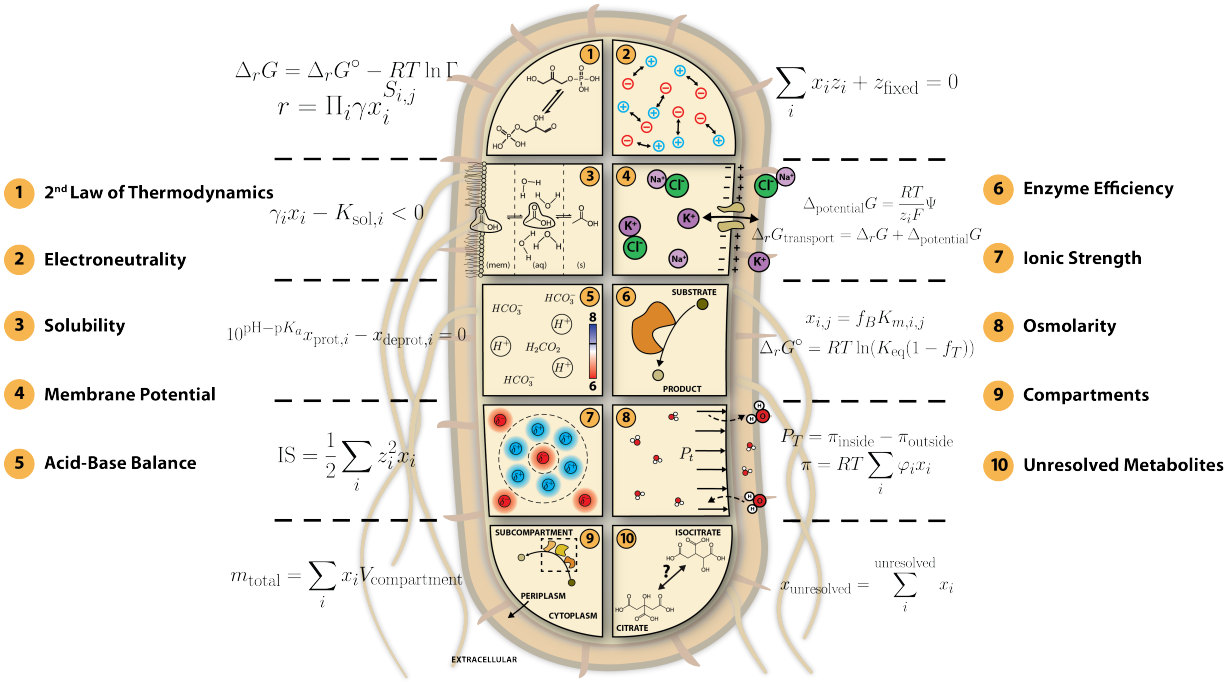

Figure 6.2: Definition of biophysical constraints. An outline of some of the biophysical constraints governing intracellular metabolite concentrations (depicted for *E. coli*).

cell type, and it represents several relevant medical interests. With the continued production of quantitative -omic data, we will soon see the development of true whole-cell computational models for the RBC and other cells (Fig. 6.3). Such a model will account for sequence variations between individuals, allowing for personalized physiological predictions. Quantitative measurements such as proteomics, lipidomics, metabolomics, and membrane proteomics will be used to inform these models, allowing for the incorporation of additional molecular mechanisms. Ultimately, the integration of such data will empower systems biologists to ask and answer new questions about the inner workings of cells.

In Chapter 5 of this dissertation, we have taken the first steps toward this whole-cell model, integrating kinetic and constraint-based modeling approaches with proteomic constraints overlaid. These preliminary results will provide a springboard for future researchers to inte-

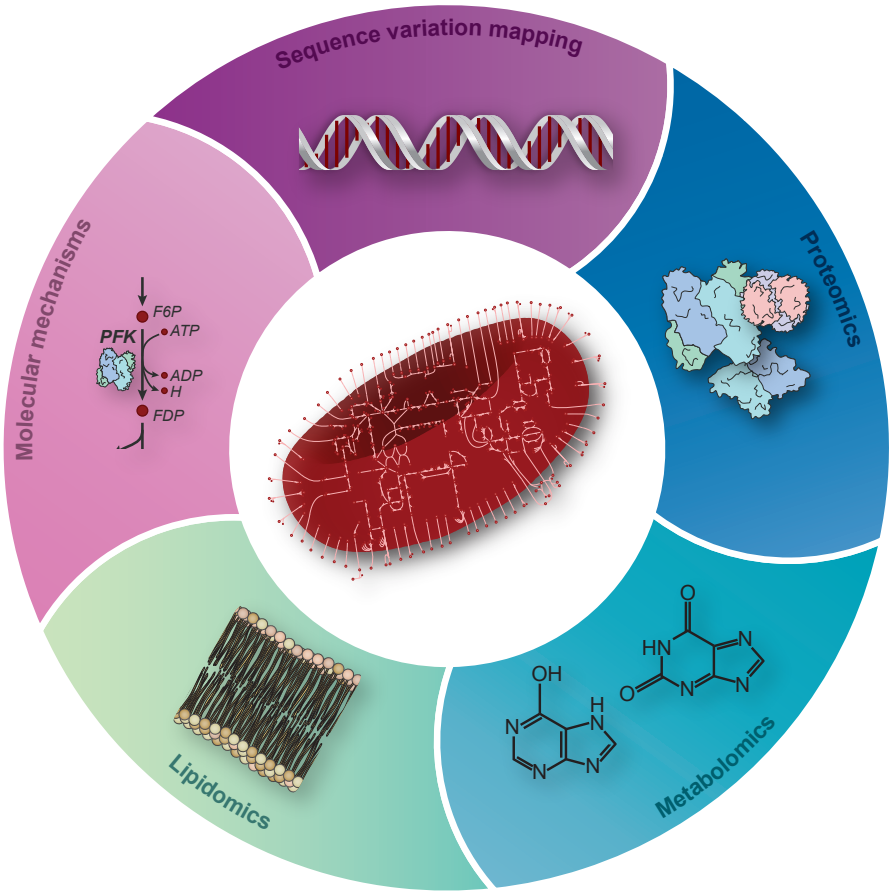

Figure 6.3: Next generation whole-cell RBC models. The next generation model of the red cell will integrate many different -omic data types, sequence variations, and various mechanisms to account for the myriad constraints on its structure and function.

grate new data types and modeling formalisms to build more powerful representations of RBC physiology.

Interpreting data

The effective dissemination of-omics data, their contextualization, and visualization to the community represent new challenges for the field. There is a growing need for software that interacts with and utilizes biological data, whether through computation, visualization, or data storage. Researchers in the life sciences have been developing their own software tools to meet

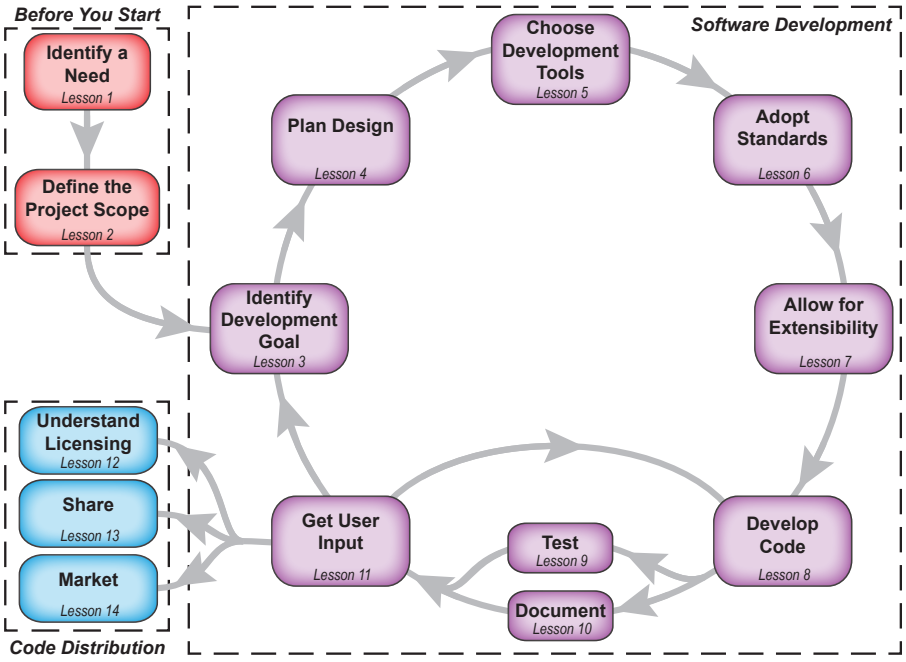

Figure 6.4: 14 lessons for software development.

these needs, often with great success. Developing and maintaining software with many outside users is difficult work. Extensive learning resources for computer programming and software development make it easy to get started, but it remains challenging to develop high-quality software that is widely used. To gain traction, academic software libraries need to be intuitive, well-documented, compatible, and extensible.

We provide a starting point and reference guide for scientists with some background in coding who are interested in developing and sharing their own software libraries with their academic communities. We present this guide through a series of lessons that have been influenced by our own experiences in developing academic software, covering topics from project scope and design to licensing and marketing (Fig. 6.4).

Collectively, these 14 guidelines provide a practical roadmap for the entire precess of

conceptualizing, designing, and distributing a software library. Lessons 1 and 2 cover the things you need to think about before starting development. Lessons 3 through 11 focus on the actual software development process, including planning, design, and implementation. Finally, Lessons 12 through 14 discuss important issues related to code distribution. These steps can be used to build software platforms that can integrate the data types we discussed above—metabolomics, proteomics, etc.—into visualization or modeling software [283].

Following these outlined steps, we built ErythroDB (http://erythrodb.org), an open source multi-omic visual knowledge base for the human red blood cell to help address the challenges of data dissemination. The transfusion medicine community is on the brink of a paradigm shift, moving into a phase in which the broad use of systems biology approaches, wide integration of disparate data types, and easy accessibility and visualization is likely to touch most of its members over the coming years.

6.4 Conclusions

In this dissertation, we have shown that the human RBC is a useful model system for applying and developing systems biology approaches. In Chapter 2, we examined how perturbing the external temperature affects the RBC metabolome during storage. The primary conclusion here was that the network-level trends observed through PCA of the metabolome are conserved. Having studied the state of the metabolome through an observational study, we then asked whether these metabolic trajectories could be predicted. Thus, we used linear statistical models in Chapter 3 to predict these longitudinal trajectories given a set of five biomarkers as input. We showed that not only were these models capable of predicting the concentration of a target metabolite at a given point but that we could also forecast future concentration values given

only measurements up to the first ten days of storage. These models were surprisingly accurate, outperforming standard metrics for the majority of measured metabolite signals.

While such predictive models are quite valuable in biology, the obvious limitation is the inability of a purely statistical model to provide any mechanistic insight into the system. Thus, we employed bottom-up mechanistic models to explore the dynamics of glycolysis and the effect of allosteric regulation in Chapter 4, finding that the addition of regulatory information improved the disturbance rejection capabilities of the system. Kinetic models do not practically scale to the full metabolic network, however, forcing us to use constraint-based models to examine the dynamics on a network-level. In order to model the behaviors observed during storage, we developed a novel flux balance analysis method that integrates quantitative time-course metabolomics data into a constraint-based model. This method allowed us to explore the cellular states observed through PCA and provided improved *in silico* predictions as compared to normal flux balance analysis.

In Chapter 5, we extended our use of mechanistic models to develop a method for the integration of quantitative proteomics data into a cell-scale metabolic model. Here, we used *E. coli* instead of the RBC due to the existence of a cell-scale model that already explicitly accounted for the proteome. This chapter provides the blueprints for moving forward with the creation of a similar cell-scale model of the RBC that will account for both the full metabolome and the full proteome.

Every physical system has infinite dimensions and is nonlinear. Certain systems, however, can be represented in low-dimensions and with linear models. Taken together, the results presented here provide empirical proof that we can effectively model the RBC metabolome using low-dimensional, linear models. The fact that the network-level trends observed in the PCA

were conserved across a 33°C change in temperature suggests that the RBC metabolic network is robust to such perturbations. Further, the statistical models utilized in Chapter 3 could not have provided accurate predictions from just five biomarkers if the system could not be represented in a reduced dimensionality. Integrating quantitative intracellular and extracellular metabolomics data into a constraint-based model allowed us to explore the physiologically within one of the reduced dimensions.

Ultimately, the implication of being able to represent the metabolome in a low-dimensional space is that more than one measured variable represents the same driving principle that governs the system. Dimensionality reduction allows us to take advantage of this redundancy of information by forming single variables that are linear combinations of measured variables in the system. These driving principles have yet to be fully elucidated, although the results in Chapter 4 provide more evidence that the depletion of the large intracellular 2,3-DPG and malate pools are important.

We can place the data presented in Chapter 2 (i.e., temperature variation) in the context of several of the data sets discussed in Chapter 1 to fully realize the robustness of the RBC metabolic network (Fig. 6.5) to perturbations. We performed analysis on the z-score of the raw metabolomics data for each individual condition; these normalized data were then concatenated into a single matrix on which the dimensionality reduction techniques were applied. The overlaid lines in the global plots are cubic polynomial fitted lines calculated from the average of all biological replicates for a given condition. The plot was rotated in the transformed space such that time zero is in the lower left corner of the plot.

Each of these studies used to PCA to examine the metabolome under each condition, but running PCA on all of the data together demonstrates that these trends are conserved across

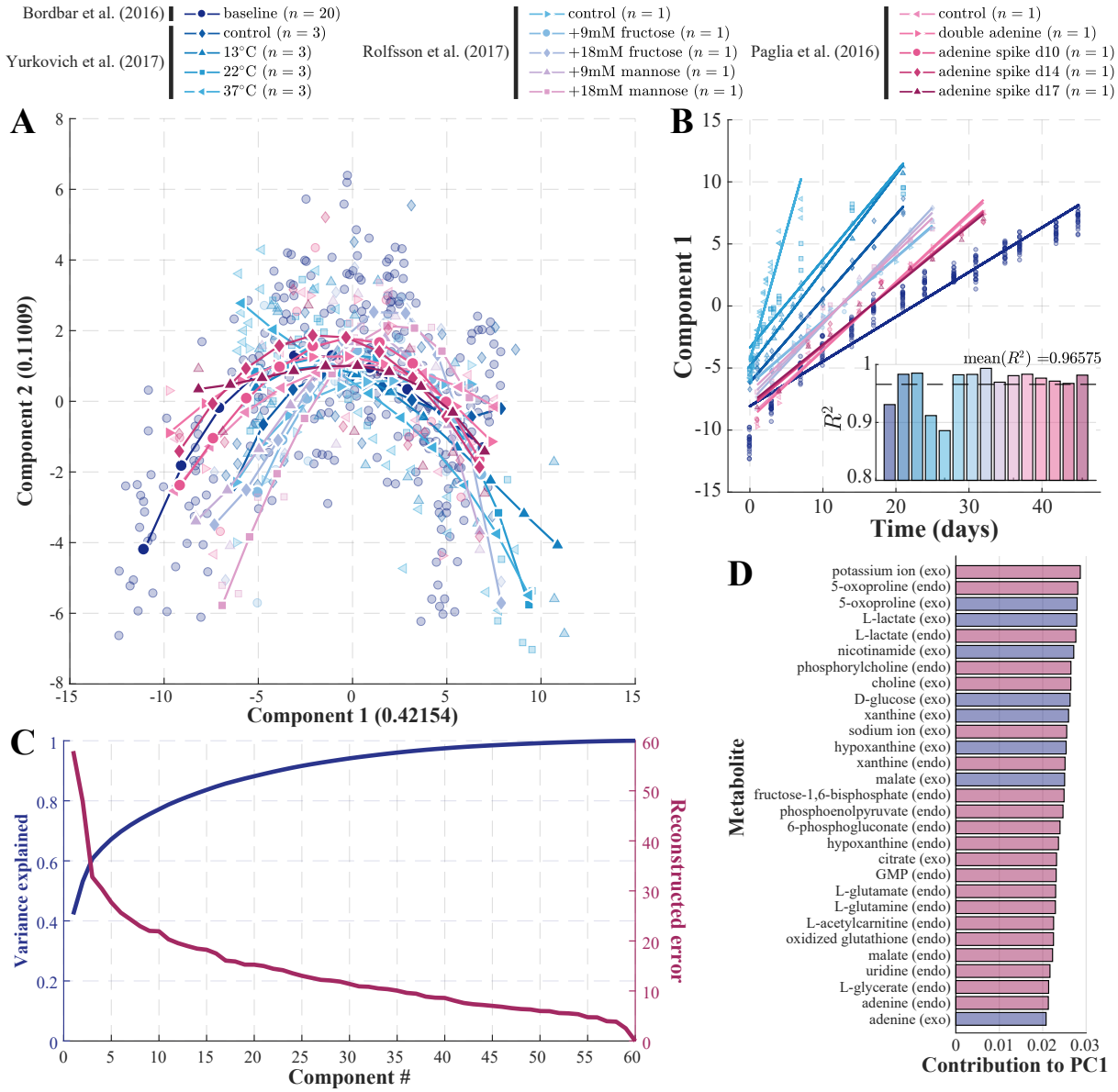

Figure 6.5: Empirically analyzing the dimensionality of the metabolome. (A) Principal component analysis on four published data sets: the baseline data (Bordbar et al. [41]), temperature variation (Yurkovich et al. [121], presented in Chapter 2), sugar supplementation (Rolfsson et al. [59]), and adenine supplementation (Paglia et al. [89]). (B) The first principal component was highly correlated with time in each data set, with an average $R^2 = 0.9657$. (C) Over 80% of the total variance is represented in the first 12 principal components, with approximately 60% represented in the first three components. (D) The magnitude of the contribution of the top metabolites to the first principal component; the storage-age biomarkers [62] are highlighted in purple.

all conditions. As the community continues to generate more data on various perturbations to the storage conditions, performing analysis on all of the data collectively will become more important. In order to elucidate the motions behind each of these modeled dimensions, we will first need to find an experimental condition that truly disrupts the network-level trends.

Transfusion medicine is a major part of healthcare. The results and insights gained from the application of -omics data sets and systems biology analytics to stored RBCs will continue to grow in scope and sophistication. As this systems view expands to include additional types of information and data, phenomena such as genetic variation in the human population is likely to come into focus. The community has built large, collaborative efforts like the REDS-III initiative [284] that directly address some of these issues.

The human RBC is not only the ideal target for systems analysis, but it also represents a system of high interest for studying human physiology and is central to transfusion medicine. The RBC is the cell type most amenable to systems analysis through the integration of multiple -omic data types into a mechanistic model. These data sets can be gathered to reflect various criteria such as gender, age, and ethnic diversity. Ultimately, there is great promise for the use of systems biology approaches to design experiments informed by a mechanistic understanding of RBC physiology [27]. Multi-scale analysis of RBC functions is needed to elucidate its role in human physiology, a fact easily demonstrated by looking at the physiologically-relevant RBC time scales: one second for capillary transit, one minute for average circulatory time, 45 minutes for ATP turnover, 24 hours for circadian rhythms, and 60 days for its half-life in circulation. Once we succeed with a refined definition of RBC systems biology, it is logical to proceed to the next simplest cell—the human platelet. The presence of organelles (e.g., mitochondria) and signaling pathways will provide challenges beyond those faced in RBC physiology. Nevertheless,

the methods and modeling formalisms used in this dissertation on RBCs can be readily applied to other cellular systems like the platelet to explore our ability to observe, predict, and understand metabolic physiology.

Acknowledgements

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

- **JT Yurkovich** and BO Palsson. 2018. "Quantitative -omics data empowers bottom-up systems biology." In press (*Current Opinion in Biotechnology*). The dissertation author was the primary author.

- **JT Yurkovich**, BJ Yurkovich, A Dräger, BO Palsson, and ZA King. 2017. "A Padawan Programmer's Guide to Developing Software Libraries." *Cell Systems*, 5(5):431-437. The dissertation author was the primary author.

- **JT Yurkovich**, A Bordbar, ÓE Sigurjónsson, and BO Palsson. 2018. "Systems biology as an emerging paradigm in transfusion medicine." *BMC Systems Biology*, 12:31. The dissertation author was the primary author.

Bibliography

[1] Kuhn TS (2012) The Structure of Scientific Revolutions: 50th Anniversary Edition. University of Chicago Press.

[2] Ideker T, Galitski T, Hood L (2001) A new approach to decoding life: Systems biology. Annual Review of Genomics and Human Genetics 2: 343–372.

[3] Thiele I, Palsson BØ (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc 5: 93–121.

[4] Palsson B (2011) Systems Biology: Simulation of Dynamic Network States. Cambridge University Press.

[5] Bordbar A, Monk JM, King ZA, Palsson BO (2014) Constraint-based models predict metabolic and associated cellular functions. Nature Reviews Genetics 15: 107–120.

[6] Vidyasagar M (2014) Hidden Markov Processes: Theory and Applications to Biology (Princeton Series in Applied Mathematics). Princeton University Press.

[7] Banga JR (2008) Optimization in computational systems biology. BMC Systems Biology 2: 47.

[8] Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? Nat Biotechnol 28: 245–248.

[9] Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, Zielinski DC, Bordbar A, Lewis NE, Rahmanian S, Kang J, Hyduke DR, Palsson BØ (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox v2.0. Nature Protocols 6: 1290–1307.

[10] Lewis NE, Nagarajan H, Palsson BO (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. Nat Rev Microbiol 10: 291–305.

[11] Shalem O, Sanjana NE, Zhang F (2015) High-throughput functional genomics using CRISPR–cas9. Nature Reviews Genetics 16: 299–311.

[12] Ebrahim A, Lerman JA, Palsson BO, Hyduke DR (2013) COBRApy: COnstraints-Based reconstruction and analysis for python. BMC Syst Biol 7: 74.

[13] Yim H, Haselbeck R, Niu W, Pujol-Baxley C, Burgard A, Boldt J, Khandurina J, Trawick JD, Osterhout RE, Stephen R, Estadilla J, Teisan S, Schreyer HB, Andrae S, Yang TH, Lee SY, Burk MJ, Dien SV (2011) Metabolic engineering of escherichia coli for direct production of 1, 4-butanediol. Nature Chemical Biology 7: 445–452.

[14] Palsson B (2015) Systems Biology: Constraint-based Reconstruction and Analysis. Cambridge University Press.

[15] Sender R, Fuchs S, Milo R (2016) Revised estimates for the number of human and bacteria cells in the body. PLoS Biol 14: e1002533.

[16] Takei T, Amin NA, Schmid G, Dhingra-Kumar N, Rugg D (2009) Progress in global blood safety for HIV. J Acquir Immune Defic Syndr 52 Suppl 2: S127–31.

[17] Simon TL, McCullough J, Snyder EL, Solheim BG, Strauss RG (2016) Rossi's Principles of Transfusion Medicine. John Wiley & Sons.

[18] Roback JD, Josephson CD, Waller EK, Newman JL, Karatela S, Uppal K, Jones DP, Zimring JC, Dumont LJ (2014) Metabolomics of ADSOL (AS-1) red blood cell storage. Transfus Med Rev 28: 41–55.

[19] Glynn SA, Klein HG, Ness PM (2016) The red blood cell storage lesion: the end of the beginning. Transfusion 56: 1462–1468.

[20] D'Alessandro A, Seghatchian J (2017) Hitchhiker's guide to the red cell storage galaxy: Omics technologies and the quality issue. Transfusion and Apheresis Science 56: 248–253.

[21] Chen D, Serrano K, Devine DV (2016) Introducing the red cell storage lesion. ISBT Sci Ser 11: 26–33.

[22] D'Alessandro A (2017) Red blood cell storage lesion. VOXS 12: 207–213.

[23] Paglia G, Palsson BØ, Sigurjonsson OE (2012) Systems biology of stored blood cells: can it help to extend the expiration date? J Proteomics 76 Spec No.: 163–167.

[24] D'Alessandro A, Kriebardis AG, Rinalducci S, Antonelou MH, Hansen KC, Papassideri IS, Zolla L (2015) An update on red blood cell storage lesions, as gleaned through biochemistry and omics technologies. Transfusion 55: 205–219.

[25] Nemkov T, Hansen KC, Dumont LJ, D'Alessandro A (2016) Metabolomics in transfusion medicine. Transfusion 56: 980–993.

[26] Patti GJ, Yanes O, Siuzdak G (2012) Innovation: Metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Biol 13: 263–269.

[27] Bordbar A (2017) Interpreting the deluge of omics data: new approaches offer new possibilities. Blood Transfus 15: 189–190.

[28] Hod EA, Francis RO, Spitalnik SL (2017) Red blood cell storage Lesion-Induced adverse effects: More smoke; is there fire? Anesth Analg 124: 1752–1754.

[29] Logan JA, Kelly ME, Ayers D, Shipillis N, Baier G, Day PJR (2010) Systems biology and modeling in neuroblastoma: practicalities and perspectives. Expert Rev Mol Diagn 10: 131–145.

[30] Verma M, Karimiani EG, Byers RJ, Rehman S, Westerhoff HV, Day PJR (2013) Mathematical modelling of miRNA mediated BCR.ABL protein regulation in chronic myeloid leukaemia vis-a-vis therapeutic strategies. Integr Biol 5: 543–554.

[31] Zhang W, Edwards A, Fan W, Flemington EK, Zhang K (2012) miRNA-mRNA correlation-network modules in human prostate cancer and the differences between primary and metastatic tumor subtypes. PLoS One 7: e40130.

[32] Mani KM, Lefebvre C, Wang K, Lim WK, Basso K, Dalla-Favera R, Califano A (2008) A systems biology approach to prediction of oncogenes and molecular perturbation targets in b-cell lymphomas. Mol Syst Biol 4: 169.

[33] Bartel J, Krumsiek J, Theis FJ (2013) Statistical methods for the analysis of high-throughput metabolomics data. Comput Struct Biotechnol J 4: e201301009.

[34] Brereton RG, Lloyd GR (2014) Partial least squares discriminant analysis: taking the magic away: PLS-DA: taking the magic away. J Chemom 28: 213–225.

[35] Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ (2012) Bayesian independent component analysis recovers pathway signatures from blood metabolomics data. J Proteome Res 11: 4120–4131.

[36] Worley B, Powers R (2013) Multivariate analysis in metabolomics. Curr Metabolomics 1: 92–107.

[37] Bordbar A, McCloskey D, Zielinski DC, Sonnenschein N, Jamshidi N, Palsson BO (2015) Personalized Whole-Cell kinetic models of metabolism for discovery in genomics and pharmacodynamics. Cell Syst 1: 283–292.

[38] Thiele I, Swainston N, Fleming RMT, Hoppe A, Sahoo S, Aurich MK, Haraldsdottir H, Mo ML, Rolfsson O, Stobbe MD, Thorleifsson SG, Agren R, Bölling C, Bordel S, Chavali AK, Dobson P, Dunn WB, Endler L, Hala D, Hucka M, Hull D, Jameson D, Jamshidi N, Jonsson JJ, Juty N, Keating S, Nookaew I, Le Novère N, Malys N, Mazein A, Papin JA, Price ND, Selkov E Sr, Sigurdsson MI, Simeonidis E, Sonnenschein N, Smallbone K, Sorokin A, van Beek JHGM, Weichart D, Goryanin I, Nielsen J, Westerhoff HV, Kell DB, Mendes P, Palsson BØ (2013) A community-driven global reconstruction of human metabolism. Nat Biotechnol 31: 419–425.

[39] Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BØ (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci U S A 104: 1777–1782.

[40] Bordbar A, Jamshidi N, Palsson BO (2011) iAB-RBC-283: A proteomically derived knowledge-base of erythrocyte metabolism that can be used to simulate its physiological and patho-physiological states. BMC Syst Biol 5: 110.

[41] Bordbar A, Johansson PI, Paglia G, Harrison SJ, Wichuk K, Magnusdottir M, Valgeirsdottir S, Gybel-Brask M, Ostrowski SR, Palsson S, Rolfsson O, Sigurjónsson OE, Hansen MB, Gudmundsson S, Palsson BO (2016) Identified metabolic signature for assessing red blood cell unit quality is associated with endothelial damage markers and clinical outcomes. Transfusion 56: 852–862.

[42] D'Alessandro A, Nemkov T, Kelher M, West FB, Schwindt RK, Banerjee A, Moore EE, Silliman CC, Hansen KC (2015) Routine storage of red blood cell (RBC) units in additive solution-3: a comprehensive investigation of the RBC metabolome. Transfusion 55: 1155–1168.

[43] D'Alessandro A, Nemkov T, Hansen KC, Szczepiorkowski ZM, Dumont LJ (2015) Red blood cell storage in additive solution-7 preserves energy and redox metabolism: a metabolomics approach. Transfusion 55: 2955–2966.

[44] Rolfsson Ó, Sigurjonsson ÓE, Magnusdottir M, Johannsson F, Paglia G, Gudmundsson S, Bordbar A, Palsson S, Brynjólfsson S, Gudmundsson S, Palsson B (2017) Metabolomics comparison of red cells stored in four additive solutions reveals differences in citrate anticoagulant permeability and metabolism. Vox Sang 112: 326–335.

[45] D'Alessandro A, D'Amici GM, Vaglio S, Zolla L (2011) Time-course investigation of SAGM-stored leukocyte-filtered red bood cell concentrates: from metabolism to proteomics. Haematologica 97: 107–115.

[46] Högman CF, Hedlund K, Sahleström Y (1981) Red cell preservation in protein-poor media. III. protection against in vitro hemolysis. Vox Sang 41: 274–281.

[47] Heaton A, Miripol J, Aster R, Hartman P, Dehart D, Rzad L, Grapka B, Davisson W, Buchholz DH (1984) Use of adsol preservation solution for prolonged storage of low viscosity AS-1 red blood cells. Br J Haematol 57: 467–478.

[48] Simon TL, Marcus CS, Myhre BA, Nelson EJ (1987) Effects of AS-3 nutrient-additive solution on 42 and 49 days of storage of red cells. Transfusion 27: 178–182.

[49] Walker WH, Netz M, Gänshirt KH (1990) [49 day storage of erythrocyte concentrates in blood bags with the PAGGS-mannitol solution]. Beitr Infusionsther 26: 55–59.

[50] Sparrow RL (2012) Time to revisit red blood cell additive solutions and storage conditions: a role for "omics" analyses. Blood Transfus 10 Suppl 2: s7–11.

[51] D'Alessandro A, Nemkov T, Yoshida T, Bordbar A, Palsson BO, Hansen KC (2017) Citrate metabolism in red blood cells stored in additive solution-3. Transfusion 57: 325–336.

[52] Bordbar A, Yurkovich JT, Paglia G, Rolfsson O, Sigurjónsson ÓE, Palsson BO (2017) Elucidating dynamic metabolic physiology through network integration of quantitative time-course metabolomics. Sci Rep 7: 46249.

[53] Dawson RB, Levine Z, Zuck T, Hershey RT, Myers C (1978) Blood preservation XXVII. fructose and mannose maintain ATP and 2,3-DPG. Transfusion 18: 347–352.

[54] Sharma V, Ichikawa M, Freeze HH (2014) Mannose metabolism: more than meets the eye. Biochem Biophys Res Commun 453: 220–228.

[55] Ha V, Jayalath VH, Cozma AI, Mirrahimi A, de Souza RJ, Sievenpiper JL (2013) Fructose-containing sugars, blood pressure, and cardiometabolic risk: a critical review. Curr Hypertens Rep 15: 281–297.

[56] Valeri F, Boess F, Wolf A, Göldlin C, Boelsterli UA (1997) Fructose and tagatose protect against oxidative cell injury by iron chelation. Free Radic Biol Med 22: 257–268.

[57] Beutler E, Duron O (1966) Studies on blood preservation. the relative capacities of hexoses, hexitols, and ethanol to maintain red cell ATP levels during storage. Transfusion 6: 537–542.

[58] Dawson RB, Hershey RT, Myers CS, Zuck TF (1980) Blood preservation. XXVIII. galactose and maltose maintain red blood cell 2,3-DPG and ATP. Transfusion 20: 110–113.

[59] Rolfsson Ó, Johannsson F, Magnusdottir M, Paglia G, Sigurjonsson ÓE, Bordbar A, Palsson S, Brynjólfsson S, Gudmundsson S, Palsson B (2017) Mannose and fructose metabolism in red blood cells during cold storage in SAGM. Transfusion .

[60] Concha II, Velásquez FV, Martínez JM, Angulo C, Droppelmann A, Reyes AM, Slebe JC, Vera JC, Golde DW (1997) Human erythrocytes express GLUT5 and transport fructose. Blood 89: 4190–4195.

[61] Paglia G, Sigurjónsson ÓE, Bordbar A, Rolfsson Ó, Magnusdottir M, Palsson S, Wichuk K, Gudmundsson S, Palsson BO (2016) Metabolic fate of adenine in red blood cells during storage in SAGM solution. Transfusion 56: 2538–2547.

[62] Paglia G, D'Alessandro A, Rolfsson Ó, Sigurjónsson ÓE, Bordbar A, Palsson S, Nemkov T, Hansen KC, Gudmundsson S, Palsson BO (2016) Biomarkers defining the metabolic age of red blood cells during cold storage. Blood .

[63] Casali E, Berni P, Spisni A, Baricchi R, Pertinhez TA (2015) Hypoxanthine: a new paradigm to interpret the origin of transfusion toxicity. Blood Transfus 14: 555–556.

[64] Hoff JH, Cohen E, Ewan T (1896) Studies in chemical dynamics. F. Muller.

[65] Kavanau JL (1950) Enzyme kinetics and the rate of biological processes. J Gen Physiol 34: 193–209.

[66] Behradek J (1930) Temperature coefficients in biology. Biol Rev Camb Philos Soc 5: 30–58.

[67] Elias M, Wieczorek G, Rosenne S, Tawfik DS (2014) The universality of enzymatic rate-temperature dependency. Trends Biochem Sci 39: 1–7.

[68] Somero GN, Hochachka PW (1968) The effect of temperature on catalytic and regulatory functions of pyruvate kinases of the rainbow trout and the antarctic fish trematomus bernacchii. Biochem J 110: 395–400.

[69] Hochachka PW (1991) Temperature: the ectothermy option. Biochemistry and molecular biology .

[70] Gillooly JF, Brown JH, West GB, Savage VM, Charnov EL (2001) Effects of size and temperature on metabolic rate. Science 293: 2248–2251.

[71] Burnside WR, Erhardt EB, Hammond ST, Brown JH (2014) Rates of biotic interactions scale predictably with temperature despite variation. Oikos 123: 1449–1456.

[72] Kirschbaum MUF (1995) The temperature dependence of soil organic matter decomposition, and the effect of global warming on soil organic C storage. Soil Biol Biochem 27: 753–760.

[73] Criddle RS, Hopkin MS, McARTHUR ED, Hansen LD (1994) Plant distribution and the temperature coefficient of metabolism. Plant Cell Environ 17: 233–243.

[74] Chaui-Berlinck JG, Alves Monteiro LH, Navas CA, J E P (2002) Temperature effects on energy metabolism: a dynamic system analysis. Proceedings of the Royal Society B: Biological Sciences 269: 15–19.

[75] Clarke A, Fraser KPP (2004) Why does metabolism scale with temperature? Funct Ecol 18: 243–251.

[76] Al-Fageeh MB, Smales CM (2006) Control and regulation of the cellular responses to cold shock: the responses in yeast and mammalian systems. Biochemical Journal 397: 247–259.

[77] Schulte PM (2015) The effects of temperature on aerobic metabolism: towards a mechanistic understanding of the responses of ectotherms to a changing environment. J Exp Biol 218: 1856–1866.

[78] D'Alessandro A, Liumbruno G, Grazzini G, Zolla L (2010) Red blood cell storage: the story so far. Blood Transfus 8: 82–88.

[79] Wallas CH (1979) Sodium and potassium changes in blood bank stored human erythrocytes. Transfusion 19: 210–215.

[80] Högman CF, Meryman HT (1999) Storage parameters affecting red blood cell survival and function after transfusion. Transfus Med Rev 13: 275–296.

[81] Hamasaki N, Yamamoto M (2000) Red blood cell function and blood storage. Vox Sang 79: 191–197.

[82] Yurkovich JT, Yang L, Palsson BO (2017) Biomarkers are used to predict quantitative metabolite concentration profiles in human red blood cells. PLoS Comput Biol 13: e1005424.

[83] Yurkovich JT, Yang L, Palsson BO (2017) Utilizing biomarkers to forecast quantitative metabolite concentration profiles in human red blood cells. In: 2017 IEEE Conference on Control Technology and Applications (CCTA). pp. 961–966.

[84] Lee JS, Kim-Shapiro DB (2017) Stored blood: how old is too old? J Clin Invest 127: 100–102.

[85] Rapido F, Brittenham GM, Bandyopadhyay S, La Carpia F, L'Acqua C, McMahon DJ, Rebbaa A, Wojczyk BS, Netterwald J, Wang H, Schwartz J, Eisenberger A, Soffing M, Yeh R, Divgi C, Ginzburg YZ, Shaz BH, Sheth S, Francis RO, Spitalnik SL, Hod EA (2017) Prolonged red cell storage before transfusion increases extravascular hemolysis. J Clin Invest 127: 375–382.

[86] Burgard AP (2004) Flux coupling analysis of genome-scale metabolic network reconstructions. Genome Research 14: 301–312.

[87] Borgmann AI, Moon TW (1976) Enzymes of the normothermic and hibernating bat, myotis lucifugus: Temperature as a modulator of pyruvate kinase. Journal of Comparative Physiology ? B 107: 185–199.

[88] Southwood AL (2003) Metabolic and cardiovascular adjustments of juvenile green turtles to seasonal changes in temperature and photoperiod. Journal of Experimental Biology 206: 4521–4531.

[89] Paglia G, Sigurjónsson ÓE, Rolfsson Ó, Valgeirsdottir S, Hansen MB, Brynjólfsson S, Gudmundsson S, Palsson BO (2014) Comprehensive metabolomic study of platelets reveals the expression of discrete metabolic phenotypes during storage. Transfusion 54: 2911–2923.

[90] Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, Djoumbou Y, Mandal R, Aziat F, Dong E, Bouatra S, Sinelnikov I, Arndt D, Xia J, Liu P, Yallou F, Bjorndahl T, Perez-Pineiro R, Eisner R, Allen F, Neveu V, Greiner R, Scalbert A (2012) HMDB 3.0–the human metabolome database in 2013. Nucleic Acids Research 41: D801–D807.

[91] Smith CA, Want EJ, Qin C, Trauger SA, Br TR, Custodio DE, Abagyan R, Siuzdak G (2005) Metlin: A metabolite mass spectral database. Drug Monit 27: 747–751.

[92] Hegarty TW (1973) Temperature coefficient (q10), seed germination and other biological processes. Nature 243: 305–306.

[93] Laidler KJ (1984) The development of the arrhenius equation. Journal of Chemical Education 61: 494.

[94] Sierra CA (2011) Temperature sensitivity of organic matter decomposition in the arrhenius equation: some theoretical considerations. Biogeochemistry 108: 1–15.

[95] Leifeld J, Fuhrer J (2005) The temperature response of CO2 production from bulk soils and soil fractions is related to soil organic matter quality. Biogeochemistry 75: 433–453.

[96] Hardy RN (1979) Temperature and Animal Life (Studies in Biology). Hodder.

[97] Larhlimi A, David L, Selbig J, Bockmayr A (2012) F2c2: a fast tool for the computation of flux coupling in genome-scale metabolic networks. BMC Bioinformatics 13: 57.

[98] Antonelou MH, Kriebardis AG, Papassideri IS (2010) Aging and death signalling in mature red cells: from basic science to transfusion practice. Blood Transfus 8 Suppl 3: s39–47.

[99] Flatt JF, Bawazir WM, Bruce LJ (2014) The involvement of cation leaks in the storage lesion of red blood cells. Front Physiol 5: 214.

[100] Beger RD, Dunn W, Schmidt MA, Gross SS, Kirwan JA, Cascante M, Brennan L, Wishart DS, Oresic M, Hankemeier T, Broadhurst DI, Lane AN, Suhre K, Kastenmüller G, Sumner SJ, Thiele I, Fiehn O, Kaddurah-Daouk R, for "Precision Medicine and Pharmacometabolomics Task Group"-Metabolomics Society Initiative (2016) Metabolomics enables precision medicine: "a white paper, community perspective". Metabolomics 12: 149.

[101] Patel S, Ahmed S (2015) Emerging field of metabolomics: big promise for cancer biomarker identification and drug discovery. J Pharm Biomed Anal 107: 63–74.

[102] Jansson J, Willing B, Lucio M, Fekete A, Dicksved J, Halfvarson J, Tysk C, Schmitt-Kopplin P (2009) Metabolomics reveals metabolic biomarkers of crohn's disease. PLoS One 4: e6386.

[103] Beger RD, Bhattacharyya S, Yang X, Gill PS, Schnackenberg LK, Sun J, James LP (2015) Translational biomarkers of acetaminophen-induced acute liver injury. Arch Toxicol 89: 1497–1522.

[104] O'Shea K, Cameron SJS, Lewis KE, Lu C, Mur LAJ (2016) Metabolomic-based biomarker discovery for non-invasive lung cancer screening: A case study. Biochim Biophys Acta 1860: 2682–2687.

[105] Aurich MK, Paglia G, Rolfsson Ó, Hrafnsdóttir S, Magnúsdóttir M, Stefaniak MM, Palsson BØ, Fleming RMT, Thiele I (2015) Prediction of intracellular metabolic states from extracellular metabolomic data. Metabolomics 11: 603–619.

[106] Ljung L (1998) System Identification: Theory for the User. Pearson Education.

[107] Tzes AP, Yurkovich S (1990) A frequency domain identification scheme for flexible structure control. J Dyn Syst Meas Control 112: 427.

[108] Steiglitz K, McBride L (1965) A technique for the identification of linear systems. IEEE Trans Automat Contr 10: 461–464.

[109] Seborg DE, Edgar TF, Shah SL (1986) Adaptive control strategies for process control: A survey. AIChE J 32: 881–913.

[110] Capan M, Hoover S, Jackson EV, Paul D, Locke R (2016) Time series analysis for forecasting hospital census: Application to the neonatal intensive care unit. Appl Clin Inform 7: 275–289.

[111] Kuepfer L, Peter M, Sauer U, Stelling J (2007) Ensemble modeling for analysis of cell signaling dynamics. Nature Biotechnology 25: 1001–1006.

[112] Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. Int J Forecast 22: 679–688.

[113] Kim-Shapiro DB, Lee J, Gladwin MT (2011) Storage lesion: role of red blood cell breakdown. Transfusion 51: 844–851.

[114] Yurkovich JT, Palsson BO (2016) Solving puzzles with missing pieces: The power of systems biology. Proceedings of the IEEE 104: 2–7.

[115] Schult DA, Swart P (2008) Exploring network structure, dynamics, and function using networkx. In: Proceedings of the 7th Python in Science Conferences (SciPy 2008). volume 2008, pp. 11–16.

[116] Smith BL, Williams BM, Oswald RK (2002) Comparison of parametric and nonparametric models for traffic flow forecasting. Transportation Research Part C: Emerging Technologies 10: 303–321.

[117] Zhang ZW, Cheng J, Xu F, Chen YE, Du JB, Yuan M, Zhu F, Xu XC, Yuan S (2011) Red blood cell extrudes nucleus and mitochondria against oxidative stress. IUBMB Life 63: 560–565.

[118] Yoshida T, Shevkoplyas SS (2010) Anaerobic storage of red blood cells. Blood Transfus 8: 220–236.

[119] Pujo-Menjouet L (2016) Blood Cell Dynamics: Half of a Century of Modelling. Mathematical Modelling of Natural Phenomena 11: 92–115.

[120] O'Brien EJ, Monk JM, Palsson BO (2015) Using genome-scale models to predict biological capabilities. Cell 161: 971–987.

[121] Yurkovich JT, Zielinski DC, Yang L, Paglia G, Rolfsson O, Sigurjónsson ÓE, Broddrick JT, Bordbar A, Wichuk K, Brynjólfsson S, Palsson S, Gudmundsson S, Palsson BO (2017) Quantitative time-course metabolomics in human red blood cells reveal the temperature dependence of human metabolic networks. J Biol Chem .

[122] Mahadevan R, Edwards JS, Doyle FJ (2002) Dynamic flux balance analysis of diauxic growth in Escherichia coli. Biophys J 83: 1331–1340.

[123] Waldherr S, Oyarzn DA, Bockmayr A (2015) Dynamic optimization of metabolic networks coupled with gene expression. Journal of theoretical biology 365: 469–485.

[124] Jamshidi N, Palsson B (2008) Formulating genome-scale kinetic models in the post-genome era. Mol Syst Biol 4: 171.

[125] Joshi A, Palsson BO (1989) Metabolic dynamics in the human red cell: Part I—A comprehensive kinetic model. J Theor Biol .

[126] Joshi A, Palsson BO (1989) Metabolic dynamics in the human red cell. part II–Interactions with the environment. J Theor Biol 141: 529–545.

[127] Joshi A, Palsson BO (1990) Metabolic dynamics in the human red cell. part III—Metabolic reaction rates. J Theor Biol 142: 41–68.

[128] Joshi A, Palsson BO (1990) Metabolic dynamics in the human red cell. part IV–Data prediction and some model computations. J Theor Biol 142: 69–85.

[129] Mulquiney PJ, Kuchel PW (1999) Model of 2, 3-bisphosphoglycerate metabolism in the human erythrocyte based on detailed enzyme kinetic equations1: computer simulation and metabolic control analysis. Biochemical Journal 342: 597–604.

[130] Nakayama Y, Kinoshita A, Tomita M (2005) Dynamic simulation of red blood cell metabolism and its application to the analysis of a pathological condition. Theor Biol Med Model 2: 18.

[131] Jamshidi N, Palsson BØ (2010) Mass action stoichiometric simulation models: Incorporating kinetics and regulation into stoichiometric models. Biophysical Journal 98: 175–185.

[132] Du B, Zielinski DC, Kavvas ES, Dräger A, Tan J, Zhang Z, Ruggiero KE, Arzumanyan GA, Palsson BO (2016) Evaluation of rate law approximations in bottom-up kinetic models of metabolism. BMC Systems Biology 10.

[133] Bode HW (1938) Variable equalizers. The Bell System Technical Journal 17: 229-244.

[134] Cavicehi TJ (1996) Phase-root locus and relative stability. IEEE Control Systems Magazine 16: 69–77.

[135] Abraham EH, Salikhova AY, Hug EB (2003) Critical ATP parameters associated with blood and mammalian cells: Relevant measurement techniques. Drug Development Research 59: 152–160.

[136] Wan J, Ristenpart WD, Stone HA (2008) Dynamics of shear-induced ATP release from red blood cells. Proceedings of the National Academy of Sciences 105: 16432–16437.

[137] Arciero JC, Carlson BE, Secomb TW (2008) Theoretical model of metabolic blood flow regulation: roles of ATP release by red blood cells and conducted responses. AJP: Heart and Circulatory Physiology 295: H1562–H1571.

[138] Berg JM, Tymoczko JL, Stryer L (2002) Biochemistry (Chapters 1-34). W. H. Freeman.

[139] Schöneberg T, Kloos M, Brüser A, Kirchberger J, Sträter N (2013) Structure and allosteric regulation of eukaryotic 6-phosphofructokinases. Biological Chemistry 394.

[140] Zanella A, Fermo E, Bianchi P, Valentini G (2005) Red cell pyruvate kinase deficiency: molecular and clinical aspects. British Journal of Haematology 130: 11–25.

[141] Atkinson DE, Walton GM (1967) Adenosine triphosphate conservation in metabolic regulation rat liver citrate cleavage enzyme. Journal of Biological Chemistry 242: 3239–3241.

[142] Purich DL, Fromm HJ (1972) Studies on factors influencing enzyme responses to adenylate energy charge. Journal of Biological Chemistry 247: 249–255.

[143] Shen L, Fall L, Walton GM, Atkinson DE (1968) Interaction between energy charge and metabolite modulation in the regulation of enzymes of amphibolic sequences. Phosphofructokinase and pyruvate dehydrogenase. Biochemistry 7: 4041–4045.

[144] Zames G (1981) Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses. IEEE Transactions on Automatic Control 26: 301–320.

[145] Liao CL, Atkinson DE (1971) Regulation at the phosphoenolpyruvate branchpoint in Azotobacter vinelandii: pyruvate kinase. Journal of bacteriology 106: 37–44.

[146] Ling KH, Byrne WL, Lardy H (1955) [38] Phosphohexokinase: Fructose-6-phosphate+ ATP Fructose-1, 6-diphosphate+ ADP. Methods in enzymology 1: 306–310.

[147] Wolfram Research Inc (2017) Mathematica 11.1. Champaign, Illinois, 11.1 edition. URL `http://www.wolfram.com`.

[148] Prankerd TAJ, Altman KI (1954) A study of the metabolism of phosphorus in mammalian red cells. Biochemical Journal 58(4): 622-633.

[149] Ponce J, Roth S, Harkness DR (1971) Kinetic studies on the inhibition of glycolytic kinases of human erythrocytes by 2, 3-diphosphoglyceric acid. Biochimica et Biophysica Acta (BBA)-Enzymology 250: 63–74.

[150] Gerber G, Preissler H, Heinrich R, Rapoport SM (1974) Hexokinase of human erythrocytes. purification, kinetic model and its application to the conditions in the cell. European Journal of Biochemistry 45: 39–52.

[151] Hoggett JG, Kellett GL (1995) Kinetics of the cooperative binding of glucose to dimeric yeast hexokinase PI. Biochemical journal 305: 405–410.

[152] Dean RB, Dixon WJ (1951) Simplified statistics for small numbers of observations. Analytical Chemistry 23: 636–638.

[153] Varma A, Palsson BO (1994) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. Appl Environ Microbiol 60: 3724–3731.

[154] Topfer N, Kleessen S, Nikoloski Z (2015) Integration of metabolomics data into metabolic networks. Frontiers in Plant Science 6.

[155] Willemsen AM, Hendrickx DM, Hoefsloot HCJ, Hendriks MMWB, Wahl SA, Teusink B, Smilde AK, van Kampen AHC (2015) MetDFBA: incorporating time-resolved metabolomics measurements into dynamic flux balance analysis. Mol BioSyst 11: 137–145.

[156] Kleessen S, Irgang S, Klie S, Giavalisco P, Nikoloski Z (2015) Integration of transcriptomics and metabolomics data specifies the metabolic response of chlamydomonas to rapamycin treatment. The Plant Journal 81: 822–835.

[157] Heise R, Fernie AR, Stitt M, Nikoloski Z (2015) Pool size measurements facilitate the determination of fluxes at branching points in non-stationary metabolic flux analysis: the case of arabidopsis thaliana. Frontiers in Plant Science 6.

[158] Bergdahl B, Heer D, Sauer U, Hahn-Hägerdal B, van Niel EW (2012) Dynamic metabolomics differentiates between carbon and energy starvation in recombinant saccharomyces cerevisiae fermenting xylose. Biotechnology for Biofuels 5: 34.

[159] McCloskey D, Gangoiti JA, King ZA, Naviaux RK, Barshop BA, Palsson BO, Feist AM (2014) A model-driven quantitative metabolomics analysis of aerobic and anaerobic metabolism in e. coli K-12 MG1655 that is biochemically and thermodynamically consistent. Biotechnol Bioeng 111: 803–815.

[160] D'Alessandro A, Righetti PG, Zolla L (2010) The red blood cell proteome and interactome: An update. Journal of Proteome Research 9: 144–163.

[161] Young JD (2014) INCA: a computational platform for isotopically non-stationary metabolic flux analysis. Bioinformatics 30: 1333–1335.

[162] Picker SM, Schneider V, Oustianskaia L, Gathof BS (2009) Cell viability during platelet storage in correlation to cellular metabolism after different pathogen reduction technologies. Transfusion 49: 2311–2318.

[163] Kilkson H, Holme S, Murphy S (1984) Platelet metabolism during storage of platelet concentrates at 22 degrees C. Blood 64: 406–414.

[164] Wasylenko TM, Stephanopoulos G (2014) Metabolomic and 13 c-metabolic flux analysis of a xylose-consuming saccharomyces cerevisiae strain expressing xylose isomerase. Biotechnology and Bioengineering 112: 470–483.

[165] Feng X, Zhao H (2013) Investigating glucose and xylose metabolism in saccharomyces cerevisiae and scheffersomyces stipitis via 13 c metabolic flux analysis. AIChE Journal 59: 3195–3202.

[166] Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, Palsson BO (2011) A comprehensive genome-scale reconstruction of escherichia coli metabolism–2011. Molecular Systems Biology 7: 535–535.

[167] Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H (2006) Construction of escherichia coli k-12 in-frame, single-gene knockout mutants: the keio collection. Molecular Systems Biology 2.

[168] Patrick WM, Quandt EM, Swartzlander DB, Matsumura I (2007) Multicopy suppression underpins metabolic evolvability. Molecular Biology and Evolution 24: 2716–2722.

[169] Bennett BD, Kimball EH, Gao M, Osterhout R, Dien SJV, Rabinowitz JD (2009) Absolute metabolite concentrations and implied enzyme active site occupancy in escherichia coli. Nature Chemical Biology 5: 593–599.

[170] Simpson RJ, Brindle KM, Campbell ID (1982) Spin ECHO proton NMR studies of the metabolism of malate and fumarate in human erythrocytes. Dependence on free NAD levels. Biochim Biophys Acta 721: 191–200.

[171] Tong LV (2008) Development and application of mass spectrometry-based metabolomics methods for disease biomarker identification. Ph.D. thesis, Massachusetts Institute of Technology.

[172] Thomas A, Rahmanian S, Bordbar A, Palsson BØ, Jamshidi N (2014) Network reconstruction of platelet metabolism identifies metabolic signature for aspirin resistance. Scientific Reports 4.

[173] Mo ML, Palsson BØ, Herrgård MJ (2009) Connecting extracellular metabolomic measurements to intracellular flux states in yeast. BMC Systems Biology 3: 37.

[174] Ishii N, Nakahigashi K, Baba T, Robert M, Soga T, Kanai A, Hirasawa T, Naba M, Hirai K, Hoque A, Ho PY, Kakazu Y, Sugawara K, Igarashi S, Harada S, Masuda T, Sugiyama N, Togashi T, Hasegawa M, Takai Y, Yugi K, Arakawa K, Iwata N, Toya Y, Nakayama Y, Nishioka T, Shimizu K, Mori H, Tomita M (2007) Multiple high-throughput analyses monitor the response of e. coli to perturbations. Science 316: 593–597.

[175] Paglia G, Sigurjónsson ÓE, Rolfsson Ó, Hansen MB, Brynjólfsson S, Gudmundsson S, Palsson BO (2015) Metabolomic analysis of platelets during storage: a comparison between apheresis- and buffy coat-derived platelet concentrates. Transfusion 55: 301–313.

[176] Millard P, Letisse F, Sokol S, Portais JC (2012) IsoCor: correcting MS data in isotope labeling experiments. Bioinformatics 28: 1294–1296.

[177] D'Alessandro A, Nemkov T, Sun K, Liu H, Song A, Monte AA, Subudhi AW, Lovering AT, Dvorkin D, Julian CG, Kevil CG, Kolluru GK, Shiva S, Gladwin MT, Xia Y, Hansen KC, Roach RC (2016) AltitudeOmics: Red blood cell metabolic adaptation to high altitude hypoxia. J Proteome Res 15: 3883–3895.

[178] Liu H, Zhang Y, Wu H, D'Alessandro A, Yegutkin GG, Song A, Sun K, Li J, Cheng NY, Huang A, Edward Wen Y, Weng TT, Luo F, Nemkov T, Sun H, Kellems RE, Karmouty-Quintana H, Hansen KC, Zhao B, Subudhi AW, Jameson-Van Houten S, Julian CG, Lovering AT, Eltzschig HK, Blackburn MR, Roach RC, Xia Y (2016) Beneficial role of erythrocyte adenosine A2B Receptor-Mediated AMP-Activated protein kinase activation in High-Altitude hypoxia. Circulation 134: 405–421.

[179] Sun K, Zhang Y, D'Alessandro A, Nemkov T, Song A, Wu H, Liu H, Adebiyi M, Huang A, Wen YE, Bogdanov MV, Vila A, O'Brien J, Kellems RE, Dowhan W, Subudhi AW, Jameson-Van Houten S, Julian CG, Lovering AT, Safo M, Hansen KC, Roach RC, Xia Y (2016) Sphingosine-1-phosphate promotes erythrocyte glycolysis and oxygen release for adaptation to high-altitude hypoxia. Nat Commun 7: 12086.

[180] Zhong R, Liu H, Wang H, Li X, He Z, Gangla M, Zhang J, Han D, Liu J (2015) Adaption to high altitude: An evaluation of the storage quality of suspended red blood cells prepared from the whole blood of tibetan plateau migrants. PLoS One 10: e0144201.

[181] Reisz JA, Wither MJ, Dzieciatkowska M, Nemkov T, Issaian A, Yoshida T, Dunham AJ, Hill RC, Hansen KC, D'Alessandro A (2016) Oxidative modifications of glyceraldehyde 3-phosphate dehydrogenase regulate metabolic reprogramming of stored red blood cells. Blood 128: e32–42.

[182] Zolla L, Timperio AM, Mirasole C, D'Alessandro A (2013) Red blood cell lipidomics analysis through HPLC-ESI-qTOF: application to red blood cell storage. J Integr OMICS 3.

[183] Thiele I, Fleming RMT, Que R, Bordbar A, Diep D, Palsson BO (2012) Multiscale modeling of metabolism and macromolecular synthesis in e. coli and its application to the evolution of codon usage. PLoS One 7: e45635.

[184] O'Brien EJ, Lerman JA, Chang RL, Hyduke DR, Palsson BØ (2013) Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol Syst Biol 9: 693.

[185] Sánchez BJ, Zhang C, Nilsson A, Lahtvee PJ, Kerkhoven EJ, Nielsen J (2017) Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. Mol Syst Biol 13: 935.

[186] Schultz A, Qutub AA (2015) Predicting internal cell fluxes at sub-optimal growth. BMC Syst Biol 9: 18.

[187] Bryk AH, Wiśniewski JR (2017) Quantitative analysis of human red blood cell proteome. J Proteome Res 16: 2752–2761.

[188] Yang L, Yurkovich JT, Lloyd CJ, Ebrahim A, Saunders MA, Palsson BO (2016) Principles of proteome allocation are revealed using proteomic data and genome-scale models. Sci Rep 6: 36734.

[189] Mih N, Brunk E, Bordbar A, Palsson BO (2016) A multi-scale computational platform to mechanistically assess the effect of genetic variation on drug responses in human erythrocyte metabolism. PLoS Comput Biol 12: e1005039.

[190] Monk J, Nogales J, Palsson BO (2014) Optimizing genome-scale network reconstructions. Nature Biotechnology 32: 447–452.

[191] Reed JL (2012) Shrinking the metabolic solution space using experimental datasets. PLoS Comput Biol 8: e1002662.

[192] Kim MK, Lun DS (2014) Methods for integration of transcriptomic data in genome-scale metabolic models. Comput Struct Biotechnol J 11: 59–65.

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

[194] Shlomi T, Cabili MN, Herrgård MJ, Palsson BØ, Ruppin E (2008) Network-based prediction of human tissue-specific metabolism. Nat Biotechnol 26: 1003–1010.

[195] Kim M, Yi JS, Lakshmanan M, Lee DY, Kim BG (2016) Transcriptomics-based strain optimization tool for designing secondary metabolite overproducing strains of streptomyces coelicolor. Biotechnol Bioeng 113: 651–660.

[196] Lerman JA, Hyduke DR, Latif H, Portnoy VA, Lewis NE, Orth JD, Schrimpe-Rutledge AC, Smith RD, Adkins JN, Zengler K, et al. (2012) In silico method for modelling metabolism and gene product expression at genome scale. Nature Communications 3: 929.

[197] Liu JK, O'Brien EJ, Lerman JA, Zengler K, Palsson BO, Feist AM (2014) Reconstruction and modeling protein translocation and compartmentalization in *Escherichia coli* at the genome-scale. BMC Syst Biol 8: 110.

[198] O'Brien E, Utrilla J, Palsson B (2016) Quantification and classification of e. coli proteome utilization and unused protein costs across environments. PLoS Comput Biol 12: e1004998.

[199] Schmidt A, Kochanowski K, Vedelaar S, Ahrné E, Volkmer B, Callipo L, Knoops K, Bauer M, Aebersold R, Heinemann M (2016) The quantitative and condition-dependent escherichia coli proteome. Nat Biotechnol 34: 104–110.

[200] Utrilla J, O???Brien EJ, Chen K, McCloskey D, Cheung J, Wang H, Armenta-Medina D, Feist AM, Palsson BO (2016) Global rebalancing of cellular resources by pleiotropic point mutations illustrates a multi-scale mechanism of adaptive evolution. Cell Systems 2: 260–271.

[201] Price M, Wetmore KM, Deutschbauer AM, Arkin AP (2016) A comparison of the costs and benefits of bacterial gene expression. bioRxiv:038851 .

[202] Oh Yg, Lee Dy, Lee SY, Park S (2009) Multiobjective Flux Balancing Using the NISE Method for Metabolic Network Analysis. Biomolecular Engineering : 999–1008.

[203] Aidelberg G, Towbin BD, Rothschild D, Dekel E, Bren A, Alon U (2014) Hierarchy of non-glucose sugars in escherichia coli. BMC Syst Biol 8: 1.

[204] Hui S, Silverman JM, Chen SS, Erickson DW, Basan M, Wang J, Hwa T, Williamson JR (2015) Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria. Mol Syst Biol 11: 784.

[205] Galperin MY, Makarova KS, Wolf YI, Koonin EV (2014) Expanded microbial genome coverage and improved protein family annotation in the cog database. Nucleic Acids Research 43: D261-D269.

[206] Liu M, Durfee T, Cabrera JE, Zhao K, Jin DJ, Blattner FR (2005) Global transcriptional programs reveal a carbon source foraging strategy by escherichia coli. Journal of Biological Chemistry 280: 15921–15927.

[207] Klumpp S, Hwa T (2014) Bacterial growth: global effects on gene expression, growth feedback and proteome partition. Curr Opin in Biotechnol 28: 96–102.

[208] Fischer E, Sauer U (2003) Metabolic flux profiling of *Escherichia coli* mutants in central carbon metabolism using gc-ms. European Journal of Biochemistry 270: 880–891.

[209] Ibarra RU, Edwards JS, Palsson BO (2002) *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. Nature 420: 186-189.

[210] Gerosa L, van Rijsewijk BRH, Christodoulou D, Kochanowski K, Schmidt TS, Noor E, Sauer U (2015) Pseudo-transition analysis identifies the key regulators of dynamic metabolic adaptations from steady-state data. Cell Systems 1: 270–282.

[211] van Rijsewijk BRH, Nanchen A, Nallet S, Kleijn RJ, Sauer U (2011) Large-scale 13c-flux analysis reveals distinct transcriptional control of respiratory and fermentative metabolism in escherichia coli. Mol Syst Biol 7: 477.

[212] Yang L, Ma D, Ebrahim A, Lloyd CJ, Saunders MA, Palsson BO (2016) solveME: fast and reliable solution of nonlinear ME models. BMC Bioinform 17: 391.

[213] O'Brien EJ, Palsson BO (2015) Computing the functional proteome: recent progress and future prospects for genome-scale models. Curr Opin Biotechnol 34: 125-134.

[214] LaCroix RA, Sandberg TE, O'Brien EJ, Utrilla J, Ebrahim A, Guzman GI, Szubin R, Palsson BO, Feist AM (2015) Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of *Escherichia coli* K-12 MG1655 on glucose minimal medium. Appl Environ Microbiol 81: 17–30.

[215] Escalante A, Cervantes AS, Gosset G, Bolívar F (2012) Current knowledge of the *Escherichia coli* phosphoenolpyruvate–carbohydrate phosphotransferase system: peculiarities of regulation and impact on growth and product formation. Applied Microbiology and Biotechnology 94: 1483–1494.

[216] Seo SW, Kim D, OBrien EJ, Szubin R, Palsson BO (2015) Decoding genome-wide gadewx-transcriptional regulatory networks reveals multifaceted cellular responses to acid stress in escherichia coli. Nat Commun 6.

[217] Seo SW, Kim D, Latif H, O'Brien EJ, Szubin R, Palsson BO (2014) Deciphering Fur transcriptional regulatory network highlights its complex role beyond iron metabolism in Escherichia coli. Nat Commun 5: 4910.

[218] Seo SW, Kim D, Szubin R, Palsson BO (2015) Genome-wide reconstruction of oxyr and soxrs transcriptional regulatory networks under oxidative stress in escherichia coli k-12 mg1655. Cell Reports 12: 1289–1299.

[219] Sun Y, Fleming RM, Thiele I, Saunders MA (2013) Robust flux balance analysis of multiscale biochemical reaction networks. BMC Bioinformatics 14: 240.

[220] Ma D, Saunders MA (2015) Solving multiscale linear programs using the simplex method in quadruple precision. In: Numerical Analysis and Optimization. pp. 223–235.

[221] Ma D, Yang L, Fleming RM, Thiele I, Palsson BO, Saunders MA (2016) Reliable and efficient solution of genome-scale models of Metabolism and macromolecular Expression. arXiv:160600054 [q-bioMN] .

[222] Wunderling R (1996) Paralleler und objektorientierter Simplex-Algorithmus. Ph.D. thesis, Technische Universität Berlin. `https://opus4.kobv.de/opus4-zib/frontdoor/index/index/docId/538`. Retrieved September 19, 2016.

[223] Volkmer B, Heinemann M (2011) Condition-dependent cell volume and concentration of *Escherichia coli* to facilitate data conversion for systems biology modeling. PloS ONE 6: e23126.

[224] Westerhoff HV, Palsson BO (2004) The evolution of molecular biology into systems biology. Nat Biotechnol 22: 1249–1252.

[225] Ge H, Walhout AJM, Vidal M (2003) Integrating 'omic' information: a bridge between genomics and systems biology. Trends Genet 19: 551–560.

[226] García-Roa M, Del Carmen Vicente-Ayuso M, Bobes AM, Pedraza AC, González-Fernández A, Martín MP, Sáez I, Seghatchian J, Gutiérrez L (2017) Red blood cell storage time and transfusion: current practice, concerns and future perspectives. Blood Transfus 15: 222–231.

[227] D'alessandro A, Nemkov T, Reisz J, Dzieciatkowska M, Wither MJ, Hansen KC (2017) Omics markers of the red cell storage lesion and metabolic linkage. Blood Transfus 15: 137–144.

[228] Nielsen J (2017) Systems biology of metabolism. Annu Rev Biochem 86: 245–275.

[229] Johnson CH, Ivanisevic J, Siuzdak G (2016) Metabolomics: beyond biomarkers and towards mechanisms. Nat Rev Mol Cell Biol 17: 451–459.

[230] Zamboni N, Saghatelian A, Patti GJ (2015) Defining the metabolome: size, flux, and regulation. Mol Cell 58: 699–706.

[231] Larance M, Lamond AI (2015) Multidimensional proteomics for cell biology. Nat Rev Mol Cell Biol 16: 269–280.

[232] Han X (2016) Lipidomics for studying metabolism. Nat Rev Endocrinol 12: 668–679.

[233] D'Alessandro A, Hansen KC, Silliman CC, Moore EE, Kelher M, Banerjee A (2015) Metabolomics of AS-5 RBC supernatants following routine storage. Vox Sang 108: 131–140.

[234] Dumont LJ, D'Alessandro A, Szczepiorkowski ZM, Yoshida T (2016) CO2 -dependent metabolic modulation in red blood cells stored under anaerobic conditions. Transfusion 56: 392–403.

[235] D'Alessandro A, Gray AD, Szczepiorkowski ZM, Hansen K, Herschel LH, Dumont LJ (2017) Red blood cell metabolic responses to refrigerated storage, rejuvenation, and frozen storage. Transfusion 57: 1019–1030.

[236] Nemkov T, Hansen KC, D'Alessandro A (2017) A three-minute method for high-throughput quantitative metabolomics and quantitative tracing experiments of central carbon and nitrogen pathways. Rapid Commun Mass Spectrom 31: 663–673.

[237] Yuan J, Bennett BD, Rabinowitz JD (2008) Kinetic flux profiling for quantitation of cellular metabolic fluxes. Nat Protoc 3: 1328–1340.

[238] Bordbar A, Palsson BØ (2012) Moving toward Genome-Scale kinetic models: The mass action stoichiometric simulation approach. In: Koyutürk M, Subramaniam S, Grama A, editors, Functional Coherence of Molecular Networks in Bioinformatics, New York, NY: Springer New York. pp. 201–220.

[239] McCloskey D, Young JD, Xu S, Palsson BO, Feist AM (2016) Modeling method for increased precision and scope of directly measurable fluxes at a Genome-Scale. Anal Chem 88: 3844–3852.

[240] Yuan P, D'Lima NG, Slavoff SA (2017) Comparative membrane proteomics reveals a nonannotated e. coli heat shock protein. Biochemistry .

[241] Wei L, Gregorich ZR, Lin Z, Cai W, Jin Y, McKiernan SH, McIlwain S, Aiken JM, Moss RL, Diffee GM, Ge Y (2017) Novel sarcopenia-related alterations in sarcomeric protein post-translational modifications in skeletal muscles identified by top-down proteomics. Mol Cell Proteomics .

[242] Mayers MD, Moon C, Stupp GS, Su AI, Wolan DW (2017) Quantitative metaproteomics and Activity-Based probe enrichment reveals significant alterations in protein expression from a mouse model of inflammatory bowel disease. J Proteome Res 16: 1014–1026.

[243] D'alessandro A, Dzieciatkowska M, Nemkov T, Hansen KC (2017) Red blood cell proteomics update: is there more to discover? Blood Transfus 15: 182–187.

[244] Low TY, Seow TK, Chung MCM (2002) Separation of human erythrocyte membrane associated proteins with one-dimensional and two-dimensional gel electrophoresis followed by identification with matrix-assisted laser desorption/ionization-time of flight mass spectrometry. Proteomics 2: 1229–1239.

[245] Pasini EM, Kirkegaard M, Mortensen P, Lutz HU, Thomas AW, Mann M (2006) In-depth analysis of the membrane and cytosolic proteome of red blood cells. Blood 108: 791–801.

[246] Goodman SR, Kurdia A, Ammann L, Kakhniashvili D, Daescu O (2007) The human red blood cell proteome and interactome. Exp Biol Med 232: 1391–1408.

[247] Roux-Dalvai F, Gonzalez de Peredo A, Simó C, Guerrier L, Bouyssié D, Zanella A, Citterio A, Burlet-Schiltz O, Boschetti E, Righetti PG, Monsarrat B (2008) Extensive analysis of the cytoplasmic proteome of human erythrocytes using the peptide ligand library technology and advanced mass spectrometry. Mol Cell Proteomics 7: 2254–2269.

[248] Shevchenko A, Simons K (2010) Lipidomics: coming to grips with lipid diversity. Nat Rev Mol Cell Biol 11: 593–598.

[249] Triebl A, Trötzmüller M, Hartler J, Stojakovic T, Köfeler HC (2017) Lipidomics by ultrahigh performance liquid chromatography-high resolution mass spectrometry and its application to complex biological samples. J Chromatogr B Analyt Technol Biomed Life Sci 1053: 72–80.

[250] Quehenberger O, Armando AM, Brown AH, Milne SB, Myers DS, Merrill AH, Bandyopadhyay S, Jones KN, Kelly S, Shaner RL, Sullards CM, Wang E, Murphy RC, Barkley RM, Leiker TJ, Raetz CRH, Guan Z, Laird GM, Six DA, Russell DW, McDonald JG, Subramaniam S, Fahy E, Dennis EA (2010) Lipidomics reveals a remarkable diversity of lipids in human plasma. J Lipid Res 51: 3299–3305.

[251] Song J, Liu X, Wu J, Meehan MJ, Blevitt JM, Dorrestein PC, Milla ME (2013) A highly efficient, high-throughput lipidomics platform for the quantitative detection of eicosanoids in human whole blood. Anal Biochem 433: 181–188.

[252] Christinat N, Morin-Rivron D, Masoodi M (2016) High-Throughput quantitative lipidomics analysis of nonesterified fatty acids in human plasma. J Proteome Res 15: 2228–2235.

[253] Gallego SF, Sprenger RR, Neess D, Pauling JK, Færgeman NJ, Ejsing CS (2017) Quantitative lipidomics reveals age-dependent perturbations of whole-body lipid metabolism in ACBP deficient mice. Biochim Biophys Acta 1862: 145–155.

[254] Hyötyläinen T, Orešič M (2015) Analytical lipidomics in metabolic and clinical research. Trends Endocrinol Metab 26: 671–673.

[255] Selvaraj S, Krishnaswamy S, Devashya V, Sethuraman S, Krishnan UM (2015) Influence of membrane lipid composition on flavonoid-membrane interactions: Implications on their biological activity. Prog Lipid Res 58: 1–13.

[256] Aoun M, Corsetto PA, Nugue G, Montorfano G, Ciusani E, Crouzier D, Hogarth P, Gregory A, Hayflick S, Zorzi G, Rizzo AM, Tiranti V (2017) Changes in red blood cell membrane lipid composition: A new perspective into the pathogenesis of PKAN. Mol Genet Metab 121: 180–189.

[257] Georgatzakou HT, Antonelou MH, Papassideri IS, Kriebardis AG (2016) Red blood cell abnormalities and the pathogenesis of anemia in end-stage renal disease. Proteomics Clin Appl 10: 778–790.

[258] Wu H, Bogdanov M, Zhang Y, Sun K, Zhao S, Song A, Luo R, Parchim NF, Liu H, Huang A, Adebiyi MG, Jin J, Alexander DC, Milburn MV, Idowu M, Juneja HS, Kellems RE, Dowhan W, Xia Y (2016) Hypoxia-mediated impaired erythrocyte lands' cycle is pathogenic for sickle cell disease. Sci Rep 6: 29637.

[259] Goozee K, Chatterjee P, James I, Shen K, Sohrabi HR, Asih PR, Dave P, Ball B, ManYan C, Taddei K, Chung R, Garg ML, Martins RN (2017) Alterations in erythrocyte fatty acid composition in preclinical alzheimer's disease. Sci Rep 7: 676.

[260] Liebisch G, Ekroos K, Hermansson M, Ejsing CS (2017) Reporting of lipidomics data should be standardized. Biochim Biophys Acta 1862: 747–751.

[261] Yang K, Han X (2016) Lipidomics: Techniques, applications, and outcomes related to biomedical sciences. Trends Biochem Sci 41: 954–969.

[262] Fahy E, Subramaniam S, Murphy RC, Nishijima M, Raetz CRH, Shimizu T, Spener F, van Meer G, Wakelam MJO, Dennis EA (2009) Update of the LIPID MAPS comprehensive classification system for lipids. J Lipid Res 50 Suppl: S9–14.

[263] Geris L, Gomez-Cabrero D (2015) Uncertainty in Biology: A Computational Modeling Approach. Springer.

[264] Hanson SM, Ekins S, Chodera JD (2015) Modeling error in experimental assays using the bootstrap principle: understanding discrepancies between assays using different dispensing technologies. J Comput Aided Mol Des 29: 1073–1086.

[265] Yizhak K, Benyamini T, Liebermeister W, Ruppin E, Shlomi T (2010) Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. Bioinformatics 26: i255–60.

[266] Ebrahim A, Brunk E, Tan J, O'Brien EJ, Kim D, Szubin R, Lerman JA, Lechner A, Sastry A, Bordbar A, Feist AM, Palsson BO (2016) Multi-omic data integration enables discovery of hidden biological regularities. Nat Commun 7: 13091.

[267] Stasi M, De Luca M, Bucci C (2015) Two-hybrid-based systems: powerful tools for investigation of membrane traffic machineries. J Biotechnol 202: 105–117.

[268] Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ, von Mering C (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res 45: D362–D368.

[269] Zhong Q, Pevzner SJ, Hao T, Wang Y, Mosca R, Menche J, Taipale M, Taşan M, Fan C, Yang X, Haley P, Murray RR, Mer F, Gebreab F, Tam S, MacWilliams A, Dricot A, Reichert P, Santhanam B, Ghamsari L, Calderwood MA, Rolland T, Charloteaux B, Lindquist S, Barabási AL, Hill DE, Aloy P, Cusick ME, Xia Y, Roth FP, Vidal M (2016) An inter-species protein-protein interaction network across vast evolutionary distance. Mol Syst Biol 12: 865.

[270] Bosman GJCGM (2016) The proteome of the red blood cell: An auspicious source of new insights into Membrane-Centered regulation of homeostasis. Proteomes 4.

[271] Kholodenko BN, Kiyatkin A, Bruggeman FJ, Sontag E, Westerhoff HV, Hoek JB (2002) Untangling the wires: A strategy to trace functional interactions in signaling and gene networks. Proceedings of the National Academy of Sciences 99: 12841–12846.

[272] Reed JL, Famili I, Thiele I, Palsson BO (2006) Towards multidimensional genome annotation. Nature Reviews Genetics 7: 130–141.

[273] Stelling J (2004) Mathematical models in microbial systems biology. Current Opinion in Microbiology 7: 513–518.

[274] Moses V, Sharp PB (1972) Intermediary metabolite levels in escherichia coli. Journal of General Microbiology 71: 181–190.

[275] Tepper N, Noor E, Amador-Noguez D, Haraldsdóttir HS, Milo R, Rabinowitz J, Liebermeister W, Shlomi T (2013) Steady-state metabolite concentrations reflect a balance between maximizing enzyme efficiency and minimizing total metabolite load. PLoS ONE 8: e75370.

[276] Yaginuma H, Kawai S, Tabata KV, Tomiyama K, Kakizuka A, Komatsuzaki T, Noji H, Imamura H (2014) Diversity in ATP concentrations in a single bacterial cell population revealed by quantitative single-cell imaging. Scientific Reports 4.

[277] Zelezniak A, Sheridan S, Patil KR (2014) Contribution of network connectivity in determining the relationship between gene expression and metabolite concentration changes. PLoS Computational Biology 10: e1003572.

[278] Lee SJ, Trostel A, Adhya S (2014) Metabolite changes signal genetic regulatory mechanisms for robust cell behavior. mBio 5: e00972–13–e00972–13.

[279] Beard DA, Qian H (2007) Relationship between thermodynamic driving force and one-way fluxes in reversible processes. PLoS ONE 2: e144.

[280] Henry CS, Broadbelt LJ, Hatzimanikatis V (2007) Thermodynamics-based metabolic flux analysis. Biophys J 92: 1792–1805.

[281] Hoppe A, Hoffmann S, Holzhütter HG (2007) Including metabolite concentrations into flux balance analysis: thermodynamic realizability as a constraint on flux distributions in metabolic networks. BMC Systems Biology 1: 23.

[282] Garg S, Yang L, Mahadevan R (2010) Thermodynamic analysis of regulation in metabolic networks using constraint-based modeling. BMC Research Notes 3: 125.

[283] Yurkovich JT, Yurkovich BJ, Dräger A, Palsson BO, King ZA (2017) A padawan programmer's guide to developing software libraries. Cell Systems 5: 431–437.

[284] Kleinman S, Busch MP, Murphy EL, Shan H, Ness P, and SAG (2013) The national heart, lung, and blood institute recipient epidemiology and donor evaluation study (REDS-III): a research program striving to improve blood donor and transfusion recipient outcomes. Transfusion 54: 942–955.