METABOLIC MODELING: INTEGRATION WITH MULTI-OMIC DATASETS, STATISTICAL EVALUATION, AND APPLICATION TO OUR UNDERSTANDING OF PHOTOSYNTHETIC CARBON ASSIMILATION

By

Joshua Akito Matthew Kaste

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Biochemistry and Molecular Biology – Doctor of Philosophy

2024

ABSTRACT

Many biotechnological efforts in plants have been proposed to address issues around climate change, sustainability, and food security. These include the modification of the oilseed crop *Camelina sativa* to improve the efficiency with which it produces oil from captured carbon, to the suppression of photorespiration and enhancement of yield in crops by engineering a Carbon Concentrating Mechanism into them. However, research and development efforts to use biotechnological interventions to improve plants and microbes have been hampered by their extreme complexity. Many of these bioengineering efforts seek to modify the rates of *in vivo* biochemical reactions – referred to hereafter as fluxes – in order to improve the efficiency or yield with which a desired product(s) is produced. Therefore, these efforts require the characterization and modification of the organism's metabolic activity. As in other areas of engineering, these processes can be aided by the use of quantitative modeling.

In the case of metabolic modeling, multiple approaches already exist. These include the use of simplified compartmental models (Chapter 2) and enzyme-based modeling (Chapter 4), as well as constraint-based approaches such as the linear-optimization-based Flux Balance Analysis (Chapter 3) and the nonlinear regression based Metabolic Flux Analysis (Chapter 2). Although these methods are rarely used in tandem, they are mathematically interrelated and can be used to validate and/or corroborate one another's findings. Indeed, previous literature in the area of metabolic modeling has frequently paid short shrift to the importance of validation and model selection in this area of study, calling into question the biological relevance and accuracy of many modeling studies.

I begin with a discussion of challenges and prospects for future development in the statistical evaluation of metabolic models (Chapter 1), emphasizing the need for cross-comparison of multiple techniques, which I put into practice in later chapters. Emphasis is put on the need for careful validation of ^{13}C-Metabolic Flux Analysis findings using multiple lines of evidence and on the usefulness of validating Flux Balance Analysis flux predictions using estimates from ^{13}C-Metabolic Flux Analysis.

I apply these methodological insights to studies of *Camelina sativa* and its relative *Arabidopsis thaliana*. I start with a ^{13}C-Metabolic Flux Analysis study of the metabolism of photosynthesizing leaves of *Camelina sativa* (Chapter 2). By modeling the stable isotopic labeling levels of Calvin-Benson intermediates with a series of polyexponential models, I corroborate the

study's ^{13}C-Metabolic Flux Analysis findings, resulting in a more detailed model of *C. sativa*'s leaf metabolism and resolving a decades-old mystery in the labeling of these metabolites. Following this, I present a novel method of incorporating multi-omic datasets into FBA predictions of metabolic fluxes in the closely related organism *A. thaliana* (Chapter 3). I demonstrate that this new method successfully improves agreement between FBA and ^{13}C-MFA flux maps of *A. thaliana*, setting the stage for improved FBA and metabolic engineering insight into the related *C. sativa*.

Finally, I turn my attention to reaction-diffusion modeling. In Chapter 4, I apply enzyme-based and spatial modeling techniques to understand the net CO_2 fixation and light-use efficiency implications of incorporating a biophysical Carbon Concentrating Mechanism into a C3 plant.

After some concluding remarks in Chapter 5, I present two additional studies that are related either to the quantitative modeling of plants or metabolic networks, but which are not directly related to the rest of the investigations in this thesis. First, I investigate the extent to which the kinds of tissue-specific gene expression patterns utilized in Chapter 3 are conserved across all flowering plants, providing evidence that such an approach may be broadly usable in plant metabolic modeling. Finally, I present interactive educational materials that teach the underlying theory for all of the metabolic modeling approaches used in the above studies. I implemented these materials into an intensive workshop series put on at Michigan State University and demonstrate that participants' self-assessed confidence in the techniques taught increased significantly.

Copyright by
JOSHUA AKITO MATTHEW KASTE
2024

To those who ask one more question when they have already been given an answer

ACKNOWLEDGEMENTS

Throughout my time at Michigan State University, I have had the exceedingly good luck to study under, learn from, and work and collaborate with a great number of kind, hard-working, and knowledgeable friends and colleagues. I also would not be here without the support of people in my life who have believed in me over the years. I cannot hope to name everyone who has helped me in my journey. Having said that, here's my best attempt:

Dr. Yair Shachar-Hill has been a steadfast supporter of my work and a wonderful advisor. He has provided me with valuable scientific and professional advice while at the same time allowing me the freedom to develop as an independent scholar. He is a model of what a scientific mentor should be, and I can only hope to provide my own mentees the same combination of patience, insight, and guidance in the future.

My committee – Dr. Shachar-Hill, Dr. Tom Sharkey, Dr. Erich Grotewold, Dr. Michaela TerAvest, and Dr. Chih-Li Sung – have provided invaluable guidance and suggestions throughout my Ph.D. and I am immensely grateful for their support.

Many members of my committee, and others here at Michigan State University, have also been wonderful collaborators. Dr. Tom Sharkey, Dr. Michaela TerAvest, Dr. Chih-Li Sung, Dr. Berkley Walker, Dr. Dan Chitwood, Dr. Bob VanBuren, Dr. Yuan Xu, Dr. Kathryne Ford, Miles Roberts, Kenia Segura-Aba, Dr. Saurabh Palande, Anne Steensma, and many more trainees from NRT-IMPACTS program have played key roles in the studies I present in this thesis, and I could not have done all of this without them.

The current and former members of the Shachar-Hill lab made me feel welcome in Michigan and helped keep me sane during the pandemic. I'd like to sincerely thank Dr. Danielle Hoffmann, Dr. Shawna Rowe, Dr. Yuan Xu, Dr. Na Pang, Anne Steensma, Peter Koroma, and Antwan Green for their camaraderie and good humor. I would like to also thank Antwan Green, who worked with me as an undergraduate researcher, for his patience as I learned how best to serve as a research mentor.

I would also like to acknowledge people whose support and passion for science inspired me to keep going and without whom I would not have ended up getting my Ph.D. here at MSU: Dr. Michael Milgroom, Dr. Mickey Drott, Dr. Carolyn Young, Dr. Mihwa Yi, Dr. Farhad Ghavami, and Brenda Johnson.

I also would not be here today if it were not for my parents, Matthew and Miki Kaste, as

well as the support of my late grandfather Hubert Kaste.

Finally, I would like to acknowledge the unwavering support of my lovely wife Veronica, who has stuck with me through the many ups and downs of my personal and professional life for close to a decade. And, of course, our wonderful pets Duchess, Gabriel, and Makoto, who have been sources of joy through good times and bad.

PREFACE

> "Two things that in my opinion reinforce one another and remain eternally true are: Do not quench your inspiration and your imagination, do not become the slave of your model; and again: Take the model and study it, otherwise your inspiration will never become plastically concrete." – Vincent van Gogh, in a letter to Theo van Gogh

There are many ways of conceptualizing what it is we do as "scientists" and what it is about the scientific process – nebulous and ill-defined as it is – that makes it so unusually effective at making predictions about the natural world and enabling new technologies. In my mind, its power derives in large part from the continuous back-and-forth between empirical measurement and model-making.

That the interplay between theory and observation is key to the scientific method is by no means a new observation, but I think we should take a moment to consider why we need both theory and observation and not just one or the other to make sense of the world. When we endeavor to understand something using reason alone, we end up writing something like Plato's *Timaeus*: admirably creative and logically constructed, but unmoored entirely from the inconvenient details of how things actually work. On the other hand, if we were to try to understand the world using observation alone, we would accomplish nothing but an accumulation of miscellaneous measurements with no ability to build out of them an understanding of natural phenomena. To summarize: there are an infinite number of "reasonable" explanations of how the world works and an infinite number of observations that could be made about it, but it is only by combining these explanations and observations that we can develop a robust understanding.

In reality, this intricate dance between theory and observation is much harder than we usually like to admit and, as scientists, we are prone to missteps. Here, I will focus on those areas of science where we gather quantitative data about natural phenomena and then build and evaluate mechanistic, quantitative models to try to explain these data. We may fit a model to a dataset, but find that this model generalizes poorly. We might conclude, based on a model fit, that our pet hypothesis about an important characteristic of a system was correct, not realizing that a simpler model that does not invoke that characteristic at all actually fits our observations even better. We might rigorously validate some aspect of a model, but then use it in ways that have not been validated. Or, we might fixate on using one specific lens to examine our

observations, when the use of multiple lenses to look at the system from different angles and at different scales might reveal more about it and, through consilience, make us more confident in our findings.

But where there are problems, there are also opportunities, so I have worked throughout my Ph.D. to identify and address such challenges and, in doing so, generate new insights in the area of the quantitative modeling of photosynthetic metabolism. Towards this end, in this thesis I present work on (i) the development and exploration of different modeling techniques, (ii) the use these different techniques as lenses through which to contextualize data, and (iii) the statistical interrogation of model fits and, more generally, the use of modeling in the study of metabolism.

The admittedly abstract motivations that inspired much of the work in this thesis has resulted in it being composed of an unusually diverse set of studies, unified loosely by a focus on rigorous development and analysis of models describing photosynthetic metabolism. Despite these abstract motivations, however, the products of the work are concrete, and include:

1. A resolution to a many-decades-old mystery in the labeling patterns of photosynthetic intermediates, pointing towards a previously underappreciated cycling between vacuolar and cytosolic sugars in photosynthesizing leaves.
2. The first demonstration of an algorithm that successfully improves the accuracy of Flux Balance Analysis predictions in a whole-plant model using transcriptomic or proteomic data.
3. Spatially-resolved reaction-diffusion models that refine our understanding of the metabolic tradeoffs involved in the use of Carbon Concentrating Mechanisms.

I felt very privileged during my time here at Michigan State University to have been allowed – by my research mentor and my funding sources – to pursue a rather unorthodox set of questions. It is my hope that you, the reader, will also appreciate how the studies described in the following chapter interrelate and contribute to our descriptions of photosynthesis and *how* we arrive at these descriptions in the first place.

TABLE OF CONTENTS

Chapter 1

Model validation and selection in metabolic flux analysis and flux balance analysis

This research was published in:

J. A. M. Kaste, Y. Shachar-Hill, Model validation and selection in metabolic flux analysis and flux balance analysis. *Biotechnology Progress*, e3413 (2023).

1.1. Preface

The conversations and ideas that gave rise to this paper started between me, Dr. Shachar-Hill, Dr. Xu, and Dr. Sharkey while we were conducting the study that ultimately became Chapter 2. During the course of that study, I became increasingly interested in the general questions surrounding the statistical evaluation, validation, and model selection practices of ^{13}C-MFA flux maps. The idea of using a simpler compartmental modeling strategy to corroborate the gross architecture of a ^{13}C-MFA model, in the absence of a well worked out general model selection strategy for ^{13}C-MFA, did make it into the paper presented in Chapter 2. However, the conversations that Dr. Shachar-Hill and I continued to have on this topic ended up ranging far beyond the context of that study. Moreover, my work on validating the predictions of a novel FBA implementation using MFA flux estimates, which I present in Chapter 3, gave me insight into validation practices on the FBA side of constraint-based modeling.

After some preliminary literature review, it became quite clear that despite the importance of validation and model selection to the area of constraint-based metabolic modeling, shockingly little had been said on the topic. Due to the paucity of discussion and analysis in this area, a straightforward review paper would have been of little use. So, we decided to write a perspective article that summarizes common practices and their drawbacks, while also presenting our original insights and perspectives on the future of this area.

The paper presented in this chapter has been published in the journal *Biotechnology Progress*. I am first and corresponding author on the study.

1.2. Abstract

^{13}C-Metabolic Flux Analysis (^{13}C-MFA) and Flux Balance Analysis (FBA) are widely used to investigate the operation of biochemical networks in both biological and biotechnological research. Both methods use metabolic reaction network models of metabolism operating at steady state so that reaction rates (fluxes) and the levels of metabolic intermediates are constrained to be invariant. They provide estimated (MFA) or predicted (FBA) values of the fluxes through the network *in vivo*, which cannot be measured directly. These fluxes can shed light on basic biology and have been successfully used to inform metabolic engineering strategies. Several approaches have been taken to test the reliability of estimates and predictions from constraint-based methods and to compare alternative model architectures. Despite advances in other areas of the statistical evaluation of metabolic models, such as the quantification of flux

estimate uncertainty, validation and model selection methods have been underappreciated and underexplored. We review the history and state-of-the-art in constraint-based metabolic model validation and model selection. Applications and limitations of the χ^2-test of goodness-of-fit, the most widely used quantitative validation and selection approach in ^{13}C-MFA, are discussed, and complementary and alternative forms of validation and selection are proposed. A combined model validation and selection framework for ^{13}C-MFA incorporating metabolite pool size information that leverages new developments in the field is presented and advocated for. Finally, we discuss how adopting robust validation and selection procedures can enhance confidence in constraint-based modeling as a whole and ultimately facilitate more widespread use of FBA in biotechnology.

1.3. Introduction

The set of biochemical reaction rates in the metabolic network of a living system (its flux map) represents an integrated functional phenotype that emerges from multiple layers of biological organization and regulation, including the genome, transcriptome, and proteome (Nielsen, 2003). The study of metabolic fluxes is therefore important for systems biology, rational metabolic engineering, and synthetic biology. A grand challenge of systems biology is building an integrated mechanistic understanding of the operation of living organisms across these levels of regulation (Spivey, 2004) – an understanding that goes beyond statistical or correlative descriptions, however useful these can be. Meeting this challenge requires fluxes to be accurately predicted from network structure using explicit rules or hypotheses and reliably estimated using experimental data. Fluxes are also critical to many biotechnological and metabolic engineering applications. Examples such as the development of lysine hyper-producing strains of *Corynebacterium glutamicum* (Koffas et al., 2003; Koffas and Stephanopoulos, 2005; Becker et al., 2011) and the rewiring of *E. coli*'s metabolism to make it grow chemoautotrophically (Gleizer et al., 2019) attest to the usefulness of these techniques. As the scale and complexity of integrative systems biology and biological engineering efforts increase, so too will the need for reliable and robust estimates of fluxes.

I*n vivo* fluxes cannot be directly measured, necessitating modeling approaches to estimate or predict them. The most commonly used approaches for metabolic modeling are the constraint-based modeling frameworks of ^{13}C-Metabolic Flux Analysis (^{13}C-MFA) and Flux Balance Analysis (FBA). Both require a metabolic network consisting of metabolites linked by

biochemical reactions to be defined using the biochemical literature, knowledge of the enzymes and transporters expressed from the genome, and physico-chemical rules. In ^{13}C-MFA, atom mappings describing the positions and interconversions of the carbon atoms in reactants and products are also included in the model. These methods assume that the system is at metabolic steady-state, such that the concentrations of all metabolic intermediates and reaction rates are constant (Antoniewicz, 2015). External fluxes, such as the uptake of a substrate or the rate of production of new cells or a product, are also measured and used to constrain the possible flux ranges. These assumptions and constraints define a "solution space" containing all flux maps consistent with them but are typically insufficient to pinpoint a unique flux map.

In ^{13}C-MFA, isotopic labeling data is used to identify a particular solution within the solution space. ^{13}C-labeled substrates are fed to the system under investigation and the endpoint labeling, or time-course labeling in Isotopically Nonstationary Metabolic Flux Analysis (INST-MFA), of metabolites is measured using mass spectrometry and/or NMR techniques (Antoniewicz, 2015; Cheah and Young, 2018). Given a metabolic network, a flux map, and information about the labeled substrate fed into the system, the label distribution through all the metabolites in a network can be solved analytically. However, ^{13}C-MFA works backwards from measured label distributions to flux maps by minimizing the differences between measured and estimated Mass Isotopomer Distribution (MID) values by varying flux estimates (Jazmin et al., 2014). For INST MFA pool size measurements can also be included in the minimization process.

In FBA, linear optimization is used to identify a flux map (or set of flux maps) from the solution space (Orth et al., 2010b). This is the map(s) for which the sum of one or more fluxes (the objective function) is maximized or minimized. Objective functions frequently represent measures of efficiency, including the maximization of growth rate or product formation or the minimization of total flux (Holzhütter, 2004). Such functions may embody hypotheses about what the *in vivo* system has been evolutionarily tuned to optimize, or questions about the operational capacity of that system under particular conditions. Since the objective function, together with the network architecture and empirical and/or theoretical constraints introduced by the modeler, is a key determinant of the flux maps generated by FBA, careful selection, justification, and, ideally, validation of objective functions is crucial. As shown in Schnitzer et al., (2022), alternative objective functions can, and should, be evaluated to identify those that result in the best agreement with experimental data. In many cases, the constraints – typically on

external fluxes – imposed during an FBA optimization result in a set of viable flux maps (a solution space) rather than a single map. In such cases, related techniques, including Flux Variability Analysis (Mahadevan and Schilling, 2003) and random sampling (Schellenberger and Palsson, 2009; Bordel et al., 2010; Megchelenbrink et al., 2014; Haraldsdóttir et al., 2017) can be used to characterize the set of flux maps consistent with the set constraints. The computational tractability and small amount of experimental data necessary to perform FBA allow the analysis of Genome-Scale Stoichiometric Models (GSSMs). These models incorporate all known reactions believed to occur in an organism based on a combination of genome annotation and manual curation. Additional linear-optimization-based methods for solving GSSMs using the FBA framework have been developed and are sometimes used together with FBA. These include Minimization of Metabolic Adjustment (MOMA) (Segrè et al., 2002), and Regulatory On/Off Minimization (ROOM) (Shlomi et al., 2005), as well as a host of methods that incorporate omic data into the optimization process [e.g., (Åkesson et al., 2004; Becker and Palsson, 2008; Tian and Reed, 2018; Pandey et al., 2019; Ravi and Gunawan, 2021)]. FBA and its related methods are sometimes used to analyze models other than true GSSMs, such as "core" models that focus on central metabolic processes that conduct the large majority of flux (Orth et al., 2010a). When discussing validation, however, the same principles apply to all of these linear optimization methods and across the different model scales. For the sake of simplicity, we will be using "FBA" to refer to this family of methods generally and will refer to the medium- to large-scale models used with these methods as "FBA models."

Progress has been made in improving the statistical rigor and reliability of flux estimates and characterizing uncertainty in estimates and predictions. For example, in MFA, the development of effective methods for flux uncertainty estimation (Antoniewicz et al., 2006) allows researchers to better quantify confidence in flux predictions and, where appropriate, to gather additional data to better support their conclusions. Bayesian techniques for the characterization of uncertainties in flux estimates derived from isotopic labeling have also been presented (Theorell et al., 2017). On the experimental side of MFA, there have been advances in designing and implementing parallel labeling experiments, wherein the labeling patterns obtained using multiple tracers are simultaneously fit to generate a single ^{13}C-MFA flux map. This enables more precise estimation of fluxes than experiments with individual tracers or tracer combinations allow (Chang et al., 2008; Crown et al., 2012; Crown and Antoniewicz, 2012;

Leighty and Antoniewicz, 2013; Millard et al., 2014; Crown et al., 2015; Crown et al., 2016; Beyß et al., 2021). Greater resolution in isotopic labeling data through the use of tandem mass spectrometry techniques, which allow for the quantification of positional labeling, can also improve the precision of modeled fluxes, as described in Choi and Antoniewicz (2019) and Wang et al., (2021). Recent years have also seen developments in FBA meant to improve the reliability of its predictions. For example, studies have characterized the impact of departures from metabolic steady state and devised methods to account for uncertainties in biomass compositions [e.g., (Dinh et al., 2022; Choi et al., 2023)]. The many sources of uncertainty when working with FBA and genome-scale models, and attempts to characterize and mitigate this uncertainty, have been reviewed elsewhere (Bernstein et al., 2021).

In this review, we specifically focus on the validation of flux predictions and estimates from constraint-based modeling studies and the selection of well-supported model architectures, which have received less attention and specific treatment in the literature. How can MFA and FBA researchers validate the accuracy of their estimates and predictions? These flux analysis methods also require researchers to make choices about the network structure of the model to be used. This leads to questions of model selection; that is, how do we select the most statistically justified model from among the alternatives? Validation and model selection are key to improving the fidelity of model-derived fluxes to the real *in vivo* ones. The fields of systems and synthetic biology have seen substantial development of model selection and validation practices (Kirk et al., 2013; Gross and MacLeod, 2017), but these topics are not frequently discussed in the metabolic modeling literature. Previous reviews and methods papers have touched on the use of tools like the χ^2-test of goodness-of-fit for the validation of MFA models (Antoniewicz, 2018; Long and Antoniewicz, 2019a). However, to our knowledge, no reviews covering the various methods for validating FBA predictions exist, nor have previous reviews discussed the various limitations of the χ^2-test. Moreover, previous reviews have not addressed the most recent improvements in model selection in ^{13}C-MFA, which have not been adequately incorporated into routine practice. Addressing these topics explicitly is important for practitioners as they carry out their work. It is also important for readers of the flux analysis literature, who must understand the assumptions, tests of validity, and model selection techniques underlying what they are reading.

Although only a subset of research groups conduct both FBA and MFA modeling, we believe most metabolic modeling practitioners and consumers read literature containing both

modeling paradigms. As we highlight in this review, some similar themes emerge when examining the validation of both FBA and MFA flux maps. Finally, one of the most robust validations that can be conducted for FBA predictions is comparison against MFA estimated fluxes, which makes simultaneously considering the validity of both FBA and MFA flux maps crucial. For these reasons, we consider both modeling approaches in this review.

We review and provide our perspective on these areas and prospects for future development, highlighting: (1) validation methods applicable to FBA flux maps; (2) approaches for validating ^{13}C-MFA flux maps; and (3) developments and prospects for model selection in ^{13}C-MFA; (4) How validation and model selection practices in ^{13}C-MFA could benefit from a greater emphasis on the isolation of training and validation datasets and; (5) the importance of corroborating flux mapping results using independent modeling and experimental techniques.

1.4. Validation techniques in FBA and ^{13}C-MFA

FBA and ^{13}C-MFA studies commonly validate the model(s) used, though there is great variation in their nature and extent. We summarize these validation strategies in **Figure 1**.

1.4.1. Validation in FBA

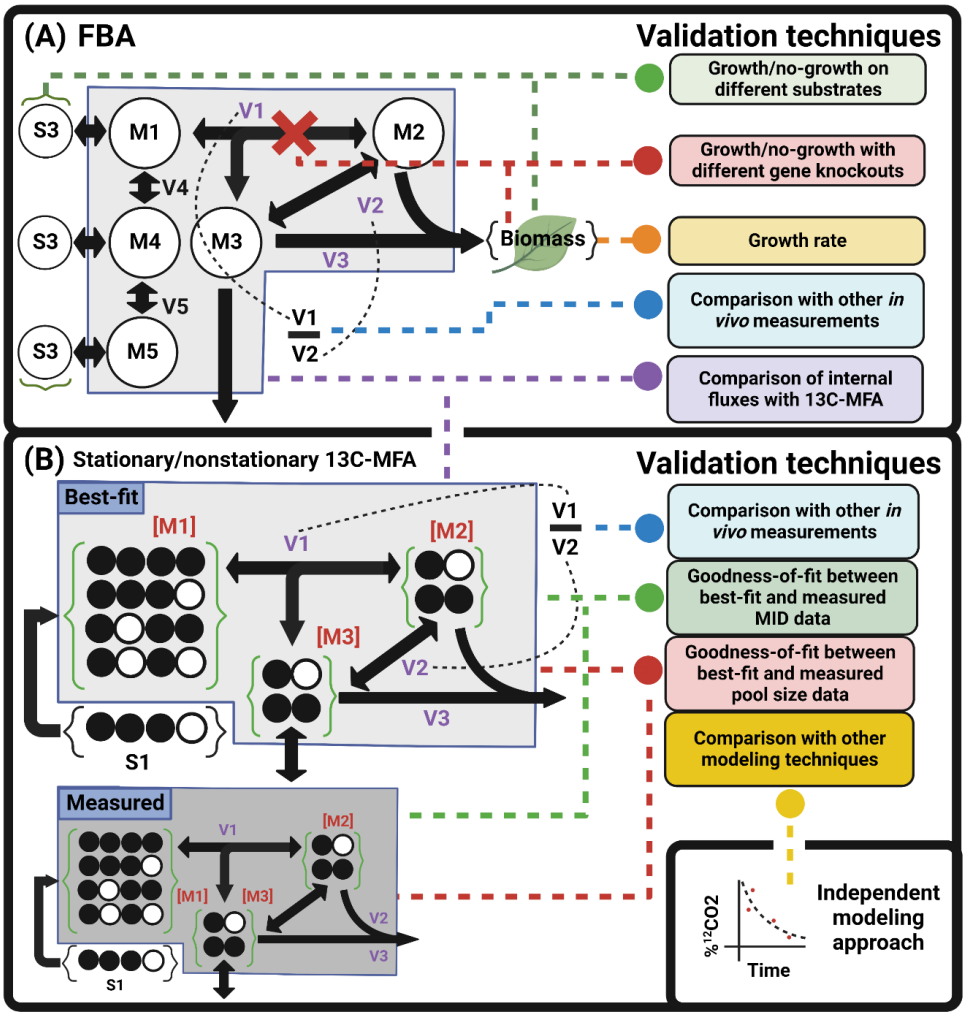


Figure 1.1: Graphical summary of validation strategies in **(A)** FBA and **(B)** ^{13}C-MFA. Dotted lines connect inputs with the associated validation technique(s). **(A)** FBA predictions can be validated by comparing growth rate or growth/no-growth phenotypes across different substrates, growth conditions, or sets of gene knockouts in silico and in vivo. Values can be calculated from flux maps and compared with experimental measurements. FBA internal flux predictions can be compared with ^{13}C-MFA fluxes. **(B)** Values can be calculated from ^{13}C-MFA flux maps and compared with an independent experimental measurement from the in vivo system. Goodness-of-fit can be assessed between simulated and measured MIDs, and simulated and measured metabolite pool sizes in INST-MFA. Flux maps can be compared with the results of independent modeling exercises. Molecules are schematically shown as connected circles of atomic positions: open circles are unlabeled, and filled circles are isotopically labeled. Abbreviations: Mn - metabolites in the metabolic network; S_n – exogenous substrates; V_i – Fluxes; $[M_n]$ – metabolite concentrations.

The COnstraint-Based Reconstruction and Analysis (COBRA) framework, implemented in software solutions such as the COBRA Toolbox (Heirendt et al., 2019) and *cobrapy* (Ebrahim et al., 2013) and widely used for FBA studies, features functions and pipelines that can be used to ensure basic functionality of models including balancing of charge, pH, and cofactors/cosubstrates, thermodynamic feasibility, and connectivity of all metabolites. Model characteristics evaluated include the inability to generate ATP without an external source of energy and the inability to synthesize biomass without adding substrates not known to be needed. Additionally, the MEMOTE (MEtabolic MOdel TEsts) pipeline contains tests to ensure, for example, that biomass precursors can be successfully synthesized in a model in a variety of growth media (Lieven et al., 2020). MEMOTE has been used to ensure appropriate stoichiometry and consistency with accepted format standards in models entered into the BiGG (Norsigian et al., 2020) model database. These forms of Quality Control are an important first step in ensuring that models are behaving appropriately and generating useful predictions. However, following these initial checks on functionality, the techniques used to validate actual model predictions are varied and not standardized. Indeed, even in the BiGG database, which is highly curated and focuses primarily on models of microbial systems, models vary in the type and extent of validation performed. Given the variety of validation procedures that appear in the literature, it is important when using an FBA model to be aware of what specific validations were used, what their limitations are, and consequently, what inferences or downstream applications are appropriate (summarized in **Table 1.1**).

Table 1.1: The most common model validation strategies in Flux Balance Analysis, what these methods tell us, limitations, and important considerations for researchers and/or readers, and examples of these methods' implementation in the literature.

Method	Information Content	Limitations	Use case	Examples
Comparison of growth/no-growth on one or more substrates	Presence/absence of reactions necessary for substrate utilization and biomass synthesis.	Validation is qualitative, only indicating the existence of metabolic routes. Does not test the accuracy of predicted internal flux values	Useful when viability/nonviability of different growth conditions is of interest. Unlike a growth-rate comparison, does not indicate whether the efficiency of biomass synthesis is realistic.	(Pinchuk et al., 2010; Ong et al., 2014; Arion et al., 2023; Coppens et al., 2023; Tec-Campos et al., 2023)
Comparison of growth rates on one or more substrates	Consistency of metabolic network, biomass composition, and maintenance costs with observed efficiency of substrate-to-biomass conversion.	Provides quantitative information on the overall efficiency of substrate conversion to biomass, but is uninformative with respect to the accuracy of internal flux predictions.	When done across multiple substrates and conditions, this validation gives confidence in the predicted efficiency with which the model produces biomass. Useful when identifying growth-limiting factors.	(Oftadeh et al., 2021; Arion et al., 2023; Coppens et al., 2023; Tec-Campos et al., 2023)
Comparison of *in vivo* and *in silico* knockout lethality	Presence/absence of biosynthetic reactions necessary for substrate use and growth.	Care is needed to reduce incorrect predictions from many different factors, including optimization method and biomass composition changes in response to knockout.	Critically important to perform when designing growth-coupled knockout strategies (Burgard et al., 2003; Tepper and Shlomi, 2009; Stanford et al., 2015).	(Gatto et al., 2015; Alzoubi et al., 2019; Oftadeh et al., 2021; Santos-Merino et al., 2023)
Comparison of FBA predictions with MFA fluxes	Accuracy of internal flux predictions.	Few MFA flux maps exist for most organisms, making this validation impossible or requiring comparison with an MFA flux map taken for very different experimental conditions.	Important when the intended use of FBA modeling requires that the predictions of specific internal flux values be accurate.	(Shinfuku et al., 2009; Machado and Herrgård, 2014; Broddrick et al., 2019; Coppens et al., 2023)

Perhaps the most common validation in FBA is comparison between FBA-predicted and empirically measured rates of growth [e.g., (Varma and Palsson, 1994; Schroeder and Saha, 2020; Feierabend et al., 2021; Arion et al., 2023; Blázquez et al., 2023; Noecker et al., 2023; Tec-Campos et al., 2023)]. One may similarly evaluate growth/no-growth in different media and/or with different carbon sources [e.g., (Ong et al., 2014; Arion et al., 2023; Blázquez et al., 2023; Heinken et al., 2023; Tec-Campos et al., 2023)]. A related approach is the comparison of *in silico* metabolite uptake/secretion with experimental measurements (Heinken et al., 2021; Blázquez et al., 2023; Heinken et al., 2023). Such evaluations give confidence in the model's basic predictions. To ensure that the accuracy of growth-rate predictions generalizes well, we strongly recommend validating growth rates on substrates or in media conditions from which biomass composition and parameters like Growth-Associated Maintenance (GAM) and Non-Growth Associated Maintenance (NGAM) costs were not experimentally derived, as done in

(Arion et al., 2023). GAM represents the energy expenditure needed to support a certain rate of biomass growth and NGAM represents the energy expenditure required for a cell or organism to survive without any net growth (Thiele and Palsson, 2010). These values may vary depending on growth conditions, so testing whether the values measured in one set of conditions generalize to others is important. Otherwise, future users may use a model with, for example, another common media composition and find – or worse yet, simply not notice – that the resulting predictions do not accurately reflect essential characteristics of the organism's actual metabolism.

A related approach involves comparing growth/no-growth of gene knockout strains to FBA predictions to address whether the metabolic pathways used in the model mirror the biological system. Experimentally verified lethal knockouts that appear nonlethal *in silico* point to alternative routes the model can use to grow. Conversely, *in silico* lethality predictions not confirmed by experiment suggest the model is missing isoforms or alternative reaction routes. Collecting the true positive, true negative, false positive, and false negative predictions from the *in silico* vs. *in vivo* lethality predictions into a confusion matrix allows for an at-a-glance evaluation of overall model accuracy and for the comparison of alternative model architectures (Santos-Merino et al., 2023). Researchers sometimes use algorithms to identify knockouts that couple biomass accumulation to flux through a reaction for biotechnological applications (Burgard et al., 2003; Tepper and Shlomi, 2009; Stanford et al., 2015). This requires that models accurately predict growth/no-growth phenotypes for gene knockouts, but previous work in a model of *Saccharomyces cerevisiae*, for example, shows that FBA performs poorly at predicting the synthetic lethality of double-knockouts, making this a serious concern (Alzoubi et al., 2019). When performing such validations, one must keep in mind that imposed constraints and decisions made during the model construction or optimization process may implicitly or explicitly add the predictions one is trying to validate into the model, rendering the exercise meaningless. This makes clear and transparent documentation of the assumptions used in the modeling process key for reviewers and readers to assess the epistemic value of the validations that are reported.

It is crucial to note that the methods discussed above do not validate the internal flux predictions made by FBA. Due to the underdetermined nature of FBA, many radically different flux maps may be compatible with, for example, the optimization of growth-rate (Mahadevan and Schilling, 2003), making validations using growth-rate or any other individual external flux

uninformative with respect to internal flux distributions. In well-characterized systems, there may be a wealth of known metabolic functionalities that an organism can carry out and evaluating whether the model can reproduce them can give some assurance of realistic model behavior. In Duarte et al., (2007) and Sigurdsson et al., (2010), 288 metabolic processes known to take place in mammalian cells were evaluated in models of human and mouse models, though it was only the ability to carry out the processes at all, and not the actual flux values, that were evaluated. In favorable cases, individual internal fluxes can be quantitatively estimated *in vivo* using independent methods and compared directly to ones from a predicted flux map to provide a powerful form of validation. For example, in a study from our group (Kaste and Shachar-Hill, 2023) the ratio of the cyclic electron flow (CEF) to linear electron flow (LEF) fluxes in photosynthesis predicted by FBA was evaluated against CEF/LEF ratios from fluorescence measurements for validation purposes. Though less specific, the sum of FBA-predicted values for fluxes that produce and/or consume a product (such as CO_2) can also be compared to experimental measurements. In addition to these approaches, there is the possibility going forward of integrating metabolomics data into the FBA prediction process [e.g., (Lee et al., 2006)] and/or comparison of FBA results against metabolomic datasets. Although, it should be noted that metabolite levels and changes in those levels in the steady-state cannot be directly interpreted in terms of fluxes, so any attempts to validate FBA results using observations in metabolomics datasets should be done with caution.

However, validations of internal flux predictions across the network require comparing FBA flux maps with high-quality ones from ^{13}C-MFA. Such validations are the most information-rich of all the methods surveyed so far and tell us the most about how well the FBA flux maps generated by a particular combination of network architecture, constraints, and objective function line up with experimental data. Unfortunately, ^{13}C-MFA flux maps are time-consuming to generate, making this “gold-standard” validation rare. To compare FBA-predicted and MFA-estimated fluxes, the model architectures must be the same, or the MFA must at least be a subnetwork of the model used for the FBA. Additionally, the empirical constraints (e.g., substrate uptake and biomass accumulation) must be the same in both cases. In cases where the growth rates predicted or constrained for an FBA flux map do not perfectly line up with those from an MFA flux map, normalization of fluxes to account for this discrepancy can be used to get an apples-to-apples comparison (Broddrick et al., 2019). The imposition of identical external

flux constraints on both the FBA and MFA models may preclude validation of the accuracy of certain external flux predictions by the FBA. However, such comparisons can be done afterwards by removing the relevant constraints. Comparison is also complicated by the underdetermined nature of most FBA optimizations, which can result in large feasible ranges for the individual fluxes being compared against the corresponding flux values obtained from ^{13}C-MFA, making the validation less stringent. FBA optimizations that assume parsimony (Holzhütter, 2004; Lewis et al., 2010) tend to yield narrower flux ranges, but this advantage may come at the cost of neglecting other plausible objective functions that might be more accurate.

Finally, when FBA-predicted and MFA-estimated flux maps disagree, assuming the experimental constraints are consistent between the two and that the person doing the comparison is confident in the MFA estimates, either the FBA network architecture or objective function could be to blame. There is not, to our knowledge, a consistent strategy for disambiguating disagreements due to architecture or objective function. If the biological/biochemical accuracy of the objective function is in question, methods for inferring objective functions using isotopic labeling data can be employed [e.g., (Gianchandani et al., 2008)], the resulting objective functions can be compared with the one being used, and discrepancies can be considered. All objective functions that relate to growth will be affected by the accuracy of the biomass composition used in the model, although in some systems central metabolic fluxes may be relatively robust to variability in the exact values of this composition (Yuan et al., 2016). In systems for which extensive biomass composition data is available, known variability in biomass composition can be incorporated during the optimization process (Choi et al., 2023). Despite these various limitations and difficulties when validating FBA using ^{13}C-MFA fluxes, some studies have evaluated the accuracy of FBA against ^{13}C-MFA-estimated flux maps [e.g., (Schuetz et al., 2007; Chen et al., 2011; Machado and Herrgård, 2014; Tian and Reed, 2018; Long and Antoniewicz, 2019b; Blázquez et al., 2023; Coppens et al., 2023)], with mixed results.

A consistent challenge when validating FBA fluxes using any method is the need to compare the FBA flux map against empirical fluxes or other measurements that were generated under similar conditions to those being simulated. For organisms or systems whose metabolic models are undergoing continual refinement, thus requiring repeated validation, community-curated and updated validation datasets generated under well-defined and carefully reported

conditions may be useful. Standards on what metabolic phenotypes and responses need to be captured by these models [e.g., the 288 known metabolic functions in human cells used in (Duarte et al., 2007)] may also help ensure that reconstructions maintain essential biological features as they grow larger and more detailed.

To summarize, we make the following recommendations for the validation of FBA-predicted flux maps:

1. When possible, comparisons between FBA-predicted and ^{13}C-MFA-estimated flux maps should be performed to validate the accuracy of FBA-predicted internal fluxes. This provides a greater wealth of information about where and to what extent the model is, and is not, lining up with experimental evidence. When performing such validations, care should be taken to ensure that the conditions under which the FBA-predictions and MFA-estimates are generated are as similar as possible and that any necessary normalizations to account for differences have been made. For an example of thorough FBA-to-MFA comparisons, see Broddrick et al., (2019) and Roell et al., (2023).
    - **Note:** FBA-predicted flux maps require definition not just of the network architecture and constraints, but also an objective function for optimization. Validation of the FBA-predicted flux maps is therefore also a validation of the selected objective function. It is possible for a poorly selected objective function to generate flux predictions that do not align with MFA-estimated fluxes; in such cases, alternative objective functions can be explored.
2. As highlighted in **Table 1.1**, different validation methods evaluate different aspects of the model's predictions. Therefore, employing a number of different validations allows for a fuller and more detailed analysis of model performance and increases the likelihood that other users of the model may be able to appropriately apply it to their research question. For an example of a study employing several different validation techniques, see Heinken et al., (2023).
3. Validations of model predictions are only valuable when the data the predictions are validated against has not already been used in the training or construction of the model. The complexity of the metabolic model reconstruction and analysis process can make it difficult to notice when contamination of the validation dataset by

training data has occurred. In order to identify contamination, one must consider the source of all data used for validation and consider whether it or a value derived from it was used at any stage of the FBA modeling process. For an example of a study that clearly and systematically validates FBA predictions while avoiding such contamination, see Arion et al., (2023).

Improving confidence in the accuracy of FBA flux maps is valuable because generating validated ^{13}C-MFA flux maps for all systems and conditions of interest is impractical. ^{13}C-MFA requires substantial experimental work for each set of conditions and is unsuitable for many multicellular tissues and organisms where the required combination of extended periods of metabolic steady state, controlled provision of informative, non-perturbing labeled substrates, and obtaining enough labeling data cannot be achieved. This FBA-empowered future for systems biology and biotechnology requires well-validated MFA flux maps, so we turn our attention to model validation and selection in MFA.

1.4.2. Validation in ^{13}C-MFA

^{13}C-MFA flux estimates are typically validated based on the goodness-of-fit between measured labeling data and the corresponding values generated by the network model after the optimization of model parameters. The goodness-of-fit is represented by the sum of squared residuals (SSR) where each residual is weighted by dividing it by its experimental variance. The χ^2-test of goodness-of-fit, which is built into commonly used ^{13}C-MFA software (Weitzel et al., 2013; Shupletsov et al., 2014; Young, 2014), is then used to test whether the SSR falls within the 95% confidence interval expected for the defined number of degrees of freedom (DOF). Since its development as a validation method in ^{13}C-MFA (Antoniewicz et al., 2006), the χ^2-test has been widely used and has been useful in the validation of ^{13}C-MFA metabolic models inferred from genome annotations (Au et al., 2014; Cordova and Antoniewicz, 2016; Cordova et al., 2017; Yu King Hing et al., 2021; Dahle et al., 2022; Imada et al., 2023; Mitosch et al., 2023).

However, as described in Sundqvist et al., (2022) and Theorell et al., (2017), the use of the χ^2-test can be problematic in ^{13}C-MFA for several reasons. When upper- and lower-bounds are imposed on estimated flux parameter values, this makes accurate estimation of the effective DOF for the χ^2-test difficult (Theorell et al., 2017). It can also be difficult to accurately determine errors in the MID measurements made for ^{13}C-MFA, resulting in distortion of the variance-weighted SSR values that are being compared against the 95% Confidence Interval

(Sundqvist et al., 2022).

In addition to these technical difficulties with properly applying the χ^2-test, problems arise from how the test is implemented into the model development process during a typical ^{13}C-MFA study. Especially for eukaryotic systems, ^{13}C-MFA flux modeling generally involves making iterative changes to the model based on how well it can explain the data – as assessed informally and by the χ^2-test – followed by refinement and assessment of the data based on this agreement. For example, if the data do not allow the fluxes between the same metabolite in different compartments to be determined, they may be merged in the model or additional measurements may be made to resolve them. Metabolites may also be excluded from the model due to inconsistency between their simulated vs. measured MIDs causing the model to fail the χ^2-test, on the assumption that biological, model-structural, or analytical uncertainties underlie these unexplained divergences (Xu et al., 2022)[1]. The difficulty of accurately quantifying MID measurement errors, mentioned earlier, may be addressed by arbitrarily increasing the assumed measurement error, which reduces the deduced precision of flux estimates to take into account the potential for error sources not accounted for by experimentally observed scatter (Xu et al., 2021b; Sundqvist et al., 2022; Xu et al., 2022)[1]. This process is a natural consequence of the diversity and uncertainty of the metabolic architecture of different systems and is a valid form of exploratory data analysis and model building. However, altering the model by excluding specific data points and adding additional fluxes or metabolites until the χ^2-test passes, and then relying on this very same test as validation is statistically dubious from a rigorous perspective. As in the case of an FBA model validation in which the prediction being validated has been implicitly introduced to the model itself, a final validation of a ^{13}C-MFA model with the same data used to make it acceptable, as quantified by the χ^2-test, does not constitute a real validation. It also can naturally lead to over- or under-fit models, which we discuss below in the section on model selection.

Due to these difficulties, we propose that the χ^2-test, as it is currently used, should be used as one of multiple lines of evidence to consider when validating a ^{13}C-MFA model, especially for less defined and/or more complex eukaryotic systems such as plants. One way to

[1] Here we primarily cite our own work because, as discussed, there are a number of sound reasons for leaving out metabolites and/or increasing MID measurement errors. We have chosen not to highlight other studies that have employed the same practices since we do not know all of the experimental and analytical details underlying them and would not want their inclusion here to be interpreted as implicit criticism.

address the issue of using the χ^2-test for both model development and validation is to reserve a portion of the dataset only for final model validation. This practice of holding out a subset of the data to be used exclusively for validation is standard statistical practice (Gross and MacLeod, 2017) in other areas of systems biology and, conveniently, can also be used for model selection (Sundqvist et al., 2022).

In the absence of direct experimentally measurable fluxes, independent measurements that can be measured or inferred from empirical measurements *in vivo* provide an important ground-truth value to compare with flux estimates and can complement the use of the χ^2-test for validation. An example of this can be found in the plant ^{13}C-MFA literature, where independent measurements of the relative rates of oxygenation and carboxylation by the enzyme rubisco can be compared with ^{13}C-MFA flux estimates (Ma et al., 2014; Xu et al., 2021b; Xu et al., 2022). In Xu et al., (2021b) for example, our group compared predicted values for the relative rates of oxygenation and carboxylation by the enzyme rubisco in photosynthesis versus inferred values from stomatal conductance and other empirical measurements. This led us to conclude that labeling data from whole tissue extracts was insufficient to accurately estimate photorespiratory fluxes without information on the compartmentation of certain metabolites. Despite the strength of this form of validation, it is infrequently practiced.

Another little-used but potentially valuable approach to validation is the corroboration of key features of ^{13}C-MFA models using independent modeling methods. In Xu et al., (2022), simplified compartmental kinetic models yielded analytical solutions predicting that overall labeling time courses should take the form of sums of exponential rate components. Fitting labeling data to these exponential models and applying statistical model selection techniques provided independent corroboration of the overall architecture of the ^{13}C-MFA model that was used to obtain a detailed flux map.

Returning to goodness-of-fit, one must also keep in mind what information is taken into consideration and the effect of the assumed network architecture. In INST-MFA, where time-course labeling data is used, metabolite pool sizes are both estimable parameters and constrainable modeling inputs. When pool sizes are not provided as empirical measurements, pool size estimates are typically imprecise and inaccurate (Zheng et al., 2022). The inaccuracy of these estimates is not usually interpreted as an impediment to publishing ^{13}C-MFA results and according to Zheng et al., (2022), leaving out pool size information does not adversely affect flux

estimate accuracy. Flux estimates are not, however, always robust against misspecifications of the network model (Sundqvist et al., 2022). The exclusion of pool size information provides greater flexibility in fitting experimental data, allowing robustness against model misspecifications at the expense of not detecting them (Zheng et al., 2022). A useful next step for this field would be to routinely measure and include pool size estimates to improve the detection of incorrect model architectures. Measurement of all metabolites in a way that allows discrimination of pools for identical metabolites in different cellular compartments requires a method like Non-Aqueous Fractionation [e.g., [Krueger et al., 2011)], which may be prohibitively difficult to implement in many studies. In such cases, use of a strategically selected set of metabolite levels may be used to allow for improved detection of incorrect model architectures. This introduces the matter of model selection.

1.5. Model selection in ^{13}C-MFA

As discussed earlier, model development in ^{13}C-MFA is an iterative process. Alternate models developed during this process may differ in their numbers of reactions and metabolites, resulting in different DOF. Adding model parameters can result in overfitting when these extra DOF lead the ^{13}C-MFA optimization to fit noise rather than biological signal. Model selection techniques can be used to avoid this overfitting and to select the most statistically supported model among alternatives. The development of FBA models can also involve deciding between alternative architectures. However, comparison and selection of such models from sets of alternatives based on their predictions' deviations from empirical measurements is uncommon, so we focus our attention on ^{13}C-MFA.

Model misspecification can result in missing important fluxes, incorrectly estimating the rates of modeled fluxes, or incorrectly estimating the precision of flux estimates. In a study our group performed of central metabolic fluxes in the oilseed crop *Camelina sativa* (Xu et al., 2022), previously published model architectures that passed the χ^2-test of goodness-of-fit (Xu et al., 2021b) were nonetheless shown to be missing an important set of metabolic reactions involving the movement of carbohydrates to and from the vacuole. In Sundqvist et al., (2022), *in silico* examples of sub-optimal model selection resulting in flux estimates that fall outside of the 95% confidence intervals for those same fluxes generated using the correct model architecture are provided, showing the potential for biased flux estimates when model selection is not properly performed. Finally, the literature on "Genome-scale-^{13}C-MFA" has provided evidence

that the exclusion of many reactions peripheral to the metabolic network under consideration (typically core metabolism) in ^{13}C-MFA can result in artificially narrow confidence intervals. Genome-scale-^{13}C-MFA involves estimating a flux map by minimizing deviation between predicted and measured isotopic labeling but using the kind of genome-scale metabolic network more typically used for FBA analyses (Gopalakrishnan and Maranas, 2015; Hendry et al., 2020). In studies on the cyanobacterium *Synechococcus elongatus* (Gopalakrishnan et al., 2018; Hendry et al., 2019), it has been shown that the substantially larger genome-scale ^{13}C-MFA models achieved better fits to the labeling data, that these reductions in SSR were statistically justified, and that the original models of core metabolism underestimated the uncertainty in a number of flux estimates by ignoring alternative metabolic pathways that could also explain patterns in the labeling data (Hendry et al., 2020). The examples above demonstrate that rather than being a statistical curiosity, model selection (or the lack thereof) can have serious implications for the accuracy and reliability of flux modeling results.

Several approaches to model selection can be found in the ^{13}C-MFA literature, with different approaches being taken in different studies. The simplest is selecting the model with the smallest SSR. This method does not work when the DOF of the compared models are different, as increasing the DOF in a model inevitably allows it to fit a given data set better. This may be accounted for informally by noting the change in DOF [e.g., (Xu et al., 2022)], or in a more statistically rigorous way using the extra-sum-of-squares test (Draper and Smith, 1998; Boyle et al., 2017) or information criteria (Schwarz, 1978; Akaike, 1998). The most common model selection approach used in ^{13}C-MFA is an informal method using the χ^2-test, wherein models are iteratively modified until a model and dataset pass the test, or where several alternative models are evaluated and the one that passes the test by the widest margin is selected (Dalman et al., 2016; Antoniewicz, 2018; Long and Antoniewicz, 2019a; Sundqvist et al., 2022). These approaches have been used, for example, to demonstrate that the isotopic labeling data of co-culture systems cannot be adequately described by modeling with a single-culture ^{13}C-MFA model (Gebreselassie and Antoniewicz, 2015; Wolfsberg et al., 2018), to provide evidence for the operation of previously undescribed fluxes in mammalian cells (Ahn et al., 2016), and to detect missing reactions in metabolic network reconstructions from genome annotations or that are needed to describe the metabolism of mutant *E. coli* strains (Au et al., 2014; Long and Antoniewicz, 2019b).

However, the previously mentioned limitations of the χ^2-test for model validation also affect its usefulness for model selection and models failing the test due to these limitations can lead to the addition of statistically unjustified metabolites or reactions to the model until it passes (Sundqvist et al., 2022). We refer to the χ^2-test-based methods as "informal" model selection because when multiple models are evaluated, they are not directly or formally compared to determine whether the additional parameters in more complex models are statistically justified, which can naturally lead to the selection of overfit models.

The general approach of avoiding overfitting by evaluating models based on their performance on a set of data not used during the fitting process is widely used in statistics [e.g., cross-validation techniques (Hastie et al., 2017)]. The validation-based approach taken in Sundqvist et al., (2022) implements this best practice, separating fitting and testing data sets to avoid the pitfalls discussed above. In our view, this represents a substantial advancement in model selection in ^{13}C-MFA. This method divides the labeling dataset into training and validation subsets and then estimates fluxes in alternative models using the training data. These alternative models' flux maps, and their accompanying predicted MIDs, are then compared based on their agreement with the validation MID data. The model whose flux map results in the smallest SSR when compared with this validation data is selected. The authors generated synthetic labeling data from a predefined "correct" model and assessed the ability of their new method and other model selection techniques to identify this correct model from a set of alternatives. The validation-based approach accomplishes this more consistently than existing model selection methods, including χ^2-test-based methods, and does so irrespective of the value of the measurement error in the labeling datasets. The incorrect models selected by other methods contain flux estimates that fall outside the 95% confidence intervals of the fluxes from the correct model, highlighting the importance of model selection for obtaining accurate flux estimates (Sundqvist et al., 2022). The generation of MID data in additional labeling experiments to precisely measure all fluxes in a network (Chang et al., 2008; Crown et al., 2012; Crown and Antoniewicz, 2012; Leighty and Antoniewicz, 2013; Millard et al., 2014; Crown et al., 2015; Crown et al., 2016; Beyß et al., 2021) provides the reserved validation datasets needed for Sundqvist et al., (2022). This means that for ^{13}C-MFA studies that already require a parallel labeling approach, implementation of this more rigorous model selection approach is simply a matter of setting aside a subset of data to evaluate alternative model architectures.

This approach can be extended in INST-MFA by using metabolite pool size measurements in the selection process. Individual pool sizes are sensitive to the local kinetic parameters and will fit poorly when reaction networks are incompletely specified (Zheng et al., 2022). We therefore suggest that validation-based model selection using pool size measurements as input measurements is a promising prospective model selection approach for INST MFA (**Figure 1.2**). Indeed, although not referred to explicitly as model selection, in Zheng et al., (2022) the authors show that inclusion of pool size information results in an incorrectly specified network architecture failing to pass the χ^2-test of goodness-of-fit, whereas a correctly specified network does pass. This corresponds to the "first to pass χ^2" method of model selection discussed by Sundqvist et al., (2022) and is subject to the various limitations of the χ^2-test as a model selection technique covered earlier. By incorporating these metabolite pool sizes into the formalized model selection framework described by Sundqvist et al., (2022), we may arrive at a more robust form of model selection that is better at detecting misspecified networks. As Sundqvist et al., (2022) note, the optimal model selected by their method should be subjected to a final validation to assess model quality. A model architecture may be selected by the model selection process but result in a substantial deviation of some metric from independently measured values. For this final validation, a combination of the χ^2-test, independent experimental measurements, and alternative modeling approaches can be used. Keeping in mind both the trade-off between goodness-of-fit and model complexity and the multiple ways in which ^{13}C-MFA model predictions can be validated will ensure that flux estimates are as accurate and robust as possible.

Model validation and selection are an integral part of the ^{13}C-MFA process. Notably, model selection practices like the use of validation-based model selection (Sundqvist et al., 2022) and the use of the extra-sum-of-squares test (Boyle et al., 2017) to compare alternative model architectures represent, in our view, a major improvement over exclusive use of the χ^2-test of goodness-of-fit test for both purposes, but are seldom practiced in the literature. We encourage the use of these techniques and believe they hold promise for improving confidence in both the fluxes and network architectures reported in studies.

With respect to validation and model selection in MFA, we recommend the following:

1. As highlighted by Antoniewicz, (2018), transparency is key in ^{13}C-MFA, given the assumptions that must be satisfied for ^{13}C-MFA modeling as well as the sensitivity of

flux estimates to model architecture. As an example of a transparently reported ^{13}C-MFA study, see Nicolae et al., (2014).

2. The validation and selection of MFA-estimated fluxes, like the validation of any model output, benefits from multiple lines of corroborating evidence. When possible, the use of alternative modeling approaches of isotopic labeling data can be a powerful tool for arriving at well-supported model architectures, as in Xu et al., (2022).
3. In INST-MFA, metabolite pool size measurements can be used to provide additional confidence in model validity and tighten flux confidence intervals (Nöh et al., 2007), as well as provide additional measurements for validation-based model selection. However, practitioners should be aware that these measurements can make model fits highly sensitive to incorrectly specified network models in ways that may or may not affect the accuracy of flux estimates (Zheng et al., 2022). Additionally determination of subcellular compartmentation of certain metabolites may be prohibitively difficult in some cases. In such cases, key metabolites with known subcellular compartmentation may be measured.
4. We recommend the use of a proper model selection framework to compare alternative, biochemically reasonable model architectures when performing ^{13}C-MFA modeling. The framework outlined in Sundqvist et al., (2022) represents the state-of-the-art in this area. Barring the application of that method, a more traditional model selection approach, such as the extra-sum-of-squares approach used in Boyle et al., (2017) can be employed.

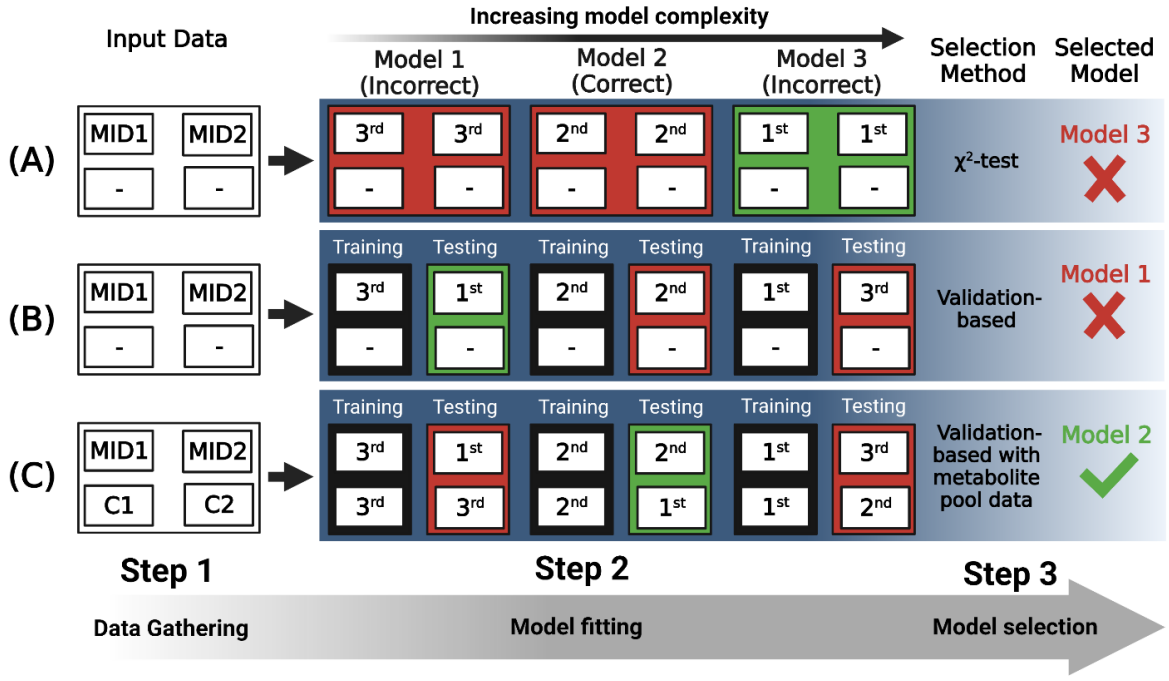


Figure 1.2: Approaches to model selection for ^{13}C-MFA. Metabolic network models 1-3 having increasing complexity are compared. Model 2 in this example is the correct description of the network. (A) Labeling data (MID1 & MID2) are gathered and, for each model, agreement between model output and these data is optimized. The χ^2-test of goodness-of-fit is used to assess each model fit and these model fits are ranked 1st, 2nd, or 3rd, with the 1st passing the test by the widest margin and being selected as the most statistically well-supported model. (B) Labeling data are split into "training" and "testing" subsets and agreement between model output and the "training" data is optimized. The Sum-of-Squared Residuals (SSR) is then calculated for each model from the deviation between its output and the "testing" data. The model fits are then ranked 1st, 2nd, and 3rd, with the 1st having the lowest SSR and being selected. (C) Labeling data and metabolite pool data (C1 and C2) are gathered and split into "training" and "testing" subsets. For each model, agreement between model output and these data is optimized. The Sum-of-Squared Residuals (SSR) is then calculated for each model from the deviation between its output and the "testing" data. The model fits are then ranked 1st, 2nd, and 3rd, with the 1st having the lowest SSR and being selected. The inclusion of metabolite pool size data into both the "fitting" and "testing" datasets provides more data to go off when evaluating goodness-of-fit, potentially increasing the likelihood of identifying the correct model from a set of alternatives.

1.6. Future directions

We believe that validation and selection deserve greater attention from the flux analysis community and suggest that implementing the approaches highlighted in this perspective will improve the accuracy and reliability of constraints-based metabolic modeling and flux estimates. However, we also recognize that some approaches suggested here, such as the use of pool size measurements, can be extremely difficult to implement in practice. A recent publication on

isotopically non-stationary MFA of *Arabidopsis thaliana* heterotrophic cell culture metabolism highlighted that although pool size data could potentially be used to improve the accuracy and precision of flux predictions, the experimental difficulty of measuring the concentrations of metabolites distributed across multiple subcellular compartments made this prohibitively difficult (Smith et al., 2022). As in all areas of science, then, the development of consensus best practices in the evaluation of and inference from data and models must arise at the intersection of rigorous statistical theory and experimental practicalities. However, we believe that researchers engaged in constraint-based metabolic modeling as well as readers of modeling studies benefit when the limitations of present validation and selection practices are clarified.

Several matters call for investigation before definitive recommendations can be made on best practices. At present, it is not clear how to appropriately weight the contributions to flux estimation of unambiguous direct flux measurements like substrate uptake, which typically have relatively large standard deviations, against MIDs, which frequently have much smaller standard deviations but whose relationship to fluxes depends on model structure and whose measured values may be offset by unknown analytical effects. Likewise, it is unclear how best to deal with those not infrequent MID measurements that have extremely small, but imprecisely measured, standard deviations, which can exert too much control over the fitting process.

Finally, we would like to conclude by emphasizing that the process of careful validation and model selection can lead to the generation of models that are not only more quantitatively sound, but that yield exciting scientific insights [e.g., (Ahn et al., 2016; Wolfsberg et al., 2018)].

1.7. Acknowledgments

This research was supported by the Office of Science (BER), U.S. Department of Energy, Grant no DE-SC0018269 (J.A.M.K., Y.S-H.). This work is supported, in part, by the NSF Research Traineeship Program (Grant DGE-1828149) to J.A.M.K. This publication was also made possible by a predoctoral training award to J.A.M.K. from Grant T32-GM110523 from National Institute of General Medical Sciences (NIGMS) of the NIH. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIGMS or NIH. Figures made using BioRender.com.

1.8. Author contributions

Conceptualization: J.A.M.K and Y.S-H. Investigation: J.A.M.K. Visualization: J.A.M.K. Writing – original draft: J.A.M.K. Writing – review and editing: J.A.M.K. and Y.S-H.

REFERENCES

Ahn WS, Crown SB, Antoniewicz MR (2016) Evidence for transketolase-like TKTL1 flux in CHO cells based on parallel labeling experiments and (13)C-metabolic flux analysis. Metabolic engineering **37**: 72–78

Akaike H (1998) Information Theory and an Extension of the Maximum Likelihood Principle. *In* E Parzen, K Tanabe, G Kitagawa, eds, Selected Papers of Hirotugu Akaike. Springer New York, New York, NY, pp 199–213

Åkesson M, Förster J, Nielsen J (2004) Integration of gene expression data into genome-scale metabolic models. Metabolic Engineering **6**: 285–293

Alzoubi D, Desouki AA, Lercher MJ (2019) Flux balance analysis with or without molecular crowding fails to predict two thirds of experimentally observed epistasis in yeast. Scientific Reports **9**: 1–9

Antoniewicz MR (2015) Methods and advances in metabolic flux analysis: a mini-review. Journal of Industrial Microbiology and Biotechnology **42**: 317–325

Antoniewicz MR (2018) A guide to ^{13}C metabolic flux analysis for the cancer biologist. Experimental and Molecular Medicine. doi: 10.1038/s12276-018-0060-y

Antoniewicz MR, Kelleher JK, Stephanopoulos G (2006) Determination of confidence intervals of metabolic fluxes estimated from stable isotope measurements. Metabolic Engineering **8**: 324–337

Arion I-S, Hiroyuki O, Matti G, Ghita G, Kapil A, X. CO, Terence H, Sebastian B (2023) A Genome-Scale Metabolic Model of Marine Heterotroph Vibrio splendidus Strain 1A01. mSystems **0**: e00377-22

Au J, Choi J, Jones SW, Venkataramanan KP, Antoniewicz MR (2014) Parallel labeling experiments validate Clostridium acetobutylicum metabolic network model for (13)C metabolic flux analysis. Metabolic engineering **26**: 23–33

Becker J, Zelder O, Häfner S, Schröder H, Wittmann C (2011) From zero to hero-Design-based systems metabolic engineering of Corynebacterium glutamicum for l-lysine production. Metabolic Engineering **13**: 159–168

Becker SA, Palsson BO (2008) Context-specific metabolic networks are consistent with experiments. PLoS Computational Biology. doi: 10.1371/journal.pcbi.1000082

Bernstein DB, Sulheim S, Almaas E, Segrè D (2021) Addressing uncertainty in genome-scale metabolic model reconstruction and analysis. Genome Biology **22**: 64

Beyß M, Parra-Peña VD, Ramirez-Malule H, Nöh K (2021) Robustifying Experimental Tracer Design for ^{13}C-Metabolic Flux Analysis. Frontiers in Bioengineering and Biotechnology. doi: 10.3389/fbioe.2021.685323

Blázquez B, San León D, Rojas A, Tortajada M, Nogales J (2023) New Insights on Metabolic Features of Bacillus subtilis Based on Multistrain Genome-Scale Metabolic Modeling. International Journal of Molecular Sciences. doi: 10.3390/ijms24087091

Bordel S, Agren R, Nielsen J (2010) Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes. PLoS Computational Biology **6**: 16

Boyle NR, Sengupta N, Morgan JA (2017) Metabolic flux analysis of heterotrophic growth in Chlamydomonas reinhardtii. PLoS ONE **12**: 1–23

Broddrick JT, Welkie DG, Jallet D, Golden SS, Peers G, Palsson BO (2019) Predicting the metabolic capabilities of Synechococcus elongatus PCC 7942 adapted to different light regimes. Metabolic Engineering **52**: 42–56

Burgard AP, Pharkya P, Maranas CD (2003) OptKnock: A Bilevel Programming Framework for Identifying Gene Knockout Strategies for Microbial Strain Optimization. Biotechnology and Bioengineering **84**: 647–657

Chang Y, Suthers PF, Maranas CD (2008) Identification of optimal measurement sets for complete flux elucidation in metabolic flux analysis experiments. Biotechnology and Bioengineering **100**: 1039–1049

Cheah YE, Young JD (2018) Isotopically nonstationary metabolic flux analysis (INST-MFA): putting theory into practice. Current Opinion in Biotechnology **54**: 80–87

Chen X, Alonso AP, Allen DK, Reed JL, Shachar-Hill Y (2011) Synergy between ^{13}C-metabolic flux analysis and flux balance analysis for understanding metabolic adaption to anaerobiosis in E. coli. Metabolic Engineering **13**: 38–48

Choi J, Antoniewicz MR (2019) Tandem mass spectrometry for ^{13}C metabolic flux analysis: Methods and algorithms based on EMU framework. Frontiers in Microbiology **10**: 31

Choi Y-M, Choi D-H, Lee YQ, Koduru L, Lewis NE, Lakshmanan M, Lee D-Y (2023) Mitigating biomass composition uncertainties in flux balance analysis using ensemble representations. Computational and Structural Biotechnology Journal **21**: 3736–3745

Coppens L, Tschirhart T, Leary DH, Colston SM, Compton JR, Hervey IV WJ, Dana KL, Vora GJ, Bordel S, Ledesma-Amaro R (2023) Vibrio natriegens genome-scale modeling reveals insights into halophilic adaptations and resource allocation. Molecular Systems Biology **19**: e10523

Cordova LT, Antoniewicz MR (2016) (13)C metabolic flux analysis of the extremely thermophilic, fast growing, xylose-utilizing Geobacillus strain LC300. Metabolic engineering **33**: 148–157

Cordova LT, Cipolla RM, Swarup A, Long CP, Antoniewicz MR (2017) (13)C metabolic flux analysis of three divergent extremely thermophilic bacteria: Geobacillus sp. LC300,

Thermus thermophilus HB8, and Rhodothermus marinus DSM 4252. Metabolic engineering **44**: 182–190

Crown SB, Ahn WS, Antoniewicz MR (2012) Rational design of ^{13}C-labeling experiments for metabolic flux analysis in mammalian cells. BMC Systems Biology **6**: 43

Crown SB, Antoniewicz MR (2012) Selection of tracers for ^{13}C-Metabolic Flux Analysis using Elementary Metabolite Units (EMU) basis vector methodology. Metabolic Engineering **14**: 150–161

Crown SB, Long CP, Antoniewicz MR (2016) Optimal tracers for parallel labeling experiments and ^{13}C metabolic flux analysis: A new precision and synergy scoring system. Metabolic engineering **38**: 10–18

Crown SB, Long CP, Antoniewicz MR (2015) Integrated ^{13}C-metabolic flux analysis of 14 parallel labeling experiments in Escherichia coli. Metabolic engineering **28**: 151–158

Dahle ML, Papoutsakis ET, Antoniewicz MR (2022) ^{13}C-metabolic flux analysis of Clostridium ljungdahlii illuminates its core metabolism under mixotrophic culture conditions. Metabolic Engineering **72**: 161–170

Dalman T, Wiechert W, Nöh K (2016) A scientific workflow framework for ^{13}C metabolic flux analysis. Journal of Biotechnology **232**: 12–24

Dinh HV, Sarkar D, Maranas CD (2022) Quantifying the propagation of parametric uncertainty on flux balance analysis. Metabolic Engineering **69**: 26–39

Draper NR, Smith H (1998) Extra Sums of Squares and Tests for Several Parameters Being Zero. Applied Regression Analysis. John Wiley & Sons, Ltd, pp 149–177

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. Proceedings of the National Academy of Sciences of the United States of America **104**: 1777–1782

Ebrahim A, Lerman JA, Palsson BO, Hyduke DR (2013) COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Systems Biology. doi: 10.1186/1752-0509-7-74

Feierabend M, Renz A, Zelle E, Nöh K, Wiechert W, Dräger A (2021) High-Quality Genome-Scale Reconstruction of Corynebacterium glutamicum ATCC 13032. Frontiers in microbiology **12**: 750206

Gatto F, Miess H, Schulze A, Nielsen J (2015) Flux balance analysis predicts essential genes in clear cell renal cell carcinoma metabolism. Scientific Reports **5**: 10738

Gebreselassie NA, Antoniewicz MR (2015) (13)C-metabolic flux analysis of co-cultures: A novel approach. Metabolic engineering **31**: 132–139

Gianchandani EP, Oberhardt MA, Burgard AP, Maranas CD, Papin JA (2008) Predicting biological system objectives de novo from internal state measurements. BMC Bioinformatics **9**: 43

Gleizer S, Ben-Nissan R, Bar-On YM, Antonovsky N, Noor E, Zohar Y, Jona G, Krieger E, Shamshoum M, Bar-Even A, et al (2019) Conversion of Escherichia coli to Generate All Biomass Carbon from CO2. Cell **179**: 1255-1263.e12

Gopalakrishnan S, Maranas CD (2015) ^{13}C metabolic flux analysis at a genome-scale. Metabolic Engineering **32**: 12–22

Gopalakrishnan S, Pakrasi HB, Maranas CD (2018) Elucidation of photoautotrophic carbon flux topology in Synechocystis PCC 6803 using genome-scale carbon mapping models. Metabolic engineering **47**: 190–199

Gross F, MacLeod M (2017) Prospects and problems for standardizing model validation in systems biology. Progress in Biophysics and Molecular Biology **129**: 3–12

Haraldsdóttir HS, Cousins B, Thiele I, Fleming RMT, Vempala S (2017) CHRR: Coordinate hit-and-run with rounding for uniform sampling of constraint-based models. Bioinformatics **33**: 1741–1743

Hastie T, Tibshirani R, Friedman J (2017) Model Assessment and Selection. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, 2nd ed. Springer, New York, NY, pp 219–260

Heinken A, Hertel J, Acharya G, Ravcheev DA, Nyga M, Okpala OE, Hogan M, Magnúsdóttir S, Martinelli F, Nap B, et al (2023) Genome-scale metabolic reconstruction of 7,302 human microorganisms for personalized medicine. Nature Biotechnology. doi: 10.1038/s41587-022-01628-0

Heinken A, Magnúsdóttir S, Fleming RMT, Thiele I (2021) DEMETER: efficient simultaneous curation of genome-scale reconstructions guided by experimental data and refined gene annotations. Bioinformatics **37**: 3974–3975

Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, Haraldsdóttir HS, Wachowiak J, Keating SM, Vlasov V, et al (2019) Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nature Protocols **14**: 639–702

Hendry JI, Dinh HV, Foster C, Gopalakrishnan S, Wang L, Maranas CD (2020) Metabolic flux analysis reaching genome wide coverage: lessons learned and future perspectives. Current Opinion in Chemical Engineering **30**: 17–25

Hendry JI, Gopalakrishnan S, Ungerer J, Pakrasi HB, Tang YJ, Maranas CD (2019) Genome-Scale Fluxome of Synechococcus elongatus UTEX 2973 Using Transient ^{13}C-Labeling Data. Plant Physiology **179**: 761–769

Holzhütter HG (2004) The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. European Journal of Biochemistry **271**: 2905–2922

Imada T, Yamamoto C, Toyoshima M, Toya Y, Shimizu H (2023) Effect of light fluctuations on photosynthesis and metabolic flux in Synechocystis sp. PCC 6803. Biotechnology Progress **39**: e3326

Jazmin LJ, Beckers V, Young JD (2014) User Manual for INCA.

Kaste JAM, Shachar-Hill Y (2023) Accurate flux predictions using tissue-specific gene expression in plant metabolic modeling. Bioinformatics: btad186

Kirk P, Thorne T, Stumpf MPH (2013) Model selection in systems and synthetic biology. Current Opinion in Biotechnology **24**: 767–774

Koffas MAG, Jung GY, Stephanopoulos G (2003) Engineering metabolism and product formation in Corynebacterium glutamicum by coordinated gene overexpression. Metabolic Engineering **5**: 32–41

Koffas MAG, Stephanopoulos G (2005) Strain improvement by metabolic engineering: Lysine production as a case study for systems biology. Current Opinion in Biotechnology **16**: 361–366

Krueger S, Giavalisco P, Krall L, Steinhauser M-C, Büssis D, Usadel B, Flügge U-I, Fernie AR, Willmitzer L, Steinhauser D (2011) A topological map of the compartmentalized Arabidopsis thaliana leaf metabolome. PLoS One **6**: e17806

Lee JM, Gianchandani EP, Papin JA (2006) Flux balance analysis in the era of metabolomics. Briefings in Bioinformatics **7**: 140–150

Leighty RW, Antoniewicz MR (2013) COMPLETE-MFA: Complementary parallel labeling experiments technique for metabolic flux analysis. Metabolic Engineering **20**: 49–55

Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, Adkins JN, Schramm G, Purvine SO, Lopez-Ferrer D, et al (2010) Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Molecular Systems Biology. doi: 10.1038/msb.2010.47

Lieven C, Beber ME, Olivier BG, Bergmann FT, Ataman M, Babaei P, Bartell JA, Blank LM, Chauhan S, Correia K, et al (2020) MEMOTE for standardized genome-scale metabolic model testing. Nature Biotechnology **38**: 272–276

Long CP, Antoniewicz MR (2019a) High-resolution (13)C metabolic flux analysis. Nature protocols **14**: 2856–2877

Long CP, Antoniewicz MR (2019b) Metabolic flux responses to deletion of 20 core enzymes reveal flexibility and limits of E. coli metabolism. Metabolic engineering **55**: 249–257

Ma F, Jazmin LJ, Young JD, Allen DK (2014) Isotopically nonstationary ^{13}C flux analysis of changes in Arabidopsis thaliana leaf metabolism due to high light acclimation. Proceedings of the National Academy of Sciences of the United States of America **111**: 16967–16972

Machado D, Herrgård M (2014) Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. PLoS Computational Biology. doi: 10.1371/journal.pcbi.1003580

Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metabolic Engineering **5**: 264–276

Megchelenbrink W, Huynen M, Marchiori E (2014) optGpSampler: An improved tool for uniformly sampling the solution-space of genome-scale metabolic networks. PLoS ONE. doi: 10.1371/journal.pone.0086587

Millard P, Sokol S, Letisse F, Portais J-C (2014) IsoDesign: A software for optimizing the design of ^{13}C-metabolic flux analysis experiments. Biotechnology and Bioengineering **111**: 202–208

Mitosch K, Beyß M, Phapale P, Drotleff B, Nöh K, Alexandrov T, Patil KR, Typas A (2023) A pathogen-specific isotope tracing approach reveals metabolic activities and fluxes of intracellular Salmonella. PLOS Biology **21**: e3002198

Nicolae A, Wahrheit J, Bahnemann J, Zeng A-P, Heinzle E (2014) Non-stationary ^{13}C metabolic flux analysis of Chinese hamster ovary cells in batch culture using extracellular labeling highlights metabolic reversibility and compartmentation. BMC Systems Biology **8**: 50

Nielsen J (2003) It Is All about Metabolic Fluxes. Journal of Bacteriology **185**: 7031–7035

Noecker C, Sanchez J, Bisanz JE, Escalante V, Alexander M, Trepka K, Heinken A, Liu Y, Dodd D, Thiele I, et al (2023) Systems biology elucidates the distinctive metabolic niche filled by the human gut microbe Eggerthella lenta. PLOS Biology **21**: e3002125

Nöh K, Grönke K, Luo B, Takors R, Oldiges M, Wiechert W (2007) Metabolic flux analysis at ultra short time scale: Isotopically non-stationary ^{13}C labeling experiments. Journal of Biotechnology **129**: 249–267

Norsigian CJ, Pusarla N, McConn JL, Yurkovich JT, Dräger A, Palsson BO, King Z (2020) BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree. Nucleic Acids Research **48**: D402–D406

Oftadeh O, Salvy P, Masid M, Curvat M, Miskovic L, Hatzimanikatis V (2021) A genome-scale metabolic model of Saccharomyces cerevisiae that integrates expression constraints and reaction thermodynamics. Nature Communications **12**: 4790

Ong W, Vu TT, Lovendahl KN, Llull JM, Serres MH, Romine MF, Reed JL (2014)

Comparisons of Shewanella strains based on genome annotations, modeling, and experiments. BMC Systems Biology **8**: 31

Orth JD, Fleming RMT, Palsson BØ (2010a) Reconstruction and Use of Microbial Metabolic Networks: the Core Escherichia coli Metabolic Model as an Educational Guide. EcoSal Plus. doi: 10.1128/ecosalplus.10.2.1

Orth JD, Thiele I, Palsson BO (2010b) What is flux balance analysis? Nature Biotechnology **28**: 245–248

Pandey V, Hadadi N, Hatzimanikatis V (2019) Enhanced flux prediction by integrating relative expression and relative metabolite abundance into thermodynamically consistent metabolic models. PLOS Computational Biology **15**: 1–23

Pinchuk GE, Hill EA, Geydebrekht OV, de Ingeniis J, Zhang X, Osterman A, Scott JH, Reed SB, Romine MF, Konopka AE, et al (2010) Constraint-based model of Shewanella oneidensis MR-1 metabolism: A tool for data analysis and hypothesis generation. PLoS Computational Biology **6**: 1–8

Ravi S, Gunawan R (2021) ΔFBA—Predicting metabolic flux alterations using genome-scale metabolic models and differential transcriptomic data. PLOS Computational Biology **17**: e1009589

Roell GW, Schenk C, Anthony WE, Carr RR, Ponukumati A, Kim J, Akhmatskaya E, Foston M, Dantas G, Moon TS, et al (2023) A High-Quality Genome-Scale Model for Rhodococcus opacus Metabolism. ACS Synth Biol **12**: 1632–1644

Santos-Merino M, Gargantilla-Becerra Á, de la Cruz F, Nogales J (2023) Highlighting the potential of Synechococcus elongatus PCC 7942 as platform to produce α-linolenic acid through an updated genome-scale metabolic modeling. Frontiers in Microbiology. doi: 10.3389/fmicb.2023.1126030

Schellenberger J, Palsson B (2009) Use of randomized sampling for analysis of metabolic networks. Journal of Biological Chemistry **284**: 5457–5461

Schnitzer B, Österberg L, Cvijovic M (2022) The choice of the objective function in flux balance analysis is crucial for predicting replicative lifespans in yeast. PLOS ONE **17**: 1–15

Schroeder WL, Saha R (2020) Introducing an Optimization- and explicit Runge-Kutta- based Approach to Perform Dynamic Flux Balance Analysis. Scientific Reports **10**: 1–28

Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. Molecular Systems Biology **3:** 199. doi: 10.1038/msb4100162

Schwarz G (1978) Estimating the Dimension of a Model. The Annals of Statistics **6**: 461–464

Segrè D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. Proceedings of the National Academy of Sciences of the United States of America **99**: 15112–15117

Shinfuku Y, Sorpitiporn N, Sono M, Furusawa C, Hirasawa T, Shimizu H (2009) Development and experimental verification of a genome-scale metabolic model for Corynebacterium glutamicum. Microbial Cell Factories **8**: 43

Shlomi T, Berkman O, Ruppin E (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. Proceedings of the National Academy of Sciences **102**: 7695–7700

Shupletsov MS, Golubeva LI, Rubina SS, Podvyaznikov DA, Iwatani S, Mashko SV (2014) OpenFLUX2: ^{13}C-MFA modeling software package adjusted for the comprehensive analysis of single and parallel labeling experiments. Microbial Cell Factories **13**: 1–25

Sigurdsson MI, Jamshidi N, Steingrimsson E, Thiele I, Palsson BØ (2010) A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1. BMC systems biology **4**: 140

Smith EN, Ratcliffe RG, Kruger NJ (2022) Isotopically non-stationary metabolic flux analysis of heterotrophic Arabidopsis thaliana cell cultures. Frontiers in plant science **13**: 1049559

Spivey A (2004) Systems biology: the big picture. Environmental health perspectives **112**: 938–943

Stanford NJ, Millard P, Swainston N (2015) RobOKoD: Microbial strain design for (over)production of target compounds. Frontiers in Cell and Developmental Biology **3**: 1–12

Sundqvist N, Grankvist N, Watrous J, Mohit J, Nilsson R, Cedersund G (2022) Validation-based model selection for ^{13}C metabolic flux analysis with uncertain measurement errors. PLOS Computational Biology **18**: e1009999

Tec-Campos D, Posadas C, Tibocha-Bonilla JD, Thiruppathy D, Glonek N, Zuñiga C, Zepeda A, Zengler K (2023) The genome-scale metabolic model for the purple non-sulfur bacterium Rhodopseudomonas palustris Bis A53 accurately predicts phenotypes under chemoheterotrophic, chemoautotrophic, photoheterotrophic, and photoautotrophic growth conditions. PLOS Computational Biology **19**: e1011371

Tepper N, Shlomi T (2009) Predicting metabolic engineering knockout strategies for chemical production: Accounting for competing pathways. Bioinformatics **26**: 536–543

Theorell A, Leweke S, Wiechert W, Nöh K (2017) To be certain about the uncertainty: Bayesian statistics for ^{13}C metabolic flux analysis. Biotechnology and Bioengineering **114**: 2668–2684

Thiele I, Palsson B (2010) A protocol for generating a high-quality genome-scale metabolic

reconstruction. Nature Protocols **5**: 93–121

Tian M, Reed JL (2018) Integrating proteomic or transcriptomic data into metabolic models using linear bound flux balance analysis. Bioinformatics **34**: 3882–3888

Varma A, Palsson BO (1994) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. Applied and Environmental Microbiology **60**: 3724–3731

Wang Y, Hui S, Wondisford FE, Su X (2021) Utilizing tandem mass spectrometry for metabolic flux analysis. Laboratory Investigation **101**: 423–429

Weitzel M, Nöh K, Dalman T, Niedenführ S, Stute B, Wiechert W (2013) ^{13}CFLUX2 - High-performance software suite for ^{13}C-metabolic flux analysis. Bioinformatics **29**: 143–145

Wolfsberg E, Long CP, Antoniewicz MR (2018) Metabolism in dense microbial colonies: (13)C metabolic flux analysis of E. coli grown on agar identifies two distinct cell populations with acetate cross-feeding. Metabolic engineering **49**: 242–247

Xu Y, Fu X, Sharkey TD, Shachar-Hill Y, Walker BJ (2021) The metabolic origins of non-photorespiratory CO2 release during photosynthesis: A metabolic flux analysis. Plant Physiology **186**: 297–314

Xu Y, Wieloch T, Kaste JAM, Shachar-Hill Y, Sharkey TD (2022) Reimport of carbon from cytosolic and vacuolar sugar pools into the Calvin-Benson cycle explains photosynthesis labeling anomalies. Proceedings of the National Academy of Sciences **119**: e2121531119

Young JD (2014) INCA: A computational platform for isotopically non-stationary metabolic flux analysis. Bioinformatics **30**: 1333–1335

Yu King Hing N, Aryal UK, Morgan JA (2021) Probing Light-Dependent Regulation of the Calvin Cycle Using a Multi-Omics Approach. Frontiers in Plant Science. doi: 10.3389/fpls.2021.733122

Yuan H, Cheung CYM, Hilbers PAJ, van Riel NAW (2016) Flux Balance Analysis of Plant Metabolism: The Effect of Biomass Composition and Model Structure on Model Predictions. Frontiers in Plant Science. doi: 10.3389/fpls.2016.00537

Zheng AO, Sher A, Fridman D, Musante CJ, Young JD (2022) Pool size measurements improve precision of flux estimates but increase sensitivity to unmodeled reactions outside the core network in isotopically nonstationary metabolic flux analysis (INST-MFA). Biotechnology Journal **17**: 1–17

Chapter 2

Reimport of carbon from cytosolic and vacuolar sugar pools into the Calvin–Benson cycle explains photosynthesis labeling anomalies

This research was published in:

Y. Xu*[2], T. Wieloch**, **J. A. M. Kaste***, Y. Shachar-Hill, T. D. Sharkey, Reimport of carbon from cytosolic and vacuolar sugar pools into the Calvin-Benson cycle explains photosynthesis labeling anomalies. *Proc. Natl. Acad. Sci.* **119**, e2121531119 (2022).

[2] ***** indicates co-first authorship.

2.1. Preface

My involvement in this study began as a simple request from Dr. Shachar-Hill and Dr. Xu to model the decrease in the proportion of ^{12}C over all carbon in the CBC intermediates of a *C. sativa* leaf in a ^{13}C labeling experiment. Dr. Sharkey had previously shown using a semi-log plot of labeling over time that the incorporation of ^{13}C label into the CBC intermediates could be fit to two distinct-looking lines, which he interpreted, as per old biochemical convention, to correspond to two processes acting on distinct time scales. These two processes were seen as mapping to the contributions of two distinct pools to the labeling time course of CBC intermediates – the contribution of the CBC intermediate pool back to itself, and the contribution of the cytosolic sugar pool via the glucose-6-phosphate shunt. This provided a convincing independent line of evidence in favor of the glucose-6-phosphate shunt's contribution to CBC labeling in prior studies, and so he, along with Dr. Shachar-Hill and Dr. Xu, were interested in demonstrating something similar in this study of *C. sativa* they were working on.

Dr. Shachar-Hill noted that, mathematically speaking, the practice of fitting straight lines to a semi-log plot does not seem to map one-to-one with the idea of identifying processes acting over distinct time scales. Because of this, he was interested in using nonlinear modeling to fit the untransformed % ^{12}C data, for which he enlisted my help. In the course of modeling the labeling data gathered by Dr. Xu, Dr. Shachar-Hill and I refined his suspicions about the problems with the semi-log plot approach. Indeed, in the area of pharmacokinetic modeling, a couple of facts had been long-since established:

1. The movement of a metabolite or label through a system of compartments interlinked by pseudo first-order kinetic processes, which the labeling of the CBC intermediate pool, influenced by cytosolic sugars, is to a first approximation, can be described by sum-of-exponential, or polyexponential, models.
2. Curve-stripping, which is a more refined version of the practice of fitting straight lines to a semi-log plot, only gives a rough approximation of the proper nonlinear fits you get from performing a nonlinear regression using such polyexponential models.

Taken together, these suggested that in order to make an inference about the number of pools contributing to CBC labeling, we really did need to use a nonlinear regression approach. But, upon implementing such an approach – and in the course of doing so, dealing with issues surrounding the heteroskedasticity of the dataset – it became apparent that one could reasonably

fit several different models (ones assuming the presence of two, three, or four compartments, and with or without metabolically inactive pools). It was at this point that I stopped viewing this all as merely a small side-project and realized that there was a substantial intellectual contribution that I could make to the work. Dr. Shachar-Hill and others had been bothered by a previous ^{13}C-MFA study of *Arabidopsis thaliana*'s central carbon metabolism due to its inclusion of metabolically inactive pools that both allowed the mathematical solver too much flexibility in fitting the data and did not cohere with our understanding of the leaf metabolism. Moreover, there had always been a substantial degree of fuzziness in the selection and justification of a specific ^{13}C-MFA network architecture, given that there are always alternative models one could posit and reasonably fit to a dataset (as discussed in Chapter 1). The polyexponential fitting using nonlinear regression and application of model selection techniques to pick a best-supported model, which could then be mapped back to a broader compartmental model of metabolism that may or may not include metabolically inactive pools, seemed like a clever way of tackling these concerns.

As I carried out the study, it became apparent that Dr. Xu's data actually fit best to a polyexponential model with three terms, corresponding to a three pool model, as opposed to the two pool model we had initially expected. Further data gathering showed convincing evidence of a third pool, the vacuolar sugars, whose inclusion in the ^{13}C-MFA network used in this study substantially improved model fits. I corroborated this finding by fitting *Nicotiana tabacum* CBC intermediate data gathered by Dr. Xinyu Fu in Dr. Berkeley Walker's lab here at MSU, which they kindly provided to us prior to its inclusion in one of their publications, which I cite later on in this chapter. This data was not published at the time that we were submitting our paper to the Proceedings of the National Academy of Sciences, and had to be excluded from the final publication as a result. However, now that those data are published, I have readded the *N. tabacum* data analysis back into the study featured in this chapter.

This study was published in the Proceedings of the National Academy of Sciences. Dr. Yuan Xu, Dr. Thomas Wieloch, and I share co-first authorship on the study. I carried out all of the regression analyses and contributed significantly to the theory and formulation of questions around the regression portion of the study. I also contributed to the writing of the methods, results, and discussion pertinent to my section of the study, as well as editing, proofreading, and formatting of the rest of the main text and the supplement. Finally, I assisted Dr. Xu in the computational implementation of uncertainty estimation in the study.

2.2. Abstract

When isotopes of carbon are fed to photosynthesizing leaves, metabolites of the Calvin–Benson cycle (CBC) are rapidly labeled initially, but then the rate of labeling slows considerably, raising questions about the integration of the CBC within leaf metabolism. We have used 2-h time courses of labeling of *Camelina sativa* leaf metabolites to test models of ^{12}C washout when the CO_2 source is rapidly switched to $^{13}CO_2$. Fitting exponential functions to the time course of CBC metabolites, we found evidence for three temporally distinct processes contributing to the labeling but none for metabolically inactive pools. We next modeled the data of all metabolites by ^{13}C isotopically nonstationary metabolic flux analysis, testing a variety of flux networks. In the model that best explains measured data, three processes determine CBC metabolite labeling. First is fixation of incoming $^{13}CO_2$; second is dilution by weakly labeled carbon in cytosolic glucose reentering the CBC following oxidative pentose phosphate pathway reactions, which forms a shunt bypassing much of the CBC. Third, very weakly labeled carbon from the vacuole further dilutes the labeling. This model predicts the shunt proceeds at about 5% of the rate of net CO_2 fixation and explains the three phases of labeling. In showing the interconnection of three compartments, we have drawn a more complete picture of how carbon moves through photosynthetic metabolism in a way that integrates the CBC, cytosolic sugar pools, glucose-6-phosphate shunt, and vacuolar sugars into a single system.

2.3. Significance statement

Photosynthesis metabolites are quickly labeled when $^{13}CO_2$ is fed to leaves, but the time course of labeling reveals additional contributing processes involved in the metabolic dynamics of photosynthesis. The existence of three such processes is demonstrated, and a metabolic flux model is developed to explore and characterize them. The model is consistent with a slow return of carbon from cytosolic and vacuolar sugars into the Calvin–Benson cycle through the oxidative pentose phosphate pathway. Our results provide insight into how carbon assimilation is integrated into the metabolic network of photosynthetic cells with implications for global carbon fluxes.

2.4. Introduction

The Calvin–Benson cycle (CBC) of photosynthesis is the source of nearly all carbon in the biosphere. CO_2 is used in a carboxylation reaction catalyzed by rubisco, and the resulting carboxylic acid, 3-phosphoglycerate (PGA), is reduced to a sugar using NADPH and helped by adenosine 5′-triphosphate (ATP) made by light-driven photosynthetic electron transport. The

reactions involve both gluconeogenesis and the nonoxidative reactions of the pentose phosphate pathway (PPP) (Sharkey, 2019). Since the first description of the CBC by Bassham et al. (Bassham et al., 1954), the core reactions have been confirmed many times. However, this metabolism is embedded in the metabolic network of photosynthesizing cells. Carbon leaves the cycle primarily by export of triose phosphate from the chloroplast for sucrose synthesis in the cytosol (Fliege et al., 1978; Flugge and Heldt, 1991), and conversion of fructose 6-phosphate (F6P) to glucose 6-phosphate (G6P) for synthesis of starch inside the chloroplast (Dietz, 1985; Sharkey et al., 1985; Dietz, 1987; Preiser et al., 2020). Many other exports from the cycle occur, especially erythrose 4-phosphate (E4P) for the phenylpropanoid pathway, pyruvate for fatty acid synthesis, and pyruvate and glyceraldehyde 3-phosphate (GAP) for the methyl erythritol 4-phosphate pathway that leads to isoprenoid synthesis (Sharkey et al., 2020).

Another set of reactions comprise the photorespiration pathway. When rubisco fixes oxygen instead of CO_2, a series of reactions involving three organelles and amino acid metabolism is initiated that results in 3/4 of the carbon first lost to 2-phosphoglycolate being returned to the CBC.

In addition to photorespiratory production of CO_2, CO_2 is released by a process originally called dark respiration in the light (Farquhar et al., 1980) but now called day respiration (Tcherkez et al., 2017), or light respiration (RL) (Xu et al., 2021a). A static analysis of label in metabolites following $^{13}CO_2$ feeding (Sharkey et al., 2020) pointed to the oxidative PPP (OPPP) as the source of the bulk of RL, for which our recent metabolic flux analysis (MFA) work provides detailed support (Xu et al., 2021a). However, there remain several puzzling observations on CBC kinetics that date back to early quantitative tracer studies (Mahon et al., 1974; McVetty and Canvin, 1981) and are reinforced by recent $^{13}CO_2$-based MFA studies.

- CBC intermediates label very quickly up to 80 to 90% of ^{13}C, but the last 10 to 20% of labeling is much slower (Hasunuma et al., 2010; Szecowka et al., 2013; Ma et al., 2014; Arrivault et al., 2017; Arrivault et al., 2019).
- The proportion of fully unlabeled molecules remains anomalously high well after most molecules are highly labeled [see Szecowka et al. (Szecowka et al., 2013) and **Appendix A, Table S2.4**, where M0 is greater than M1].
- To achieve acceptable fits, previous MFA studies assumed large metabolically inactive pools of central metabolites including metabolites of the CBC (Szecowka et al., 2013; Ma

et al., 2014; Arrivault et al., 2017; Arrivault et al., 2019). However, there is little biochemical evidence for their existence.

- Previous studies fixed numerous fluxes, including starch and sucrose biosynthesis, according to independently measured experimental values (Ma et al., 2014; Xu et al., 2021a). Recently, it was recommended to minimize fixed fluxes and imposed constraints in MFA analyses and compare independent experimental values with model outputs rather than using them as model inputs (Wieloch, 2021).
- Estimates of the relative rate of photorespiration, that is, the ratio of velocities of oxygenation/carboxylation (v_o/v_c), in MFA, are low (Xu et al., 2021a) or light dependent (Ma et al., 2014).

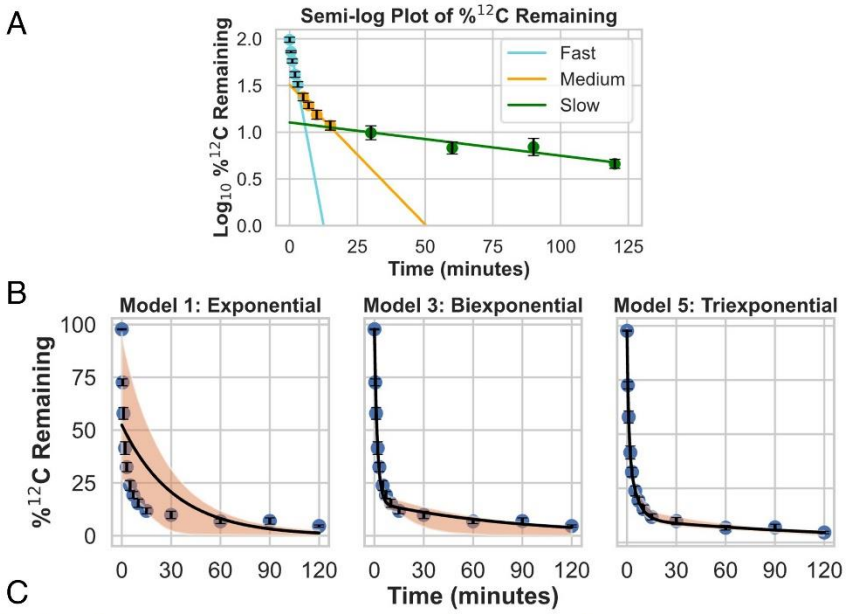
These anomalies indicate that we do not fully understand how the CBC is integrated into the metabolic network of photosynthetic cells. To explore them, we have extended a previously published dataset of leaf isotope labeling (Xu et al., 2021a) to 2 h, added data for neutral sugars, and examined the processes underlying labeling behavior. We applied several statistical tests of the interpretation of three linked processes. We also have made modifications to the isotopically nonstationary (INST)-MFA of photosynthetic metabolism (Ma et al., 2014; Young, 2014; Allen and Young, 2020). We find that three kinetic processes of labeling in CBC metabolites can be defined, and we propose pathways for each. The proposed network of carbon flow eliminates the need to hypothesize metabolically inactive pools and explains both the observed labeling of neutral sugars due to slow dynamic turnover of these products and the high ratio of unlabeled molecules (M0 isotopologue) to singly labeled ones (M1 isotopologues).

2.5. Results

2.5.1. The CBC shows three kinetic components

Following a switch from $^{12}CO_2$ to $^{13}CO_2$, a semilog plot of ^{12}C levels for the CBC intermediates RUBP, PGA, E4P, S7P, GAP, dihydroxyacetone phosphate (DHAP), and FBP **(Appendix A, Dataset S1)** shows three straight lines **(Fig. 2.1A)**. This practice of fitting straight lines on a semilog plot and/or curve stripping is borrowed from pharmacokinetics and serves as an approximation of a polyexponential model with N terms, where N is the number of decay processes acting on distinct time scales (Gibaldi and Perrier, 1982; Dunne, 1985). Interestingly, if a metabolic network is represented as a kinetic model with first-order or pseudo-first-order kinetics and M compartments or pools, the analytical solutions for the isotopic labeling in the

different compartments correspond to polyexponentials containing M terms (**Appendix, A Supplementary Text T1** and **Fig. S2.1**).



#	K	Model Formula	SoS	CV	AIC	BIC
1	2	$f(t) = Ae^{b*t}$		[-2.03,-1.98]	1579	1586
2	3	$f(t) = Ae^{b*t} + C$	0.00	[-0.212,-0.201]	1007	1017
3	4	$f(t) = Ae^{b*t} + Ce^{d*t}$	0.00	[-0.121,-0.119]	868	882
4	5	$f(t) = Ae^{b*t} + Ce^{d*t} + E$	0.00	[-0.108,-0.105]	827	845
5	6	$f(t) = Ae^{b*t} + Ce^{d*t} + Ee^{f*t}$	0.00	[-0.0958,-0.0931]	807	828
6	7	$f(t) = Ae^{b*t} + Ce^{d*t} + Ee^{f*t} + G$	1.00	[-0.0973,-0.094]	809	833
7	8	$f(t) = Ae^{b*t} + Ce^{d*t} + Ee^{f*t} + G^{h*t}$	1.00	[-0.0972,-0.0954]	810	838

Figure 2.1: Modeling of exponential decay of ^{12}C in photosynthesis metabolites. (A) A semilog plot showing the log transformed $\%^{12}C$ remaining in a time course dataset of aggregated CBC intermediates (DHAP, E4P, FBP, GAP, PGA, RUBP, and S7P) (n = 254). Error bars represent mean ±2 SE in A and B. Measured time points of labeling levels fitted by alternative models in the early, middle, and late periods of the labeling time course show evidence for three distinct processes. (B) The exponential, biexponential, and triexponential model fits to the $\%^{12}C$ remaining time course for CBC intermediates in the linear domain. The orange shaded area represents the 95% CI of the regression line obtained via bootstrap resampling (resampling n = 1,000). (C) A table summarizing the nested models we fitted to our data using nonlinear regression and model selection results. K: number of model parameters; SoS: extra sum of squares; CV: cross-validation. Green cells indicate that the model selection criterion results for a given model support it as statistically superior to the previous model, orange cells indicate that they do not support it as superior to the previous model, and gray cells indicate that the criterion cannot be evaluated. Details about the calculation and interpretation of these model selection criteria can be found in Appendix A, Supplemental Methods T2. These results uniformly point to the triexponential model without a constant reflecting an inactive pool as the best supported description of our aggregated CBC labeling dataset.

This indicates that we can fit our metabolite labeling data directly to polyexponential models and, by using model selection techniques to find the model that best describes our data, relate this to an underlying network architecture.

Nonlinear regression and model selection strongly support the existence of three distinct processes controlling the labeling of CBC intermediates but do not support inactive pools. **Fig. 2.1B (see also Appendix A, Fig. S2.2)** shows the results of fitting the measured ^{12}C levels in the aforementioned aggregated CBC intermediates of *Camelina sativa* to models with one, two, or three exponential components (corresponding to one to three processes controlling labeling kinetics). We evaluated which model best describes the data, using four statistical model selection criteria. Each represents a different measure of overfitting and approach to model selection. All four statistical criteria support the existence of three exponential components in the CBC labeling time course **(Fig. 2.1C)**, corresponding to an overall metabolic network involving fluxes among three compartments/pools. The model did not show statistically significant improvement in the fit by including constant terms, which would correspond to metabolically inactive pools. Labeling of the aggregate and individual CBC intermediates—as well as ADP glucose (ADPG)—shares similar kinetic parameter values **(Appendix A, Dataset S2)**, consistent with their high rates of interconversion and turnover during photosynthesis, resulting in rapid "mixing" of carbon between them.

Our data are best described by a triexponential model without constants (**Fig. 2.1C**; model 5 approximates the data significantly better than model 4; model 6 provides no statistically significant improvement). This corresponds to a network in which three interlinked pools contribute to ^{13}C labeling and argues against inactive metabolite pools. We do note that the model selection results for model fits to individual metabolite datasets **(Figures S2.3 – 2.11)** do not uniformly support a triexponential model, though we believe that the aggregated dataset, with its substantially larger sample size, represents a stronger indicator of the overall behavior and structure of the system, which is what we are interested in.

Is this network architecture unique to the CBC in *Camelina sativa* or does it generalize to other plant species? We performed the same exercise of fitting the ^{13}C labeling of aggregated and individual CBC intermediates using a labeling dataset gathered from *Nicotiana tabacum* (Fu et al., 2023). We find that the model selection results for the aggregated CBC intermediates as well as some of the individual metabolites (RUBP, PGA, and GAP/DHAP) also support the

triexponential model **(Figure 2.2).** This suggests that the gross architecture of three interlinked pools contributing to ^{13}C labeling is a general feature of CBC activity in higher plants, rather than a quirk of *C. sativa*'s metabolism. To elucidate these pools and their interconnectivity, we now model carbon metabolism by ^{13}C isotopically nonstationary MFA.

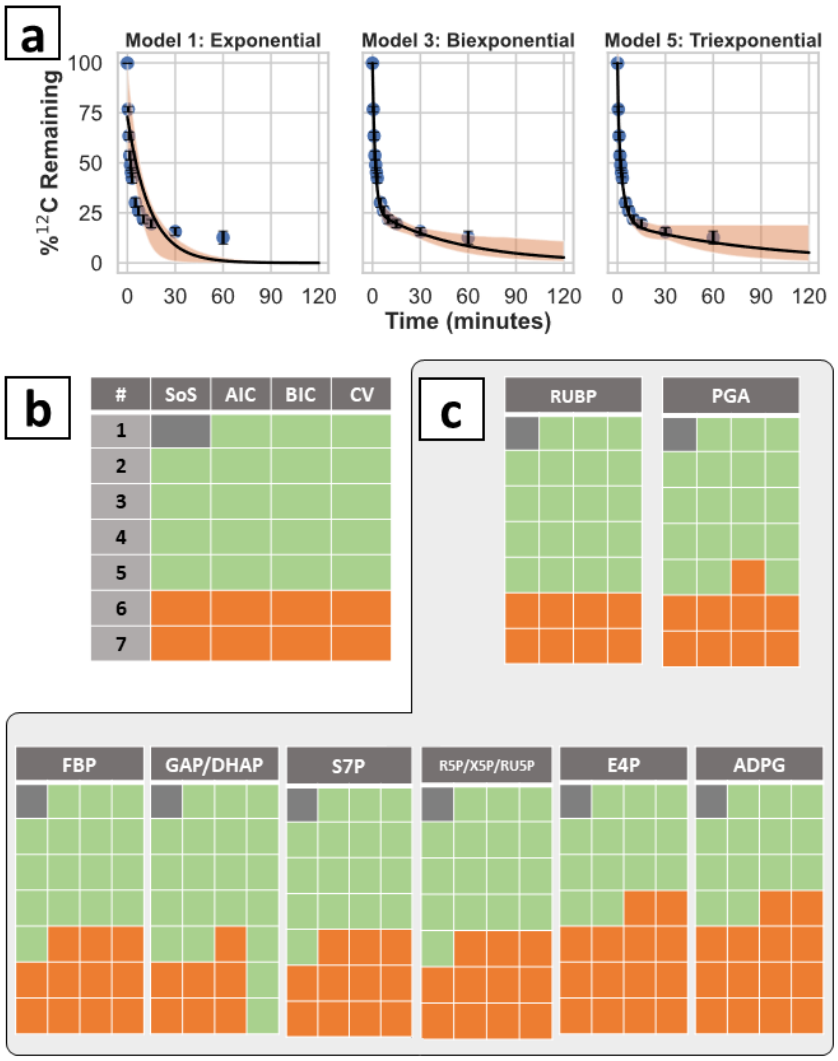


Figure 2.2: Nonlinear regression fits for single, biexponential, and triexponential models fitted to an aggregated CBC intermediate dataset from Nicotiana tabacum along with a summary of model selection results for the aggregated and individual metabolite datasets. (a) Nonlinear regression fits for polyexponential models with the aggregated CBC intermediate dataset from N. tabacum. The orange line represents the best-fit line and the shaded region represents the 95% CI estimated by bootstrap resampling. (b) Model selection results for the aggregated CBC dataset. Green squares represent model selection results supporting the model indicated by that row representing a statistical improvement over a simpler model. Orange squares represent model selection results that do not support adding the additional parameters needed for the model in that row. (c) Model selection results in the same format as that in panel (b) for individual CBC intermediates as well as ADP-glucose.

2.5.2. Network model of three pools of metabolites connected to photosynthesis.

Since we found evidence for three phases of exponential decay and against the contribution of inactive metabolite pools, we looked for processes that might account for the three phases. We began with the hypothesis that unlabeled carbon enters photosynthetic metabolism (Sharkey et al., 2020). We tested four alternatives: 1) entry of ^{12}C glucose into the cytosolic hexose-phosphate pool, which can reach the chloroplast via the cytosolic OPPP shunt and pentose phosphate transmembrane transport on either the xylulose phosphate/phosphate transporter or the triose phosphate/phosphate transporter (Hilgers et al., 2018a); 2) entry of ^{12}C glucose into the chloroplastic hexose phosphate pool to look at the possible contribution of starch turnover; 3) injection of $^{12}CO_2$ into the internal CO_2 pool to simulate an unknown source of older C being broken down; and 4) addition of ^{12}C triose phosphate to the plastid triose phosphate pool to simulate entry via the triose phosphate transporter from an unknown source in the cytoplasm (**Table 2.1** and **Appendix A, Table S2.3**). To do so, we increased the time span and range of metabolites over which labeling was measured and updated our previously developed metabolic model to include reversibility of several reactions for which there is biochemical evidence **(Appendix A, Dataset S3)**. We assessed these alternatives comparing sum of squared residuals (SSR), a measure of the goodness of fit between modeled and measured data (Young, 2014). However, SSR will be affected by how many data points are used and other factors. For this reason, we do not compare SSRs found in this study with those from our previous studies but only look for large reductions in SSRs when datasets and degrees of freedom are similar.

Table 2.1: Comparison of goodness of fit between data and best-fit simulations from alternative models.

Model	Reactions	Flux	SSR	ΔDOF
No inactive pools			1,340	0
No inactive pools + unlabeled carbon source	CO_2.u -> CO_2	0	1,340	1
	Glucose.u -> G6P.p	0.5	1,300	1
	TP.u -> TP.p	0.3	1,273	1
	Glucose.u -> G6P.c	1.9	1,126	1
No inactive pools + sucrose recycling reactions + sugar vacuole pool reactions	Glc.v <-> Glc.c	2.11	968	5*
	Suc.c-> Glc.c + Fru.c	0.05		
	Glc.c -> G6P.c	2.16		

Starting model with no inactive pools, model with unlabeled glucose source, and model with sucrose recycling and sugar vacuole pool reactions were compared with fluxes for key reactions, SSR, and Δ degree of freedom (ΔDOF); 5* denotes ΔDOF in terms of net fluxes. The lowest value of SSR is shown in blue, the 50th percentile of SSR is shown in yellow, and the highest value of SSR is shown in red. Except for the reactions described the table. all measured metabolites involved in the reactions in Fig. 3 were included in the SSR values.

The data were consistent with ^{12}C entry from intact unlabeled glucose via the OPPP shunt at a rate of 1.9 μmol·g^{-1} FW·h^{-1}, with the best SSR improvement from 1,340 to 1,126. The second possible ^{12}C entry flux is 0.3 μmol·g^{-1} FW·h^{-1} from the triose phosphate transporter from an

unknown source in the cytosol, with a decrease in SSR from 1,340 to 1,273. The third possible ^{12}C entry is from starch turnover flux of 0.5 μmol·g^{-1} FW·h^{-1}, with a decrease in SSR from 1,340 to 1,300. We also tested other models with variations in starch metabolism to test 1) whether addition of reactions representing starch turnover to the metabolic model meaningfully improves the agreement between the measured and simulated labeling and other flux data and 2) whether the fitting of such models indicates biologically significant fluxes through starch turnover. We tested six such models with different representations of how starch turnover might act to influence the carbon fluxes and expected labeling patterns. Other models were tested in which either the whole starch pool or an intermediate pool (such as might represent either oligoglucans or short- versus long-term starch pools) can turn over while maintaining the measured net starch accumulation rates.

No unknown source of older C being broken down is indicated, with $^{12}CO_2$ entry flux of 0 with no change of SSR **(Table 2.1)**. This result is consistent with the M0 abundance results (see below), as the assimilation of $^{12}CO_2$ would not selectively increase the proportion of unlabeled molecules, because it does not inject intact carbon skeletons. The starch model with the largest improvement in the fit, as defined by the SSR, was no more than a 1% improvement, with a best fit value for a starch turnover flux of no more than 11% of the G6P dehydrogenase activity **(Appendix A, Table S2.3)**.

2.5.3. Examination of labeling in key CBC intermediates supports the hypothesis that intact unlabeled molecules enter the CBC.

In our study's later time points, anomalously high values for fully unlabeled isotopologues (M0) were found well after the singly labeled (M1) isotopologues had decayed to very low levels **(Fig. 2.3)**. Since the percentage of M2 was always bigger than M1, the percentage of M1 should also be bigger than M0. However, we found the reverse pattern. The ratio of the measured percentages of M1 to M0 ranged from 0.1 to 0.4, much smaller than the predicted ratio range of 48 to 175 **(Appendix A, Table S2.4)**. If inactive metabolite pools cause the lack of complete labeling, then, at later time points, for example, in G6P, only M0 and M6 should be observed. However, the amount of M0 could account for only one-third of the ^{12}C in G6P at 2 h **(Appendix A, Table S2.5)**. We suggest that the high amount of M0 comes from a large metabolic pool, such as fully unlabeled glucose that enters the CBC at a low rate.

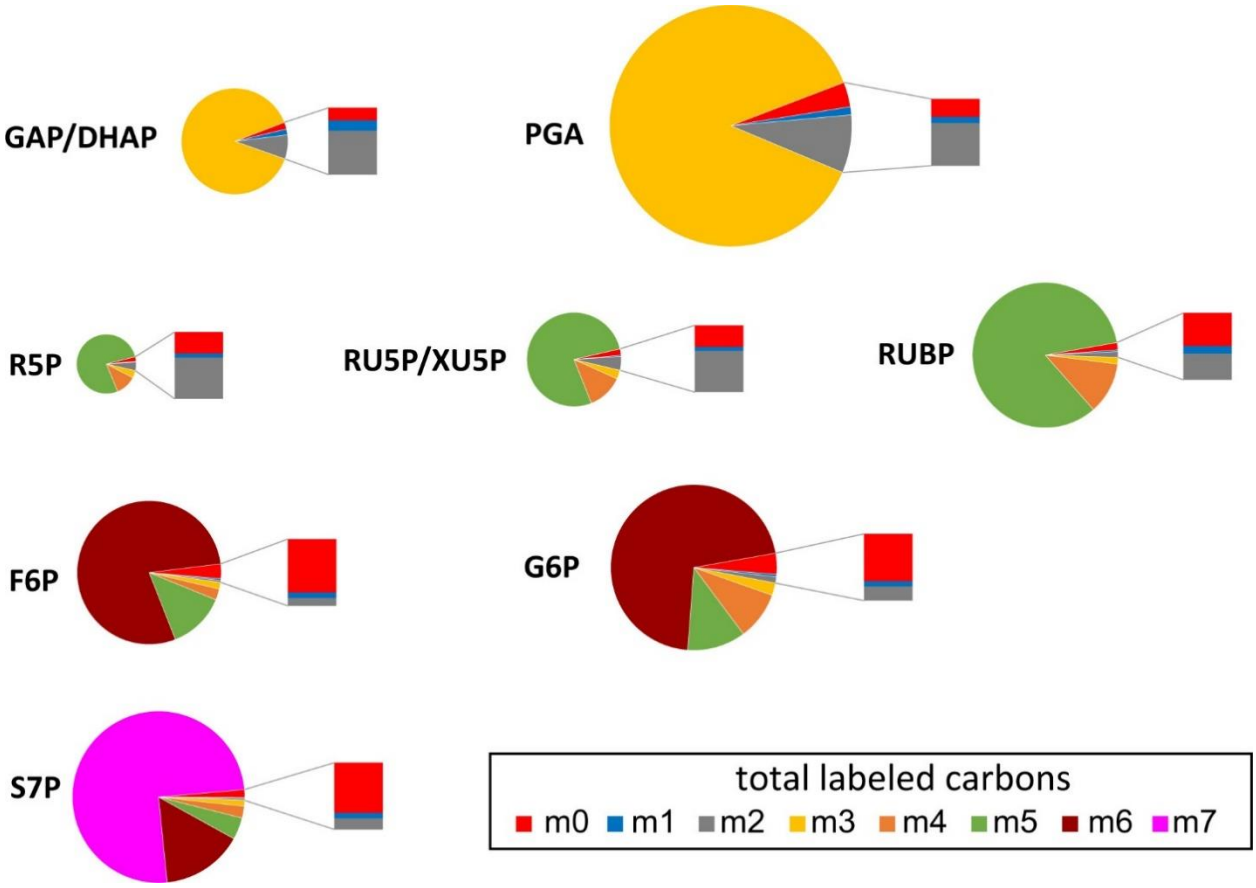


Figure 2.3: Mass isotopologue distributions of CBC intermediates showing the overabundance of M0 isotopologue at the latest time points. Percentages of relative abundance of each isotopologue for key CBC intermediates at 1 h are shown, with different colors corresponding to different isotopologues (figure legend). The size of each pie chart corresponds to the pool size of that metabolite. An expanded bar next to each pie chart shows proportions of M0, M1, and M2 isotopologues, highlighting the overabundance of the M0 relative to the M1 isotopologue. Abbreviations (see also Table S2.4): GAP, glyceraldehyde 3-phosphate; DHAP, dihydroxyacetone phosphate; PGA, 3-phosphoglyceric acid; R5P, ribose 5-phosphate; RU5P, ribulose 5-phosphate; XU5P, xylulose 5-phosphate; RUBP, ribulose 1,5-bisphosphate; F6P, fructose 6-phosphate; G6P, glucose 6-phosphate; S7P, sedoheptulose 7-phosphate.

We also observed slow turnover of neutral sugars, which suggests that a dilution flux of largely or wholly unlabeled hexose enters the CBC over an extended period during labeling experiments (**Appendix A, Fig. S2.12**; labeling kinetics for other metabolites are shown in **Appendix A, Figs. S2.13 and S2.14**). At 60 min, the glucosyl and fructosyl moieties of sucrose contained 49% and 46% ^{13}C, respectively **(Appendix A, Fig. S2.12)**. Sucrose recycling through invertase and fructokinase yields F6P that would distribute between sucrose resynthesis and G6P, but this alone is insufficient to account for a prolonged dilution flux. By contrast, at 60 min, glucose and fructose were only 12% and 20% labeled with ^{13}C, respectively **(Appendix A, Fig.**

S2.12), consistent with previous evidence that vacuolar sucrose turns over due to invertase activity (Uys et al., 2007; Nägele et al., 2010; Patrick et al., 2013). If a modest proportion of cytosolic G6P originates from the action of hexokinase on glucose leaving the vacuole, then there would be an additional source of unlabeled carbon in the cytosolic G6P pool. Sucrose recycling and turnover of vacuolar sugars could therefore slow the ^{12}C decay in CBC intermediates and correspond to the additional carbon pool attested to by the polyexponential modeling.

2.5.4. An integrated flux model with three compartments.

In light of the above results, we included sucrose recycling and sugar vacuole pool transport reactions in the model with known biochemical reactions that can mediate such slow turnover processes of sucrose/glucose/fructose. Inclusion of sucrose recycling and sugar vacuole pool reactions markedly reduced overall SSR from 1,340 to 968 and reduced individual SSRs for labeling in the least well-fitted metabolites F6P, ribose 5-phosphate (R5P), G6P, ADPG, and UDPglucose (UDPG) (**Table 2.1** and **Appendix A, Table S2.3**).

Fig. 2.4 shows the flux map for photosynthetic carbon metabolism for the model, with sucrose recycling reactions and sugar vacuole pool reactions in orange. The nonphotorespiratory CO_2 release during photosynthesis from the cytosolic G6P shunt was estimated at 5% of net CO_2 fixation compared to a photorespiratory CO_2 release of 18% of net CO_2 fixation **(Appendix A, Table S2.6)**. While intermediates of the CBC, photorespiration, and starch and sucrose biosynthesis pathways showed substantial ^{13}C labeling, the tricarboxylic acid (TCA) cycle intermediates, and most amino acids derived from them, showed very little labeling after 120 min. The flux map is consistent with previous reports of low TCA fluxes and operation of the OPPP shunt as a source of RL (Xu et al., 2021a). The 95% CIs of the flux values were estimated by both parameter continuation and Monte Carlo methods. These CI estimates showed that the net fluxes whose magnitude approaches or exceeds 1% of the rate of photosynthesis are well defined, with ranges less than ±5% of their values **(Appendix A, Dataset S2.4).** Exchange fluxes are less well defined, especially for reactions with modest net fluxes.

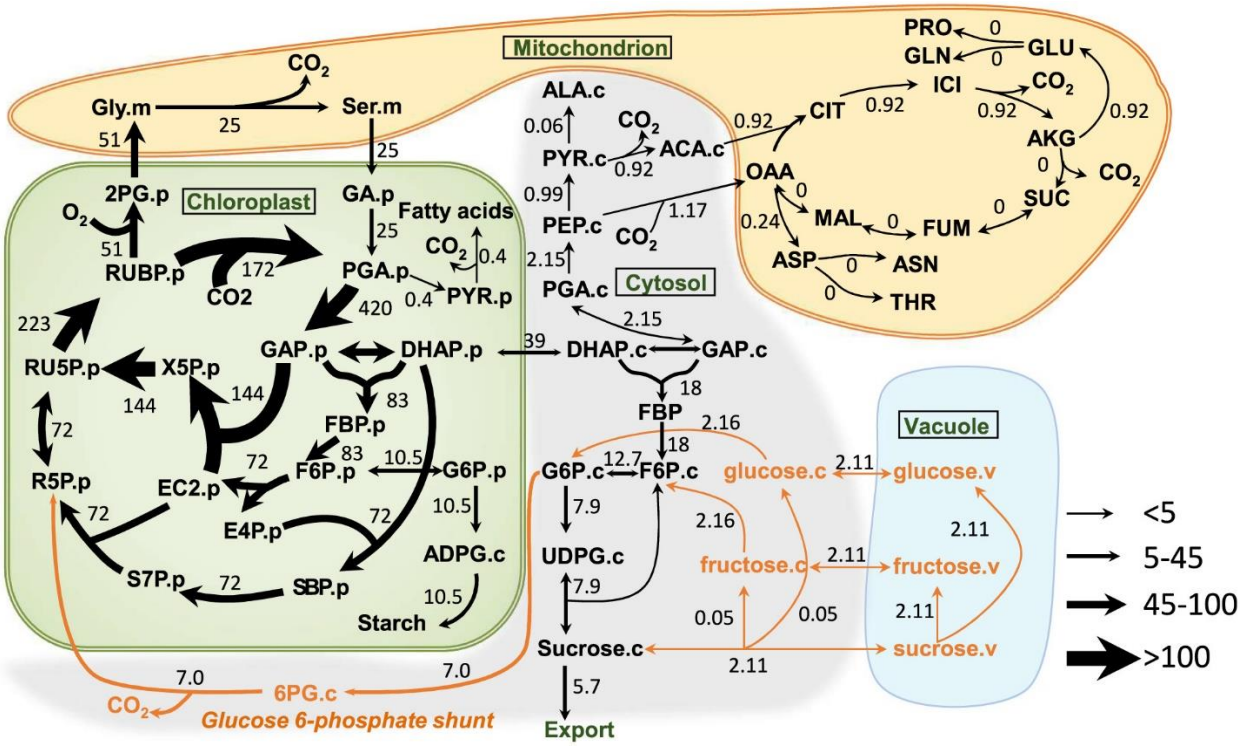


Figure 2.4: Central carbon metabolic fluxes in photosynthetic *C. sativa* leaves. Fluxes are shown as numbers and depicted by the variable width of arrows. Orange arrows highlight the carbon flow from neutral sugars through the G6P shunt, entering the CBC. Fluxes were estimated by ^{13}C INST-MFA using the INCA software suite constrained by the metabolic network model and experimental inputs including mass isotopologue distributions of measured metabolites, net CO_2 assimilation, sucrose and amino acid export rate, and measured v_o/v_c ratio. Flux units are expressed as micromoles metabolite per gram FW per hour. The model network is compartmentalized into cytosol (".c"), chloroplast (".p"), mitochondrion (".m"), and vacuole (".v"). Abbreviations: ACA, acetyl-CoA; AKG, α-ketoglutarate; ALA, alanine; ASN, asparagine; ASP, aspartate; CIT, citrate; DHAP, dihydroxyacetone phosphate; EC2, transketolase-bound-2-carbon-fragment; FBP, fructose-1,6-bisphosphatase; FUM, fumarate; GA glycerate; GLN, glutamine; GLY, glycine; ICI, isocitrate; MAL, malate; OAA, oxaloacetate; PEP, phosphoenolpyruvate; PYR, pyruvate; RU5P, ribulose-5-phosphate; RUBP, ribulose-1,5-bisophosphate; S7P, sedoeheptulose-7-phosphate; SBP, sedoheptulose-1,7-bisophosphate; SER, serine; SUC, succinate; THR, threonine.

2.5.5. Model prediction of photorespiration.

The estimation of the photorespiration rate in leaves by ^{13}C MFA is complicated by the presence of multiple subcellular pools of serine and glycine and the multiple reactions and interconversions that they can undergo in different compartments (Hanson et al., 2000), and the challenges in obtaining reliable measurements of levels, compartmentation, and labeling of other photorespiratory metabolites (Ma et al., 2017). Here we measured labeling in 2-phosphoglycolate (2PG) but were not able to reliably measure labeling in glycolate, glyoxylate, hydroxypyruvate, or

glycerate. In the absence of such additional measurements, the reliability of photorespiratory flux estimates is low, with a substantial range of possible rates, which increases if realistic compartmentation of glycine and serine is included. We therefore estimated v_o/v_c using gas exchange measurements **(Appendix A, Supplementary Information Text T2.3)**. The value obtained (0.31) was used as input to the MFA model instead of relying on fitting the labeling measured in glycine and serine **(Appendix A, Table S2.7)**. Using measurements of serine, glycine, 2PG, and glycerate without compartmentation, Ma et al. (Ma et al., 2014) obtained a v_o/v_c ratio of 0.28 to 0.43 in *Arabidopsis* under low and high light levels, which is consistent with the value estimated here.

2.5.6. No metabolically inactive CBC metabolites.

The inclusion of inactive metabolite pools was made in previous studies to account for the persistence of unlabeled carbon in CBC intermediates (Szecowka et al., 2013; Ma et al., 2014; Arrivault et al., 2019; Xu et al., 2021a). Whole shoots may include enough photosynthetically inactive tissues to account for significant inactive pools, while single mature leaves used here should have very little photosynthetically inactive tissue. We therefore eliminated model terms accounting for inactive metabolite pools included in previous studies (Szecowka et al., 2013; Ma et al., 2014; Xu et al., 2021a) for all metabolites except glycine, serine, and alanine, for which significant vacuolar pools with long turnover times are plausible (Fürtauer et al., 2019). The model without inactive pools failed to adequately explain the labeling dataset, with particularly poor agreement for F6P, R5P, G6P, ADPG, and UDPG **(Appendix A, Table S2.3)**.

To test the model shown in **Fig. 2.4**, we added the inactive pools removed earlier back into the model to see whether introducing our mechanistic explanations for the labeling dynamics of the metabolites in this network changed the inactive pool size estimates. Compared to the previous study, we found this model substantially lowered the estimated inactive pool sizes in the best-fit simulations **(Appendix A, Fig. S2.15)** compared to previous studies (Szecowka et al., 2013; Ma et al., 2014; Xu et al., 2021a). Among them, the inactive pools for RUBP, PGA, hexose 6-phosphates, RU5P, 2PG, ADPG, and UDPG were decreased to almost zero, indicating that the turnover of sugars better explains the proportion of unlabeled molecules in these metabolites than the idea of inactive pools.

2.6. Discussion

A key finding from this study is that the kinetics of the CBC is best described as a function of three interconnected processes, as indicated both by our modeling analysis of the time course of ^{12}C decay during $^{13}CO_2$ labeling experiments **(Fig. 2.1)** and by our MFA modeling results (**Appendix A, Fig. S2.15**). Our model included three inputs of carbon into the CBC: 1) 172 $\mu mol \cdot g^{-1}$ FW$\cdot h^{-1}$ by carboxylation by rubisco, 2) 75 (25 × 3 carbons per glycerate) $\mu mol \cdot g^{-1}$ FW$\cdot h^{-1}$ returned from photorespiration, and 3) 35 (7 × 5) $\mu mol \cdot g^{-1}$ FW$\cdot h^{-1}$ returned from the G6P shunt. The carbon paths in photorespiration and the G6P shunts require an extra 110 (75 + 35) carbon atoms to be processed for 172 carboxylations, adding more than 50% to the required flux through reactions in the CBC.

In previous work, we allowed only a stromal shunt (Xu et al., 2021a). When we allowed both a stromal and a cytosolic shunt with our expanded dataset, all shunt carbon flow was assigned to the cytosolic shunt, and other work based on label in 6-phosphogluconate indicated that, in unstressed plants, only the cytosolic shunt operates (Sharkey et al., 2020). Therefore, we left the stromal shunt out of the final model.

The model includes release of CO_2 in photorespiration at a rate of 25 $\mu mol \cdot g^{-1}$ FW$\cdot h^{-1}$ and, from the G6P shunt, at a rate of 7 $\mu mol \cdot g^{-1}$ FW$\cdot h^{-1}$. The rate of glucose entry into the shunt was estimated to be about 5% of the rate of net CO_2 fixation. The cost of the shunt is three ATP per glucose. Therefore, this shunt would increase the energy requirement for CO_2 fixation from three ATP and two NADPH to ~3.15 ATP and two NADPH (photorespiration also affects the energy cost of CO_2 fixation) (Sharkey and Weise, 2016; Sharkey et al., 2020).

The cost of the G6P shunt may be offset by benefits of refilling the CBC when intermediates fall during transients in light or other factors (Sharkey and Weise, 2016). This has also been proposed by Makowka et al., (2020) for glycolytic pathways in cyanobacteria.

2.6.1. MFA model fits.

The use of multiple statistical tests specifically designed for model selection and the comparison of nested model series shows the potential for improvement of statistical rigor in this important aspect of MFA modeling. Although MFA software packages like INCA (Young, 2014) can report out 95% CIs for SSRs, allowing researchers to flag overfit or underfit models, these expected ranges are not appropriate for comparing alternative model architectures. This study demonstrated that, by directly modeling ^{13}C labeling time course data, we can test models of the

general structure of the underlying network and corroborate or contradict assumed or proposed MFA models. This attests to the possible utility of these kinds of statistical tools in constraint-based modeling, and we believe advancement in this area could encourage use of MFA models to gain insight into photosynthetic metabolism.

2.6.2. Reaction network improvement.

This new model improves on previous efforts on several fronts. Comparisons of the model in this work with previous models (Ma et al., 2014; Xu et al., 2021a) are shown in Appendix A, Dataset S3.3. The reversibility of reactions in the CBC has been corrected. Reactions present in the previous models, representing inactive pools for all the CBC intermediates, ADPG, UDPG, 2PG, phosphoenolpyruvate, and glycerate have been removed. The inactive pools for alanine, glycine, and serine have been retained because of their compartmentation complexity. Reactions newly added in this study, including cytosolic OPPP shunt, sucrose recycling reactions, and sugar vacuole pool reactions, explain the longstanding puzzle of the slow labeling phase of CBC intermediates and the overabundance of fully unlabeled isotopologues. These improvements to the metabolic network have largely eliminated the need for hypothesizing inactive pools. In showing the interconnection of these three compartments, we have drawn a picture of how carbon moves through photosynthetic carbon assimilation in a way that integrates the CBC, cytosolic sugar pools, the glucose-6-phosphate shunt, and vacuolar sugars into a single system.

The data are consistent with a cytosolic G6P shunt. A stromal shunt would be undetectable, since the carbon source for a stromal shunt would have the same labeling kinetics as the rest of the CBC, as indicated by the similarity of labeling of ADPG and CBC intermediates. Measurements of the label in 6-phosphogluconate indicated that, in unstressed conditions, only the cytosolic shunt was active, while, in high temperature stress, a stromal shunt also occurs (Sharkey et al., 2020). When models that included both shunts were tested, no flux was assigned to the stromal shunt. The modified model used here predicts that the cytosolic shunt would proceed at a rate that is consistent with measurements of RL made using $^{12}CO_2$ emission into a $^{13}CO_2$-containing atmosphere (Loreto et al., 2001).

2.6.3. Sources of unlabeled carbon.

Our conclusion is that the source of unlabeled carbon that reenters the CBC is sucrose, glucose, and fructose in the cytosol and vacuole. It has been shown that SUC4-type sucrose transporters can allow sucrose release from vacuoles (Payyavula et al., 2011; Schneider et al.,

2012; Anaokar et al., 2021), and SWEET17 can mediate fructose transport across the tonoplast in leaves, although its primary activity may be in roots (Guo et al., 2014). Our model allows chloroplasts to take up pentose phosphates. A xylulose 5-phosphate transporter has been described (Eicks et al., 2002), but we found that plants lacking this gene have no growth or photosynthetic phenotype. The xylulose 5-phosphate transporter will also transport triose phosphates, and it is very possible that the triose phosphate/phosphate transporter is also bifunctional. Plants lacking both the xylulose phosphate-phosphate transporter (XPT) and triose phosphate transporter (TPT) accumulate pentose phosphates and show a much stronger reduction in growth than plants lacking the TPT alone (Hilgers et al., 2018a).

In the past, starch recycling was proposed as a possible source (Sharkey, 2019). We have abandoned that idea, because a source in starch recycling would require that 36% of the carbon going to starch comes back into metabolism, but without any label. This is unrealistic. The results of various models described above **(Appendix A, Table S2.3)** provided clear-cut evidence against a biologically significant contribution of starch turnover to labeling patterns or carbon balances in central metabolism.

2.6.4. Previously puzzling observations explained.

With the insight gained here, we address the metabolism issues raised in the Introduction.

- The CBC intermediates label very quickly up to 80 to 90% of ^{13}C, but the last 10 to 20% of labeling is much slower (Hasunuma et al., 2010; Szecowka et al., 2013; Ma et al., 2014; Arrivault et al., 2017; Arrivault et al., 2019).

The CBC in leaves shows three phases, indicating three components. The slower two components account for the apparent slow-to-label pool. This is well-explained by carbon in unlabeled pools of glucose, fructose, and sucrose reentering the CBC by way of the glucose-6-phosphate shunt in the cytosol. No evidence was found for separate active and inactive pools. Hendry et al. (Hendry et al., 2017) proposed that glycogen could supply unlabeled carbon back to the CBC intermediates in Synechococcus to explain a similar lack of complete labeling.

- The proportion of fully unlabeled molecules remains anomalously high well after most molecules are highly labeled [see Szecowka et al. (Szecowka et al., 2013) and **Appendix A, Table S2.4**, where M0 is greater than M1].

The abundance of M0 over M1 isotopologues was confirmed here. If metabolically inactive pools explained the lack of complete labeling, then the M0 isotopologues should account for all the

unlabeled carbon atoms. However, using G6P as an example, 2.9% of the molecules were fully unlabeled, but this accounts for only about one-third of the missing label (**Appendix A, Table S2.5**). Entry of carbon from relatively unlabeled free sugars into active pools accounts for the preponderance of M0 isotopologues.

- To achieve acceptable fits, previous MFA studies assumed large metabolically inactive pools of central metabolites including metabolites of the CBC (Szecowka et al., 2013; Ma et al., 2014; Arrivault et al., 2017; Arrivault et al., 2019). However, there is little biochemical evidence for their existence.

The new model of metabolism does not predict inactive pools. For all the CBC intermediates, the data fit well assuming carbon reentry through the shunt, eliminating any need to invoke inactive pools (**Appendix A, Fig. S2.15** and **Table S2.3**). The exception is SBP as reported in Arrivault et al. (Arrivault et al., 2017). High levels of M0 were found. This could result from E4P export on the XPT transporter (Hilgers et al., 2018b) followed by attachment of DHAP catalyzed by aldolase. Since there is no SBPase in the cytosol, this would be a metabolic dead end and result in a significant inactive pool of SBP.

- Previous studies fixed numerous fluxes, including starch and sucrose biosynthesis, according to independently measured experimental values (Ma et al., 2014; Xu et al., 2021a). Recently, it was recommended to minimize fixed fluxes and impose constraints in MFA analyses and compare independent experimental values with model outputs rather than using them as model inputs (Wieloch, 2021).

The final model had no fixed fluxes, although the ratio of v_o/v_c was constrained (but not fixed) based on gas exchange data (**Appendix A, Supplementary Information Text T3.3**). Fatty acid synthesis and RL were constrained (but not fixed) based on previous measurements (Xu et al., 2021a). The model returned physiologically reasonable values for starch and sucrose synthesis (Sharkey et al., 1985).

- Estimates of the relative rate of photorespiration, that is, velocity of rates of oxygenation/carboxylation (v_o/v_c), in MFA are low (Xu et al., 2021a) or light dependent (Ma et al., 2014).

We found that v_o/v_c is not well estimated by the model, requiring use of other estimates. Use of MFA to estimate photorespiration rates is less reliable than other methods (Sharkey, 1988; Busch, 2013).

Several models of plant behavior, including isotopic disequilibrium methods for measuring RL (Gong et al., 2018) and isoprene studies [reviewed in (Sharkey et al., 2020)], assume that carbon in photosynthesis is fully labeled after 10 min of feeding air with a different carbon isotopic composition and that other processes contribute "old" carbon that does not become labeled. The results presented here will allow more-refined models that include both the lack of complete labeling of CBC intermediates and the occurrence of some label in the sources for these other processes. Our results indicate that isotopic methods for measuring RL underestimate its rate because the source carbon (G6P in the cytosol) has some label at the time RL is assessed. The results presented here provide a framework for more detailed RL measurements. Measuring RL is a very difficult task but very important for understanding global carbon cycles (Tcherkez et al., 2017).

In summary, labeling of CBC intermediates by fixation of incoming $^{13}CO_2$ is diluted by weakly labeled carbon in glucose reentering the CBC. We predict that reentry of weakly labeled molecules occurs at a rate of 5% of the rate of net CO_2 fixation. The model explains three phases of labeling. In showing the interconnection of three compartments, this model provides a more complete picture of how carbon moves through photosynthetic metabolism in a way that integrates the CBC, cytosolic sugar pools, the glucose-6-phosphate shunt, and vacuolar sugars into a single system.

2.7. Methods

2.7.1. Plant growth, gas exchange, and $^{13}CO_2$ labeling.

Plant growth and gas exchange methods were used as described previously (Xu et al., 2021a). The $^{13}CO_2$-labeled leaf samples were collected at time points of 0, 0.5, 1, 2, 2.5, 3, 5, 7, 10, 15, 30, 60, 90, and 120 min as described in detail in **Appendix A, Supplemental Methods T3.4** and **T3.5**.

2.7.2. Mass spectrometry.

Mass spectrometry for anion exchange LC-MS/MS and GC-EI-MS were carried out using the methods described in ref. (Xu et al., 2021a) and detailed in Appendix A, Dataset S5. Reverse-phase LC-MS/MS and GC-chemical ionization (CI)-MS had the following changes: Samples for reverse-phase liquid chromatography-tandem mass spectrometry were analyzed by an ACQUITY UPLC pump system (Waters) coupled with Waters XEVO TQ-S ultra-performance liquid chromatography tandem mass spectrometry (Waters) by the method described in ref. (Xu et al.,

2021a). Samples for gas chromatography-electron ionization-mass spectrometry were analyzed by an Agilent 7890B GC system (Agilent) coupled to an Agilent 7010B triple quadrupole gas chromatography-electron ionization-mass spectrometer with an autosampler (CTC PAL) (Agilent). An Agilent VF5ms GC column, 30 m × 0.25 mm × 0.25 m with 10-m guard column was used. One microliter of the derivatized sample was injected with helium carrier gas at a flow rate of 1.2 $mL \cdot min^{-1}$. The oven temperature gradient was: 40 °C (1-min hold), increased at 40 °C/min to 150 °C, then a 10 °/min to 250 °C, then a 40 °C/min to 320 °C, and finally held at 320 °C for 4.5 min. CI was used, and the mass scan range was 150 amu to 650 amu with step size 0.1 amu. The ionization source temperature was set at 300 °C, and the transfer line temperature was 300 °C.

2.7.3. Nonlinear regression and model selection.

A nonlinear ordinary least-squares algorithm implemented in the Python package SciPy was used to fit models 1 to 7 **(Fig. 2.1C)** to our dataset (Virtanen et al., 2020). Briefly, best-fit lines for each model were generated by initializing and estimating model parameters 100 times with randomly selected initial parameters and then selecting the fit with the smallest SSR. CIs for parameters and fitted values were determined using bootstrap resampling (n = 1,000). Extra sum-of-squares, cross-validation, Akaike information criterion (AIC), and Bayesian information criterion (BIC) model selection criteria were evaluated for all models and model comparisons, and the Bonferroni–Holm multiple testing correction was applied for the P values generated by the extra sum-of-squares hypothesis testing (Holm, 1978; Schwarz, 1978; Akaike, 1998; Draper and Smith, 1998; Hastie et al., 2017). Further details can be found in **Appendix A, Supplemental Information Text T3.1** and **Supplemental Methods T3.2**.

2.8. Acknowledgments

This work was supported by the Division of Chemical Sciences, Geosciences and Biosciences, Office of Basic Energy Sciences of the US Department of Energy, Grant DE-FOA-0001650 (Y.X. and Y.S.-H.) and Grant DE-FG02-91ER20021 (T.W. and T.D.S.). T.D.S. receives partial salary support from Michigan State University (MSU) AgBioResearch. This work is supported, in part, by the NSF Research Traineeship Program (Grant DGE-1828149) to J.A.M.K. This publication was also made possible by a predoctoral training award to J.A.M.K. from Grant T32-GM110523 from National Institute of General Medical Sciences (NIGMS) of the NIH. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIGMS

or NIH. We thank the staff of the MSU Research Technology Support Facility Mass Spectrometry Core for excellent support of mass spectrometric analysis. We thank Chih-Li Sung for statistical advice.

2.9. Author contributions

Author contributions: Y.X., T.W., Y.S-H., and T.D.S. conceived the project and designed the experiments; T.W. and T.D.S. conducted exploratory analyses using published labeling data and developed a preliminary INST-MFA model accounting for photosynthesis labeling lags; Y.X. performed the metabolic flux analysis experiments; Y.X. and T.W. modified the kinetic MFA model used here; J.A.M.K. performed the nonlinear modeling and associated statistical tests of ^{12}C labeling data; Y.S.-H. obtained analytical solutions for simple models and provided guidance for the experimental and computational analyses; and all authors contributed to writing the manuscript.

REFERENCES

Akaike H (1998) Information Theory and an Extension of the Maximum Likelihood Principle. *In* E Parzen, K Tanabe, G Kitagawa, eds, Selected Papers of Hirotugu Akaike. Springer New York, New York, NY, pp 199–213

Allen DK, Young JD (2020) Tracing metabolic flux through time and space with isotope labeling experiments. Curr Opin Biotechnol **64**: 92–100

Anaokar S, Liu H, Keereetaweep J, Zhai Z, Shanklin J (2021) Mobilizing Vacuolar Sugar Increases Vegetative Triacylglycerol Accumulation. Frontiers in Plant Science. doi: 10.3389/fpls.2021.708902

Arrivault S, Alexandre Moraes T, Obata T, Medeiros DB, Fernie AR, Boulouis A, Ludwig M, Lunn JE, Borghi GL, Schlereth A, et al (2019) Metabolite profiles reveal interspecific variation in operation of the Calvin-Benson cycle in both C4 and C3 plants. J Exp Bot **70**: 1843–1858

Arrivault S, Obata T, Szecówka M, Mengin V, Guenther M, Hoehne M, Fernie AR, Stitt M (2017) Metabolite pools and carbon flow during C4 photosynthesis in maize: 13CO2 labeling kinetics and cell type fractionation. J Exp Bot **68**: 283–298

Bassham JA, Benson AA, Kay LD, Harris AZ, Wilson AT, Calvin M (1954) The Path of Carbon in Photosynthesis. XXI. The Cyclic Regeneration of Carbon Dioxide Acceptor1. J Am Chem Soc **76**: 1760–1770

Busch FA (2013) Current methods for estimating the rate of photorespiration in leaves. Plant Biol (Stuttg) **15**: 648–655

Dietz K-J (1985) A possible rate-limiting function of chloroplast hexosemonophosphate isomerase in starch synthesis of leaves. Biochimica et Biophysica Acta (BBA) - General Subjects **839**: 240–248

Dietz K-J (1987) Control Function of Hexosemonophosphate Isomerase and Phosphoglucomutase in Starch Synthesis of Leaves. *In* J Biggins, ed, Progress in Photosynthesis Research: Volume 3 Proceedings of the VIIth International Congress on Photosynthesis Providence, Rhode Island, USA, August 10–15, 1986. Springer Netherlands, Dordrecht, pp 329–332

Draper NR, Smith H (1998) Extra Sums of Squares and Tests for Several Parameters Being Zero. Applied Regression Analysis. John Wiley & Sons, Ltd, pp 149–177

Dunne A (1985) JANA: A new iterative polyexponential curve stripping program. Computer Methods and Programs in Biomedicine **20**: 269–275

Eicks M, Maurino V, Knappe S, Flügge U-I, Fischer K (2002) The plastidic pentose phosphate translocator represents a link between the cytosolic and the plastidic pentose phosphate pathways in plants. Plant Physiol **128**: 512–522

Farquhar GD, Caemmerer S, Berry JA (1980) A biochemical model of photosynthetic CO2 assimilation in leaves of C3 species. Planta **149**: 78–90

Fliege R, Flügge UI, Werdan K, Heldt HW (1978) Specific transport of inorganic phosphate, 3-phosphoglycerate and triosephosphates across the inner membrane of the envelope in spinach chloroplasts. Biochim Biophys Acta **502**: 232–247

Flugge U, Heldt HW (1991) Metabolite Translocators of the Chloroplast Envelope. Annu Rev Plant Physiol Plant Mol Biol **42**: 129–144

Fu X, Gregory LM, Weise SE, Walker BJ (2023) Integrated flux and pool size analysis in plant central metabolism reveals unique roles of glycine and serine during photorespiration. Nat Plants **9**: 169–178

Fürtauer L, Küstner L, Weckwerth W, Heyer AG, Nägele T (2019) Resolving subcellular plant metabolism. The Plant Journal **100**: 438–455

Gibaldi M, Perrier D (1982) Pharmacokinetics. M. Dekker, New York, NY

Gong XY, Tcherkez G, Wenig J, Schäufele R, Schnyder H (2018) Determination of leaf respiration in the light: comparison between an isotopic disequilibrium method and the Laisk method. New Phytologist **218**: 1371–1382

Guo W-J, Nagy R, Chen H-Y, Pfrunder S, Yu Y-C, Santelia D, Frommer WB, Martinoia E (2014) SWEET17, a Facilitative Transporter, Mediates Fructose Transport across the Tonoplast of Arabidopsis Roots and Leaves. Plant Physiology **164**: 777–789

Hanson AD, Gage DA, Shachar-Hill Y (2000) Plant one-carbon metabolism and its engineering. Trends Plant Sci **5**: 206–213

Hastie T, Tibshirani R, Friedman J (2017) Model Assessment and Selection. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, 2nd ed. Springer, New York, NY, pp 219–260

Hasunuma T, Harada K, Miyazawa S-I, Kondo A, Fukusaki E, Miyake C (2010) Metabolic turnover analysis by a combination of in vivo 13C-labelling from 13CO2 and metabolic profiling with CE-MS/MS reveals rate-limiting steps of the C3 photosynthetic pathway in Nicotiana tabacum leaves. J Exp Bot **61**: 1041–1051

Hendry JI, Prasannan C, Ma F, Möllers KB, Jaiswal D, Digmurti M, Allen DK, Frigaard N-U, Dasgupta S, Wangikar PP (2017) Rerouting of carbon flux in a glycogen mutant of cyanobacteria assessed via isotopically non-stationary (13) C metabolic flux analysis. Biotechnol Bioeng **114**: 2298–2308

Hilgers EJA, Schöttler MA, Mettler-Altmann T, Krueger S, Dörmann P, Eicks M, Flügge U-I, Häusler RE (2018a) The Combined Loss of Triose Phosphate and Xylulose 5-Phosphate/Phosphate Translocators Leads to Severe Growth Retardation and Impaired Photosynthesis in Arabidopsis thaliana tpt/xpt Double Mutants. Front Plant Sci **9**: 1331

Hilgers EJA, Staehr P, Flügge U-I, Häusler RE (2018b) The Xylulose 5-Phosphate/Phosphate Translocator Supports Triose Phosphate, but Not Phosphoenolpyruvate Transport Across the Inner Envelope Membrane of Plastids in Arabidopsis thaliana Mutant Plants. Front Plant Sci **9**: 1461

Holm S (1978) Board of the Foundation of the Scandinavian Journal of Statistics A Simple Sequentially Rejective Multiple Test Procedure Author (s): Sture Holm Published by : Wiley on behalf of Board of the Foundation of the Scandinavian Journal of Statistics Stable U. Scandinavian Journal of Statistics **6**: 65–70

Loreto F, Velikova V, Di Marco G (2001) Respiration in the light measured by 12CO 2 emission in 13CO 2 atmosphere in maize leaves. Functional Plant Biol **28**: 1103–1108

Ma F, Jazmin LJ, Young JD, Allen DK (2014) Isotopically nonstationary 13C flux analysis of changes in Arabidopsis thaliana leaf metabolism due to high light acclimation. Proceedings of the National Academy of Sciences of the United States of America **111**: 16967–16972

Ma F, Jazmin LJ, Young JD, Allen DK (2017) Isotopically Nonstationary Metabolic Flux Analysis (INST-MFA) of Photosynthesis and Photorespiration in Plants. *In* AR Fernie, H Bauwe, APM Weber, eds, Photorespiration: Methods and Protocols. Springer New York, New York, NY, pp 167–194

Mahon JD, Fock H, Canvin DT (1974) Changes in specific radioactivity of sunflower leaf metabolites during photosynthesis in 14CO2 and 12CO2 at three concentrations of CO2. Planta **120**: 245–254

Makowka A, Nichelmann L, Schulze D, Spengler K, Wittmann C, Forchhammer K, Gutekunst K (2020) Glycolytic Shunts Replenish the Calvin–Benson–Bassham Cycle as Anaplerotic Reactions in Cyanobacteria. Molecular Plant **13**: 471–482

McVetty PBE, Canvin DT (1981) Inhibition of photosynthesis by low oxygen concentrations. Can J Bot **59**: 721–725

Nägele T, Henkel S, Hörmiller I, Sauter T, Sawodny O, Ederer M, Heyer AG (2010) Mathematical Modeling of the Central Carbohydrate Metabolism in Arabidopsis Reveals a Substantial Regulatory Influence of Vacuolar Invertase on Whole Plant Carbon Metabolism. Plant Physiology **153**: 260–272

Patrick JW, Botha FC, Birch RG (2013) Metabolic engineering of sugars and simple sugar derivatives in plants. Plant Biotechnol J **11**: 142–156

Payyavula RS, Tay KHC, Tsai C-J, Harding SA (2011) The sucrose transporter family in Populus: the importance of a tonoplast PtaSUT4 to biomass and carbon partitioning. The Plant Journal **65**: 757–770

Preiser AL, Banerjee A, Weise SE, Renna L, Brandizzi F, Sharkey TD (2020) Phosphoglucoisomerase Is an Important Regulatory Enzyme in Partitioning Carbon out

of the Calvin-Benson Cycle. Frontiers in Plant Science. doi: 10.3389/fpls.2020.580726

Schneider S, Hulpke S, Schulz A, Yaron I, Höll J, Imlau A, Schmitt B, Batz S, Wolf S, Hedrich R, et al (2012) Vacuoles release sucrose via tonoplast-localised SUC4-type transporters. Plant Biol (Stuttg) **14**: 325–336

Schwarz G (1978) Estimating the Dimension of a Model. The Annals of Statistics **6**: 461–464

Sharkey T, Preiser AL, Weraduwage SM, Gog L (2020) Source of 12C in Calvin Benson cycle intermediates and isoprene emitted from plant leaves fed with 13CO2. Biochemical Journal **477**: 3237–3252

Sharkey TD (2019) Discovery of the canonical Calvin–Benson cycle. Photosynthesis Research **140**: 235–252

Sharkey TD (1988) Estimating the rate of photorespiration in leaves. Physiologia Plantarum **73**: 147–152

Sharkey TD, Berry JA, Raschke K (1985) Starch and Sucrose Synthesis in Phaseolus vulgaris as Affected by Light, CO2, and Abscisic Acid 1. Plant Physiology **77**: 617–620

Sharkey TD, Weise SE (2016) The glucose 6-phosphate shunt around the Calvin-Benson cycle. J Exp Bot **67**: 4067–4077

Szecowka M, Heise R, Tohge T, Nunes-Nesi A, Vosloh D, Huege J, Feil R, Lunn J, Nikoloski Z, Stitt M, et al (2013) Metabolic fluxes in an illuminated Arabidopsis rosette. Plant Cell **25**: 694–714

Tcherkez G, Gauthier P, Buckley TN, Busch FA, Barbour MM, Bruhn D, Heskel MA, Gong XY, Crous KY, Griffin K, et al (2017) Leaf day respiration: low CO2 flux but high significance for metabolism and carbon balance. New Phytologist **216**: 986–1001

Uys L, Botha FC, Hofmeyr J-HS, Rohwer JM (2007) Kinetic model of sucrose accumulation in maturing sugarcane culm tissue. Phytochemistry **68**: 2375–2392

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods **17**: 261–272

Wieloch T (2021) The next phase in the development of 13C isotopically non-stationary metabolic flux analysis. Journal of Experimental Botany **72**: 6087–6090

Xu Y, Fu X, Sharkey TD, Shachar-Hill Y, Walker and BJ (2021) The metabolic origins of non-photorespiratory CO2 release during photosynthesis: a metabolic flux analysis. Plant Physiology 1–18

Young JD (2014) INCA: A computational platform for isotopically non-stationary metabolic flux analysis. Bioinformatics **30**: 1333–1335

APPENDIX A: Supplemental Material for Chapter 2

SUPPLEMENTAL TEXT

T1. Derivation of Polyexponential Models from Analytical Solutions of Compartmental Models

In **Figure S2.1**, we show a simplified model of photosynthetic carbon assimilation with three compartments and rates V1-V5. Assuming first-order or pseudo-first-order kinetics:

$$V1 = k_1 * [X] \quad (E1)$$

$$V2 = k_2 * [Y] \quad (E2)$$

$$V3 = k_3 * [Y] \quad (E3)$$

$$V4 = k_4 * [Z] \quad (E4)$$

$$V5 = k_5 * [Z] \quad (E5)$$

We define the differential operator D, where [F] is a stand-in for any compartment's concentration:

$$\frac{d^n[F]}{dt} = D^n[F] \quad (E6)$$

Given definitions (E1-E5) and notation from E6, the rates of change of compartments X, Y, and Z are:

$$D[X] = -k_1[X] + k_2[Y] \quad (E7)$$

$$D[Y] = k_1[X] - k_2[Y] - k_3[Y] + k_4[Z] \quad (E8)$$

$$D[Z] = k_3[Y] - k_4[Z] - k_5[Z] \quad (E9)$$

Through a series of substitutions, it can be shown that this system of differential equations simplifies to a linear homogenous differential equation of the 3rd order:

$$D^3 + aD^2[X] + bD[X] + c[X] = 0 \quad (E10)$$

Where the coefficients *a*, *b*, and *c* are combinations of rate constants such that:

$$a = k_1 + k_2 + k_3 + k_4 + k_5 \quad (E11)$$

$$b = (k_1 + k_2)(k_3 + k_4 + k_5) \quad (E12)$$

$$c = k_1 k_3 k_5 \quad (E13)$$

Which are all constants. The general solution to a linear homogenous differential equation with constant coefficients is of the form:

$$[X](t) = e^{m*t} \quad (E14)$$

Where *m* is some constant. From E10 and E14, we get the characteristic polynomial:

$$r^3 + ar^2 + br + c = 0 \quad (E15)$$

This cubic polynomial has three roots, including repeating and complex roots. Due to the linearity of the system, its general solution is a linear combination of its roots, such that:

$$[X](t) = c_1 e^{r_1 t} + c_2 e^{r_2 t} + c_3 e^{r_3 t} \quad (E16)$$

Solving this cubic polynomial for biochemically reasonable estimates of k_1 through k_5 results in three real and negatively valued roots, making the general result of an identical form as the triexponential decay models we fit our data to in this study. The analytical solutions to the differential equations or systems of differential equations describing single and two-compartment models, likewise, correspond to single exponential and biexponential functions, respectively.

T2. Supplemental Methods

Nonlinear regression and bootstrapping

Fitting of $\%^{12}C$ remaining data to polyexponential models was performed in Python using the *curve_fit()* function implemented in the *SciPy* package (Virtanen et al., 2020). We performed all regressions 100 times with uniformly sampled initial parameter values and selected the fit with the lowest SSR for further analysis. N = 1000 bootstrap resampling with replacement was performed using functions from the Python package *recombinator*. Due to the time-course structure of the data, circular block bootstrapping was used to preserve some of the dependence structure between subsequent measurements (Politis and Romano, 1991). Bootstrap samples were fitted using the same general procedure as that used to generate the best-fit lines, with the exception that the initial guesses for the parameter values for the regression of the bootstrap samples were set to the best-fit parameter values. 95% confidence intervals for each parameter were derived by taking the 2.5^{th} and 97.5^{th} percentile values of the resulting distributions of all successful fits.

Data treatment for heteroskedastic residuals and outlier identification

Heteroskedastic residuals from our nonlinear regressions were corrected using a logit transformation (Johnson, 1949). Specifically, we performed nonlinear regression on models of the form:

$$logit\left(\frac{f(t)}{100}\right) = logit\left(\frac{Ae^{b*t} + \cdots}{100}\right) \quad (E17)$$

This preserves the relationship between our response, independent variables, and estimated parameters, allowing for straightforward interpretation while substantially reining in the heteroskedasticity of the residuals. Due to the presence of % ^{12}C remaining values very close to 100% in the tobacco dataset, a constant value of 0.1 was subtracted to avoid inflated values in the first time point exerting too much influence over the nonlinear regression results (and therefore resulting in heteroskedastic residuals). Specifically, the following model was fitted for the tobacco datasets:

$$logit\left(\frac{f(t) - 0.1}{100}\right) = logit\left(\frac{(Ae^{b*t} + \cdots) - 0.1}{100}\right) \quad (E18)$$

Studentized residuals were calculated for all model fits and datapoints whose studentized residuals exceeded an absolute value of 3 were excluded (**N = 5).** Due to the substantial impure heteroskedasticity in the Model 1 fits studentized residuals greater than 3 in Model 1 fits were ignored for the purposes of outlier removal.

Model selection criteria

Extra-sum-of-squares: For each nested pair of models we calculated the probability that, given the null hypothesis that the simpler of the two models is true, we would see the observed improvement in model fit as measured by the sum-of-squared residuals (SSR) (Draper and Smith, 1998). We calculate an F statistic as follows:

$$F = \frac{\frac{SSR_{simple} - SSR_{complex}}{SSR_{complex}}}{\frac{DF_{simple} - DF_{complex}}{DF_{complex}}} \quad (E19)$$

Where SSR_{simple} and $SSR_{complex}$ are the SSR values for the simpler and complex – i.e., fewer and more parameters – models, respectively, and DF_{simple} and $DF_{complex}$ are the degrees of freedom for the two models. The F statistic resulting from E19 was then compared to the F-distribution to derive a p-value representing the probability of observing this F statistic given our null hypothesis, which is that our simpler model is correct. For this study, we set $\alpha = 0.05$ and used the Holm-Bonferroni correction (Holm, 1978) to adjust our p-value cutoff to one that

corresponds to a family-wise α of 0.05. For each p-value *Pk* in the family of hypothesis tests being tested, we evaluate the following expression:

$$P_k < \frac{\alpha}{m + 1 - k} \qquad (E20)$$

where α is the family-wise α we are adjusting to, m is the number of hypothesis tests being conducted, and k is the rank of the p-value Pk in a ranked list of increasing p-values.

We selected the best-supported model for a given dataset by starting with the single exponential model and adding more parameters until we got to a model comparison that did not meet our adjusted p-value cutoff, in which case we went with the simpler model in the comparison. In cases where there was a comparison of two more complex models than the one we arrived at using the method just described that yielded a low p-value, we calculated p-value for the F-statistic comparison between the more complex of those two and the accepted model. If we were justified in rejecting the null hypothesis that the simpler model is better in this case, we went with the more complex model.

Cross-validation: For this study we used the *cross_validate*() function from the *SciKitLearn* package to perform between 5 and 10 iterations of 5-fold cross-validation on our datasets (Pedregosa et al., 2011; Hastie et al., 2017). The same non-linear ordinary least squares fitting procedure used for our best-fit parameter estimation on the full datasets was used for our cross-validation, with the only difference being that the fitting was done 5 times with different randomly selected bins of data for training and testing, resulting in 5 estimates of prediction error for each alternative model at each iteration. After 5-10 iterations, we took all the negative mean squared error estimates for each model for a given metabolite or aggregated metabolite dataset and then calculated their mean value and 95% confidence interval (± 1.96 SE). The model with the lowest average error and whose 95% CI does not overlap with the next simplest model in terms of the number of fitted parameters was chosen as the best-performing model for each dataset.

AIC/BIC: For each best-fit of Models 1-7, the AIC (Akaike, 1998) and BIC (Schwarz, 1978) were calculated as follows:

$$AIC = 2k + n \ln SSR \qquad (E21)$$

$$BIC = k \ln n + n \ln SSR \quad (E22)$$

where *k* is the number of estimated parameters in the model, and *n* is the sample size. The best-supported model for each dataset was chosen by identifying the model with the lowest AIC/BIC value that is not within two absolute units of a simpler (i.e., fewer parameters) model.

T3. Calculation of v_o/v_c

We begin with the equation from Farquhar et al., (1980)

$$A = v_C - 0.5v_O - R_L \quad (E23)$$

where *A* is the net rate of CO_2 assimilation (uptake), v_C is the velocity of carboxylation, v_O is the velocity of oxygenation, and R_L is all other sources of CO_2 release in the light, possibly primarily CO_2 released by the glucose 6-phosphate shunt (Xu et al., 2021a). Next, we define

$$\Phi = \frac{v_O}{v_C} \quad (E24)$$

and so

$$A = v_C(1 - 0.5\Phi) - R_L \quad (E25)$$

Rearranging

$$v_C = \frac{(A+R_L)}{(1-0.5\Phi)} \quad (E26)$$

We can also estimate v_O.

$$A = v_O\left(\frac{1}{\Phi} - 0.5\right) - R_L \quad (E27)$$

and so

$$v_O = \frac{(A+R_L)}{\left(\frac{1}{\Phi}-0.5\right)} \quad (E28)$$

Taking the ratio of equations and canceling (A+R_L)

$$\frac{v_O}{v_C} = \frac{(1-0.5\Phi)}{\left(\frac{1}{\Phi}-0.5\right)} \quad (E29)$$

We can expand Φ as in Farquhar et al., (1980)

$$\Phi = \frac{2\Gamma_*}{C} \quad (E30)$$

where Γ_* is the CO_2 compensation point in the absence of R_L. Therefore,

$$\frac{v_O}{v_C} = \frac{\left(1 - {}^{\Gamma_*}/_{C}\right)}{0.5\left(\frac{C}{\Gamma_*} - 1\right)} \quad (E31)$$

Where C is the CO_2 partial pressure equivalent at the sites of carboxylation. This is determined by

$$C = C_i - \frac{A}{g_m} \quad (E32)$$

where C_i is the partial pressure of CO_2 in the intercellular air spaces of the leaf (estimated from gas exchange) and g_m is the mesophyll conductance for CO_2 diffusion. In the absence of a direct measurement g_m can be estimated as

$$g_m = 0.3 + 0.11 \cdot A \quad (E33)$$

Based on multiple measurements reported in Caemmerer and Evans, (1991)

We can parameterize as follows based on measured gas exchange of the leaves used for this data set

A = 17.4 ± 1.9 μmol m^{-2} s^{-1} (avg ± SD) (measured)

Γ^*= 3.18 μmol m^{-2} s^{-1} Pa^{-1} (for tobacco, from Sharkey, (2016), adjusted to 22°C)

C = 20.5 Pa (measured C_i and corrected for g_m using *E32*)

$$\frac{v_o}{v_c} = \frac{\left(1 - \frac{3.18}{20.5}\right)}{0.5\left(\frac{20.5}{3.18} - 1\right)} = 0.31 \quad (E34)$$

T4. Plant Growth, Gas Exchange, and $^{13}CO_2$ Labeling.

Wild-type *Camelina sativa* ecotype Suneson was grown under 8/16-h day/night cycles, under a light intensity of 500 μmol m^{-2} s^{-1}, temperature of 22°C, and 50% relative humidity for 4 weeks. The youngest fully expanded leaves were used for gas exchange and labeling experiments. a LI-COR 6800 portable photosynthesis system (LI-COR Biosciences, Lincoln, NE, USA) was used to measure carbon assimilation. The reference [CO_2] was set to 400 ppm, light intensity was 500 μmol m^{-2} s^{-1}, temperature was 22°C, and relative humidity was 70% to ensure that the leaf vapor pressure deficit was ~0.85 kPa. After 10-15 min acclimation, net CO_2 assimilation rate was logged and then the CO_2 source was switched to $^{13}CO_2$ with all other parameters held constant. Gases were mixed with mass flow controllers (Alicat Scientific, Tucson AZ, USA) controlled by a custom-programmed Raspberry Pi touchscreen monitor (Raspberry Pi foundation, code available upon request). Labeled leaf samples were collected at time points of 0, 0.5, 1, 2, 2.5, 3, 5, 7, 10, 15, 30, 60, 90, and 120 min. Liquid nitrogen was directly sprayed on the leaf surface via

a customized fast quenching (0.1-0.5 s to <0°C) labeling system (13). Leaf temperature fell below 0°C between. The frozen leaf sample was stored at -80°C. There were three biological replicates for data points from 0-90 min, and two biological replicates at 120 min.

T5. Analysis of Mass Spectrometry Data.

Data from LC-MS/MS were acquired with MassLynx 4.0 (Agilent, Santa Clara, CA, USA). Data from GC-EI-MS was acquired with Agilent GC/MSD Chemstation (Agilent, Santa Clara, CA, USA). Data from GC-CI-MS was acquired with Agilent MassHunter Workstation (Agilent, Santa Clara, CA, USA). Metabolites were identified by retention time and mass to charge ratio (m/z), in comparison with authentic standards. Both LC-MS and GC-MS data were converted to MassLynx format and processed with QuanLynx software for peak detection and quantification. Parameters for transitions of measured metabolites in multiple reaction monitoring (MRM) with LC-MS/MS and selected ion monitoring (SIM) with GC-MS are shown in Appendix A, Dataset S5. Experimentally measured mass isotopomer distributions of measured metabolites are shown in Appendix A, Dataset S1.

Isotopologue Network and Flux Determination.

The metabolic network model with all reactions and their respective carbon atom transitions describing photosynthetic central metabolism in *Camelina sativa* was constructed based upon the previous studies (Ma et al., 2014; Xu et al., 2021a) and KEGG database. A list of the reactions and abbreviations are provided in **Table S2.2**. INST-MFA was performed to estimate metabolic fluxes using the Isotopomer Network Compartmental Analysis software package (INCA2.0, http://mfa.vueinnovations.com, Vanderbilt University) (Young, 2014) implemented in MATLAB 2018b. The fit for all the tested models were accepted based on $\chi 2$ test of the sum-of-squared residuals (SSR). Global best fit SSR were calculated by parameter continuation analysis. Fatty acid synthesis rate is constrained to 0.0329-0.4405 μmol CO2 g^{-1}FW h^{-1} by combining the previous measurements of 0.049-0.067 μmol CO_2 m^{-2} s^{-1} with 0.005-0.012 μmol CO_2 m^{-2} s^{-1} (Tcherkez et al., 2005; Xu et al., 2021a). R_L is constrained to 8.1-10.7 μmol CO_2 g^{-1}FW h^{-1} based on previous measurement (Xu et al., 2021a). v_o/v_c is constrained to 0.3-0.32 based on measurement in this study.

Assessment of Flux Precision

Both parameter continuation method and Monte Carlo method were independently estimated the 95% confidence intervals of the estimated flux values as shown in Appendix A, Dataset S4. 10,000 sets of perturbed data were used for Monte Carlo analysis. The resulting distribution of flux values enabled the estimation of confidence intervals. The computation-intensive parameter continuation and Monte Carlo simulations were computed in parallel using a SLURM job scheduler to distribute jobs to hundreds of compute nodes within a high-performance computing cluster provided by the Institute for Cyber-Enabled Research at Michigan State University. The two approaches gave similar results of confidence intervals for each flux solution.

Calculation of predicted percentage of isotopologues (f_{mn})

Predicted percentage of isotopologues **(f_{mn})** is calculated by the equation of:

$$fmn = (p13C)n \ * \ (p12C)m - n \ * \ mCn \qquad (E35)$$

p_{13C} is the measured ^{13}C enrichment; p_{12C} is the measured ^{12}C enrichment; n is the number of ^{13}C carbon; m is the number of total carbons; ${}_mC_n$ is the combination for choosing objects of n from the total number of objects of m.

FIGURES

Figure S2.1: Simplified compartmental model used in "T3.1: Derivation of Polyexponential Models from Analytical Solutions of Compartmental Models" showing the metabolite compartments and rates interconnecting them. Note that we are modeling the depletion of ^{12}C here, not the enrichment of ^{13}C, hence the lack of external input to the CBC under the assumption that we are working with pure $^{13}CO_2$.

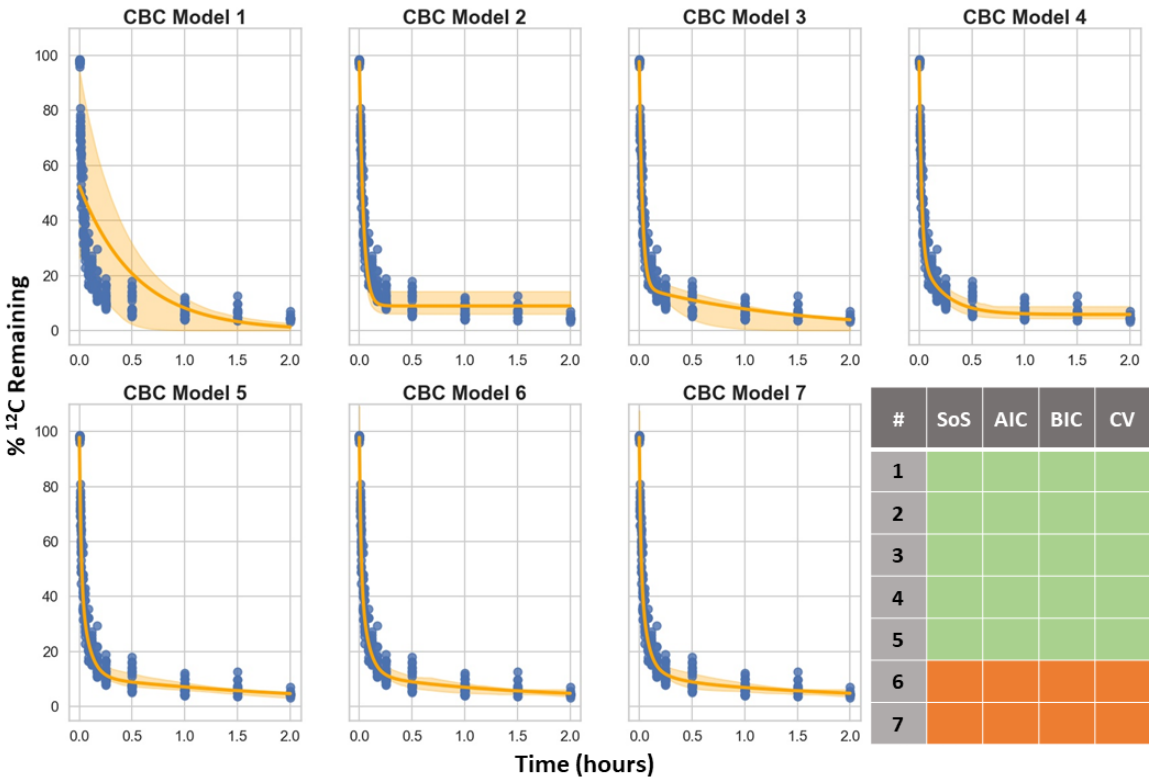


Figure S2.2: Nonlinear regression fits for all polyexponential models fitted to the aggregated Calvin-Benson Cycle intermediate dataset along with a summary of model selection results. The orange line represents the best-fit line and the shaded region represents the 95% CI estimated by bootstrap resampling. In the bottom-right table, green squares represent model selection results supporting the model indicated by that row representing a statistical improvement over a simpler model. Orange squares represent model selection results that do not support adding the additional parameters needed for the model in that row. Figure 2.1 is a subset of these data.

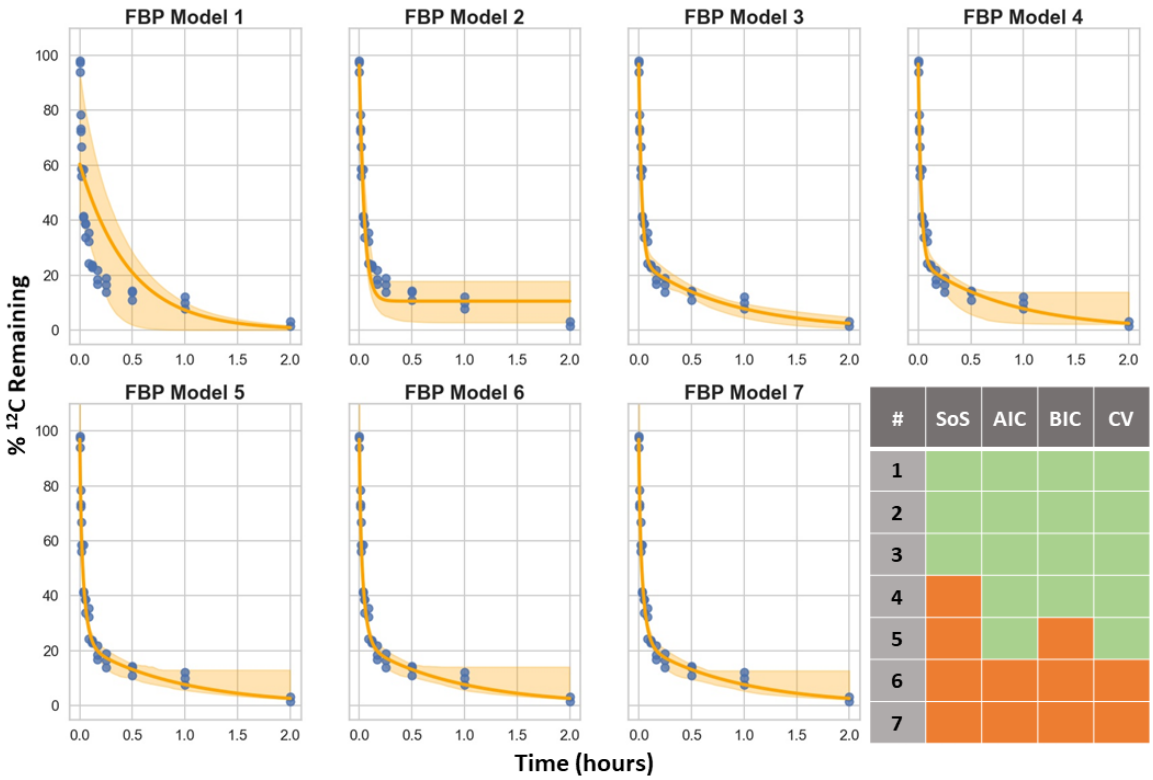


Figure S2.3: Nonlinear regression fits for all polyexponential models fitted to the fructose 1,6-bisphosphate dataset along with a summary of model selection results. The orange line represents the best-fit line and the shaded region represents the 95% CI estimated by bootstrap resampling. In the bottom-right table, green squares represent model selection results supporting the model indicated by that row representing a statistical improvement over a simpler model. Orange squares represent model selection results that do not support adding the additional parameters needed for the model in that row.

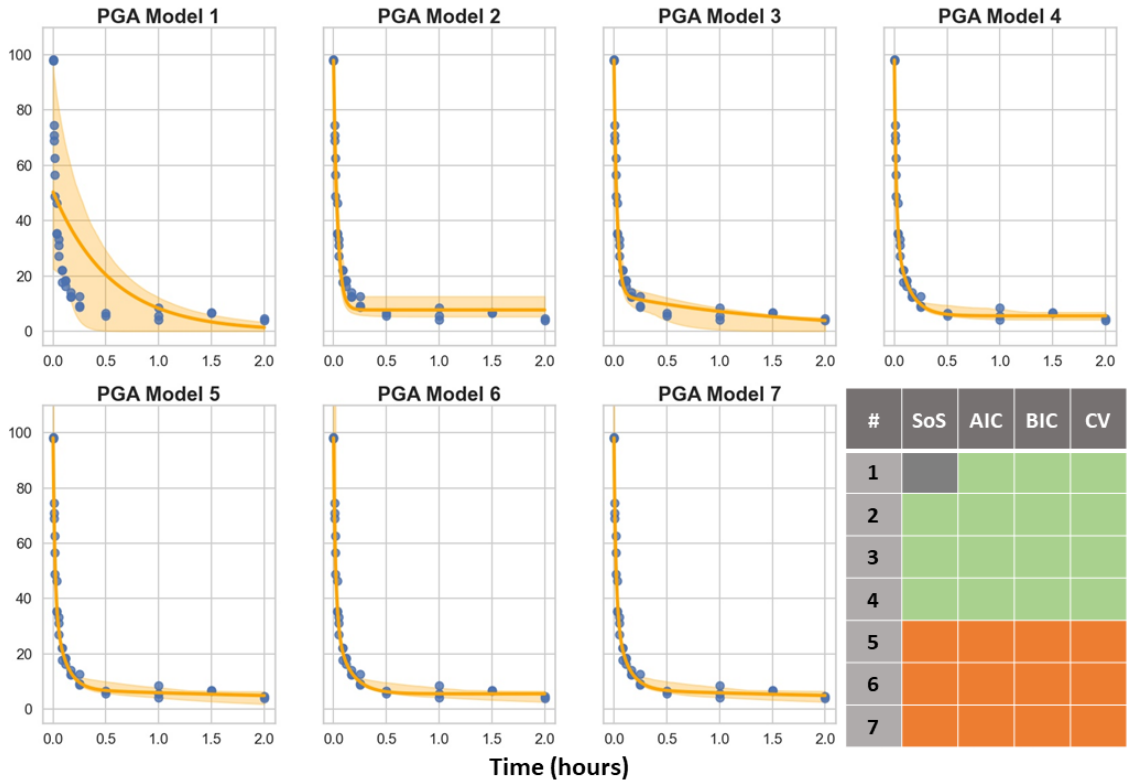


Figure S2.4: Nonlinear regression fits for all polyexponential models fitted to the 3-phosphoglycerate dataset along with a summary of model selection results. The orange line represents the best-fit line and the shaded region represents the 95% CI estimated by bootstrap resampling. In the bottom-right table, green squares represent model selection results supporting the model indicated by that row representing a statistical improvement over a simpler model. Orange squares represent model selection results that do not support adding the additional parameters needed for the model in that row.

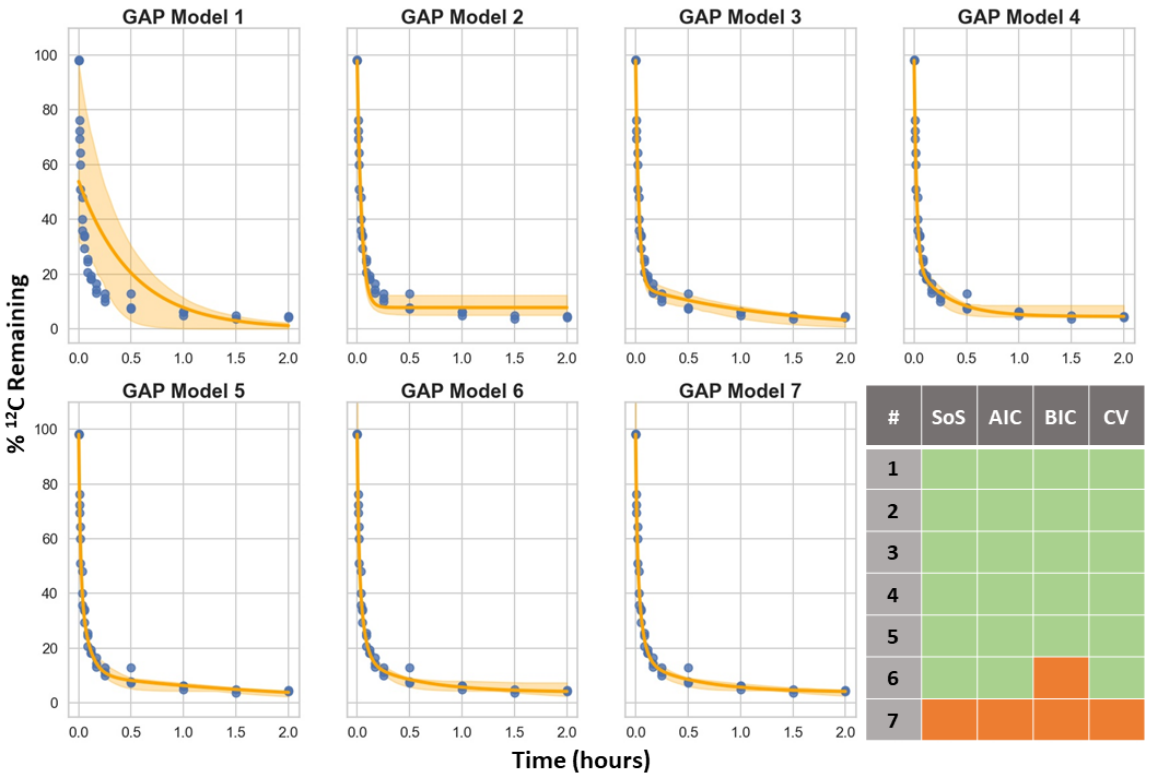


Figure S2.5: Nonlinear regression fits for all polyexponential models fitted to the glyceraldehyde-3-phosphate dataset along with a summary of model selection results. The orange line represents the best-fit line and the shaded region represents the 95% CI estimated by bootstrap resampling. In the bottom-right table, green squares represent model selection results supporting the model indicated by that row representing a statistical improvement over a simpler model. Orange squares represent model selection results that do not support adding the additional parameters needed for the model in that row.

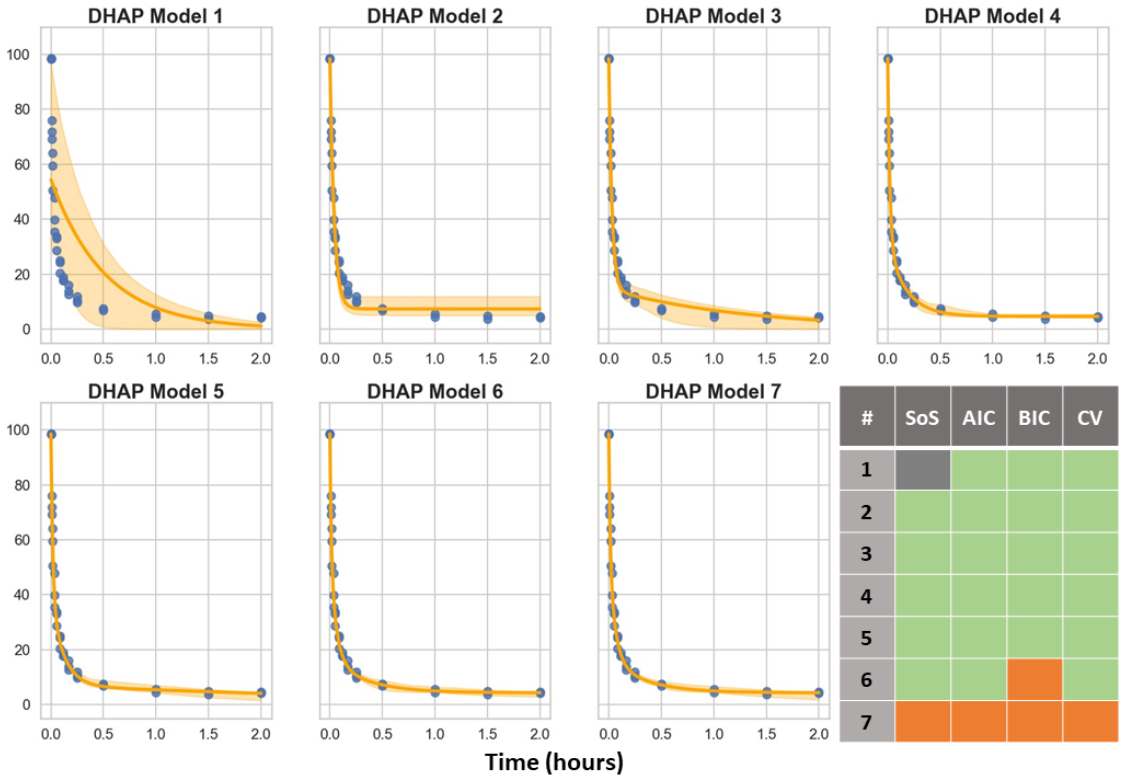


Figure S2.6: Nonlinear regression fits for all polyexponential models fitted to the dihydroxyacetone phosphate along with a summary of model selection results. The orange line represents the best-fit line and the shaded region represents the 95% CI estimated by bootstrap resampling. In the bottom-right table, green squares represent model selection results supporting the model indicated by that row representing a statistical improvement over a simpler model. Orange squares represent model selection results that do not support adding the additional parameters needed for the model in that row.

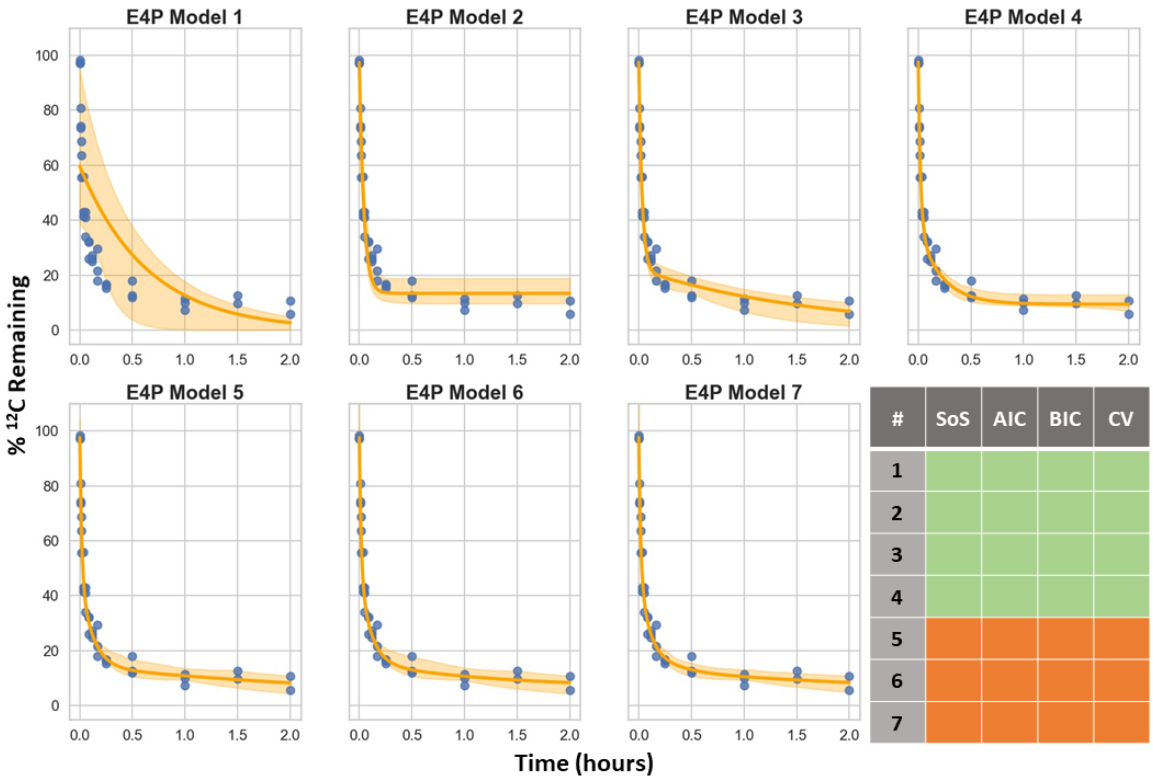


Figure S2.7: Nonlinear regression fits for all polyexponential models fitted to the erythrose-4-phosphate dataset along with a summary of model selection results. The orange line represents the best-fit line and the shaded region represents the 95% CI estimated by bootstrap resampling. In the bottom-right table, green squares represent model selection results supporting the model indicated by that row representing a statistical improvement over a simpler model. Orange squares represent model selection results that do not support adding the additional parameters needed for the model in that row.

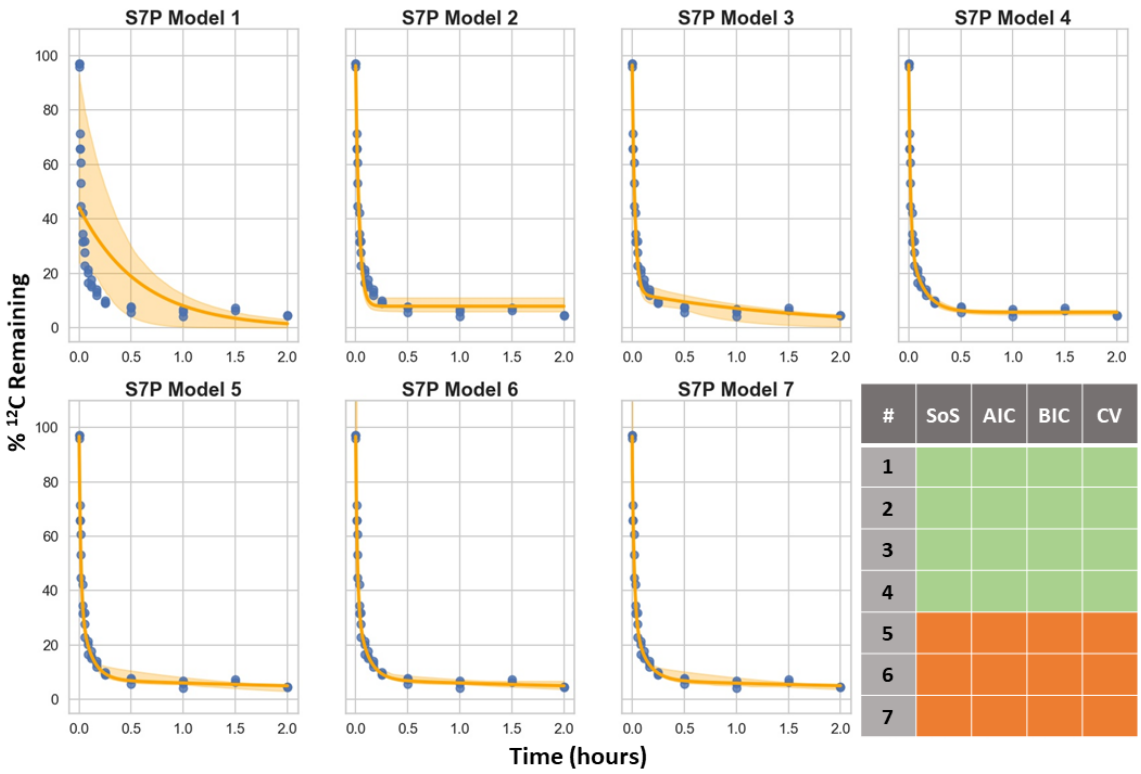


Figure S2.8: Nonlinear regression fits for all polyexponential models fitted to the sedoheptulose-7-phosphate dataset along with a summary of model selection results. The orange line represents the best-fit line and the shaded region represents the 95% CI estimated by bootstrap resampling. In the bottom-right table, green squares represent model selection results supporting the model indicated by that row representing a statistical improvement over a simpler model. Orange squares represent model selection results that do not support adding the additional parameters needed for the model in that row.

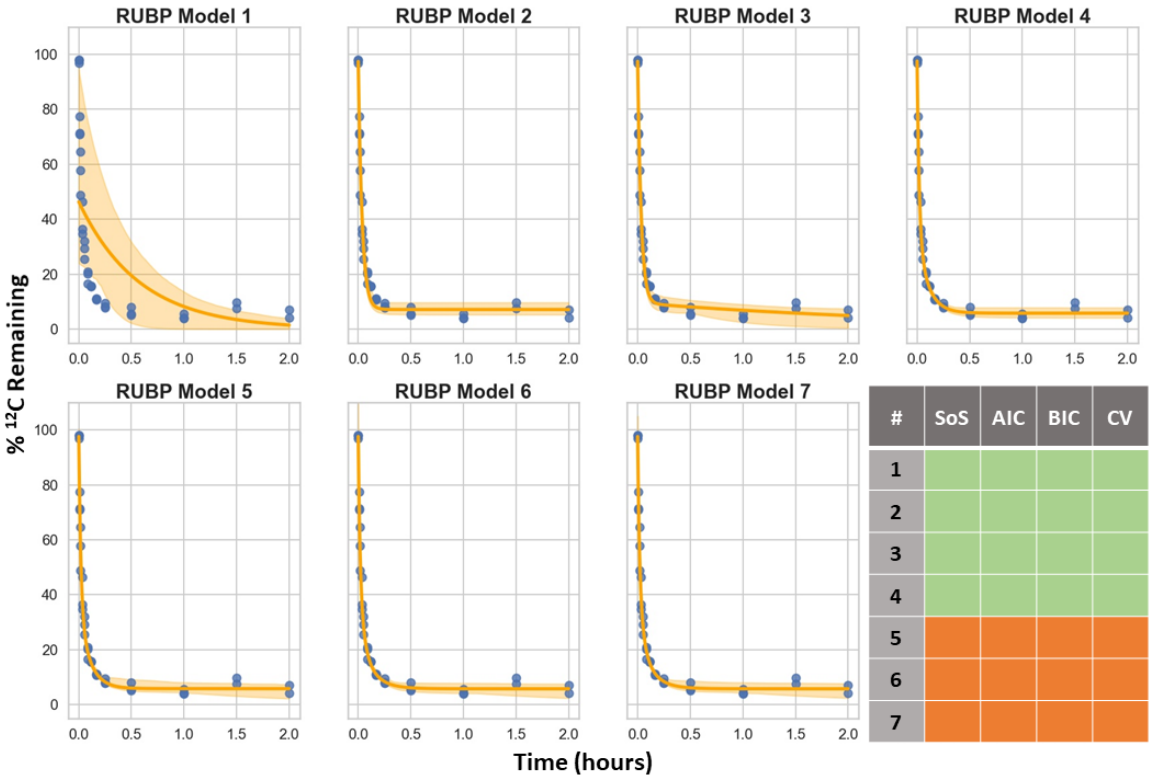


Figure S2.9: Nonlinear regression fits for all polyexponential models fitted to the ribulose 1,5-bisphosphate dataset along with a summary of model selection results. The orange line represents the best-fit line and the shaded region represents the 95% CI estimated by bootstrap resampling. In the bottom-right table, green squares represent model selection results supporting the model indicated by that row representing a statistical improvement over a simpler model. Orange squares represent model selection results that do not support adding the additional parameters needed for the model in that row.

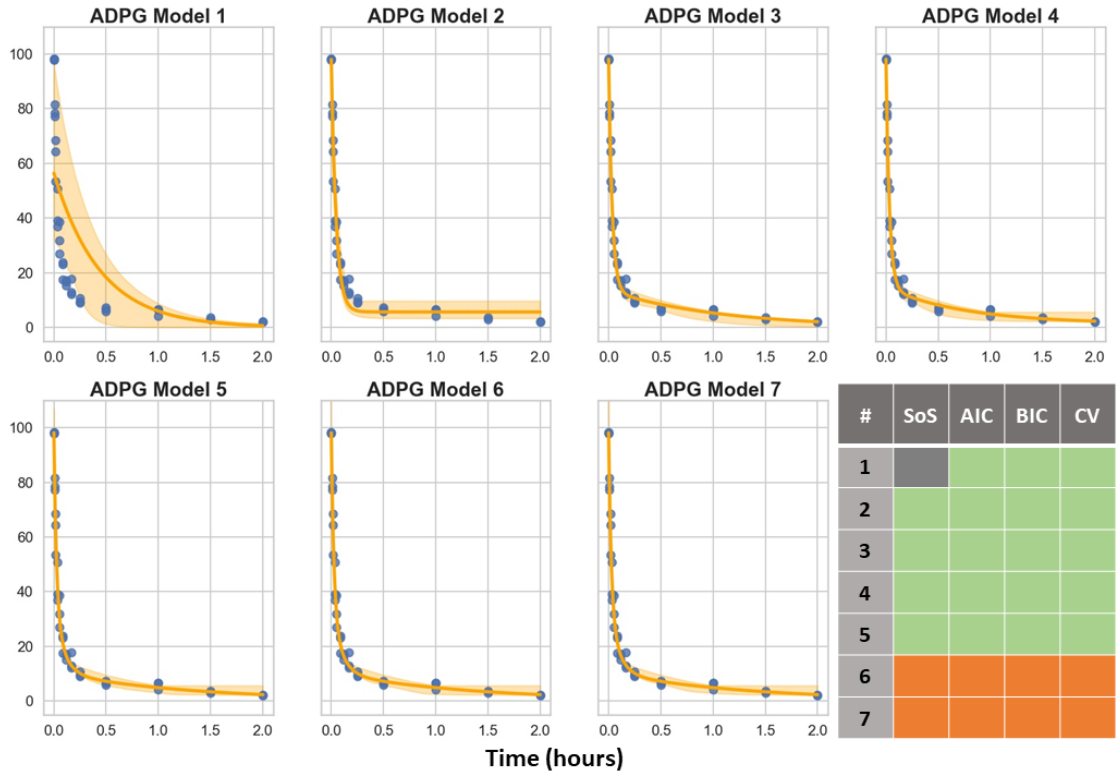


Figure S2.10: Nonlinear regression fits for all polyexponential models fitted to the ADP-glucose dataset along with a summary of model selection results. The orange line represents the best-fit line and the shaded region represents the 95% CI estimated by bootstrap resampling. In the bottom-right table, green squares represent model selection results supporting the model indicated by that row representing a statistical improvement over a simpler model. Orange squares represent model selection results that do not support adding the additional parameters needed for the model in that row.

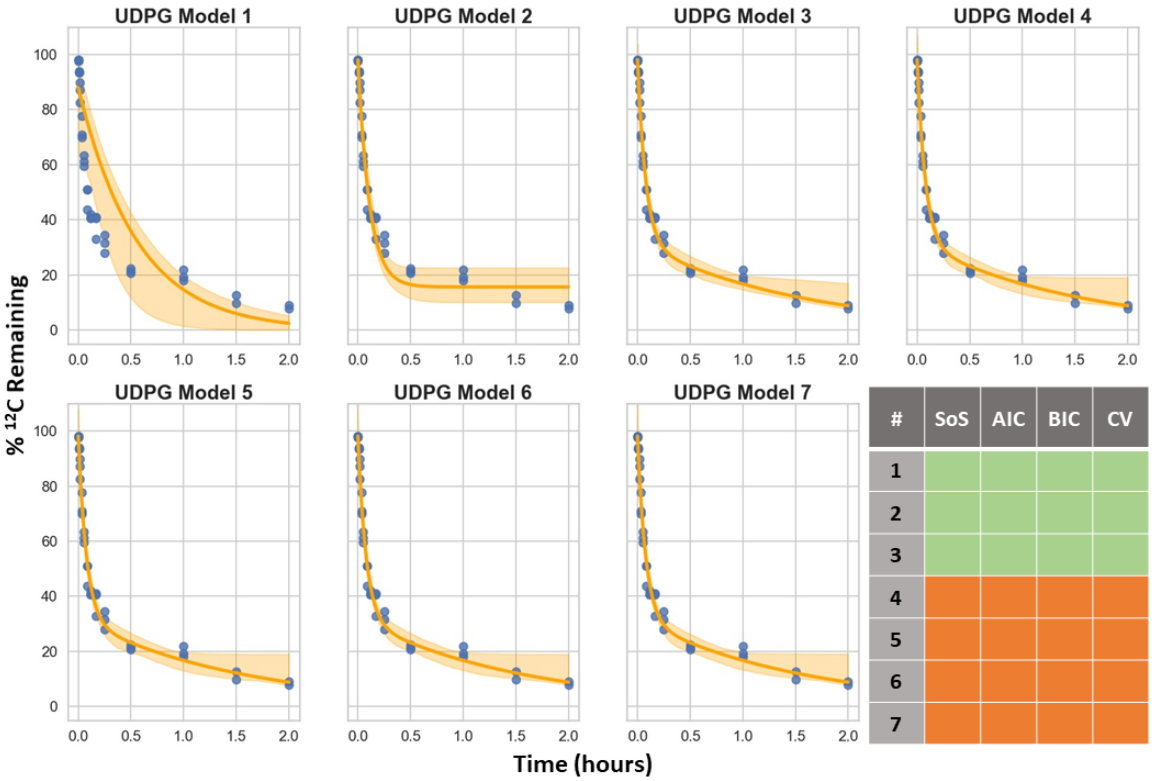


Figure S2.11: Nonlinear regression fits for all polyexponential models fitted to the UDP-glucose dataset along with a summary of model selection results. The orange line represents the best-fit line and the shaded region represents the 95% CI estimated by bootstrap resampling. In the bottom-right table, green squares represent model selection results supporting the model indicated by that row representing a statistical improvement over a simpler model. Orange squares represent model selection results that do not support adding the additional parameters needed for the model in that row.

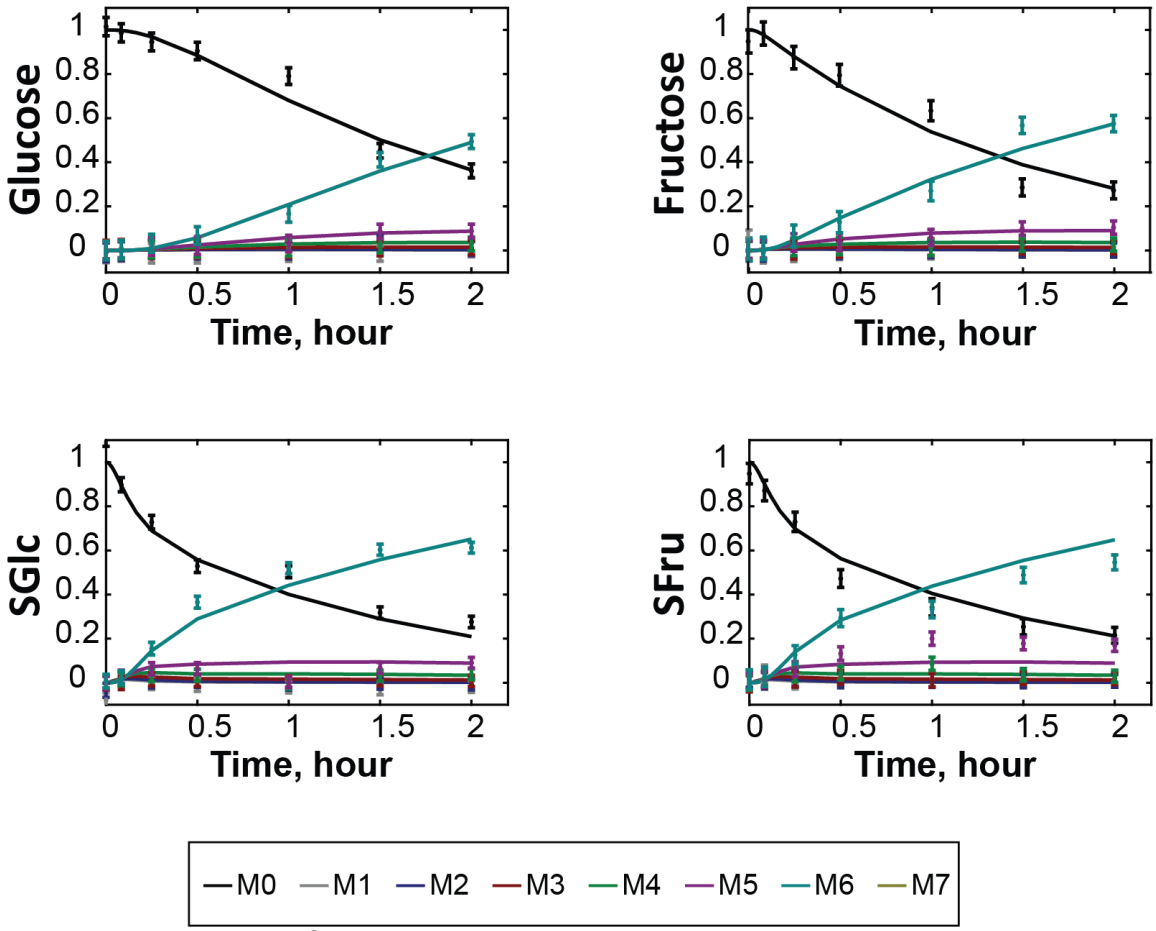


Figure S2.12: Transient $^{13}CO_2$ labeling in glucose, fructose, sucrose glucosyl moiety, and sucrose fructosyl moiety. Experimentally determined isotope labeling measurements are shown as points with error bars (n=3, ± stdev). INST-MFA fitted mass isotopologue distributions are shown as solid lines. Nominal masses of M0 mass isotopologues are shown in parentheses. Error bars represent standard errors.

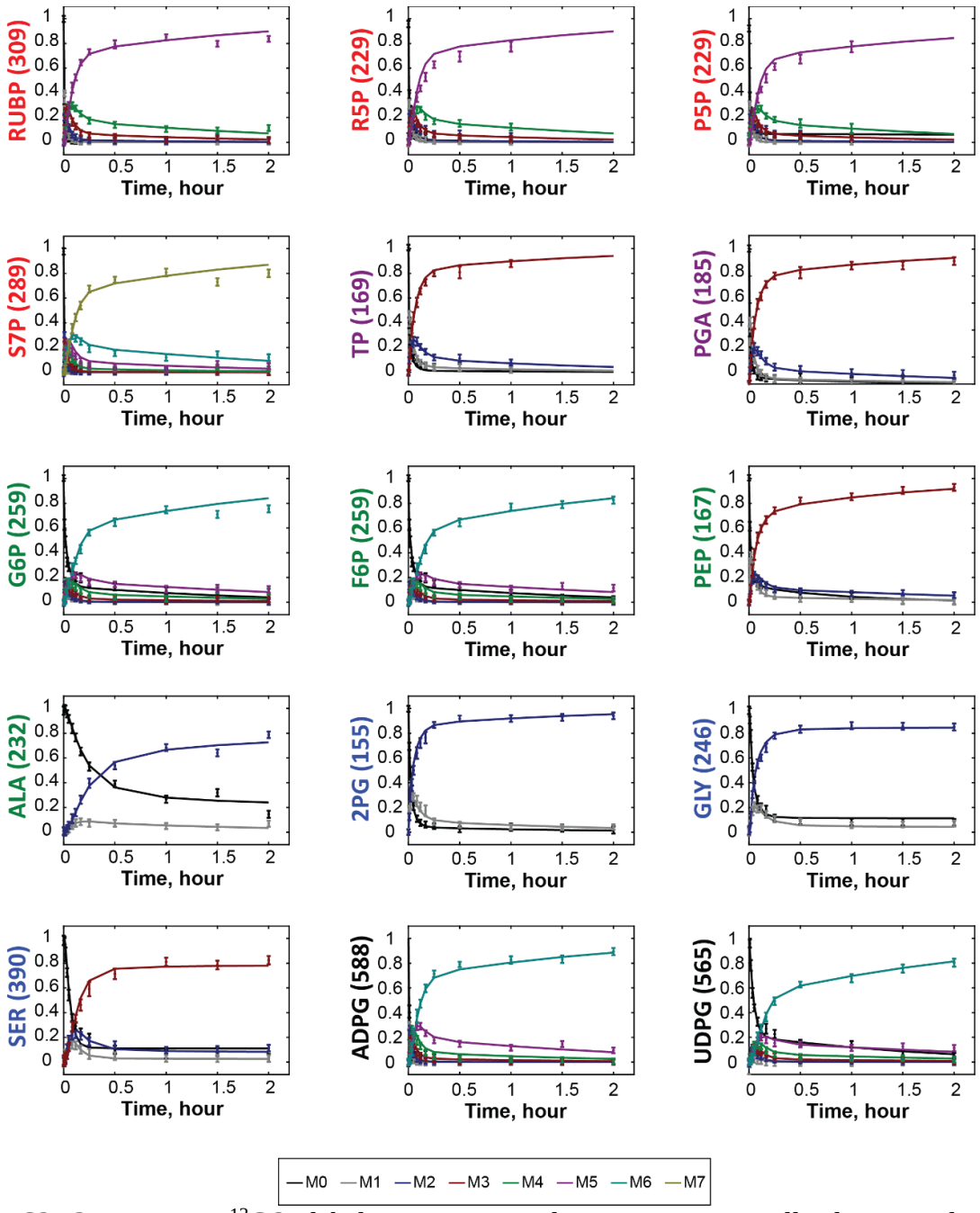


Figure S2.13: Transient $^{13}CO_2$ labeling in measured ions. Experimentally determined isotope labeling measurements are shown as points with error bars (n=3, ± stdev). INST-MFA fitted mass isotopologue distributions are shown as solid lines. Nominal masses of M0 mass isotopologues are shown in parentheses. Error bars represent standard errors. (A) C_3 and Glycolysis related metabolites. Core C_3-only intermediates [labeled in red]; intermediates shared with glycolysis [purple]; core glycolysis metabolites and products [green]; photorespiratory intermediates [blue]; then carbohydrate building substrates [black]. (B) TCA cycle related metabolites. OAA derived AA's [labeled in red]; and more slowly Thr which is made from Asp at a slower rate than Asn [purple]; Citrate [green]; Glu and Gln ions [labeled Glx in blue]; Malate Fumarate and Succinate [black].

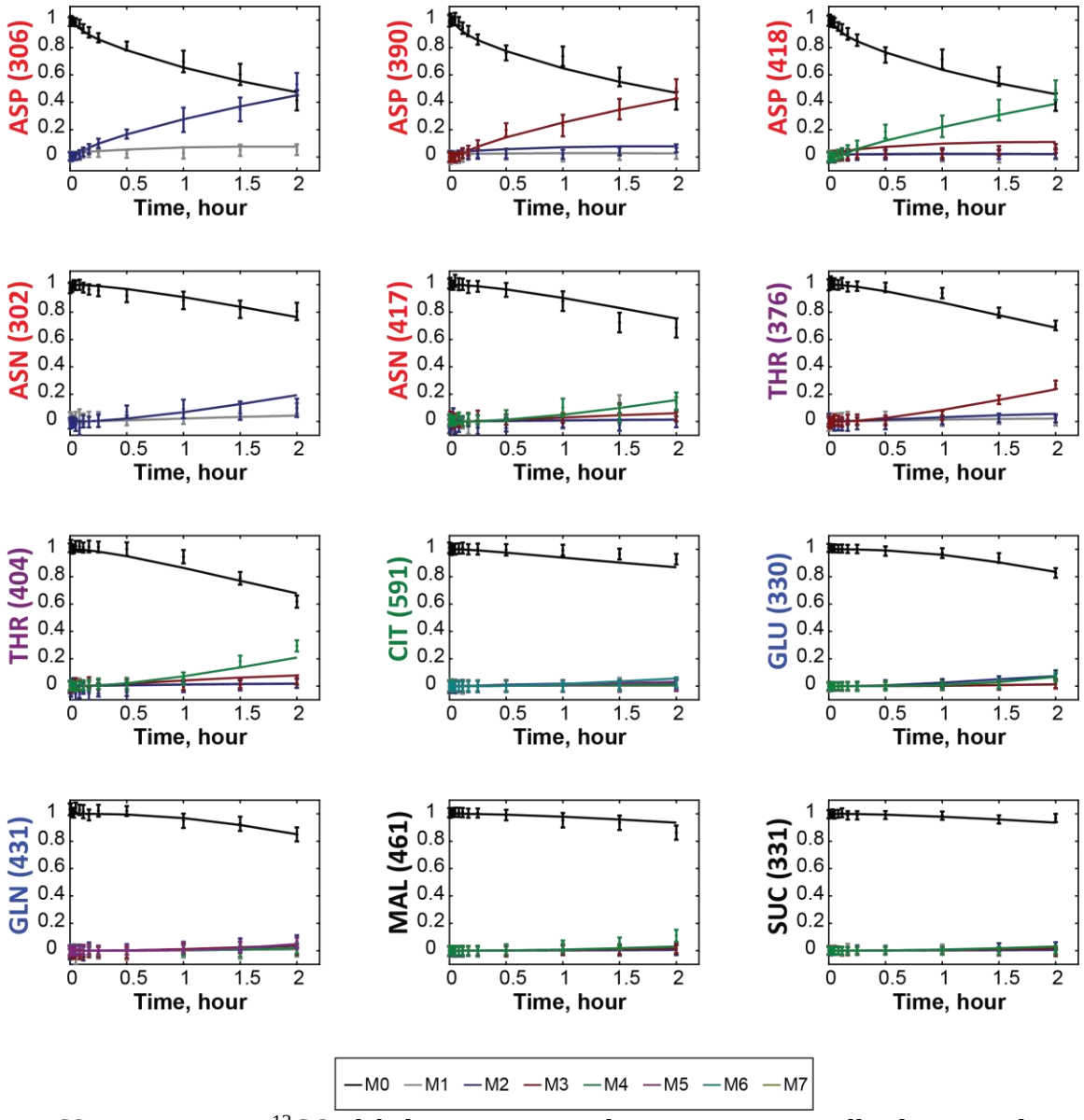


Figure S2.14: Transient $^{13}CO_2$ labeling in measured ions. Experimentally determined isotope labeling measurements are shown as points with error bars (n=3, ± stdev). INST-MFA fitted mass isotopologue distributions are shown as solid lines. Nominal masses of M0 mass isotopologues are shown in parentheses. Error bars represent standard errors. (A) C_3 and Glycolysis related metabolites. Core C_3-only intermediates [labeled in red]; intermediates shared with glycolysis [purple]; core glycolysis metabolites and products [green]; photorespiratory intermediates [blue]; then carbohydrate building substrates [black]. (B) TCA cycle related metabolites. OAA derived AA's [labeled in red]; and more slowly Thr which is made from Asp at a slower rate than Asn [purple]; Citrate [green]; Glu and Gln ions [labeled Glx in blue]; Malate Fumarate and Succinate [black].

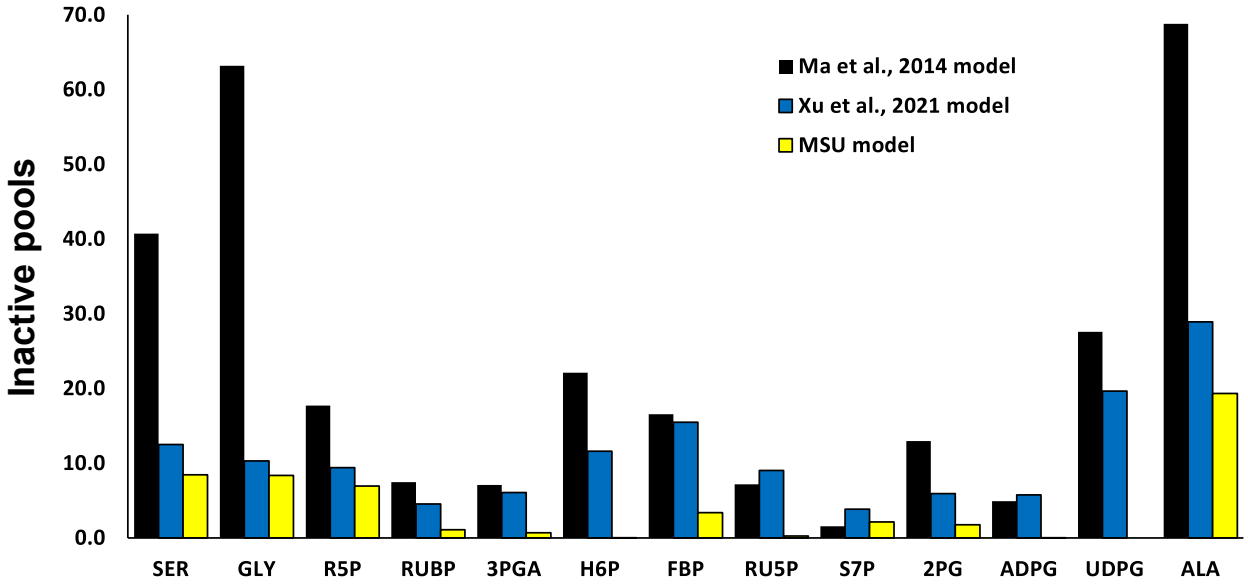


Figure S2.15: The INST-MFA estimated inactive pools for serine, glycine, R5P, RUBP, 3-PGA, H6P, FBP, RU5P, S7P, 2PG, ADPG, UDPG, and alanine were compared with Xu et al., (2021a) and Ma et al., (2014). MSU model lowered the inactive pool sizes for all the above metabolites. Among them, the inactive pools for RUBP, 3-PGA, H6P, RU5P, 2PG, ADPG, UDPG dramatically lowered to almost zero.

TABLES

Table S2.1: Rate parameters for CBC intermediates, ADPG, and UDPG. The top row is data derived from the average of all the individual metabolites and following the data for each metabolite, the time constants for each is averaged (CBC average not included) and standard deviation is shown.

Metabolite(s)	Slopes (min^{-1})		
	Fast	**Middle**	**Slow**
CBC average	-1.071	-0.203	-0.007
PGA	-1.007	-0.161	-0.003
S7P	-1.078	-0.163	-0.003
GAP	-1.050	-0.196	-0.008
DHAP	-0.950	-0.140	-0.005
FBP	-1.690	-0.371	-0.018
RUBP	-0.802	-0.141	0.002
ADPG	-0.665	-0.179	-0.013
Average - CBC	-1.04	-0.194	-0.007
Std Dev	0.30	0.075	0.006

Table S2.2: Abbreviations for metabolites and reactions.

Abbreviations	Full name
2PG	2-phosphoglycolate
3PGA	3-phosphoglycerate
ACA	acetyl-CoA
acetyl-CoA	acetyl-coenzyme A
ADPG	adenosine diphosphate glucose
AGP	ADP-glucose phosphorylase
AKG	α-ketoglutarate
ALA	alanine
ALD	aldolase
ALT	alanine transaminase
AS	asparagine synthase
ASN	asparagine
ASP	aspartate
ASPT	aspartate transaminase
C_3 **cycle**	Calvin–Benson–Bassham cycle
CIT	citrate
CO_2	carbon dioxide
CS	citrate synthase
DOF	degrees of freedom
E4P	erythrose-4-phosphate
EC2	transketolase-bound-2-carbon-fragment
ESI	electrospray ionization
F6P	fructose-6-phosphate
FBA	fructose-bisphosphate aldolase
FBP	fructose-1,6-bisphosphatase
Fru	fructose
FUM	fumarate
FVCB	Farquhar, von Caemmerer and Berry
G1P	glucose-1-phosphate
G6P	glucose-6-phosphate
G6PDH	glucose-6-phosphate dehydrogenase
GA	glycerate
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
GC-MS	gas chromatography-mass spectrometry
GDC	glycine decarboxylase
GK	glycerate kinase
Glc	glucose
GLN	glutamine
GLY	glycine
GPU	UDP-glucose pyrophosphorylase
GS	glutamine synthetase
ICI	isocitrate
IDH	isocitrate dehydrogenase
INST-MFA	isotopically nonstationary metabolic flux analysis
LC-MS/MS	liquid chromatography-tandem mass spectrometry
MAL	malate
M1P	mannose 1-phosphate
MDH	malate dehydrogenase
ME	malic enzyme
MFA	metabolic flux analysis

Table S2.2 (cont'd)

MID	mass isotopologue distribution
MRM	multiple reaction monitoring
netA	net CO_2 assimilation
OAA	oxaloacetate
OPP	oxidative pentose phosphate
PCR	pyrroline-5-carboxylate reductase
PDH	pyruvate dehydrogenase
PEP	phosphoenolpyruvate
PFP	phosphofructokinase pyrophosphate
PGAM	phosphoglycerate mutase
PGI	phosphoglucose isomerase
PGM	phosphoglucomutase
PGP	phosphoglycolate phosphatase
PK	pyruvate kinase
PPC	phosphoenolypyruvate carboxylase
PPE	phosphopentose epimerase
PPI	phosphopentose isomerase
PRK	phosphoribulokinase
PRO	proline
PYR	pyruvate
R5P	ribose-5-phosphate
R_L	respiration in the light
RU5P	ribulose-5-phosphate
RUBISCO_CO2	ribulose-1,5-bisphosphate carboxylase (oxygenase)
RUBISCO_O2	ribulose-1,5-bisphosphate (carboxylase) oxygenase
RUBP	ribulose-1,5-bisophosphate
S6P	sucrose-6-phosphate
S7P	sedoeheptulose-7-phosphate
SBP	sedoheptulose-1,7-bisophosphate
SBPase	sedoheptulose-1,7-bisphosphatase
SCA	succinyl-CoA
SER	serine
SFrc	sucrose fructosyl moiety
SGA1	serine:glyoxylate aminotransferase
SGlc	sucrose glucosyl moiety
SIM	selected ion monitoring
SPS	sucrose-phosphate synthase
SRES	squared residual
SS	starch synthase
SSR	sum-of-squared residuals
Suc	sucrose
SUC	succinate
T_3PGA	3PGA transporter
T_TP	TP transporter
TBDMS	*tert*-butyldimethylsilyl
TCA	tricarboxylic acid
THR	threonine
TK1	transketolase
TMS	trimethylsilyl
TP	triose phosphate
TS	threonine synthase

Table S2.2 (cont'd)

UDPG	uridine diphosphate glucose
v_c	velocity of rates of carboxylation
v_o	velocity of rates of oxygenation
Vpr	photorespiratory CO_2 release
X5P	xylulose-5-phosphate

Table S2.3: A comparison of the goodness of fit between data and best-fit simulations from alternative models. Starting model with no inactive pools, model with unlabeled glucose source, and model with sucrose recycling reactions and sucrose vacuole pool reactions were compared with fluxes for key reactions, SSR, top five most different SSR, and DOF. 5* DOF in terms of fluxes. The lowest value of SSR is shown in blue, the 50th percentile of SSR is shown in yellow, the highest value of SSR is shown in red. The starting model with no pools had the biggest overall SSR (1340) and highest individual SSR for R5P, FBP, UDPG, G6P, and F6P. The model with an unlabeled glucose source had both lower overall SSR and individual SSR for R5P, FBP, UDPG, G6P, and F6P. The model with sucrose recycling reactions and sucrose vacuole pool reactions had both lowest overall SSR and individual SSR for R5P, FBP, UDPG, G6P, and F6P. All abbreviations are shown in **Table S2.2**.

Model	Reactions	Flux	SSR	TOP5 most different SSR		ΔDOF
No inactive pools			1340	UDPG	215	0
				R5P	115	
				FBP	112	
				G6P	123	
				F6P	109	
No inactive pools + unlabeled glucose source	CO2.u -> CO2	0	1340	UDPG	218	1
				R5P	118	
				FBP	114	
				G6P	112	
				F6P	98	
	Glucose.u -> G6P.p	0.5	1300	UDPG	216	1
				R5P	116	
				FBP	113	
				G6P	85	
				F6P	102	
	TP.u -> TP.p	0.3	1273	UDPG	209	1
				R5P	112	
				FBP	62	
				G6P	109	
				F6P	96	
	Glucose.u -> G6P.c	1.9	1126	UDPG	109	1
				R5P	117	
				FBP	101	
				G6P	62	
				F6P	59	
No inactive pools + sucrose recycling reactions + sucrose vacuole pool reactions	Suc.v <-> Suc.c	2.11	968	UDPG	53	5*
	Glc.v <-> Glc.c	2.11		R5P	101	
	Suc.c-> Glc.c + Fru.c	0.05		FBP	76	
	Glc.c -> G6P.c	2.16		G6P	32	
	Fru.c -> F6P.c	2.16		F6P	19	

Lowest value — 50 percentile — highest value

Table S2.4: Predicted and measured ratios between M1 to M0 of CBC intermediates based on their predicated and measured percentage of isotopologues.

Metabolites	Isotopologue	Percentage of isotopologue		Ratio between M1/M0	
		Predicted	Measured	Predicted	Measured
GAP/DHAP	M0	0.01	2.4	65	0.2
	M1	0.6	0.5		
	M2	12.1	5.0		
	M3	87.4	92.1		
PGA	M0	0.01	1.6	67	0.4
	M1	0.5	0.7		
	M2	11.7	6.5		
	M3	87.7	91.1		
R5P	M0	0.001	2.4	48	0.2
	M1	0.04	0.4		
	M2	0.7	4.6		
	M3	6.6	4.0		
	M4	31.7	11.2		
	M5	61.0	77.3		
RU5P/XU5P	M0	0.001	2.2	51	0.2
	M1	0.03	0.4		
	M2	0.6	4.7		
	M3	6.1	3.1		
	M4	30.9	12.2		
	M5	62.4	77.4		
RUBP	M0	0.0001	1.6	85	0.2
	M1	0.005	0.4		
	M2	0.2	1.3		
	M3	2.6	1.5		
	M4	22.2	11.5		
	M5	75.1	83.7		
F6P	M0	0.000004	2.3	97	0.1
	M1	0.0004	0.3		
	M2	0.02	0.4		
	M3	0.3	1.2		
	M4	4.0	1.9		
	M5	25.9	11.4		
	M6	69.7	82.6		
G6P	M0	0.0001	2.9	61	0.1
	M1	0.003	0.4		
	M2	0.1	0.3		
	M3	1.1	2.1		
	M4	8.3	8.3		
	M5	33.6	10.4		
	M6	57.0	75.6		
S7P	M0	0.00000001	1.2	175	0.2
	M1	0.000002	0.3		
	M2	0.0002	0.2		
	M3	0.01	1.6		
	M4	0.2	2.1		
	M5	2.6	2.3		
	M6	21.3	12.1		
	M7	76.0	82.1		

Table S2.5: Contributions of fully unlabeled and partially labeled isotopologues to the lack of complete labeling in glucose 6-phosphate after two hours of labeling with $^{13}CO_2$. Relative abundances are from Table S5. Fully unlabeled G6P accounts for only 0.174 / (0.174+0.365) = 32% of the labeling deficit.

	Relative abundance	^{12}C in M0	^{12}C in M1 to M6
M0	0.029	0.174	-
M1	0.004	-	0.02
M2	0.003	-	0.012
M3	0.021	-	0.063
M4	0.083	-	0.166
M5	0.104	-	0.104
M6	0.756	-	0
Sum	1	0.174	0.365

Table S2.6: Carbon accounting for the model. Values in the absolute columns are fluxes from the model (Fig. 2.3) converted to a carbon basis. The last two columns are absolute values divided by the net rate of CO_2 assimilation.

	Absolute		**Relative to net assimilation**	
	μmol g^{-1} FW hr^{-1}		**%**	
	In	**Out**	**In**	**Out**
Calvin-Benson cycle carbon inputs and outputs				
Rubisco	172		123%	
Photorespiration	75	102	54%	73%
TPT		117		84%
Starch synthesis		63		45%
G6P shunt	35		25%	
Total	**282**	**282**	**202%**	**202%**
CO_2 budget				
Rubisco	172		123%	
Photorespiration		25		18%
G6P shunt		7		5%
Fatty acids		0.4		0.3%
In minus out	**139.6**		**100%**	
End Products				
Starch		63.0		45%
Sucrose		68.4		49%
Other cytosolic		6.5		5%
Fatty acids		0.8		1%
Total end products		**138.7**		**99%**

Table S2.7: v_o/v_c for models with and without labeling input for serine and glycine, with and without constraints of v_o/v_c. Four scenarios were tested: 1) with serine and glycine labeling input, unconstrained v_o/v_c; 2) with serine and glycine labeling input, constrained v_o/v_c = 0.31 +/- 5%; 3) without serine and glycine labeling input, unconstrained v_o/v_c; 4) without serine and glycine labeling input, constrained v_o/v_c = 0.31 +/- 5%.

	with serine and glycine		**without serine and glycine**	
	Unconstrained v_o/v_c	**Constrained v_o/v_c**	**Unconstrained v_o/v_c**	**Constrained v_o/v_c**
v_o	161.7	215.1	167.0	167.0
v_c	33.2	65.0	50.6	50.6
v_o/v_c	0.21	0.30	0.30	0.30

DATASET LEGENDS

All supplemental datasets can be found at the following link:

https://doi.org/10.1073/pnas.2121531119.

Dataset S1 (separate file). Experimentally measured mass isotopologue distributions of measured metabolites.

Dataset S2 (separate file). Parameter value estimates and model selection results for aggregated CBC intermediate datasets and individual metabolites. Parameters in exponential terms are sorted in terms of the absolute magnitude of their decay term.

Dataset S3 (separate file). Comparisons of the model in this work with previous models (Ma et al., 2014; Xu et al., 2021a). Reactions that are different from Ma et al., (2014) are labeled in red. Reactions from Xu et al., (2021a) are shown in yellow. Reactions newly added in this publication are shown in blue. Reactions have been removed from Ma et al., (2014) and Xu et al., (2021a) are shown in green. Note that the parameters for alanine, glycine, and serine have been kept in the model because of their compartmentation complexity.

Dataset S4 (separate file). Estimated flux values and 95% confidence intervals by parameter continuation. Values are absolute fluxes (μmol metabolites gFW-1 hr-1) based on the measured net CO2 uptake rate. The net flux is the difference between influx and efflux of metabolites moved in or out of the cell. The exchange flux is the minimum of the forward and backward fluxes of a reversible reaction. Some confidence intervals of exchange fluxes are unidentifiable or infinite. Subcellular fluxes are shown by metabolites spatially separated in the plastid (.p) and cytosol (.c).

Dataset S5 (separate file). Parameters for transitions of measured metabolites in multiple reaction monitoring (MRM) with LC-MS/MS and selected ion monitoring (SIM) with GC-MS. LC-MS/MS dwell time was set at 20 ms for each transition. Q1, m/z of the precursor ion; Q3, m/z of the product ion. Cone and collision energy were optimized by direct infusion of standards. Amino and organic acids were measured by GC-MS by tert-butyldimethylsilyl (TBDMS) derivatization whereas glucose, fructose, and sucrose were derivatized by trimethylsilyl (TMS).

REFERENCES

Akaike H (1998) Information Theory and an Extension of the Maximum Likelihood Principle. *In* E Parzen, K Tanabe, G Kitagawa, eds, Selected Papers of Hirotugu Akaike. Springer New York, New York, NY, pp 199–213

Caemmerer S, Evans J (1991) Determination of the Average Partial Pressure of CO2 in Chloroplasts From Leaves of Several C3 Plants. Functional Plant Biology **18**: 287

Draper NR, Smith H (1998) Extra Sums of Squares and Tests for Several Parameters Being Zero. Applied Regression Analysis. John Wiley & Sons, Ltd, pp 149–177

Farquhar GD, Caemmerer S, Berry JA (1980) A biochemical model of photosynthetic CO2 assimilation in leaves of C3 species. Planta **149**: 78–90

Hastie T, Tibshirani R, Friedman J (2017) Model Assessment and Selection. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, 2nd ed. Springer, New York, NY, pp 219–260

Holm S (1978) Board of the Foundation of the Scandinavian Journal of Statistics A Simple Sequentially Rejective Multiple Test Procedure Author (s): Sture Holm Published by : Wiley on behalf of Board of the Foundation of the Scandinavian Journal of Statistics Stable U. Scandinavian Journal of Statistics **6**: 65–70

Johnson NL (1949) Systems of frequency curves generated by methods of translation. Biometrika **36**: 149–176

Ma F, Jazmin LJ, Young JD, Allen DK (2014) Isotopically nonstationary 13C flux analysis of changes in Arabidopsis thaliana leaf metabolism due to high light acclimation. Proceedings of the National Academy of Sciences of the United States of America **111**: 16967–16972

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al (2011) Scikit-learn: Machine learning in Python. the Journal of machine Learning research **12**: 2825–2830

Politis DN, Romano JP (1991) A circular block-resampling procedure for stationary data. Purdue University. Department of Statistics

Schwarz G (1978) Estimating the Dimension of a Model. The Annals of Statistics **6**: 461–464

Sharkey TD (2016) What gas exchange data can tell us about photosynthesis. Plant Cell and Environment **39**: 1161–1163

Tcherkez G, Cornic G, Bligny R, Gout E, Ghashghaie J (2005) In vivo respiratory metabolism of illuminated leaves. Plant Physiology **138**: 1596–1606

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al (2020) SciPy 1.0: fundamental algorithms

for scientific computing in Python. Nature Methods **17**: 261–272

Xu Y, Fu X, Sharkey TD, Shachar-Hill Y, Walker and BJ (2021) The metabolic origins of non-photorespiratory CO2 release during photosynthesis: a metabolic flux analysis. Plant Physiology 1–18

Young JD (2014) INCA: A computational platform for isotopically non-stationary metabolic flux analysis. Bioinformatics **30**: 1333–1335

Chapter 3

Accurate flux predictions using tissue-specific gene expression in plant metabolic modeling

This research was published in:

J. A. M. Kaste, Y. Shachar-Hill. Accurate flux predictions using tissue-specific gene expression in plant metabolic modeling. *Bioinformatics* **39(5)**, btad186 (2023).

3.1. Preface

As discussed in Chapter 1, FBA flux predictions are often some combination of imprecise and inaccurate owing to the small amount of empirical data brought to bear in most FBA studies. One attractive method of improving FBA flux accuracy is to come up with methods of incorporating omic – particularly transcriptomic – data into the prediction process. Unfortunately, this is easier said than done, with many previous attempts failing to consistently generate more accurate predictions than parsimonious FBA. These benchmarking studies were done in unicellular systems where the methods evaluated are generally some variation on the idea "If gene A encoding reaction A* is expressed more highly than gene B encoding reaction B*, reaction A* should, all else being equal, have higher flux than reaction B*." For what should be obvious reasons, this is not a very good assumption to make.

Early on in my Ph. D., it occurred to me that when modeling a multi-tissue system, it seems likely that the relationship between transcript abundance and flux between tissues of the same organism is probably more consistent than this same relationship across different gene-to-reaction pairings within a tissue or organism. This was merely a hunch, but it was compelling enough for me to convince Dr. Shachar-Hill to pursue the idea of making this into an algorithm and assessing whether it can make our flux predictions more accurate, as benchmarked by comparison against ^{13}C-MFA (the gold-standard, as we argue in Chapter 1). Although I was ultimately interested in doing this in *C. sativa*, I decided to do this study initially with *A. thaliana* since it has multiple tissue-atlas RNA-seq datasets, a publicly available quantitative proteome dataset, and a lot of prior genome-scale models. As it turns out, the method I developed has been shown to be quite effective, as we present in this chapter. Due to the similarity between *A. thaliana* and *C. sativa*'s genomes, we believe this represents a significant step in the direction of more accurate and useful flux modeling in *C. sativa* for the purposes of engineering this organism for improved biofuel production.

I carried out all of the model building, computational analysis, and manuscript writing for this study in consultation with Dr. Shachar-Hill. I am first author on the manuscript featured in this chapter, which has been published in the journal *Bioinformatics*.

3.2. Abstract

Motivation: The accurate prediction of complex phenotypes such as metabolic fluxes in living systems is a grand challenge for systems biology and central to efficiently identifying

biotechnological interventions that can address pressing industrial needs. The application of gene expression data to improve the accuracy of metabolic flux predictions using mechanistic modeling methods such as Flux Balance Analysis (FBA) has not been previously demonstrated in multi-tissue systems, despite their biotechnological importance. We hypothesized that a method for generating metabolic flux predictions informed by relative expression levels between tissues would improve prediction accuracy.

Results: Relative gene expression levels derived from multiple transcriptomic and proteomic datasets were integrated into Flux Balance Analysis predictions of a multi-tissue, diel model of *Arabidopsis thaliana*'s central metabolism. This integration dramatically improved the agreement of flux predictions with experimentally based flux maps from ^{13}C Metabolic Flux Analysis (^{13}C-MFA) compared with a standard parsimonious FBA approach. Disagreement between FBA predictions and MFA flux maps, as measured by weighted averaged percent error values, dropped from be-tween 169-180% and 94-103% in high light and low light conditions, respectively, to between 10-13% and 9-11%, depending on the gene expression dataset used. The incorporation of gene ex-pression data into the modeling process also substantially altered the predicted carbon and energy economy of the plant.

Availability: Code is available from
https://github.com/Gibberella/ArabidopsisGeneExpressionWeights.

3.3. Introduction

A grand challenge for systems biology is the ability to accurately predict complex phenotypes from omic datasets based on functional principles and mechanisms. Patterns of cellular metabolism – flux maps – are one such complex phenotype (Ratcliffe and Shachar-Hill, 2006), for which tools to predict phenotypes from basic assumptions have proven useful in exploring and designing metabolic capabilities (Burgard et al., 2003; Orth et al., 2010b; Chen et al., 2011). Methods to quantify flux maps from labeling data now allow the testing of such predictions in both simpler and multicellular systems. However, the integration of omic data to improve the accuracy of flux predictions is still at an early stage.

Metabolic flux predictions are also important for real world applications since modifying an organism's metabolic activity in order to achieve some practical aim, such as overproducing a specific metabolite, is central to many biotechnology projects. As in other areas of engineering, metabolic engineering can benefit from mathematical models that describe and predict the

behavior of the relevant system(s). Researchers have developed two major modeling approaches to address this need: (1) ^{13}C-Metabolic Flux Analysis (^{13}C-MFA) and (2) Flux Balance Analysis (FBA) (Orth et al., 2010b; Antoniewicz, 2015). With ^{13}C-MFA, steady-state or kinetic isotopic labeling data for metabolites in a small- to medium-sized network are used to obtain estimates of the net and exchange fluxes through that network (Antoniewicz, 2015). These metabolic flux maps are regarded as the most reliable measures of *in vivo* metabolic fluxes; however, the throughput of this technique is limited by the large amounts of isotopic labeling data and other measurements needed to generate each flux map. FBA, which is based on applying conservation principles to a network of reactions using one or more assumptions about the functional objective(s) driving biological organization, requires substantially less experimental input data, and is therefore an attractive and commonly used metabolic modeling technique.

FBA and related metabolic modeling methods in microbial systems, together with Genome-Scale Models (GEMs) that represent the biochemical reactions encoded in an organism's genome, have enabled radical modification of microbial central metabolism (Gleizer et al., 2019) and substantial improvements in bioproduct yields (Lee et al., 2007; Park et al., 2007). These methods can, for example, allow bioengineers to predict the behavior of their system and identify interventions, such as gene knock-outs or knock-ins, that will help them modify the organism's phenotype (Burgard et al., 2003; Tepper and Shlomi, 2009b). However, many metabolic engineering applications require the modification not of microorganisms, but of multicellular eukaryotes like plants. Most GEMs of plants to date [e.g., (Poolman et al., 2009; Dal'Molin et al., 2010a; Saha et al., 2011; Arnold and Nikoloski, 2014)], have treated plants, which are composed of multiple tissues with substantial functional diversity, as single-tissue aggregated metabolic networks. This has motivated the creation of "multi-tissue" GEMs to investigate source-sink dynamics and resource allocation, with the earliest efforts in this space focusing on the interplay between mesophyll and bundle-sheath cells in C4 photosynthesis (Dal'Molin et al., 2010b; Shaw and Cheung, 2020).

Re-engineering of plant metabolism on the scale seen in microbial systems has not, to-date, been possible and predictive modeling has been neither validated in detail nor applied to successful plant metabolic engineering. This is partly due to the ease and high throughput of microbial transformation relative even to model plant systems. In addition to the greater experimental demands, the metabolic modeling of these systems is also substantially harder.

There is, consequently, a relative lack of MFA datasets with which to compare the predicted flux maps from FBA in plants. This contrasts with the availability of rich multi-omic datasets combining flux estimates with transcript and protein data for a number of different genotypes and growth conditions in systems like *E. coli* (Ishii et al., 2007). The challenges involved in generating ^{13}C-MFA flux maps for plants make improvement of plant FBA flux predictions an attractive path towards replicating the biotechnological successes seen in microbes.

An appealing approach to improving the quality of plant FBA predictions is the integration of additional network-wide data from transcriptomic and proteomic datasets. Gene expression data – particularly transcript data – is substantially easier to generate than ^{13}C-MFA flux maps. Previous attempts at integrating gene expression datasets into metabolic flux predictions have been reviewed elsewhere (Machado and Herrgård, 2014; Vijayakumar et al., 2017). Methods developed before 2014 were evaluated on the basis of their ability to improve upon parsimonious FBA (pFBA) (Lewis et al., 2010) in terms of their predictions' agreement with MFA-estimated fluxes in microorganisms and were found to not do so reliably (Machado and Herrgård, 2014). A key limitation of these studies was a lack of comparison of FBA-predictions against ^{13}C-MFA derived flux estimates. This lack of comparison against ^{13}C-MFA is shared by the plant FBA literature, in which we are aware of only a small number of evaluations under heterotrophic conditions in green algae (Boyle et al., 2017), *Arabidopsis* cell cultures (Williams et al., 2010; Cheung et al., 2013), and *Brassica napus* embryos (Hay and Schwender, 2011). Since then, several studies have developed algorithms benchmarked by their ability to make predictions in agreement with empirical flux maps derived from MFA studies (Tian and Reed, 2018; Pandey et al., 2019; Ravi and Gunawan, 2021). These studies have focused on unicellular organisms or animal tissues modeled in isolation. Their application to FBA in more complex systems is limited by the large number of resource-intensive MFA datasets needed to calibrate them (Tian and Reed, 2018) or their need for a reference expression dataset paired with an assumed-correct flux map (Pandey et al., 2019; Ravi and Gunawan, 2021).

To improve the accuracy of FBA in multicellular systems, particularly plants with their complex metabolic networks, we developed a method that integrates tissue-atlas data from multi-tissue systems into the flux-minimization procedure employed in pFBA. This method incorporates evidence from gene expression datasets into FBA metabolic flux predictions by applying weights to individual reactions according to the relative transcript or protein expression

of the gene(s) assigned to those reactions between different modeled tissues. The method is evaluated on its ability to make predictions in agreement with MFA flux maps. We demonstrate substantial improvements in the agreement of our FBA predicted fluxes with flux estimates from a ^{13}C-MFA study on *Arabidopsis thaliana* rosette leaf central metabolism (Ma et al., 2014). Finally, we show that multiple gene expression datasets, when used as inputs, result in similar improvements in agreement and that this result generalizes across different MFA flux maps. This approach has particular potential for plant and animal systems for which there are only a limited number of experimental flux maps.

3.4. Methods

3.4.1. Overview of approach

Our method makes two key assumptions: (1) Metabolic flux maps predicted from pFBA (Lewis et al., 2010), minimizing the sum total of flux through the network, are more likely to reflect real flux maps than ones not subject to this constraint, and (2) A reaction present in two tissues *A* and *B* catalyzed by an enzyme encoded by a gene that is highly expressed in *A* and poorly expressed in *B* is likely to carry higher flux in tissue *A*.

We incorporate assumption 1 by making the objective function of our FBA optimization the minimization of total flux, the same as pFBA (Lewis et al., 2010). This is represented mathematically as finding the minimum value of the linear combination of all fluxes in the network, with each flux v_j multiplied by a corresponding coefficient c_i:

$$\min_{v_j} \sum_{j \in Reactions} c_j * v_j \quad (1)$$

Where *Reactions* is the list of all reactions j in the network, v_j is the flux through a reaction j, and c_j is the coefficient – hereafter referred to as a *penalty weight* since it represents a penalization of the likelihood of using a reaction j to carrying flux. When c_j is 1 for all reactions, our method reduces to pFBA, which can be seen as the limiting case of gene expression having no influence in predicting network flux patterns. We incorporate assumption 2 by calculating, for each reaction in our network model, a coefficient derived from the relative expression of genes encoding the enzyme(s) that catalyze that reaction between the different tissues in the gene expression dataset. Using the coefficient vector to account for relative expression resembles the approach taken by Jenior et al., (2020). However, our method compares gene expression across tissues within a multi-tissue model to generate more accurate flux predictions, rather than

comparing the expression of genes to the most expressed gene in a dataset as a proxy for transcriptional investment as a way of generating context-specific models. Reactions and genes are associated by the Gene-Protein-Reaction (GPR) terms in the model. This results in reactions mapped to relatively highly expressed genes receiving small values of c_j and reactions mapped to minimally expressed genes receiving large ones.

3.4.2. Model construction and dataset selection

The *Arabidopsis thaliana* core metabolism model developed by Arnold and Nikoloski, (2014) was used as the basis for a multi-tissue diel model. This model was chosen due to its rich GPR annotation and focus on central metabolism. The core model was duplicated six times to create leaf, stem, and root versions of the model for both day and night, which were interconnected by transporters allowing the movement of specific compounds and metabolites. The substrates, products, and constraints applied to the model can be found in the **Appendix B, Supplementary Methods**. The model used in this study can be found in **Appendix B, Dataset S2.**

^{13}C-MFA flux maps were obtained *in planta* in *Arabidopsis thaliana* by Ma et al., (2014), and these were used as the empirical best estimates of flux distributions. Although there are not any other ^{13}C-MFA flux maps available of autotrophic *A. thaliana* leaves, Szecowka et al., (2013) provides estimates of select fluxes in autotrophic *A. thaliana* leaf central metabolism, which we used for additional confirmation of our method's efficacy. The pairing of fluxes in both flux studies to the FBA network are described in **Appendix B, Dataset S1.**

We searched the literature for high-quality, high-coverage RNA-seq and quantitative proteomic tissue atlases and found two suitable datasets meeting these criteria: Mergner et al., (2020) and Klepikova et al., (2016). The proteomic dataset from Mergner et al., (2020) is a mass-spectrometry-based quantitative proteome that reports IBAQ values, which are an accurate measure of protein abundances (Krey et al., 2014). For bioinformatic processing details, see **Appendix B, Supplementary Methods**. For dataset IDs, growth conditions and key parameters from each study, see **Tables S3.4-S3.5**.

3.4.3. Penalty weight vector calculation

We calculated the expression weight for each gene in each tissue on the basis of how the expression of a reaction in a particular tissue, as measured by transcriptomic or proteomic abundance, compared to the expression of that same gene in the other tissues.

$$W_{it} = \frac{Max(E_i)}{E_{it}} \quad (2)$$

Where W_{it} is the expression weight for a given gene *i* in a tissue *t*, E_i is the list of expression values of gene *i* for each tissue, E_{it} is the expression of gene *i* in tissue *t*, and Max() is the maximum value from a set of one or more elements. Many GPRs in the model consist of multiple genes that represent isozymes or members of protein complexes. The former are denoted by OR terms and the latter by AND terms in the GPR formulation. This results in many reactions having more than one expression weight due to being mapped to multiple genes. We combine these multiple weights into a single penalty weight value for each reaction by averaging the expression weights of isozymes and taking the "worst" (i.e., largest, most penalizing value) when genes form subunits of a protein complex. As an example, the penalty weight for a reaction *R* in the leaf subnetwork of our model with a GPR of the form (Gene1 OR Gene2) AND (Gene3), corresponding to a protein complex made of the product of Gene 3 and the product of Gene 1 or Gene 2, would be represented by:

$$c_{R,lf} = SF\left(Max\left(\frac{\left(W_{gene1,lf} + W_{gene2,lf}\right)}{2}, W_{gene3,lf}\right) - 1\right) + 1$$

Where $c_{R,lf}$ represents the overall penalty weight in the leaf (lf) for reaction R, SF (or the scaling factor) is a coefficient that modulates the magnitude of the calculated penalty weights and $W_{gene1,lf}$, $W_{gene2,lf}$, and $W_{gene3,lf}$ are the penalty weights for the individual genes Gene1, Gene2, and Gene3. Note that in the present implementation of this method, stoichiometric coefficients in GPR terms are ignored. When the one or more genes contained in a GPR for a reaction/tissue combination are all more highly expressed than the same genes in the other tissues, the scale for that reaction/tissue combination will be 1. For reaction/tissue combinations that have no corresponding GPR, we explored setting the penalty weights to 1 or a value calculated from the median penalty weight assigned to reactions in the same tissue **(for details, see Appendix B, Supplementary Methods)**.

3.4.4. Optimization

The optimization done in this paper is a variation on pFBA, which finds the flux map(s) satisfying imposed constraints with minimum total flux through the network (Lewis et al., 2010). The minimization of total flux (**Eq. 1)** is subject to the following constraints:

$$Sv = 0 \quad (5)$$

$$LB_j \leq v_j \leq UB_j \quad (6)$$

$$v_{biomass_{(tissue)}} = v_{fixed\ biomass_{tissue}} \quad (7)$$

Where **S** is the stoichiometric matrix of the metabolic network being modeled, ***v*** is the vector of all fluxes, *LB* and *UB* are the vectors of all upper and lower bound constraints, and $v_{\text{biomass(tissue)}}$ and $v_{\text{fixed biomass(tissue)}}$ are the biomass flux for a given tissue and the defined biomass constraint for that tissue, respectively. **Eq. 5** represents the steady state of all internal metabolites, **Eq. 6** represents the bounds and reversibility constraints, and **Eq. 7** represents the definition of biomass accumulation rates. All optimizations were done in the COnstraint-Based Reconstruction and Analysis (COBRA) Toolbox in MATLAB (Heirendt et al., 2019) using the Gurobi™ optimizer version 8.1.1 (Gurobi Optimization, LLC, 2021).

3.4.5. Error evaluation

We make the assumption that the ^{13}C-MFA fluxes reported by Ma et al., (2014b) are the true *in vivo* metabolic fluxes and therefore regard the discrepancy between FBA-predicted fluxes and these ^{13}C-MFA fluxes as a measure of error. Biomass accumulation (i.e., the difference in dry weight between a timepoint t and another timepoint t_{-1}) was not reported by Ma et al., (2014b), but is the basis for the flux through the biomass equation in FBA. To allow a comparison between our FBA-predicted fluxes and the MFA-estimated fluxes in Ma et al., (2014b), we set an arbitrary biomass flux of 0.01 g/hr through the leaf, stem, and root biomass reactions in both the day and night, similar to the approach taken by de Oliveira Dal'Molin et al., (2015). We then normalized our fluxes by multiplying them by a factor A calculated as the ratio of the measured leaf CO_2 uptake from Ma et al., (2014b) and the net leaf CO_2 uptake in our FBA flux map. A weighted average error for each FBA-predicted flux map was then obtained using the following expression:

$$\sum_{j \in Measured} \left(\frac{\left| (v_j^p * A) - v_j^m \right|}{\left| v_j^m \right|} * \frac{\left| v_j^m \right|}{\sum_{j \in Measured} \left| v_j^m \right|} \right) \quad (8)$$

Where v_j^p and v_j^m are the FBA-predicted and MFA-estimated fluxes of a flux j and A is the normalization factor previously described. We calculated weighted average errors rather than just average errors because small absolute differences between FBA-predicted and MFA-estimated

flux values can correspond to extremely large % error values when the MFA-estimated fluxes are small. Additional details on the error evaluation can be found in the Appendix B, Supplementary Methods. We quantified the maximum/minimum weighted average errors of each flux map using Flux Variability Analysis (FVA) (Mahadevan and Schilling, 2003). For details, see Appendix B, Supplemental Methods.

3.5. Results

3.5.1. The application of gene expression penalty weights reliably reduces discrepancies between FBA-predicted and MFA-estimated fluxes

Table 3.1. Weighted average % error values calculated from weighted vs. unweighted flux maps for transcriptomic and proteomic datasets from Mergner et al., (2020) and Klepikova et al., (2016). Values in brackets represent the lowest and highest possible error values given the results of Flux Variability Analysis. Weighted average error values were calculated from flux maps generated using a scaling factor of 1.

Dataset	Light Level	Weighted average error (%)	
		No gene expression weights	With penalty weights
Mergner et al. Transcriptome	High	169 – 180	14.7 – 17.1
	Low	93.8 – 103	14.9 – 18.1
Mergner et al. Proteome	High	169 – 180	10.9 – 13.4
	Low	93.8 – 103	8.74 – 10.9
Klepikova et al. Transcriptome	High	169 – 180	14.8 – 17.4
	Low	93.8 – 103	19.3 – 21.7

Predicted flux maps were generated for a multi-tissue diel model of *Arabidopsis thaliana*'s central metabolism using flux balance analysis in which the sum of all the metabolic and transport fluxes required for steady state growth is minimized, with each flux being multiplied by a penalty weight that was derived from the relative expression of the gene(s) involved in conducting that flux (see methods). Penalty weights for each reaction were calculated from RNA-seq (Klepikova et al., 2016; Mergner et al., 2020) and proteomic (Mergner et al., 2020) datasets using the relative expression of each gene in the different tissues. The weighted average % error between these flux maps and ^{13}C-MFA estimates from Ma et al., (2014b) were used to quantify the accuracy of these FBA predictions, as compared to the accuracy of flux maps generated by pFBA (Lewis et al., 2010) alone. The flux maps arrived at after the application of either transcriptomic or proteomic penalty weights show greater agreement, as measured by the weighted average % error, with ^{13}C-MFA estimates than the results from pFBA

alone **(Table 3.1)**. These reductions in error are substantial and statistically significant at $\alpha = 0.01$; they are consistent across comparisons against two different flux maps (high-light and low-light conditions) and are sustained across a range of assumed ratios of starch to sucrose production and carboxylase to oxygenase fluxes through rubisco (v_o/v_c). Marked reductions in error are seen whether one uses the transcriptomic or proteomic tissue-atlas datasets from Mergner et al., (2020) or Klepikova et al., (2016), so that the improvement in flux predictions is not dependent on the values obtained in a specific gene expression dataset or type.

We wanted to confirm that these reductions in error are in fact dependent on penalty weights calculated from gene expression data and not an artifact of the weighting procedure itself. Indeed, previous studies have used the application of randomized weights as a method of exploring different possible flux modes in a plant metabolic network (Cheung et al., 2015). We found that substituting the leaf for the root proteomic dataset, and vice-versa, resulted in no reduction in weighted average error **(Appendix B, Table S3.1)** compared to pFBA. Neither did randomization of the penalty weight vector and subsequent optimization. The mean of the weighted average errors of 50 high-light condition flux maps generated with independent randomized penalty weight vectors at a scaling factor of 1 was 201%, versus the unweighted error value of 169-180% for that condition.

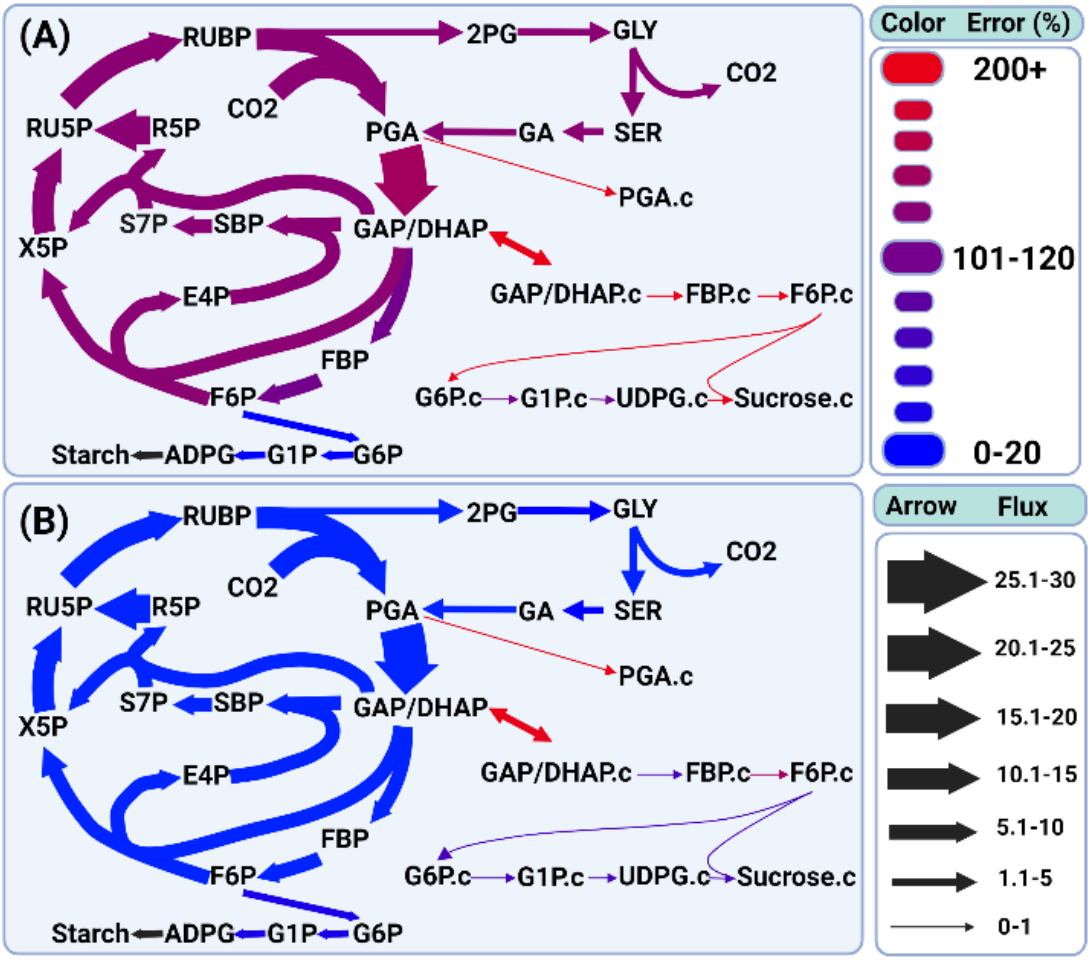


Figure 3.1: Percent errors of specific reactions in central metabolism before **(A)** and after **(B)** gene expression weight application. The error values in **(A)** are the **lowest** possible given FVA results and the values in **(B)** are the **highest** possible given FVA results. We see substantial decreases in errors associated with central carbon assimilation, as well as starch and sucrose synthesis. Since the ^{13}C-MFA estimated fluxes from Ma et al., (2014) do not feature the flux from ADPG to Starch, this flux lacks an estimated error and is therefore shown in black. Flux values are relative to the lowest flux in the network.

3.5.2. Increases in agreement between FBA-predicted and MFA-estimated fluxes are broadly distributed across central metabolism

Although there is variation among individual fluxes in the degree to which omic data integration improves agreement between predicted and experimentally derived values, the reduction in weighted error as a result of penalty weight application is distributed broadly across the fluxes for which ^{13}C-MFA estimates are available. If, for example, the improvement were due to a substantial decrease in one or a small number of high-flux reactions and a negligible decrease or even increase in error for other reactions **(Fig. 3.1)** the overall finding would be less striking and potentially less broadly applicable. The reductions in error are consistent not only

across metabolic subsystems within a single FBA flux map, but also across alternative stoichiometric network structures. Initial pFBA-derived solutions for a model identical to that used to generate the other predictions except with unconstrained uptake and discharge of protons from root tissue show similar reductions in error **(Appendix B, Table S3.2)**. Upon application of penalty weights, this model converges to a similar value of weighted average error and linear correlation as other model configurations.

Table 3.1: Measures of carbon and energy utilization derived from the predicted flux maps with and without penalty weights applied. Reference values: a, (Ma et al., 2014b); b, (Kramer et al., 2004); c, (Weraduwage et al., 2015). (b) and (c) reference values are not associated with a particular light level.

Dataset used for weighting	Light	Rubisco flux ÷ net CO_2 assimilation	Photorespiratory CO_2 loss / net CO_2 assimilation (%)	Cyclic/Linear Electron Flow	% of leaf daytime CO_2 assimilation going to biomass
None	High	2.86	62	24%	43
	Low	1.85	26	31%	54
Mergner *et al.* Protein	High	1.29	26	20%	18
	Low	1.17	14	15%	26
Mergner *et al.* Transcripts	High	1.20	25	21%	18
	Low	1.15	14	27%	33
Klepikova *et al.*Transcripts	High	1.30	27	17%	19
	Low	1.25	15	14%	31
Reference values	High	1.28[a]	28[a]	13%[b]	56%[c]
	Low	1.17[a]	16[a]		

3.5.3. Error reductions are a function of the scaling factor parameter and are improved by the application of a tissue-specific median weight for reactions lacking Gene Protein Reaction terms

The magnitude of the penalty weights calculated and applied by the present method depend on the magnitude of the scaling factor term, **(Eq. 2)**. The increased agreement between the FBA-predicted and MFA-estimated flux maps only manifests in the majority of cases for scaling factors of 0.05-0.1 or greater **(Fig 3.2)**. We also note that the relationship between the scaling factor value and the improved agreement is monotonic – that is, we do not see erratic increases and decreases as we increase the scaling factor value and, by extension, the strength of the assumed relationship between flux and gene expression. The necessity of a non-negligible scaling factor, the consistency of error improvement as the scaling factor is increased, and the similarity in the pattern of error improvement across multiple datasets as seen in **Fig 3.2** all suggest that real biological signal related to the partitioning of metabolic activity across the plant's tissues is being extracted from the gene expression datasets. Finally, we observe that the

flux maps generated using penalty weight derived from the Mergner et al., (2020) proteomic dataset have noticeably better weighted average errors than flux maps generated using transcriptomic dataset **(Table 3.1; Fig. 3.2)**. This is consistent with the closer relationship between measured protein levels and metabolic fluxes than between transcripts and fluxes. It is also consistent with at least one other study's attempts at integrating gene expression data into FBA in *E. coli* (Tian and Reed, 2018).

Although the presented method does not involve fitting the scaling factor parameter using goodness-of-fit to the ^{13}C-MFA fluxes, in **Fig 3.1** and **Tables 3.1**-**3.2**, we show results from a scaling factor of 1 because it falls in the plateau of low average error values we see in **Fig 3.2.** There are no independent ^{13}C-MFA datasets of this system against which to evaluate whether a scaling factor value of 1 generalizes well outside of the datasets considered in the present study. However, Szecowka et al., (2013) do report fluxes from illuminated *A. thaliana* leaves estimated by kinetic flux profiling. The flux map generated using v_o/v_c and starch:sucrose synthesis constraints from that study without any omic weighting has a weighted average error of 108%; this error drops to between 6-9% when penalty weights generated with a scaling factor of 1 are applied **(Appendix B, Table S3.6; Dataset S5)**.

In our initial formulation of the algorithm for generating gene expression derived penalty weights, the weight of all reactions with no associated GPR was set to 1, since this is the implicit value of the coefficient for all reactions in a standard pFBA optimization. Since this runs the risk of introducing a systematic bias against using reactions that have associated GPRs, we attempted to counteract this risk by assigning all reactions lacking a GPR a penalty weight corresponding to the median penalty weight of all weighted reactions in the tissue in which those reactions are found. Comparing the results with and without the tissue-specific median penalty weights for reactions without GPRs, we see modest improvements in the weighted average errors from a scaling factor of 1 onwards when using the transcriptomic and proteomic datasets from Mergner et al., (2020) **(Fig. 3.2)**, though the effect is not large, indicating that our method is robust to including or omitting the tissue-specific median weight correction.

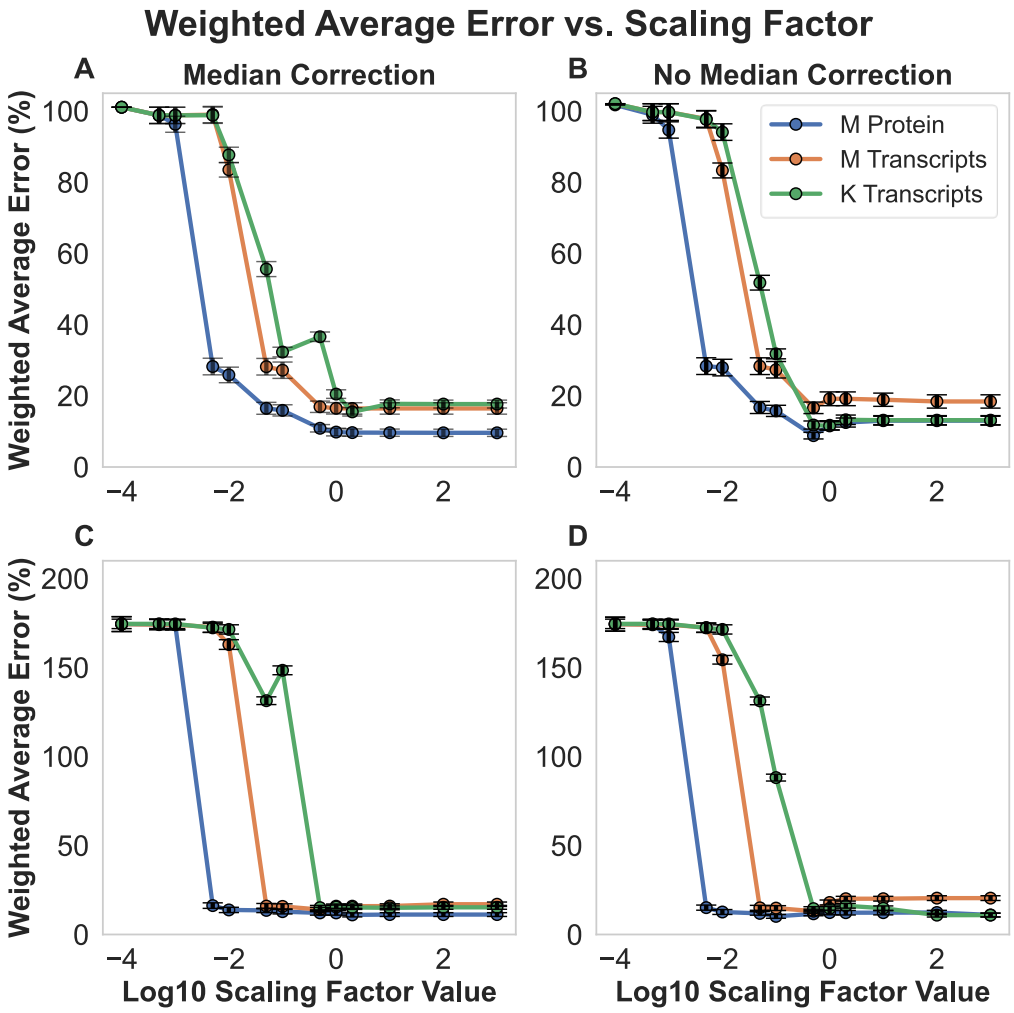


Figure 3.2: Weighted average errors of FBA predictions compared with MFA-estimated flux maps as a function of scaling factor value, light-level, and application of a tissue-specific median weight correction. Panels show weighted average errors of flux maps generated using **(A)** low-light constraints and a tissue-specific median correction applied, **(B)** low-light constraints and without a tissue-specific median correction applied, **(C)** high-light constraints and with a tissue-specific median correction applied, and **(D)** high-light constraints and without a tissue-specific median correction applied. "M Protein" and "M Transcripts" refer to flux maps generated using proteomic- and RNA-seq-derived weights from Mergner et al., (2020). "K Transcripts" refers to flux maps generated using RNA-seq derived weights from Klepikova et al., (2016). Upper and lower bars on each point represent the highest and lowest possible weighted average errors given FVA results, and the points themselves represent the average of these values.

3.5.4. Changes in the carbon and energy economy upon application of gene expression weights

In addition to improving quantitative agreement between the FBA-predicted and MFA-estimated flux maps, the gene expression weighting procedure also generates flux maps that present a substantially different picture of carbon and energy metabolism in *Arabidopsis* leaves.

In both high and low light FBA-predicted fluxes there is a substantial decrease in leaf mitochondrial Electron Transport Chain (ETC) activity and overall flux in mitochondria-localized reactions in the light relative to nighttime ETC activity and overall flux **(Appendix B, Table S3.3)**. MFA and other recent work further points to low TCA cycle fluxes in photosynthesizing leaves (Tcherkez et al., 2005; Xu et al., 2021a; Xu et al., 2022). This decrease in mitochondrial activity goes hand-in-hand with a predicted decrease in the use of unusually high fluxes related to proline metabolism to indirectly support the consumption of excess reductant produced via the light reactions of photosynthesis. Alongside this decrease in mitochondrial activity is a decrease in the ratio of cyclic electron flow (CEF) to linear electron flow (LEF) in the chloroplast **(Table 3.2)**. Although reliable empirical measurements of this CEF/LEF ratio are difficult to obtain, previous studies have shown that a C3 plant like *Arabidopsis* relying on cyclic electron flow to bring the ratio of ATP/NADPH produced up to that needed for normal growth would have a CEF amounting to ~13% of LEF (Kramer et al., 2004). Due to the presence of other balancing mechanisms, such as the malate valve (Selinski and Scheibe, 2019), this 13% value would represent an upper bound on stoichiometrically predicted values for CEF/LEF. Application of gene expression data decreases the CEF/LEF ratios in all but one FBA-predicted flux map to values much closer to the expected ~13% upper bound than are predicted using conventional pFBA **(Table 3.2)**.

Ma et al., (2014) reported MFA-derived estimates of %vpr, or the rate of photorespiratory CO_2 release via glyoxylate decarboxylation as a % of CO_2 assimilation, as well as the ratio of rubisco carboxylation flux to net CO_2 assimilation in the leaf. The unweighted flux predictions for the high and low light conditions disagree substantially with these estimates **(Table 3.2).** However, application of gene expression weights consistently brings estimates of these parameters into close agreement with MFA-derived values. The integration of gene expression also changes the predicted efficiency with which *Arabidopsis* converts atmospheric CO_2 into biomass **(Table 3.2)**. For comparison with these predicted efficiencies, we used the empirical *A.*

thaliana biomass, leaf area, and gas exchange data reported by Weraduwage et al., (2015) to calculate that approximately 56% of the net CO_2 assimilation in illuminated leaves ends up incorporated into biomass, which is closer to the value in our unweighted flux predictions than our weighted flux predictions, although it should be noted that these data were gathered from a hydroponic system.

3.6. Discussion

^{13}C-MFA is broadly accepted as being the most reliable method for estimating metabolic flux maps *in vivo* due to its ability to make use of substantial amounts of isotopic labeling data to arrive at well-supported flux maps in small- to medium-scale networks (Antoniewicz, 2015). However, the technique's utility is limited by the substantial experimental effort that goes into the generation of each individual flux map. FBA, with its requirement of much less experimental data, has become the method of choice for more exploratory or predictive metabolic modeling studies. The implicit assumption is usually that the predictions of FBA – or at least the range of its predictions in cases where a unique solution is not provided – agree with those we would arrive at if we were able to conduct a ^{13}C-MFA study. This makes our optimization procedures when performing FBA and validation of FBA models against MFA results of vital importance. The method presented here, by bringing FBA-predicted fluxes into line with MFA-estimates represents a step in the direction of higher-confidence FBA flux maps.

One limitation, as well as motivation, for the present study is the lack of a large set of ^{13}C-MFA datasets in plants and other multi-tissue eukaryotic systems. Systems like *E. coli* have multi-omic datasets consisting of transcriptomic, proteomic, and fluxomic measurements (Ishii et al., 2007) that have been utilized to empirically infer the relationship between gene expression and metabolic fluxes. This empirical training can then be used to more accurately predict fluxes in new contexts (Tian and Reed, 2018). The sparsity of ^{13}C-MFA data in more complex systems makes such an approach currently impossible.

A noteworthy theoretical aspect of the present approach is its simplicity, the only variable parameter being a single scaling factor that controls the magnitude of the penalty weights. That the assumption of a consistent value relating the relative abundances of transcripts or proteins in different tissues to the "preference" of an organism to partition flux among particular reactions can result in substantial improvement in error was of great interest in light of the complexity of the relationship between measures of gene expression – transcriptomic and proteomic

abundances – and flux. Particularly when making biotechnological interventions in a system to modify its metabolism, there is often an assumed strong linear relationship between transcription, translation, and, ultimately, metabolic flux, but the reality is rarely so simple. Although moderate correlations between transcript and protein abundances have been demonstrated across many systems, the degree of correlation varies across systems and experimental contexts (Maier et al., 2009; Liu et al., 2016). The correlation between these data types and rates of central metabolic reactions, which carry the large majority of total metabolic flux, is weaker still (Kuile and Westerhoff, 2001). Some previous studies found that changes in the gene expression related to individual reactions typically do not correlate well with changes in fluxes (Schwender et al., 2014; Tian and Reed, 2018), with some central metabolic fluxes in particular showing a negative correlation between changes in gene expression and flux. In both cases, gene expression data related to reactions were compared within the same cell type or tissue; in our study, we instead compare inter-tissue abundances, mirroring the long-standing practice in the literature of inferring relative metabolic activity in different tissues by their transcript and protein investment in relevant pathway steps. It may be that only by considering gene expression on an inter-tissue basis in the context of the entire complex stoichiometric network underlying metabolism can predictive gains from including gene expression evidence be properly realized.

Future work should aim to expand the number of available datasets, and the experimental conditions and genotypes for which they are gathered, in order to enable more thorough evaluation of methods like the one presented in this paper. Indeed, evaluating the presented method requires ^{13}C-MFA fluxes, multi-tissue omic data, and a genome-scale model all for the same biological system, which, to our knowledge, is only possible for *A. thaliana*. Building on the work of Ma et al., (2014), experimental improvements and refinements of the underlying network architecture of central carbon metabolism have been introduced in the context of ^{13}C-MFA in *Camelina sativa* (Xu et al., 2021a; Xu et al., 2022) and *Nicotiana tabacum* (Chu et al., 2022). In the present study the Ma et al. 2014 flux maps are used without change and we adopted a highly curated *A. thaliana* genome-scale model from which to construct the whole-plant model. This approach precluded the possibility of our reanalyzing the MFA-estimated flux map or biasing the construction of a purpose-built genome-scale model, making the MFA-to-FBA comparison more favorable. However, in future studies a combination of MFA network refinements, expanded datasets, and further improvements in the flux estimation procedures

holds promise for improving the fidelity of the ^{13}C-MFA comparison data. On the FBA side, the use of more detailed growth and composition measurements for FBA along with more detailed representation of different tissue types will potentially allow for more biologically accurate and representative FBA flux map predictions. These improvements in both MFA-estimation and FBA-prediction of flux maps, along with an expansion in the number of available ^{13}C-MFA datasets against which to compare FBA predictions, will allow for more extensive validation of the method described in this paper as well as other methods aiming to incorporate omic datasets into flux prediction.

A distinct aspect of the proposed method is its demonstrated ability to bring FBA-predicted fluxes in line with MFA-estimated fluxes across multiple input datasets, model architectures, and using multiple independent gene expression datasets. Our hope is that methods for incorporating transcriptomic and proteomic data may advance this field to the point where FBA-predicted flux maps can be used with high confidence for practical engineering goals. This, combined with the automated reconstruction of GEMs from genomic and biochemical databases (Saha et al., 2014) suggests a future with rapid turnaround from the initial identification of an organism of interest to metabolic flux predictions and rational genetic engineering to achieve biotechnological aims.

3.7 Acknowledgments

We would like to thank Dr. Doug Allen for permission to adapt Figure 3 from Ma et al., (2014) for use in Figure 3.1 in this publication. Figure 3.1 was created with BioRender.com.

3.8 Funding

This research was supported by the Office of Science (BER), U.S. Department of Energy, Grant no DE-SC0018269 (J.A.M.K. and Y.S-H.). This work is supported, in part, by the NSF Research Traineeship Program (Grant DGE-1828149) to J.A.M.K. This publication was also made possible by a predoctoral training award to J.A.M.K. from Grant T32-GM110523 from National Institute of General Medical Sciences (NIGMS) of the NIH. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIGMS or NIH.

REFERENCES

Antoniewicz MR (2015) Methods and advances in metabolic flux analysis: a mini-review. Journal of Industrial Microbiology and Biotechnology **42**: 317–325

Arnold A, Nikoloski Z (2014) Bottom-up metabolic reconstruction of arabidopsis and its application to determining the metabolic costs of enzyme production. Plant Physiology **165**: 1380–1391

Boyle NR, Sengupta N, Morgan JA (2017) Metabolic flux analysis of heterotrophic growth in Chlamydomonas reinhardtii. PLoS ONE **12**: 1–23

Burgard AP, Pharkya P, Maranas CD (2003) OptKnock: A Bilevel Programming Framework for Identifying Gene Knockout Strategies for Microbial Strain Optimization. Biotechnology and Bioengineering **84**: 647–657

Chen X, Alonso AP, Allen DK, Reed JL, Shachar-Hill Y (2011) Synergy between 13C-metabolic flux analysis and flux balance analysis for understanding metabolic adaption to anaerobiosis in E. coli. Metabolic Engineering **13**: 38–48

Cheung CYM, Ratcliffe RG, Sweetlove LJ (2015) A Method of Accounting for Enzyme Costs in Flux Balance Analysis Reveals Alternative Pathways and Metabolite Stores in an Illuminated Arabidopsis Leaf. Plant physiology **169**: 1671–82

Cheung CYM, Williams TCR, Poolman MG, Fell DA, Ratcliffe RG, Sweetlove LJ (2013) A method for accounting for maintenance costs in flux balance analysis improves the prediction of plant cell metabolic phenotypes under stress conditions. Plant Journal **75**: 1050–1061

Chu KL, Koley S, Jenkins LM, Bailey SR, Kambhampati S, Foley K, Arp JJ, Morley SA, Czymmek KJ, Bates PD, et al (2022) Metabolic flux analysis of the non-transitory starch tradeoff for lipid production in mature tobacco leaves. Metabolic Engineering **69**: 231–248

Dal'Molin CG de O, Quek LE, Palfreyman RW, Brumbley SM, Nielsen LK (2010a) AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. Plant Physiology **152**: 579–589

Dal'Molin CG de O, Quek LE, Palfreyman RW, Brumbley SM, Nielsen LK (2010b) C4GEM, a genome-scale metabolic model to study C4 plant metabolism. Plant Physiology **154**: 1871–1885

Gleizer S, Ben-Nissan R, Bar-On YM, Antonovsky N, Noor E, Zohar Y, Jona G, Krieger E, Shamshoum M, Bar-Even A, et al (2019) Conversion of Escherichia coli to Generate All Biomass Carbon from CO2. Cell **179**: 1255-1263.e12

Gurobi Optimization, LLC (2021) Gurobi Optimizer Reference Manual.

Hay J, Schwender J (2011) Metabolic network reconstruction and flux variability analysis of storage synthesis in developing oilseed rape (Brassica napus L.) embryos. Plant Journal **67**: 526–541

Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, Haraldsdóttir HS, Wachowiak J, Keating SM, Vlasov V, et al (2019) Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nature Protocols **14**: 639–702

Ishii N, Nakahigashi K, Baba T, Robert M, Soga T, Kanai A, Hirasawa T, Naba M, Hirai K, Hoque A, et al (2007) Multiple high-throughput analyses monitor the response of E. coli to perturbations. Science **316**: 593–597

Jenior ML, Moutinho TJ, Dougherty BV, Papin JA (2020) Transcriptome-guided parsimonious flux analysis improves predictions with metabolic networks in complex environments. PLoS Computational Biology. doi: 10.1371/journal.pcbi.1007099

Klepikova AV, Kasianov AS, Gerasimov ES, Logacheva MD, Penin AA (2016) A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. Plant Journal **88**: 1058–1070

Kramer DM, Avenson TJ, Edwards GE (2004) Dynamic flexibility in the light reactions of photosynthesis governed by both electron and proton transfer reactions. Trends in Plant Science **9**: 349–357

Krey JF, Wilmarth PA, Shin J-B, Klimek J, Sherman NE, Jeffery ED, Choi D, David LL, Barr-Gillespie PG (2014) Accurate label-free protein quantitation with high- and low-resolution mass spectrometers. Journal of proteome research **13**: 1034–1044

Kuile BH, Westerhoff HV (2001) Transcriptome meets metabolome: Hierarchical and metabolic regulation of the glycolytic pathway. FEBS Letters **500**: 169–171

Lee KH, Park JH, Kim TY, Kim HU, Lee SY (2007) Systems metabolic engineering of Escherichia coli for L -threonine production. Molecular Systems Biology. doi: 10.1038/msb4100196

Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, Adkins JN, Schramm G, Purvine SO, Lopez-Ferrer D, et al (2010) Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Molecular Systems Biology. doi: 10.1038/msb.2010.47

Liu Y, Beyer A, Aebersold R (2016) On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell **165**: 535–550

Ma F, Jazmin LJ, Young JD, Allen DK (2014a) Isotopically nonstationary 13C flux analysis of changes in Arabidopsis thaliana leaf metabolism due to high light acclimation. Proceedings of the National Academy of Sciences of the United States of America **111**: 16967–16972

Ma F, Jazmin LJ, Young JD, Allen DK (2014b) Isotopically nonstationary 13C flux analysis of changes in Arabidopsis thaliana leaf metabolism due to high light acclimation. Proceedings of the National Academy of Sciences of the United States of America **111**: 16967–16972

Machado D, Herrgård M (2014) Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. PLoS Computational Biology. doi: 10.1371/journal.pcbi.1003580

Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metabolic Engineering **5**: 264–276

Maier T, Güell M, Serrano L (2009) Correlation of mRNA and protein in complex biological samples. FEBS Letters **583**: 3966–3973

Mergner J, Frejno M, List M, Papacek M, Chen X, Chaudhary A, Samaras P, Richter S, Shikata H, Messerer M, et al (2020) Mass-spectrometry-based draft of the Arabidopsis proteome. Nature **579**: 409–414

de Oliveira Dal'Molin CG, Quek LE, Saa PA, Nielsen LK (2015) A multi-tissue genome-scale metabolic modeling framework for the analysis of whole plant systems. Frontiers in Plant Science **6**: 1–12

Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? Nature Biotechnology **28**: 245–248

Pandey V, Hadadi N, Hatzimanikatis V (2019) Enhanced flux prediction by integrating relative expression and relative metabolite abundance into thermodynamically consistent metabolic models. PLOS Computational Biology **15**: 1–23

Park JH, Lee KH, Kim TY, Lee SY (2007) Metabolic engineering of Escherichia coli for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation. Proceedings of the National Academy of Sciences of the United States of America **104**: 7797–7802

Poolman MG, Miguet L, Sweetlove LJ, Fell DA (2009) A genome-scale metabolic model of Arabidopsis and some of its properties. Plant Physiology **151**: 1570–1581

Ratcliffe RG, Shachar-Hill Y (2006) Measuring multiple fluxes through plant metabolic networks. Plant Journal **45**: 490–511

Ravi S, Gunawan R (2021) ΔFBA—Predicting metabolic flux alterations using genome-scale metabolic models and differential transcriptomic data. PLOS Computational Biology **17**: e1009589

Saha R, Chowdhury A, Maranas CD (2014) Recent advances in the reconstruction of metabolic models and integration of omics data. Current Opinion in Biotechnology **29**: 39–45

Saha R, Suthers PF, Maranas CD (2011) Zea mays irs1563: A comprehensive genome-scale metabolic reconstruction of maize metabolism. PLoS ONE. doi: 10.1371/journal.pone.0021784

Schwender J, König C, Klapperstück M, Heinzel N, Munz E, Hebbelmann I, Hay JO, Denolf P, De Bodt S, Redestig H, et al (2014) Transcript abundance on its own cannot be used to infer fluxes in central metabolism. Frontiers in Plant Science **5**: 1–16

Selinski J, Scheibe R (2019) Malate valves: old shuttles with new perspectives. Plant Biology **21**: 21–30

Shaw R, Cheung CYM (2020) Multi-tissue to whole plant metabolic modelling. Cellular and Molecular Life Sciences **77**: 489–495

Szecowka M, Heise R, Tohge T, Nunes-Nesi A, Vosloh D, Huege J, Feil R, Lunn J, Nikoloski Z, Stitt M, et al (2013) Metabolic fluxes in an illuminated Arabidopsis rosette. Plant Cell **25**: 694–714

Tcherkez G, Cornic G, Bligny R, Gout E, Ghashghaie J (2005) In vivo respiratory metabolism of illuminated leaves. Plant Physiology **138**: 1596–1606

Tepper N, Shlomi T (2009) Predicting metabolic engineering knockout strategies for chemical production: Accounting for competing pathways. Bioinformatics **26**: 536–543

Tian M, Reed JL (2018) Integrating proteomic or transcriptomic data into metabolic models using linear bound flux balance analysis. Bioinformatics **34**: 3882–3888

Vijayakumar S, Conway M, Lio P, Angione C (2017) Seeing the wood for the trees: A forest of methods for optimization and omic-network integration in metabolic modelling. Briefings in Bioinformatics **19**: 1218–1235

Weraduwage SM, Chen J, Anozie FC, Morales A, Weise SE, Sharkey TD (2015) The relationship between leaf area growth and biomass accumulation in Arabidopsis thaliana. **6**: 1–21

Williams TCR, Poolman MG, Howden AJM, Schwarzlander M, Fell DA, Ratcliffe RG, Sweetlove LJ (2010) A Genome-scale metabolic model accurately predicts fluxes in central carbon metabolism under stress conditions. Plant Physiology **154**: 311–323

Xu Y, Fu X, Sharkey TD, Shachar-Hill Y, Walker and BJ (2021) The metabolic origins of non-photorespiratory CO2 release during photosynthesis: a metabolic flux analysis. Plant Physiology 1–18

Xu Y, Wieloch T, Kaste JAM, Shachar-Hill Y, Sharkey TD (2022) Reimport of carbon from cytosolic and vacuolar sugar pools into the Calvin-Benson cycle explains photosynthesis labeling anomalies. Proceedings of the National Academy of Sciences **119**: e2121531119

APPENDIX B: Supplemental Material for Chapter 3

SUPPLEMENTARY METHODS

Datasets and Omic Data Processing

Sample IDs and SRR numbers for all transcriptomic and proteomic datasets used in this study can be found in Appendix B, **Table S3.5**.

The raw RNA-seq data for all 137 samples in Klepikova et al., (2016) was trimmed using the *fastp* algorithm (Chen et al., 2018) and then aligned to the TAIR10 *Arabidopsis thaliana* genome obtained from ensembl plants using the *salmon* algorithm (Lamesch et al., 2012; Patro et al., 2017; Howe et al., 2020). RNA-seq reads from Mergner et al., (2020) were taken directly from the published supplemental material. Library normalization was performed on the RNA-seq datasets from both Klepikova et al., (2016) and Mergner et al., (2020) using the DeSeq2 procedure (Love et al., 2014) and averages of the normalized transcript abundance values across replicates from Klepikova et al., (2016) were used. Intensity-based Absolute Quantification (IBAQ) (Schwanhüusser et al., 2011) values for the proteomic data from Mergner et al., (2020) were divided by the sum total intensity across all protein signals measured for a given sample to normalize them.

Error Evaluation and Statistical Analysis

To evaluate whether the error values for measured reactions in individual flux maps generated using gene expression weights were statistically significantly different from the errors without application of these weights, the Wilcoxon signed-rank test was used (Wilcoxon, 1992). The Bonferroni-Holm multiple testing correction (Holm, 1978) was used to correct the family-wise α of all hypothesis tests to 0.01, where each hypothesis test is asking, by the Wilcoxon signed-rank test, whether the differences between a given FBA-predicted flux map (e.g., the flux map generated using protein-derived gene expression weights and a Scaling Factor of 1) and our MFA-estimated flux map could be attributed to random chance. In the high light condition ^{13}C-MFA flux map from Ma et al., (2014b), the flux through the malate dehydrogenase reaction was reported as exactly 0 – as this made its error undefined, it was excluded from the high light condition's error calculation. Fluxes carrying zero flux were likewise excluded from error calculations when comparing FBA predictions against fluxes from the Szecowka et al., (2013) dataset.

Flux Variability Analysis

Flux Variability Analysis (FVA) (Mahadevan and Schilling, 2003) was used to determine the maximum and minimum values possible for each of the fluxes included in the error calculation, subject to the following constraint:

$$c \cdot v = opt \quad (1)$$

Where *c* is the vector of all weight coefficients, *v* is the vector of all fluxes, and *opt* is the value of the objective function determined by the initial optimization procedure. As shown in Appendix B, Dataset S3.3**,** some of the fluxes included in the error function are not uniquely defined, such that they can vary up and down without violating **Eq. 1**. To account for this variation, maximum and minimum weighted average errors were calculated, where the minimum and maximum errors correspond to the smallest and largest weighted average errors possible for a flux map given the maximum/minimum values for all evaluated fluxes. FVA was performed in MATLAB using the COBRA Toolbox (Heirendt et al., 2019) and Gurobi™ 8.1.1 (Gurobi Optimization, LLC, 2021). Note that we encountered infeasible solutions in some cases when using a scaling factor of 1000 and omic datasets with the leaf and root data swapped – in such cases, the corresponding columns of the FVA results have been left blank.

Model Constraints

Light uptake and photosynthetic activity were restricted to the leaf tissue and mineral uptake was restricted to root tissue. Inter-tissue transport and day/night continuity of metabolites were defined and constrained as in Cheung & Shaw 2018 (Shaw and Cheung, 2018) as were ATP and NADPH maintenance flux values. Biomass compositions of leaf, stem, and root were taken from de Oliveira Dal'Molin et al., (2015), based on Poorter and Bergkotte, (1992). Reactions were added to produce biomass components that appear in the de Oliveira Dal'Molin et al., (2015) biomass equations but not in the core metabolic model (Arnold and Nikoloski, 2014); this involved adding subnetworks of missing reactions for several components and single summary reactions for others. Cytosolic pentose phosphate pathway reactions were also added to the model. All reactions were converted to irreversible form, wherein all reversible reactions were converted to independent forward and reverse reactions, prior to solving. This is simply to ensure that all fluxes, including those representing the reverse flux of a reversible reaction, take values that are zero or positive.

In order to generate predictions corresponding to the high-light and low-light flux maps

reported by Ma et al., (2014), the v_o/v_c, or ratio of ribulose 1,5-bisphosphate carboxylase/oxygenase (rubisco) oxygenation activity to its carboxylation activity, and the ratio of starch to sucrose synthesis were both constrained to the values estimated in that study. v_o/v_c and starch to sucrose synthesis values were likewise constrained to the values estimated by Szecowka et al., (2013) when comparing FBA fluxes against that study.

TABLES

Table S3.1: Reductions in weighted average error with application of gene expression weights derived from gene expression data with incorrect tissue specification or randomized values.

Dataset	Light Level	Weighted Average Error (%)	
		Without Gene Expression Weights	**With Leaf/Root Flipped Gene Expression Weights**
Mergner et al. Transcriptome	High	168 – 180	181 - 215%
	Low	93.8 – 103	155 - 185
Mergner et al. Proteome	High	168 – 180	249 - 295
	Low	93.8 – 103	131 - 160
Klepikova et al. Transcriptome	High	168 – 180	87.9 - 109
	Low	93.8 – 103	97.2 – 120

[a]Weighted average errors are calculated from flux maps generated using a scaling factor of 1.

Table S3.2: Reductions in weighted average errors with an alternate model architecture allowing free uptake and discharge of protons from the root tissue compartment.

Dataset	Light Level	Weighted Average Error (%)	
		Without Gene Expression Weights (%)	With Gene Expression Weights (%)[a]
Mergner et al. Transcriptome	High	127 - 135	11.8 – 14.2
	Low	66.1 - 73.8	16.8 – 19.4
Mergner et al. Proteome	High	127 - 135	10.5 - 12.8
	Low	66.1 - 73.8	8.88 - 11.1
Klepikova et al. Transcriptome	High	127 - 135	13.7 - 16.5
	Low	66.1 - 73.8	20.8 - 23.2

[a]Weighted average errors are calculated from flux maps generated using a scaling factor of 1.

Table S3.3: Ratios of day vs. night leaf mitochondrial fluxes and Electron Transport Chain fluxes in flux maps with and without integration of gene expression evidence.

Dataset	Light Level	Ratio of total leaf mitochondrial flux in day vs. night		Ratio of leaf mitochondrial ATP synthase flux in day vs. night	
		Without Gene Expression Weights	With Gene Expression Weights[a]	Without Gene Expression Weights	With Gene Expression Weights[s]
Mergner et al. Transcriptome	High	1.17	0.144	1.15	0.0888
	Low	1.23	0.0625	1.20	1.94 * 10^-5
Mergner et al. Proteome	High	1.17	0.0693	1.15	6.17 * 10^-5
	Low	1.23	0.0981	1.20	1.87 * 10^-5
Klepikova et al. Transcriptome	High	1.16	0.179	1.15	0.121
	Low	1.23	0.615	1.20	0.503

[a]Values for weighted cases calculated from flux maps generated using a scaling factor of 1.

Table S3.4: Growth conditions and key constraints from transcriptomic, proteomic, ^{13}C-MFA, and kinetic flux profiling datasets used in the present study.

Dataset	Type	Growth conditions	Tissue	Age	v_o/v_c	Starch/sucrose biosynthesis rate
Klepikova et al. (Klepikova et al., 2016)	Transcriptomic	Philips Master TL5 HO 54 W/840 lamps light source at a 27cm distance; 22°C; 50% relative humidity; 16/8-h day/night cycle	Leaf	Anthesis of first flower	N/A	N/A
			Stem	Anthesis of first flower	N/A	N/A
			Root	7th day after germination	N/A	N/A
Mergner et al. (Mergner et al., 2020)	Transcriptomic	Continuous white light; 22°C	Leaf	22 days old	N/A	N/A
			Stem	30 days old	N/A	N/A
			Root	22 days old	N/A	N/A
	Proteomic	Continuous white light; 22°C	Leaf	22 days old	N/A	N/A
			Stem	30 days old	N/A	N/A
			Root	22 days old	N/A	N/A
Ma et al. Low Light Conditions (Ma et al., 2014)	^{13}C-MFA	200 µmol m^{-2} s^{-1}; 22/18°C; 50% relative humidity; 16/8-h day/night cycles	Leaf	28 days old	0.29	0.26
Ma et al. High Light Condition (Ma et al., 2014)	^{13}C-MFA	500 µmol m^{-2} s^{-1}; 22/18°C; 50% relative humidity; 16/8-h day/night cycles	Leaf	28 days old	0.43	0.16
Szecowka et al. (Szecowka et al., 2013)	Kinetic flux profiling	120 µmol m^{-2} s^{-1} irradiance; 22/20°C; 50% relative humidity; 8/16-h day/night cycles	Leaf	35 days old	0.4	0.45

Table S3.5: Reductions in weighted average errors when using constraints from the Szecowka et al. 2013 dataset and gene expression weights from different sources.

Gene expression dataset	Weighted Average Error (%)	
	Without Gene Expression Weights (%)	With Gene Expression Weights (%)[a]
Mergner et al. Transcriptome	107.6 – 107.8	6.09
Mergner et al. Proteome	107.6 – 107.8	7.55
Klepikova et al. Transcriptome	107.6 – 107.8	8.65

REFERENCES

Arnold A, Nikoloski Z (2014) Bottom-up metabolic reconstruction of arabidopsis and its application to determining the metabolic costs of enzyme production. Plant Physiology **165**: 1380–1391

Chen S, Zhou Y, Chen Y, Gu J (2018) Fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics **34**: i884–i890

Gurobi Optimization, LLC (2021) Gurobi Optimizer Reference Manual.

Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, Haraldsdóttir HS, Wachowiak J, Keating SM, Vlasov V, et al (2019) Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nature Protocols **14**: 639–702

Holm S (1978) Board of the Foundation of the Scandinavian Journal of Statistics A Simple Sequentially Rejective Multiple Test Procedure Author (s): Sture Holm Published by : Wiley on behalf of Board of the Foundation of the Scandinavian Journal of Statistics Stable U. Scandinavian Journal of Statistics **6**: 65–70

Howe KL, Contreras-Moreira B, De Silva N, Maslen G, Akanni W, Allen J, Alvarez-Jarreta J, Barba M, Bolser DM, Cambell L, et al (2020) Ensembl Genomes 2020-enabling non-vertebrate genomic research. Nucleic Acids Research **48**: D689–D695

Klepikova AV, Kasianov AS, Gerasimov ES, Logacheva MD, Penin AA (2016) A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. Plant Journal **88**: 1058–1070

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al (2012) The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. Nucleic Acids Research **40**: 1202–1210

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology **15**: 1–21

Ma F, Jazmin LJ, Young JD, Allen DK (2014a) Isotopically nonstationary 13C flux analysis of changes in Arabidopsis thaliana leaf metabolism due to high light acclimation. Proceedings of the National Academy of Sciences of the United States of America **111**: 16967–16972

Ma F, Jazmin LJ, Young JD, Allen DK (2014b) Isotopically nonstationary 13C flux analysis of changes in Arabidopsis thaliana leaf metabolism due to high light acclimation. Proceedings of the National Academy of Sciences of the United States of America **111**: 16967–16972

Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-

based genome-scale metabolic models. Metabolic Engineering **5**: 264–276

Mergner J, Frejno M, List M, Papacek M, Chen X, Chaudhary A, Samaras P, Richter S, Shikata H, Messerer M, et al (2020) Mass-spectrometry-based draft of the Arabidopsis proteome. Nature **579**: 409–414

de Oliveira Dal'Molin CG, Quek LE, Saa PA, Nielsen LK (2015) A multi-tissue genome-scale metabolic modeling framework for the analysis of whole plant systems. Frontiers in Plant Science **6**: 1–12

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nature methods **14**: 417–419

Poorter H, Bergkotte M (1992) Chemical composition of 24 wild species differing in relative growth rate. Plant, Cell & Environment **15**: 221–229

Schwanhüusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M (2011) Global quantification of mammalian gene expression control. Nature **473**: 337–342

Shaw R, Cheung CYM (2018) A dynamic multi-tissue flux balance model captures carbon and nitrogen metabolism and optimal resource partitioning during arabidopsis growth. Frontiers in Plant Science **9**: 1–15

Szecowka M, Heise R, Tohge T, Nunes-Nesi A, Vosloh D, Huege J, Feil R, Lunn J, Nikoloski Z, Stitt M, et al (2013) Metabolic fluxes in an illuminated Arabidopsis rosette. Plant Cell **25**: 694–714

Wilcoxon F (1992) Individual comparisons by ranking methods. Breakthroughs in statistics. Springer, pp 196–202

Chapter 4

Biophysical carbon concentrating mechanisms in land plants: insights from reaction-diffusion modeling

A preprint of this study is available:

J. A. M. Kaste, B.J. Walker, Y. Shachar-Hill. Biophysical carbon concentrating mechanisms in land plants: insights from reaction-diffusion modeling. *bioRxiv* (2023). doi: https://doi.org/10.1101/2024.01.04.574220

4.1. Preface

The project described in this chapter was born out of conversations with Anne Steensma, a fellow graduate student in the Shachar-Hill laboratory, and Dr. Berkley Walker. Anne was interested in using metabolic modeling as a way of exploring a hypothetical setup for a Carbon-Concentrating Mechanism (CCM) in the red alga *Cyanidioschyzon merolae*. I provided substantial assistance in the early stages of the project, including writing the code for the initial versions of the model architecture, coming up with a computational approach for estimating CO_2 compensation points *in silico*, and setting up the code for distribution to MSU's High-Performance Computing Cluster (HPCC) for parameter exploration and analysis. The analysis for this project is proceeding steadily and I plan on writing it up as a manuscript before the end of the year, on which I am anticipated to be co-first author. The conversations our group was having about modeling biophysical CCMs led us to asking some broader questions about the efficiency of such mechanisms in land plants, such as:

1. Previous modeling studies make the addition of a carboxysome to a land plant seem very energetically favorable, but is this finding robust?
2. Why do land plants not pump any bicarbonate from apoplastic water when there is a substantial concentration of bicarbonate available?
3. Is there something about the physiology and/or morphology of hornworts that has caused them to repeatedly evolve a pyrenoid-based biophysical CCM when such biophysical CCMs are entirely absent in all other land plant lineages?

I took up answering these questions by building spatially-resolved reaction-diffusion models of inorganic carbon and O_2 movement in land plant and algal systems using the *Virtual Cell* platform. The results of this analysis have been written up as a manuscript that has been deposited as a preprint and is currently under peer review.

4.2. Abstract

Carbon Concentrating Mechanisms (CCMs) have evolved numerous times in photosynthetic organisms. They elevate the concentration of CO_2 around the carbon-fixing enzyme rubisco, thereby increasing CO_2 assimilatory flux and reducing photorespiration. Biophysical CCMs, like the pyrenoid-based CCM of *Chlamydomonas reinhardtii* or carboxysome systems of cyanobacteria, are common in aquatic photosynthetic microbes, but in land plants appear only among the hornworts. To predict the likely efficiency of biophysical CCMs in C3 plants, we used

spatially resolved reaction-diffusion models to predict rubisco saturation and light use efficiency. We find that the energy efficiency of adding individual CCM components to a C3 land plant is highly dependent on the permeability of lipid membranes to CO_2, with values in the range reported in the literature that are higher than used in previous modeling studies resulting in low light use efficiency. Adding a complete pyrenoid-based CCM into the leaf cells of a C3 land plant is predicted to boost net CO_2 fixation, but at higher energetic costs than those incurred by photorespiratory losses without a CCM. Two notable exceptions are when substomatal CO_2 levels are as low as those found in land plants that already employ biochemical CCMs and when gas exchange is limited such as with hornworts, making the use of a biophysical CCM necessary to achieve net positive CO_2 fixation under atmospheric CO_2 levels. This provides an explanation for the uniqueness of hornworts' CCM among land plants and evolution of pyrenoids multiple times.

4.3. Introduction

Ribulose-1,5-bisphosphate carboxylase/oxygenase (rubisco) catalyzes the fixation of CO_2 as part of the Calvin-Benson Cycle (CBC) but is also capable of fixing O_2. The fixation of O_2 results in the formation of 2-phosphoglycolate (2PG), with the photorespiratory pathway being necessary to detoxify and recover the carbon in 2PG and recycle it back into the CBC. Although rubisco shows selectivity for CO_2 relative to O_2, significant photorespiratory flux still occurs in photosynthetic systems due to the much higher partial pressure of O_2 in the earth's atmosphere relative to CO_2. Photorespiratory flux lowers net carbon assimilation and incurs substantial energetic costs, in the form of ATP, redox equivalents, and ultimately photons. Although the costs associated with photorespiration vary between plant species and environmental conditions, it has been estimated that photorespiration accounts for crop yield decreases of 20 and 36% for soybean and wheat respectively under current climate conditions (Walker et al., 2016).

Carbon Concentrating Mechanisms (CCMs) increase the concentration of CO_2 around rubisco, competitively inhibiting the oxygenation reaction, suppressing photorespiration, and increasing carboxylation flux (Raven et al., 2017). In biochemical CCMs, such as C4 and CAM photosynthesis, inorganic carbon is fixed into an intermediate form of organic carbon, before eventually being released around rubisco (Ludwig, 2013; Bräutigam et al., 2017). Biophysical or "inorganic" CCMs, on the other hand, do not rely on any additional intermediate organic carbon species, but instead use pumps, diffusional barriers, carbonic anhydrases, and pH differences

between cellular compartments to increase the CO_2 concentration near rubisco (Raven et al., 2008). Such CCMs are common in cyanobacteria and algae (Raven et al., 2008), but are conspicuously absent in C3 plants, including almost all land plants. This has motivated researchers to look into the possibility of introducing a CCM, either in its entirety or individual components, into these plants to improve carbon fixation, reduce photorespiratory CO_2 and energy losses, and ultimately boost yields (Ermakova et al., 2020; Hennacy and Jonikas, 2020).

The seemingly substantial benefits of CCMs raise the question of why they are not already more widespread in land plants. Despite their lack of a CCM, C3 plants are still the most abundant group of land plants in terms of vegetation coverage and gross photosynthetic productivity (Still et al., 2003; Raven et al., 2017). In the case of C4 photosynthesis, the large number of anatomical and biochemical features required has been invoked as a reason why, rather than being universally adopted in land plants, it has instead repeatedly evolved only in lineages exposed to the kinds of hot, arid conditions that limit water availability and exacerbate the losses associated with photorespiration (Sage et al., 2018). However, such an explanation is less satisfactory in the case of biophysical CCMs because they are present in the hornworts. It also raises the question of why biophysical CCMs are uniformly absent in all land plant lineages except for the hornworts (Villarreal and Renner, 2012).

Have inefficiencies associated with biophysical CCMs precluded their successful emergence in C3 plants and can we examine the presence and absence of these biophysical CCMs in different groups of organisms using these inefficiencies? The efficiency of intermediate photosynthetic configurations, featuring some but not all of the essential parts of a CCM, may also represent a barrier to the emergence of CCMs in land plant lineages. Anatomical and life history details of hornworts may explain why, among the land plants, only hornworts have evolved pyrenoid-based biophysical CCMs (PCCMs), and have done so repeatedly (Villarreal and Renner, 2012). The poikilohydric life history of hornworts makes it necessary for them to have highly desiccation-tolerant cell walls which, together with bryophytes' generally thicker cell walls (Flexas et al., 2021) and hornworts' simpler tissue architecture, may explain their extremely low gas conductance (Meyer et al., 2008; Carriquí et al., 2019). We hypothesized that the distinct morphologic characteristics and habitat of hornworts may explain why they, uniquely among the land plants, evolved biophysical CCMs. It is possible that the different paths that inorganic carbon has to take from the environment into a C3 land plant cell versus an algal cell

can similarly explain why the former never uses pyrenoids to concentrate carbon and the latter frequently does.

A closer examination of the costs of a CCM may also inform the viability and strategy of biotechnological projects focused on introducing them to C3 crops. Prior quantitative modeling work argues that incorporating individual CCM components – in particular, bicarbonate transporters at the chloroplast membrane – and entire CCMs into land plant systems may boost net CO_2 fixation as well as improve the efficiency of photosynthetic carbon assimilation by reducing the energetic costs associated with photorespiration (McGrath and Long, 2014; Fei et al., 2022). Similar arguments have been made in favor of engineering biochemical – e.g. C4 – photosynthesis into C3 plants (Walker et al., 2016). These models represent sophisticated, integrative descriptions of photosynthetic carbon assimilation. For the purposes of the questions we are interested in, however, we needed models of both land plant and algal systems that represent photo-assimilatory processes at the whole-cell level. We also needed models that allow us to explore substantial uncertainties in certain key parameters, and that include energy costs associated with carbonic anhydrase (CA) activity in the thylakoid lumen.

Here we developed spatially-resolved reaction-diffusion models of land plants and green algae with and without PCCMs in the *Virtual Cell* platform (Schaff et al., 1997; Cowan et al., 2012). These models represent, to our knowledge, the first such models of C3 land plants containing pyrenoid-based biophysical CCMs, as well as the first models of algal systems containing biophysical CCMs going beyond the scale of the chloroplast and including the whole cell in an aqueous environment. We highlight the substantial uncertainty in reported or predicted values of the permeability of lipid membranes to CO_2 and explore how this uncertainty can give rise to qualitatively different conclusions as to the efficiency and effectiveness of adding chloroplast envelope bicarbonate pumps in particular. Finally, we find that despite the near-ubiquity of biophysical CCMs in algae, modeling suggests that lower levels of external inorganic carbon (DIC) are needed to make CCMs energetically favorable for land plants.

4.4. Methods

4.4.1. Model details

Spatially-resolved reaction-diffusion models of carbon assimilation were developed in the *Virtual Cell* platform, a software suite that allows for the creation and analysis of chemical reaction diffusion dynamics in the context of 3D models (Schaff et al., 1997; Cowan et al.,

2012). Baseline parameters for simulations can be found in **Table 4.1** and diagrams of two of the models used in this study, showing the representative features of the land plant and algal models, as well as the differences between the with- and without-PCCM models, can be seen in **Figure 4.1.**

Systems were represented as spatially symmetrical, with spherical concentric compartments that were converted into volumetric pixels (voxels) according to the simulations' spatial resolution. All results presented are from simulations containing either 9,261 voxels or 12,167 voxels. Due to the large parameter explorations done in this study, minor geometrical modifications were made to make efficient numerical simulation feasible. Specifically, the radius of the apoplast water layer in the land plant models was extended out from the 9.41um it should be based on a cell wall thickness of 0.32μm plus an apoplast water layer of equivalent thickness to 10μm. We also modeled the thylakoid tubules of with-PCCM models as a set of six cylinders of radius 0.5μm extending into the pyrenoid, with exchange between the tubules and the pyrenoid occurring at the end of these cylinders, in contrast to the larger number of finer tubules used in Fei et al., (2022).

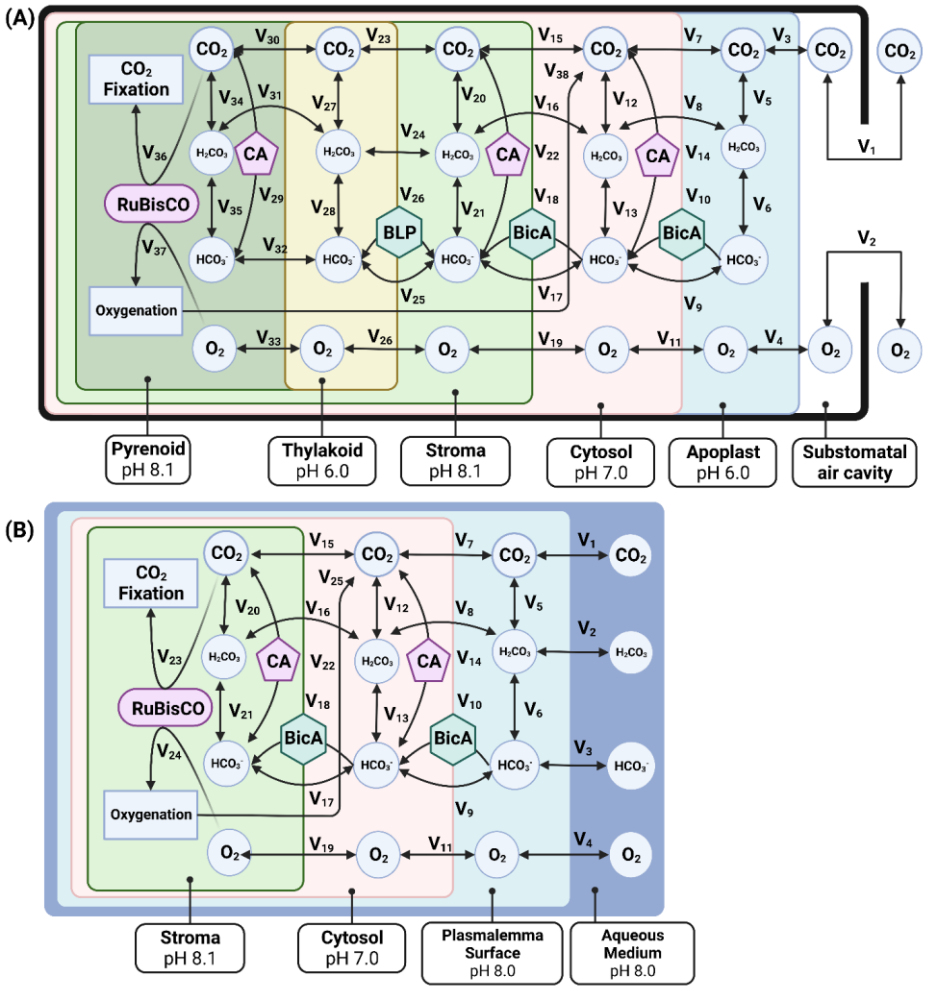


Figure 4.1: Diagrammatic representations of (A) a model of photosynthetic carbon assimilation in a land plant mesophyll cell containing a *C. reinhardtii* style PCCM, and (B) a model of an algal cell that does not contain a pyrenoid. CA refers to carbonic anhydrase, BLP refers to bestrophin-like proteins that serve as membrane channels for passive bicarbonate transport, and BicA is a cyanobacterial active bicarbonate transporter. In the VCell implementation of the model, some strongly linked steps are combined for the sake of numerical computability. Exact specifications for all flux equations used can be found in the publicly shared model implementations in VCell (see code and data availability statement). Note that for the sake of numerical tractability, the carbonic-anhydrase catalyzed interconversion of CO_2 and HCO_3 in the thylakoid in models featuring a CCM (v29) is localized to the pyrenoid but uses the pH value of the thylakoid; in the real biological system, the carbonic-anhydrase is inside the thylakoid tubules that penetrate into the pyrenoid.

Table 4.1: Model parameter definitions with source references and, where applicable with notes on derivation. When parameters were derived from parameterization of a previous modeling study, both the modeling study and the original literature reference for the parameter are cited. References in "Ref." column: (1) Mazarei & Sandall 1980; (2) Fei et al. 2022; (3) Xiang & Anderson 1994; (4) Walker, Smith & Cathers 1980; (5) Bentley & Pittman 1997; (6) Gutknecht, Bisson & Tosteson 1977; (7) Missner et al. 2008; (8) Hopkinson et al. 2011; (9) Widomska, Raguz & Subczynski 2007; (10) Mitchell et al. 2010; (11) Larsson et al. 1997; (12) Pocker & Ng 1973; (13) Pocker & Miksch 1978; (14) McGrath & Long 2014; (15) Bernacchi et al. 2005; (16) Badger & Andrews 1974; (17) Farquhar, von Caemmerer & Berry 1980; (18) von Caemmerer 2000; (19) Price et al. 2004; (20) Bernacchi et al. 2006; (21) Kump 2008; (22) Pritchard, Grout & Short 1986; (23) Flexas et al. 2021; (24) Ouk, Oi & Taniguchi 2020; (25) Slaton & Smith 2002; (26) Yu, Tang & Kuo (2000); (27) Feely, Doney & Cooley (2009); (28) Felle 2001; (29) Kramer, Sacksteder & Cruz 1999.

Name	Value(s)	Units	Notes	Ref.
Diffusion coefficient of CO_2 in water	1.88×10^3	$\mu m^2 s^{-1}$		(1, 2)
Diffusion coefficient of H_2CO_3 in water	1.2×10^3	$\mu m^2 s^{-1}$	Assumed in Fei et al., (2022) to be identical to diffusion coefficient of acetic acid	(2, 3)
Diffusion coefficient of HCO_3^- in water	1.15×10^3	$\mu m^2 s^{-1}$		(2, 4)
Diffusion coefficient of O_2 in water	2.42×10^3	$\mu m^2 s^{-1}$		(5)
Membrane permeability to CO_2	3.50e-03; 3.20e-02	$m\ s^{-1}$	Parameter scanned between reported values	(6, 7)
Membrane permeability to H_2CO_3	30	$\mu m\ s^{-1}$		(2)
Membrane permeability to HCO_3^-	0.05	$\mu m\ s^{-1}$		(8)
Membrane permeability to O_2	75	$cm\ s^{-1}$		(9)
Besotrophin-like channel mediated permeability of thylakoid to HCO_3^-	1×10^{-2}	$m\ s^{-1}$		(2)
Chloroplast membrane permeability to HCO_3^- mediated by LCIA	1×10^{-8}	$m\ s^{-1}$		(2)
Rate constant of spontaneous hydration of CO_2	6×10^{-2}	s^{-1}		(10)
Rate constant of spontaneous dehydration of H_2CO_3	2×10^1	s^{-1}		(10)
Rate constant of spontaneous deprotonation of H_2CO_3	1×10^7	s^{-1}		(10)
Rate constant of spontaneous protonation of HCO_3^-	5×10^{10}	$M^{-1}\ s^{-1}$		(10)
Carbonic anhydrase k_{cat}	0.3×10^6	s^{-1}		(11)
Carbonic anhydrase K_m for CO_2	1.5	$mol\ m^{-3}$		(12)
Carbonic anhydrase K_{eq}	0.56×10^{-6}	$mol\ m^{-3}$		(13)
Carbonic anhydrase K_m for HCO_3	34	$mol\ m^{-3}$		(13)
Carbonic anhydrase concentration in stroma	270	µM		(14)
Carbonic anhydrase concentration in cytosol	135	µM	Assumed to be half the stroma value	(14)

Table 4.1 (cont'd)

Carbonic anhydrase concentration in lumen	**135**	**µM**	**Assumed to be half the stroma value**	**(14)**
Rubisco V_{max} of carboxylation	7600	µmol L^{-1} s^{-1}		(15)
Rubisco V_{max} of oxygenation	1596	µmol L^{-1} s^{-1}	Calculated from ratio of k_{cat} values of carboxylation and oxygenation.	(16, 17)
Rubisco K_m O_2	8.6	µmol L^{-1}		(18)
Rubisco K_m CO_2	215	µmol L^{-1}		(18)
BicA V_{max}	1.85×10^{-4}	mol m^{-2} s^{-1}	Parameter scanned	(19)
BicA K_m HCO_3^-	0.217	mol m^{-3}		(19)
Stomatal conductance	0.4375	mol m^{-2} s^{-1}		(20)
Atmospheric concentration of CO_2	412	ppm		Assumed
Atmospheric concentration of O_2	0.21	Partial pressure		(21)
Thickness of cell wall in angiosperms	0.32	µm		(14, 22)
Thickness of cell wall in bryophytes	1.6	µm		(23)
Effective porosity of C3 plant cell wall	0.2	Unitless		(14)
Effective porosity of hornwort cell wall	0.0001		Parameter scanned	Calculated (23)
Thickness of unstirred boundary layer in algal model	0.32	µm	Assumed to be the same as cell wall thickness	Assumed
Thickness of unstirred apoplast water layer in land plant models	0.32	µm	Assumed to be the same as cell wall thickness	Assumed
Permeability of pyrenoid starch sheath to dissolved inorganic carbon	0.1 * P_{CO2}	µm s^{-1}	From range of permeabilities that allow effective carbon concentration in modeling done by (Hopkinson et al., 2011)	(2)
Permeability of pyrenoid starch sheath to oxygen	0.1 * P_{O2}	µm s^{-1}	Assumed to behave similarly to dissolved inorganic carbon	Assumed
Radius of pyrenoid	1.0	µm		(2)
Radius of thylakoid	0.5	µm	Multiplied by 10X to account for simpler thylakoid architecture	(2)
Height of thylakoid	4	µm		Assumed
Radius of chloroplast	4.63	µm	Calculated from stroma volume fraction and assuming spherical geometry	(14)
Radius of cytosol	8.77	µm	Assuming spherical geometry	(24)
Radius of plasmalemma surface	9.23	µm	Calculated from radius of cytosol, cell wall thickness, and assumed apoplast water thickness	Calculated
Radius of substomatal space in land plant model	11.63	µm		(14)
Proportion of cell wall adjacent to intercellular airspace in land plant	0.5	Unitless		(25)
pH of land plant apoplast	6.0	pH		(26)
pH of ocean water	8.1	pH		(27)
pH of cytosol	7.2	pH		Calculated (28)
pH of stroma	8.0	pH		(29)
pH of lumen	6.0	pH		(29)

4.4.2. Reaction equations

Carboxylation flux by rubisco is calculated as in Farquhar et al., (1980) **(E1)**. The rate of carboxylation by rubisco is normally taken to be the minimum of v_c and J, where J describes the rate of ribulose-1,5-bisphosphate regeneration enabled by photosynthetic electron transport and a function of J_{max}, a maximum rate of RuBP regeneration, among other parameters (Farquhar et al., 1980). Estimates of the relevant parameters are available for land plants but, to our knowledge, not for algae. We are also specifically examining CO_2-limiting conditions where rubisco reaction rate limitations dominate. For these reasons, we calculate the carboxylation and oxygenation rates assuming that the system is not limited by RuBP regeneration as in Fei et al., (2022).

$$v_c = \frac{(V_{\max}^{carboxylation} * [CO_2])}{\left([CO_2] + K_m^{CO_2}\left(1 + \frac{[O_2]}{K_m^{O2}}\right)\right)} \quad (\boldsymbol{E1})$$

The ratio of oxygenation to carboxylation Vmax is:

$$\frac{Vmax^{oxygenation}}{Vmax^{carboxylation}} = \frac{k_{cat}^{oxygenation}}{k_{cat}^{carboxylation}} \quad (\boldsymbol{E2})$$

Using a $\frac{k_{cat}^{oxygenation}}{k_{cat}^{carboxylation}}$ value of 0.21 as in Farquhar et al., (1980), we can thereby calculate the $Vmax^{oxygenation}$ of our systems. The oxygenation flux by rubisco is then calculated as:

$$v_o = \frac{(V_{\max}^{oxygenation} * [O_2])}{\left([O_2] + K_m^{O_2}\left(1 + \frac{[CO_2]}{K_m^{CO2}}\right)\right)} \quad (\boldsymbol{E3})$$

Interconversion of CO_2 with bicarbonate via carbonic anhydrase is described as in McGrath and Long, (2014):

$$\frac{[CA] * CA_{kcat} * \left([CO_2] - \frac{[HCO_3][H^+]}{K_{eq}}\right)}{K_m^{CO_2} + [HCO_3]\left(\frac{K_m^{CO_2}}{K_m^{HCO_3}}\right) + [CO_2]} \quad (\boldsymbol{E4})$$

In the land plant models, the flux density of dissolution of gaseous CO_2 or O_2 into the water layer is as in Hemond and Fechner, (2022):

$$FluxDensity_{WaterLayer} = -\frac{D_w\left(C_w - \frac{C_a}{H}\right)}{\delta_w} \quad (\boldsymbol{E5})$$

Where D_w is the diffusion rate of the dissolving species in water C_w *and* C_a are the concentrations of that species in the air and in the water layer, H is the dimensionless Henry's Law constant, and δ_w is the length of the unstirred water layer into which the gas is dissolving. In our models, we assume the presence of a thin layer of water on top of the plant's cell wall that is the same thickness as the cell wall itself into which CO_2 is dissolving.

Permeation of aqueous species through the cell wall is given by the following equation, as in McGrath and Long, (2014):

$$FluxDensity_{CellWall} = \frac{D_w}{\delta_{CellWall}} * EffectivePorosity \quad (\boldsymbol{E6})$$

Where *EffectivePorosity* is the porosity of the cell wall divided by the tortuosity of the cell wall.

For computational tractability, we combine the processes of gases dissolving into water and the aqueous species passing through the cell wall. Note that in the above equation D_w / δ_w and D_w **EffectivePorosity* / $\delta_{CellWall}$ gives permeability (in units of μm/s) of the water layer and the cell wall, respectively. Multiplying these values by surface area (SA) gives conductivities (in units of $\mu m^3/s$). The inverses of these values are resistances, which can be summed to give the total resistance of the water layer plus the cell wall. The inverse of this, again, will be the conductivity of the overall system, which can be multiplied by the concentration gradient from the air to the surface of the plasmalemma to give the total flux.

$$J = \left(\left(\frac{D_w * SurfaceArea}{\delta_w}\right)^{-1} + \left(\frac{D_w * EffectivePorosity * \mathrm{SA}}{\delta_{CellWall}}\right)^{-1}\right)^{-1} * \left(\frac{C_a}{H} - C_w\right) \quad (\boldsymbol{E7})$$

Permeation through lipid membranes is given by:

$$P * ([Outside] - [Inside]) \quad (\boldsymbol{E8})$$

Active transport by bicarbonate transporter *BicA* is described using Michaelis-Menten kinetics:

$$\frac{Vmax_{BicA} * [HCO_3]}{K_m * [HCO_3]} \quad (\boldsymbol{E9})$$

4.4.3. Efficiency calculations

Net CO_2 fixation is described as:

$$NetFixation = Flux_{carboxylation} - \frac{Flux_{oxygenation}}{2} \quad (\boldsymbol{E10})$$

2 NADPH equivalents are expended per carboxylation or oxygenation reaction based on the stoichiometry of the CBC cycle and photorespiration. 3 ATP and 3.5 ATP are used for a single carboxylation or oxygenation event, respectively (Edwards and Walker, 1983).

In models featuring a PCCM, there is a lumenal carbonic anhydrase that catalyzes the following reaction:

$$CO_2 + H_2O \longleftrightarrow HCO_3^- + H^+ \quad (\boldsymbol{E11})$$

Due to the acidic pH of the lumen (Kramer et al., 1999) the net flux of this reaction is overwhelmingly in the direction of CO_2 and H_2O, so that entry of bicarbonate depletes the proton motive force (pmf) that is maintained by the light reactions of photosynthesis, which imposes an indirect ATP cost on CCM activity by requiring additional proton pumping to maintain the pmf (Mukherjee et al., 2019). Based on a 14:3 ratio of pumped protons to ATP synthesis via the thylakoid membrane ATP synthase, inferred from the number of c-subunits in such ATP synthases (Seelert et al., 2000), we can calculate the indirect ATP cost of this lumen CA activity as:

$$ATP_{cost} = J_{CA_{lumen}} * \frac{3}{14} \quad (\boldsymbol{E12})$$

This is added to the other ATP consumption in the model (due to the metabolic costs of carboxylation and oxygenation) to give total ATP use. This can be compared with NADPH use due to carboxylation and oxygenation to get an estimate of the total ATP, NADPH, and the ATP:NADPH ratio needed to support the activity in the model. From the values provided in Walker et al., (2020) we estimate the amount of either Cyclic Electron Flow (CEF) or Malate Valve activity needed to rebalance the ATP/NADPH ratio needed for a particular model, which we can then convert into an additional demand for photons and, therefore, a the number of

photons needed on a per reaction (carboxylation or oxygenation) basis (**Appendix C, Figure S4.1**). From this, we can calculate the number of photons needed to support model fluxes and then compare this to the net fixation achieved by a model to get an estimate of light use efficiency.

$$\varphi_{CEF} = \frac{V_c - \frac{1}{2}V_o}{(V_c + V_o)\left(Photons_{base} + \left((Ratio * NADPH_{base}) - ATP_{base}\right) * 0.43\frac{photons}{ATP}\right)} \quad (\boldsymbol{E13})$$

$$\varphi_{malate} = \frac{V_c - \frac{1}{2}V_o}{(V_c + V_o) * \left(Photons_{base} + \frac{(Ratio * NADPH_{base}) - ATP_{base}}{\frac{5.45\ ATP}{2\ NADPH}} * 4\frac{photons}{NADPH}\right)} \quad (\boldsymbol{E14})$$

Where *Vc and Vo* are the modeled rates of carboxylation and oxygenation, *Ratio* refers to the modeled ATP/NADPH ratio necessary to support the fluxes in the model, and $Photons_{base}$, ATP_{base} and $NADPH_{base}$ refer to the photons used and the ATP and NADPH generated in the process of making two NADPH molecules via Linear Electron Flow (LEF) (Walker et al., 2020).

4.4.4. Concentration calculations

All concentrations in the models used in this study are in units of µM. To calculate the µM concentrations of CO_2 and O_2 in the atmosphere, we used the following conversion:

$$\frac{412\ \mu\text{mol}\ CO_2}{\text{mol air}} * \frac{1\ \text{mol air}}{24.79\ L\ \text{air}} = \sim \frac{16.62\ umol\ CO_2}{L\ air}$$

$$\frac{0.2095\ \text{mol}\ O_2}{\text{mol air}} * \frac{1\ \text{mol air}}{24.79\ L\ \text{air}} = \sim \frac{8450.98\ umol\ O_2}{L\ air}$$

4.5. Results

4.5.1. Validation of compensation point predictions and sensitivity analysis

The land plant and algal carbon assimilation models were validated by comparing a key estimated result (CO_2 compensation point) with experimentally measured values from the literature. The CO_2 compensation point is the external CO_2 level at which net CO_2 assimilation by a photosynthesizing organism is zero (i.e., carbon assimilation by rubisco is balanced out by

CO_2 losses to photorespiration and respiration in the light, denoted as R_L). Low compensation points are also a defining feature of organisms with CCMs since they maintain net positive carbon assimilation at lower CO_2 concentrations, making this a useful indicator of whether land plant and algal models with and without CCMs reasonably recreate the carbon assimilation dynamics of real systems.

As shown in **Figure 4.2** and **Table 4.2**, the models with CCMs have substantially lower compensation points than the models lacking CCMs. Moreover, as shown in **Table 4.2**, these estimated compensation point values fall within the ranges of values reported in the literature for angiosperm land plants and algae with and without CCMs **(Table 4.2).** Note that the reported compensation points of hornworts with pyrenoids (11-13 ppm) are lower than those of closely related C3 liverworts, but higher than typical estimates for C4 plants and pyrenoid-containing algae (Villarreal and Renner, 2012).

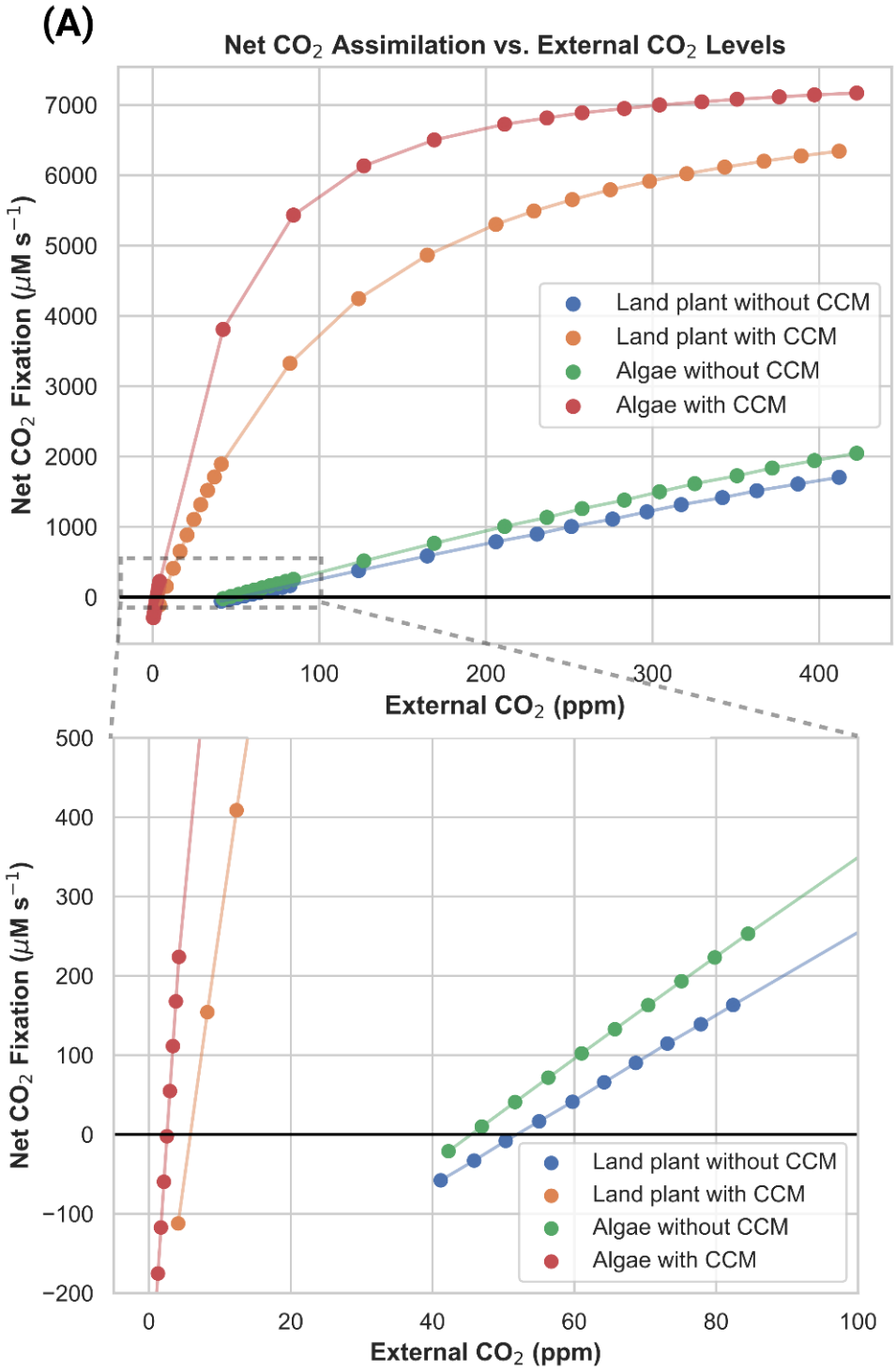


Figure 4.2: Net CO_2 assimilation versus external CO_2 concentrations in carbon assimilation models. The point at which net CO_2 assimilation is zero defines the compensation point. **(A)** The full range of saturation and external CO_2 concentrations, and **(B)** a zoomed-in panel showing the point at which each curve reaches 0% rubisco saturation (i.e., the compensation point).

Table 4.2: Predicted compensation points for different models from the present study compared with reference values from the literature. Reference column numbers refer to their numbering in the bibliography.

Model	Compensation point (ppm CO2)	Reference Values (ppm CO2)	Measurement References
Land plant with CCM	6.2	1.3; 4.3; 0.7 – 9.0	(Fladung and Hesselbach, 1987; Lee et al., 2022)
Land plant without CCM	52.7	48; 57; 48.2 – 53.4; 65-100	(Fladung and Hesselbach, 1987; Tolbert et al., 1995; Peixoto et al., 2021)
Algal model with CCM	2.7	0.75 – 2.5; 6.0	(Coleman and Colman, 1980; Raven et al., 1982)
Algal model without CCM	44.6	43.5 – 58; 64.5	(Raven et al., 1982; Steensma et al., 2023)

The sensitivity analysis results shown in **Figure 4.3** show that simulated net CO_2 assimilation and quantum yield values from the land plant models are relatively robust to local variations in all parameters, providing us with confidence that these results are not merely the result of a very particular selection of parameters. In both the land plant and algal models without PCCMs, rubisco V_{max}, cell and chloroplast radii, and membrane permeability to CO_2 are the most influential determinants of net CO_2 assimilation and quantum yield. In the land plant model, stomatal conductance also stands out. The addition of a PCCM reduces the sensitivity of net CO_2 assimilation to changes in any input parameter but increases the sensitivity of the predicted quantum yield to input parameter values. The local stability of our results to perturbations in key parameters is comparable with previous studies, being more variable than the models presented in Fei et al., (2022), which spatially modeled a smaller system (algal chloroplasts), and significantly less variable than the models presented in McGrath and Long, (2014), which modeled land plant CO_2 assimilation at a similar scale. We also characterized the sensitivity of our modeling results to the spatial resolution of the numerical simulations. Our results **(Appendix C, Figures S4.2-3)** show that rubisco saturation - the percentage of maximum rubisco activity achieved – and quantum yield in an algal model lacking a CCM are robust to the simulation resolution. Increasing the resolution all the way down to 0.32um, well beyond what could feasibly be done given the amount of parameter exploration done in this study, does result

in noticeable changes in pyrenoid [CO_2] and [HCO_3^+], resulting in small increases in rubisco saturation and small decreases in quantum yield **(Appendix C, Figures S4.4-5)**.

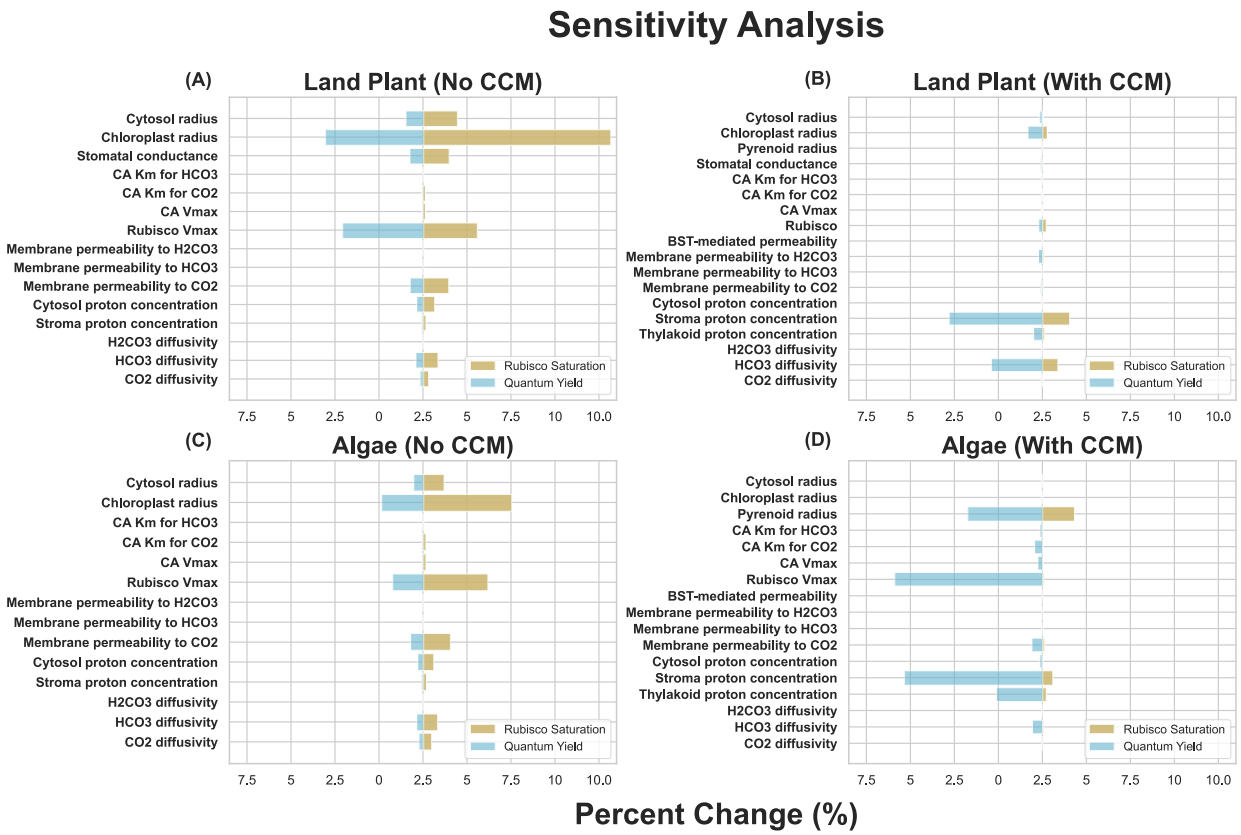


Figure 4.3: Sensitivity analysis results for **(A)** the land plant model lacking a CCM, **(B)** the algal model lacking a CCM, **(C)** the land plant model with a CCM, and **(D)** the algal model with a CCM. Orange bars indicate the absolute % change of quantum yield resulting from a 10% change in the indicated parameter, and blue bars represent the same for rubisco saturation. For both of the land plant models, increasing the cytosol radius by 10% resulted in problems with solving the systems numerically, so the cytosol radius was increased by 1% instead and, assuming a linear relationship between the size of radius increase and the change in rubisco saturation and quantum yield, multiplied by 10 to get the values shown in **(A-B)**.

4.5.2. Efficiency of chloroplast membrane bicarbonate channel is strongly dependent on assumed permeability of chloroplast membrane to CO_2

Previous studies (Price et al., 2010; McGrath and Long, 2014) have suggested that the incorporation of bicarbonate transporters into the chloroplast membrane of a land plant could improve net fixation and/or the efficiency of carbon assimilation, and that this could represent a reasonable intermediate stage in a broader biotechnological effort to implement a full CCM in a land plant. Modeling studies on CCM systems typically assume the lipid membrane permeability

of 0.35 cm/s, which was experimentally measured and reported in Gutknecht et al., (1977). However, there is substantial uncertainty as to the value of parameter, with experimental estimates ranging over many orders of magnitude (Evans et al., 2009). The permeability may be as much as an order of magnitude higher than the Gutknecht *et al* value, as reported by Missner et al., (2008). We hypothesized that the apparent favorability of employing a chloroplast membrane bicarbonate pump may be highly sensitive to the assumed chloroplast membrane CO_2 permeability.

To test this hypothesis, we performed a parameter exploration from an order of magnitude lower than the widely cited Gutknecht et al., (1977) value up to the Missner et al., (2008) value in both land plant and algal systems, calculating net fixation as well as ATP/CO_2 and light-use efficiency, as shown in **Figure 4.4**.

These results show that the light use efficiency of a chloroplast membrane bicarbonate transporter is highly sensitive to the value of the chloroplast envelope's permeability to CO_2, with a large range of permeabilities resulting in 2X more ATP usage per unit of CO_2 fixed. In the land plant model, we see increases in both rubisco saturation and quantum yield as BicA pumping activity increases when lipid membrane permeability values are equivalent to, or below that reported in Gutknecht et al., (1977) **(Figure 4.4A-B)**. At permeabilities higher than this, increased BicA activity actually decreases quantum yield, though net fixation still increases **(Figure 4.4A-B).** We see a similar picture in the algal model **(Figure 4.4E-F)**, suggesting that the differences in DIC form, concentration, and diffusivity do not greatly impact the sensitivity of this strategy to the specific value of lipid membrane permeability to CO_2. The decrease in quantum yield in models with high lipid membrane permeability to CO_2 is driven by increased leakage of CO_2 from the chloroplast back into the cytosol after it interconverts with the bicarbonate just pumped by BicA **(shown as flux V_{15} in Figure 4.1).** As lipid membranes become more permeable to CO_2, its tendency to escape the chloroplast before being fixed by rubisco increases. Lowering the external CO_2 concentration does, however, change the energy efficiency penalty of increased BicA activity significantly **(Figure 4.4C-D;G-H)**. Even at higher lipid membrane permeability values, we see only minimal decreases in quantum yield with increased BicA bicarbonate pumping.

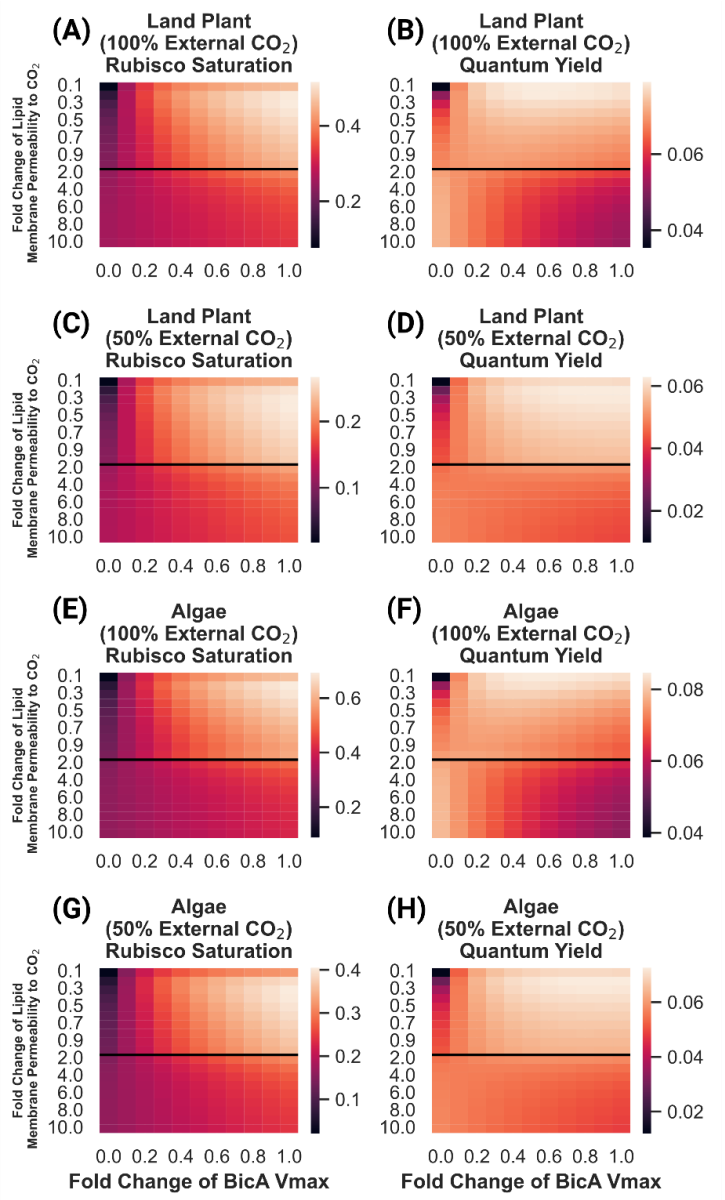


Figure 4.4: Rubisco saturation and quantum yield of land plant and algal models of CO_2 assimilation under 100% and 50% external CO_2 levels, as a function of lipid membrane permeability to CO_2 and BicA bicarbonate transporter V_{max}. Fold change of lipid membrane permeability is relative to the value reported in Gutknecht et al., (1977). **(A)** Predicted rubisco saturation of a land plant model under 100% external CO_2. **(B)** Predicted quantum yield of a land plant model under 100% external CO_2. **(C)** Predicted rubisco saturation of a land plant model under 50% external CO_2. **(D)** Predicted quantum yield of a land plant model under 50% external CO_2. **(E)** Predicted rubisco saturation of an algal model under 100% external CO_2. **(F)** Predicted quantum yield of an algal model under 100% external CO_2. **(G)** Predicted rubisco saturation of an algal model under 50% external CO_2. **(H)** Predicted quantum yield of an algal model under 50% external CO_2. The black lines in each plot indicate the Gutknecht et al., (1977) value for lipid bilayer permeability to CO_2 as well as a transition in the y-axis from increments of 0.1X to 1X fold changes.

4.5.3. Efficiency of a plasmalemma bicarbonate channel is strongly dependent on external DIC levels and limited by the rate of equilibration between CO_2 and bicarbonate

We found that although the strategy of pumping bicarbonate from the cytosol to the chloroplast may incur substantial energy costs, implementing a bicarbonate pump at the plasmalemma may be more effective. This makes sense considering that in aqueous systems at near-neutral pH, most of the DIC in the system is in the form of bicarbonate. We incorporated a plasmalemma bicarbonate transporter and explored the efficiency of such a system across different external DIC concentrations and activities of the transporter in both algal and land plant systems **(Figure 4.5)**.

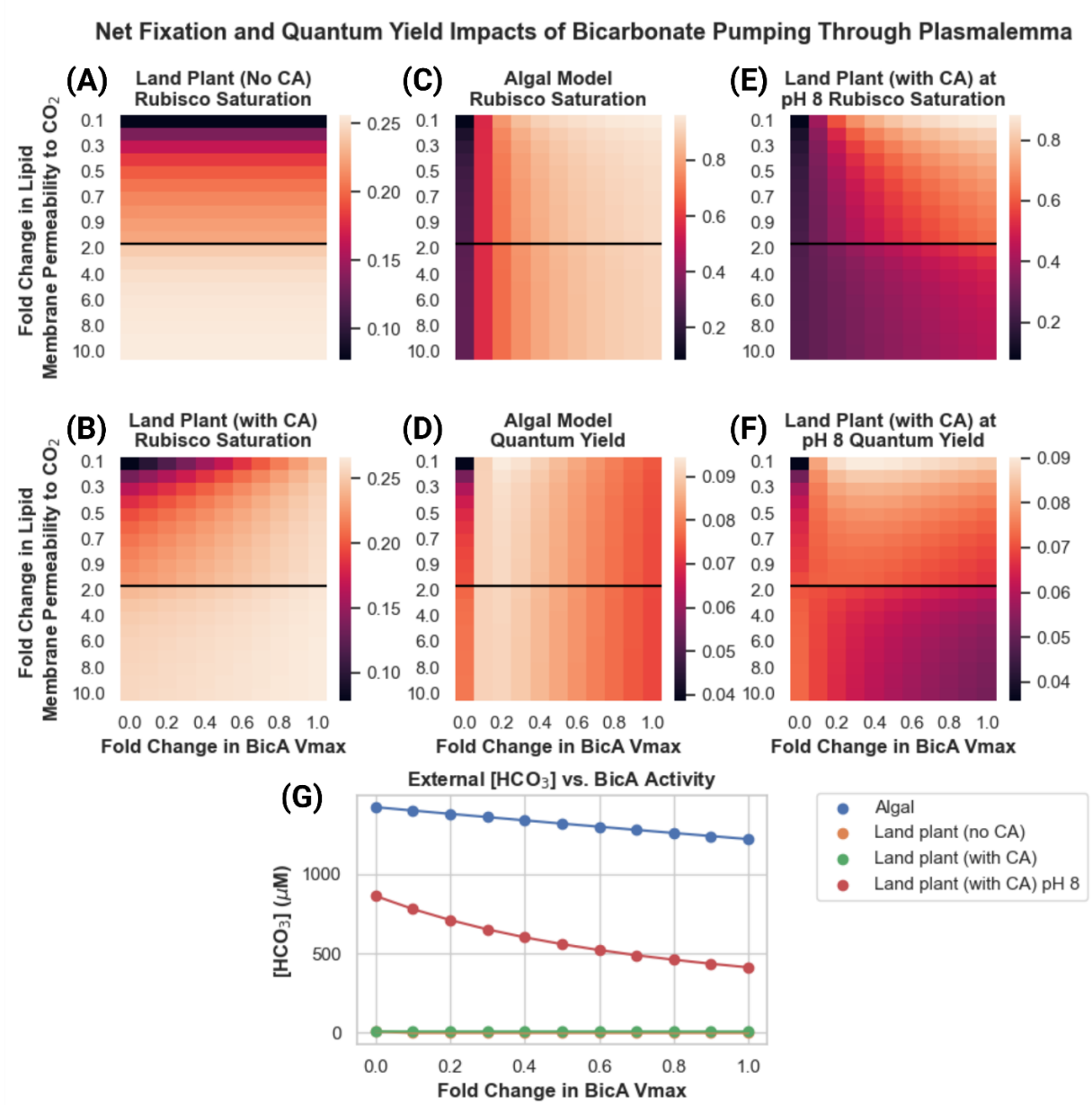


Figure 4.5: Predicted rubisco saturation and quantum yield in land plant and algal models with a BicA bicarbonate pump present in the plasmalemma membrane, as a function of assumed lipid membrane permeability to CO_2 and BicA V_{max}. Fold change of lipid membrane permeability is relative to the value reported in Gutknecht et al., (1977). **(A)** Predicted rubisco saturation of a land plant model lacking an apoplastic carbonic anhydrase. **(B)** Predicted rubisco saturation of a land plant model with an apoplastic carbonic anhydrase. **(C)** Predicted rubisco saturation of an algal model. **(D)** Predicted quantum yield of an algal model. **(E)** Predicted rubisco saturation of a land plant model with an apoplastic carbonic anhydrase and an apoplast pH of 8. **(F)** Predicted quantum yield of a land plant model with an apoplastic carbonic anhydrase and an apoplast pH of 8. The black lines in each plot indicate the Gutknecht et al., (1977) value for lipid bilayer permeability to CO_2 as well as a transition in the y-axis from increments of 0.1X to 1X fold changes.

In the land plant model, the plasmalemma bicarbonate pump is not an effective means of increasing either net fixation or energy efficiency. As anticipated, the pump does work in the algal case (**Figure 4.5**). The key difference appears to be that the external environment in the algal system, which is suffused with bicarbonate ions, can maintain reasonably high steady-state concentrations in the vicinity of the cell to support the bicarbonate pumping activity **(Figure 4.5C-D).** In contrast, in the land plant system all dissolved bicarbonate available to the cell must first enter the system as CO_2 in the intercellular airspace, dissolve into the water in the apoplast, and then spontaneously hydrate to H_2CO_3 and deprotonate into bicarbonate. Although the protonation/deprotonation between H_2CO_3 is extremely fast, the hydration/dehydration is not [first-order rate constant of hydration of CO_2 to H_2CO_3 is $6 \times 10^{-2}\ s^{-1}$ (Mitchell et al., 2010)]. The result is an almost instantaneous depletion of the HCO_3^- concentration in the apoplast space, with insufficient spontaneous hydration flux to replenish it **(Figure 4.5G)**. Adding carbonic anhydrase activity to the apoplast allows for much faster regeneration of the external HCO_3^- concentration, allowing BicA to impact rubisco saturation **(Figure 4.5A-B)**. However, the pH of the apoplast, although variable, tends to be slightly to moderately acidic (Yu et al., 2000), resulting in low HCO_3^- concentrations in the land plant model even with the apoplast carbonic anhydrase included **(Figure 4.5G)**. It is only when the apoplast pH is made substantially more basic (pH of 8) and a carbonic anhydrase is included that the land plant model can replicate the algal model's rubisco saturation and quantum yield gains by using a plasmalemma bicarbonate pump **(Figure 4.5E-F)**.

4.5.4. PCCM integration results in greater marginal cost of CO_2 fixation improvements in land plants vs. algal systems and switches from decreasing to increasing light-use efficiency around a C_i typical of C4 plants

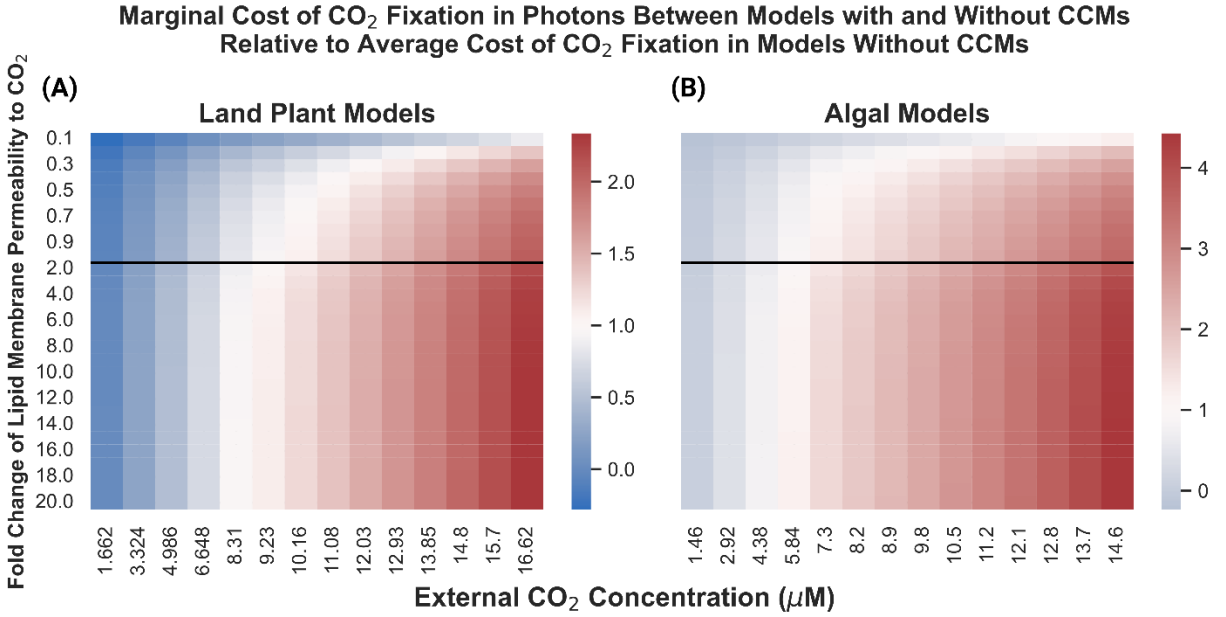


Figure 4.6: The ratio of the marginal cost in photons of one unit of net CO_2 fixation in land plant **(A)** and algal **(B)** models resulting from adding a PCCM relative to the average cost of fixing one molecule of CO_2 in those same models without CCMs, as a function of lipid membrane permeability and external CO_2 concentrations. Fold change of lipid membrane permeability is relative to the value reported in Gutknecht et al., (1977). Blue indicates that for a given lipid membrane permeability / external CO_2 concentration combination, the model containing a CCM has a lower marginal cost of CO_2 fixation – i.e.., is more light-efficient – than the average cost of CO_2 fixation in the model lacking a CCM. Red indicates that for a given parameterization, the model containing a CCM has a higher marginal cost of CO_2 fixation than the average cost of CO_2 fixation in its CCM lacking counterpart. The black lines in each plot indicate the Gutknecht et al., (1977) value for lipid bilayer permeability to CO_2 as well as a transition in the y-axis from increments of 0.1 to 1 in the X-fold changes.

We compared the energy-use efficiency of PCCM integration by comparing the predicted cost in photons of fixing CO_2 molecule in four different models: (i) a land plant model with a PCCM, (ii) a land plant model without a PCCM, (iii) an algal model with a PCCM, and (iv) an algal model without a PCCM. By dividing the increase in net CO_2 fixation in models (i) and (iii) relative to models (ii) and (iv) we estimated the marginal cost of in photons of fixing an additional CO_2 molecule using a PCCM in our land plant and models **(Figure 4.6)**. As we observed when examining the efficiency of the plasmalemma and chloroplast envelope BicA bicarbonate pumps, the assumed permeability of lipid membranes can have an impact on

efficiency; in this case, however, the relative marginal cost values do not change dramatically between an assumed permeability equivalent to that used in previous studies (1.0 in **Figure 4.6A-B)** and the higher value closer to that reported in Missner et al., (2008).

In the algal models, the use of the PCCM appears to only become marginally efficient with respect to light usage below an external [CO_2] of 4.38 μM. In contrast, the CCM is efficient in the land plant model below a substomatal [CO_2] of 243 ppm.

4.5.5. As cell wall thickness increases and cell wall effective porosity decreases, PCCMs become more favorable in land plant models

Given the findings regarding PCCMs in land plants highlighted above, it is interesting that many species of hornworts have pyrenoids – are there any meaningful biophysical differences between hornworts and other land plants that could explain these differences? As highlighted in Meyer et al., (2008) and Flexas et al., (2021) hornworts and other bryophytes have cell walls that are both substantially thicker and less porous compared to other land plants. From the mesophyll conductance values reported for angiosperms and bryophytes reported in Meyer et al., (2008) and Flexas et al., (2021), and with the assumption that other internal resistances to CO_2 diffusion are similar between bryophytes and embryophytes, we can estimate that the effective porosity of a bryophyte like a hornwort must be on the order of four orders of magnitude smaller than in a typical C3 angiosperm. We explore parameters within this range of possible porosity values and across multiple external CO_2 concentrations (**Figure 4.7**).

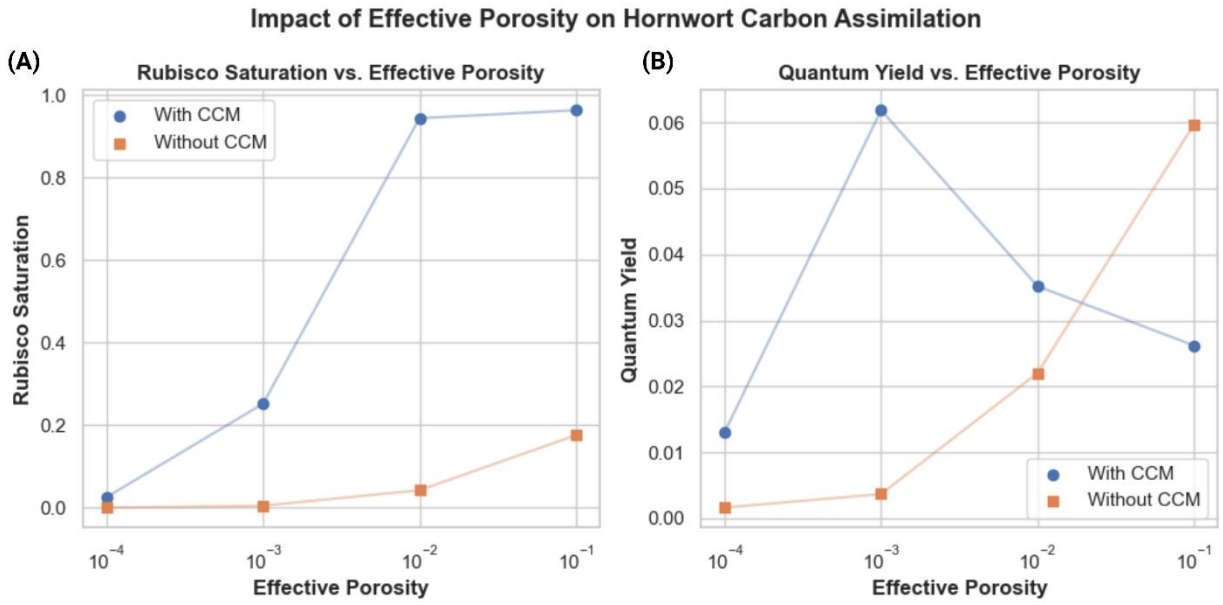


Figure 4.7: Rubisco saturation **(A)** and quantum yield **(B)** of a land plant model with varying effective porosity values. Blue points / lines represent predicted rubisco saturation or quantum yield in models including a PCCM; orange points/ lines represent predicted saturation or quantum yield in models not including a PCCM.

Below effective porosities on the order of 10^{-1}, which fall in the range we would expect of angiosperms, our model shows that the plant struggles to fix CO_2 without a CCM. With a PCCM, however, the model can achieve some level of net CO_2 fixation all the way down to effective porosities of 10^{-3}. Below porosities of 10^{-3}, we do not observe net CO_2 fixation in the model without a PCCM, and at a porosity of 10^{-4}, both models with and without PCCMs struggle to fix carbon. In terms of light-use efficiency, the model with a PCCM achieves a greater quantum yield of photosynthesis than the model without a PCCM below effective porosities of 10^{-2}.

4.6. Discussion

We initially hypothesized that the conspicuous absence of biophysical CCMs in almost all land plant lineages, in contrast to algae where they are widespread (Raven et al., 2005), may be the result of lower efficiency of such systems in land plants relative to algae, and that this results from their different biophysical contexts. To our surprise, we found that PCCMs appear to result in qualitatively similar improvements in quantum yield and net CO_2 assimilation in land plant and algal models. In the algal model, the fact that addition of a PCCM does not result in efficiency gains until relatively low external DIC levels are reached is surprising, given that

Chlamydomonas reinhardtii cells appear to concentrate carbon even at recent "air-level" – approximately 330 ppm – CO_2 concentrations (Badger et al., 1980). This implies that algae may routinely run their CCMs even when this incurs a quantum yield penalty. In contrast, the intercellular CO_2 concentration at which the CCM improves quantum yield in the land plant model (~243 ppm) is higher than reported estimates of C_i in C4 plants under laboratory, greenhouse, and field conditions (Bunce, 2005). Previous work has described the evolutionary history of C4 photosynthesis (Sage et al., 2018) and identified certain anatomical features – namely Kranz anatomy – and environmental factors such as hot, arid conditions that lead to increased transpirational water loss and factors such as Water-Use Efficiency (WUE) as key predictors of C4 emergence. If the estimated quantum yield gains resulting from the introduction of a biophysical CCM to a land plant in this study apply to biochemical CCMs like C4 and CAM photosynthesis, this may represent an additional evolutionary driver towards such systems.

Hornworts are the only land plant lineage that has evolved a biophysical CCM and they have done so multiple times (Villarreal and Renner, 2012). Hornworts, as well as some other bryophytes, are noteworthy for having substantially slower gas exchange between their surroundings and their photosynthetic tissues when compared with vascular land plants (Meyer et al., 2008). Our results show that a land plant with the low effective cell wall porosities we might expect given their extremely poor gas exchange characteristics, the use of a CCM becomes necessary to achieve net CO_2 fixation, which would impose a strong selective pressure for adopting one. The fact that hornworts represent the earliest-diverging extant branch of the land plants, and therefore may have maintained the genes and regulatory networks necessary to adopt a PCCM, may explain why this biophysical CCM strategy has been adopted by hornworts and not other land plants growing in conditions where biochemical CCMS have been selected for. We should note that in the models presented in this study, at effective porosities below 10^{-3}, only single digit values of rubisco saturation are achieved even with a biophysical CCM present and active, which may not be sufficient for viability, especially since we do not have or include estimates of respiration in the light in the models. This is despite the fact that mesophyll conductance to CO_2 in hornworts, which we are using effective porosity as a proxy for in this study, has been measured to be four-to-five orders of magnitude lower than in angiosperms (Flexas et al., 2021). This suggests that our model underestimates the strength of the hornwort CCM or otherwise does not properly describe some aspect of hornwort CO_2 assimilation. The

ratio of chloroplast-to-thallus surface area has not been explored in our modeling, but was found in a previous study to be a potentially important determinant of hornwort mesophyll conductance (Carriquí et al., 2019). Future work might aim to incorporate an exploration of chloroplast position and surface area to better account for this in the modeling.

These results shed light on potential challenges associated with improving crop productivity via the introduction of biophysical CCMs. The specific value chosen for the permeability of lipid bilayers to CO_2 has a large effect on the predicted energy efficiency of our models, with values higher than those used in previous modeling studies (McGrath and Long, 2014; Fei et al., 2022) but within the range of previously reported literature values (Gutknecht et al., 1977; Missner et al., 2008) resulting in qualitatively different conclusions. We see this in our consideration of BicA-mediated $HCO3^-$ pumping, which had been previously flagged as a promising intermediate step in introducing a biophysical CCM to a C3 plant (Price et al., 2010; McGrath and Long, 2014). As noted in Fei et al., (2022), barriers to CO_2 diffusion form a key component of known functional CCMs, so the finding that the chloroplast membrane may provide enough of a diffusion barrier for the transport of HCO_3^- into the stroma and subsequent conversion to CO_2 to meaningfully improve net fixation and carbon assimilatory efficiency was surprising. Our results show that at or below the permeability reported in Gutknecht et al., (1977), which is used in other modeling studies, increasing BicA pumping activity leads to improvements in quantum yield, indicating more efficient CO_2 fixation with respect to light use. However, above this value, we see uniform decreases in quantum yield with increased BicA activity. Net CO_2 fixation increases with BicA pumping in all cases; therefore, in situations where light is abundant relative to CO_2, this decrease in efficiency may not impact plant fitness. However, recent modeling work suggests that J_{max}, the maximum rate of ribulose-1,5-bisphosphate (RuBP) regeneration enabled by photosynthetic electron transport, is more limiting to crop yield than limits to the maximum rate of carboxylation (V_{max} of rubisco carboxylation) under the projected elevated atmospheric CO_2 levels of 2050 and 2100 (He and Matthews, 2023). In this study, improved quantum yields correspond to a combination of (i) lower expenditures of ATP for each CO_2 molecule fixed, and (ii) a more favorable ATP/NADPH ratio needed for fixation, resulting in less energy loss from the use of Cyclic Electron Flow during ATP/NADPH rebalancing (Walker et al., 2020). Under conditions of J_{max} limitations, differences in quantum yield may become a critical factor in determining yield, making the sensitivity of quantum yield

in this and other studies to assumed lipid bilayer permeability to CO_2 a matter of critical importance.

Interestingly, previous studies in this area (McGrath and Long, 2014; Fei et al., 2022) have performed sensitivity analyses that include this permeability as a surveyed parameter and its modeled effect is small compared to other parameters. These small local sensitivity values are estimated by observing the change in an output value like light-saturated CO_2 assimilation with a ± 10% change in the permeability parameter. This ignores the fact that the uncertainty in this value is in the range of at least an order of magnitude (Evans et al., 2009), and so despite low local sensitivity, the overall change that can result from varying it within reasonable bounds is substantial. The substantial uncertainty in this critical parameter could be reined in by future experimental measurements, though this will still be complicated by the potentially large variation between different plant systems, dynamic remodeling of lipid bilayers in response to developmental and environmental cues, etc. In the absence of well-defined values for this parameter, we encourage future groups modeling such systems to explore a range of values and to characterize the robustness of their conclusions to its variation.

In the near-neutral or slightly basic conditions that most photosynthetic organisms in aqueous environments find themselves in, HCO_3^- represents the primary form of Dissolved Inorganic Carbon (DIC) in their surroundings. Due to the impermeability of lipid bilayers to passive diffusion of HCO_3^-, the use of this pool of DIC requires organisms to employ an active transport mechanism [e.g., cyanobacterial HCO_3^- pumps like BicA (Price et al., 2004)] to move it from the extracellular to the intracellular space, which may often make sense due to the sheer quantity of DIC that is present in the environment. Although land plants ultimately obtain CO_2 from the atmosphere, this CO_2 must dissolve into water prior to entering photosynthesizing cells, at which point this aqueous CO_2 interconverts with other DIC species. This raises the possibility of a similar strategy – pumping HCO_3^- from a land plant's apoplast water into the intracellular environment to increase net CO_2 fixation – potentially viable. However, our results indicate that the limited spontaneous rate of CO_2 and HCO_3^- interconversion without the activity of carbonic anhydrase means that this strategy does not work.

Of note here is the fact that a quantitatively very similar system arises in algae growing in acidic environments where external HCO_3^- levels are negligible, such as the red alga *Cyanidioschyzon merolae* (De Luca et al., 1978). In such systems, all DIC must first enter the

cell passively as aqueous CO_2, at which point it will interconvert primarily between CO_2 and HCO_3, with the ratio of CO_2:HCO_3 determined by the cytosolic pH. There is strong evidence that *C. merolae* has a non-pyrenoid based CCM (Steensma et al., 2023). Such a system could use HCO_3^- pumping across the chloroplast envelope as a method of concentrating carbon, but our results suggest that this system would require maintenance of a near-neutral cytosolic pH along with the presence of carbonic anhydrases in the cytosol to be viable. The maintenance of this near-neutral pH in an acidic environment may, in turn, represent a substantial energetic cost to the organism.

4.7. Data and Code Availability

All results generated as part of this study can be found in the Supplemental Material. Models used for generating the results can all be found under the account *kastejos* in the Virtual Cell interface. Specific model names can be found for each dataset in the corresponding Supplemental Material tables.

4.8. Acknowledgments

The Virtual Cell, the software platform used for the reaction-diffusion simulations in this study, is supported by NIH Grant R24 GM137787.

This research was supported by the U.S. Department of Energy, Office of Science Biological and Environmental Research Grant no DE-SC0018269 (J.A.M.K., Y.S-H.) and Basic energy Sciences Grant no DE- FG02-91ER20021 (B.J.W.). This work is supported, in part, by the NSF Research Traineeship Program (Grant DGE-1828149) to J.A.M.K. This publication was also made possible by a predoctoral training award to J.A.M.K. from Grant T32-GM110523 from National Institute of General Medical Sciences (NIGMS) of the NIH. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIGMS or NIH.

4.9. Author Contributions

J.A.M.K, B.J.W, and Y.S-H. conceptualized the study. J.A.M.K. developed the models, ran the simulations, and analyzed the results. J.A.M.K. wrote the first draft of the manuscript. All authors contributed to revising and editing the final manuscript.

REFERENCES

Badger MR, Kaplan A, Berry JA (1980) Internal inorganic carbon pool of Chlamydomonas reinhardtii: evidence for a carbon dioxide-concentrating mechanism. Plant physiology **66**: 407–413

Bräutigam A, Schlüter U, Eisenhut M, Gowik U (2017) On the Evolutionary Origin of CAM Photosynthesis. Plant Physiol **174**: 473–477

Bunce J (2005) What is the usual internal carbon dioxide concentration in C4 species under midday field conditions? Photosynthetica **43**: 603–608

Carriquí M, Roig-Oliver M, Brodribb TJ, Coopman R, Gill W, Mark K, Niinemets Ü, Perera-Castro AV, Ribas-Carbó M, Sack L, et al (2019) Anatomical constraints to nonstomatal diffusion conductance and photosynthesis in lycophytes and bryophytes. New Phytologist **222**: 1256–1270

Coleman JR, Colman B (1980) Effect of oxygen and temperature on the efficiency of photosynthetic carbon assimilation in two microscopic algae. Plant Physiol **65**: 980–983

Cowan AE, Moraru II, Schaff JC, Slepchenko BM, Loew LM (2012) Spatial modeling of cell signaling networks. Methods in cell biology **110**: 195–221

De Luca P, Taddei R, Varano L (1978) Cyanidioschyzon merolae: a new alga of thermal acidic environments. Webbia **33**: 37–44

Edwards G, Walker D (1983) C3 ,C4: Mechanisms, Cellular and Environmental Regulation of Photosynthesis. Univ of California Press

Ermakova M, Danila FR, Furbank RT, von Caemmerer S (2020) On the road to C4 rice: advances and perspectives. Plant J **101**: 940–950

Evans JR, Kaldenhoff R, Genty B, Terashima I (2009) Resistances along the CO_2 diffusion pathway inside leaves. Journal of Experimental Botany **60**: 2235–2248

Farquhar GD, Caemmerer S, Berry JA (1980) A biochemical model of photosynthetic CO= assimilation in leaves of C3 species. Planta **149**: 78–90

Fei C, Wilson AT, Mangan NM, Wingreen NS, Jonikas MC (2022) Modelling the pyrenoid-based CO2-concentrating mechanism provides insights into its operating principles and a roadmap for its engineering into crops. Nature Plants **8**: 583–595

Fladung M, Hesselbach J (1987) Developmental Studies on Photosynthetic Parameters in C3, C3 - C4 and C4 Plants of Panicum. Journal of Plant Physiology **130**: 461–470

Flexas J, Clemente-Moreno MJ, Bota J, Brodribb TJ, Gago J, Mizokami Y, Nadal M, Perera-Castro AV, Roig-Oliver M, Sugiura D, et al (2021) Cell wall thickness and composition are involved in photosynthetic limitation. Journal of experimental botany **72**:

3971–3986

Gutknecht J, Bisson MA, Tosteson FC (1977) Diffusion of carbon dioxide through lipid bilayer membranes: effects of carbonic anhydrase, bicarbonate, and unstirred layers. The Journal of general physiology **69**: 779–794

He Y, Matthews ML (2023) Seasonal climate conditions impact the effectiveness of improving photosynthesis to increase soybean yield. Field Crops Research **296**: 108907

Hemond HF, Fechner EJ (2022) Chemical fate and transport in the environment. Academic Press

Hennacy JH, Jonikas MC (2020) Prospects for Engineering Biophysical CO_2 Concentrating Mechanisms into Land Plants to Enhance Yields. Annu Rev Plant Biol **71**: 461–485

Hopkinson BM, Dupont CL, Allen AE, Morel FMM (2011) Efficiency of the CO_2-concentrating mechanism of diatoms. Proceedings of the National Academy of Sciences **108**: 3830–3837

Kramer DM, Sacksteder CA, Cruz JA (1999) How acidic is the lumen? Photosynthesis Research **60**: 151–163

Lee M-S, Boyd RA, Ort DR (2022) The photosynthetic response of C3 and C4 bioenergy grass species to fluctuating light. GCB Bioenergy **14**: 37–53

Ludwig M (2013) Evolution of the C4 photosynthetic pathway: events at the cellular and molecular levels. Photosynth Res **117**: 147–161

McGrath JM, Long SP (2014) Can the cyanobacterial carbon-concentrating mechanism increase photosynthesis in crop species? A theoretical analysis. Plant physiology **164**: 2247–2261

Meyer M, Seibt U, Griffiths H (2008) To concentrate or ventilate? Carbon acquisition, isotope discrimination and physiological ecology of early land plant life forms. Philosophical transactions of the Royal Society of London Series B, Biological sciences **363**: 2767–2778

Missner A, Kügler P, Saparov SM, Sommer K, Mathai JC, Zeidel ML, Pohl P (2008) Carbon dioxide transport through membranes. The Journal of biological chemistry **283**: 25340–25347

Mitchell MJ, Jensen OE, Cliffe KA, Maroto-Valer MM (2010) A model of carbon dioxide dissolution and mineral carbonation kinetics. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences **466**: 1265–1290

Mukherjee A, Lau CS, Walker CE, Rai AK, Prejean CI, Yates G, Emrich-Mills T, Lemoine SG, Vinyard DJ, Mackinder LCM, et al (2019) Thylakoid localized bestrophin-like proteins are essential for the CO_2 concentrating mechanism of

Chlamydomonas reinhardtii. Proc Natl Acad Sci U S A **116**: 16915–16920

Peixoto MM, Sage TL, Busch FA, Pacheco HDN, Moraes MG, Portes TA, Almeida RA, Graciano-Ribeiro D, Sage RF (2021) Elevated efficiency of C3 photosynthesis in bamboo grasses: A possible consequence of enhanced refixation of photorespired CO_2. GCB Bioenergy **13**: 941–954

Price GD, Badger MR, von Caemmerer S (2010) The Prospect of Using Cyanobacterial Bicarbonate Transporters to Improve Leaf Photosynthesis in C3 Crop Plants. Plant Physiology **155**: 20–26

Price GD, Woodger FJ, Badger MR, Howitt SM, Tucker L (2004) Identification of a SulP-type bicarbonate transporter in marine cyanobacteria. Proceedings of the National Academy of Sciences **101**: 18228–18233

Raven JA, Ball LA, Beardall J, Giordano M, Maberly SC (2005) Algae lacking carbon-concentrating mechanisms. Can J Bot **83**: 879–890

Raven JA, Beardall J, Johnston AM (1982) Inorganic Carbon Transport in Relation to H+ Transport at the Plasmalemma of Photosynthetic Cells. Plasmalemma and Tonoplast: Their Functions in the Plant Cell. Elsevier Biomedical Press, Amsterdam, pp 41–47

Raven JA, Beardall J, Sánchez-Baracaldo P (2017) The possible evolution and future of CO_2-concentrating mechanisms. Journal of Experimental Botany **68**: 3701–3716

Raven JA, Cockell CS, De La Rocha CL (2008) The evolution of inorganic carbon concentrating mechanisms in photosynthesis. Philos Trans R Soc Lond B Biol Sci **363**: 2641–2650

Sage RF, Monson RK, Ehleringer JR, Adachi S, Pearcy RW (2018) Some like it hot: the physiological ecology of C4 plant evolution. Oecologia **187**: 941–966

Schaff J, Fink CC, Slepchenko B, Carson JH, Loew LM (1997) A general computational framework for modeling cellular structure and function. Biophysical journal **73**: 1135–1146

Seelert H, Poetsch A, Dencher NA, Engel A, Stahlberg H, Müller DJ (2000) Proton-powered turbine of a plant motor. Nature **405**: 418–419

Steensma AK, Shachar-Hill Y, Walker BJ (2023) The carbon-concentrating mechanism of the extremophilic red microalga Cyanidioschyzon merolae. Photosynth Res **156**: 247–264

Still CJ, Berry JA, Collatz GJ, DeFries RS (2003) Global distribution of C3 and C4 vegetation: Carbon cycle implications. Global Biogeochemical Cycles **17**: 6–1

Tolbert NE, Benker C, Beck E (1995) The oxygen and carbon dioxide compensation points of C3 plants: possible role in regulating atmospheric oxygen. Proceedings of the National Academy of Sciences **92**: 11230–11233

Villarreal JC, Renner SS (2012) Hornwort pyrenoids, carbon-concentrating structures, evolved and were lost at least five times during the last 100 million years. Proceedings of the National Academy of Sciences **109**: 18873–18878

Walker BJ, Kramer DM, Fisher N, Fu X (2020) Flexibility in the Energy Balancing Network of Photosynthesis Enables Safe Operation under Changing Environmental Conditions. Plants (Basel, Switzerland). doi: 10.3390/plants9030301

Walker BJ, VanLoocke A, Bernacchi CJ, Ort DR (2016) The Costs of Photorespiration to Food Production Now and in the Future. Annu Rev Plant Biol **67**: 107–129

Yu Q, Tang C, Kuo J (2000) A critical review on methods to measure apoplastic pH in plants. Plant and Soil **219**: 29–40

APPENDIX C: Supplemental Material for Chapter 4

FIGURES

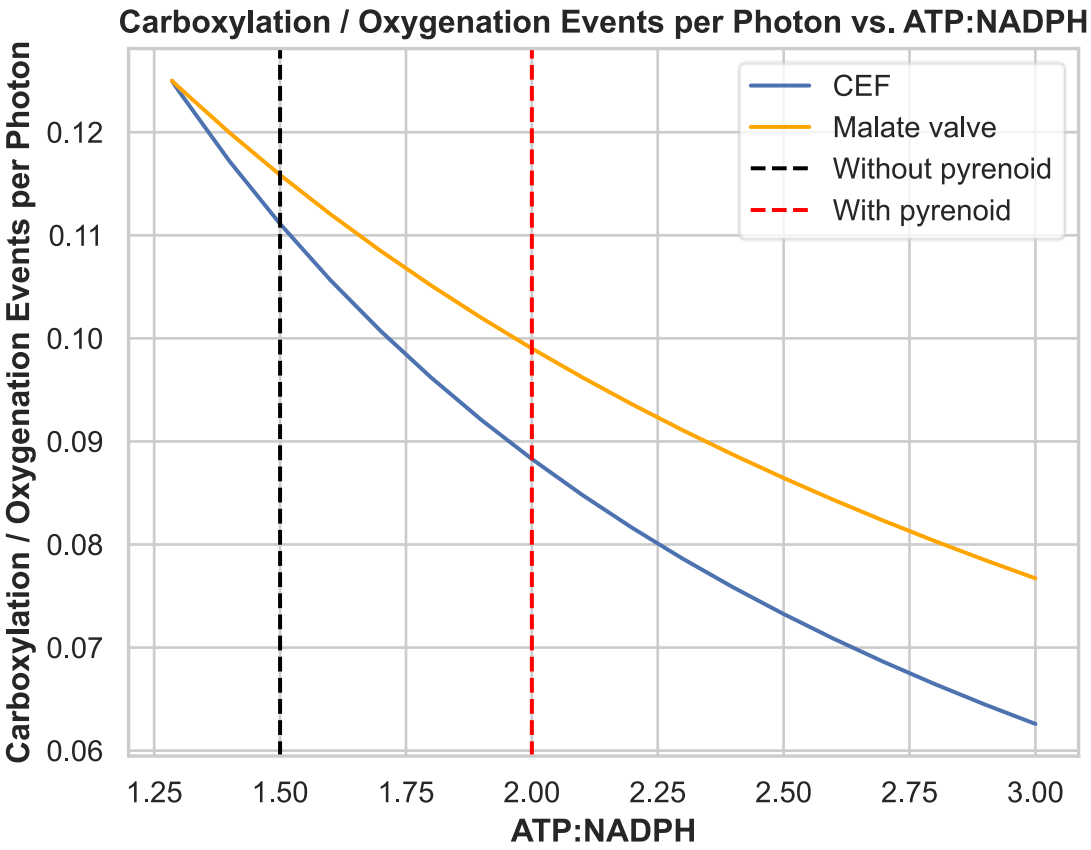


Figure S4.1: Carboxylation / oxygenation events per photon as a function of varying ATP:NADPH ratios. Costs associated with using either Cyclic Electron Flow or the malate valve for increasing ATP:NADPH ratio from the products of the light reactions are taken from Walker et al., (2020).

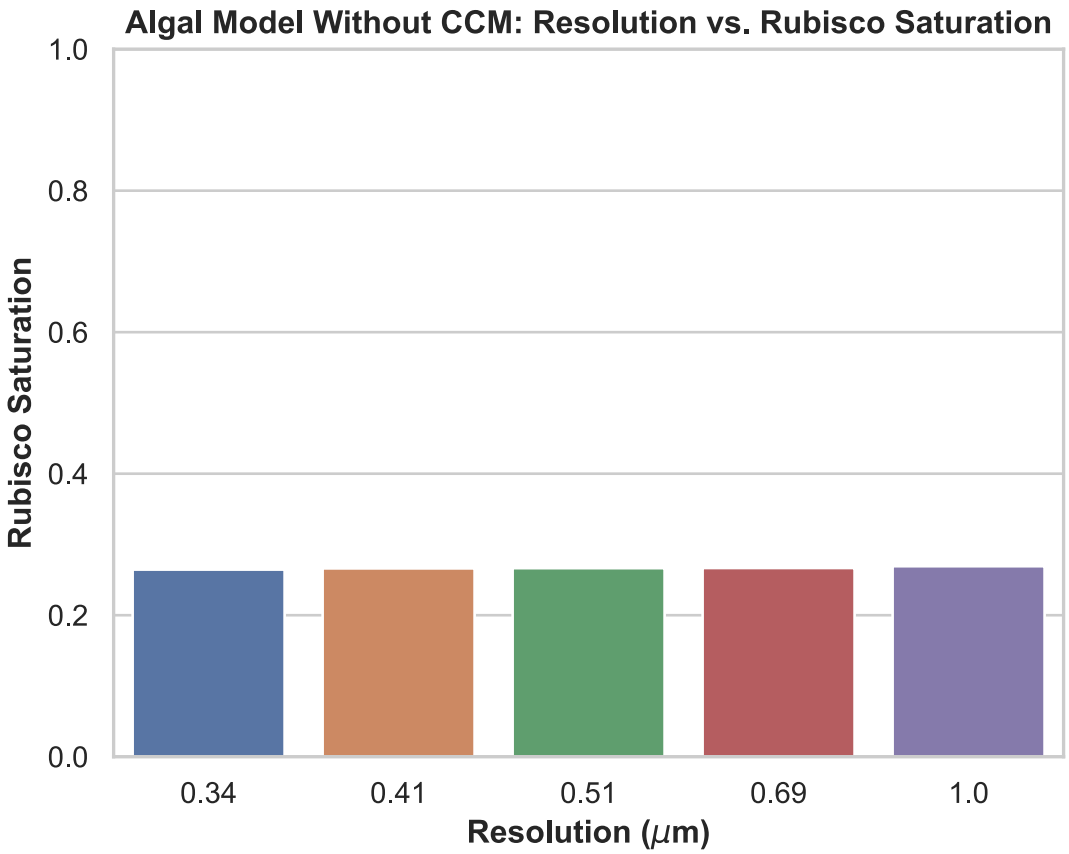


Figure S4.2: Effect of simulation spatial resolution on rubisco saturation. Simulation results are taken from the model of an algal cell without a CCM.

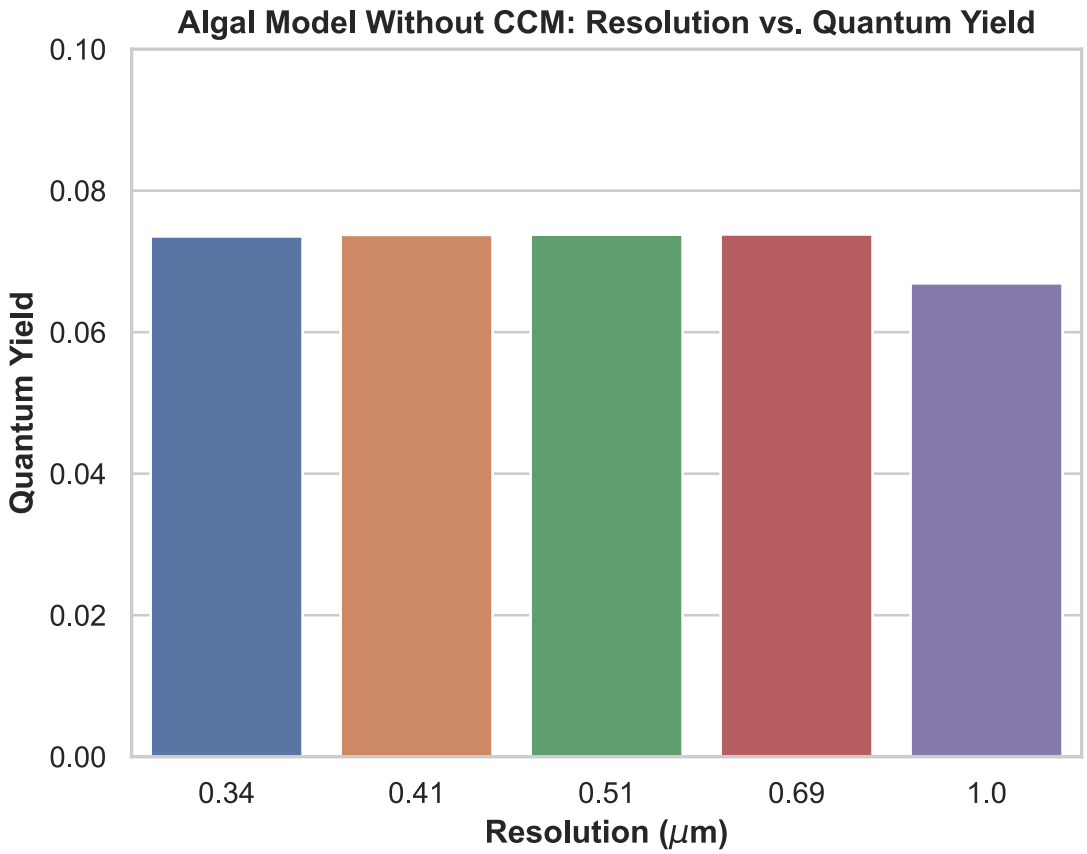


Figure S4.3: Effect of simulation spatial resolution on quantum yield. Simulation results are taken from the model of an algal cell without a CCM.

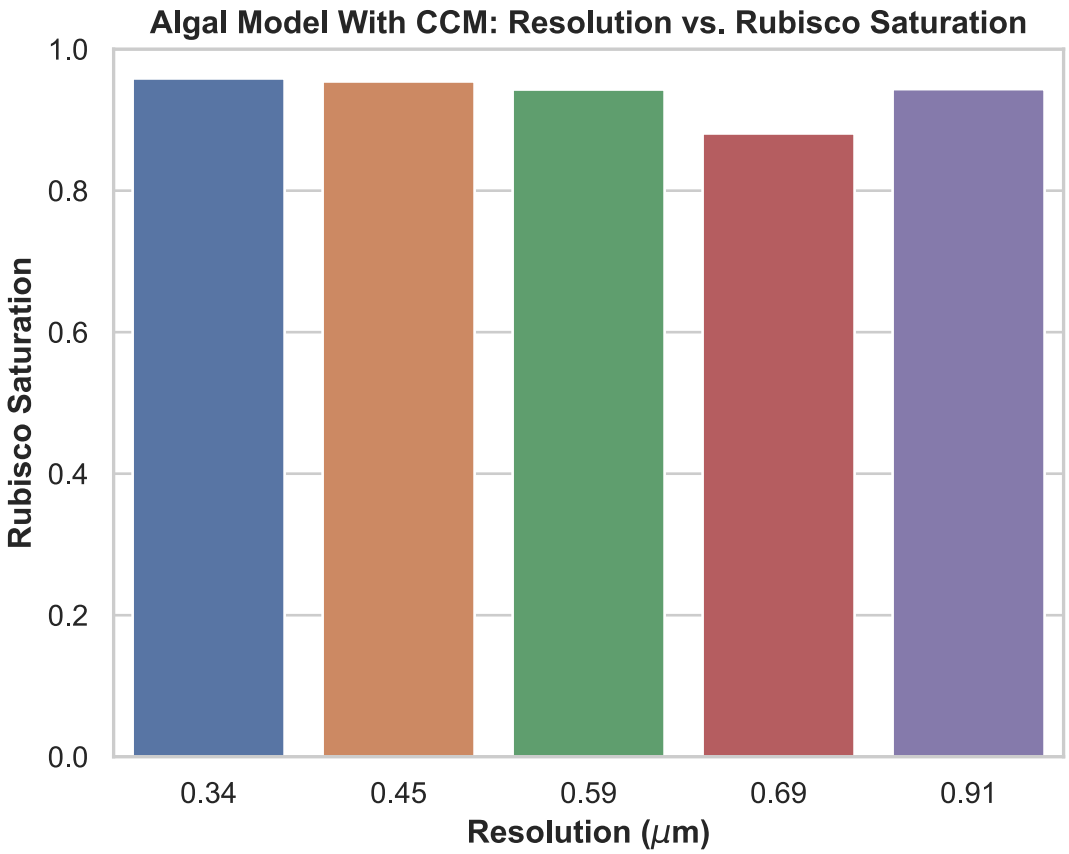


Figure S4.4: Effect of simulation spatial resolution on rubisco saturation. Simulation results are taken from the model of an algal cell with a CCM.

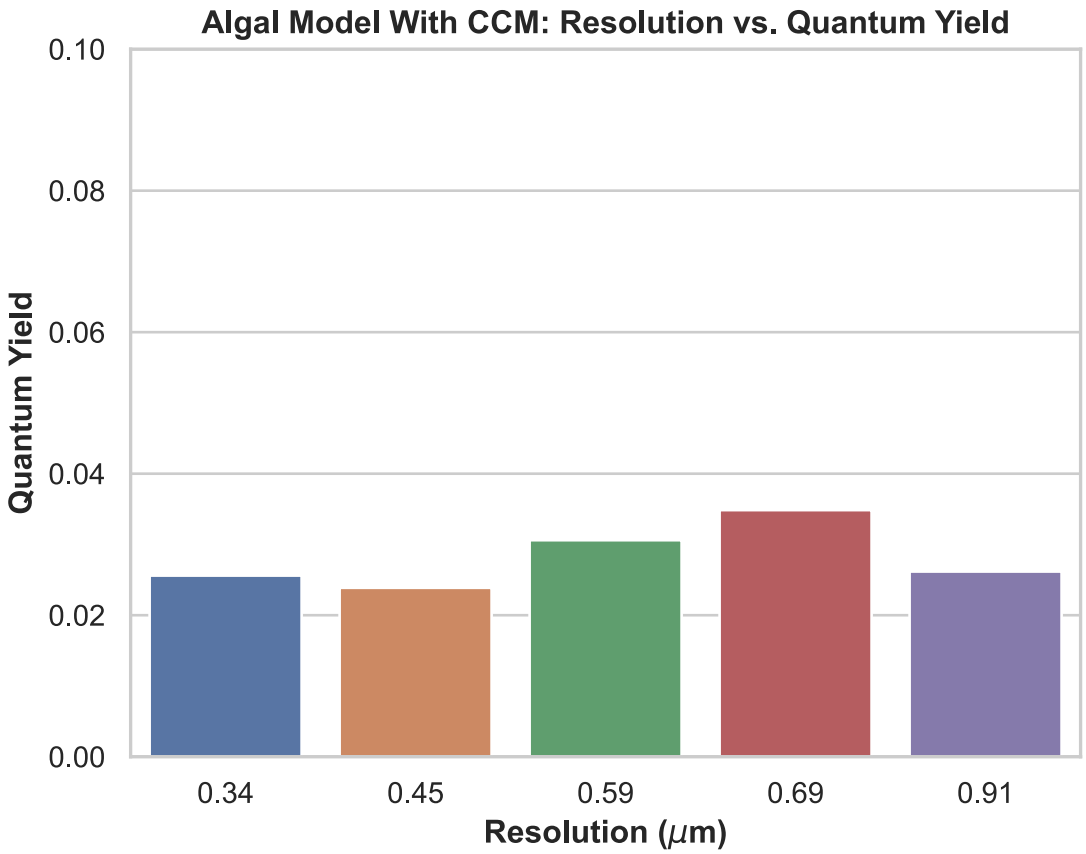


Figure S4.5: Effect of simulation spatial resolution on quantum yield. Simulation results are taken from the model of an algal cell with a CCM.

Supplementary Datasets Descriptions

All supplemental datasets can be found at the following link:

https://doi.org/10.1101/2024.01.04.574220

Supplemental Tables.xlsx: Contains results from VCell simulations discussed and analyzed in the manuscript.

Chapter 5

Concluding Remarks

5.1. Introduction

Taken as a whole, the studies presented in this thesis represent an attempt to improve and refine our understanding of photosynthetic metabolism by interrogating and improving the techniques we use to model it. I articulate the importance of careful statistical evaluation of metabolic models and the utility of using multiple, independent modeling approaches in Chapter 1 (Kaste and Shachar-Hill, 2023a), and provide an example of putting these ideas into practice Chapter 2 (Xu et al., 2022). Again in Chapter 3, I make use of validation principles and implement the "gold-standard" validation of an FBA flux map – comparison against an MFA flux map – that I argue for in Chapter 1 (Kaste and Shachar-Hill, 2023b). Finally, in Chapter 4, I look at reaction-diffusion modeling of photosynthetic systems, identifying sources of model uncertainty that affect prior work's conclusions with regards to the efficiency of using Carbon-Concentrating Mechanisms.

In this chapter, I will review some of the takeaway results and conclusions from the studies that I have presented. I will also highlight limitations and future directions of this overall research program.

5.2. Takeaway messages and future work

5.2.1. Analyzing systems using multiple modeling paradigms can help reveal new aspects of these systems, but may be redundant with refinements or extensions of existing paradigms

In Chapter 2, I showed that complementing our ^{13}C-MFA study with a pharmacokinetics-derived polyexponential modeling approach allowed us to further refine the ^{13}C-MFA model and discover new properties of the system under consideration. However, inasmuch as both approaches are fundamentally just different mathematical formalisms describing the same biological phenomena, there should exist underlying mathematical connections between the two that, with sufficient exploration, unify them or render one or the other redundant. As referenced in Chapter 1 and described in Nöh and Wiechert, (2011) and Zheng et al., (2022), fitting time-course isotopic labeling data to incomplete metabolic network model specifications can lead to unmodeled reactions contributing unaccounted-for labeled/unlabeled atoms (referred to in Zheng et al., (2022) as "time constants"). When fitting the isotopic labeling data to generate a flux map without any metabolite pool sizes constrained by experimental measurements, the pool sizes estimates essentially capture the error introduced by these time constants (Zheng et al., 2022). In Xu et al., (2022), I use the polyexponential modeling appoach to essentially reveal these

unmodeled processes without the use of pool size data. It is possible that the routine inclusion of pool size measurements as a way of detecting model misspecifications, which I advocate for in Chapter 1, would make the polyexponential modeling approach as a way of detecting these unmodeled factors redundant, but further investigation will be needed to confirm whether this is the case.

Future work should also look into whether this polyexponential modeling method can be fruitfully applied to other systems. I use the polyexponential modeling approach in Chapter 2 to characterize how many processes acting over different time scales are influencing the labeling of CBC intermediates. I then corroborate the findings from the polyexponential modeling with our ^{13}C-MFA results. In that same study, I applied the same polyexponential modeling approach to *Nicotiana tabacum* data gathered by Xinyu Fu and colleagues in Fu et al., (2023) and found similar patterns. This corroborated our findings by suggesting that the cycling of cytosolic and vacuolar sugars may occur in *N. tabacum* as well. However, since I did not do a ^{13}C-MFA incorporating the vacuolar sugar exchange and demonstrate that this resulted in a statistically significantly better model fit, I have not yet provided equivalent evidence of the operation of such a cytosolic-to-vacuolar sugar recycling pathway in any plant other than *C. sativa*. Future work should attempt to provide such evidence, with *N. tabacum* and *A. thaliana* representing the obvious candidates for such follow-up studies due to the presence of leaf CBC intermediate isotopic labeling datasets in these systems. In order to gauge how conserved this recycling phenomenon is, though, *N. tabacum* would be the system of greater biological significance. This is because *A. thaliana* and *C. sativa* are very closely related, so if the phenomenon was found in *A. thaliana* as well, it would raise the question of whether it is conserved broadly in land plants, or just in this very specific lineage of the Brassicaceae.

5.2.2. Considering metabolic network structure in addition to omic datasets can result in drastically improved predictive power, but further work is necessary to demonstrate general applicability of this principle

As noted by Schwender et al., (2014), there is a very poor correlation between changes in transcript abundance and changes in flux when comparing a particular tissue – in the case of Schwender et al., (2014), plant embryo tissues in different growth media – under different conditions. Indeed, the large number of biochemical and regulatory processes that intervene between transcript, or even protein, accumulation make the use of transcripts or proteins as an

input data type for predicting fluxes questionable. Despite this, the work presented in Kaste and Shachar-Hill, (2023b) and Chapter 3 suggests that it is possible to extract signal from transcriptomic and proteomic datasets for flux prediction. One key difference between the approaches taken by Schwender et al., (2014) and Kaste and Shachar-Hill, (2023b) is that the latter places omic abundances in the context of the entire metabolic network, constraining the influence of the omic data by other metabolic necessities like the accumulation of measured amounts of biomass. However, the datasets analyzed by Schwender et al., (2014) and Kaste and Shachar-Hill, (2023b) are also quite different. In order to provide stronger evidence that it is the consideration of the omic data in the context of a network that allows these data, despite low correlation with differences in flux, to generate accuracy improvements, it should be demonstrated that there *is* actually a poor correlation between flux differences and omic differences in Kaste and Shachar-Hill, (2023b). This could be done with the FBA predictions in the multiple modeled tissues alone, or as part of a broader study where MFA flux maps are generated for the non-photosynthetic tissues as well.

Although successful, this method has thus far only been shown to work in *A. thaliana*, and a crucial next step to demonstrate its utility would be to show efficacy in another system. Moreover, the MFA-to-FBA flux map comparison was only possible for leaf tissues using values reported by Ma et al., (2014). If MFA flux maps of the non-photosynthetic stem and root tissues of *A. thaliana* could be generated under similar conditions, I might be able to evaluate the FBA flux maps for those tissues as well.

A number of simplifications were employed when developing and evaluating the algorithm described in Chapter 3 and Kaste and Shachar-Hill, (2023b), which could benefit from further evaluation. Although the base model of *A. thaliana* used to build the multi-tissue model evaluated in the study (Arnold and Nikoloski, 2014) contained GPR terms with detailed enzyme complex stoichiometries, this stoichiometric detail was ignored by the algorithm. Rewriting the algorithm to incorporate these stoichiometric ratios and evaluating whether it results in improved accuracy could be a fruitful future research project. Additionally, I ran into computational constraints when attempting to perform uniform random sampling of the flux solution spaces generated by our FBA optimizations. In short, the multi-tissue models I was optimizing were too large to efficiently sample. Because of this, when reporting weighted average error values, I calculated best- and worst-case (i.e., maximum and minimum possible errors) using the

maximum and minimum possible for each flux, given the model constraints, using FVA (Mahadevan and Schilling, 2003). This approach is suboptimal because not every linear combination of maximal and minimal values for each flux from FVA will represent a valid solution to the optimization problem. As a result, the real range of possible weighted average error values is almost certainly narrower than what is reported in (Kaste and Shachar-Hill, 2023b). The use of stronger computational resources to overcome the numerical difficulties posed by the uniform random sampling method, or an improved process by which the FVA-derived upper and lower bounds for each value are iteratively perturbed until a valid flux solution is generated, could be used in a future study to overcome this limitation.

5.2.3. Spatially-resolved reaction-diffusion modeling allows for powerful investigations of photosynthetic metabolism, but is limited by computational power

Along similar lines, computational limitations also affected the depth of analysis possible in the work reported in Chapter 4 and Kaste et al., (2024). As described in that chapter, a number of geometric simplifications were made to make solving the spatial reaction-diffusion models in that study numerically tractable, given the large parameter explorations I performed. In addition, these parameter explorations were limited to two-dimensional, or at most very coarse three-dimensional spaces. This stands in contrast with an extensive 10-plus-dimensional parameter exploration I am currently performing together with my colleague Anne Steensma on a forthcoming study on which I am co-first author. The relative computational simplicity of compartmental models allowed for a substantially more thorough parameter exploration. In order to achieve something similar, a follow-up on the work presented in Chapter 4 could derive analytical solutions for the models. One limitation of this approach is that derivation of such analytical solutions often requires some mathematical simplifications, as demonstrated in Fei et al., (2022), where the spontaneous (i.e., not CA-mediated) interconversion of CO_2 and HCO_3 was omitted because it was incompatible with getting an analytical solution. An alternative, if too many such simplifications would be necessary, would be to distribute the *Virtual Cell* platform's calculations to a local High Performance Computing Cluster (HPCC). By default, the Virtual Cell distributes simulation jobs to the HPCC at the University of Connecticut. However, the software imposes a limit of forty jobs per user. By implementing the *Virtual Cell* on a local university cluster, substantially larger numbers of jobs could be run simultaneously, opening the door for larger parameter explorations and deeper analysis of the model.

Even relatively sophisticated models of similar systems have had to employ simplifications and idealizations (McGrath and Long, 2014; Fei et al., 2022) and the parameter explorations and sensitivity analyses they employ are heavily hypothesis-driven, with only a small number of parameters varied and investigated. As the computational power available to research groups continues to grow, there may be great value in revisiting these existing models and rerunning analyses to better characterize the robustness of previous results and conclusions. Such investigations may reveal surprising interactions between geometric or enzymatic parameters and deepen our understanding of what factors contribute to photosynthetic efficiency and productivity.

REFERENCES

Arnold A, Nikoloski Z (2014) Bottom-up metabolic reconstruction of Arabidopsis and its application to determining the metabolic costs of enzyme production. Plant Physiology **165**: 1380–1391

Fei C, Wilson AT, Mangan NM, Wingreen NS, Jonikas MC (2022) Modelling the pyrenoid-based CO2-concentrating mechanism provides insights into its operating principles and a roadmap for its engineering into crops. Nature Plants **8**: 583–595

Fu X, Gregory LM, Weise SE, Walker BJ (2023) Integrated flux and pool size analysis in plant central metabolism reveals unique roles of glycine and serine during photorespiration. Nat Plants **9**: 169–178

Kaste JAM, Walker BJ, Shachar-Hill Y (2024) Biophysical carbon concentrating mechanisms in land plants: insights from reaction-diffusion modeling. bioRxiv 2024.01.04.574220

Kaste JAM, Shachar-Hill Y (2023a) Model validation and selection in metabolic flux analysis and flux balance analysis. Biotechnology Progress: e3413

Kaste JAM, Shachar-Hill Y (2023b) Accurate flux predictions using tissue-specific gene expression in plant metabolic modeling. Bioinformatics: btad186

Ma F, Jazmin LJ, Young JD, Allen DK (2014) Isotopically nonstationary ^{13}C flux analysis of changes in Arabidopsis thaliana leaf metabolism due to high light acclimation. Proceedings of the National Academy of Sciences of the United States of America **111**: 16967–16972

Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metabolic Engineering **5**: 264–276

McGrath JM, Long SP (2014) Can the cyanobacterial carbon-concentrating mechanism increase photosynthesis in crop species? A theoretical analysis. Plant physiology **164**: 2247–2261

Nöh K, Wiechert W (2011) The benefits of being transient: isotope-based metabolic flux analysis at the short time scale. Applied Microbiology and Biotechnology **91**: 1247–1265

Schwender J, König C, Klapperstück M, Heinzel N, Munz E, Hebbelmann I, Hay JO, Denolf P, De Bodt S, Redestig H, et al (2014) Transcript abundance on its own cannot be used to infer fluxes in central metabolism. Frontiers in Plant Science **5**: 1–16

Xu Y, Wieloch T, Kaste JAM, Shachar-Hill Y, Sharkey TD (2022) Reimport of carbon from cytosolic and vacuolar sugar pools into the Calvin-Benson cycle explains photosynthesis labeling anomalies. Proceedings of the National Academy of Sciences **119**: e2121531119

Zheng AO, Sher A, Fridman D, Musante CJ, Young JD (2022) Pool size measurements improve precision of flux estimates but increase sensitivity to unmodeled reactions

outside the core network in isotopically nonstationary metabolic flux analysis (INST-MFA). Biotechnology Journal **17**: 1–17

Chapter 6

Additional Studies: Integrative Teaching of Metabolic Modeling and Flux Analysis with Interactive Python Modules

This research was published in:

J. A. M. Kaste, A. Green, Y. Shachar-Hill, Integrative Teaching of Metabolic Modeling and Flux Analysis with Interactive Python Modules. *Biochemistry and Molecular Biology Education* 51(6): 653-661 (2023).

6.1. Preface

My interest in teaching metabolic modeling – the subject of this chapter – stems from two sources, one practical and one theoretical. The practical reason is that I found learning the underlying theory quite difficult and many concepts of central importance to metabolic modeling can take a great deal of time and effort to properly grasp. The theoretical reason is that the interrelationship between different modeling approaches – enzyme-based simulations, FBA, and MFA – is not emphasized or discussed very strongly in the literature, and the communities that have formed around these different techniques do not seem to interact extensively. As I have discussed in earlier chapters, there is a lot to be gained from comparing and integrating these different ways of looking at the same systems, so I saw great value in developing educational resources that puts this idea front and center for learners right from the get-go.

Dr. Shachar-Hill has run an annual intensive workshop series on metabolic modeling at Michigan State University for many years now. Previous iterations have used Excel spreadsheets and proprietary programs like INCA for demonstrating enzyme-based kinetic and constraint-based modeling to learners. I decided to develop interactive Python notebooks that allowed learners to more easily interface with and manipulate their metabolic modeling simulations. I enlisted the help of a bright undergraduate researcher in our lab, Antwan Green, in doing this work. I set up pre- and post-workshop surveys to assess learners' experiences and we found that the combination of the workshop lecture material and these interactive simulations resulted in positive outcomes. We decided to package these simulations together with lesson plans and lecture notes into a freely available GitHub repository for teachers and learners to access, and also wrote a manuscript describing what we put together, which I present in this chapter. I came up with the concept for these simulations and wrote most of the code for the project, with assistance from Antwan Green. I also ran all logistics related to the survey component of the study, including getting our IRB exemption for the study approved, and wrote up all of the lecture notes and lesson plans. I wrote the manuscript with input and editing from Antwan Green and Dr. Shachar Hill. The manuscript presented in this chapter has been published in the journal *Biochemistry and Molecular Biology Education* (Kaste et al., 2023).

Rather than a one-and-done study, I see this as a first step towards building a robust set of metabolic modeling learning resources. In future iterations of the workshop and in these learning materials, Dr. Shachar-Hill and I plan on interweaving the lecture and interactive sections more

seamlessly so that learners are running examples and engaging in active learning, reinforcing concepts they were just exposed to. Although the materials, as presented, already represent a big step-up from existing learning resources for these topics, there is always room for improvement.

6.2. Abstract

The modeling of rates of biochemical reactions – fluxes – in metabolic networks is widely used for both basic biological research and biotechnological applications. A number of different modeling methods have been developed to estimate and predict fluxes, including kinetic and constraint-based (Metabolic Flux Analysis and Flux Balance Analysis) approaches. Although different resources exist for teaching these methods individually, to-date no resources have been developed to teach these approaches in an integrative way that equips learners with an understanding of each modeling paradigm, how they relate to one another, and the information that can be gleaned from each. We have developed a series of modeling simulations in Python to teach kinetic modeling, Metabolic Control Analysis, ^{13}C-Metabolic Flux Analysis and Flux Balance Analysis. These simulations are presented in a series of interactive notebooks with guided lesson plans and associated lecture notes. Learners assimilate key principles using models of simple metabolic networks by running simulations, generating and using data, and making and validating predictions about the effects of modifying model parameters. We used these simulations as the hands-on computer laboratory component of a four-day metabolic modeling workshop and participant survey results showed improvements in learners' self-assessed competence and confidence in understanding and applying metabolic modeling techniques after having attended the workshop. The resources provided can be incorporated in their entirety or individually into courses and workshops on bioengineering and metabolic modeling at the undergraduate, graduate, or postgraduate level.

6.3. Introduction

Metabolic modeling provides scientists with a quantitative description of the *in vivo* rates of biochemical reactions in biological networks. These rates of biochemical reactions – fluxes – are a function of many layers of cellular regulation (transcriptional, translational, post-translational, etc.) and relate directly to the living system's functional phenotype. Understanding metabolic flux thus provides important insights into biological systems and underlies efforts to rationally modify their metabolism to suit our biotechnological needs (Nielsen, 2003).

Fluxes in metabolic pathways and networks cannot be directly measured, necessitating

the use of mathematical modeling approaches to estimate or predict them. These approaches can be broadly categorized into kinetic and constraint-based methods. Within both categories, methods exist both for predicting fluxes and for estimating them from experimental data. Kinetic methods involve simulating the dynamically changing fluxes and metabolite concentrations in a metabolic network over time (Saa and Nielsen, 2017), whereas constraint-based methods like Flux Balance Analysis (FBA) (Orth et al., 2010b) and Metabolic Flux Analysis (MFA) (Antoniewicz, 2015) estimate steady-state fluxes using linear optimization principles or experimentally-measured isotopic labeling data.

Metabolic modeling, and particularly constraint-based modeling approaches, have been used productively to aid in biotechnological applications. For example, Metabolic Flux Analysis techniques using isotopic labeling informed the engineering of the bacterium *Corynebacterium glutamicum* to produce high concentrations of lysine (Koffas et al., 2003; Koffas and Stephanopoulos, 2005; Becker et al., 2011). Flux Balance Analysis has been deployed to improve the microbial production of a number of bioproducts, including threonine (Lee et al., 2007) and valine (Park et al., 2007), and in ambitious reengineering efforts like that described in (Gleizer et al., 2019) where FBA and related methods including (Burgard et al., 2003) were used to enable engineering of normally heterotrophic *Escherichia coli* to incorporate CO2 into its biomass using a heterologously expressed Calvin-Benson Cycle. These and an increasing number of other metabolic modeling applications indicate that this is an area that is of great value to learners and practitioners in biology, biochemistry, and chemical engineering.

Related to kinetic metabolic analysis, Metabolic Control Analysis (MCA) provides mathematical tools for understanding how control over flux and internal metabolite concentrations are distributed between the enzymes in a biochemical network (Fell, 1992; Moreno-Sánchez et al., 2008). Like metabolic flux modeling and mapping the questions addressed by Metabolic Control Analysis have major biotechnological implications. We believe it therefore makes sense to introduce and teach concepts in MCA along with kinetic and constraint-based metabolic modeling techniques.

Although previous studies have described and provided resources for teaching kinetic metabolic modeling (Armando et al., 2009), FBA (Orth et al., 2010b; Chaves et al., 2022), MFA (Wong et al., 2004; Wong and Barford, 2010), and MCA (Snoep et al., 1999; Rodríguez-Caso et al., 2002; Angelani et al., 2018), there are not any published and freely available instructional

resources for introducing these toolsets to learners in an integrative and interactive fashion. Moreover, although papers and books exist describing how to experimentally approach ^{13}C-MFA (Crown et al., 2012; Dieuaide-Noubhani and Alonso, 2014; Krömer et al., 2014; Antoniewicz, 2018) or the theoretical background behind the technique (Stephanopoulos et al., 1998; Ratcliffe and Shachar-Hill, 2006), we are not aware of any dedicated and published educationally focused resources for introducing learners to the theoretical background behind label-assisted MFA. We believe introducing learners to all of these major areas of metabolic modeling together allows them to appreciate their interconnections and better evaluate what approach(es) may be useful to their own research and/or engineering goals than if they encounter them in isolation.

To address this gap in the biochemistry education literature, we developed a series of interactive Python-based Jupyter notebooks featuring exercises that give learners hands-on experience with kinetic modeling, FBA, MFA, and MCA. These notebooks were used as the hands-on laboratory exercises for the 2022 iteration of an annual metabolic modeling workshop at Michigan State University. To assess the efficacy of the workshop and the interactive exercises, surveys were distributed to participants – a mix of graduate students and postdoctoral researchers – before, immediately after, and four months after the workshop to measure self-assessed competence and confidence in metabolic modeling techniques and in the application of these techniques to learners' own research questions. Although the materials are structured with a particular sequence and timeline, the individual notebooks, paired with appropriate lecture material, contain sufficient explanation to be flexibly incorporated into different course or workshop structures.

6.4. Methods

6.4.1. Exercise development

All simulation code was written in Python and packaged and presented in Jupyter notebooks (Kluyver et al., 2016). Numpy (Harris et al., 2020) and SciPy (Virtanen et al., 2020) were used to handle data import and export and calculate control coefficients for MCA. Interactive elements were incorporated into the notebooks using the *ipywidgets* package**.** MFA simulations were run in Python using the package *mfapy* (Matsuda et al., 2021) and FBA simulations were run using *cobrapy* (Ebrahim et al., 2013). For the FBA exercises, the genome-scale model of *E. coli*'s metabolic network iJO1366 (Orth et al., 2011) was used along with a smaller "core" model of *E. coli*'s metabolic network (Orth et al., 2010a). Several example networks from (Ratcliffe and

Shachar-Hill, 2006) were adopted for demonstration purposes throughout the notebooks.

Time-courses of metabolite concentrations, fluxes, and labeling were generated in kinetic simulations featuring reversible or irreversible first-order and Michaelis-Menten kinetics. Euler's method was used to generate all concentration, flux, and labeling values. In most of the simulations that feature labeling, all metabolites are treated as having only one labelable position, so the proportion of labeled and unlabeled metabolite is tracked. In the simulations in the notebook for Day 4 (see **Table 6.1**), both one- and two-carbon molecules are present, so the quantities of unlabeled, half-labeled, and fully-labeled species for each metabolite are calculated and tracked independently to allow for comparison with ^{13}C-MFA flux map results.

6.4.2. Survey ethics and analysis

The survey component of this study was deemed exempt by the Michigan State University Office of Research Regulatory Support. Survey respondents were asked to self-assess their confidence in and understanding of kinetic and constraint-based metabolic modeling methods and the application of these methods to their own research goals on a Likert scale (Likert, 1932). Survey responses were gathered from workshop participants before, immediately after, and four-months following the workshop. The survey instruments can be found in the supplemental materials**.** One-sided Mann-Whitney U tests (Neuhäuser, 2011) were used to compare pre- and post-workshop responses, where our null hypothesis was that there is no difference between the pre- and post-workshop responses and our alternative hypothesis was that the post-workshop responses were higher than the pre-workshop responses. We evaluated each question with $\alpha = 0.05$.

6.5. Results and discussion

6.5.1 Educational Jupyter notebooks

We developed a series of four Jupyter notebooks covering various aspects of kinetic and constraint-based metabolic modeling and metabolic control analysis. A graphical summary of the different areas of metabolic modeling covered and their relationships is shown in **Figure 6.1**. In addition to learning the theory behind these methods, learners are exposed to the key concepts for successful applications of flux modeling listed below. We also note in **Table 6.1** and the lesson plans when an exercise can be used to teach one of these concepts.

- **Concept 1:** The relationship between the noise and time resolution of experimental data and the confidence one can have in parameter estimates and assumed model architectures.
- **Concept 2:** The uniqueness and identifiability of flux estimates in FBA and ^{13}C-MFA and their relationship to model complexity.
- **Concept 3:** The distribution of control over fluxes and concentrations in a network across the reactions of that network.

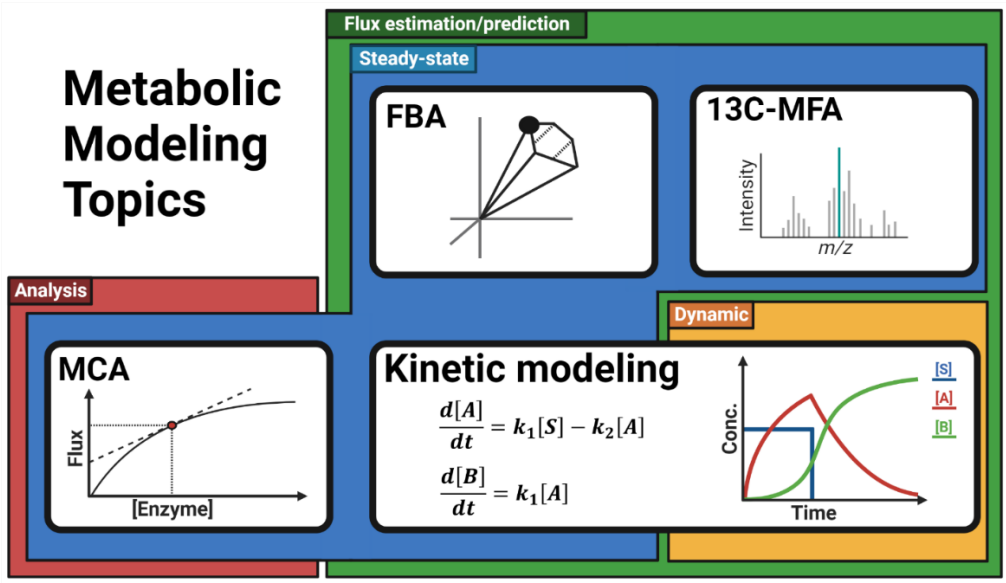


Figure 6.1: Metabolic modeling topics covered in the resources presented in this study. A majority of the techniques covered – kinetic modeling, FBA, and ^{13}C-MFA – are used to estimate or predict fluxes through a metabolic network. MCA, on the other hand, is used to analyze the relationship between enzyme activities/concentrations and metabolite or regulator concentrations on the flux through the network. Within the flux estimation/prediction techniques, kinetic modeling can be used to estimate fluxes and metabolite concentrations in systems whether they are in steady-state or not (dynamic systems where concentrations are still changing). The constraint-based modeling techniques of FBA and ^{13}C-MFA, on the other hand, rely on an assumption of metabolic steady-state, as does MCA.

These concepts are necessary both to effectively conduct any experiment or study involving flux analysis and to understanding the primary metabolic modeling literature. They are often not intuitively obvious, and the first two also receive rather little attention in the teaching or research literature. The concepts are therefore explained in the lecture notes, revisited throughout the Jupyter notebooks and demonstrated with hands-on exercises. For example, in **Exercises 4.0 – 4.2** in the **Day 4** Jupyter Notebook, learners gain insight into **Concept 1** by first using a kinetic model to generate simulated labeling data and then attempting to fit it using both

correctly and incorrectly specified network models using ^{13}C-MFA. By doing so, the learners can observe the difference in ^{13}C-MFA fits when using the correct or incorrect model specification and how this difference can be obscured even by low levels of experimental noise. This allows instructors to highlight important issues concerning data quality and to discuss model selection, which is rarely addressed in the literature (Sundqvist et al., 2022).

Table 6.1: A table describing the contents of the interactive exercises presented in this publication. Descriptions of key concepts are outlined in the text.

Day	Section(s)	Contents	Concept
1	**1.0 – 1.2**	Introduction to the Jupyterlab Interface.	
	2.0 – 2.2	Exploration of a simulation demonstrating first-order kinetics.	
	3.0 – 3.1	Exercise on inferring kinetic parameters from example datasets.	1
	4.0	Exercise demonstrating the relationship between model architecture and the information contained in each datapoint.	1
	5.0 – 5.1	Introduction to metabolic steady-state and the utility of labeling data.	
2	**1.0**	Introduction to reversible first-order kinetic models	
	2.0	Exercise on inferring model parameters in the presence of reversibility	1
	3.0 – 3.4	Exploration of metabolic control analysis, including calculation of flux and concentration control coefficients as well as elasticities.	3
	4.0	Comparison of results gathered in **3.0 – 3.4** "by hand" with results from an automated MCA script.	1
3	**1.0 – 1.2**	Metabolic control analysis with branching networks, negative control coefficients, and modeling a system with an incomplete network description.	3
	2.0 – 2.2	Kinetic modeling with Michaelis-Menten kinetics.	
	3.0	Fitting a dataset using either first-order or Michaelis-Menten kinetics in the presence or absence of noise.	1
	4.0	Kinetic modeling with reversible Michaelis-Menten kinetics.	
	5.0	Using MCA to calculate response coefficients.	3
4	**1.0 – 1.2**	A kinetic simulation that incorporates labeling dynamics, for comparison with ^{13}C-MFA and FBA.	
	2.0	Introduction to FBA modeling.	2
	3.0 – 3.3	Introduction to FVA and randomized sampling methods in FBA.	2
	4.0 – 4.2	Introduction to ^{13}C-MFA and comparison with results from **1.0 – 1.2.**	2, 1
	5.0	Discussion about incorporating metabolic modeling into one's own work and/or research.	

The subjects covered in the sections of each notebook with the timeline for a 4 day workshop are given in **Table 6.1.** On the first and second days, learners are given an extensive introduction to kinetic modeling theory and exercises before learning about MCA, FBA, and ^{13}C-MFA. We do this to allow learners to gain both a theoretical and practical understanding of the dynamic ways that matter moves through biochemical networks. The hands-on experience exposes learners to the sometimes surprisingly complex behavior of even simple networks governed by systems of Ordinary Differential Equations (ODEs). This is aimed at giving learners

a strong sense of the dynamics of metabolic systems before learning about steady-state approaches, in which simplifications of the kinetic state allow powerful analyses in ^{13}C-MFA and FBA. MCA is explored in the second and third days and MCA calculations of flux- and concentration- control coefficients are discussed. Control coefficients are connected to the understanding of reversible first-order kinetics participants gained from the preceding kinetic modeling exercises. Lastly, participants are introduced to constraint-based methods by analyzing the same network structure using kinetic modeling, FBA, and ^{13}C-MFA. This highlights the different inputs needed and the resulting outputs from each technique. To our knowledge, this is the first such cross-comparison of different metabolic modeling techniques presented in the teaching literature, and we believe this will be of value to instructors introducing this material to their students and trainees.

Interactive sliders and drop-down menus were incorporated into all of the notebooks to allow learners to modify parameters, run simulations and visualize their results. This allows learners to expose the underlying simulation code and for those with a modest background in Python or general coding to see how the simulations function and potentially to modify the model structures. By default the code is not visible, making the notebooks approachable for participants interested in using metabolic modeling without engaging with the underlying code. We believe that the incorporation of these interactive modules into the notebooks will make the resources presented in this publication useable by learners with little to no coding knowledge.

In writing the notebooks, special attention was given to commenting the Python code used to run the simulations and interactive interface elements. We believe the extensive commenting used in these notebooks, together with the use of intuitive and easy-to-understand methods for implementing the simulations will make the notebooks both easy for instructors to adopt and for learners interested in the underlying code to understand it. This is in contrast to software like COPASI that, while very powerful, obscure the underlying simulation logic (Hoops et al., 2006). Installation and compatibility issues are commonplace when using computational resources, particularly when workshop or class participants are asked to run code or software on their own computers. To further ensure maximal useability of these resources by instructors, detailed installation instructions for Windows, MacOS, and Linux systems with the specific version numbers needed to successfully run all of the notebooks provided with the notebooks.

Table 6.2: Quantitative pre- and post-workshop survey results evaluating learners' self-assessed confidence and competence in metabolic modeling techniques.

Question	Pre-workshop Median	Post-workshop Median	Significant improvement?[a]
I feel confident in applying and incorporating metabolic modeling techniques to my research question(s).	2	3	Not Significant
I feel confident in evaluating the results of a metabolic modeling study or a study that incorporates metabolic modeling.	2	3	Significant
I feel confident in identifying metabolic modeling techniques and software that I can apply to my research question(s).	2	4	Significant
I understand the purpose(s) of metabolic modeling.	4	4	Significant
I can describe kinetic metabolic modeling, what information it can provide, and its limitations.	3	4	Significant
I can describe Metabolic Flux Analysis, what information it can provide, and its limitations.	3	4	Significant
I can describe Flux Balance Analysis, what information it can provide, and its limitations.	2	4	Significant
I understand the data types I would need to carry out kinetic metabolic modeling.	2.5	4	Significant
I understand the data types I would need to carry out Metabolic Flux Analysis.	2.5	4	Significant
I understand the data types I would need to carry out Flux Balance Analysis.	2	4	Significant
I can name the language(s) or software package(s) I would use to incorporate metabolic modeling into my own research.	2	4	Significant
I can critically evaluate the application and results of metabolic modeling in publications and presentations relevant to my area of research.	3	4	Significant

[a]Statistically significant improvement was defined by rejection of the null hypothesis by the one-sided Mann-Whitney U test (Neuhäuser, 2011) at $\alpha = 0.05$.

6.5.2. Implementation in workshop and survey results

The Jupyter notebooks were incorporated into a four-day workshop held at Michigan State University in May 2022. Participants in the workshop included graduate students and postdoctoral researchers. Each day of the workshop consisted of three hours of lecture in the morning and a three-hour hands-on period for computational exercises. Due to time constraints and interest among the participants in constraint-based modeling approaches – particularly label-assisted flux mapping using MFA – the third day's notebook exercises were omitted and replaced with the fourth day's exercises on constraint-based modeling. The last day of the workshop was used for an open-ended discussion about participants' research aims and how they could incorporate what they learned in the workshop into their own work. For instructors interested in incorporating not only the computational resources developed for the workshop, but

also all or portions of the lecture material, detailed lecture notes have been provided online at https://github.com/Gibberella/Metabolic-Modeling-Lessons.

The pre- and post-workshop survey results suggest that participants felt they gained greater confidence in and knowledge of metabolic modeling over the course of the workshop **(Table 6.2)**. Our survey evaluated participants' self-assessed confidence and competence but did not ask participants to attribute their comprehension gains to the lecture or hands-on components. In a free-response question ("What did you find useful about the workshop?"), one participant responded, "Understanding what goes into metabolic modeling, learning how to critically appraise these models in published literature, and beginning to learn how to implement them into our own projects." In response to that same question, another participant focused more specifically on FBA: "The hands-on use of *cobrapy* was very helpful. This helped me understand how one goes about metabolic modeling." It should be noted, however, that the sample sizes for the study were small and we had fewer respondents in the post-workshop survey than the pre-workshop survey (N = 12 in the pre-workshop survey and N = 7 in the post-workshop survey). Because of this, the results may be skewed due to survivorship bias from learners who were either no longer interested in the topic or unhappy with the presentation of the material leaving and not participating in the post-workshop survey.

Multiple respondents noted that they would have liked to have worked with real datasets in the exercises rather than simulated ones. Given the modifiability and extensive annotation of the notebooks provided, we encourage instructors using the provided resources to add analyses of real datasets that are relevant to their specific audience. We believe this will help provide real-world context for learners as they carry out the exercises.

Although we have packaged and used the materials presented in the context of an intensive workshop, we believe the materials can be adapted to a variety of teaching circumstances. The Jupyter-based simulations could be used for computer lab sessions in a semester-long course, for example, or used as an interactive demonstration in a lecture setting. With the relevant theory taught beforehand, these resources may also be appropriate for undergraduate learning. As noted, the extensive annotation of the code paired with the easy-to-use graphical interface for the exercises also makes them suitable for both learners with extensive and with no prior knowledge of programming.

6.6. Conclusions

Recognizing the absence of resources for teaching the major areas and techniques of metabolic modeling and flux analysis in an integrative fashion, we have developed a set of resources that should be readily adoptable by instructors, students, and researchers alike to teach and learn. By emphasizing the legibility and cross-platform useability of our code, we hope the resources presented in this study can be used and incorporated by the broader teaching community into other workshop and class settings.

6.7. Data and code availability statement

All code, documentation, lecture notes, and lesson plans developed for the present study can be found at https://github.com/Gibberella/Metabolic-Modeling-Lessons.

6.8. Acknowledgements

This research was supported by the Office of Science (BER), U.S. Department of Energy, Grant no DE-SC0018269 (J.A.M.K., A.G., Y.S-H.). This work is supported, in part, by the NSF Research Traineeship Program (Grant DGE-1828149) to J.A.M.K. This publication was also made possible by a predoctoral training award to J.A.M.K. from Grant T32-GM110523 from National Institute of General Medical Sciences (NIGMS) of the NIH. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIGMS or NIH. We would like to acknowledge the Pathways to Research program at Michigan State University for its support of A.G. We would also like to acknowledge Veronica Greve for her input on the survey component of this study.

REFERENCES

Angelani CR, Carabias P, Cruz KM, Delfino JM, de Sautu M, Espelt MV, Ferreira-Gomes MS, Gómez GE, Mangialavori IC, Manzi M, et al (2018) A metabolic control analysis approach to introduce the study of systems in biochemistry: the glycolytic pathway in the red blood cell. Biochemistry and Molecular Biology Education 46: 502–515

Antoniewicz MR (2015) Methods and advances in metabolic flux analysis: a mini-review. Journal of Industrial Microbiology and Biotechnology 42: 317–325

Antoniewicz MR (2018) A guide to ^{13}C metabolic flux analysis for the cancer biologist. Experimental and Molecular Medicine. doi: 10.1038/s12276-018-0060-y

Armando RP, Francisca SJ, Medina MÁ (2009) First steps in computational systems biology: A practical session in metabolic modeling and simulation. Biochemistry and Molecular Biology Education 37: 178–181

Becker J, Zelder O, Häfner S, Schröder H, Wittmann C (2011) From zero to hero-Design-based systems metabolic engineering of Corynebacterium glutamicum for l-lysine production. Metabolic Engineering 13: 159–168

Burgard AP, Pharkya P, Maranas CD (2003) OptKnock: A Bilevel Programming Framework for Identifying Gene Knockout Strategies for Microbial Strain Optimization. Biotechnology and Bioengineering 84: 647–657

Chaves GL, Batista RS, de Sousa Cunha J, Altmann DL, da Silva AJ (2022) Teaching cellular metabolism using metabolic model simulations. Education for Chemical Engineers 38: 97–109

Crown SB, Ahn WS, Antoniewicz MR (2012) Rational design of ^{13}C-labeling experiments for metabolic flux analysis in mammalian cells. BMC Systems Biology 6: 43

Dieuaide-Noubhani M, Alonso AP (2014) Plant metabolic flux analysis. Springer

Ebrahim A, Lerman JA, Palsson BO, Hyduke DR (2013) COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Systems Biology. doi: 10.1186/1752-0509-7-74

Fell DA (1992) Metabolic control analysis: A survey of its theoretical and experimental development. Biochemical Journal 286: 313–330

Gleizer S, Ben-Nissan R, Bar-On YM, Antonovsky N, Noor E, Zohar Y, Jona G, Krieger E, Shamshoum M, Bar-Even A, et al (2019) Conversion of Escherichia coli to Generate All Biomass Carbon from CO_2. Cell 179: 1255-1263.e12

Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, et al (2020) Array programming with NumPy. Nature 585: 357–362

Hoops S, Gauges R, Lee C, Pahle J, Simus N, Singhal M, Xu L, Mendes P, Kummer U (2006) COPASI - A COmplex PAthway SImulator. Bioinformatics 22: 3067–3074

Kaste JAM, Green A, Shachar-Hill Y (2023) Integrative teaching of metabolic modeling and flux analysis with interactive python modules. Biochemistry and Molecular Biology Education 51: 653–661

Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, Kelley K, Hamrick J, Grout J, Corlay S, et al (2016) Jupyter Notebooks—a publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas - Proceedings of the 20th International Conference on Electronic Publishing, ELPUB 2016 87–90

Koffas MAG, Jung GY, Stephanopoulos G (2003) Engineering metabolism and product formation in Corynebacterium glutamicum by coordinated gene overexpression. Metabolic Engineering 5: 32–41

Koffas MAG, Stephanopoulos G (2005) Strain improvement by metabolic engineering: Lysine production as a case study for systems biology. Current Opinion in Biotechnology 16: 361–366

Krömer JO, Nielsen LK, Blank LM (2014) Metabolic Flux Analysis. Methods in Molecular Biology. New York, NY

Lee KH, Park JH, Kim TY, Kim HU, Lee SY (2007) Systems metabolic engineering of Escherichia coli for L -threonine production. Molecular Systems Biology. doi: 10.1038/msb4100196

Likert R (1932) A technique for the measurement of attitudes. Archives of Psychology 140: 1–55

Matsuda F, Maeda K, Taniguchi T, Kondo Y, Yatabe F, Okahashi N, Shimizu H (2021) mfapy: An open-source Python package for ^{13}C-based metabolic flux analysis. Metabolic Engineering Communications 13: e00177

Moreno-Sánchez R, Saavedra E, Rodríguez-Enríquez S, Olín-Sandoval V (2008) Metabolic Control Analysis: A tool for designing strategies to manipulate metabolic pathways. Journal of Biomedicine and Biotechnology. doi: 10.1155/2008/597913

Neuhäuser M (2011) Wilcoxon–Mann–Whitney Test BT - International Encyclopedia of Statistical Science. *In* M Lovric, ed,Springer Berlin Heidelberg, Berlin, Heidelberg, pp 1656–1658

Nielsen J (2003) It Is All about Metabolic Fluxes. Journal of Bacteriology 185: 7031–7035

Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, Palsson B (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism-2011. Molecular Systems Biology 7: 1–9

Orth JD, Fleming RMT, Palsson BØ (2010a) Reconstruction and Use of Microbial Metabolic Networks: the Core Escherichia coli Metabolic Model as an Educational Guide. EcoSal Plus. doi: 10.1128/ecosalplus.10.2.1

Orth JD, Thiele I, Palsson BO (2010b) What is flux balance analysis? Nature Biotechnology 28: 245–248

Park JH, Lee KH, Kim TY, Lee SY (2007) Metabolic engineering of Escherichia coli for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation. Proceedings of the National Academy of Sciences of the United States of America 104: 7797–7802

Ratcliffe RG, Shachar-Hill Y (2006) Measuring multiple fluxes through plant metabolic networks. Plant Journal 45: 490–511

Rodríguez-Caso C, Sánchez-Jiménez F, Medina MÁ (2002) A modeling and simulation approach to the study of metabolic control analysis. Biochemistry and Molecular Biology Education 30: 169–171

Saa PA, Nielsen LK (2017) Formulation, construction and analysis of kinetic models of metabolism: A review of modelling frameworks. Biotechnology Advances 35: 981–1003

Snoep JL, Mendes P, Westerhoff HV (1999) Teaching Metabolic Control Analysis and kinetic modelling: Towards a portable teaching module. The Biochemist 25–28

Stephanopoulos G, Aristidou AA, Nielsen J (1998) Metabolic engineering: principles and methodologies. Academic Press.

Sundqvist N, Grankvist N, Watrous J, Mohit J, Nilsson R, Cedersund G (2022) Validation-based model selection for ^{13}C metabolic flux analysis with uncertain measurement errors. PLOS Computational Biology 18: e1009999

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods 17: 261–272

Wong KW, Barford JP, Porter JF (2004) Understanding the practical consequences of metabolic interactions - A software package for teaching and research. IFAC Proceedings Volumes (IFAC-PapersOnline) 37: 315–320

Wong KWW, Barford JP (2010) Metstoich: Teaching quantitative metabolism and energetics in biochemical engineering. Chemical Engineering Education 44: 147–156

APPENDIX D: Supplemental Material for Chapter 6

SURVEY INSTRUMENTS

Pre-workshop survey

1. I am a …
    a. Undergraduate
    b. Graduate student
    c. Postdoctoral researcher
    d. Faculty member
    e. None of the above
2. I would describe myself as a …
    a. Biologist
    b. Biochemist
    c. Computational Scientist
    d. None of the above
3. I feel confident in applying and incorporating metabolic modeling techniques to my research question(s)
    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)
4. I feel confident in evaluating the results of a metabolic modeling study or exercise.
    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)
5. I feel confident in identifying metabolic modeling software and techniques that I can apply to my research question(s)
    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)
6. I understand the purpose(s) of metabolic modeling.
    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)
7. I can describe kinetic metabolic modeling and its limitations.
    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)
8. I can describe Metabolic Flux Analysis and its limitations.
    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)
9. I can describe Flux Balance Analysis and its limitations.
    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)
10. I understand the data types I would need to carry out kinetic metabolic modeling.

a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)

11. I understand the data types I would need to carry out Metabolic Flux Analysis.
    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)
12. I understand the data types I would need to carry out Flux Balance Analysis.
    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)
13. I can name the language(s) or software package(s) I would use to incorporate metabolic modeling into my own research
    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)
14. I can critically evaluate the application and results of metabolic modeling in publications and presentations relevant to my area of research.
    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)

Post-workshop survey

1. I feel confident in applying and incorporating metabolic modeling techniques to my research question(s)
    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)
2. I feel confident in evaluating the results of a metabolic modeling study or exercise.
    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)
3. I feel confident in identifying metabolic modeling software and techniques that I can apply to my research question(s)
    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)
4. I understand the purpose(s) of metabolic modeling.
    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)
5. I can describe kinetic metabolic modeling and its limitations.
    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)
6. I can describe Metabolic Flux Analysis and its limitations.
    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)
7. I can describe Flux Balance Analysis and its limitations.
    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)
8. I understand the data types I would need to carry out kinetic metabolic modeling.
    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)
9. I understand the data types I would need to carry out Metabolic Flux Analysis.

    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)
10. I understand the data types I would need to carry out Flux Balance Analysis.
    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)
11. I can name the language(s) or software package(s) I would use to incorporate metabolic modeling into my own research
    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)
12. I can critically evaluate the application and results of metabolic modeling in publications and presentations relevant to my area of research.
    a. (Strongly Disagree), (Disagree), (Neutral), (Agree) (Strongly Agree)
13. What did you find useful about the workshop?
    a. Free response
14. What did you not find useful about the workshop?
    a. Free response
15. What changes to the workshop do you think would improve it in future iterations?
    a. Free response

4-months after survey

1. If you have had one or more opportunities to apply any of the knowledge you gained in the metabolic modeling workshop, please share your experience(s). If you have not applied any of the knowledge you gained in the metabolic modeling workshop and there are specific reasons why, please share.
        i. Free response

Chapter 7

Additional Studies: Topological data analysis reveals a core gene expression backbone that defines form and function across flowering plants

This research was published in:

S. Palande, **J. A. M. Kaste**, M. D. Roberts, K. S. Aba, C. Claucherty, J. Dacon, R. Doko, T. B. Jayakody, H. R. Jeffery, N. Kelly, A. Manousidaki, H. M. Parks, E. M. Roggenkamp, A. M. Schumacher, J. Yang, S. Percival, J. Pardo, A. Y. Husbands, A. Krishnan, B. L. Montgomery, E. Munch, A. M. Thompson, A. Rougon-Cardoso, D. H. Chitwood, R. VanBuren, *Model validation and selection in metabolic flux analysis and flux balance analysis*. *PLoS Biology*, 21(12): e3002397 (2023).

7.1. Preface

The work described in this chapter was born out of a project initially conceived by Dr. Bob VanBuren and Dr. Dan Chitwood, who introduced it as a class project to the students of HRT841: Foundations in Computational and Plant Sciences, the first of a two-part course series given to students in the NRT-IMPACTS fellowship program at Michigan State University. The basic premise put forward to us students was that Topological Data Analysis (TDA) techniques could be used to analyze gene expression patterns in publicly available plant RNA-seq datasets. After the whole class deliberated, we decided that we would delimit our study to flowering plants. What followed was extensive data collection from the NCBI SRA followed by alignment, quantification, curation, and meta data organization to put together a coherent and expansive dataset. For the TDA portion of the study, which was helmed by postdoctoral researcher Dr. Saurabh Palande, we modeled our approach after a previous study that looked at organ-specific gene expression patterns in diverse animal lineages. This study used a statistical technique called Surrogate Variable Analysis (SVA) to help minimize the impact of unmodeled technical variables on their analyses. Myself and three other graduate students in the class – Miles Roberts, Kenia Segura-Aba, and Andriana Manousidaki – took on the task of applying SVA to our dataset.

SVA turned out to not be a suitable technique to apply to the dataset gathered for this study, although the process of attempting to use it was nonetheless highly informative. Despite the failure to use SVA, TDA applied to the "uncorrected" expression dataset we had gathered yielded some very interesting results. After the conclusion of both HRT841: Foundations in Computational and Plant Sciences and CSS844: Frontiers in Computational and Plant Sciences, I continued to work on interpreting and contextualizing the specific genes identified by the TDA analysis. This analysis ended up comprising a substantial portion of the results section of the study, which has been published in PLoS Biology and on which I am co-first author (Palande et al., 2023).

Although this chapter may seem like a non-sequitur from the rest of my work and came out of a class project, as you will see, the study concerns itself greatly with conserved patterns of gene expression across different plant tissues and stresses. This aspect of the study was of great interest to me due to the importance of tissue-specific expression patterns to the method I developed and presented in Chapter 3. Indeed, if such patterns can be consistently identified and then incorporated into metabolic modeling predictions using the method from Chapter 3 or

something similar, multi-tissue FBA flux predictions in novel plant systems could potentially be improved. Although the work presented in this chapter does not go so far as to identify any such patterns, it represents a first step in this direction, and as such relates to the broader aims of this thesis.

7.2. Abstract

Since they emerged approximately 125 million years ago, flowering plants have evolved to dominate the terrestrial landscape and survive in the most inhospitable environments on earth. At their core, these adaptations have been shaped by changes in numerous, interconnected pathways and genes that collectively give rise to emergent biological phenomena. Linking gene expression to morphological outcomes remains a grand challenge in biology, and new approaches are needed to begin to address this gap. Here, we implemented topological data analysis (TDA) to summarize the high dimensionality and noisiness of gene expression data using lens functions that delineate plant tissue and stress responses. Using this framework, we created a topological representation of the shape of gene expression across plant evolution, development, and environment for the phylogenetically diverse flowering plants. The TDA-based Mapper graphs form a well-defined gradient of tissues from leaves to seeds, or from healthy to stressed samples, depending on the lens function. This suggests that there are distinct and conserved expression patterns across angiosperms that delineate different tissue types or responses to biotic and abiotic stresses. Genes that correlate with the tissue lens function are enriched in central processes such as photosynthetic, growth and development, housekeeping, or stress responses. Together, our results highlight the power of TDA for analyzing complex biological data and reveal a core expression backbone that defines plant form and function.

7.3. Introduction

Over 300,000 gene expression datasets have been collected for thousands of diverse plant species spanning over 900 million years of divergence (Lim et al., 2022). This wealth of publicly available datasets spans ecological niches, species, developmental stages, tissues, stresses, and even single cells, providing a largely untapped reservoir of biological information. These diverse datasets provide an opportunity to link insights from various biological disciplines, including ecology, development, physiology, genetics, evolution, biochemistry, and cell biology through a common computational and mathematical framework. These gene expression datasets have been

analyzed individually for specific experiments and hypotheses, but large-scale meta-analyses across the publicly available expression datasets are largely nonexistent for plants.

Beyond a common currency that links the subdisciplines of biology, gene expression links its emergent levels. Below gene expression, the genome gives rise to transcriptional networks and protein interactions that are directly responsible for the complexity of gene expression. Above it, gene expression orchestrates cell-specific expression and the development of the organism itself, impacting phenotypes ranging from physiology to plasticity that propagate further to the population, community, and ecological levels. These features, from molecular (DNA, promoter sequences, -omics datasets) to the organismal, population, and ecological levels (life history traits, climatic data from species distributions, etc.) have been used in the past as labels and predicted outputs of machine learning models (Washburn et al., 2019; Azodi et al., 2020). The structure—the shape—of gene expression in flowering plants is therefore a constraint that is formed by and impacts biological phenomena below and above it, respectively.

Data visualization lies at the heart of exploratory data analysis and provides us with a powerful tool for generating hypotheses that can later be examined using standard statistical techniques. In the era of Big Data, the development of new data visualization pipelines has become increasingly important due to the high dimensionality of the datasets generated and the need to identify patterns and structures that can then become targets for more focused studies. Just as we can look upon the shape of a leaf and derive insights into how it functions from multiple perspectives (developmental, physiological, and evolutionary), we can visualize the shape of any type of data using a Mapper graph (Singh et al., 2007). The Mapper algorithm takes as input a filter function that describes a biological aspect of the data and uses mathematical ideas of shape to return a graph that reveals the underlying structure of the data. Even abstract data types like gene expression datasets, therefore, have a shape that we can visualize and derive insights from. For example, Nicolau and colleagues visualized the structure of breast cancer gene expression, identifying 2 distinct branches with differing underlying genotypes and prognostic outcomes that traditional statistical and bioinformatic approaches fail to resolve (Nicolau et al., 2011). This structure was revealed using a pairwise correlation distance matrix as input and modeling of the residuals of each sample from a vector of healthy gene expression as a measure of disease severity. In a second example, using a lens of developmental stage on single-cell RNASeq data, Rizvi and colleagues visualized the underlying structure of gene expression

during murine embryonic stem cell differentiation, revealing transient states as well as asynchronous and continuous transitions between cell types (Rizvi et al., 2017). In both examples, Mapper allowed the shape of data, through a selected lens, to be visualized. The resulting topology of the graph—in the form of loops, branch points, or flares—allowed previously hidden structures to be seen and novel insights to be derived. Loops, branch points, and flares in topological data analysis (TDA)-based Mapper graphs are visual representations of patterns, transitions, and outliers in the data. They provide insights into the topological structure and organization of the data, helping to identify clusters, subgroups, and potential anomalies. Loops represent recurring patterns or relationships in the data, branch points occur when different subsets of data points exhibit distinct topological characteristics, and flares typically indicate outliers or subgroups within a larger cluster and can help identify regions of interest or anomalous behavior in the data.

Surveys of gene expression capture tens of thousands of data points per sample, and this high dimensionality can be represented by a unique shape that underlies emergent biological features. This shape explains gene expression along evolutionary, developmental, and environmental trajectories, leading to innovations that have marked the successful adaptation and proliferation of plant species. To visualize this shape is to better understand what transcriptional profiles are possible and to know the boundaries or constraints that permit or limit gene expression. Here, we analyzed publicly available gene expression profiles across diverse flowering plant families and visualized the underlying structure of gene expression in plants as a graph using the Mapper algorithm. We identified unique topological shapes of plant gene expression when viewed through lenses that delineate different tissue or stress responses. These complex, emergent patterns were largely hidden by biological complexity and sample heterogeneity. Our results demonstrate the ability of Mapper to uncover these patterns in high-dimensional plant gene expression datasets and its potential as a powerful tool for biological hypothesis generation.

7.4. Results

7.4.1. A representative catalog of flowering plant gene expression

The vast number of gene expression datasets in plants provides a unique opportunity to search for patterns of conservation and divergence throughout angiosperm evolution, across developmental time, tissues, and stress response axes. Previous studies have tried to find

common signatures that define different plant tissues or responses to abiotic/biotic stresses, but these have been limited in species breadth (Proost and Mutwil, 2018), depth (Julca et al., 2021), or had limited downstream analyses (Zhang et al., 2020a). Here, we reanalyzed public expression data on the NCBI sequence read archive (SRA) and applied a topological data analysis method to map the shape of gene expression in plants. We included 54 species that captured the broadest phylogenetic diversity within angiosperms while maximizing the breadth of expression at the tissue and stress levels **(Fig 7.1A)**. This includes 44 eudicots across 13 families and 9 monocot species across 2 families, as well as *Amborella trichocarpa*, which is sister to the rest of angiosperms. Raw reads were downloaded, cleaned, and reprocessed through a common RNAseq pipeline to remove artifacts related to the different algorithms and downstream analyses used by each group. After filtering datasets with low read mapping, our final set of expression data includes 2,671 samples across 7 distinct developmental tissues and 9 stress classifications for 54 species.

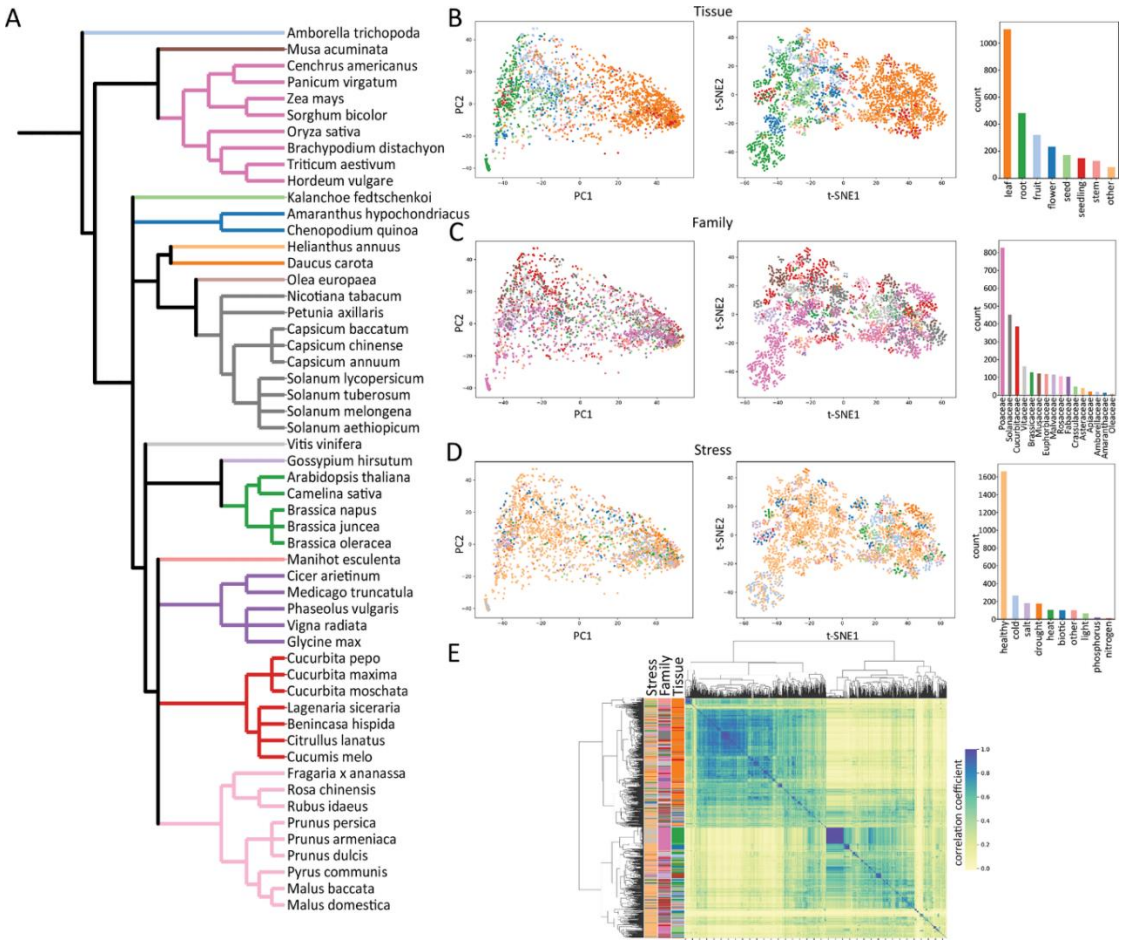


Figure 7.1: Dimensional space of plant gene expression across evolution, development, and stress. (**A**) Representative phylogeny of the 54 plant species included in this study. Nodes (species) are colored by plant family as denoted in Fig **7.1C**. Dimensionality reduction of all samples by principal components (left) and t-SNE (right) are shown for tissue type (**B**), plant family (**C**), and abiotic/biotic stress (**D**). Individual samples are quantified and colored by tissue, family, and stress as shown in the respective bar plots. (**E**) Hierarchical clustering of samples with various biological features highlighted (stress, family, and tissue). Raw expression data underlying the graphs in this figure can be found in S7 Dataset, and code to regenerate analyses can be found in https://zenodo.org/records/8428609 (Palande, 2023).

To facilitate comparisons of gene expression across species, we limited our analysis to a set of 6,328 orthologous low-copy genes that were conserved across all 54 plant species using Orthofinder (Emms and Kelly, 2015). These sets of orthologous genes or orthogroups are mostly single copy in our diploid species and scale with ploidy in polyploid species. The orthogroups are conserved across a diverse selection of Angiosperm lineages and correspond to well-conserved biological processes. Gene ontology (GO) term enrichment analysis on the *Arabidopsis thaliana* loci associated with these orthogroups show enrichment for basic

biological functions like “DNA replication initiation” and “tRNA methylation” at the top of the list of enriched GO terms, as well as functions specific to photosynthetic organisms like “photosystem II assembly,” and “tetraterpenoid metabolic process.” Although the remaining orthogroups contain significant biological information, they were excluded from analysis as multigene families typically have diverse functions with divergent expression profiles that would conflate downstream comparative analyses.

The transcript per million (TPM) counts were summed for all genes within an orthogroup for a given species and merged into a single dataframe to create a final matrix of 6,335 orthologs by 2,671 samples. Principal component analysis (PCA) (Pearson, 1901) and t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten and Hinton, 2008) based dimensionality reduction show some separation of samples by different biological factors **(Fig 7.1)**. The sample space is most clearly delineated by tissue, where both PC1 (explaining 25.4% variation) and t-SNE1 separate the samples into a gradient from root to leaf tissues with other plant tissues sandwiched in between **(Fig 7.1B and 7.1D).** This distribution largely correlates with tissue function, as the sink tissues of flowers, seeds, and fruits resolve closer to the root samples along t-SNE1 and PC1. No tissue type is separated fully by either dimensionality reduction approach. Samples from the 16 plant families are distributed throughout the dimensional space, suggesting that family- or species-level traits are not masking emergent features of distinct tissues **(Fig 7.1C)**. Interestingly, abiotic and biotic stresses are similarly distributed throughout the dimensional space, with no clear grouping of the same stress across species or individual experiments. This could be due to intrinsic differences in how individual species respond to stress or to differences in the way stress experiments are carried out by different research groups. To account for batch effects and the influence of unmodeled factors, we applied surrogate variable analysis (SVA) to generate estimates of surrogate variables and their effects on our expression matrices. We identified 24 surrogate variables within the dataset, but these latent variables were intrinsically linked to the primary factors in our study (e.g., stress, tissue, and family). Removing surrogate variables would have masked much of the biology we were attempting to quantify, so we chose not to use these “data cleaning” approaches (see **Appendix D, Text S7.2.A** for more details).

7.4.2. Topological data analysis and the shape of plant gene expression

Traditional dimensionality reduction and hierarchical clustering provided some degree of separation, but they were unable to delineate samples by stress or to identify expression patterns related to biological function. This may be related to residual heterogeneity, noise, or because of the inherent biological complexity that underlies plant evolution and function. To test these possibilities, we used a topological data analysis approach to map the shape of our data. TDA was implemented using Mapper (Tauzin et al., 2021), which provides a compact, multiscale representation of the data that is well suited for visual exploration and analysis. Mapper is particularly well suited for genomics data as these datasets typically have extremely high dimensionality and sparsity (Nicolau et al., 2011). To construct mapper graphs from our gene expression data, we created 2 different lenses of tissue and stress, adopting an approach similar to Nicolau and colleagues' **(Fig 7.2A–2E)**. To create the stress lens, we first identified all the healthy samples from the dataset and fit a linear model to them **(Fig 7.2; see Methods)**. This model serves as the idealized healthy orthogroup expression. We then projected all the samples onto this linear model and obtained the residuals. These residuals measure the deviation of the sample gene expression from the modeled healthy expression, and the lens function is simply the length of the residual vector.

The obvious separation between leaf and root samples in the dimension reduction plots supports a strong photosynthetic versus nonphotosynthetic divide. We used this observation to create a binary tissue lens in the same way as the stress lens. We identified all the photosynthetic samples (i.e., leaf tissue) and created an idealized expression profile by fitting a linear model to these expression profiles **(Fig 7.2)**. We then projected all the samples onto this linear model and obtained the residuals to establish the lens function by tissue. To define the cover for each lens, we divided the range of the lens function into intervals of uniform length, with the same amount of overlap between adjacent intervals. We experimented with a range of value lengths of the intervals and the size of the overlap to identify the values that produced relatively stable mapper graphs. The clustering was performed using DBSCAN, a commonly used clustering algorithm in Mapper (Pathak et al., 2021).

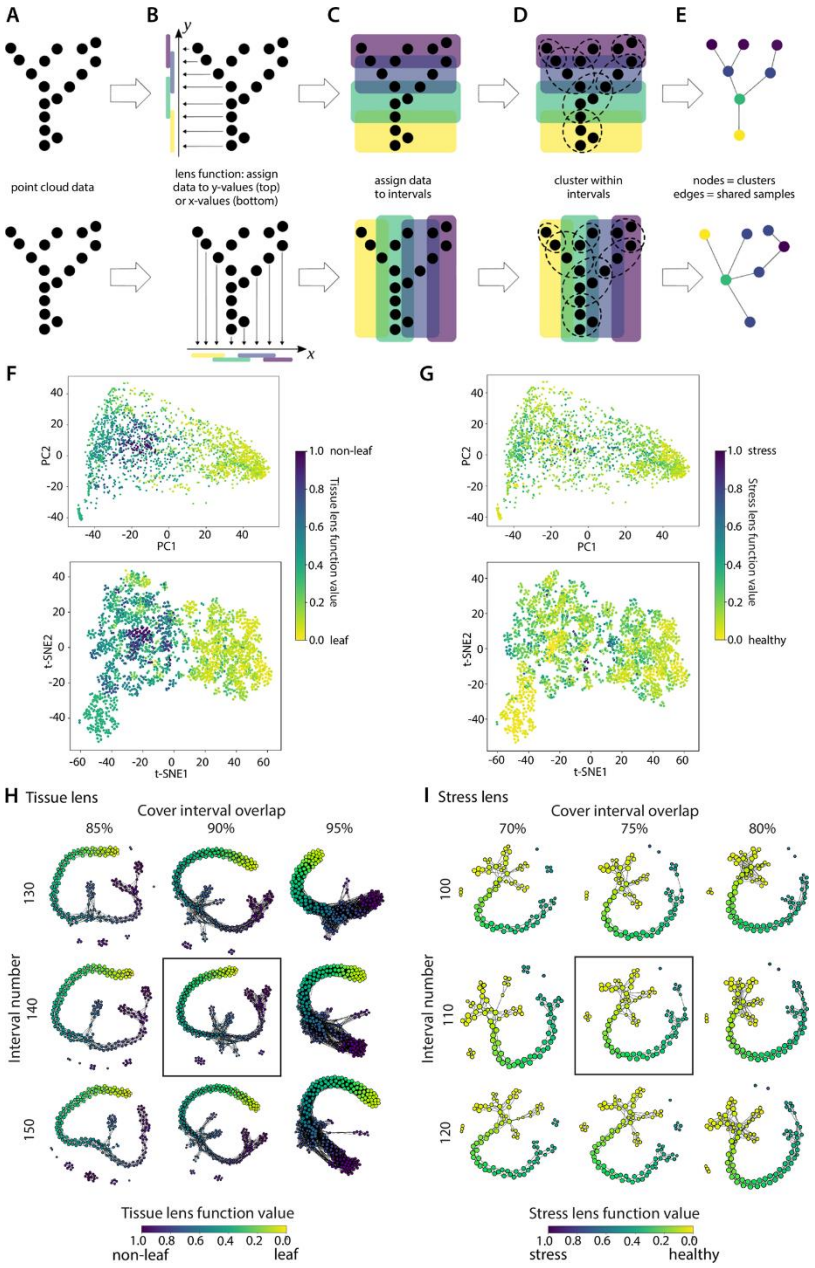


Figure 7.2: Topology-based Mapper graphs and the shape of gene expression in plants. Overview of Mapper graph construction and lens functions (**A**-**E**). The lens function value of each sample is shown in the principal component (top) and t-SNE (bottom) based dimensional reduction from **Fig 7.1** for the tissue (**F**) and stress lens (**G**). Mapper graphs across variable cover intervals and interval number for the tissue (**H**) and stress (**I**) lens function. The Mapper graph constructions we chose for further analysis are enclosed within a box. Raw expression data underlying the graphs in this figure can be found in S7 Dataset, and code to regenerate analyses can be found in https://zenodo.org/records/8428609 (Palande, 2023).

Overlaying the tissue lens value of each sample over the PCA and t-SNE dimensional space reveals a clear gradient across PC1 and t-SNE1, with the highest lens function values

found in seed, fruit, and flower tissues **(Fig 7.2F)**. For the stress lens function, samples are distributed across the dimensional space, with no obvious correlation between healthy and stressed lens values, similar to the observation from individual abiotic/biotic stresses **(Figs 7.1D and 2G)**.

Mapper graphs for the tissue and lens functions reflect an emergent and striking topological shape of plant expression **(Fig 7.2H and 7.2I)**. Each node in the Mapper graphs corresponds to a bin of similar RNAseq samples with color representing the average lens value of samples within each node. Edges (connections) show common samples between overlapping bins. Changing the cover interval overlap and interval number has marginal effects on the core graph structure but changes the shape and connectivity of sparse nodes on the outskirts of the graphs **(Fig 7.2H and 2I)**. This central stability highlights the robustness of our input data and significance of the underlying features defining the graph shape (Carriere and Oudot, 2018). The Mapper graphs for both the tissue and stress lens functions show a backbone structure with numerous embedded nodes and flares that form a well-defined gradient from leaf to seed or healthy to stressed, respectively. This suggests that there are distinct and conserved expression patterns across angiosperms that delineate different tissues or responses to biotic and abiotic stresses.

Our input dataset is unbalanced, with large discrepancies in the number of input samples for different species, stresses, or tissue types. We tested if biases in the distribution of samples could explain the topological shape we observed. We downsampled the most frequent factor combinations and surveyed the effect it had on the Mapper graph topology. Our study has 3 factors: family, tissue, and stress with 16 families, 8 tissue types, and 10 stresses. In total, 1,280 unique 3-way combinations are possible (family + tissue + stress), but in our dataset, only 195 unique combinations are present and they have a heavily skewed distribution **(Appendix E, Fig S7.1)**. Based on this distribution, we chose a cutoff of 30 and downsampled the 30 most common factor combinations. This significantly reduced the sampling bias for family, tissue, and stress, but it did not eliminate them **(Appendix E, Fig 7.2.B)**. We then reran the Mapper algorithm using this downsampled dataset. The topology is quite similar, suggesting that biases in sample representation are not the major factor underlying the patterns we observed **(Appendix E, Fig 7.2.C).**

7.4.3. Topological shape reflects the underlying biological features of gene expression

To identify and characterize these conserved biological patterns, we first simplified the Mapper graphs into 18 nodes for both the tissue and stress lens functions **(Figs 7.3 and 7.4)**. The core tissue-based Mapper graph has discrete nodes for each surveyed plant tissue with a gradual transition of leaves (node 1), to roots (2), fruits (11 and 13), and, finally, seeds (14, 15, and 16; **Fig 7.3A**). At the fourth node, the Mapper graph proliferates into terminal branches of flower (node 9), stem (10), fruit (12), and mixtures of uncategorized tissue types (5 and 8). RNAseq samples from the 16 angiosperm families are largely dispersed across nodes by tissue, with some notable exceptions **(Fig 7.3B)**. Most fruit samples are found along the gradient of the core graph structure, but fruits from the rose (Rosaceae) family form a separate node (node 12). Flowers from the eudicot species are mixed with fruit tissues in nodes along the core graph structure, but monocot flowers from the grass family (Poaceae) are found in discrete, branching nodes (9 and 17). The biotic and abiotic stress RNAseq samples are dispersed by tissue across the Mapper graph **(Fig 7.3C)**, supporting the complexity and heterogeneity of these samples.

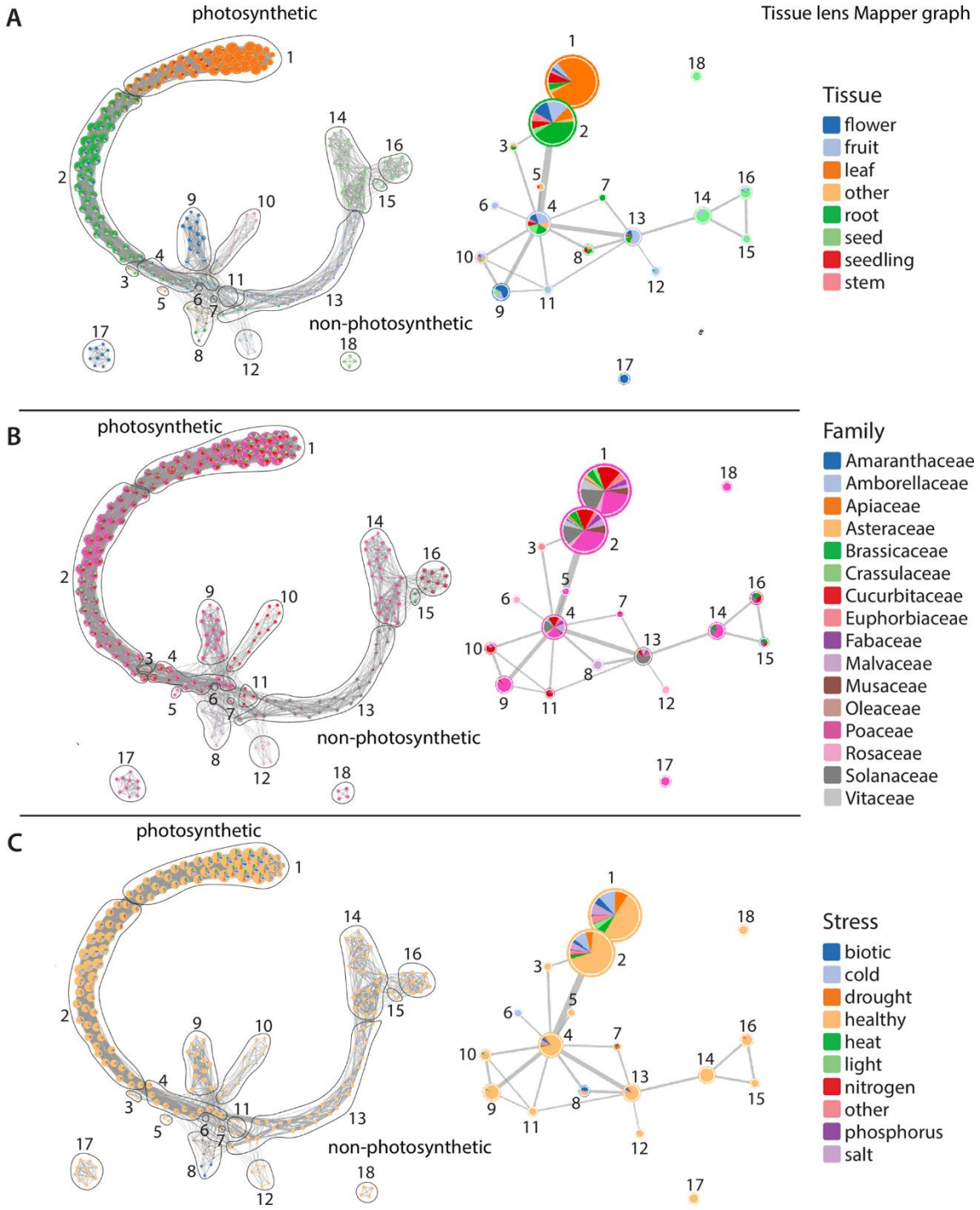


Figure 7.3: Simplified Mapper graphs detailing the distribution of samples along the tissue lens. Nodes along the full Mapper graphs (left) are clustered into simplified Mapper graphs (right), and samples are colored by tissue (A), family (B), and stress category (C). Photosynthetic and nonphotosynthetic ends of the Mapper graph are indicated.

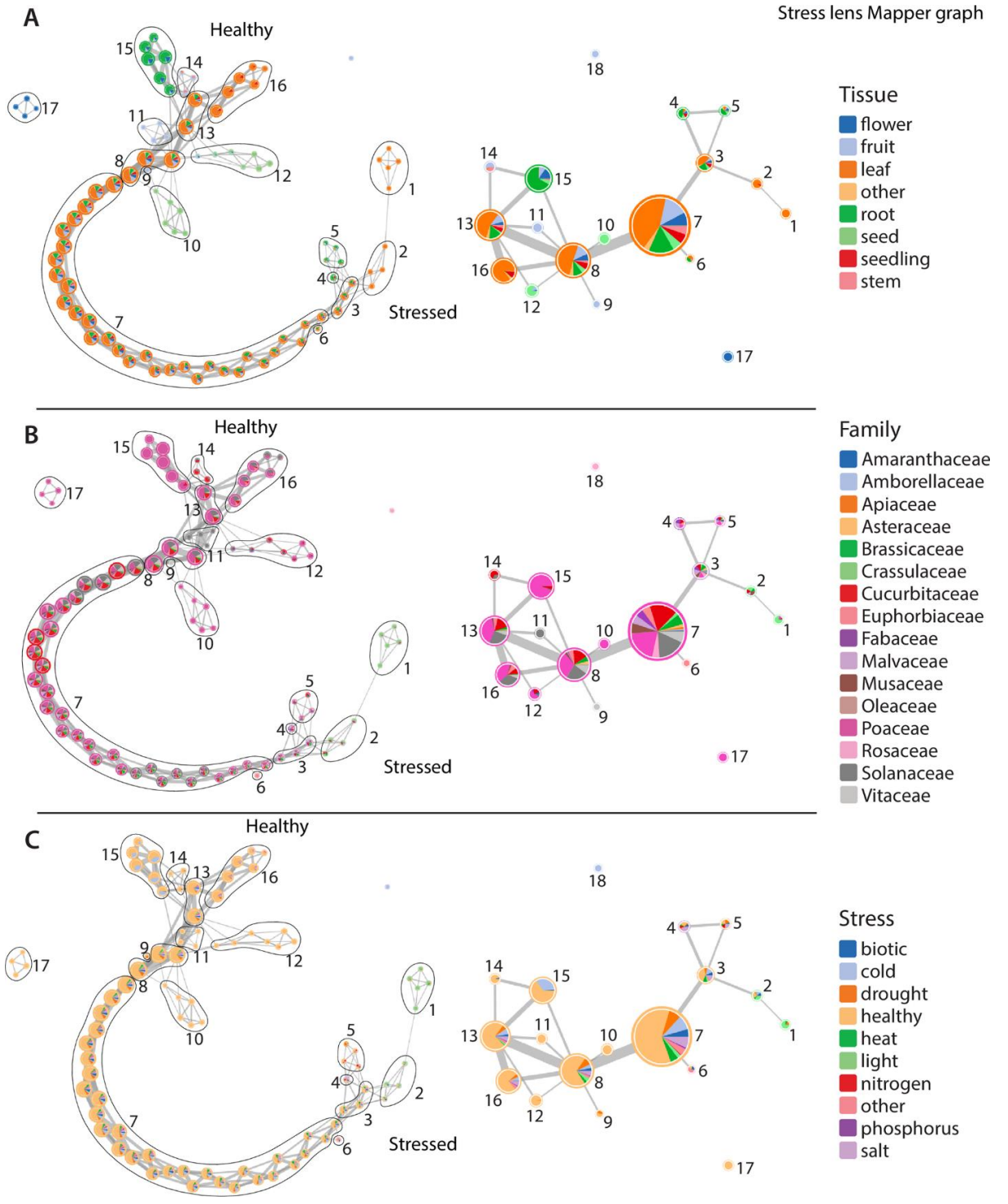


Figure 7.4: Simplified Mapper graphs detailing the distribution of samples along the stress lens. Nodes along the full Mapper graphs (left) are clustered into simplified Mapper graphs (right) and samples are colored by tissue (A), family (B), and stress category (C). Healthy and stressed ends of the Mapper graph are indicated.

Mapper graphs clearly distinguish tissues across plant taxa, but what are the biological features that underlie this topology? We surveyed the expression patterns of the 6,328 orthogroups used to generate our Mapper graphs to see if they are enriched in certain biological processes related to evolutionarily conserved, tissue-specific functions. We classified genes as positively or negatively correlated with the tissue lens and conducted GO enrichment in these groups of genes. We expect negatively correlated genes to be characteristic of leaf gene expression and positively correlated genes to be characteristic of non-leaf gene expression. Supporting this, Mapper graphs and GO terms associated with the tissue lens–correlated genes point to photosynthetic versus nonphotosynthetic metabolism as a key factor in the overall gene expression patterns of plant tissues **(Fig 7.3 and Appendix E, S1 Dataset)**. Enriched negatively correlated GO terms are mostly related to photosynthesis and include response to red and blue light, chloroplast and thylakoid organization, carotenoid metabolic process, and regulation of photosynthesis among others **(Appendix E S1 Dataset)**. Plants and green algae are characterized by a set of well-conserved genes that are not found in nonphotosynthetic organisms termed "the GreenCut2 inventory" (Karpowicz et al., 2011). Most of the GreenCut2 genes (421 out of 677) are found within the 6,328 orthogroups in our analysis, and we tested if these are enriched among correlated genes. Genes from the GreenCut2 inventory are overrepresented in this set of genes, with 26.7% of the tissue-correlated (positively or negatively) genes being in the GreenCut2 resource versus 6.7% of the entire set of orthogroups **(Appendix E, Table 7.1)**. This overrepresentation is even more stark if we delimit our analysis to only the genes negatively correlated with the tissue lens, of which 50.3% are in the GreenCut2 inventory. The overlapping loci between the 2 sets contain genes encoding protein products involved in various aspects of photosynthesis, including pigment biosynthesis and binding (e.g., AT4G10340, AT1G04620, AT1G44446) (Murray and Kohorn, 1991; Andersson et al., 2001; Meguro et al., 2011), the operation of the photosynthetic light reactions (e.g., AT4G05180, AT5G44650, AT3G17930) (Schubert et al., 2002; Albus et al., 2010; Xiao et al., 2012), or the operation of the Calvin–Benson Cycle (AT1G32060) (Harmon et al., 2001).

Enriched GO terms that are positively correlated with the tissue lens are largely related to housekeeping and core metabolic processes including ubiquitination, macromolecule catabolism, the electron transport chain, peptide biosynthesis, and Golgi vesicle–mediated transport among many others **(Appendix E, S2 Dataset)**. Enriched genes include proteins involved in the TCA

cycle and respiration (e.g., AT1G47420, AT2G18450, AT4G26910) (Kruft et al., 2001; Millar et al., 2001; Menges et al., 2002) and in the development of specific nonphotosynthetic tissue types like seeds (e.g., AT2G40170, AT2G38560) (Leon-Kloosterziel et al., 1996; Wang et al., 2008) and pollen/pollen tubes (e.g., AT2G03120, AT2G41630) (Han et al., 2009; Zhou et al., 2013). However, many of the tissue lens–correlated genes do not intuitively relate to the photosynthetic versus nonphotosynthetic tissue distinction, and further examination of these loci on a gene-by-gene basis may shed light on conserved differences between plant tissues.

The simplified Mapper graph from the stress lens has 18 nodes that form a continuous gradation of healthy to stressed tissues **(Fig 7.4)**. Individual tissue types, regardless of stress condition, are enriched in certain nodes but are less defined than under the tissue lens **(Fig 7.4A)**. RNAseq samples related to light and heat stress are found in discrete nodes (1 and 2, respectively) at the terminus of the Mapper graph across all species where these data were available **(Fig 7.4C)**. Other stress RNAseq samples are found in nodes with healthy tissues but are generally concentrated toward the stress end of the Mapper graph. An interesting exception is a group of cold stressed root samples from the grass (Poaceae) family (node 15). Clustering of distinct stresses within the same node suggests a core stress response conserved across Angiosperms for all abiotic and biotic factors. The gradient of sample distribution from healthy to stressed across the Mapper graph may be related to the severity of stress experienced by plants in each individual experiment.

To explore what constitutes these conserved stress-related expression patterns, we searched for GO enrichment of genes that are positively correlated with the stress lens. This group of genes is heavily enriched in functions related to stress, including responses to water deprivation, chitin, reactive oxygen species, fungi, wounding, bacteria, and general defense mechanisms **(Appendix E, S3 Dataset)**. Genes positively correlated with the stress lens include loci related to the biosynthesis of compounds with diverse stress-related activities like jasmonic acid and jasmonic acid derivatives (AT2G35690, AT2G46370) (Staswick and Tiryaki, 2004; Schilmiller et al., 2007) and ascorbic acid (AT3G09940) (Lisenbee et al., 2005). Negatively correlated genes are enriched in functions related to growth and reproduction such as DNA replication, mitosis, and rRNA processing, among others **(Appendix E, S4 Dataset)**. This includes genes involved in regulation of the cell cycle (AT3G54650, AT4G12620, AT2G01120) (Collinge et al., 2004; Masuda et al., 2004; Kim et al., 2008), chromatin organization

(AT1G15660, AT1G65470) (Kaya et al., 2001; Ogura et al., 2004), and the development of reproductive structures (AT1G34350, AT2G41670, AT4G27640, AT3G52940) (Broadhvest et al., 2000; Dou et al., 2016; Huang et al., 2017; Liu et al., 2019). This pattern points towards an intuitive distinction between the stressed and unstressed samples in our dataset in terms of their investment in cell proliferation and reproduction. Most of these genes are involved in core biological functions with conserved roles across eukaryotes, and their coordinated perturbation could be predictive of stress responses in diverse lineages.

7.5. Discussion

Genome-scale datasets have high dimensionality, and even the simplest pairwise experiment has hundreds or thousands of complex and interconnected cellular pathways in dynamic flux between conditions. Comparisons across plant lineages are similarly complex, as each species has its own evolutionary history with thousands of duplicated, lost, or new genes enabling its unique and elegant biology. This complexity presents major challenges for characterizing underlying biological mechanisms and identifying shared and distinct properties across evolutionary timescales. Here, we leveraged the wealth of public gene expression datasets across diverse flowering plants and used a set of deeply conserved genes to search for patterns of conservation across tissue types, stress responses, and evolution. We first tested traditional dimensionality reduction and clustering-based approaches but found that they were largely ineffective and unable to clearly resolve samples. Instead, we used a novel topological framework to compare samples and test for evolutionary conservation.

Topological data analysis has been applied to complex, high dimensionality biological datasets including gene expression profiles correlated with human cancers and other diseases (Nicolau et al., 2011; Mandal et al., 2020; Rabadán et al., 2020). To our knowledge, TDA has not been used for plant science datasets outside of shape (Li et al., 2018; Zeng et al., 2021; Amézquita et al., 2022). Flowering plants have tremendous phylogenetic, developmental, phenotypic, and genomic scale diversity, creating additional layers of complexity compared to other lineages. Despite this, Mapper was able to capture hidden and emergent signatures of gene expression at the tissue and stress scales that were missed using traditional approaches. Most developmental tissues or stress responses are not perfectly separated but instead fall within a gradient along a central shape. The central shape of the tissue lens Mapper graph represents the life cycle of a plant with transitions from the vegetative tissues of leaves and roots to

reproductive flowers, fruit, and, eventually, seeds. Nodes along the Mapper graphs that contain mixtures of tissues such as fruits and flowers, leaves and stems, or even leaves and roots reflect developmental plasticity, heterogeneity, and overlapping functions between different organs. Flowers give rise to fruits and the complex processes of fertilization, seed, and fruit development blur the lines between distinct tissue types. This complexity and interconnectivity is central to biological processes but is masked by traditional dimensionality reduction approaches, which can oversimplify nonlinear datasets.

The stressed and healthy samples are less clearly delineated in the Mapper graphs than samples from different plant tissues. This may reflect artifacts stemming from variation in the severity, duration, or method of applying stresses across different experiments and species. For example, mildly stressed samples might have expression signatures that mirror healthy tissues with comparatively few differentially expressed genes. Despite this issue, we observed a strong gradient of sample distribution from healthy to stressed across the graph. Distinct stresses were generally found within the same nodes, and genes that were positively correlated with the stress lens show enrichment in classical stress pathways. This includes the core stress-responsive hormones jasmonic acid and abscisic acid and their corresponding transcriptional network as well as broader shifts in metabolic processes geared toward defense. Taken together, this suggests that plants have deeply conserved expression signatures across evolution and for different stresses. Abiotic and biotic stress responses have been mostly studied in isolation, but they typically co-occur in natural environments, and they have overlapping signaling, hormonal, and network responses in plants (reviewed in Rejeb et al. (2014)]. The topological shape of gene expression points to a shared set of pathways or perturbations that define if a tissue is healthy or stressed. Environmental stresses broadly disrupt photosynthesis and core metabolic and cellular functions either as a direct response to physical trauma or in preparation for defense or resilience. These changes may serve as the backbone of the topological shape we observed for the stress lens.

Although we observed a deeply conserved pattern of gene expression underlying plant form and function, our analyses capture a snapshot of the evolutionary innovations found in flowering plants. We used a set of low-copy, conserved genes to enable comparisons of expression across species, and we had to exclude around approximately 70% of all plant genes. This includes most enzymes, transcription factors, and regulatory elements, which are mostly found in large, rapidly

evolving, or lineage-specific gene families that cannot be resolved to high-confidence orthologs across eudicots and monocots. Duplication and subsequent sub- or neofunctionalization of these genes drive the evolution of new plant traits and developmental differences of plant organs. Single-copy genes by contrast have deeply conserved functions in core metabolism, photosynthesis, and housekeeping processes that typically transcend tissue, species, and environmental changes. Given these limitations, it is somewhat surprising that our analyses were able to clearly separate tissue types and stresses despite missing information from most of the genes that should underlie these biological differences. Applying TDA with a full set of genes in a single species with well-curated gene expression profiles could uncover complex or emergent biological signatures that were previously hidden.

Here, we provide a proof of concept for studying complex biological traits using TDA, and a similar analytical framework could be applied to numerous areas of plant science research and beyond. Compared to the approximately 300,000 published plant gene expression datasets (Lim et al., 2022), our study has a somewhat sparse sampling of species and a subset of expressed genes, yet we were able to detect a number of hidden trends. TDA of high-resolution sampling over narrower phenotypic spaces such as drought responses in a single species or tissue divergence across 900 million years of plant evolution could yield transformative insights that were previously overlooked. However, researchers should exercise caution when applying TDA to gene expression data as the lack of a robust hyperparameter tuning procedure could potentially result in misleading conclusions. This reflects a broader problem in machine learning and data science, but hyperparameter search, cross-validation, and feature selection can enable data-driven tuning of the appropriate hyperparameters. With the appropriate datasets and sufficient sampling, TDA can be widely applicable for developing a deeper understanding of complex, emergent biological phenomena.

7.6. Methods

7.6.1. Assembling a representative catalog of flowering plant expression data

We selected species that captured the broadest phylogenetic diversity within angiosperms and species that had a breadth of expression at the tissue and stress levels. We also selected only species with a high-quality reference genome to enable accurate read mapping and downstream comparative genomics. Metadata including species, accession, tissue type, experimental treatments, replicate number, and sequencing platform were collected manually for each sample

using the NCBI BioProject and SRAs, as well as the primary data publications **(Appendix E, S6 Dataset)**. Raw RNAseq reads were downloaded from the NCBI SRA and quantified using a pipeline developed in the VanBuren lab to trim, quantify, and identify differentially expressed genes (https://github.com/pardojer23/RNAseqV2). Using a common analytical pipeline helped reduce noise between experiments that used different algorithms in the original publications. Raw Illumina reads from various platforms were first quality trimmed using fastp (v0.23) (Chen et al., 2018) with default parameters. The quality filtered reads were pseudoaligned to the corresponding transcripts (gene models) for each species using Salmon (v1.6.0) (Patro et al., 2017) with the quasi-mapping mode. Transcript-level estimates were converted to gene-level transcript per million counts using the R package tximport (Soneson et al., 2015).

7.6.2. Comparing expression across species

To facilitate detailed cross-species comparisons, we first clustered proteins from all 54 species into orthogroups using Orthofinder (v2.3.8) (Emms and Kelly, 2015). Genomes and proteomes were downloaded for each species from Phytozome v13 (Goodstein et al., 2012). Orthofinder was run using default parameters and the reciprocal DIAMOND search (v2.0.11) (Buchfink et al., 2021) was used for sequence alignment, and groups of similar proteins were clustered using the Markov Cluster Algorithm. In total, 2,317,289 genes (94% of input genes) were clustered into 86,185 orthogroups across the 54 species. Of these, 33,585 orthogroups are found in only a single species and 7,742 are found in at least 52 out of 54 species. This set of broadly conserved orthogroups was further refined by filtering out orthogroups with an average of >2 genes per ortholog for the diploid species to avoid including multigene families with diverse functions in the analysis. This set of 6,335 orthogroups was used as a common framework to allow comparison of expression across species. For orthogroups where a species had more than one gene, the total TPM for all genes in that orthogroup was summed and the raw TPM was used for single-copy genes. Expression data for each sample across all species were combined into a single expression matrix **(Appendix E, S7 Dataset)**, and SVA was used to characterize the potential impacts of unmodeled technical variables on the dataset **(see Appendix E, Text 7.A)**. PCA was performed using built-in functions in Scikit-learn (Pedregosa et al., 2011) on the log2+1 or z-score transformed gene expression data (raw TPMs) to reduce dimensionality and capture the main sources of variation within the datasets.

7.6.3. Surrogate variable analysis

To account for batch effects and the influence of unmodeled factors on the expression matrix used for the present study, we applied SVA to generate estimates of surrogate variables and their effects on our expression matrices (Leek and Storey, 2007; Leek et al., 2012). Briefly, SVA assumes that the expression of a particular gene *i* across *j* independent RNA-seq experiments can be described by the following linear equation:

$$x_{ij} = u_i + f_i(y_j) + e_{ij} \quad (\mathbf{1})$$

where u_i is the baseline expression level of gene *i*, $f_i(y_j)$ represents the effect of a measured variable y_j, and e_{ij} is the error term (Leek and Storey, 2007). However, if there are a number of *L* unmodeled factors affecting the expression of gene *i*, then the error term e_{ij} contains both randomly distributed experimental error as well as the effects of unmodeled factors. That is:

$$e_{ij} = \sum_{l}^{L} y_{li} g_{ij} + e'_{ij} \quad (\mathbf{2})$$

where $g_l = (g_l = (g_{l1},\ldots,g_{ln})$ is a function describing the effect of all unmodeled factors up to *L*, y_{li} is the coefficient describing the influence of an unmodeled factor *l* on the expression of gene *i*, and e'_{ij} is the true randomly distributed noise term (Leek and Storey, 2007). Combining (1) and (2) yields:

$$x_{ij} = u_i + f_i(y_i) + \sum_{l}^{L} y_{li} g_{ij} + e'_{ij} \quad (\mathbf{3})$$

By using the svaseq() method implemented in the R package sva (v. 3.36.0) (Leek et al., 2012; Leek, 2014), we identified and estimated the values of 24 separate surrogate variables. These surrogate variables, which correspond to vectors of values for each expression value x_{ji}, in the $\sum_l^L y_{li} g_{ij} + e'_{ij}$ term in (3).

To determine the amount of variation due to a proxy batch variable (bioproject), 3 biological primary variables (stress, tissue, and family), and the pairwise interactions each surrogate variable explains, we regressed all the estimated surrogate variables on each variable (either batch or biological) or on a pairwise interaction. McNemar's formula was used to calculate the adjusted R2 values for each surrogate variable.

7.6.4. Mathematical basis of topological data analysis

The flexibility of Mapper allows us to apply it to various types of data. Here, we will describe the Mapper construction in the simplest setting of point cloud data and then explain how it was applied to the gene expression data.

Consider a point cloud $\boldsymbol{X} \subset \mathbf{R}^d$ equipped with a function f: $\boldsymbol{X} \rightarrow \mathbf{R}$. An open cover of $\boldsymbol{X}$ is a collection $\boldsymbol{U} = \{\boldsymbol{U}_i\}_{i \in \mathbf{I}}$ of open sets in $\mathbf{R}^d$, such that $\boldsymbol{X} \subset \bigcup_{i \in \mathbf{I}} \boldsymbol{U}_i$, where $\mathbf{I}$ is an index set. The 1-dimensional nerve of the cover $\boldsymbol{U}$, denoted as $\boldsymbol{M} := \boldsymbol{N}_1(\boldsymbol{U})$, is called the Mapper graph of ($\boldsymbol{X}$, f). In this graph, each open set $\boldsymbol{U}_i$ is represented as a vertex $\boldsymbol{i}$, and 2 vertices, $\boldsymbol{i}$ and $\boldsymbol{j}$, are connected by an edge if and only if the intersection of $\boldsymbol{U}_i$ and $\mathbf{U}_j$ is nonempty.

To construct a Mapper graph, we start by defining a cover $\boldsymbol{V} = \{\boldsymbol{V}_j\}_{j \in \mathbf{J}}$ of the image $f(\boldsymbol{X}) \subset \mathbf{R}$ of f, where $\mathbf{J}$ is a finite index set, by splitting the range of $f(\boldsymbol{X})$ into a collection of overlapping intervals. Next, for each $\boldsymbol{V}_j$, we identify the subset of points $\boldsymbol{X}_j$ in $\boldsymbol{X}$ such that $f(\boldsymbol{X}_j) \subset \boldsymbol{V}_j$ and apply a clustering algorithm to identify clusters of points in $\boldsymbol{X}_j$. The cover $\boldsymbol{U}$ of $\boldsymbol{X}$ is the collection of such clusters induced by $f^{-1}(\boldsymbol{V}_j)$ for each j. Once we have the cover $\boldsymbol{U}$, we compute its 1-dimensional nerve $\boldsymbol{M}$ and visualize it in the form of a weighted graph.

For example, consider Fig **7.2A–2E**. The point cloud $\boldsymbol{X}$ in this case consists of points in the 2-dimensional plane, in the shape of a "Y". The function f simply maps each point to its y-coordinate. We divide the range of f into 4 overlapping intervals, represented by the 4 colored segments along the y-axis in **Fig 7.2**. For each interval $\boldsymbol{V}_j$, the colored rectangles in the center panel of the figure show the subsets of points $\boldsymbol{X}_j \in \boldsymbol{X}$ such that $\boldsymbol{X}_j = f^{-1}(\boldsymbol{V}_j)$. Then, we apply clustering to each $\boldsymbol{X}_j$ separately to obtain the cover $\boldsymbol{U}$ of $\boldsymbol{X}$. The 1-dimensional nerve of $\boldsymbol{U}$, i.e., the mapper graph $\boldsymbol{M}$, is shown in the rightmost panel. The color of each vertex corresponds to the cover interval it belongs to. **Fig 7.2A–2E** illustrates mapper graph construction from the same set of points, but with x-coordinate used as the lens. We can observe that the 2 lens functions produce 2 slightly different mapper graphs.

7.6.5. Constructing Mapper graphs and lens functions

To construct Mapper graphs from our gene expression data, we create 2 different lenses, adopting an approach similar to the one used in Nicolau and colleagues' paper (Nicolau et al., 2011). We refer to these lenses as the tissue lens and the stress lens, respectively. To create the stress lens, we first identified all the healthy samples from the dataset and fit a linear model to them. This model serves as the idealized healthy orthogroup expression. Then, we project all the

samples (healthy as well as stressed) onto this linear model and obtain the residuals. These residuals measure the deviation of the sample gene expression from the modeled healthy expression. The lens function is simply the length of the residual vector. To define the cover, we divide the range of the lens function into intervals of uniform length, with the same amount of overlap between adjacent intervals. We experimented with a range of values length of the intervals and the size of the overlap to identify the values that produced relatively stable Mapper graphs. The clustering was performed using DBSCAN, a commonly used clustering algorithm for Mapper.

The construction of Mapper graph relies on several user-defined parameters: the lens function f, the cover $\boldsymbol{V}$, and the clustering algorithm. Optimizing these parameters is an interesting open problem in TDA research (Chalapathi et al., 2021). The function f plays the role of a lens, through which we look at the data, and different lenses provide different insights (Singh et al., 2007). The choice of f is typically driven by the domain knowledge and the data under consideration. In this study, the data under consideration are very similar to the dataset studied by Nicolau and colleagues (Nicolau et al., 2011). Therefore, we followed similar methods to define the lenses. Our choice of lenses is further justified by the observations from the dimension reduction plots.

The cover $\boldsymbol{V} = \{\boldsymbol{V}_j\}_{j \in J}$ of $f(\boldsymbol{X})$ consists of a finite number of open intervals as cover elements. To define $\boldsymbol{V}$, we use the simple strategy of defining intervals of uniform length and overlap. Adjusting the interval length and the overlap increases or decreases the amount of aggregation provided by the Mapper graph. The optimal choice was made by visually inspecting Mapper graphs over a range of parameter values. The parameters resulting in the most stable structure were selected. Any clustering algorithm can be employed to obtain the cover $\boldsymbol{U}$. We use the density-based clustering algorithm, DBSCAN (Ester et al., 1996), which is commonly used in Mapper because it does not require a priori knowledge of the number of clusters. Instead, DBSCAN requires 2 input parameters: the number of samples in a neighborhood for a point to be considered as a core point, and the maximum distance between 2 samples for one to be considered in the neighborhood of the other.

7.6.6. Functional annotation of orthogroups

The correlation between expression values and tissue lens and stress lens values was calculated for each orthogroup. The top 2.5% most positively and negatively correlated orthogroups for

each lens were selected to represent the tissue lens or stress lens correlated orthogroups. *Arabidopsis* gene IDs were used to identify the overlap between the GreenCut2 (Karpowicz et al., 2011) inventory with *Arabidopsis* orthologs in our overall set of orthogroups, as well as our sets of tissue lens and stress lens correlated orthogroups. The *binom_test()* function from SciPy (Virtanen et al., 2020) was used to apply one-sided binomial tests to check for enrichment of GreenCut2 loci in the overall, tissue lens, and stress lens correlated orthogroup sets. GO term enrichment of the sets of genes mapped to orthogroups and correlated with the tissue lens or stress lens was done using GOATOOLS (Klopfenstein et al., 2018). Data on gene function and biochemical reactions associated with specific loci were derived from TAIR (Lamesch et al., 2012), KEGG (Kanehisa and Goto, 2000), and a genome-scale metabolic model of *Arabidopsis* metabolism from (de Oliveira Dal'Molin et al., 2015).

7.7. Data availability statement

The code, metadata, raw datasets from this project are available on a dedicated GitHub page: https://github.com/PlantsAndPython/plant-evo-mapper and Zenodo: https://zenodo.org/records/8428609.

7.8. Funding

This work was funded primarily by National Science Foundation Research Traineeship training grant (NSF 1828149 to ATM, DHC, and RV) which established the Integrated training Model in Plant And CompuTational Sciences (IMPACTS) program at Michigan State University. This grant funded fellows within this program (JAMK, MDR, KSA, CC, JD, RD, TBJ, HRJ, AM, EMR, AMS, JY) as well as the project-based curriculum for the Plants and Python Course that formed the backbone of this manuscript. This work is also supported by NSF Plant Genome Research Program awards IOS-2310355 to EM, DHC, and RV, IOS-2310356 to AH, and IOS-2310357 to AK, NSF Developmental Mechanisms award IOS-2039489 to AH, and NSF Biological Integration Institute award (DBI-2213983 to RV). Several students (JAMK, MDR, KSA, HMP, JP) were supported by predoctoral training award (T32-GM110523 to RV) from the National Institute of General Medical Sciences of the NIH. This project was supported by the USDA National Institute of Food and Agriculture, and by Michigan State University AgBioResearch to AMT, DHC, and RV. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

7.9. Author contributions

Conceptualization: Sourabh Palande, Joshua A. M. Kaste, Miles D. Roberts, Kenia Segura Aba´, Jamell Dacon, Aman Y. Husbands, Arjun Krishnan, Beronda L Montgomery, Elizabeth Munch, Addie M. Thompson, Alejandra Rougon-Cardoso, Daniel H. Chitwood, Robert VanBuren.

Data curation: Sourabh Palande, Joshua A. M. Kaste, Miles D. Roberts, Kenia Segura Aba´, Carly Claucherty, Jamell Dacon, Rei Doko, Thilani B. Jayakody, Hannah R. Jeffery, Nathan Kelly, Andriana Manousidaki, Hannah M. Parks, Emily M. Roggenkamp, Ally M. Schumacher, Jiaxin Yang, Sarah Percival, Jeremy Pardo, Alejandra Rougon-Cardoso, Daniel H. Chitwood, Robert VanBuren.

Formal analysis: Sourabh Palande, Joshua A. M. Kaste, Miles D. Roberts, Kenia Segura Aba´, Carly Claucherty, Rei Doko, Thilani B. Jayakody, Hannah R. Jeffery, Nathan Kelly, Andriana Manousidaki, Hannah M. Parks, Emily M. Roggenkamp, Ally M. Schumacher, Jiaxin Yang, Sarah Percival, Jeremy Pardo, Alejandra Rougon-Cardoso, Daniel H. Chitwood, Robert VanBuren.

Project administration: Alejandra Rougon-Cardoso, Daniel H. Chitwood, Robert VanBuren.

Software: Jeremy Pardo.

Supervision: Daniel H. Chitwood, Robert VanBuren.

Visualization: Daniel H. Chitwood.

Writing – original draft: Joshua A. M. Kaste, Alejandra Rougon-Cardoso, Daniel H. Chitwood, Robert VanBuren.

Writing – review & editing: Aman Y. Husbands, Arjun Krishnan, Beronda L Montgomery, Elizabeth Munch, Addie M. Thompson, Alejandra Rougon-Cardoso, Daniel H. Chitwood, Robert VanBuren.

REFERENCES

Albus CA, Ruf S, Schöttler MA, Lein W, Kehr J, Bock R (2010) Y3IP1, a nucleus-encoded thylakoid protein, cooperates with the plastid-encoded Ycf3 protein in photosystem I assembly of tobacco And Arabidopsis. Plant Cell **22**: 2838–2855

Amézquita EJ, Quigley MY, Ophelders T, Landis JB, Koenig D, Munch E, Chitwood DH (2022) Measuring hidden phenotype: quantifying the shape of barley seeds using the Euler characteristic transform. in silico Plants **4**: diab033

Andersson J, Walters RG, Horton P, Jansson S (2001) Antisense Inhibition of the Photosynthetic Antenna Proteins CP29 and CP26: Implications for the Mechanism of Protective Energy Dissipation. The Plant Cell **13**: 1193–1204

Azodi CB, Pardo J, VanBuren R, de los Campos G, Shiu S-H (2020) Transcriptome-Based Prediction of Complex Traits in Maize[OPEN]. The Plant Cell **32**: 139–151

Broadhvest J, Baker SC, Gasser CS (2000) SHORT INTEGUMENTS 2 Promotes Growth During Arabidopsis Reproductive Development. Genetics **155**: 899–907

Buchfink B, Reuter K, Drost H-G (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. Nature Methods **18**: 366–368

Carriere M, Oudot S (2018) Structure and stability of the one-dimensional mapper. Foundations of Computational Mathematics **18**: 1333–1396

Chalapathi N, Zhou Y, Wang B (2021) Adaptive covers for mapper graphs using information criteria. 2021 IEEE International Conference on Big Data (Big Data). IEEE, pp 3789–3800

Chen S, Zhou Y, Chen Y, Gu J (2018) Fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics **34**: i884–i890

Collinge MA, Spillane C, Köhler C, Gheyselinck J, Grossniklaus U (2004) Genetic interaction of an origin recognition complex subunit and the Polycomb group gene MEDEA during seed development. Plant Cell **16**: 1035–1046

Dou XY, Yang KZ, Ma ZX, Chen LQ, Zhang XQ, Bai JR, Ye D (2016) AtTMEM18 plays important roles in pollen tube and vegetative growth in Arabidopsis. Journal of integrative plant biology **58**: 679–692

Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology **16**: 157

Ester M, Kriegel H-P, Sander J, Xu X, others (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. kdd. pp 226–231

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res **40**: D1178-1186

Han S, Green L, Schnell DJ (2009) The Signal Peptide Peptidase Is Required for Pollen Function in Arabidopsis. Plant Physiology **149**: 1289–1301

Harmon AC, Gribskov M, Gubrium E, Harper JF (2001) The CDPK superfamily of protein kinases. New Phytologist **151**: 175–183

Huang B, Qian P, Gao N, Shen J, Hou S (2017) Fackel interacts with gibberellic acid signaling and vernalization to mediate flowering in Arabidopsis. Planta **245**: 939–950

Julca I, Ferrari C, Flores-Tornero M, Proost S, Lindner A-C, Hackenberg D, Steinbachová L, Michaelidis C, Gomes Pereira S, Misra CS, et al (2021) Comparative transcriptomic analysis reveals conserved programmes underpinning organogenesis and reproduction in land plants. Nature Plants **7**: 1143–1159

Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research **28**: 27–30

Karpowicz SJ, Prochnik SE, Grossman AR, Merchant SS (2011) The GreenCut2 resource, a phylogenomically derived inventory of proteins specific to the plant lineage. Journal of Biological Chemistry **286**: 21427–21439

Kaya H, Shibahara K ichi, Taoka K ichiro, Iwabuchi M, Stillman B, Araki T (2001) FASCIATA genes for chromatin assembly factor-1 in Arabidopsis maintain the cellular organization of apical meristems. Cell **104**: 131–142

Kim HJ, Oh SA, Brownfield L, Hong SH, Ryu H, Hwang I, Twell D, Nam HG (2008) Control of plant germline proliferation by SCFFBL17 degradation of cell cycle inhibitors. Nature **455**: 1134–1137

Klopfenstein DV, Zhang L, Pedersen BS, Ramírez F, Warwick Vesztrocy A, Naldi A, Mungall CJ, Yunes JM, Botvinnik O, Weigel M, et al (2018) GOATOOLS: A Python library for Gene Ontology analyses. Scientific Reports **8**: 10872

Kruft V, Eubel H, Jänsch L, Werhahn W, Braun HP (2001) Proteomic approach to identify novel mitochondrial proteins in Arabidopsis. Plant Physiology **127**: 1694–1710

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al (2012) The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. Nucleic Acids Research **40**: 1202–1210

Leek JT (2014) Svaseq: Removing batch effects and other unwanted noise from sequencing data. Nucleic Acids Research **42**: e161

Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD (2012) The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics **28**: 882–883

Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genetics **3**: 1724–1735

Leon-Kloosterziel KM, van de Bunt GA, Zeevaart JAD, Koornneef M (1996) Arabidopsis Mutants with a Reduced Seed Dormancy. Plant Physiology **110**: 233–240

Li M, An H, Angelovici R, Bagaza C, Batushansky A, Clark L, Coneva V, Donoghue MJ, Edwards E, Fajardo D, et al (2018) Topological Data Analysis as a Morphometric Method: Using Persistent Homology to Demarcate a Leaf Morphospace. Frontiers in Plant Science 9:

Lim PK, Zheng X, Goh JC, Mutwil M (2022) Exploiting plant transcriptomic databases: Resources, tools, and approaches. Plant Communications **3**: 100323

Lisenbee CS, Lingard MJ, Trelease RN (2005) Arabidopsis peroxisomes possess functionally redundant membrane and matrix isoforms of monodehydroascorbate reductase. Plant J **43**: 900–914

Liu HH, Xiong F, Duan CY, Wu YN, Zhang Y, Li S (2019) Importinβ4 mediates nuclear import of grf-interacting factors to control ovule development in arabidopsis. Plant Physiology **179**: 1080–1092

Mandal S, Guzmán-Sáenz A, Haiminen N, Basu S, Parida L (2020) A Topological Data Analysis Approach on Predicting Phenotypes from Gene Expression Data. *In* C Martín-Vide, MA Vega-Rodríguez, T Wheeler, eds, Algorithms for Computational Biology. Springer International Publishing, Cham, pp 178–187

Masuda HP, Ramos GBA, De Almeida-Engler J, Cabral LM, Coqueiro VM, Macrini CMT, Ferreira PCG, Hemerly AS (2004) Genome based identification and analysis of the pre-replicative complex of Arabidopsis thaliana. FEBS Letters **574**: 192–202

Meguro M, Ito H, Takabayashi A, Tanaka R, Tanaka A (2011) Identification of the 7-hydroxymethyl chlorophyll a reductase of the chlorophyll cycle in arabidopsis. Plant Cell **23**: 3442–3453

Menges M, Hennig L, Gruissem W, Murray JAH (2002) Cell cycle-regulated gene expression in Arabidopsis. Journal of Biological Chemistry **277**: 41987–42002

Millar AH, Sweetlove LJ, Giegé P, Leaver CJ (2001) Analysis of the Arabidopsis Mitochondrial Proteome. Plant Physiology **127**: 1711–1727

Murray DL, Kohorn BD (1991) Chloroplasts of Arabidopsis thaliana homozygous for the ch-1 locus lack chlorophyll b, lack stable LHCPII and have stacked thylakoids. Plant Molecular Biology **16**: 71–79

Nicolau M, Levine AJ, Carlsson G (2011) Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proceedings of the National Academy of Sciences **108**: 7265–7270

Ogura Y, Shibata F, Sato H, Murata M (2004) Characterization of a CENP-C homolog in Arabidopsis thaliana. Genes and Genetic Systems **79**: 139–144

de Oliveira Dal'Molin CG, Quek LE, Saa PA, Nielsen LK (2015) A multi-tissue genome-scale metabolic modeling framework for the analysis of whole plant systems. Frontiers in Plant Science **6**: 1–12

Palande S (2023) PlantsAndPython/plant-evo-mapper: plant-evo-mapper-first-release.

Palande S, Kaste JAM, Roberts MD, Segura Abá K, Claucherty C, Dacon J, Doko R, Jayakody TB, Jeffery HR, Kelly N, et al (2023) Topological data analysis reveals a core gene expression backbone that defines form and function across flowering plants. PLOS Biology **21**: e3002397

Pathak S, Agarwal A, Ankita A, Gurve MK (2021) Restricted Randomness DBSCAN: A faster DBSCAN Algorithm. 2021 Thirteenth International Conference on Contemporary Computing (IC3-2021). pp 7–12

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nature methods **14**: 417–419

Pearson K (1901) LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2**: 559–572

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al (2011) Scikit-learn: Machine learning in Python. the Journal of machine Learning research **12**: 2825–2830

Proost S, Mutwil M (2018) CoNekT: an open-source framework for comparative genomic and transcriptomic network analyses. Nucleic Acids Research **46**: W133–W140

Rabadán R, Mohamedi Y, Rubin U, Chu T, Alghalith AN, Elliott O, Arnés L, Cal S, Obaya ÁJ, Levine AJ, et al (2020) Identification of relevant genetic alterations in cancer using topological data analysis. Nature Communications **11**: 3808

Rejeb IB, Pastor V, Mauch-Mani B (2014) Plant Responses to Simultaneous Biotic and Abiotic Stress: Molecular Mechanisms. Plants (Basel) **3**: 458–475

Rizvi AH, Camara PG, Kandror EK, Roberts TJ, Schieren I, Maniatis T, Rabadan R (2017) Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. Nature Biotechnology **35**: 551–560

Schilmiller AL, Koo AJK, Howe GA (2007) Functional diversification of acyl-coenzyme A

oxidases in jasmonic acid biosynthesis and action. Plant Physiology **143**: 812–824

Schubert M, Petersson UA, Haas BJ, Funk C, Schröder WP, Kieselbach T (2002) Proteome map of the chloroplast lumen of Arabidopsis thaliana. Journal of Biological Chemistry **277**: 8354–8365

Singh G, Memoli F, Carlsson G (2007) Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. Eurographics Symposium on Point-Based Graphics. doi: 10.2312/SPBG/SPBG07/091-100

Soneson C, Love MI, Robinson MD (2015) Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res **4**: 1521

Staswick PE, Tiryaki I (2004) The oxylipin signal jasmonic acid is activated by an enzyme that conjugate it to isoleucine in Arabidopsis W inside box sign. Plant Cell **16**: 2117–2127

Tauzin G, Lupo U, Tunstall L, Pérez JB, Caorsi M, Medina-Mardones AM, Dassatti A, Hess K (2021) giotto-tda: A topological data analysis toolkit for machine learning and data exploration. The Journal of Machine Learning Research **22**: 1834–1839

Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. Journal of machine learning research 9:

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods **17**: 261–272

Wang C, Wang H, Zhang J, Chen S (2008) A seed-specific AP2-domain transcription factor from soybean plays a certain role in regulation of seed germination. Science in China Series C: Life Sciences **51**: 336–345

Washburn JD, Mejia-Guerra MK, Ramstein G, Kremling KA, Valluru R, Buckler ES, Wang H (2019) Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. Proceedings of the National Academy of Sciences **116**: 5542–5549

Xiao J, Li J, Ouyang M, Yun T, He B, Ji D, Ma J, Chi W, Lu C, Zhang L (2012) DAC is involved in the accumulation of the cytochrome b6/f complex in Arabidopsis. Plant Physiology **160**: 1911–1922

Zeng D, Li M, Jiang N, Ju Y, Schreiber H, Chambers E, Letscher D, Ju T, Topp CN (2021) TopoRoot: a method for computing hierarchy and fine-grained traits of maize roots from 3D imaging. Plant Methods **17**: 127

Zhang H, Zhang F, Yu Y, Feng L, Jia J, Liu B, Li B, Guo H, Zhai J (2020) A Comprehensive Online Database for Exploring ~20,000 Public Arabidopsis RNA-Seq Libraries. Mol Plant **13**: 1231–1233

Zhou J-J, Liang Y, Niu Q-K, Chen L-Q, Zhang X-Q, Ye D (2013) The Arabidopsis general transcription factor TFIIB1 (AtTFIIB1) is required for pollen tube growth and endosperm development. Journal of Experimental Botany **64**: 2205–2218

APPENDIX E: Supplemental Material for Chapter 7

Text S7.A: Confounder discussion from the Surrogate Variable Analysis

We used Surrogate Variable Analysis (SVA) (Leek et al., 2012) to explore the effects of confounding technical variables on the publicly available SRA data assembled for this study. Briefly, we identified three primary variables of interest (tissue, stress, and family), which were fixed in the model used to estimate "surrogate variables" to minimize the amount of variability attributable to these primary variables captured by the estimated surrogate variables. These surrogate variables represent unaccounted for technical variables impacting the dataset. Due to the breadth of families, stresses, and tissues analyzed, we do not have a full factorial design (i.e., there are combinations of family, stress, and tissue factor values for which there are no expression datasets). Because of this, SVA would remove variability due to our primary variables and their interactions. To get a sense of what kind of impact the surrogate variables might have on the dataset when removed, we estimated the correlation between the first order interactions between our primary variables and the surrogate variables identified by SVA. We identified 24 surrogate variables which individually captured between 53% and 98% of variation between BioProjects **(Fig S7.4)**. We also estimated the interaction terms between the tissue, family, and stress factor combinations that were present in the dataset and estimated how much of their variation was getting captured by the surrogate variables. Individual surrogate variables captured up to 14% of variation between stress conditions, up to 66% of variation between tissue conditions, and up to 63% of variation between families. For the interaction terms between primary variables, individual surrogate variables captured up to 83% of the variation between tissue and family combinations, up to 65% of the variation between stress and family combinations, and up to 71% of the variation between tissue and stress combinations. This suggests that even though stress, tissue, and family are treated as protected primary variables, there are underlying latent variables related to our primary variables and their interactions that may be important sources of biological variation being captured by the surrogate variables. Although individual surrogate variables could be selectively accounted for in downstream analyses in such a way that minimizes the removal of biological signal, this would be a highly subjective process. Moreover, due to our inability to precisely calculate the true correlation

between our surrogate variables and interaction terms due to the fact that many factor combinations are missing, this would be statistically dubious as well.

Because the surrogate variables show substantial linear correlation with our primary variables and their interaction terms, the application of SVA would require eliminating substantial amounts of biological signal. Since the goal of our study is to identify heterogeneous patterns due to stress, tissue, and family within a high-dimensional gene expression dataset, SVA may not be appropriate for us to use. Alternatively, one could potentially minimize the loss of this signal by cherry-picking individual surrogate variables to include in downstream analysis, which would naturally introduce human bias. A third option would be to use an algorithm like ComBat-seq (Zhang et al., 2020b) that relies on explicitly defined batches, which is problematic for the present study since the closest metadata for batch available for the studies gathered on SRA is the BioProject ID's, but these are, at best, a proxy for batches of samples and are not sufficient to assess the technical variability or noise in the data. More broadly, as discussed in (Jaffe et al., 2015), such genomic data "cleaning" methods, by their very nature, delimit the observable features of the resulting datasets to those prespecified by the investigator. In our view, this limits their utility for broad exploratory analyses of the kind described in this study. For all the above reasons, we opted to not use SVA, ComBat, or related techniques.

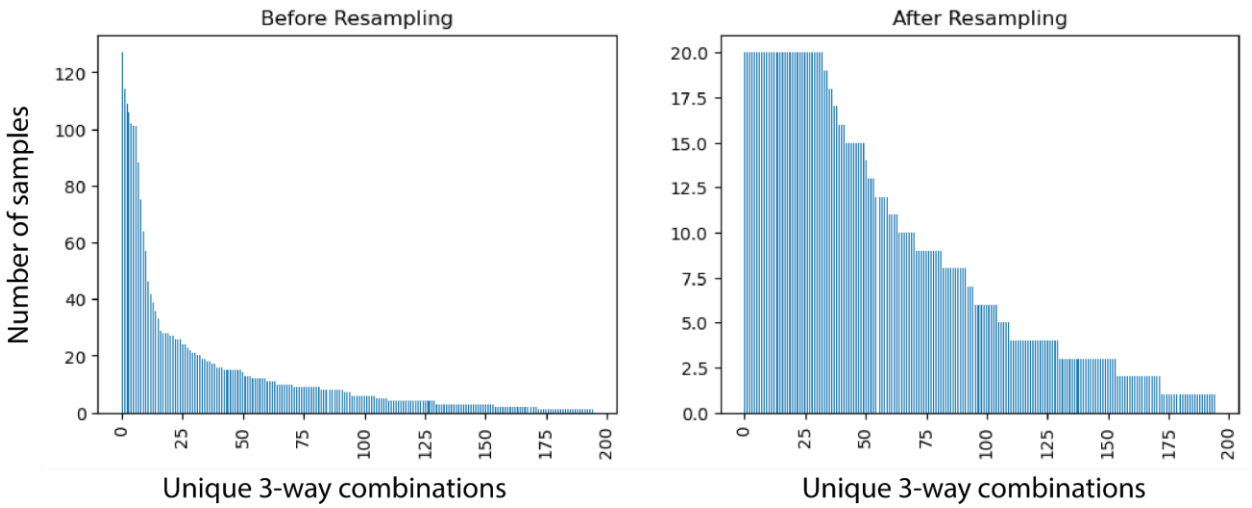


Figure S7.1: Histogram of 3-way factors of the RNA seq samples before and after downsampling. The distribution of 3-way factors for family, tissue, and stress are plotted. The 16 families, 8 tissue types and 10 stresses equate to 1280 unique 3-way combinations, but we only observed 195 unique combinations in our dataset. The distribution of samples from the entire dataset is shown on the left and the distribution of samples when downsampling the 30 most common 3-way combinations is shown on the right.

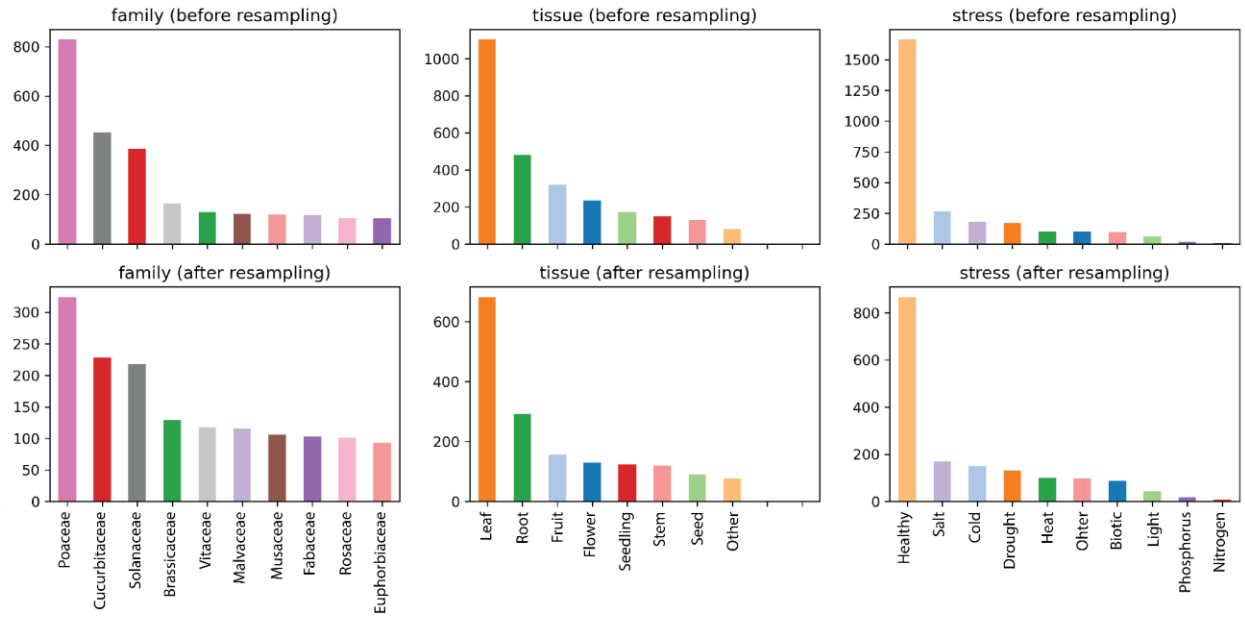


Figure S7.2: Factor-wise frequency plots of RNAseq samples before and after subsampling. The number of samples in each family, tissue type, or stress are plotted before (top) and after (bottom) subsampling.

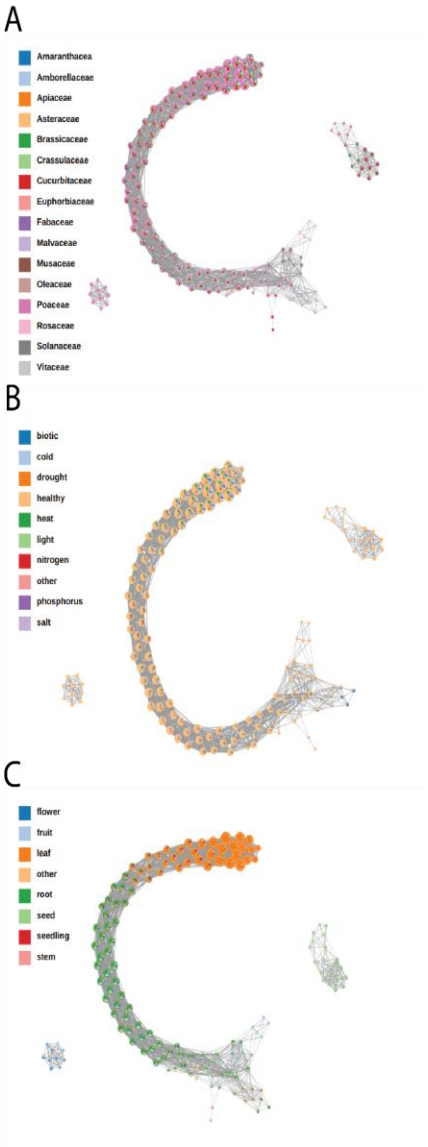


Figure S7.3: Topology of Mapper graphs generated from the subsampled data. Samples from each node in the mapper graph are colored by plant family (A), stress (B), or tissue type (C), using the subsampled data. The overall topology and sample distribution are similar to the Mapper graphs constructed with the full, unbalanced dataset, suggesting sample distribution is not a major factor in our analyses.

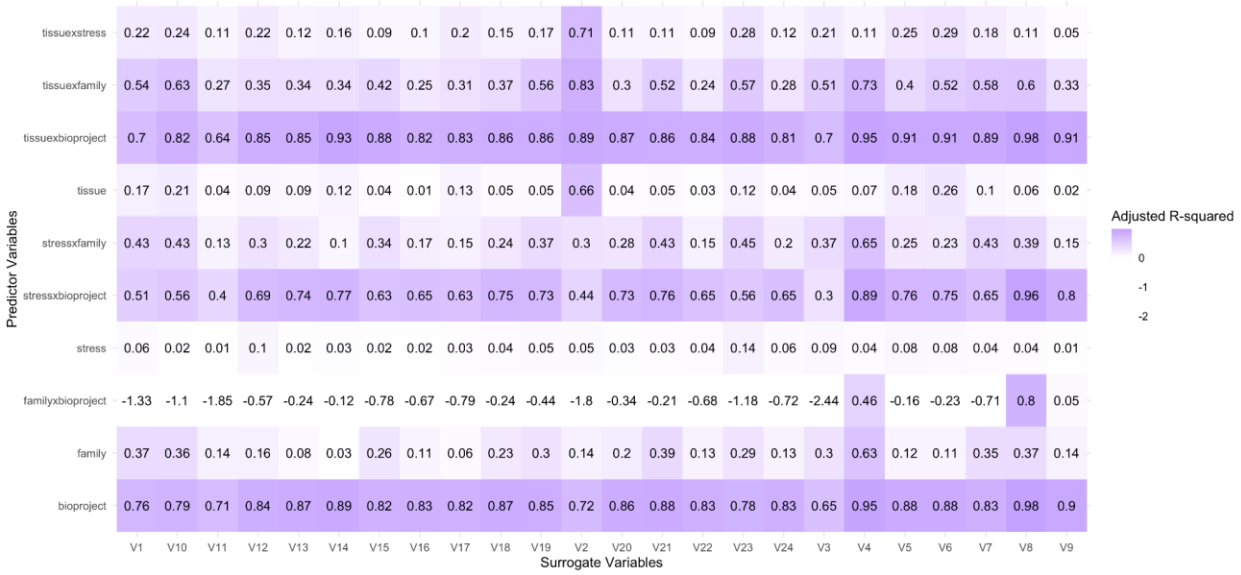


Figure S7.4: Linear regression analysis of association of surrogate variables to one batch variable (bioproject), our biological variables of interest (stress, tissue, family), and their pairwise interactions. All surrogate variables were regressed on either each variable or interaction individually to calculate adjusted R^2 values.

TABLES

Table S7.1: Enrichment of GreenCut2 genes in orthogroup-mapped *Arabidopsis thaliana* genes and stress-/tissue- correlated orthogroup-mapped genes. The proportion of GreenCut2 genes in the all the orthogroups used in this study was compared against the proportion of GreenCut2 genes in a list of all A. thaliana genes using a one-sided binomial test. The proportion of tissue-lens and stress-lens correlated orthogroup-mapped genes in GreenCut2 was compared against the proportion of GreenCut2 genes in the entire set of orthogroup-mapped genes using one-sided binomial tests. Tissue-correlated genes were hypothesized to be more likely to be in GreenCut2 than a random selection of orthogroup-mapped genes, and the stress-correlated genes were hypothesized to be less likely.

Dataset	# of Genes in Dataset	# of Genes in GreenCut2	% GreenCut2	p-value
All Arabidopsis Genes	**27662**	**677**	**2.45**	
All Orthogroup-Mapped Genes	**6328**	**421**	**6.65**	**$2.76 * 10^{-96}$**
All Tissue-lens Correlated Genes	**318**	**85**	**26.7**	**$9.18 * 10^{-29}$**
Stress-lens Correlated Genes	**318**	**7**	**2.20**	**0.000252**

Dataset Descriptions

All supplemental datasets can be found at the following link:
https://doi.org/10.1371/journal.pbio.3002397

S1 Dataset. GO Term enrichment results on genes negatively correlated with the tissue lens (XLSX)

S2 Dataset. GO term enrichment results on genes positively correlated with the tissue lens (XLSX)

S3 Dataset. GO term enrichment results on genes positively correlated with the stress lens (XLSX)

S4 Dataset. GO term enrichment results on genes positively correlated with the stress lens (XLSX)

S5 Dataset. Overlap between orthogroup-mapped genes and tissue lens and stress lens correlated genes with the GreenCut2 resource (Karpowicz) (XLSX)

S6 Dataset. Metadata of the raw data used in this experiment (CSV)

S7 Dataset. Expression matrix of TPMs for the normalized orthogroups (CSV)

REFERENCES

Jaffe AE, Hyde T, Kleinman J, Weinbergern DR, Chenoweth JG, McKay RD, Leek JT, Colantuoni C (2015) Practical impacts of genomic data "cleaning" on biological discovery using surrogate variable analysis. BMC Bioinformatics **16**: 1–10

Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD (2012) The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics **28**: 882–883

Zhang Y, Parmigiani G, Johnson WE (2020) ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genomics and Bioinformatics. doi: 10.1093/nargab/lqaa078